



Jimma University

Jimma Institute of Technology

Faculty of Computing and Informatics

Attention-based Amharic-to-Afaan Oromo Neural Machine Translation

By: Meron Gashaw

**A Thesis Submitted to Faculty of Computing and Informatics,
Institute of Technology, Jimma University, in Partial Fulfillment for Degree
of Master of Science in Information Technology**

Jimma, Ethiopia

January 2021


JIMMA UNIVERSITY
JIMMA INSTITUTE OF TECHNOLOGY
FACULTY OF COMPUTING AND INFORMATICS

Attention-based Amharic-to-Afaan Oromo Neural Machine Translation

By: Meron Gashaw Asfaw

This is to certify that the thesis prepared by Meron Gashaw titled: Attention-based Amharic-to-Afaan Oromo Neural Machine Translation and submitted in partial fulfillment of the requirements for the degree of masters of Science in information technology complies with the regulation of the university and meets the accepted standards concerning originality and quality.

Signed by

| | Name | Signature | Date |
|--------------------------|------------------------------|--|---------------------|
| Principal Advisor | Yaregal Assabie (PhD) |  | January 2021 |
| Co-Advisor | Mizanu Zelalem(MSc) | | January 2021 |

JIMMA, ETHIOPIA

January 2021

Declaration


I, the undersigned, declare that this thesis entitled “Attention-based Amharic- Afaan Oromo Neural Machine Translation” is my original work and has not been presented for a degree in this or any other universities, and all sources of references used for the thesis work have been appropriately acknowledged.

Name: Meron Gashaw Asfaw

Signature: _____ **Date:** _____

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

Advisor Name: Yaregal Assabie (PhD.)

Signature: _____  _____ **Date:** _____

This thesis has been submitted for examination with my approval as a University Co-advisor.

Co-Advisor Name: Mizanu Zelalem (MSc.)

Signature: _____ **Date:** _____

Acknowledgment

First and foremost, I would like to thank God for helping me achieve this. I would also like to express my deep gratitude to my principal advisor Dr. Yaregal Assabie and my co-advisor Mr. Mizanu Zelalem for their sincere support, patience, and meaningful guidance when doing this work. Next I prefer to say thank you to workineh Wogaso who courses me while conducting this thesis and At last, not least, my family, your support, and prayer made all of this possible; thank you.

Dedication

This thesis work is dedicated to my husband, Nahil, who has been a constant source of support and encouragement during the challenges of school and life. I am truly thankful for having you in my life. This work is also dedicated to my parents, Asres and Gashaw, who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve.

Abstract

Humans use a natural language (NL) in order to convey meanings from one entity or group to another. This NL should be mutually understandable by both communicating entities, which is not imaginary living in a world having a population of more than 7 billion. So, there should be a translator between the two, of which most of the time is a human. But, human translator is expensive and inconvenient. The emergence of Natural Language Processing (NLP) and Machine Translation (MT) has solved this issue. MT is an automatic translation of a source language to a target language. Allowing the use of neural network models to learn a statistical model for MT, Neural Machine Translation (NMT), aims at building a single neural network that can be jointly tuned to maximize translation performance. In this work, an attention based Amharic to Afaan Oromo NMT system has been developed. The system is developed based on Encoder-Decoder model by using a Bi-directional Gated Recurrent Unit (BGRU). In order to compare the performance of the system, we have also implemented our system before applying attention mechanism. The non-attention based system with basic encoder-decoder architecture, have some limitations. As the length of the sentence increase the interdependency of words will loosely increase. This shows that the non-attention based architecture works well with shorter sentences but highly suffer to translate longer sentences. Moreover, as each word in the sentence is visited, it must be assigned a new identity number in order to identify a word by a unique index at the time it encountered it in the data. But when the length of the dictionary increases, the dimension of word vector needed becomes higher. We have observed that these problems have been solved by applying an attention mechanism to the system. Prior to this work there was no other NMT system translating Amharic to Afaan Oromo. So, in this work attention based Amharic to Afaan Oromo neural machine translation has been developed and its performance was compared to that of non-attention based Amharic to Afaan Oromo neural machine translation system.

For the evaluation purpose, BLEU score evaluation was used. We have recorded a BLEU score of 61.49 for the non-attention based system and 67.82 for the attention based Amharic to Afaan Oromo Neural Machine Translation.

Keywords: Natural Language Processing, Machine Translation, Neural Machine Translation, Recurrent Neural Network, Local Attention

TABLE OF CONTENTS

| | |
|---|----|
| CHAPTER 1: INTRODUCTION..... | 1 |
| 1.1 Background..... | 1 |
| 1.2 Motivation | 3 |
| 1.3 Statement of the Problem | 4 |
| 1.4 Objectives..... | 6 |
| 1.5 Scope and Limitations..... | 6 |
| 1.6 Methods..... | 6 |
| 1.7 Application of Results..... | 7 |
| 1.8 Organization of the Thesis | 8 |
| CHAPTER 2: LITERATURE REVIEW | 9 |
| 2.1 Introduction | 9 |
| 2.2 Overview of Amharic Language..... | 9 |
| 2.2.1 Amharic Punctuations..... | 10 |
| 2.2.2 Amharic Morphology..... | 11 |
| 2.2.3 Amharic Phrases | 16 |
| 2.2.4 Amharic Sentences..... | 18 |
| 2.3 A Brief Overview of Afaan Oromo | 21 |
| 2.3.1 Afaan Oromo Punctuation Marks | 22 |
| 2.3.2 Afaan Oromo Morphology..... | 23 |
| 2.3.3 Afaan Oromo Phrases | 33 |
| 2.3.4 Afaan Oromo Sentences..... | 34 |
| 2.4 Machine Translation | 35 |
| 2.4.1 History of Machine Translation..... | 37 |
| 2.4.2 Approaches to Machine Translation | 39 |
| 2.4.3 Evaluation of Machine Translation..... | 46 |
| 2.5 System and Language Modeling in NMT | 48 |
| 2.5.1 System Modeling (Encoder-Decoder) | 48 |
| 2.5.2 Language Modeling | 50 |
| CHAPTER 3: RELATED WORKS..... | 52 |
| 3.1 Introduction | 52 |

| | | |
|---|--|-----|
| 3.2 | Machine Translation for foreign language pairs..... | 52 |
| 3.3 | Machine Translation involving Ethiopian language | 56 |
| CHAPTER 4: DESIGN OF ATTENTION BASED AMHARIC TO AFAN OROMO NMT | | 64 |
| 4.1 | Introduction | 64 |
| 4.2 | System Design | 64 |
| 4.2.1 | Language model training..... | 64 |
| 4.2.2 | Language model testing | 66 |
| 4.3 | System design Components..... | 68 |
| 4.3.1 | Data Pre-processing | 68 |
| 4.3.2 | Indexing..... | 69 |
| 4.3.3 | Word Embedding..... | 70 |
| 4.3.4 | Padding..... | 71 |
| 4.3.5 | Encoder | 71 |
| 4.3.6 | Attention..... | 74 |
| 4.3.7 | Decoder | 76 |
| CHAPTER 5: Experimentation and Discussion | | 77 |
| 5.1 | Introduction | 77 |
| 5.2 | Dataset collection and preparation..... | 77 |
| 5.3 | System environment/ tools used for the development | 78 |
| 5.4 | Parameter optimization and training the experimental systems | 78 |
| 5.5 | Experimental results..... | 82 |
| 5.6 | Discussion on the result of the study | 82 |
| CHAPTER 6: Conclusion and Future Work | | 85 |
| 6.1 | Introduction | 85 |
| 6.2 | Conclusion..... | 85 |
| 6.3 | Future work | 86 |
| References | | 87 |
| Appendix I: Sample of parallel corpus..... | | 97 |
| Appendix II: The last epoch results with loss level and a time to taken for 163 batches. | | 100 |
| Appendix III: Sample output | | 104 |

List of Tables

| | |
|---|----|
| Table 2.1 Gender, number, and case marker suffixes | 12 |
| Table 2.2 Definiteness marker suffixes | 12 |
| Table 2.3 verbal markers for person, gender, and time | 13 |
| Table 2.4 Passive voice markers | 14 |
| Table 2.5 Afaan Oromo Writing Scripts | 22 |
| Table 2.6 The most common plural morphs | 24 |
| Table 2.7 Plural form of kinship terms | 25 |
| Table 2.8 Gender inflection of noun | 26 |
| Table 2.9 different types of case markers | 27 |
| Table 2.10 Afaan Oromo Active and Passive Voices | 29 |
| Table 2.11 Number Inflection in Afaan Oromo | 30 |
| Table 2.12 marked gender on adjectives | 31 |
| Table 2.13 Adjectival Case markers | 31 |
| Table 2.14 Example of Derivational Noun | 32 |
| Table 4.1 Vocabulary Formation and Sentence Tokenization | 69 |
| Table 5.1 Example Translation performance for long sentence | 83 |

List of Figures

| | |
|---|----|
| Figure 2.1 Amharic Letters..... | 10 |
| Figure 2.2 The Vauquouis triangle for rule-based MT systems [12]..... | 41 |
| Figure 2.3 Design of Encoder-Decoder language modeling [85]..... | 45 |
| Figure 2.4 System Modelling | 50 |
| Figure 4.1 Language Model Training | 65 |
| Figure 4.2 Language Model Testing | 66 |
| Figure 4.3 Amharic (Left) and Afaan Oromo (Right) Words Sample Indexing | 70 |
| Figure 4.4 Internal Structure of GRU | 72 |
| Figure 4.5 Bidirectional GRU Network Structure | 73 |
| Figure 4.6 Global (Left) and Local (Right) Attention Model [121]. | 75 |
| Figure 4.7 Encoder-Decoder model with local attention | 75 |
| Figure 5.1 Number of Epoches vs Loss Level..... | 80 |
| Figure 5.2 Loss level for each batch size and embedding dimension with respect to epochs..... | 80 |
| Figure 5.3 Learning rate vs Loss level | 81 |
| Figure 5.4 Learning rate vs Time taken to train | 81 |
| Figure 5.5 Number of words in sentence vs BLEU Score | 84 |

List of Acronyms

| | |
|---------|---|
| BGRU | Bidirectional Gated Recurrent Unit |
| BLEU | Bi-Lingual Evaluation Understudy |
| BPE | Byte Pair Encoding |
| CBOW | Continuous Bag of Words |
| CNN | Convolutional Neural Network |
| CPU | Central Processing Unit |
| DNN | Deep Neural Network |
| EBMT | Example Based Machine Translation |
| EoS | End of Sentence |
| FDRE | Federal Democratic Republic of Ethiopia |
| GPU | Graphics Processing Unit |
| GRU | Gated Recurrent Unit |
| HMT | Hybrid Machine Translation |
| LM | Language Model |
| LSTM | Long Short Term Memory |
| MT | Machine Translation |
| NL | Natural Language |
| NLP | Natural Language Processing |
| NMT | Neural Machine Translation |
| PoS | Part of Speech |
| RBMT | Rule Based Machine Translation |
| RNN | Recurrent Neural Network |
| Seq2Seq | Sequence to Sequence Learning |
| SMT | Statistical Machine Translation |
| TM | Translation Model |
| WER | Word Error Rate |
| WMT | Workshop on Statistical Machine Translation |
| BGRU | Bidirectional Gated Recurrent Unit |

CHAPTER 1: INTRODUCTION

1.1 Background

Humans use a natural language (NL) in order to convey meanings from one entity or group to another. This NL should be mutually understandable by both communicating entities, which is not imaginary living in a world having a population of more than 7 billion. So, there should be a translator between the two, of which most of the time is a human. But, human translator is expensive and inconvenient. The advancement of technology and the rise of the internet as a means of communication led to an ever-increasing demand for Natural Language Processing [1]. Natural Language Processing (NLP), also called computational linguistics is widely regarded as a promising and critically important endeavor in the field of computer science research [2]. The general goal for most computational linguists is to let the computer have the ability to understand and generate natural language so that eventually people can address their computers through text and speech as though they were addressing another person [2]. NLP is the field of Artificial Intelligence that gives the machines the ability to read, understand, and derive meaning from human languages [3]. NLP applications are useful in facilitating human-human, human-computer, computer-human, and computer-computer communication via computing systems. It is also used to analyze text, allowing machines to understand how humans speak. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts of speech tagging, relationship extraction, stemming, text mining, machine translation (MT), and automated question answering [4].

MT is one of the oldest sub-disciplines of computer science to study the natural language and has significantly contributed to the development of fields such as computational linguistics and artificial intelligence [5]. MT is concerned with automatic translations of natural language texts, in contrast to e.g. machine-aided human translation. It is considered to be the most substantial way in which machines could actually communicate to humans and vice versa [6]. To process any translation, human or automated, the meaning of a text in the source language must be fully restored in the target language, i.e. the translation. While on the surface this seems straightforward, it is far more complex [7]. Translation is not a mere word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know how each word

may influence another [7]. This requires extensive expertise in grammar, syntax, semantics, etc., in the source and target languages, as well as familiarity with each local region in which syntax and semantic mean sentence structure and meanings respectively. In recent years with the development of the internet, the need for MT has greatly increased [8]. Search engines, data mining, social media, and education are application areas where MT can provide substantial value. In a more than ever-connected world where the main form of communication is a natural language, the demand for translation is only rising. Manual translation of a text can both be very time consuming and expensive, providing another reason to improve automatic MT and evaluation [9]. In addition to these, issues related to confidentiality are another matter which makes MT favorable. Giving sensitive data to a professional translator might be risky. Many people use MT systems to translate their sensitive data because no one would agree to give a private correspondence to a translator who is not known, or no one would entrust documents to other people. Moreover, for the sake of universality, a machine translator usually translates a text which is in any language whereas a professional translator specializes in one particular field. Online translation and translation of web page content is also a favorable advantage of a machine translator, in that these services are at hand and one can translate information quickly. Furthermore, you can translate any web page content and query of a search engine by the use of MT systems [10, 11].

Several approaches are there to be used in MT, such as; Rule-Based MT (RBMT), Statistical MT (SMT), Hybrid MT (HMT), and Neural MT (NMT) [23].

RBMT is also known as the Knowledge-Based MT or Classical Approach of MT. It is a general term that denotes MT systems based on linguistic information about the source and target languages basically retrieved from (bilingual) dictionaries and a collection of rules called grammar rules covering the main semantic, morphological, and syntactic regularities of each language respectively [13]. In this approach, human experts specify a set of rules to describe the translation process, so that an enormous amount of input from human experts is required [13]. It consists of a bilingual or multilingual lexicon, and software programs to process the rules. The rules play a major role in various stages of translation such as syntactic processing, semantic interpretation, and contextual processing of language. On the other hand, SMT is a method for translating text from one natural language to another based on the knowledge and statistical models extracted from bilingual corpora. A supervised or unsupervised statistical machine

learning algorithm is used to build statistical tables from the corpora. This process is called learning or training. The statistical tables consist of statistical information such as the characteristics of well-formed sentences and the correlation between the languages. In SMT, the core process (transfer) includes a translation model that takes as input source language words or word sequences (phrases) and produces target language words or word sequences as an output. The other approach which is, Hybrid MT approach is developed taking advantage of both statistical and rule-based translation methodologies.

NMT is a recently proposed framework for MT-based purely on neural networks. Neural networks are making in-roads into the MT industry, providing major advances in translation quality over the existing industry-standard SMT technology. Because of how the technology functions, neural networks better capture the context of full sentences before translating them, providing much higher quality and more human-sounding output [14]. This technique has begun to show promising results when compared to other approaches [11], [15], [16], [17], [18]. NMT being an end-to-end trained model meets the approach of modeling the whole MT process via a single artificial neural network [19]. Unlike the traditional phrase-based translation system, NMT attempts to build and train a single large neural network. The neural network reads a sentence and outputs a correct translation [11]. Its goal is to design a fully trainable model of which every component is tuned based on training corpora to maximize its translation performance. People have been turning their heads towards NMT systems, which after being introduced seriously in 2014 have seen many refinements. However, the language model of NMTs Encoder-Decoder is generally unable to accurately process long input sequences since only the last hidden state of the Encoder RNN is used as the context vector for the Decoder. The Attention Mechanism directly addresses this issue as it retains and utilizes all the hidden states of the input sequence during the decoding process.

1.2 Motivation

Afaan Oromo is one of the languages of the lowland East Cushitic within the Cushitic family of the Afro-Asiatic Phylum [20, 21]. It is also one of the major languages spoken in Ethiopia. According to [22], Afaan Oromo is the sixth most widely spoken language in Africa preceded by Swahili, Arabic French, Hausa, and Yoruba. The Oromo language also referred to as Afaan Oromo or Oromiffaa has more than 30 million speakers across Africa. It is spoken mainly in the

Horn of Africa, Ethiopia, Kenya and Somalia. It is also a medium of instruction and a school subject in primary and secondary schools in the Oromiya regional state, one of the administrative regions that have the largest population of all the regions of the Federal Democratic Republic of Ethiopia (FDRE) [23].

Amharic is one of the languages in the Semitic family which is widely spoken in Ethiopia [24]. Which makes it the second-largest spoken language in Ethiopia, next to Afaan Oromo, the second most-spoken Semitic language in the world (after Arabic), and one of the eight largest languages on the African continent [22]. Amharic is an official working language of the Federal Democratic Republic of Ethiopia [24].

Even though Afaan Oromo speakers are large in number, because Amharic is the official working language of the FDRE, most of the official documents, letters, newspapers, and vacancies are written and produced in Amharic [24]. Speakers of Afaan Oromo language who are unable to speak and understand Amharic cannot communicate and interact with Amharic speakers and documents in an easy way without finding translators. Thus, non-Amharic speakers of Afaan Oromo language speakers face problems of lack of information. The constitution [24] recommends it is better if every regional official document is translated and documented in Amharic language parallel with the local language. Thus, it is a good contribution if there is a way to translate federal written Amharic documents and literal news into Afaan Oromo languages which can prevent non-Amharic speakers of Afaan Oromo speakers from lack of information and face unwanted expenses, such as time and cost. Amharic-to-Afaan Oromo MT can solve the aforementioned problems. This has motivated us to work on attention-based Amharic-to-Afaan Oromo NMT.

1.3 Statement of the Problem

MT systems have been developed by using different methodologies and approaches for pairs of foreign languages [12], [25]. Since English language is the most dominant language in the world, most of these foreign language translations have been done between English and other foreign language pairs. These include Arabic-to-English Neural MT [26], English-Japanese MT [27], and French to English Statistical MT system [28], etc. In contrast to this, there is relatively little MT system among English and Ethiopian languages. Some of the studies are carried out on

English-Amharic language pair [29], [30], and English-Afaan Oromo language pair [23], [31]. Some of the MTs done between Ethiopian languages are Amharic-to-Tigrigna MT using a hybrid approach [32], Bi-directional Ge'ez-Amharic MT [13].

Since Afaan Oromo language is used as a means of communication in different government and non-government institutions and serving as a working language at the regional level, it benefits non-Amharic speakers, speaking Afaan Oromo, people if documents, news, and articles are automatically translated into Afaan Oromo.

To the best of the researcher's knowledge, there has been one research that is done to translate Amharic document to Afaan Oromo by Gelan Tulu [36], using a hybrid approach. As it is in [123], Hybrid approach has its share of drawbacks, the greatest of which is the need for extensive editing. Human translators will be required. Because of the increased implementation of NMT within various translation apps and services, the technology is able to translate text from grammatically complex languages, by learning and utilizing their specifics. Unlike the traditional MT systems, using tedious steps like preparation of language modeling, preparation of translation modeling, tuning and decoding steps encoder-decoder based machine translation became the better choice for the simplicity of the steps of modeling. In encoder-decoder modeling, the steps are interconnected. Therefore, human intervention at training time is not necessary like that of statistical approach. Moreover, NMT is considered more effective in handling word ordering, morphology, and syntax. The technology comes with the promise of cost-effective translation for under-resourced languages, which makes it beneficial for all businesses, regardless of their location or spoken language.

As the researcher's knowledge is concerned, there is no prior study conducted on the development of the Amharic-to-Afaan Oromo Neural MT system. We found that NMT is also a very important NLP task that has to be done for Afaan Oromo. Thus we propose attention-based Amharic-to-Afaan Oromo Neural MT system, believing the attention mechanism can even take this advantage of NMT to a higher level.

This study was attempted to answer the research question:

- What are state of the art methods to implement machine translation which is in neural network approaches to overcome linguistic barriers and to address the knowledge among Amharic language and Afaan Oromo language speakers and users?

- Dose the enhancement mechanisms like attention can improve the performance of the translation for longer sentences between the two morphologically reach languages?

1.4 Objectives

General objective

The general objective of this study is to design and develop an Attention-based Amharic-to-Afaan Oromo Neural MT system.

Specific objectives

The specific objectives are:

- ✓ Exploring and understanding related systems and literature.
- ✓ Developing parallel bilingual corpus for Amharic and Afaan Oromo languages.
- ✓ Designing a general architecture for Attention-based Amharic-to-Afaan Oromo Neural MT.
- ✓ Develop prototype.
- ✓ Testing and evaluating the performance of the system.

1.5 Scope and Limitations

Attention-based Amharic to Afaan Oromo Neural MT is designed to perform Neural MT of texts written in Amharic text to Afaan Oromo. We have collected and prepared a dataset from legal sources in addition to the one collected by Solomon Teferra et.al. [1], which is from religious domain. But we believe that, if it was not for the time constraint to prepare other datasets, these datasets are still not adequate.

1.6 Methods

Literature Review

To find up-to-date methodologies in the MT domain, a thorough literature review will be conducted. For this study, secondary data sources, like books, articles, publications, and other

resources related to the topic will be reviewed. This helps to have a better understanding of the subject of the study. Studies related to this study will be compiled to know the pros and cons of various NMT techniques. MT systems in different languages will be studied with respect to the closeness and difference among the languages. The details of the approaches and algorithms followed to build the translation system will be reviewed. The linguistic behavior of Amharic and Afaan Oromo languages will also be investigated and identified.

Data collection

To conduct NMT, a parallel corpus of source and the target language is required. The translation system we proposed tries to generate translations using the Amharic-Afaan Oromo corpus, based on neural network methods. The sources for both languages are a holy bible and legal document specifically Magalata Oromia.

Prototype development

In order to develop a prototype for NMT, some approaches and techniques are needed. Word alignment, reordering, and language modeling can be performed with the help of a well-trained deep neural network. Word2vec generates the word-vectors that are used by recurrent auto-encoder in reconstruction task. RNN has the capability to implement reordering rules on sentences.

Evaluation mechanism

NMT system can be evaluated either using a human (manual) or automatic evaluation methods. Manual evaluation is time-consuming and expensive to perform, BLEU score will be used to evaluate the performance of the system, which is an automatic evaluation technique.

1.7 Application of Results

After its completion, the results of this research work can be applied and can be used in different areas. The system can be used for translating different texts, letters, historical and cultural books, different online resources, and also teaching and learning materials from Amharic to Afaan Oromo. This study can be used to simplify the barrier of language difficulty among language users. It enables to access information and interaction easily and fills the communication gap between peoples using the two languages; moreover, the study can be used as a component for other NLP applications such as speech translation.

1.8 Organization of the Thesis

The rest of the thesis is organized as follows. In chapter two a literature review which includes an overview of both the source and target languages and MT approaches especially NMT is discussed. Chapter Three presents different related works in the MT domain. The design of the attention-based Amharic to Afaan Oromo NMT system is presented in chapter four. The experiments and results are discussed in Chapter Five while Chapter Six will take the thesis to an end by presenting our conclusions, and recommendations regarding the work.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In this chapter, a brief overview of Amharic and Afaan Oromo, and MT in general, and Neural MT are discussed in detail. Additionally, the advancement of Neural MT over other MT approaches: Statistical MT (SMT), Rule-Based MT (RBMT), Example-Based MT (EBMT), and Hybrid MT (HMT) are described in detail.

2.2 Overview of Amharic Language

Amharic is one of the languages in the Semitic family which is widely spoken in Ethiopia [24]. It is the second-largest spoken language in Ethiopia, next to Afaan Oromo, the second most-spoken Semitic language in the world (after Arabic), and one of the eight largest languages on the African continent [22]. Amharic is an official working language of the Federal Democratic Republic of Ethiopia [24]. In addition to this, in Ethiopia, it is used in commerce, government, media, national education, military, and religion since the late 12th century. Amharic uses an alphabet referred to as Fidel (ፊደል/fidälə/ in Amharic). According to [42], it adopts all Ge'ez alphabet symbols and added some new symbols of its own. Fidel essentially consists of consonant and vowel characters in different sequences. It is written from left to right, using primary and derived nouns as well as prefixes and suffixes to make sentences agree with two genders and various quantities, similar to Latin languages. But also uses different prefixes and suffixes to conjugate verbs in different tenses. It has 34 basic characters with each having 7 forms for each consonant-vowel combination, giving $7 \times 34 = 238$ syllable patterns, or fidels [43]. The first form is the basic form; the other forms are derived from it by more or less regular modifications indicating the different vowels [44]. In addition to the 238 characters, there are other non-standard alphabets that contain special features usually representing labialization. Each alphabet represents a consonant together with its vowel. The vowels are fused to the consonant form in the form of diacritic markings. The diacritic markings are strokes attached to the base characters to change their order [45]. Figure 2.1 shows Amharic letters:

| | | | | | | | | |
|---|---|---|---|---|---|---|--|---|
| ሀ | ሁ | ሂ | ሃ | ሄ | ሀ | ሁ | | |
| ለ | ሉ | ሊ | ላ | ሌ | ል | ሎ | | ሏ |
| ሐ | ሑ | ሒ | ሓ | ሔ | ሕ | ሐ | | ሑ |
| መ | ሙ | ሚ | ሚ | ሚ | ም | ሞ | | ሟ |
| ወ | ወ | ዐ | ዐ | ዐ | ሥ | ሦ | | ሧ |
| ረ | ሩ | ሪ | ሪ | ሪ | ር | ሮ | | ሯ |
| ሰ | ሱ | ሲ | ሳ | ሴ | ሰ | ሱ | | ሴ |
| ሸ | ሹ | ሺ | ሻ | ሼ | ሸ | ሹ | | ሺ |

| | | | | | | | | |
|---|---|---|---|---|---|---|--|---|
| ከ | ከ | ከ | ካ | ከ | ከ | ከ | | ካ |
| ኸ | ኸ | ኸ | ኹ | ኸ | ኸ | ኸ | | ኹ |
| ዐ | ዐ | ዐ | ዐ | ዐ | ዐ | ዐ | | |
| ዐ | ዐ | ዐ | ዐ | ዐ | ዐ | ዐ | | |
| ዘ | ዘ | ዘ | ዘ | ዘ | ዘ | ዘ | | ዘ |
| ዠ | ዠ | ዠ | ዠ | ዠ | ዠ | ዠ | | ዠ |
| የ | የ | የ | የ | የ | የ | የ | | |
| ደ | ደ | ደ | ደ | ደ | ደ | ደ | | ደ |

Figure 0.1 Amharic Letters

2.2.1 Amharic Punctuations

Punctuation marks are symbols that are used in sentences and phrases to make the meaning clearer. Amharic has its own punctuation marks. The most commonly used punctuation marks in Amharic are:

- ፡ is used to separate words. Nowadays, it is uncommon to see the punctuation mark ‘:’ in Amharic electronic or paper-based writings instead white spaces are used to demarcate words.
- ። is used to show the end of a sentence.
- ፣ is used to separate comparative and sequential lists of names, phrases, or numbers as well as to separate parts of a sentence that are not complete on their own.
- ፤ is used to separate equivalent main phrases in one idea. Even though it is not placed at the end of a paragraph, it can be used to separate sentences with similar ideas in a paragraph.
- ፥ is used to introduce speech from a descriptive prefix.
- ? Indicates an interrogative clause or phrase.
- ! Is used to emphasize strong feelings and placed after a word or at the end of a sentence.
- ፡- is used following clarification of a certain subject. It will preface validation statements and examples that support the clarification.

2.2.2 Amharic Morphology

Amharic is one of the morphologically rich languages. Like other Semitic languages, Amharic exhibits a root-pattern morphological phenomenon. That is, the Amharic word units exhibit, phoneme, morpheme, root, stem, and word [46]. A phoneme is every glyph or consonant form and a morpheme is the smallest meaningful unit in a word which is a phoneme or collection of phonemes. A root is a set of consonants (also called radicals) that has a basic lexical meaning. It is the basis for the derivation of verbs. A pattern consists of a set of vowels that are inserted among the consonants of a root to form a stem. The pattern is combined with a particular prefix or suffix to create a single grammatical form [41] or another stem [51]. Stems are, therefore, formed by intercalating the vowels among root consonants. A stem can be free or bound. A free stem can stand as a word on its own whereas a bound stem has a bound morpheme affixed to it. A morpheme can be free or bound, where a free morpheme can stand as a word on its own whereas a bound morpheme cannot. For instance, ቤሰውነት/bäsäwənätə/ can be a good example of a morpheme. This word can be separated into three separate morphemes: ቤ-/bä/ the prefix, ሰው/säwə/ the root, and -ነት/nätə/ the suffix. A collection of phonemes or sounds creates a word, which can be as simple as a single morpheme or contain several of them. According to [46], the Amharic language makes use of both prefixing and suffixing to create inflectional and derivational word forms. These word forms are discussed below:

Inflectional Morphology

Inflection is a morphological variation that does not change the part of speech category and general meaning, but the grammatical function. Morphemes or affixes can be added to a word to mark the syntactic function of the word by keeping its original class. This way of forming a word is called inflectional morphology [47]. According to [32], since the Amharic language is highly inflectional, a given root of a language word can be found in different forms. Below are the highly inflected parts of speech in Amharic:

Noun (ሰም/səmə/): Noun is any member of a class of words that typically can be combined with determiners to serve as the subject of a verb, can be interpreted as singular or plural, can be replaced with a pronoun, and refer to an entity, quality, state, action, or concept. Amharic nouns inflect for case, number, definiteness, and gender marker affixes [48]. Therefore, a noun is in the form of stem + {Gender Marker Suffix, Number Marker Suffix, Case Marker Suffix,

Definiteness Marker Suffix}. The gender marker suffix comes from the set {-ኢት'it'}, the number marker suffix comes from the set {-ኣች'oc',-ዎች 'woc',-ኣን 'an',-ኣየ 'Iye', -ኣት'at'}, the case marker suffix comes from the set {-ን'n', -ዩ'yE',-ኤ 'E',-ዎ 'wo',-ሀ 'h',-ኧ 'x',-ኡ 'u', ዋ'wa',-ኣችን 'acn',-ኣችሁ 'achu', -ኣቸው 'acew'}, whereas the definiteness marker suffix comes from the set {-ኡ'u',-ዋ'wa',-ው'wu',-ኢቱ 'itu', -ይቱ'ytu'} [46]. These marker suffixes are elaborated with examples in the following tables:

Table 0.1 Gender, number, and case marker suffixes

| Word(amha ric,afan oromo) | Gender Marker | | Number | | Case | |
|---------------------------------|---------------|------------------------|-------------|------------|-------------|-------------------------|
| | Masculine | Feminine | Singular | Plural | Nominative | Accusative |
| ፍየል/fəyälə/ Re'ee | ፍየል/fəyälə/ | ፍየል-ኢት/fəyälə- 'itə | ፍየል/fəyälə/ | ፍየል- ኣች | ፍየል/fəyälə/ | ፍየል-ን/ fəyälə-nə |
| ልጅ/ləğə/ aa | ልጅ/ ləğə/ | ልጅ-ኢት/ləğə- 'itə | ልጅ/ləğə/ | ልጅ-ኣች | ልጅ/ ləğə/ | ልጅ-ን/ləğə- nə |
| ግመል/gəməlä/ gaala | ግመል/gəməlä/ | ግመል- ኢት/gəməlä-'itə | ግመል/gəməlä | ግመል- ኣች | ግመል/gəməlä/ | ግመል- ን/gəməlä- nə |

Table 0.2 Definiteness marker suffixes

| Word | Number | Gender | Definiteness |
|-------------|----------|-----------|--------------|
| ፍየል/fəyälə/ | Singular | Feminine | ፍየል-ዋ |
| | | Masculine | ፍየል-ኡ |
| | Plural | | ፍየሎች-ኡ |
| ልጅ/ləğə/ | Singular | Feminine | ልጅ-ዋ |
| | | Masculine | ልጅ-ኡ |
| | Plural | | ልጆች-ኡ |
| ግመል/gəməlä/ | Singular | Feminine | ግመል-ዋ |
| | | Masculine | ግመል-ኡ |
| | Plural | | ግመሎች-ኡ |

Verb (ግስ/gəsə/): are words that indicate action. They take place at the end of clause positions. They are derived from roots and use a combination of prefixes and suffixes to indicate the

person, number, voice (active/passive), tense, and gender. Amharic verbs take subject markers as a suffix like -ሁ /-hu/ for the subject ‘I’, -ህ /-h/ for the subject ‘You’, ኝ /-c/ for the subject ‘She’ and so on, to agree with the subject of the sentence [49]. Similar to nouns and adjectives verbs are also derived from Verbal Roots by affixing the vowel ኣ, Verbal Stems by affixing morphemes, and compound Words of stems with verbs. According to [49] and [50], the followings are some properties of Amharic verbs:

- Among the seven Amharic writing symbols order, the majority of the verb words use the first order Amharic writing system
- Verbs can use ‘እየ-’ prefix morpheme
- Verbs can change their last symbol to the Amharic seven order writing system. Finally, these changed verbs may take ‘-አል’ suffix morpheme
- Verbs are placed at the end of an Amharic sentence, and
- Verbs have a suffix attached to them indicating the subject of the sentence

For instance in ግብጽ ድርድሩን ተረጎሞኝ/gəbəs’ə dəɾədərənə tə-räta-čə/ Egypt loses the negotiation/, the underlined word is a verb with prefix ተ- and the suffix -ኝ.

Verbs are inflected for person, gender, number, and time with the basic verb form being the third person masculine singular. The perfect tense (basic form) normally expresses the past tense. Prefixes such as እ‘I’-, ት‘t’-, and ይ‘y’- are used for the first-, second-, and third-person future forms [51], and the suffixes ኣ‘ea’, and ኣኝ‘eac’ are used to indicate masculine and feminine subjects respectively. Verbs in passive voice are marked by a suffix that depends on person and number, whereas both prefixes and suffixes are used to inflect verbs to express intention [51]. The following table summarizes verbal markers for a person, gender, and time as adapted from [46].

Table 0.3 verbal markers for person, gender, and time

| Person | Number Marker | | Time Marker | |
|---------------------------|---------------|-------------|--------------|-----------------|
| | Singular | Plural | Past | Present/ Future |
| 1 st | -ከ/ሁ ‘ku/hu’ | -ን ‘n’ | -ከ/ሁ ‘ku/hu’ | እ/እን- ‘I/In’ |
| 2 nd masculine | -ከ/ህ ‘k/h’ | -አኝሁ ‘achu’ | -ከ/ህ ‘k/h’ | ት- ‘t’ |
| 2 nd feminine | -ሽ ‘x’ | -አኝሁ ‘achu’ | -ሽ ‘x’ | ት- ‘t’ |

| | | | | |
|---------------------------|-----------|-------------|-----------|--------|
| 2 nd plural | -ኡ ‘u’ | -ኡኸሁ ‘achu’ | -ኡ ‘u’ | ት- ‘t’ |
| 3 rd plural | -ኡ ‘u’ | -ኡ ‘u’ | -ኡ ‘u’ | ይ- ‘y’ |
| 3 rd masculine | -ኡ ‘ea’ | -ኡ ‘u’ | -ኡ ‘ea’ | ይ- ‘y’ |
| 3 rd feminine | -ኡኸ ‘eac’ | -ኡ ‘u’ | -ኡኸ ‘eac’ | ት- ‘t’ |

As it is described in [51], verbs in the passive voice are marked by suffixes that depend on person and number, whereas both prefixes and suffixes are used to inflect verbs to express intention. The following table illustrates this.

Table 0.4 Passive voice markers

| Person | 1 st | 2 nd masculine | 2 nd feminine | 3 rd masculine | 3 rd feminine |
|----------|-----------------|---------------------------|--------------------------|---------------------------|--------------------------|
| Singular | -ኸ‘N’ | -ሀ‘h’ | -ኸ‘x’ | -ው‘w’ | -ኡት‘at’ |
| Plural | -ን‘n’ | -ኡኸሁ‘achu’ | -ኡኸሁ‘achu’ | -ኡኸው‘acew’ | -ኡኸው‘acew’ |

Adjectives (ቅጽል/ቃስ’ልፀ):- Amharic adjectives modify nouns or pronouns by describing, identifying, or quantifying words. Nouns tell about things nature, but adjectives tell about things behavior or characteristics, like shape, size, color, type, and/or property. Adjectives are also inflected for gender, number, and case similar to nouns. However, as it is in [51], some categories of adjectives can be marked for numbers by repeating the second consonant so that the forth vowel form is used between consonants that are to be repeated. For instance, the singular form ቀይ/qäyā/ (red) become ቀይይ/qäyayā/ ‘Qeyay’ (ቀይኣይ) in the plural. Adjectives always come before nouns or pronouns which they modify. But all the words that come before nouns cannot always be an adjective. For example: - ይህ መጽሐፍ/yəhə mäs’əhafə/ “*This book*” In this example ይህ/ “This” precede the noun መጽሐፍ/mäs’əhafə/ “book” but this doesn’t mean ይህ/yəhə/ “this” is an adjective, it is a pronoun. In Amharic, some of the morphemes that are used to inflect given adjectives are, ‘-ኡ/o’, ‘-ኡት/it’, ‘-ኡኸ/oc’ and ‘ት-/t’.

Derivational morphology

Derivational Morphology is a morphology concerned with how words are derived from morphemes through processes such as affixation or compounding. Affix is a morpheme fastened to a stem or base form of a word, and modifies its meaning or creates a new word [52]. In Amharic affixes can be prefix, suffix, and infix. A prefix is a morpheme added at the beginning of a word whereas suffixes are added at the end to form derivatives. Infixes are inserted in the body of a word causing a change in meaning, which can be easily observed in the iterative and reciprocal aspect of a root word in the Amharic language [52], [50], and [53]. Some of the Amharic suffixes are: -ነት 'net', -ኧት 'eat', -ኤ 'E', -አ 'o', -አት 'at', -አሽ 'ox', -አታ 'ota', -ኤት 'Et', -ኢያ 'iya', -ኛ 'Na', -ኧ 'ea', -ኢ 'i', -አዋ 'awa', -አዊ 'awi', -ኢት 'it', -እና 'Ina', -አ 'a', -አት 'ot', -ኤታ 'Eta', and -ታ 'ta'.

Noun: unlike inflection, nouns can be derived and have a new structure and meaning as well as word classification [51], [53]. The following are examples of derivational Amharic nouns: ቤተኛ, ሰውነት, ልጅነት, ከብረት/betäña, säwänätə, ləğänätə, kəbərätə/ respectively etc. are derived from the original nouns, ቤት, ሰው, ልጅ,/betə, säwə, ləğə,/ and ከብር/kəbərə/ by attaching the morphemes like ኧኛ, ነት, and ኧት respectively. Nouns can be derived from the root and stem forms of verbs by infixing vowels between consonant and affixing morphemes respectively. It can be also derived from adjectives.

Verbs: Verbs are words that indicate action and they take place at the end of clause positions. Amharic verbs take subject markers as a suffix like -ሁ /-hu/ for the subject 'I', -ህ /-h/ for the subject 'You', ኧ /-c/ for the subject 'She' and so on, to agree with the subject of the sentence [49]. As it is described in [46], verbal roots follow the pattern CCC, where C is a consonant. When a verb is derived from such a root, vowels are fused with one or more of the consonants; thus if [v] denotes an optional vowel, then the root CCC changes into C[v]C[v]C[v]. For example, ንግር/nəgərə/ 'ngr' changes into ነገረ/nägärä/ 'negere' (ን-ኧ-ግ-ኧ-ር-ኧ 'n-ea-g-ea-r-ea'). More generally, the last consonant can either remain the same (C) or add a vowel (C[v]) or add a longer suffix.

Adjectives: Amharic adjectives modify nouns or pronouns by describing, identifying, or quantifying words [49]. They are derived from verbs, nouns, verbal roots, and stems by adding suffixes. The most commonly used suffixes include; -ኛ 'Na', -አማ 'ama', -አም 'am', -አዊ 'awi', -አ 'a'.

For example, አመድ/’ämädə/ ’amed’ for “ash” is converted to አመዳም (አመድ-አም) /’ämädamə/ ‘amedam’ for ‘ashy’.

2.2.3 Amharic Phrases

A phrase is a structure in a language that is constructed from one or more words in the language. It is a syntactic structure that consists of one or more than one word but lacks the subject-predicate organization of a clause. A phrase is composed of either only headword or other words or phrases with the head combination. The other words or phrases that are combined with the head in phrase construction can be specifiers, modifiers, and complements [32], and [49].

Modifiers are used to specifically point out the amount, time, place, type, etc. of the headword or phrase in the phrase construction. They can be adjectival phrases, noun phrases, prepositional phrases, or sentences [49]. By taking ጥቁር መኪና/t’əqurə mäkina/ “black car” as an example, ጥቁር/t’əqurə/ “black” is an adjectival phrase. This shows one word can be a phrase. This adjectival phrase ጥቁር/t’əqurə/ “black” specifically point out the type or color of መኪና/ mäkina/ “car”. Modifiers can also be sentenced as in ትናንት የተመረቀው ፓርክ/tənanətə yätämäräqäwə parəkə/ “the park which inaugurated yesterday”. Here, ትናንት የተመረቀው/ tənanətə yätämäräqäwə/ “which inaugurated yesterday” is a sentence that points out the headword ፓርክ/ parəkə/ “park”. In addition to this, more than one modifier may come in phrase construction.

On the other hand, **Specifiers** are used to specifying the identity, location, number, and possession, etc. of the head. They can be genitive, deictic, or quantifier. They may be either primitive or derived. The derived specifiers can be derived from the noun by prefixing 'የ'/yä/ morpheme to indicate possession [49]. Example: - የጫላ ላም/yäč'ala lamə/ “Chala’s Cow” Here the specifier is የጫላ/ yäč'ala/ “Chala’s” that shows the owner of the cow. As of [49], the derived specifiers can also be derived from modifiers (especially prepositional phrases) by combining specifiers.

Complements are used to make ideas complete [49]. For instance, in ጫማ ገዛሁ/ č'ama gäzahu/ “I bought shoes” and የቆዳ ጫማ ገዛሁ/yäqoda č'ama gäzahu/ “I bought leather shoes”, the first sentence does not give full information about the shoes but in the second sentence የቆዳ“leather” is a compliment that indicates from what the shoes which shows full information about the shoes.

In Amharic, phrases are categorized into five categories, namely noun phrase, verb phrase, adjectival phrase, adverbial phrase, and prepositional phrase [51], [53], [32].

Noun phrase

A noun phrase (NP) is a phrase that has a noun as its head. In this phrase construction, the head of the phrase is always found at the end of the phrase. This type of phrase can be made from a single noun or combination of a noun with either other word classes including noun word class or phrases. That means one noun can be a noun phrase [49]. For example, in ፍየሏ ሁለት ግልገል ወለደች/fəyäləwa hulätə gələgälə wälädäčə/ “The goat delivered two kids”, there are two parts: the subject ፍየሏ/ fəyäləwa/ “the goat” and the object with the verb ሁለት ግልገል ወለደች/ hulätə gələgälə wälädäčə/ “delivered two kids”. Thus, the first part is a noun phrase and the second is a verb phrase. Therefore, the noun phrase in the above example is only the noun “the goat”.

A noun phrase can be simple or complex. The simplest noun phrase consists of a single noun or pronoun such as እሱ/’əsü/ (he), እሷ/’əšä/ (she), እነሱ (they), etc. A complex noun phrase can consist of a noun (called the head) and other constituents (like complements, specifiers, adverbial and adjectival modifiers) that modify the head from different aspects [32].

Verb Phrase

Amharic verb phrase is constructed with a verb as a head and other constituents such as complements, modifiers, and specifiers [36]. For example: in the following Amharic verb phrase, መርጌ ወደ መናፈሻ ሄደች/märəge wädä mänafähä hedäčə/ “Mergie went to park”, ወደመናፈሻ/wädämänafähä/ (to park) is a prepositional phrase modifying the verb ሄደች/hedäčə/ (went).

Adjectival phrase

Amharic Adverbial phrases (AdvP) are made up of one adverb as a headword and one or more other lexical categories including adverbs themselves as modifiers. The head of the AdvP was also found at the end. Unlike other phrases, AdvPs do not take compliments. Most of the time, the modifiers of AdvPs are PPs that comes always before adverbs [53], [51], [49]. For instance, in መርጌ እንደ አክሱት በጣም ቆንጆ ነች/märəge ’ənädä ’äkəsəta bät’amə qonəğə näčə/ “Mergie is so beautiful like her aunt”, the phrase እንደ አክሱት በጣም ቆንጆ/’ənädä ’äkəsəta bät’amə qonəğə/ “so

beautiful like her aunt” is an adjectival phrase. In this phrase, እንደ አክሱት/’ənədä ’äkəsəta/ “like her aunt” is a modifier, while በጣም/ bät’amə/ “very” is a specifier.

Prepositional phrase

Amharic prepositional phrase (PP) is made up of a preposition (Prep) head and other constituents such as nouns, noun phrases, prepositional phrases, etc. [51], [53], and [49]. Unlike other phrase constructions, prepositions cannot be taken as a phrase instead they should be combined with other constituents and the constituents may come either previous to or after the preposition which is the head of the phrase. Broadly speaking, if the complements are nouns or NPs, the position of prepositions are in front of the complements [49]. For instance, in እንደ ትልቅ/’ənədä tələqə/ ሀገር/hägärə/ “like big country”, እንደ/’ənədä / “like” is a preposition which is combined with the noun ሀገር/hägärə/ “country” and it comes in front of complement ትልቅ/tələqə/ “big”; whereas if the complements are PPs, the position will shift to the end of the phrase [49]. For instance, in ዛፍ ስር/zafu sərə/ “under the tree” ስር/ zafu/ “under” is a prepositional phrase that is combined with the noun ዛፍ/sərə/ “tree” and it came at the end.

Adverbial phrase

Amharic Adverbial phrases (AdvP) are made up of one adverb as a headword and one or more other lexical categories including adverbs themselves as modifiers. The head of the AdvP was also found at the end. Unlike other phrases, AdvPs do not take compliments. Most of the time, the modifiers of AdvPs are PPs that comes always before adverbs [51], [53], [49]. For instance, in መርጌ እንደ አክሱት በጣም ቆንጆ ነች/märəge ’ənədä ’äkəsəta bät’amə qonəgo näčə/ “Mergie is very beautiful like her aunt” እንደ አክሱት በጣም/’ənədä ’äkəsəta bät’amə/ “like her aunt” is an adverbial phrase and the headword is በጣም/bät’amə /(very). The modifier that is found in the AdvP is እንደ አክሱት/’ənədä ’äkəsəta/, (like her aunt) which is comparative PP.

2.2.4 Amharic Sentences

In Amharic grammar, the groups of phrases that together express ideas are called a sentence. The sentence structure for the Amharic language, like that of Afaan Oromo, is a Subject-object-verb (SOV) structure, contrasting to English with a subject-verb-object combination [2], [53], [51]. For example in አበበ በሶ በላ/’äbäbä bäsö bälä/ “Abebe ate Beso”, the Amharic sentence is composed of አበበ/’äbäbä/ the subject, በሶ/bäsö/ the object, and በላ/bälä/ the verb which is in

contrast to English. Amharic sentences are constructed from simple or complex NP and simple or complex VP but NP always comes first as a subject [49].

Examples:-

- a) ብዙ ቀያይ በጎች/bəzu qäyayə bāgočə/ “many red sheep”
- b) ወደ ዱር ሄዱ/wädä durə hedu / “went to wild”
- c) ብዙ ቀያይ በጎች ወደ ዱር ሄዱ/bəzu qäyayə bāgočə wädä durə hedu/ “many red sheep went to wild”

As we can see here, the first two constructions do not express the full idea but the last one does. Because the last one expresses full information such as; who did go wild? Where sheep did went? etc. All these questions have been answered by the last word construction. In the last construction, there are NP and VP which build the sentence and these are NP ብዙ ቀያይ በጎች/bəzu qäyayə bāgočə/ “many red sheep” and VP ወደ ዱር ሄዱ/wädä durə hedu/ “went to wild”. The remaining phrases (other than NP and VP) are being constructed in NPs or VPs that are found in a sentence. Based on this construction, sentences can be simple or complex.

Simple Sentences

Simple sentences are sentences, which contain only one verb. A simple sentence can be constructed from NP followed by VP which only contains a single verb [49], [53], and [51]. For instance in አስቴር መጽሐፍ ገዛች/’äsäterə mäs’əhafə gäzačə/, “Aster bought a book”, there is only one verb ገዛች/ gäzačə/, “bought”. Here, the verb ገዛች/ gäzačə/, “bought” is transitive which takes the object መጽሐፍ/mäs’əhafə/ “book”. On the other hand, in ዳምጠው ወደ አዲስ አበባ ሄደ/damət’äwə wädä ’ädisə ’äbäba hedä/, “Damtew went to Addis Ababa”, the verb ሄደ/ head/ “went” does not take any object so that it is an intransitive verb.

According to [51], simple sentences can be declarative sentences, interrogative sentences, and imperative sentences based on the purpose for which they are spoken.

Declarative sentences: Declarative sentences have the nature of making a declaration or they are just simple statements. Declaration sentences are used to convey ideas and feelings of the speaker about things, and happenings. In Amharic, declarative sentences always end with the Amharic punctuation mark አራት ነጥብ (“::”) /’äratə nä’təbə/ which is equivalent to full-stop (.) in

the English language [35]. For example, sentences like ስጋው ጭማ ነው/səgawə č'oma näwə / “the meat is fatty”, አስቴር አስተማሪ ሆነች/’äsäterə ’äsätämarī honäčə / “Aster became a teacher” are some types of declarative sentences.

Interrogative sentences: In Amharic Interrogative sentences are sentences that can form a question. A sentence that asks about the subject, complement, or the action specified by the verb is said to be an interrogative sentence. The question can be the one that asks about known things to be sure or the one that asks the unknown thing. In order to ask the unknown thing, the enquirer can use interrogative pronouns like ማን/manə/, “who”, ምን/mənə/ “what”, የት/yätə/ “where”, ስንት/sənətə/ “how many”, and መቼ/mäčä / “when” whereas in order to make sure the known thing the enquirer can use assurance words or he can change the way of his speech. An interrogative sentence always ends with a question mark (?) [35], [51], and [2]. For instance: ምሳህን በላህ? /məsaḥənə bälähə? / “Did you eat your lunch?” መቼ መጣህ?/mäčä mät’ahə? / “When did you come?” are examples of interrogative sentences for making sure what is known and to inquiry unknown things respectively.

Imperative sentences: Imperative sentences are used to convey commands or instructions. Most of the time, the subject (that is a second-person pronoun) of the sentence is omitted but since Amharic words are highly inflected, subject marker prefixes indicate the specific subject. However, sometimes when the command is passed for a third person, the subject of the sentence can be third person pronoun or noun [35], and [51]. For example in ቡና አፍዩ/buna ’äfäyi/ “make coffee”, the subject is አንቺ/’anəči/ “you” which is second person feminine singular. On the other hand, in ካሰች ቡና ታፍላ/kasäčə buna tafəla/ “let Kasech make coffee”, the command is for the third person that doesn’t exist at the time of speech so the subject will be አሷ/’əሷ/ “she” which is the third person feminine singular.

Complex Sentences

Complex sentences are formed from either complex noun phrases or complex verb phrases or both. In other words, a complex sentence can have a complex NP and a simple VP, a simple NP, and a complex VP, or both complex NP and complex VP. Complex NPs contain at least one embedded sentence which can be a complement or another type of phrase. On the other hand, complex VPs contain at least one sentence or more than one verb [35], [53], and [51].

For instance: ካሳ የገባበት የሳር ቤት በጣም ትልቅ ነው/kasa yägäbabätə yäsarə betə bät'amə täləqə näwə/ “the thatched house that Kassa has entered is so big”. In this sentence, the head of the noun phrase is ካሳ የገባበት የሳር ቤት/ kasa yägäbabätə yäsarə betə/ “the thatched house that Kassa has entered”. The head with the complement የሳር/yäsarə/ “thatched” forms simple noun phrase የሳር ቤት/ yäsarə betə / “thatched house” and this noun phrase are combined with the embedded sentence or clause ካሳ የገባበት/kasa yägäbabätə/ “that Kassa has entered” to form a complex noun phrase. But the clause that makes the complex phrase is dependent which is identified by the morpheme የ/yä/ “that”.

2.3 A Brief Overview of Afaan Oromo

Afaan Oromo is one of the languages of the lowland East Cushitic within the Cushitic family of the Afro-Asiatic Phylum like Somali, Sidama, Hadiya, and Afar-Saho [20], [21], and [55]. Even though it is mainly spoken by the Oromo people, it is also one of the major languages spoken in Ethiopia. According to [22], Afaan Oromo is the sixth most widely spoken language in Africa preceded by Swahili, Arabic, French, Hausa, and Yoruba. The Oromo language also referred to as Afaan Oromo or Oromiffaa has more than 30 million speakers across Africa. It is spoken mainly in countries of the Horn of Africa including Kenya and Somalia and of course Ethiopia. It is also a medium of instruction and a school subject in primary and secondary schools in the Oromiya regional state, one of the administrative regions that have the largest population (33,692,000 as of [56]) of all the regions of the Federal Democratic Republic of Ethiopia (FDRE) [23].

Afaan Oromo Writing System

Afaan Oromo uses a Latin-based alphabet known as Qubee that consists of thirty-two symbols in general. Of these, eighteen basic consonants (dubbifamaa) with three additional consonants, that are included to write the borrowed/ foreign words and five vowels (10 if double vowels are needed to be considered) (dubbachiiftuu). Double consonant letters are derived from a combination of two consonant letters. Qubee is characterized by capital and small letters which is known in the English alphabet. As in English, vowels are sound makers and are sounds by themselves. Vowels in Afaan Oromo are characterized as short and long vowels. Since there are not any indigenous Afaan Oromo words that contain ‘p’, ‘v’, and ‘z’, the basic consonant also does not include these. However, when foreign words such as ‘post’ (poostaa in Afaan Oromo)

are referred to, these characters are used. Thus, the extended alphabet of the Afaan Oromo includes these characters (‘p’, ‘v’, and ‘z’) to support foreign words. In addition to these symbols, an apostrophe is used to represent a sound in addition to its use as a punctuation mark. It represents a hiccup-like sound (called hudhaa) as in qe’ee and fal’aana. [57]. Table 5 as adopted from [37], depicts the Afaan Oromo writing script.

2.3.1 Afaan Oromo Punctuation Marks

The most commonly used punctuation marks in Afaan Oromo are:

- . The **period** is placed at the end of declarative sentences, statements thought to be complete, and after many abbreviations.
 - Example:
 - Ani kitaaba bitadhe. “I bought a book.”
 - A.L.A. “G.C.”
- ? **Question mark** is used to indicate a direct question when placed at the end of a sentence. Example: Na wajjin dhuftaa? (Can you come with me?)
- ! **Exclamation mark** is used at the end of the command and exclamatory sentences.
 - Example: ol ka’i! “Stand Up!”
- , **Comma** is used to show a separation of ideas or elements within the structure of a sentence. Example: Caalaa, Margee, Dachaa fi Olhiqaan as turan. “Chala, Margie, Dacha and Olhika were here”
- : **Colon** is used to separate and introduce lists, clauses, and quotations, along with several conventional uses.
- ; **Semicolon** is used to connect independent clauses. It shows a closer relationship between the clauses than a period would show.

Table 0.5 Afaan Oromo Writing Scripts

| The basic consonant letters | | | | | | |
|--------------------------------|-----------|----------|-----------|-----------|-----------|------------|
| Capital | B | C | D | F | G | H |
| Small | b | C | D | F | G | H |
| Sound relative to English word | as in bad | ch sound | As in dad | As in far | As in gap | As in hall |

| | | | | | | | | | | |
|--|--------------|------------|--------------|--------------|-----------------|-----------------|----------|------------|------------|------------|
| Capital | J | K | L | M | N | Q | | | | |
| Small | j | K | L | M | N | Q | | | | |
| Sound relative to English word | As in jacket | As in car | As in lab | As in man | As in narrow | Glottal sound k | | | | |
| Capital | R | S | T | W | X | Y | | | | |
| Small | r | S | T | W | X | Y | | | | |
| Sound relative to English word | As in rat | As in sand | As in task | As in wallet | Glottal sound t | As in yard | | | | |
| The three consonant letters to write the borrowed foreign words | | | | | | | | | | |
| Capital | P | | V | | Z | | | | | |
| Small | P | | V | | Z | | | | | |
| Sound relative to English word | As in pasta | | As in virus | | As in zigzag | | | | | |
| The vowel letters in longer and shorter sound | | | | | | | | | | |
| Capital | A | AA | E | EE | I | II | O | OO | U | UU |
| Small | a | Aa | E | Ee | i | Ii | O | oo | U | Uu |
| Sound relative to English word | As in cut | As in far | As in in egg | As in eight | As in hip | As in teeth | As in or | As in hole | As in bull | As in book |

2.3.2 Afaan Oromo Morphology

Morphology is a branch of linguistic that studies and describes how words are formed in a language [60]. Morphemes are the minimal meaningful units which may constitute words or parts of words. There are two categories of morphemes in Afaan Oromo; free and bound morphemes. Free morpheme can stand as a word on its own whereas bound morpheme does not occur as a word on its own.

In Afaan Oromo roots are bound as they cannot occur on their own like dhug- “drink” and beek- “know”, which are pronounceable only when other completing affixes are added to them. In other words, these roots serve as base stems in Afaan Oromo since they possess non-verbalized glosses [61]. The same to the root, affix is also a morpheme that cannot occur independently. It is attached in some manner to the root, which serves as a base. Affixes are of three types: prefix, infix, and suffix. Prefix and suffix occur at the beginning and end of a root respectively. For instance in beekumsa “knowledge”, -umsa is a suffix while beek- “know” is a stem. Suffixes are

the predominant affix in the language [62]. As far as the researcher could ascertain from the existing literature, there is no sufficient evidence that shows infixes are used in Afaan Oromo.

As discussed by [63], the same to other languages like Amharic and English, there are two kinds of morphology (inflectional and derivational) in Afaan Oromo. Derivational morphology deals with those changes that result in changing classes of words (changes in the part of speech). For instance, noun or an adjective may be derived from a verb.

Inflectional Morphology

Inflectional morphology is concerned with the inflectional changes in words where word stems are combined with grammatical markers for things like person, gender, number, tense, case, and mode. Inflectional changes do not result changes in parts of speech. Highly inflected word classes in Afaan Oromo are as follows:

Noun

Almost all nouns in Afaan Oromo end with a vowel except for a few of them which end in specific consonants like n, l, t. Afaan Oromo nouns inflect for gender, number, definiteness, and case marker affixes [64], and [63].

Number

Afaan Oromo has different suffixes to form the plural of a noun. The use of different suffixes differs from dialects to dialects. The majority of noun plural forms were formed by using the suffix –oota/ -ota, followed by -lee,-wwan,-een,-olii,-olee and -a(n) [65]. In the other way in collective nouns, some exist in plural form only (for example; hamaamota, “ሠርገኛ” “bridegroom”) whereas some others have the same singular and plural forms like ilkaan “ጥርስ/ጥርሶች” “tooth/teeth”, or quba “ጣት/ጣቶች” “finger/fingers”. In some cases, the plural marker varies based on the semantic nature of the noun. The following table shows the most common plural morphs:

Table 0.6 The most common plural morphs

| Base forms | Inflected form | Amharic/ English Meaning |
|-------------------|-----------------------|---------------------------------|
| Fira | Fir-oota | ዘመዶች/ relatives |

| | | |
|----------|--------------|-----------------|
| Laga | Lag(g)-een | ወንዞች/ rivers |
| Barruu | barruu-lee | መጽሔቶች/ journals |
| Jaarsa | Jaars-olii | ሽማግሌዎች/ elders |
| Barmaata | Barmaat-ilee | ልምዶች/ habits |
| Qaama | Qaam-olee | አካላት/ bodies |

Nouns of kinship terms are marked for plurality by the morpheme -an, which follows either geminate consonant or short /a/. The morpheme triggers the consonant to be geminate when it is appended. It may also occur by lengthening the final short /a/ on the base word when the base noun ends in short /a/ followed by gemination or consonant cluster [64]. The following table illustrates this:

Table 0.7 Plural form of kinship terms

| Base forms | Inflected form | Amharic/ English Meaning |
|------------|----------------|----------------------------------|
| Ilma | Ilma-an | ልጆች /Sons |
| Eessuma | Eessum(m)-an | አጎትበእናት/ Uncles (through mother) |
| Wasiila | Wasiil(l)-an | አጎትበአባት/ Uncles (through father) |

As described in [64], ± animate nouns may take the plural morph -an but the phonological form of the nouns, that they end in vowel length, make the morph preceded by the geminate phoneme /-ww-/ for settling occurrence of an impermissible number of vowels. The consonant is doubled because of that the morph -an occurs following geminate consonant. + abstract nouns are members of this morpheme which is preceded by the epenthetic element /-ww-/. All the suffixes in the following table are utilized for a noun ending in a vowel(s). For instance, obboleessa/ “ወንድም”/ “brother” can have a plural form of obboleewwan/ obboleewwan-an “ወንድሞች” “brothers”.

Gender

The same to other Ethiopian and foreign languages, two types of gender, that is, masculine and feminine, exist in Afaan Oromo. These are identified through gender marking suffixes, or

lexically by using different words for masculine and feminine forms [66], and [67]. The distinct words for masculine and feminine like adaadaa “አከሰት” “aunt” and eessuma “አጎት” “uncle” are also used in Afaan Oromo. Gender indicating words can be used for animals and they are placed immediately after or before the nouns they belong to. The most common contrastive pair of words used in this way is kormaa “ወንድ (ለ እንሰሳ)” “male (M)” and dhaltuu “እንሰት” “female (F)” [64]. In general feminine and masculine differs in their ending. Most of the time feminine ends with -tuu, -tii, and -ttii, while masculine ends with -saa, and -aa. The following table depicts the gender inflection of noun in general.

Table 0.8 Gender inflection of noun

| Base Form | Masculine | | Feminine | |
|-----------|----------------|--------------------------|----------------|--------------------------|
| | Inflected form | Amharic/ English Meaning | Inflected form | Amharic/ English Meaning |
| Mararaa | Marar-aa | አሳዛኝ /sorrowful | Marar-tuu | አሳዛኝ /sorrowful |
| Diimaa | Diim-aa | ቀይ/Red | Diim-tuu | ቀይ/Red |
| Sooressa | Sooressa | ሁብታም/rich | Soorettii | ሁብታም/Rich girl |
| Gurbaa | Gurbaa | ወንድ/Boy | Intala | ሴት/Girl |
| Eessuma | Eessuma | አጎት/uncle | Adaadaa | አከሰት/Aunt |

As it can be seen from the above table, the last two examples are distinguished for gender lexically. On the other hand, the first two examples are derived from verbs indicate that the long -aa suffixed to the verb root or to a consonant-final stem marks masculine gender whereas the suffix -tuu makes verbal nouns in the feminine gender.

Case

Case is a grammatical relationship of nouns or pronouns to other words in a sentence. It is a morpho-syntactic category that is construed in its syntagmatic occurrence. Languages differ especially in a morphological case rather than a syntactic case [64]. According to [68] there are six types of cases in Afaan Oromo: nominative, accusative, genitive, dative, ablative, and vocative ones; he considers the absolutive case as the primitive form of nouns.

The nominative case is marked by four different morphs of allomorphic variation occurring in complementary distribution. The allomorphs for the nominative case are -n, -ni, -i, and Ø. The dative case signals a noun that takes the position before or after the direct object with the

function of telling „for whom“ or „to whom“ the action is done as semantic criteria. The dative case can be marked either by –f or –tti suffixes. The case marked on nominals for the indication of possession is known as a genitive case [64]. According to [69], the genitive case is formed in two ways: by prefixing “kan” and lengthening the last vowel (or suffixing -i to the final consonant of the possessor noun; and by juxtaposing the thing possessed and the possessor in that order and lengthening the final vowel of the possessor if it is short (or suffixing -i after -C). An ablative case is used to express origin or from where a movement begins. For nouns that end in a long vowel, long -aa and –ii following copular elements -dha and -ti respectively are used to show ablative case. Vocative cases mark the noun representing the entity (animate) we address. It is a verbal means of calling attention. In Afaan Oromo, “yaa”, and “-na” are most commonly used to mark vocative cases. The following table shows these different types of case markers:

Table 0.9 different types of case markers

| Base form | Type | Inflected forms | Amharic/ English Meaning |
|-----------|------------|-----------------|--------------------------------|
| Siree | Nominative | Siree-n | አለጋው/ the bed |
| Hirriba | Nominative | Hirrib-ni | አንቅልፍ/ sleep |
| Morma | Nominative | Morm-i | አንገቱ(ቷ)/ the neck |
| Nama | Dative | Namaa-f | ለሰው/ to someone |
| Abbaa | Dative | Abbaa-tti | ለአባት/to father |
| Farda | Genitive | Fard-aa | ለፈረስ/of horse |
| Bishaan | Genitive | Bishaan-ii | ለውሃ/of water |
| Jimma | Ablative | Jimmaa | ከጅማ/ from Jimma |
| Adaamaa | Ablative | Adaamaa-dhaa | ከአዳማ/ from Adama |
| Nama | Vocative | Nama-na | አንተ!/you/ guy |

Verbs

According to [111], verb (xumura) is a word that expresses action, state of being in or relationship between two things. In Afaan Oromo verbs mostly appear at the end of a sentence. For example: Biyyaan Finfinnee deeme (Biya went to Finfine). Deeme (went) is the verb of the sentence. Like Amharic, Afaan Oromo verbs can be modified to indicate person, gender, tense and number. Different inherent and agreement grammatical categories account for the inflection

of verbs in Afaan Oromo. The inherent ones are aspect, mood, and voice whereas the agreement properties include person, number, gender, and case. According to [64], several studies, especially the earlier ones, tense are also considered as the inflectional categories of verbs in Afaan Oromo. The inflected forms of Afaan Oromo verb are discussed below:

Aspect:

Aspect is context related which morphologically distinguishes between completeness and incompleteness of an action. It is bound with situation and duration, unlike tense which is just about the time of an event in relation to the speech time. In the Oromo language the roots or stems of verbs, usually ending in a consonant, take inflectional morphemes showing a distinction between perfective and imperfective aspects [112]. These two aspects are distinguished primarily by their suffix vowel, which is -a (and its allomorphs -i and -u) for the imperfect and -e (and its allomorph -i) for the perfective. In addition, the continued actions are categorized as an imperfective aspect whereas a short and completed action can be considered as a perfective aspect. In general, as in [111], the concept of perfectness is that an action is prior to a specific moment in time whereas the imperfectness is connected with action in process or in progress.

Example: Isheen uffata miiccite “እሷ ልብስ አጠበች”/ “she has washed the cloth” is in perfect form whereas Caalaan deema jira “ጫላ እየሄደ ነው”/ “Chala is going” is in an imperfect form.

Mood

Mood is the attitude of the speaker towards an utterance. It is originally from the word mode“ which means a specific way of doing something. In connection with the styles of speech which arises from the involvement of feeling, Afaan Oromo has several types of moods from which some of the modal forms are discussed below:

Indicative mood: which involves making statements and asking questions constitutes the most common clause type in Oromo. In its construction, yes/no question is similar to a declarative sentence except the final vowel length along with intonational relevance on the question form.

For example, Caalaan kitaaba bite “ጫላ መጽሐፍ ገዛ”/ “Chala bought a book”

According to [112], the subject is placed at the beginning of declarative sentences as in examples [18]. However, in interrogative sentences also, the subject is placed at the beginning. What

makes interrogative is intonational variation. Actually, the subject can be placed either at the beginning or at the end of a sentence in both declarative and interrogative sentences. This is a kind of topicality shift from subject to object or theme in syntactic consideration.

Imperative Mood: In Afaan Oromo, the imperative begins with the object as it precedes the verb in the word order of the language. Intransitive verbs are used at the beginning of the sentence in the form of the subject ‘you’ understood. However, it may happen following motion verbs like deemuu ‘to go’ or ka’uu ‘to stand’ in their converbal forms. The motion verbs often precede the objects of transitive verbs, and they happen in the terminating vowel length.

Example: Kitaaba sana fidi “ፆንን መጽሐፍ አምጣ”/“Bring that book”

Jussive Mood: [112] states that imperative and jussive have semantic and morphological features in common. The jussive mood is marked by the pre-verbal particle haa and the dependent suffix -u or -i on the verb. They co-occur in a sentence to mark mood and aspect. The suffixes mainly mark the imperfective aspect. This construction is, however, rather syntactic.

Example: Barattoonni haa deeman “ተማሪዎቹ ይሂዱ”/ “Let the students go”

Voice

Voice is a verb form that relates the action of a verb with its participants (or arguments). It tells us if the subject performs or receives the action indicated by the verb. When the subject performs the action the voice is active whereas the form in which the subject receives the action is passive voice. Using sentence types in which the verb form is changed for the purpose of such grammatical function is inflectional. Voice involves all valency changing verb forms including causative and middle; however, here we will discuss the most common ones – active and passive forms. Here are few examples in the table below:

Table 0.10 Afaan Oromo Active and Passive Voices

| Voice | Root | Marker | Inflected form | Amharic/ English Meaning |
|---------|------|--------|----------------|--------------------------|
| Active | Mur- | - | Mur-e | ‘ቆረጠ’/ ‘cut’ |
| | Bit- | - | Bit-e | ‘ገዛ’/‘bought’ |
| Passive | Mur- | -am- | Mur-am-e | ‘ተቆረጠ’/‘was cut’ |
| | Bit- | -am- | Bit-am-e | ‘ተገዛ’/‘was bought’ |

Adjective

According to [36], in Afaan Oromo, adjectives (addeessa) come after the nouns they qualify. They are inflected following the nouns they modify in a sentence. The inflectional categories or properties of adjectives are the same with that of nouns. Adjectives are inflected for number, gender, singulative and case like nouns; however, sometimes they are marked differently from nouns. For instance, adjectives, unlike nouns, are inflected by reduplication to mark plurality.

Number:

When adjectives occur with nouns in sentences, number is marked on both of them. Nouns are marked for plurality, but adjectives are marked for number by reduplication of its initial syllable (CV, CVC), or by the plural suffix -(o)ota. In the former way of marking plurality, the initial syllable reduplication co-occurs with the final vowel shift from -aa to -oo when the adjectives end in long -aa. The latter way of marking a number in adjectives is the same as that of nouns. According to [113], the suffix -(o)ota shows plurality in adjectives. The following table shows some of the examples:

Table 0.11 Number Inflection in Afaan Oromo

| Form of Inflectional | Singular | Plural |
|----------------------|------------------------|-------------------------------|
| Lexical Coding | Hiyyeessa “ድሃ”/“poor” | Hiyyeeyyii “ድሃኛቶ”/ “poors” |
| Reduplication | Cimaa “ጠንካራ”/ “strong” | ci-ccimoo “ጠንካራኛቶ”/ “strongs” |
| -(o)ota | Gamna “ብልጥ”/ “wise” | Gamn-oota “ብልጥኛቶ”/ “wises” |

Gender:

In Afaan Oromo, it is very common for the base forms of adjectives to be used with the masculine. Inflection occurs when we make them fit for the feminine. To inflect adjective in gender, the suffixes like -ittii, -icha(a), -tuu, -oo, -aa, -duu, etc. can be used. The following table shows marked gender on adjectives

Table 0.12 marked gender on adjectives

| Masculine | Feminine |
|---------------------------|----------------------------|
| Ham-aa “ከፉ”/ “Bad” | Ham-tuu “ከፉ”/ “Bad” |
| Gabaab-aa “አጭር”/ “Short” | Gabaab-duu “አጭር”/ “short” |
| Gurraa-cha “ጥቁር”/ “Black” | Gurraa-ttii “ጥቁር”/ “Black” |

Case:

If a noun is marked for a nominative case, an adjective following it will also be marked for the same case. As it is indicated in the following table, the nominative case is marked by three allomorphs whereas the absolutive case is the unmarked object form. The dative case is marked by the underlying suffixes -f. The genitive case is marked by vowel length on the noun or pronoun possessor. The instrumental and beneficiary cases are marked by -n and -f respectively. These are suffixed to the nouns or pronouns under the case.

Table 0.13 Adjectival Case markers

| Case | Marker | Example |
|-------------|--------------|---|
| Nominative | -n, -ni, -i | Qal’aa-n “ቀጭን”/ “thin” Gamn-i “ብልጠ/ጧ”/ “wise” |
| Absolutive | No marker | Gamna “ብልጥ”/ “wise” Xiqqaa “ትንሽ”/ “small” |
| Dative | -f | Diimaa-f “ለቀይ”/ “for red” Dadhabaa-f “ለደካማው”/ “for weak” |
| Genitive | Vowel length | Hamaa “ከፉ”/ “of bad” |
| Beneficiary | -f | Baay’ee-f “ለብዙ”/ “for most” Guddaa-f “ለትልቅ”/ “for big/respected” |

Derivational Morphology in Afaan Oromo

Nouns

According to [63], in Afaan Oromo, derivational suffixes are added to the root or stem of the word to create a derived noun. The following suffixes play an important role in oromoo word

Example: Deemuu “መሄድ”/ “to go”

deemsisuu “ማስኬድ”/ “to cause to go”

Intensive: It is formed by duplication of the initial consonant and the following vowel, geminating the consonant.

Example: Waamuu “መጥራት”/ “to call, invite”

wawwaamuu “መጠራራት”/ “to call

intensively”

Adjectives

Derivational adjective is the creation of new adjectives from nouns or from other adjectives or from verbs.

Example: jaba [base form] “ጠንካራ”/ “strong” jabaachuu [derivated form] “መጠንከር”/ “to be strong”

2.3.3 Afaan Oromo Phrases

In Afaan Oromo there are five different kinds of phrases, namely noun phrase, verb phrase, prepositional phrase, adjectival phrase, and adverbial phrase. They are discussed below:

Noun Phrases

A noun phrase is a phrase that has a noun or indefinite pronoun as its head. For example: in the sentence, Kitaabni Caalaa sun tarsa’e “ያ የጫላ መጽሐፍ ተቀደደ”/ “That book of Chala is torn”, Kitaabni Caalaa is a noun phrase, and the head (noun) of the noun phrase is “Kitaabni”.

Verb Phrases

In a verb phrase the word that the phrase about is the verb. For example: in the sentence, Margeen uffata adii bitte “መርጌ ነጭ ልብስ ገዛች”/ “Marge bought white cloth”, ‘bitte’ is the head of the verb phrase “uffata adii bitte”. The verb phrase tells what Marge did.

Prepositional Phrases

A preposition links a noun to an action or to another noun. A prepositional phrase is a phrase that has a preposition as its head. For example: in the sentence, Erga bokkaan caamee, gara magaalaa deemne “ዝናቡ ካባራ በኋላ ወደ ከተማ ሄድን”/ “After the rain stops, we went to the city”, gara magaalaa

“ወደ ከተማ” “to the city” is a prepositional phrase and the head of the prepositional phrase is gara “ወደ”/ “to”

Adjective Phrases

In an adjective phrase, one or more words work together to give more information about the adjective. For example: in the sentence, Caaltun barnoota ishiitiin daran cimuudha “ጫለቱ በትምህርቷ በጣም ጎበዝ ነች”/ “Chaltu is very clever in her education”, the phrase “barnoota ishiitiin daran cimuudha” is adjectival phrase.

Adverbial Phrases

Adverbs may modify the manner of an action, indicate the time of an action, give location or indicate a degree. Consider the following Afaan Oromo adverbial phrases:

- Margeen jarjaraan dhufte “መርጌ በችኮላ መጣች”/ “Marge came in hurry”; jarjaraan “በችኮላ” indicates the manner of an action.
- Firri isaanii turee dhufa “ዘመዳቸው ቆይቶ መጣ”/ “their relative comes late” turee indicates the time of an action.
- Dabbara keessan bakka kana ka’aadha deema “ደብተራቸን እዚጋ አድርጋቸ ሂዱ”/ “Put your exercise book here and go” bakka kana “እዚጋ” indicates location.

2.3.4 Afaan Oromo Sentences

Afaan Oromo and Amharic are the same in sentence structure order. Like Amharic, the sentence structure of Afaan Oromo is Subject-Object-Verb (SOV). In terms of their structure, Afaan Oromo sentences are classified into simple, compound, complex, and compound-complex sentences.

Simple Sentence

A simple sentence is a sentence with one main clause or it is a sentence that consists of only a single verb in its structure no matter how many subjects are there. No matter how the sentence lengthened, if it has a single verb and if it has not a dependent clause in its construction, then it is a simple sentence.

Example: Caalaan Sangaa bite. “ጫላ ሰንጋ ገዛ”/ “Chala bought an ox”

Compound Sentence

In compound sentence, two or more simple sentences or two or more independent clauses combined together to form a sentence. Each main clause of a compound sentence has its own subject and predicate. These clauses are usually combined by coordinating conjunctions, semicolon, and adding markers like {-e} on the verb to generate long sound.

Example: Namni mana ijaareefi namni barumsa barate gaariidha “ቤት የሰራ እና ትምህርት የተማረ ሰው ጥሩ ነው”/ “the man who built a house and learned education is good”

Complex Sentence

A complex sentence can be formed from one independent clause and one or more than one dependent clause.

Example: Yoo soromtes, yoo hiyyoomtes gorsa abbaa kee hin dagatin “ሁብታም ብትሆንም ድሃ የአባትህን ምክር አትርሳ”/ “don’t forget your father’s advice whether you get poor or rich”

Compound-Complex Sentence

This kind of sentence has both characteristics of a compound and complex sentences. This means: in one sentence there are simple sentences, or two or more independent clauses, also one or more dependent clause(s).

Example: Yommu dhaqes, yommuu gales, natti goree na gaafatee darbe “ሲሄድም ሲመጣም ጎራ ብሎ ጠይቆኝ ሄደ”/ “he visited me when he had gone and come”.

2.4 Machine Translation

MT is a subfield of computational linguistics that is focused on translating text from one language (source language) to another (target language) using computers with or without human assistance [54]. MT can be viewed as a system that builds a representation of the same content in the form of different languages. It is based on the idea that the same content can be expressed by different languages. Ideally, MT is a batch process that is applied to a given text which produces a perfect translated text which then only needs to be printed out [70]. Translation is not a mere word-to-word substitution. A translator must interpret and analyze all elements of a text and know how each word may influence another and this requires extensive expertise in grammar,

syntax (sentence structure/word order), semantics, etc., in the source and target languages, as well as familiarity with each local region in which syntax and semantic means of sentence structure and meanings respectively [2].

MT systems can be sub-language MT or it can be general-purpose [47]. Sub-language MT systems are designed particularly for some specific domains for some specialized purposes. The specialized language is referred to as a sub-language. A sub-language is used by experts in certain fields of area for communication purposes. It contains words that are only known by those experts of that specific field of study or words which can be used in different ways. Sublanguages are also characterized by special grammatical patterns. The general purpose MT systems are designed for translation of texts and speech from the entire domain without any domain restriction. MT systems can be bilingual systems or multilingual systems depending on the number of languages involved in the translation process. Bilingual systems are designed specifically for two languages (single pair of languages) and multilingual systems are designed for more than two languages. The translation can be unidirectional or bidirectional [47]. In case of unidirectional, the system translates from the source language into the target language only in one direction. Bidirectional systems work in both directions in a way that one language can act as a source language or a target language. Bilingual systems can be unidirectional or they can be bidirectional, but multilingual systems are usually designed to be bidirectional.

MT has its own advantage in allowing communication between users who speak different languages which advances globalization of the information highway [24]. Socio-political importance is one of them. Many countries are well known for having communities with multiple languages. The social or political importance of MT, according to Arnold (1995), arises from the importance of translation of concepts from one language to another and to keep the social and political stability of the country. From another perspective, the other issue is the vanishing of unique culture associated with the language speakers and the language itself, and the way of thinking will matter to society. Hence we can conclude that translation is the remedy to avoid these problems by facilitating the ordinary human transaction and for gathering the information one needs to play a full part in society. The commercial importance of MT is a result of factors that are directly related to its social and political importance. To justify some of the commercial importance of MT the following can be considered as fact to the idea. One can select a product without language constraint (for example, if we need material to a specific topic in the

language of Amharic, and if the best material is available in English, we can take it and it can be translated to the language of our interest).The scientific importance of MT is an intellectual exercise application and testing ground for many ideas in Computer Science, Artificial Intelligence, and Linguistics, and some of the most important developments in these fields have begun in MT. This is illustrated using the first widely available logic programming language called Prolog. Prolog was formed as a key part of the Japanese "Fifth Generation" program of research in the late 1980s, can be found in the 'Q-systems' language, originally developed for MT. Philosophically, MT represents an attempt to automate an activity that can require the full range of human knowledge that is, for any piece of human knowledge, it is possible to think of a context where the knowledge is required. Scientists on MT agree that the extent to which one can automate translation is an indication of the extent to which one can automate ' thinking'. Nowadays, there are several MT systems that are used in day-to-day use around the world. Some of these are METEO (since 1977 used at the Canadian Meteorological Centre in Dorval, Montreal), SYSTRAN (in use at the CEC and elsewhere), LOGOS, ALPS, ENGSPAN (and SPANAM), METAL, and GLOBALINK [71].

2.4.1 History of Machine Translation

Interest in automatic MT started in the late forties after World War II and it may be dated to a memorandum written in March 1947 from Warren Weaver of the Rockefeller Foundation to cyberneticist Norbert Wiener [75] which contains the following two sentences as has been taken from [109].

“I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text.” Warren Weaver

From the mid-1970s onwards, the demand for MT came from quite different sources with different needs and different languages. The demand was towards cost-effective machine-aided translation systems that could deal with commercial and technical documentation in the principal languages of international commerce.

The 1980s witnessed the emergence of a wide variety of MT system types, and from a widening number of countries. Examples include Systran operating in many pairs of languages, Logos

(German-English and English French), the internally developed systems at the Pan American Health Organization (Spanish-English), the Metal system (German-English), and major systems for Japanese-English translation from Japanese computer companies. The end of the 1980s was a major turning point for MT for two reasons. Firstly, a group from IBM published the results of experiments on a system called Candide based purely on statistical methods. Secondly, certain Japanese groups began to use methods based on corpora of translation examples, i.e. using the approach now called ‘example-based’ translation. In both approaches, the distinctive feature was that no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents; both approaches differed from earlier ‘rule-based’ methods in the exploitation of large text corpora instead of using a set of hard-coded rules [23].

In the 1990s, the use of MT and translation aids by large corporations has grown rapidly. The first commercial MT system for Russian/English/German-Ukrainian is developed at Kharkov State University (1991). MT on the web starts with Systran offering a free translation of small texts (1996), followed by AltaVista Babelfish, which racked up 500,000 requests a day (1997). A particularly impressive increase is seen in the area of software localization (i.e. the adaptation and translation of equipment and documentation for new markets). On the research front, the principal areas of growth are seen in example-based and statistical MT approaches, in the development of speech translation for specific domains, and in the integration of translation with other language technologies [23], [58].

In 2000s, the move is towards combining the rule-based and SMT paradigms with the goal of improving the quality of the output as well as the performance of the system. Recent years have seen significant step advancements in MT technology with Google’s research on Neural MT implying an optimistic future for the industry. It has become clear that MT is moving away from being the high speed, untenable quality option for translating organizations, toward offering a reasonable alternative for translating low visibility content. In just a few years, a whole host of MT vendors have emerged promising acceptable quality at a fraction of the cost of professional linguists. The race for a competitive edge in quality is in full effect and MT providers are beginning to take differentiated approaches to ‘boost’ the quality that their systems are capable of producing [58], and [23].

2.4.2 Approaches to Machine Translation

MT is one of the research areas under computational linguistics. Various methodologies have been devised to automate the translation process. However, the objective has been to restore the meaning of the original text in the translated verse. In general, the process of translation has two levels known as Metaphrase and Paraphrase. Metaphrase means word-to-word translation. It relates to formal equivalence, i.e., the translated version will have a literal translation for each word in the text. However, the translated text may not necessarily convey the meaning of the original text. That means sometimes the semantics may differ from the original text. Paraphrase relates to dynamic equivalence, i.e., the translated text would contain the gist of the original text but may not necessarily contain the word-to-word translation [59]. Different methods of MT are explained in the following sections.

Dictionary based Machine Translation

This method of translation is based on entries of a language dictionary. The word's equivalent is used to develop the translated verse. The first generation of MT (the late 1940s to mid-1960s) was entirely based on machine-readable or electronic dictionaries. To some extent, this method is still helpful in translation of phrases but not sentences. Most of the translation approaches developed later on more or less utilize bilingual dictionaries with grammatical rules [59].

Rule-based Approaches

Rule-based MT also is known as “Knowledge-based MT”, systems are fundamentally based on formulated rules for translation which is based on linguistic information about the source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. It consists of a collection of rules called grammar rules, lexicon, and software programs to process the rules. However, According to [47] there are three classical systems categorized by how they perform translation, namely: direct approach, transfer approach, and Interlingua approach. The following sections briefly discuss the different approaches used in rule-based MT.

Direct approach

Words of Source Language are translated without passing through an additional/intermediary representation. It carries out word-by-word translation with the help of a bilingual dictionary usually followed by some syntactic rearrangement. Due to this direct mapping, such systems are highly dependent on both the source and target languages [64] Anusaarka is a MT system based on a direct approach. It has been developed at the Indian Institute of Information Technology, Hyderabad, and covers all major Indian languages. This approach needs only a little syntactic and semantic analysis and it is basically bilingual and unidirectional [72], and [59].

Transfer based

The transfer model belongs to the second generation of MT (mid-60s to 1980s). In this, the source language is transformed into an abstract, less language-specific representation. An equivalent representation (with the same level of abstraction) is then generated for the target language using bilingual dictionaries and grammar rules. These systems have three major components: Analysis, Transfer, and Synthesis

Analysis of the source text is done based on linguistic information such as morphology, part-of-speech, syntax, semantics, etc. Heuristics as well as algorithms are applied to parse the source language and derive either the syntactic structure (for language pair of the same family) of the text to be translated; or the semantic structure (for language pair of different families). Transfer The syntactic/semantic structure of the source language is then transferred into the syntactic/semantic structure of the target language. In Synthesis the module replaces the constituents in the source language to the target language equivalents. This approach, however, has a dependency on the language pair involved. Thus, two independent monolingual dictionaries were suggested in the Eurotra project. Also, there are different representations for different languages. PaTrans (Translation for Patents) is based on a transfer-based approach and is one of the outcomes of Eurotra Research. Mantra is also a translation model for Indian Languages based on the transfer approach. It is a Government of India funded project and the parser used for language processing is known as Vyakarta[59].

Interlingua

This is considered to belong to the third generation of MT. It is an inherent part of a branch called Inter-linguistics. Interlingua aims to create linguistic homogeneity across the globe.

Interlingua is a combination of two Latin words *Inter* and *Lingua* which means between/intermediary and language respectively. In Interlingua, the source language is transformed into an auxiliary/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 auxiliary representation. Hence, only two modules i.e., analysis and synthesis are required in this type of system. Also, because of its independence on the language pair for translation, this system has much relevance in multilingual MT. This emphasizes on single representation for different languages. The main advantage of the interlingua approach is that it creates an economical multilingual environment that requires $2n$ translation systems to translate among n languages wherein the other case, the direct approach requires $n(n-1)$ translation systems.[73] and [59]

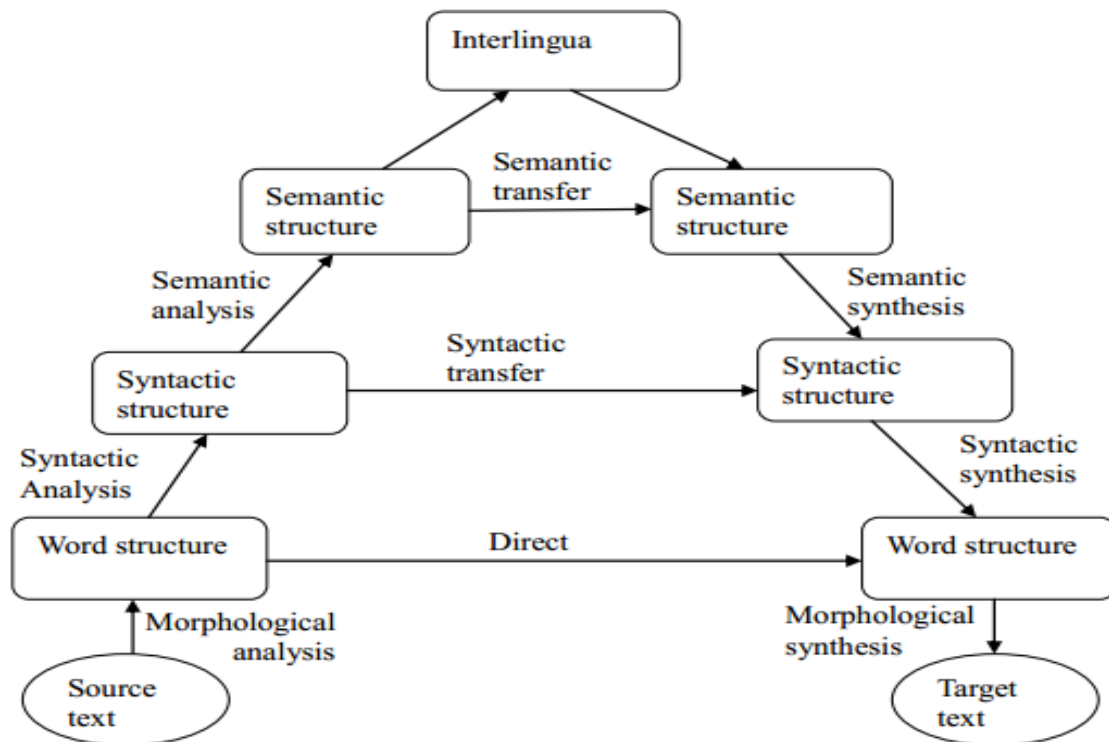


Figure 0.2 The Vauquois triangle for rule-based MT systems [12]

Corpus-based Machine Translation Approach

The corpus-based (Empirical) MT has been dominating the traditional rule-based (Classical) ones since the late 1980s. This system is data-driven as opposed to rule-driven. The rule-based has been requiring human encoded linguistic knowledge and intensive representation of the

languages through different structural and language rules. The relative failure of rule-based approaches, the growing availability of machine-readable parallel corpus (collection of source language document with its counterpart target language documents), and the increase in the capability of hardware (CPU, memory, disk space) with decreasing in cost are among the critical factors for the flourishing of corpus-based MT systems[73]. Hence, corpus-based MT is based fundamentally on the principle of using existing translations as a prime source of information for the production of new ones (i.e., it believes in the fact that large amounts of data contain essential knowledge for making a functional system [47]. Despite the rule-based models that require explicit linguistic knowledge, data-driven ones rectify the lack of such knowledge in such a way that the knowledge can be retrieved and used automatically [23]. CBMT approach is further classified into two major approaches: Statistical MT (SMT) and Example-Based MT (EBMT) approach.

Example-based Machine Translation

Example-based translation (also known as Memory based translation) is based on recalling/finding analogous examples (of the language pairs). This concept of “Translation by Analogy” was first proposed by Makoto Nagao in 1981[75]. An Example-Based MT (EBMT) system is given a set of sentences in the source language (from which one is translating) and corresponding translations of each sentence in the target language with a point to point mapping. These examples are used to translate a similar type of sentence of source-language to the target language. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. Advantages of an EBMT system over SMT system as put forth by [74] are:

- This can work with a small set of data (even with one sentence pair)
- Trains translation program and decodes more quickly
- Less principled (at least in theory)

Statistical Machine Translation (SMT)

Warren Weaver, in 1949, had introduced the idea of Statistical MT (SMT). In this, statistical methods are applied to generate a translated version using bilingual corpora. Example: n-gram based SMT⁸; Occurrence based SMT⁹, etc. Macherey¹⁰ has experimented statistical methods for spoken language understanding for SMT. the goal of statistical MT is to translate a source language sequence into a target language sequence by maximizing the posterior probability of the target sequence given the source sequence. Probabilities that describe correspondences between the words in the source language and words in the target language are learned from a bilingual parallel corpus and language models are learned from a monolingual text in the target language. As the available training corpus becomes large, the performance of the system increases. Statistical MT tries to generate translations using statistical methods based on bilingual text corpora. Where such corpora are available, impressive results can be achieved by translating texts of a similar kind, but such large corpora are still very rare. The general architecture of SMT includes three components: language model, translation model, and decoder. The language model ensures that words come in the right order [2], [59], [76].

Statistical word-based translation model: in this model the Fundamental unit is Word and the Reordering is Algorithms related to the alignment of words are required to achieve utmost accuracy in sentence translation. Compound words, idioms, homonyms create complexity for simple word-based translation. In a Statistical phrase-based model [77], [78] the Fundamental unit is a phrase or sequence of words. A sequence of words in the source and the target language is developed and Decoding is done based on the vector of features with matching values for the language sequence pair. The statistical syntax-based model has a Fundamental unit the translation rule. Translation rule consists of a sequence of words and variables in the source language, a syntax tree in the target language (having words or variables at leaves), and a vector of feature values that describes the language pair's likelihood [79], [30].

Hybrid Machine Translation Approach (Hybrid MT)

The Hybrid MT (HMT) approach combines the strength of both the statistical MT approach and that of a Rule-based approach to get better results of translation accuracy. The system combines linguistic information and statistical information. By doing this, the problems of ambiguity on the rule-based system are solved by the statistical part and complex sentences, which are difficult for SMT, are solved by rule-based systems. In order to implement MTs by using HMT, we have to select from two well-known approaches. The first approach is rule post-processed by statistics

in which translations are performed by using rules of both languages like syntax rule of categorizing word into either subject or verb or object, and statistics are then used to adjust the candidate words in the correct order of target language sentence [81].

The second HMT approach is statistics guided by rules in which rules are used to preprocess data by grouping words into its class of either subject or verb or object groups, while statistics are used to generate the correct translation of words based on a statistical computation of matching word and latter rules are applied on candidate words to drive translated target language sentence. Therefore, the selection between these two approaches of HMT translation depends on whether statistics is preferred for word translation or rules are preferred for translation of each word of language under study [81].

Neural Machine Translation (NMT) approach

Neural MT or NMT is a newly emerging approach to MT, recently proposed by [83], [84], [85]. At the beginning of neural network MT, a neural network was introduced as a supportive tool for SMT to facilitate the computation of statistical probability that was assigned to each word in a sequence. After further work of different authors, the pure neural network MT was modeled with the idea of jointly training and translating of data from one language to another language [9], [10]. Unlike the traditional phrase-based translation system which consists of many small sub-components that are tuned separately, neural MT attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation [82]. As such, neural MT systems are said to be end-to-end systems as only one model is required for the translation in which its strength lies in the mapping from input text to associated output text.

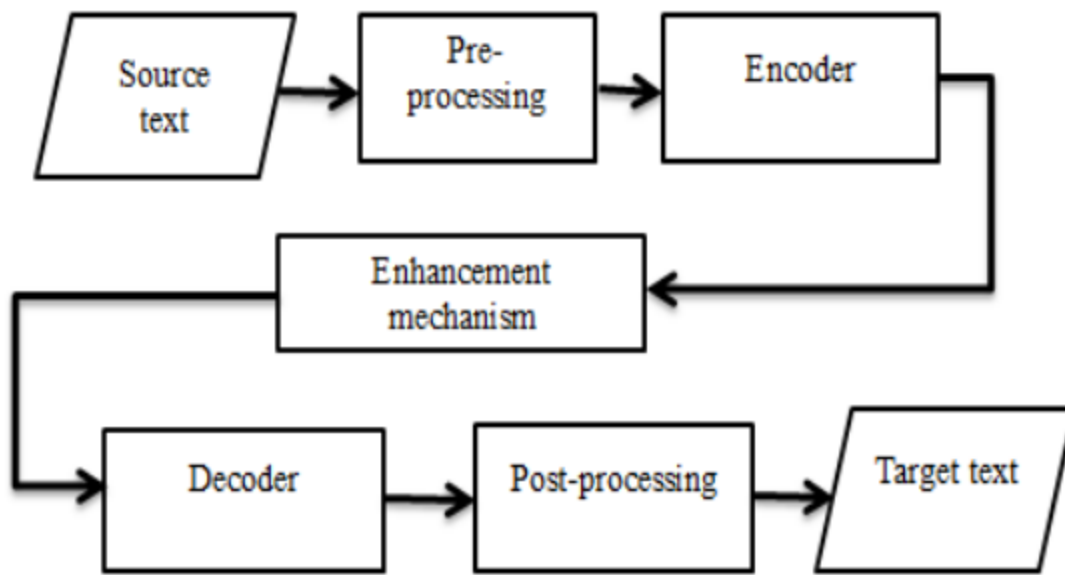


Figure 0.3 Design of Encoder-Decoder language modeling [85]

In recent times, neural MT becomes the application of deep neural networks (DNNs) to build end-to-end encoder-decoder models in which a translation system consists of subcomponents that are separately optimized [86]. Therefore, figure 2.3 shows this encoder-decoder language modeling when the source sentence pre-processed to be fed into the encoder, then the encoder generates a contextual relationship and feeds a data enhancement mechanism which separately shows the decoder the relevant context for and generates target sentence that matches the given context [85].

The neural MT varies in terms of the exact architecture of the neural network used for MT which is used to implement the encoder and decoder part of the encoder-decoder end-to-end language modeling [85]. The original sequential data must be changed into the form which is suitable to read for a neural network. This format is a word embedding format which changes sequential data into vector form representation. The encoder part receives this embedded form of data and generates a contextual relation between words. The decoder part receives the context generated by the encoder and searches for matching words within the target language. When the sequence includes a longer sentence, an enhancement mechanism like attention is needed to help to store the longer context within the sentence [87], [88].

Attention mechanism can solve problems in encoder-decoder approaches which is to recover performance degrade as the number of sequence length increases. The context within a sentence is derived as the inter-dependency of nearby words in a given sequence word in the sentence. But as the length of the sentence increase, the inter-dependency of words at the beginning of the sentence and the end of the sentence is loosely related. This problem also high in the statistical MT approach, as statistical MT is weak on the translation of longer sentences than the short sentence [87], [89].

The other problem found in the encoder-decoder mechanism is how to handle a larger number of vocabulary sizes available within the data. While each word in the sentence is visited, there must be assigned a new identity number to identify words by unique id number at the time it encountered in the data. But when the length of the dictionary increases, the number used for word representation becomes higher and the dimension of the word vector needed becomes higher. So, adding an attention mechanism to encoder-decoder language modeling solves this problem by minimizing the changing of a higher dimensional vector into a lower dimension vector [88], [90].

2.4.3 Evaluation of Machine Translation

MT is the task to translate a text from a source language to a target language. As MT emerges as an important mode of translation, its quality is becoming more and more important. Evaluating MT results lacks an appropriate, consistent, and easy to use criterion [91]. The quality of a translation is inherently subjective; there is no objective or quantifiable good. Therefore, any metric must assign quality scores so they correlate with the human judgment of quality. That is, a metric should score highly translations that humans score highly, and give low scores to those humans give low scores. Human judgment is the benchmark for assessing automatic metrics, as humans are the end-users of any translation output. The measure of evaluation for metrics is a correlation with human judgment. This is generally done at two levels, at the sentence level, where scores are calculated by the metric for a set of translated sentences, and then correlated against human judgment for the same sentences. And at the corpus level, where scores over the sentences are aggregated for both human judgments and metric judgments, and these aggregate scores are then correlated [92]. Some automatic evaluation methods are discussed below.

BLEU

BLEU was one of the first metrics to report a high correlation with human judgments of quality. The metric is currently one of the most popular in the field. The central idea behind the metric is that "the closer a MT is to a professional human translation, the better it is"[93].The metric calculates scores for individual segments, generally sentences—then average these scores over the whole corpus for a final score. It has been shown to correlate highly with human judgments of quality at the corpus level [94]. BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations. The metric modifies simple precision since MT systems have been known to generate more words than appear in a reference text. No other MT metric is yet to significantly outperform BLEU with respect to correlation with human judgment across language pairs [95].

NIST

The NIST metric is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST also calculates how informative a particular n-gram is. That is to say, when a correct n-gram is found, the rarer that n-gram is, and the more weight it is given [96]. For example, if the bigram "on the" correctly matches, it receives lower weight than the correct matching of bigram "interesting calculations," as this is less likely to occur. NIST also differs from BLEU in its calculation of the brevity penalty, insofar as small variations in translation length do not impact the overall score as much [92].

Word error rate (WER)

The Word error rate (WER) is a metric based on the Levenshtein distance, where the Levenshtein distance works at the character level, WER works at the word level. It was originally used for measuring the performance of speech recognition systems but is also used in the evaluation of MT. The metric is based on the calculation of the number of words that differ between a piece of machine-translated text and a reference translation. A related metric is the Position-independent word error rate (PER), which allows for re-ordering of words and sequences of words between a translated text and a reference translation [92].

METEOR

The METEOR metric is designed to address some of the deficiencies inherent in the BLEU metric. The metric is based on the weighted harmonic mean of unigram precision and unigram

recall. The metric was designed after research by Lavie (2004) into the significance of recall in evaluation metrics. Their research showed that metrics based on recall consistently achieved higher correlation than those based on precision alone, cf. BLEU and NIST [97]. METEOR also includes some other features not found in other metrics, such as synonymy matching, where instead of matching only on the exact word form; the metric also matches on synonyms. For example, the word "good" in the reference rendering as "well" in the translation counts as a match. The metric also includes a stemmer, which lemmatizes words and matches on the lemmatized forms. The implementation of the metric is modular insofar as the algorithms that match words are implemented as modules, and new modules that implement different matching strategies may easily be added [92].

LEPOR

A new MT evaluation metric LEPOR was proposed as the combination of many evaluation factors including existing ones (precision, recall) and modified ones (sentence-length penalty and n-gram based word order penalty). The experiments were tested on eight language pairs from ACL-WMT2011 including English-to-other (Spanish, French, German, and Czech) and the inverse, and showed that LEPOR yielded higher system-level correlation with human judgments than several existing metrics such as BLEU, Meteor-1.3, TER, AMBER and MP4IBM1 [98]. An enhanced version of LEPOR metric, hLEPOR, is introduced in paper [99]. hLEPOR utilizes the harmonic mean to combine the sub-factors of the designed metric. Furthermore, they design a set of parameters to tune the weights of the sub-factors according to different language pairs. The ACL-WMT13 Metrics shared task[100] results show that hLEPOR yields the highest Pearson correlation score with human judgment on the English-to-Russian language pair, in addition to the highest average-score on five language pairs (English-to-German, French, Spanish, Czech, Russian).

2.5 System and Language Modeling in NMT

2.5.1 System Modeling (Encoder-Decoder)

With the growing importance of global communications and international business relations, new and more efficient translation methods have been developed. The latest technological advancements have brought more opportunities for automated language translation. These

innovative methods work towards eliminating the language barriers businesses oftentimes face when communicating with foreign partners. Most recently, a new approach to MT has emerged, NMT, which uses Deep Learning to translate text. The new technology not only uses a fraction of the memory needed by its predecessor, SMT, but it is also modeled around the neural frameworks of the human brain. By employing both deep learning and representation learning, NMT allows for contextually precise and fast translation of whole sentences, reducing translation errors by 60%. NMT has already been adopted by the best MT service providers, including Microsoft, Google, and Yandex [114].

Because of the increased implementation of NMT within various translation apps and services, the technology is able to translate text from grammatically complex languages, by learning and utilizing their specifics. Unlike SMT, using tedious steps like preparation of language modeling, preparation of translation modeling, tuning and decoding steps encoder-decoder based machine translation became the better choice for the simplicity of the steps of modeling. In encoder-decoder modeling, the steps are interconnected. Therefore, human intervention at training time is not necessary like that of statistical approach. NMT is considered more effective in handling word ordering, morphology, and syntax. The technology comes with the promise of cost-effective translation for under-resourced languages, which makes it beneficial for all businesses, regardless of their location or spoken language.

In recent time, the Encoder-Decoder end-to-end Language Model (LM) is becoming attractive LM for machine translation task which is based on a deep learning algorithm. Sequence-to-sequence learning (Seq2Seq) is about training models to convert sequences from one domain (e.g. sentences in Amharic) to sequences in another domain (e.g. the same sentences translated to Afaan Oromo). There are multiple ways to handle this task, either using RNNs or using 1D convnet. A natural choice for sequential data is the recurrent neural network (RNN), used by most of the recent NMT work and for both the Encoder and Decoder. The used RNN models, however, differ in terms of (a) directionality – unidirectional or bidirectional; (b) depth – single or deep multi-layer; and (c) type – often either a vanilla RNN, an LSTM, or a gated recurrent unit (GRU). For the Encoder, almost any architecture can be used since we have fully observed the source sentence. Choices on the Decoder side are more limited since we need to be able to generate a translation.

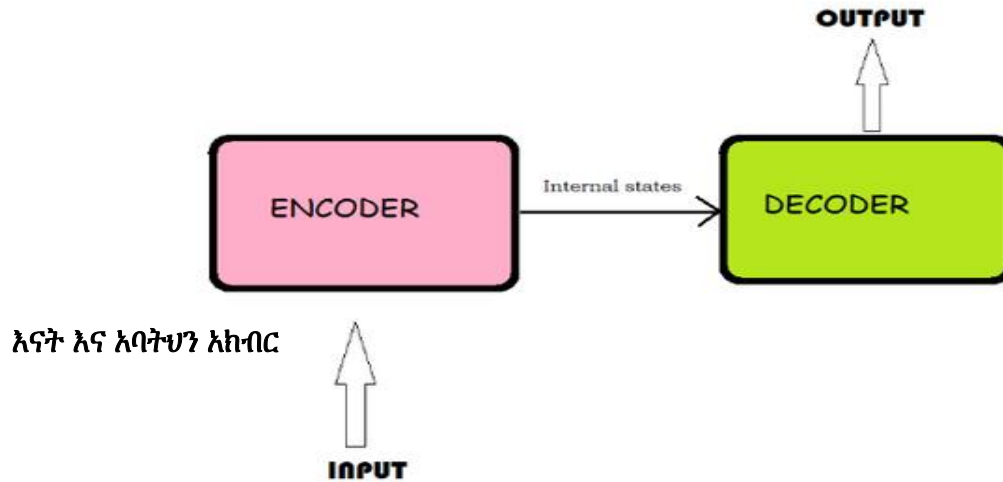


Figure 0.4 System Modelling

2.5.2 Language Modeling

Language modeling (LM) is the task of assigning a probability to sentences in a language [105] and it provides context to distinguish between words and phrases that sound similar. Estimating the relative likelihood of different phrases is useful in many NLP applications, especially those that generate text as an output. Language modeling is used in speech recognition, MT, part of speech tagging, parsing, optical character recognition, hand writing recognition, information retrieval, and other applications. Data sparsity is a major problem in building language models. Most possible word sequences are not observed in training. One solution is to make the assumption that the probability of a word only depends on the previous n words. This is known as an n -gram model or unigram model when $n = 1$. The unigram model is also known as the bag of words model. N -gram was a traditional language model which can only handle short contexts of about 4 to 6 words and does not generalize well to unseen n -grams. In A statistical language model a probability distribution is taking over in sequences of words. Given such a sequence, say of length m , it assigned a probability to the whole sequence

According to the work of different researchers, in the early emergence of NMT, the RNN was used to generate word co-occurrence probabilities for machine translation tasks. Besides assigning a probability to each sequence of words, language model also assigns a probability for the likelihood of a given word (or a sequence of words) to follow a sequence of words [106].

Neural language models (or continuous space language models) use continuous representations or embedding's of words to make their predictions. These models make use of neural networks. Continuous space embedding's help to alleviate the curse of dimensionality in language modeling: as language models are trained on larger and larger texts, the number of unique words (the vocabulary) increases. The number of possible sequences of words increases exponentially with the size of the vocabulary, causing a data sparsity problem because of the exponentially many sequences. Thus, statistics are needed to properly estimate probabilities. Neural networks avoid this problem by representing words in a distributed way, as non-linear combinations of weights in a neural net.

An alternate description is that a neural net approximates the language function. The neural net architecture might be feed-forward or recurrent, and while the former is simpler the latter is more common. The difference between Recurrent and feed-forward neural networks, in feed-forward networks, history is represented by context of $N - 1$ word - it is limited in the same way as in N -gram back-off models. In recurrent networks, history is represented by neurons with recurrent connections - history length is unlimited. Also, recurrent networks can learn to compress whole history in low dimensional space, while feed-forward networks compress (project) just single word. Recurrent networks have possibility to form short term memory, so they can better deal with position invariance; feed-forward networks cannot do that. In RNN as history length is unlimited when we have a high dimensional space our dependencies become longer; layers in the unrolled RNN also increase. As the network becomes deeper, the gradients flowing back in the back-propagation step become smaller. As a result, the learning rate becomes slow and makes it infeasible to expect long term dependencies of the language. And we used the GRU unit both in Encoder and Decoder part as the solution to this issue.

The GRU by [107] is the most widely used solution for gradient vanishing problem which occurs when gradient values start approaching zero as we BPTT. GRU networks are just an advanced version of plain RNNs that we discussed above. These networks are designed to remember information for long periods without having to deal with the vanishing gradient problem. In the GRU cell, the vanishing gradient problem is solved by writing the current state as a memory of the network. This writing process is a regulated reset gates. The difference between LSTM (Long Short Term Memory) and GRU is that LSTM has three gates (input, output and forget gate) and

GRU has two gates (reset and update gate). GRU couples forget as well as input gates. GRU use less training parameters and therefore use less memory, execute faster and train faster than LSTM's whereas LSTM is more accurate on dataset using longer sequence.

The standard Seq2Seq model is generally unable to accurately process long input sequences since only the last hidden state of the Encoder RNN is used as the context vector for the Decoder [115], [116]. The Attention Mechanism directly addresses this issue as it retains and utilizes all the hidden states of the input sequence during the decoding process. It does this by creating a unique mapping between each time step of the Decoder output to all the Encoder hidden states. This means that for each output that the Decoder makes, it has access to the entire input sequence and can selectively pick out specific elements from that sequence to produce the output [117].

CHAPTER 3: RELATED WORKS

3.1 Introduction

There are different studies conducted on MT approaches, strategies, techniques, and implementations that have been documented. In this chapter, we try to review works that have been done in MT in the following subtopic.

3.2 Machine Translation for foreign language pairs

Pure rule-based English to Tamil translation was developed by M.Kasthuri et.al. in [34]. English and Tamil follow different word ordering; the subject-object-verb pattern is followed by Tamil and subject-verb-object by English. This is the main reason for the work to follow the syntax transfer-based approach. A parser is used as the translation engine of the proposed system that analyzes English text and then by using the transfer lexicon target Tamil text is generated. The proposed rule-based translation system is composed of five core modules. Root words and feature equations of the source text are obtained by a morph analyzer which is the first module of the system. Parts of speech tagger and word sense disambiguation is the second module. The tagger assigns word class to each word of the source sentence and the disambiguation process identifies in which sense a word was meant in the given text. The parser is the third module for analyzing source text. From a single source structure, one or more target sentences are generated from the fourth module, i.e., target generator. The fifth and last module is the morph generator that handles target text morphology.

Nithya Bet. al. in [101] develop A Hybrid Approach to English to Malayalam MT. This hybrid approach extends the baseline statistical machine translator with a translation memory. A statistical machine translator performs translation by applying machine learning techniques on the corpus. The translation memory caches the recently performed translations in memory and eliminates the need for performing redundant translations. The proposed system has two main components. They are statistical machine translator and translation memory. In the statistical component Language model, the Translation Model and Decoding are done and A TM should have four main components. They are a mechanism to store sentences and their translations, a search mechanism to find input sentence matches from TM, Provision for post-editing the translator output, and Provision for updating the TM The input sentence to be translated is first searched in the TM. If the sentence has a match in the TM, then the translation is performed by directly copying the TM match's translation. But if there is no match available in TM, the input is sent to the SMT system. The SMT system performs translation and provides a rough output. This output can be modified by a human expert in case it is not perfect. The final output is fed to the TM too so as to cache the translation. The system was tested with a test set consisting of 70 English sentences. Both manual and automatic evaluation techniques were employed to measure the efficiency of the new approach and the work achieved 69.33% in the BLEU score.

A Novel Approach for English to South Dravidian Language Statistical MT System by [102] is developed motivated by the fact that even though there are efforts towards building such as English to South Dravidian translation system, unfortunately, there is no an efficient translation system till now so to achieve the modeling of the research using statistical approach The first and most important step is creating a well-aligned parallel corpus for training the system As experimental research shows that the existing methodology for bilingual parallel corpus creation is not efficient for English to South Dravidian language in the SMT system. In order to increase the performance of the translation system, the researchers have introduced a new approach to creating a parallel corpus. The main ideas which the authors have implemented and proven very effective for English to south Dravidian languages SMT system are: (i)reordering the English source sentence according to Dravidian syntax, (ii) using the root suffix separation on both English and Dravidian words and iii) use of morphological information which substantially reduce the corpus size required for training the system. Since the unavailability of full-fledged parsing and morphological tools for Malayalam and Kannada languages, sentence synthesis was

done both manually and the existing morph analyzer created by Amrita University. From the experiment, the authors found that the performance of the systems is significantly well and achieves a very competitive accuracy for small-sized bilingual corpora. The training and testing sentence size was limited to a maximum of twelve words and According to the structure of sentence, the researchers have written reordering rules. The performance of the translation system was evaluated with BLEU evaluation metric and score 24.9 % for Malayalam and 24.5 % Kannada languages.

Charu Verma et.al.in Hindi-English Neural MT Using Attention Model [103] developed Neural MT by jointly learning to Align and Translate. Here, attention reads as a neural extension of the Encoder-Decoder model. The Encoder-Decoder model contains several limitations which are resolved by Attention. Neural network work on vectors, so it compresses all important information of source sentence in encoder-decoder approach this make neural network difficult to work with long sentences, for mainly those sentences which are longer compared to training corpus sentence. they used BLEU metric to evaluate the performance of their MT system. they also performed human evaluation using adequacy measures on a 5 point score. They did their evaluation on 500 sentences which were divided into 5 documents of 100 sentences each. The results of BLEU score are 0.375784 in baseline NMT and 0.552926 in NMT with the attention model.

D. Do, et.al. developed “MT from Japanese and French to Vietnamese” [104]. This study was done by using the SMT approach. They also conducted the experiments on parallel corpora collected from TED talks. They used phrase-based and tree-to-string models and have shown that the SMT system trained on French to Vietnamese obtains better results than the system of Japanese to Vietnamese because French and Vietnamese have more similarities in the structures of sentences than between Japanese and Vietnamese.

Ilya Sutskever et. al. conducted research in “English to French MT using Sequence to Sequence Learning with Neural Networks” [105]. The study was carried out by using the Neural MT approach. They applied a Deep Neural Network (DNN) approach to the previous study carried out in a phrase-based SMT approach. The model they used is a recurrent neural network language model. They used a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of fixed dimensionality, and then another deep LSTM to decode the target

sequence from the vector. They used WMT'14 English to the French dataset. They trained their models on a subset of 12M sentences consisting of 348M French words and 304M English words, which is a clean “selected” subset from [106]. They chose this translation task and this specific training set subset because of the public availability of a tokenized training and for comparing the performance from the baseline SMT [106]. They evaluated their models using the standard BLEU score metric. On the WMT'14 English to french translation task, they obtained a BLEU score of 34.81 using a simple left-to-right beam-search decoder. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset [94]. This result shows that neural network architecture outperforms a phrase-based SMT system. When they reverse the order of the words in all sources, the BLEU score increased to 36.5. Finally, they found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier. Additionally, they confirmed LSTM did not have difficulty in long sentences.

Yonghui Wu et. alin [107] conducted research in “Google's Neural MT System: Bridging the Gap between Human and MT”. The study was done by using the Neural MT approach for the purpose of overcoming many of the weaknesses of previous Google's conventional phrase-based translation systems. In addition to this, they improved the NMT robustness problem (particularly when input sentences contain rare words) which is stated as a problem by many previous researchers [105] and by the authors of 'N-gram counts and language models from the common crawl'. Their model is a sequence-to-sequence learning framework with attention. It has three components: an encoder network, a decoder network, and an attention network. It consists of a deep LSTM network with 8 encoder and 8 decoder layers. The encoder transforms a source sentence into a list of vectors, one vector per input symbol. Given this list of vectors, the decoder produces one symbol at a time, until the special end-of-sentence symbol (EOS) is produced. A decoder is implemented as a combination of an RNN network and a softmax layer. The encoder and decoder are connected through an attention module which allows the decoder to focus on different regions of the source sentence during the course of decoding. To improve parallelism and therefore decrease training time, their attention mechanism connects the bottom layer of the decoder to the top layer of the encoder. To improve the handling of rare words, they divided words into a limited set of common sub-word units (“word pieces”) for both input and output.

They followed a beam search technique for implementation purposes. For testing their system, they used WMT¹⁴ English-to-French and English-to-German benchmarks as a dataset. They evaluated their models using the standard BLEU score metric. Specifically, on WMT¹⁴ English-to-French, their single model scores 38.95 BLEU, an improvement of 7.5 BLEU from a single model without an external alignment model reported in ‘Addressing the rare word problem in NMT’ and an improvement of 1.2 BLEU from a single model without an external alignment model from that research. Likewise, on WMT¹⁴ English-to-German, their single model scores 24.17 BLEU, which is 3.4 BLEU better than a previous competitive baseline done in ‘N-gram counts and language models from the common crawl’. They also reported as, on production data, their implementation is even more effective. Finally, they reported as human evaluations show that their system has reduced translation errors by 60% compared to their previous phrase-based system on many pairs of languages: English↔French, English↔Spanish, and English↔Chinese. Additionally, their experiments suggest the quality of the resulting translation system gets closer to that of average human translators.

In [40] the authors developed Attention-based English to Punjabi neural MT. To build a decent NMT system, they crawled data from various freely available websites. The entire corpus was further cleaned and grammatical errors/typos were corrected. They restricted every sentence of their parallel corpus to have at least 4 words. They have used a test and development dataset of size 488 and 300 sentences respectively. To evaluate the system, BLEU scores on tokenized translations were used. They have tried two variants of parallel corpora; one with simple tokenized sentences and another with Byte Pair Encoding (BPE). The BPE was learned on the concatenation of train data with 40000 merge-operations. Finally, they got BLEU score of 24.48 and 26.07 for tokenized and BPE corpora respectively.

3.3 Machine Translation involving Ethiopian language

Eleni Teshome in [2] develop Bidirectional English-Amharic MT and conduct the experiment by using Constrained Corpus the main problem that the author of this paper states was that there is only one system, English Amharic Statistical MT, is being developed in Ethiopia. For the fact that there is a need for a MT because people use human translation and they tend to be slower as compared to machines. Sometimes it can be hard to get a precise translation that reveals what the text is about without everything being translated word-by-word. In addition, it can be more

important to get the result without delay which is hard to accomplish with a human translator. That is when MT comes in, which solves most of the problems caused by a human translator. And even English-Amharic MT was developed as 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 .the general objective of the research was to design and develop a bidirectional English Amharic MT system using the constrained corpus. The scope of the study is restricted to the implementation of bidirectional translation on the languages English and Amharic using the constrained corpus. It is called “Constrained” because some of the corpus used is prepared manually and the other is collected and examined carefully. The translation was performed mainly on simple sentences and based on those sentences it experiments on how the system translates from English to Amharic as well as Amharic to English. It was also tested on complex sentences to see and identify its applicability. Experiments were carried out based on the dataset and results were recorded. The experiments were taken separately, one for the simple sentences and the other for complex sentences. The result obtained for the simple sentence using BLEU Score had an average of 82.22% accuracy for the English to Amharic, 90.59% for the Amharic to English and using the manual questionnaire preparation method, the accuracy from English to Amharic was 91% and from Amharic to English was 97%. For the complex sentences, the result acquired from the BLEU Score was approximately 73.38% for the English to Amharic, 84.12% for the Amharic to English, and from the questionnaire method from English to Amharic was 87% and from Amharic to English was 89%. From this, the researcher can see that the difference with the BLEU score and the questionnaire preparation method is not that visible so we can use both methods as reference. As a result, the researcher assumes that with a corpus that is very large and appropriately examined, a better translation could be achieved since more words will be available in the provided corpus with a higher probability of a particular word preceding another. The limitation of the study was the prototype parses only simple Amharic sentences that are not more than four words and it uses a small sample prepared corpora due to a lack of large annotated corpora in the language pairs.

Samrawit Zewgneh in [108] developed English-Amharic Document Translation Using a Hybrid Approach based on the problem faced by previously done English-Amharic MT using SMT by Ambaye and Eleni [2] which is the problem of accuracy and understandability for complex

sentences. The main objective of this study is to design and develop the English-Amharic document translation system by using a hybrid approach. This research achieved 15% and 20% accuracy improvement for simple and complex sentences over the statistical MT approach. This study also identified preprocessing the inputs of SMT is more suitable to improve accuracy for complex sentences while post-processing the outputs of SMT is more suitable for simple sentences.

Sisay Adugna in [23] develops English–Afaan Oromo MT by using a statistical approach the researcher is motivated by as here are abundant documents in English on the Internet lack of English language knowledge creates a problem of fully utilizing these documents. The researcher believes that studying how to make these documents available in local languages (such as Afaan Oromo) is vital in addressing the language barrier thereby reducing the effect of the digital divide as no research has been conducted on English-Afaan-Oromo MT. The objective of the study was to develop a prototype English-Afaan Oromo MT system using a statistical approach, i.e., without explicit formulation of linguistic rules and by using a limited corpus of about 20,000 bilingual sentences, and the researcher achieves a translation accuracy of 17.74%.

Jabesa Daba and Yaregal Assabie in [31] developed Bidirectional English – Afaan Oromo MT using the hybrid approach initiated by the experiment which was conducted on English – Afaan Oromo language pair done by using statistical methodology[23]. The experiment was intended to translate English sentences into Afaan Oromo only in one direction which is still not a solution for the need of Afaan Oromo to English translation. The other thing is that the accuracy of the translation which has a BLEU score of 17.74% is not satisfactory. The general objective of this research work was to develop a bidirectional English – Afaan Oromo MT system using a hybrid approach. Since the system is bidirectional, two language models are developed; one for English and the other for Afaan Oromo. Translation models that assign a probability that a given source language text generates a target language text are built and a decoder that searches for the shortest path is used. Two major experiments are conducted by using two different approaches and their results are recorded. The first experiment is carried out by using a statistical approach. The result obtained from the experiment has a BLEU score of 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation. The second experiment is carried out by using a hybrid approach and the result obtained has a BLEU score of 37.41% for English to Afaan Oromo translation and 52.02% for Afaan Oromo to English translation.

From the result, the researcher deduces that the hybrid 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.

Arfaso Birhanu in [37] develops Bi-Directional English-Afaan Oromo MT Using Convolutional Neural Network by standing from the problem of the study of MT between English and Afaan Oromo was studied by using a statistical MT approach but has not been studied by using a neural network-based approach like that of MT between English and other European languages like German, French or Romanian. However, the neural network-based approach has better learning ability which captures and applies the features of human learning behavior and the previous studies done for translating English-Afaan Oromo language don't show satisfactory accuracy. The general objective of this study is to design and implement a bi-directional MT between English and Afaan Oromo based on a convolutional neural network. Three systems were implemented where the first system uses a word-based statistical approach that used as a baseline, while the second system with a recurrent neural network approach is used as a competitive model and lastly, the third system with convolutional neural networks for the bi-directional translation between Afaan Oromo and English languages the convolutional neural network achieved 3.86 BLEU score improvement on translation from English to Afaan Oromo and 3.32 BLEU score on translation from Afaan Oromo to English translation than baseline system. The researcher also deduces that the convolutional neural network approach has shown an improvement of 1.58 BLEU score on translation from English to Afaan Oromo and 1.51 BLEU score on translation from Afaan Oromo to English translation than the recurrent neural network approach and the convolutional neural network approach is faster on training than recurrent neural network approach.

Solomon Teferra et al. in [1] conduct research and develop Parallel Corpora for bi-lingual English-Ethiopian Languages Statistical MT motivated by the fact that In order to enable Ethiopians to use the documents and information produced in technologically favored languages, the documents need to be translated. Since manual translation is expensive, a promising alternative is the use of MT, particularly SMT as Ethiopian languages suffer from a lack of basic linguistic resources such as morphological analyzer, syntactic analyzer, morphological synthesizer, etc. The major and basic resource required for SMT is parallel corpora, which are not available for Ethiopian languages. The collection and preparation of parallel corpora for

Ethiopian languages is, therefore, an important endeavor to facilitate future MT research and development. The researchers presented some Ethiopian language researches conducted by graduate students and mainly raised the unavailability of linguistic resources stated by students which in turn affects the results that they obtain. The researcher collected and prepared parallel corpora for English and Ethiopian Languages that fall under the Semitic, Cushitic, and Omotic language families. They have considered Amharic, Tigrigna, and Ge'ez from the Semitic, Afaan-Oromo from the Cushitic, and Wolaita from Omotic language families. In the experimental setup, they used Moses with GIZA++ alignment tool for aligning words and phrases. SRILM toolkit was used to develop language models using semi-automatically prepared corpora from the training and tuning corpora of target languages. They achieved a BLEU score of 13.31 for English-Amharic translation while the Amharic-English has a 22.68. Similarly, the English-Tigrigna and Tigrigna-English have BLEU scores of 17.89 and 27.53, respectively. Likewise, English-Afaan Oromo has a 14.68 BLEU and Afaan Oromo-English has an 18.88 BLEU score. In a similar way, the English-Wolaita translation has BLEU of 10.49 while Wolaita-English has 17.39. Finally, The English-Ge'ez and Ge'ez-English translation has BLEU score of 6.67 and 18.01, respectively. Finally, they concluded as the English-Ethiopian languages SMT systems have less BLEU scores than that of Ethiopian languages-English ones. The reason they raised is the fact that when the Ethiopian languages are used as a target language, the translation from English as a source language is challenged by many-to-one alignment. They recommended investigating the effect of domains on SMT performance as to future work.

Mulubrhan Hailegebreal in [110] develops A Bidirectional Tigrigna – English Statistical MT by the motivation of the fact that Tigrigna is the official and working language of the Tigray region of Ethiopia and one of the official languages of Eritrea. And the English language is now the most widely used language in the world. So the Tigrigna speakers face a Lack of information and Communication gap and there is no prior investigation on the development of the MT system on the pair of English-Tigrigna or Tigrigna-English. As a result, the way of developing a MT 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. The general objective of this research is to design a bidirectional Tigrigna – English MT system

using a statistical approach. The work is done on MOSES statistical MT framework, and 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 that assign a probability that a given source language text generates a target language text are built and a decoder that searches for the shortest path is used. As a result, the researcher obtained a BLEU score of 53.35 % for Tigrigna – English and 22.46 % for English – Tigrigna translations.

Akubazgi Gebremariam in [32] develops Amharic-to-Tigrigna MT Using Hybrid Approach for the problem of 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-to-word. In addition, it can be more important to get the result without delay which is hard to accomplish with a human translator and there is no prior study conducted on the development of the Amharic-to-Tigrigna MT system which is crucial for the fulfillment of the lack of information and communication gap. The general objective of this study was to design and develop the Amharic-to-Tigrigna MT system using a hybrid approach and the study proposes a syntactic reordering approach that aligns the structural arrangement order of words in the source sentence to be more similar to the target sentences. The reordering rules are developed that fulfills both simple and complex Amharic sentences that have a difference in the structural arrangement order of words. One language model is developed since the system is unidirectional i.e. Amharic-to-Tigrigna. Translation model which assigns a probability that a given source language sentence generates a target language sentence is built and a decoder that searches for the best sequence of translation probability is used. Two major experiments are conducted using two different approaches. The first experiment is carried out using a statistical approach and the result obtained from the experiment has a BLEU score of 7.02%. The second experiment is carried out using a hybrid approach and the result obtained has a BLEU score of 17.47% s. From the result, the researcher concluded that the hybrid approach is better than the statistical approach for the Amharic-to-Tigrigna MT system.

Biruk Abel in [76] develops Geez to Amharic MT motivated by the fact that the distinctive attainment of Ethiopian history lies in the vast collection of manuscripts, compiled and preserved in the monasteries and churches. Almost all these scriptures and other religious works of the religion are done by the Geez language. Therefore, the Geez language means more to the

Ethiopian Orthodox Tewahido Church and Ethiopia. Overall, there are numerous church scriptures in the church which were written in Geez and not yet translated into the Amharic Language. Hence, an automatic translation system that translates Geez to a language being spoken at the national level like Amharic is of paramount importance. The general objective of this research was to design and implement an MT system that automatically translates Geez sentences to corresponding Amharic sentences using hybrid MT techniques. The system is composed of two main components a Rule-Based Geez Corpus Preprocessor and a Baseline SMT. The Rule-Based Preprocessor takes the manually Part of Speech (POS) tagged Geez corpus and produces another corpus that contains reordered Geez sentences having a similar structure with that of Amharic sentences. This component contains a set of activities that process each Geez sentence in the input corpus one by one to determine POS pattern and subsequently apply the corresponding reordering rule. It first reads all sentences from the input file and iterates through all sentences and it first determines POS pattern and applies the corresponding reordering rule. After each sentence is processed the output corpus along with the Amharic corpus will be supplied as an input to the Baseline SMT. Then using the input corpora the actual translation of Geez sentence to Amharic sentences will be performed by the Decoder of the Baseline SMT by using the Language model of Amharic and Translation model. The research is evaluated using BLEU evaluation metrics and compared with that of the Baseline SMT. Two experiments were conducted; one to test the Baseline SMT and the other to test the proposed system. To test the Baseline SMT both Geez and Amharic corpus without POS were used while to test the proposed system Geez corpus with POS and Amharic corpus with no POS were used. Based on the test results the Baseline SMT scored a BLEU of 72% and the proposed system outscored it by 4% and scored 76% owing to the reordering rules applied on Geez corpus.

Gelan Tulu in [36] develop Bidirectional Amharic-Afaan Oromo MT Using Hybrid Approach initiated by Afaan Oromo is the language with the largest number of speakers and even though there are a lot of historical, cultural, and religious documents available in Amharic and Afaan Oromo languages To address the knowledge to every citizen, there is a need to translate these documents to other Ethiopian languages especially from Amharic to Afaan Oromo and vice versa and there is no MT study conducted on Amharic-Afaan Oromo language pairs. With the fact that Amharic and Afaan Oromo are widely used in media, industries, and offices, there is a huge electronic data available in both languages. The general objective of this research work was

to design and develop a bidirectional Amharic – Afaan Oromo MT system using a hybrid approach. The scope of the research was restricted to translating simple sentences. The system has four components: sentence reordering, language model, decoding, and translation model. The sentence reordering is used to pre-process the structure of the source language to be more similar to the structure of the target language by using their Part of Speech (POS) tagging and to better guide the statistical engine. The Language models are done by using IRSTLM tool and translation models by using GIZA++ have been developed for Afaan Oromo and Amharic languages because the system is bidirectional. A decoder has been used to find the best translation in the target language (Amharic/Afaan Oromo) for a given source language (Afaan Oromo/Amharic) based on the translation and language models. The accuracy of the system is checked by two experiments using two different approaches. The first experiment is conducted by using a statistical approach to translate Amharic to Afaan Oromo and vice versa and has a BLEU score of 89.39% and 80.33% respectively. The second experiment is carried out by using a hybrid approach and has a BLEU score of 91.56% and 82.24% for Amharic to Afaan Oromo and Afaan Oromo to Amharic translation respectively. The result shows that the hybrid approach is slightly better than the statistical approach.

As we can see in the related works discussed above, in general, most of the papers done for different language pairs in recent years are based on NMT and they resulted in promising results when replacing SMT and hybrid approach with NMT [41], [42]. It is the current state of MT technology and also many researchers are joining this approach lately. Additionally, the attention mechanism can take this performance even to a higher level. Thus, in this work, we will apply the attention-based NMT approach for Amharic to Afaan Oromo MT.

CHAPTER 4: DESIGN OF ATTENTION BASED AMHARIC TO AFAN OROMO NMT

4.1 Introduction

In this chapter, we will discuss the architecture of attention based Amharic-Afaan Oromo neural machine translation in detail. Our model is based on the attention-based Seq2Seq model, in which we have an Encoder that learns the source (Amharic) language and a Decoder which learns about the target (Afaan Oromo) language and decodes the encoded source sentence. It utilizes GRU layers for both an Encoder and a Decoder which simplifies the stacked LSTM layer. We used a bi-directional and deep-multilayer GRU, which simplifies the beam-search decoding algorithm by producing translations from left to right. Section 4.2 discusses the system design of our proposed work by discussing the language model training and language model testing. The components of the system design are discussed in detail in section 4.3.

4.2 System Design

As the human translator needs to learn the structures of language pairs before it starts to translate from one language to another, the MT system should also take training on the structure of both source and target languages deeply and repetitively before translating from one to another language. Taking this fact into our work, in our design, the designing phase takes consideration of two-step processes. These are the training phase and the testing phase. In encoder-decoder language modeling, the training phase is followed by the testing phase, and we have used RNN for both training and testing. Sections 4.2.1 and 4.2.2 have discussed them in detail.

4.2.1 Language model training

To train our model, we first prepared a parallel corpus of Amharic and Afaan Oromo sentences. This corpus is then preprocessed before it is given to one-hot representation so that word embedding is formed. To do this, we tokenize the text and stem it to the root word after cleaning and normalizing it. After this, the word will be indexed by using the one-hot representation and padding will be applied to it. From these sequences of one-hot vector representation, word embedding will be formed. The output of the word embedding will be fed to the encoder, for

Amharic sentence, until the maximum length of the sentence is reached. The output of the encoder will be in turn given to the attention layer, which will process attention weight and outputs the context vector. This context vector will be given to the decoder. In case of Afaan Oromo sentence, the output of word embedding will be given directly to the decoder, and sentence matching evaluation will be performed with that of Amharic. This way, the machine will learn by updating parameters and doing the same thing repetitively to finally produce a trained model. Figure 4.1 illustrates this.

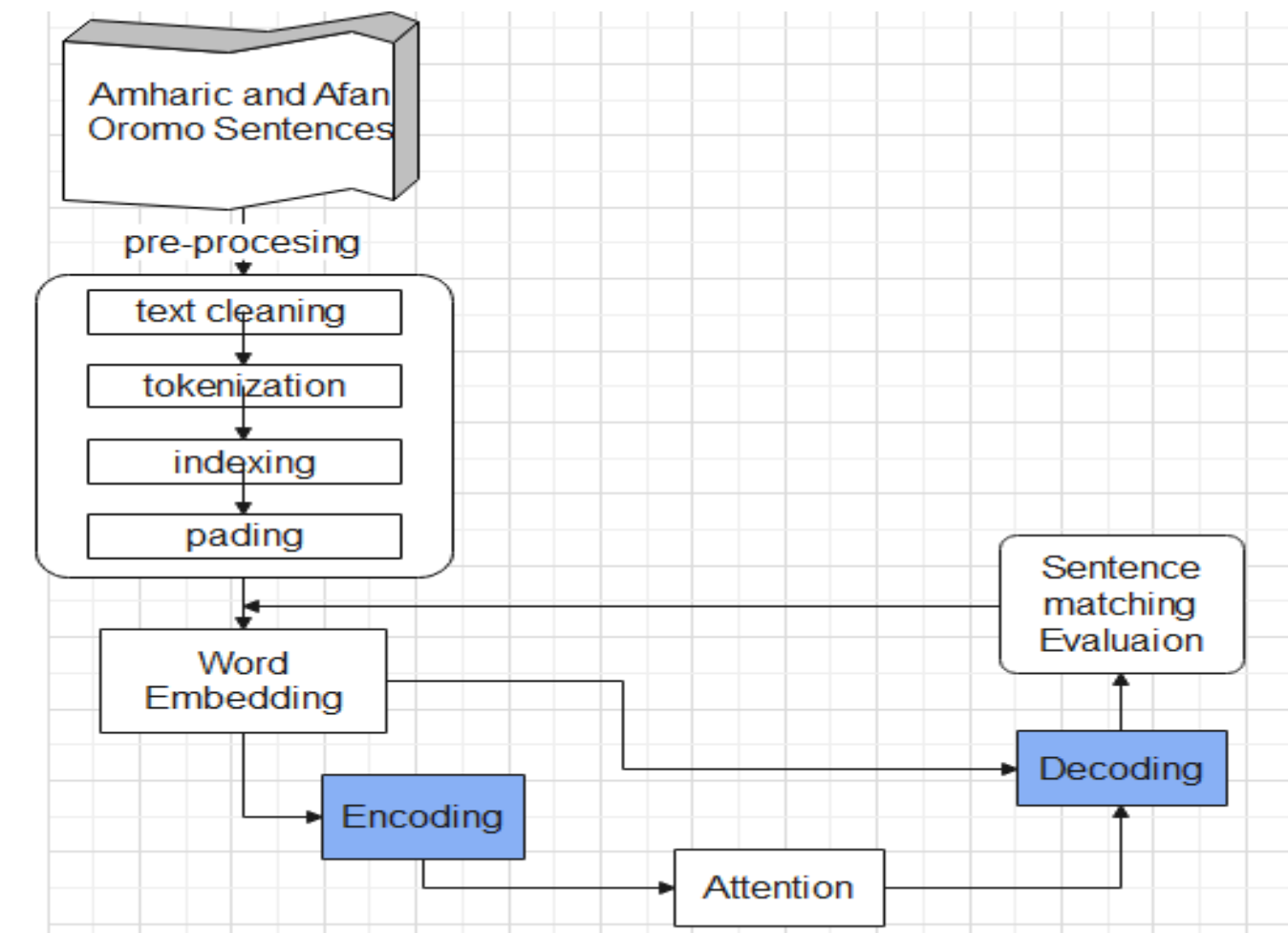


Figure 0.1 Language Model Training

To train our system model using RNN, forward pass, and back-propagation algorithms are used which uses a deep multi-layer GRU architecture. For each new input at a given time, GRU unit updates its memory to produce a hidden state. This hidden state is computed from the current input, future information, and a hidden state generated for the previous input, which will be often set to 0 for the starting hidden state. In our work, the input sequence is formed from the Amharic

sentence, Afaan Oromo sentence, and the end of sentence marker, <end>, because the Encoder and Decoder share many operations in common in forward pass algorithm.

4.2.2 Language model testing

To test the model, the same procedure as that of training model will be followed, in that the preprocessing, indexing and word embedding should be applied to our source language, i.e. Amharic. The output of the word embedding will be fed to the encoder, until the maximum length of the sentence is reached. The output of the encoder will be in turn given to the attention layer, which will process attention weight and outputs the context vector. This context vector will be given to the decoder.

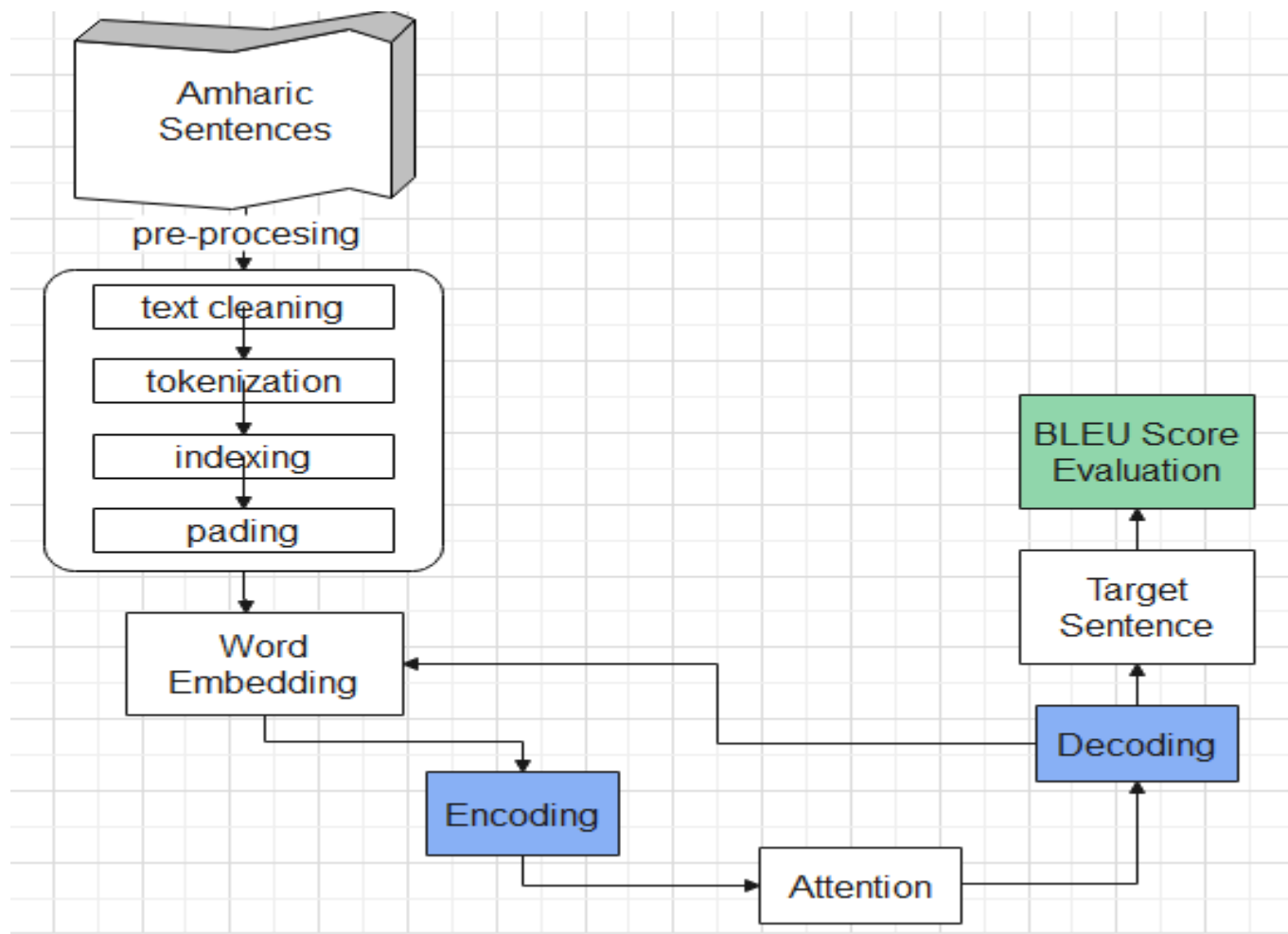


Figure 0.2 Language Model Testing

The decoder receives contextual vector from the encoder network to predict the translated target word of the target language. To do this task, the RNN-based decoder uses both currently

received input word context, and previously predicted word in combination with the trained model to give the prediction of target word for the current position. Here if the current wordcontext is at the beginning, the start of sentence indicator, <start>, will be used as the previous word. But if current word-context is not at the beginning and previous is not relevant to the candidate words, the teaching force technique which forces the neural network to ignore this irrelevant input and considers other alternatives. Finally the BLEU evaluation will be used to compute the accuracy of the translation. The language model testing is illustrated in figure 4.2.

It is common for models developed for machine translation problems to output a probability distribution over each word in the vocabulary for each word in the output sequence. It is then left to a decoder process to transform the probabilities into a final sequence of words. It's obvious to encounter this when working with recurrent neural networks on natural language processing tasks where text is generated as an output. The final layer in the neural network model has one neuron for each word in the output vocabulary and a softmax activation function is used to output a likelihood of each word in the vocabulary being the next word in the sequence. Decoding the most likely output sequence involves searching through all the possible output sequences based on their likelihood. The size of the vocabulary is often tens or hundreds of thousands of words. Therefore, the search problem is exponential in the length of the output sequence and is intractable (NP-complete) to search completely.

We have used beam searching to translate the Amharic sentence to Afaan Oromo sentence. Beam search is a relative search algorithm that is widely used to locate the output sequence from the recurrent neural network. It is based on the breadth-first search algorithm to reduce memory requirements. While the Greedy search method has produced a pretty good result, Beam-search is a more elaborated one with better results. Although it is not a necessary component for NMT, beam-search has been chosen by most NMT models to get the best performance. It's also the conventional technique of MT task that has been used for years in finding the most appropriate translation result [118]. After the source sentence is encoded, the decoding will immediately start after the end of sentence marker, <end>, for the source sentence is fed as input to the encoder.

4.3 System design Components

4.3.1 Data Pre-processing

To facilitate the machine translation by simplifying the input data both for training and testing, we have to preprocess it. This will help in obtaining a structured representation of the original corpus and to reduce the vocabulary size without introducing meaning change. This task includes normalizing the words, removing punctuations, since we do not need them for translation, expanding abbreviations and tokenizing the sentence.

Text Cleaning

To preprocess our corpus, we have performed text cleaning. This is done based on the properties of the languages. This involves removing quotes, removing unnecessary characters, removing spaces, and punctuation marks like: #, : and ;; adding a space between the word for Amharic sentence. Besides, the characters having the same role in a given word, for instance: ሰ ሠ, have been replaced with a single character, for example ሰ. In addition to this, words of Amharic sentence are normalized with labialized Amharic characters, for instance: ወጥቲዋል is replaced by ወጥቲል. In case of Afaan Oromo, special punctuation mark, which is hudhaa (‘), “apostrophe”, is leave as it is while other punctuation are removed and converting of text to lower case is done. In addition to removing special characters, “<start>” and “<end>”tag at the start and the end of the sentence, respectively, is added to the source and target sentences. After all this, the output will be given to the tokenizer. For instance, consider the following sentence:

ጁሮ ያለው ይስማ። which is equivalent to “Kan Gurra qabu haa dhaga’u”. After text cleaning the output which will be fed to the tokenizer will be: <start> ጁሮ ያለው ይስማ <end> for Amharic sentence and <start> Kan Gurra qabu haa dhaga’u <end> for Afaan Oromo sentence.

Tokenizing Source and Target Sentence

The machine cannot directly work on a sequence of sentences. Rather, these sentences must be split into a sequence of words, and then these sequences of words must be represented in integer form. This integer form representation represents each word in the sentence with its unique integer number derived at the time of vocabulary formation. This process is called tokenization. In our work, the process of segmenting the sentence into words is based on whitespace. For

instance, considering the output of the text cleaning from the previous section, the following table illustrates tokenization.

Table 0.1 Vocabulary Formation and Sentence Tokenization

| Cleaned text input | Word counting and vocabulary formation | | | | Sentence Tokenization |
|--|--|---|---------|----|-----------------------|
| <start>ይሁን እንጂ ከራሳችሁ ፀጉር እንደ እንኳ አትጠፋም <end> | <start> | 1 | ፀጉር | 6 | [3 4 5 6 7 8 9] |
| | <end> | 2 | እንደ | 7 | |
| | ይሁን | 3 | እንኳ | 8 | |
| | እንጂ | 4 | አትጠፋም | 9 | |
| | ከራሳችሁ | 5 | | | |
| <start>Ta'us rifeensa mataa keessanii keessaa tokko illee hin badu <end> | <start> | 1 | Keessaa | 7 | [3 4 5 6 7 8 9 10 11] |
| | <end> | 2 | Tokko | 8 | |
| | Ta'us | 3 | Illee | 9 | |
| | Rifeensa | 4 | Hin | 10 | |
| | Mataa | 5 | Badu | 11 | |
| | Keessanii | 6 | | | |

4.3.2 Indexing

After the dataset is tokenized it has to be changed into some form of numerical representation in which each word of the sentence in the data must be identified and represented by a unique index. This representation is called one-hot vector representation. The first word in a vocabulary is represented with the first index zero (0) and the last word is represented with index which is total number of unique words minus one. In our work, we have used an `index_word` attribute to represent vocabulary word by a unique ID. An `index_word` is a word-to-index dictionary where words are the keys and the corresponding integers are the values. Thus, each word in a sentence corresponds to a vocabulary item in a vocabulary. Figure 4.3 shows indexing sample:

| Input(Amharic) Language; index to word mapping | Target(Oromic) Language; index to word mapping |
|--|--|
| 1 ----> <start> | 1 ----> <start> |
| 41 ----> ሌላ | 205 ----> barattoonni |
| 68 ----> ደቀ | 211 ----> kaan |
| 93 ----> መዛመርት | 17 ----> garuu |
| 17 ----> ግን | 103 ----> lafa |
| 160 ----> ባህር | 24 ----> irraa |
| 608 ----> ዳርቻ | 158 ----> baay4ee |
| 53 ----> ብዙ | 43 ----> utuu |
| 340 ----> ራቀ | 3 ----> hin |
| 1152 ----> ጫትር | 1293 ----> fagachuu |
| 590 ----> ገደማ | 16 ----> gara |
| 1128 ----> ርቀት | 1574 ----> meetira |
| 14 ----> ላይ | 191 ----> qofa |
| 31 ----> ክብር | 955 ----> fagaachuu |
| 442 ----> አላ | 21 ----> waan |
| 416 ----> ተሞላ | 997 ----> turaniif |
| 814 ----> መረብ | 428 ----> bidiruu |
| 888 ----> ጉተተ | 365 ----> xinnoo |
| 434 ----> ትንሽ | 46 ----> irra |
| 320 ----> ጅልባ | 1013 ----> taa4anii |
| 22 ----> መጣ | 1021 ----> kiyoo |
| 2 ----> <end> | 563 ----> qurxummii |
| | 592 ----> guutamuu |
| | 28 ----> sana |
| | 7314 ----> harkisaa |
| | 486 ----> dhufan |
| | 2 ----> <end> |

Figure 0.3 Amharic (Left) and Afaan Oromo (Right) Words Sample Indexing

4.3.3 Word Embedding

One-hot representation results in a higher-dimensional vector, which will be taking a lot of time for the machine to learn. So once the sequence of vector representation is delivered by the one-hot representation, this vector should be converted to a lower-dimensional vector, for which we have used Word2Vec. Word embedding changes a higher dimensional one-hot vector into a lower-dimensional vector by keeping related words in a closed area of vector space by using algebraic vector representation. It allows similar words to have similar encodings. One of the basic advantages of word embedding is the reduction of out-of-vocabulary impact. This is possible because words will not be completely unknown as far as they have feature vectors even if they may not be seen in the training dataset. There are two types of Word2Vec, namely Skip-gram and Continuous Bag of Words (CBOW). The CBOW model learns to predict a target word leveraging all words in its neighborhood. The sum of the context vectors is used to predict the target word. The neighboring words taken into consideration are determined by a pre-defined

window size surrounding the target word. The SkipGram model, on the other hand, learns to predict a word based on a neighbor word. To put it simply, given a word, it learns to predict another word in its context. Even though CBOW is much faster than SkipGram, it is not able to represent rare words well. In addition to this, SkipGram is more efficient and accurate than CBOW, for which we have chosen SkipGram for our work.

The encoder will use this embedded form of data to generate a contextual relation between words. The decoder, on the other side will receive this context generated by the encoder and searches for a word that matches it in Afaan Oromo.

4.3.4 Padding

After word embedding is done, padding will be applied to the sentences. Padding is a process of adding a pad character to a given string so that all strings should maintain the same length. Since strings are by nature different in length, to let them have the same length, a pad character will be added either at the beginning, the end or in both directions. But, most of the time, a pad character will be added at the end of the string.

4.3.5 Encoder

After the process of padding and word embedding is done, produced tensor will be given as an input to the encoder. The encoder will take tensor as input and produce a single valued vector representation of the tensor, i.e. context vector. The design of encoder-decoder in our model is based on RNN. The RNN based encoder will encode the complete information of the source sequence, i.e. Amharic sentence, into a single-valued vector, which will be passed to the decoder, which will in turn produces an output sequence, i.e. Afaan Oromo sentence. That means, to work on textual data, for the encoder, an embedded form of sequential data should be given as an input. Assume we have n sequence length, the encoder will accept $A = \{a_1, a_2, \dots, a_n\}$, and produces a context of $B = \{b_1, b_2, \dots, b_n\}$ by maintaining the order of the input.

Even though there are different deep neural network architectures on which the encoder of RNN based system can be designed, GRU, and LSTM are the one who have gained popularity in MT. As it is stated in [30], and [119], LSTM has its own memory, which stores information outside the learning flow of the neural network. Thus, problems of long-term dependencies, i.e., in this case, for example, rarely frequented word combinations that are far apart, can be better

controlled. The LSTM cell corresponds to a node of a recurrent network and has, in addition to the input and output, a forget gate that avoids overfeeding of the vanishing gradient. GRU, on the other hand, is performed via an update and reset gates. The advantage of GRU cells over LSTM is that they are just as powerful as LSTM for moderately spaced combinations, even with small data sets; but they need less computing power. This shows that, with the same equipment, larger networks are feasible. The main difference is the presence or absence of an output gate, which tells how much of the content is presented to the next layer of the network. For LSTM cells, the whole memory can be limited by the output gate; which is false in case of GRU. In addition to this in the LSTM, there is no control of information flow in the cell as there is no reset gate. The internal structure of GRU is depicted in figure:

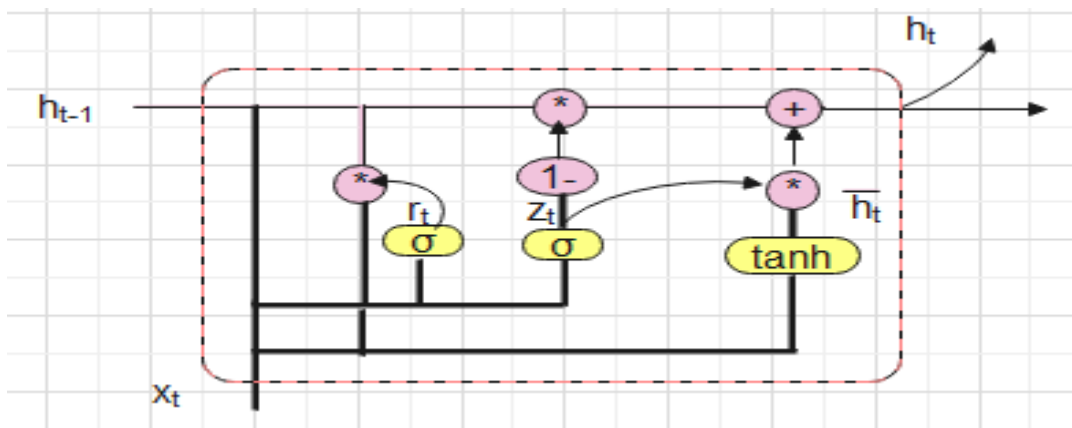


Figure 0.4 Internal Structure of GRU

At time t , the GRU calculates the new state as:

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \hat{h}_t$$

Where, z_t represents update gate and r_t represents reset gate.

This is to compute a linear interpolation between the previous state h_{t-1} and the current candidate state \hat{h}_t with the new sequence information. The update gate z_t decides to keep how much past information and to add how much new information. It controls the extent to which the information of the previous state is brought into the current state. The larger the value of z_t , the more information of the previous state is brought in. The state of z_t is updated as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

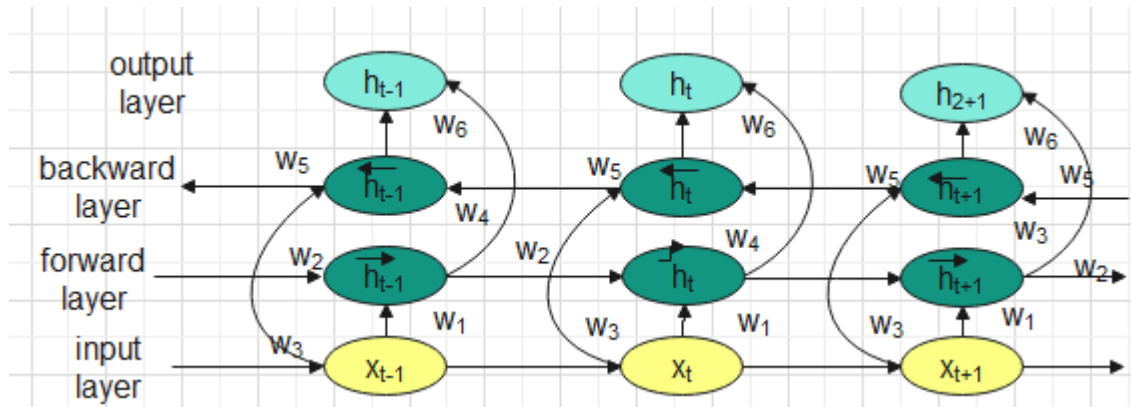
Where x_t is the sample vector at time t , \hat{h}_t is the candidate state computed in the same way as the hidden layer of traditional RNN network:

$$\hat{h}_t = \tanh(W_h x_t + r_t \otimes (U_h h_{t-1}) + b_h)$$

Where r_t denotes a reset gate which controls how much the previous state contributes to the current candidate state \hat{h}_t . The smaller the r_t value, the smaller the contribution from the previous state. If $r_t=0$, it will forget the previous state. The reset gate is updated as:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

For many sequence modeling tasks, it is beneficial to have access to future as well as past context. However, standard GRU networks process sequences in temporal order and they ignore future context. Bidirectional GRU networks extend the unidirectional GRU networks by introducing a second layer, where the hidden to hidden connections flow in opposite temporal order. The model is therefore able to exploit information both from the past and the future. Due to the advantages of GRU, in general, and bidirectional GRU, in specific, addressed above, we preferred to use bidirectional GRU in our model both for the encoder and decoder. The following figure depicts the bidirectional GRU network structure:



/Figure 0.5 Bidirectional GRU Network Structure

After all these, the Encoder outputs context vector as its final output from current *input data*, previous output of the system and accessing the future output. This context vector will be fed into the attention layer or directly to the decoder.

4.3.6 Attention

In the standard encoder-decoder architecture, the encoder outputs a summarized single vector for all input sequence. This summarized vector will be given to the decoder to decode it. This way, the decoder will have access to only the last layer of the encoder. Representing the entire input sequence with single vector is inefficient. This problem will even get worse if the size of the vocabulary is huge. To tackle this problem we have applied attention mechanism to our encoder-decoder model. In attention based encoder decoder model, the target word is predicted based on the context vectors associated with the source position as well as the previously generated target word. The attention mechanism aligns the input and output sequences, with an alignment score parameterized by a feed-forward network.

Attention could be either global or local [120, 121]. Global attention calculates the context vector by considering the relevance order of all words in the source sentence. This method considers all the source word in the decoding period. The main drawback is calculation speed deteriorates when the sequence is very long since one hidden state will be generated in one time-step in the Encoder, the cost of score function would be linear with the number of time-steps on the Encoder. When the input is a long sequence like a compound sentence or a paragraph, it may affect the decoding speed. Even if covering more information would generally get a better result, practically, in human language the current word would naturally have a high dependency with some of its nearby words. So, the idea of local attention has emerged. In contrast to global attention, local attention will just calculate the relevance with a subset of the source sentence. This fixes the length of attention vector by giving a scope number, thus avoiding the expensive computation in getting context vectors. Given the current target words position p_t , the local attention fixes the context vector in scope D . The context vector c_t is then derived as a weighted average over the set of source hidden states within the range $[p_t - D, p_t + D]$. In our work we have implemented the local attention mechanism. The following figure shows the attention mechanisms.

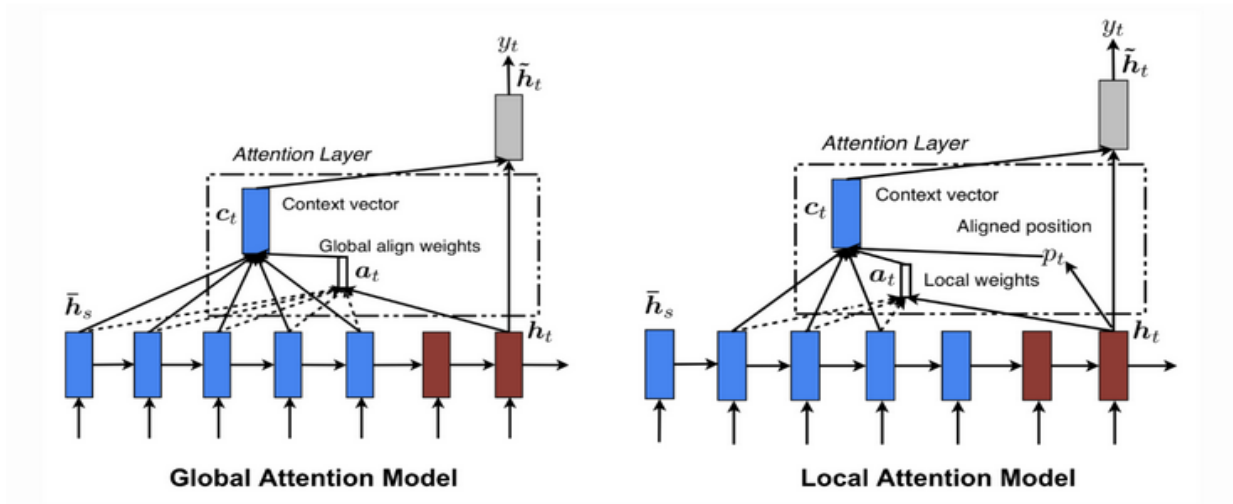


Figure 0.6 Global (Left) and Local (Right) Attention Model [121].

After this, context vector with attention will be given as an input to the decoder.

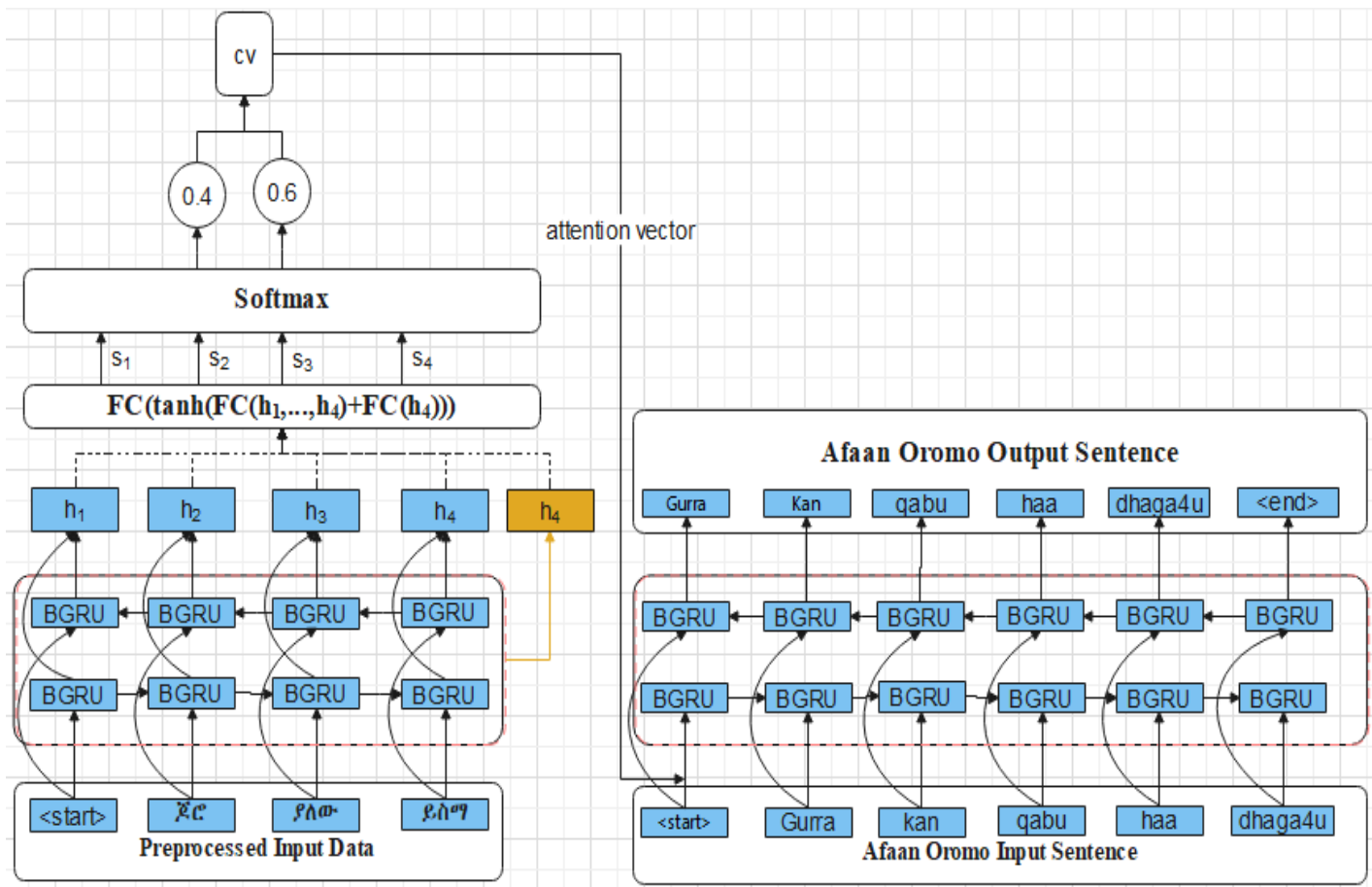


Figure 0.7 Encoder-Decoder model with local attention

4.3.7 Decoder

Decoder decodes what it has received from the attention layer and produces the equivalent meaning of what has been encoded. In our work we have used RNN based model to design our decoder. There are different choices for the decoder architecture like GRU, LSTM, bidirectional RNN, etc. As we have described in the encoder part, we will use a bidirectional GRU for the decoder too. The output of the attention layer is fed to the decoder network as input. This work will be repeated for all layers of the GRU, and the decoder will output word prediction and generate new hidden state, and new output word prediction until the end of the text is reached. Here, the number of GRU we used in a single layer is set as the maximum length of the sequence in the target sentence data. When the shorter sentence entered into the network, the end of sentence must be marked with end of string indicator symbol and the other units are fed with padding which is used to change the variable length vector into fixed length vector by completing the length of shorter sentence by using zero value. In order to find the matching word, Decoder uses the beam searching mechanism with the use of softmax function which searches for the best matching word. Therefore, the Decoder generates one word at a time and repeats the searching until encounter end of string indicator <end> which shows the last word of the sentence. Figure 4.7 illustrates the encoder-decoder model with local attention.

CHAPTER 5: EXPERIMENTATION AND DISCUSSION

5.1 Introduction

This chapter discusses the results of this work by showing the data preparation, environmental setups, parameter optimization, and performance of the research using BLEU score metrics. Finally, we have compared the result of our work with and without applying attention mechanism.

5.2 Dataset collection and preparation

As it is known, NMT needs a bunch of dataset so that it provides better results in terms of translation accuracy. So, we have tried to prepare as much data as possible. In this study, we have prepared a total of 11,727 sentence pairs (Amharic (source) and Afaan Oromo (target) from religious books like the holy bible and legal sources like The Oromia National Regional State Magalata Oromia. Most of this corpus has been collected by Solomon Teferra [1] and we have collected the rest of it. After we have stored it in text format so that we can feed it to our system after aligning the collected parallel sentences separated by tab(" "). For the purpose of fastening training time, we have used the google colaboratory since it uses Graphics Processing Unit (GPU) rather than a CPU which takes significantly very large training time. GPU is composed of hundreds of cores that can handle thousands of threads simultaneously. We have classified our data into training and test sets by taking 90% of the dataset for training set and the rest 10% for test set. So, in our case, we have 10,554 training set and 1173 test set.

We have used different techniques to prepare the data we have collected. This includes character normalization, and splitting the sentence to word level (tokens). After we have made tokens of the sentences, these sequences of words are represented by integers and feed to the system. To do this, the vocabularies are assigned with a unique index number while they are visited and then are passed to word to Vec.

5.3 System environment/ tools used for the development

To implement our system, choosing a programming language, and preparing the required environment is necessary. A programming language chosen to implement the system is python programming language. Python programming language supports a set of a freely available library in the deep learning algorithm. We used Keras library, TensorFlow library, and NumPy library which is freely available. We have chosen colab as a baseline for implementation. Colab is the only tool which supports graphical processing unit (GPU) which speeds up the training time hundreds time faster than central processing unit (CPU).

5.4 Parameter optimization and training the experimental systems

In our experimental study, we have trained and tested the proposed system on our parallel corpus of 11,727 sentences in two columns (one for Amharic and the other for Afaan Oromo) separated by a tab. To get the desired result, we have done different experiments on different issues to adjust the parameters of the model. The first issue is the selection of neuron units to the number of dense layers and batch sizes. For training and testing dataset, we have to choose one of 16, 32, 64, 128, 256, 512, 1024, 2048 neuron units with batch sizes based on the number of data size we are using. Batch size determines how many samples should be processed before the model is updated. Thus, the size of a batch must be more than or equal to one and less than or equal to the number of samples in the training dataset. In our case, we have chosen a batch size of 64. This means, the training set we have (10,554), is divided by 64 and gives us the number of batches in each epoch, which is 146. The algorithm takes the first 64 samples from the training dataset and trains the network. Next, it takes the second 64 samples and trains the network again. We keep doing this procedure until we have propagated all samples through the network. The problem might happen with the last set of samples since we've used 10,554 which are not divisible by 64 without remainder. After training for 164 times for 64 batches, the system just gets the final 58 samples (the remainder of 10,554 divided by 64) and train the network. For each epoch 55 epochs in our case) it iterates for 164 times for 64 batches and 1 time for 58 batches). The number of batches in each epoch is inversely proportional to batch size. Training a system by small batch sizes (8, 16, and 32), results in less accuracy, while training a system by large batch sizes (128, 256, and 512) requires more memory and even make the training slower. Thus, we preferred 1024 neuron units for 2 dense layers with 64 batch sizes for our training dataset since our corpus is medium size.

The second issue after setting neuron units and batch size is the selection of word embedding size. For this purpose, we trained 1024 neuron units with 64 batch size in three different number of word embedding size 64, 128 and 256 respectively for 55 epochs. From these different word embedding sizes, we selected 1024 neuron units with 64 batch size and 256 word embedding dimension for 55 epochs because of it minimizes a loss level when compared with 1024 neuron units with 128 batch size and 128 embedding dimension, 1024 neuron units with 256 batch size and 256 embedding dimension and 1024 neuron units with 64 batch size and 128 embedding dimension with the same number of epochs. The number of epochs is the number of complete passes (number of looking back into a single dataset element) through the training dataset.

After adjusting neuron units, batch size, and word embedding, choosing a learning rate for training is important. In order to select optimum learning rate, we trained our system with 0.001 learning rate which takes 7,562.5 secs to train with loss level of 0.011708, with 0.1 which takes 70,183 secs with loss level of 0.016 and with 0.0001 learning rate which takes 30,684 secs and with loss level of 0.012312. As shown in Figure 5.2 and Figure 5.3, Adam's optimizer (which defaults learning rate to 0.001) minimizes a data's training time highly in optimum loss level compared to 0.1 and 0.0001 learning rate. So we initialized the defaults learning rate to 0.001 and we selected sigmoidal optimizer to optimize parameter. Therefore, the training of our architecture took 7,562.5 seconds (2 hours 06 minutes and 03 secs) on 90% of our corpus (10,554) training sentence pairs. To our model, we have used batch training with a batch size of 64 and trained for 55 epochs on GPU based core i3 with 4GB RAM computer. As we have seen in figure 5.1 the loss level significantly decreased until it reaches 35th epoch and then it decreased with insignificant numbers.

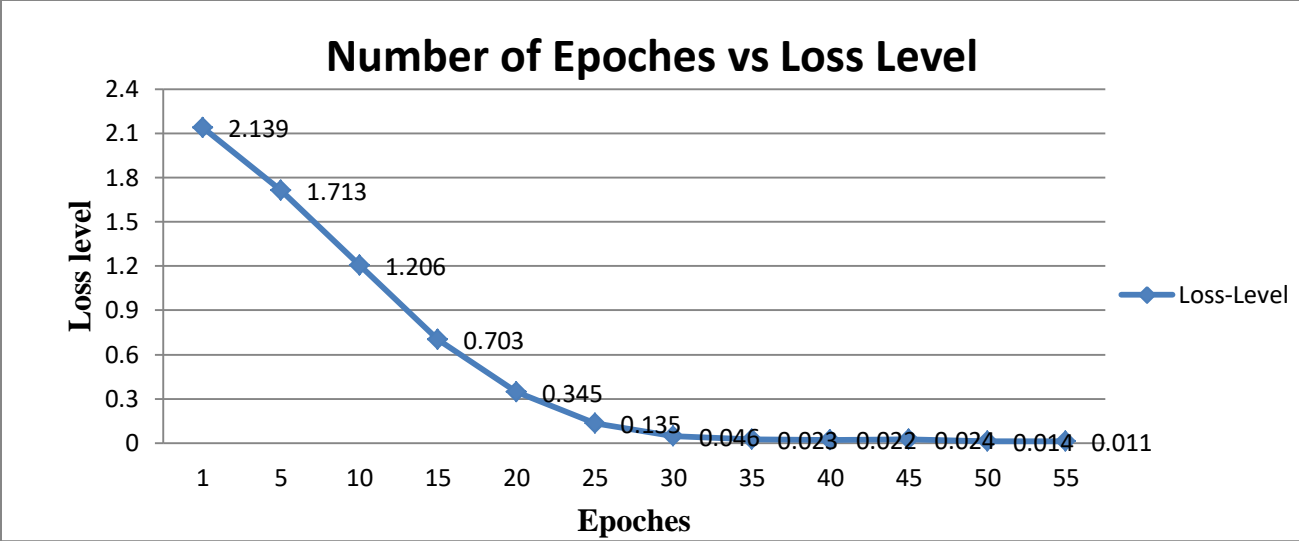


Figure 0.1 Number of Epoches vs Loss Level

The above diagram shows the loss level in different number of epochs the loss decreases significantly until the 35th epochs and show slight improvement for the next number of epochs so we choose epoch 55 which we get low loss level value.

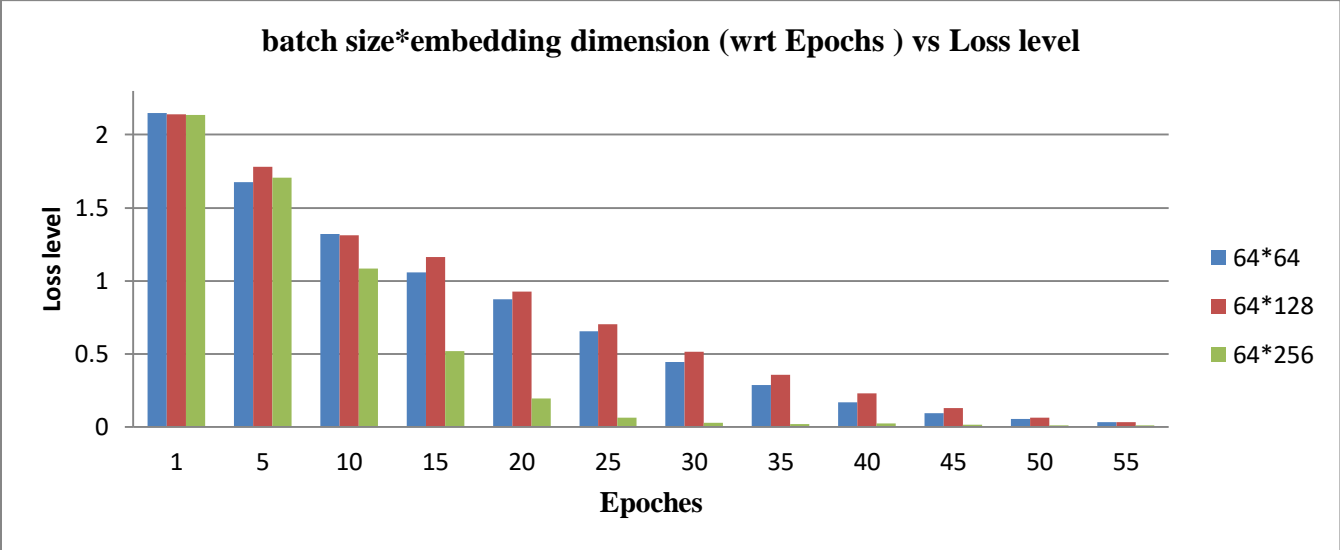


Figure 0.2 Loss level for each batch size and embedding dimension with respect to epochs

As we illustrate in the above diagram the batch size*embedding dimension combination with respect to different number of epoch to get the minimum number of loss level batch size 64 and embedding dimension 256 give us the smallest loss level value in almost all number of epochs except epoch number 1 and 5 so we choose 64 batch size and 256 embedding dimension with 55 epoch for our system.

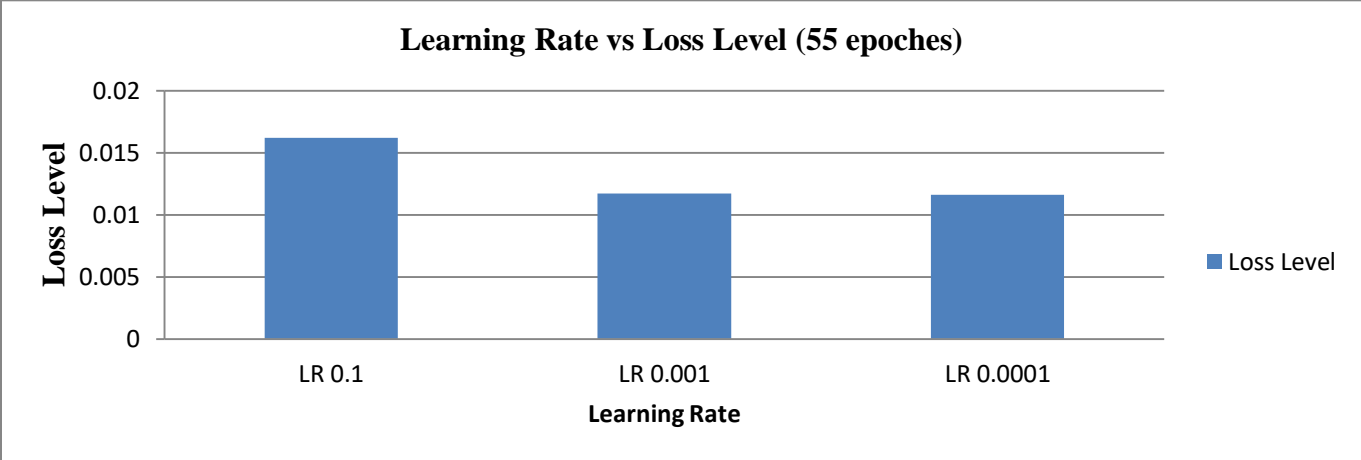


Figure 0.3 Learning rate vs Loss level

The above diagram shows learning rate versus loss level for 55 epochs as we can see in the diagram the smallest learning rate gives us the lowest loss level value which is the loss level we get the highest accuracy for but the value decrease in insignificant way after learning rate 0.001 and when we see learning rate versus time taken for each epoch in the diagram below the time taken to finish each epoch per second drop dramatically for the learning rate 0.001 so we choose the default Adams optimizer value learning rate 0.001.

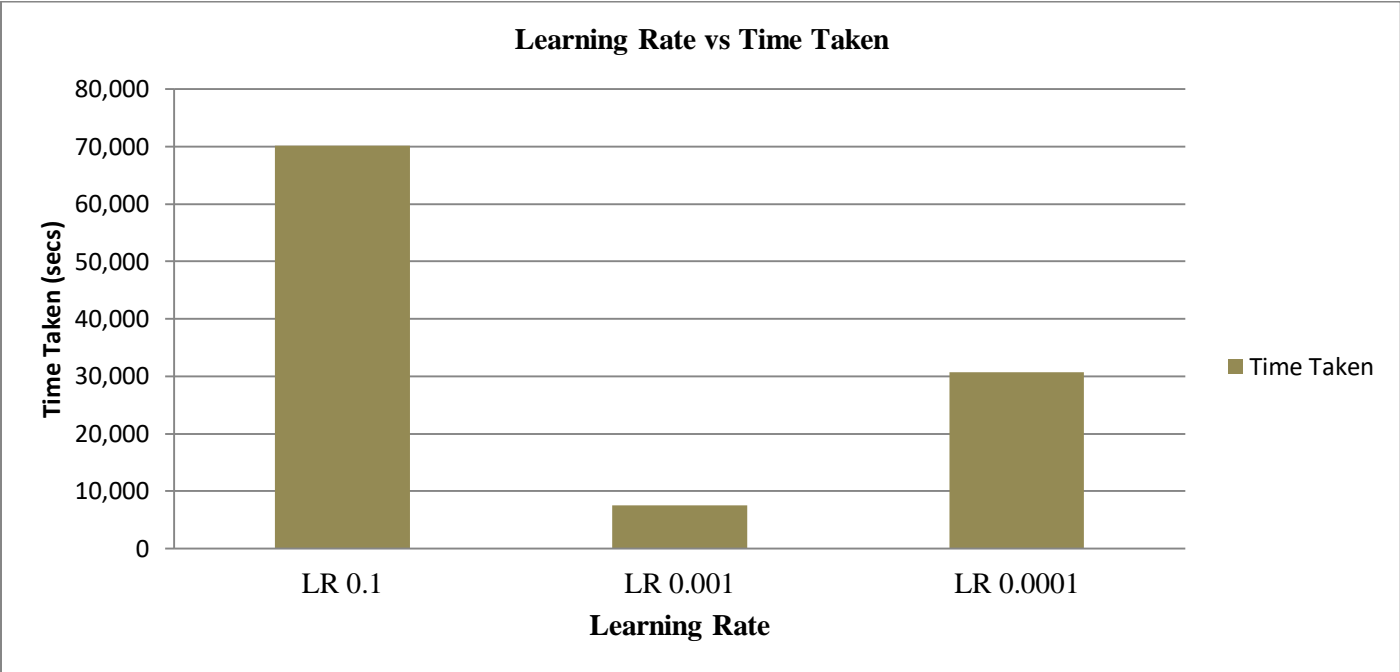


Figure 0.4 Learning rate vs Time taken to train

5.5 Experimental results

In our experiment, we have implemented Amharic to Afaan Oromo with and without attention mechanism. We have used the same train and test data set with the same parameters for both mechanisms. After training and testing we measured the result by using BLEU score metrics and report the result by taking average BLEU score of 30 sample output result of 55 epochs of training and testing the systems. When translating input sentence to output sentence, we use human translated reference sentence to get BLEU score of the translation. The average result of Amharic to Afaan Oromo NMT system without attention in BLEU score is 61.49%, and Attention-based Amharic to Afaan Oromo NMT system is 67.82%.

5.6 Discussion on the result of the study

In this work, we have designed and implemented MT between Amharic and Afaan Oromo languages by using attention based NMT approaches which is said to be way of the future by solving a problem found in rule based and statistical MT to exploit the language structure.

Regardless of its advantage, NMT, with basic encoder-decoder architecture, the encoder reads the whole input sentences, encodes the complete information into a single real-valued vector which is passed to the Decoder and then Decoder network generates the target translation. This basic architecture is well suited with a translation of shorter sentences. But as the length of a given sentence become larger, the neural network with basic encoder-decoder architecture will get difficulty in deriving a context since context within a sentence is derived as the inter-dependency of nearby words in a given sequences of words within a sentence. This is because as the sentences become longer, the inter-dependency of words at the beginning and at the end of a sentence will be very loose. Moreover, NMT with basic encoder-decoder model has limitation in a way it handles a larger number of vocabulary. That is, since a unique index will be assigned to a word that is visited, to uniquely identify it within a dataset, the number used for word representation will become larger and results in larger word vector dimension. In our work, we have conducted two experiments using two different mechanisms, i.e. non-attention-based and attention-based, and compared them to each other with an expectation to get a better result in terms of both BLEU score and time efficiency and also expecting that the imitation of the basic encoder-decoder based NMT will be solved by the attention mechanism.

According to the result of our experiment, by testing our work without adding attention mechanism we have achieved a BLEU score of 61.49 while a BLEU score of 67.82 is achieved by employing attention mechanism. Moreover, the training time that is required to train the machine is also significantly improved for the case of attention-based mechanism compared to that of Amharic to Afaan Oromo NMT without attention.

In addition to this, as we are able to conclude our experiment, the ability of neural networks to store longer contextual relevancies of words found in longer sentence, can be even enhanced by applying attention mechanism. We have illustrated this by the following example table:

Table 0.1 Example Translation performance for long sentence

| | |
|--|---|
| Attention Based Amharic-Afaan Oromo NMT | Input: Long Amharic sentence: ጳውሎስ ወንድሞቼን ኃጢአት ለሰውነታችሁ ምኞት ተገብሮ እንድትሆኑ በማድረግ ሚች በሆነው ሰውነታችሁ ላይ መንገሡን እንዲቀጥል አትፍቀዱ በማለት አሳስቧቸዋል። |
| | Reference: Phaawulos obboolota isaa yommuu gorsu, Isin akka kajeellaa fooniitiif hin ajajamneef, cubbuun qaama keessan isa du’u irratti mootii akka ta’u hin godhinaa jedheera. |
| | Output: phaawulos obboolota keessan yommuu gorsu isin akka kajeellaa ajaja cubbuun qaama keessan isa mootii akka jedheera |
| | BLEUScore: 59.41 |
| Non Attention Based Afaan Oromo NMT | Input: Long Amharic sentence ጳውሎስ ወንድሞቼን ኃጢአት ለሰውነታችሁ ምኞት ተገብሮ እንድትሆኑ በማድረግ ሚች በሆነው ሰውነታችሁ ላይ መንገሡን እንዲቀጥል አትፍቀዱ በማለት አሳስቧቸዋል። |
| | Reference: Phaawulos obboolota isaa yommuu gorsu, Isin akka kajeellaa fooniitiif hin ajajamneef, cubbuun qaama keessan isa du’u irratti mootii akka ta’u hin godhinaa jedheera. |
| | Output: phaawulos warri kaan hundi akka isaaniif hin dandeenye ni beektu |
| | BLEU Score: 49.17 |

But, in general, both mechanisms delivers higher performance for shorter sentences compared to longer sentences. This is illustrated in figure 5.5.

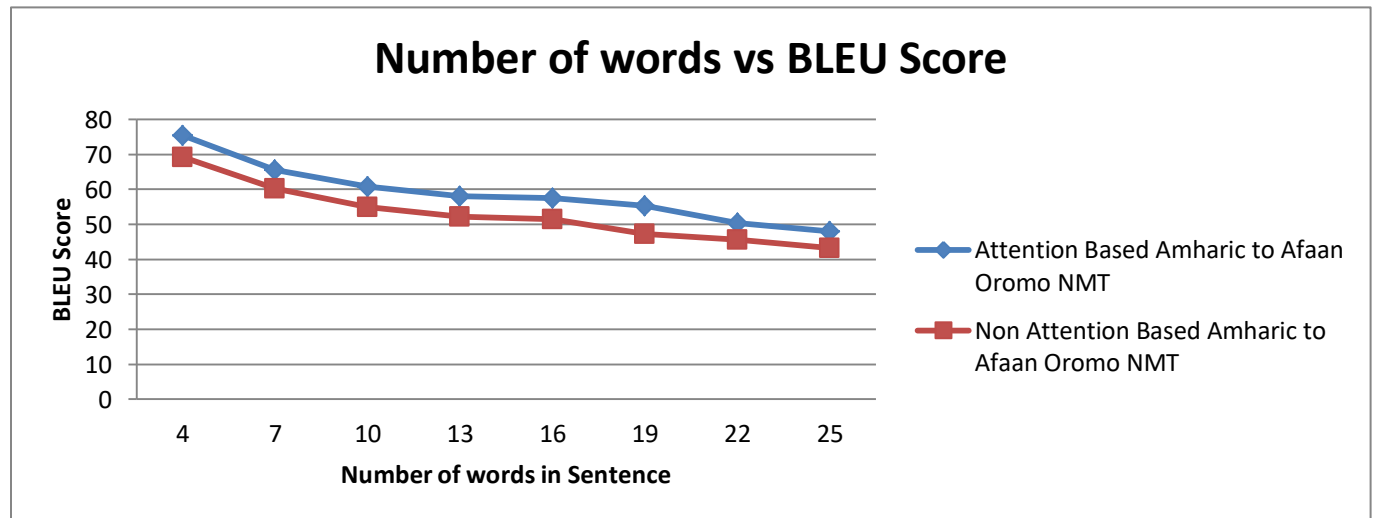


Figure 0.5 Number of words in sentence vs BLEU Score

The above diagram shows that attention mechanism improves the basic encoder decoder mechanisms for all short and long sentences but we observe that the attention mechanism outperform for longer sentence even though the blue score result decreases no matter what we give our system longer sentences.

The other interesting feature is that the system delivers good performance in translating sub sentences/ phrases. For example considering the sentence “መንፈስ ለጉባኤዎቹ የሚናገረውን ጆሮ ያለው ይስማ።” which will have a reference of “**Inni gurra qabu wanta hafuurri gumiiwwanitti dubbatu haa dhaga’u**”, we have tested the system with a source sub-sentence/ phrase of “ጆሮ ያለው ይስማ”. The system has produced the equivalent translation “**namni gurra qabu haa dhaga’u**” of this phrase even though it is not trained with this exact sentence. Beside this, the other interesting feature seen on using attention-based translation with RNN model is that the RNN based model keeps contextual relatedness in which related words are used for translation rather than only using exact target word for translation.

In general, all the experiment results that we have discussed above shows that considering the morphologically complexity of the languages and with the limited dataset that we have compared to foreign languages, NMT with attention shows promising performance.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Introduction

In this chapter, we will conclude the work that we have done so far. In this chapter, the conclusion driven from this research work and the recommendation for any one that interested to work on the machine translation between Amharic and Afaan Oromo language pair (or any other language pair) or related task will be discussed. We will discuss our contribution, and finally we will try to forward our recommendation regarding the future work.

6.2 Conclusion

The main goal of this study is to develop attention-based Amharic to Afaan Oromo NMT system. After exploring the morphological structure, similarities, and differences of the source and target languages, data sets for both languages are prepared. For the matter of this work, we have prepared a total of 11,727 parallel corpus collected mainly from religious book and legal books.

As the objective of the work is to improve the machine translation between the language pair by implementing RNN based architecture, two experiments were conducted by using the collected data set to check the accuracy of the system using two different (with attention and non-attention) approaches.

The first experiment is conducted by using the non-attention-based approach and delivers a BLEU score of 61.49. The second experiment is carried out by using an attention-based approach and it has a BLEU score of 67.82. Therefore, the results of our experiments have shown that attention-based Amharic to Afaan Oromo NMT system has shown better BLEU score value and improved the Amharic to Afaan Oromo NMT without attention by a BLEU score of +6.33. Moreover, our experiment on attention based mechanism has shown less training time in comparison with the non-attention-based system.

More importantly the attention-based system shows the great result in case of translating longer sentence than the non-attention-based system. We have also seen as using attention-based translation with RNN model is that the RNN based model keeps contextual relatedness in which related words are used for translation rather than only using exact target word for translation.

6.3 Contribution

- ✓ The main contribution of our work is the system we developed is the first NMT system developed from Amharic language to Afaan Oromo language pair. Thus it will help the upcoming researchers being as the input to try other enhancements in NMT.
- ✓ We contribute a small amount of corpus, as we used a corpus prepared by Solomon et al. we try to prepare some additional corpus from legal source,
- ✓ The other contribution of this work is that this work can be used as the beginning point for other NLP applications like Chat → Response, Question → Answering, Human → Machine (e.g. IOT commands), text classification and other related research areas in the Amharic Afaan Oromo language pair.
- ✓ The other contribution of this thesis is the discussion of how much attention-based mechanism outperforms non-attention-based system with the same parameters. Thus one can easily join to attention-based mechanism after deeply understanding this work.

6.4 Future work

This work is only used for translation Amharic to Afaan Oromo NMT. So, we recommend making this system bi-lingual as a future work in order to easily distribute resources written by using both languages as a source and target languages. This work is limited to NMT of text to text sentences; but we recommend that speech to text NMT between these language pair will be studied. Lastly, we recommend that more datasets will be prepared and used with more layers to increase the accuracy of the system.

REFERENCES

- [1] Solomon Teferra, Martha Yifru, Michael Melese, Million Meshesha, Solomon Atinafu, Yaregal Assabie, Biniyam Ephrem, Wondimagegnhue Tsegaye, Tsegaye Andargie, Wondwossen Mulugeta, Hafte Abera, Tewodros Abebe, Amanuel Lemma, Seifedin Shifaw “Parallel Corpora for bi-Directional Statistical MT for Seven Ethiopian Language Pairs”, Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 83–90 Santa Fe, New Mexico, USA, August 20, 2018.
- [2] Eleni Teshome, “Bidirectional English-Amharic MT: An Experiment using constrained corpus”, MSc thesis, Addis Ababa University, Ethiopia, 2013.
- [3] Towards Data Science. 2020. Towards Data Science. [online] Available at: <<https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp/>> [Accessed 9 October 2020].
- [4] NewGenApps. 2020. NLP | Enterprise Search | Automatic Summarization - Newgenapps. [online] Available at: <<https://www.newgenapps.com/technology/natural-language-processing/>> [Accessed 9 October 2020].
- [5] S. Nirenburg and H. L. Somers, “Readings in MT” MITPress, 2003.
- [6] VaniK, “Example-Based MT”, [online] Available at: <http://dSPACE.cusat.ac.in/jspui/bitstream/3623/1/EBMTorginal.htm> [Accessed 9 October 2020].
- [7] Slocum, Jonathan, "A survey of MT: its history, current status, and future prospects", 1985.
- [8] Diego Moussallema, Matthias Wauera, Axel-Cyrille Ngonga Ngomo, “MT using Semantic Web Technologies: A Survey”, Journal of Web Semantics, 2018
- [9] Mirjam Sepesy Maučec and Gregor Donaj, “MT and the Evaluation of Its Quality”, 2019
- [10] 2019. [online] Available at: <<https://www.omniglot.com/language/articles/MT/>> [Accessed 9 October 2019].
- [11] Bahdanau, D., Cho, K., & Bengio, Y. “Neural MT by jointly learning to align and translate”, ICLR, 2015

- [12] Dorr, B.J., Jordan, P.W., & Benoit, J.W. “A survey of Current Paradigms in MTs”, (Technical Report: LAMP-TR-027/UMIACS-TR-98-72/CS-TR- 3961). University of Maryland, College Park, 1998
- [13] Tadesse Kassa, “Morpheme-Based Bi-directional Ge’ez Amharic MT”, Master’s thesis, Addis Ababa University, Ethiopia, 2018.
- [14] Microsoft.com. 2020. Microsoft Translator Blog: Microsoft Translator Launching Neural Network Based Translations For All Its Speech Languages. [online] Available at: <<https://www.microsoft.com/en-us/translator/blog/2016/11/15/microsoft-translator-launchingneural-network-based-translations-for-all-its-speech-languages>> [Accessed 9 October 2020].
- [15] Sutskever, I., Vinyals, O., & Le, Q. V., “Sequence to sequence learning with neural networks”, In Advances in neural information processing systems (pp. 3104-3112), 2014
- [16] H. Luong, K. Cho, and Ch. Manning, “Neural MT” - Tutorial ACL, 2016.
- [17] Kalchbrenner, N., & Blunsom, P., “Recurrent Continuous Translation Models”. In EMNLP (Vol. 3, No. 39, p. 413), 2013.
- [18] K. Cho et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical MT,” In Proceedings of the Conference on EMNLP, October 25-29, Doha, Qatar, 2014, pp. 1724-1734.
- [19] Sites.google.com. 2020. Neural MT - Tutorial ACL 2016. [online] Available at: <<https://sites.google.com/site/acl16nmt/>> [Accessed 9 October 2020].
- [20] M. Bulcha, "Oromo Writing", Nordic Journal of African Studies, pp. 36-59, 1995.
- [21] G. B. Gene ,“Students in Ancient oriental civilization”, No.60, S. leslie and U. G. Thomas, Eds., chicago: university of chicago, 1982.
- [22] Marstranlation.com. 2020. 10 Most Spoken Languages In Africa | Mars Translation. [online] Available at: <<https://www.marstranlation.com/blog/10-most-spoken-languages-in-africa>> [Accessed 9 October 2020].

- [23] Sisay Adugna, “English – Afaan Oromoo MT: an experiment using statistical approach”, Master’s thesis, Addis Ababa University, 2009
- [24] Constitution of the Federal Democratic Republic of Ethiopia
- [25] Sloculn, Jonathan, "A survey of MT: its history, current status, and future prospects," Jonathan Sloculn, 1985.
- [26] Almahairi, A., Cho, K., Habash, N., & Courville, A., “First Result on Arabic to English Neural MT”, 2016.
- [27] Takahashi S, Wada H, Tadenuma R, and Watanabe S., “English-Japanese MT”, Readings in MT. 2003
- [28] Holger Schwenk, Jean-Baptiste Fouet and Jean Senellart, “First Steps towards a generalpurpose French/English Statistical MTSystem ”, In Proceedings of the third workshop on statistical MT, pages 119-122, Ohio, 2008.
- [29] Mulu Gebreegziabher Teshome and Laurent Besacier."Preliminary experiments on EnglishAmharic Statistical MT.", In SLTU, pp. 36-41, 2012.
- [30] Michael Gasser, “Toward a Rule-Based System for English-Amharic Translation”, In Proc of the 8th Int. 4th workshop on African LT, 2012.
- [31] Jabesa Daba. “A Hybrid Approach to the Development of Bidirectional English-Oromiffa MT”, In Proceedings of the 9th Int. Conference on NLP (PoITAL2014), Springer Lecture Notes in Artificial Intelligence (LNAI), Vol. 8686, pp. 228-235, Warsaw, Poland, 2014.
- [32] Akubazgi Gebremariam, “Amharic-Tigrigna MT using hybrid approach”, Master’s thesis, Addis Ababa University, Ethiopia, 2017.
- [33] Banklesstimes.com. 2020. Neural MT: Now And Into The Future – Bankless Times. [online] Available at: <<https://www.banklesstimes.com/2018/02/24/neural-machine-translation-now-future>> [Accessed 9 October 2020].

- [34] M.Kasthuri and S.Britto Ramesh Kumar, “Rule Based MT System from English to Tamil”, In Proceedings of IEEE World Congress on Computing and Communication Technologies, 2014.
- [35] Abdurehman Dawud Mohammed. “A Top-Down Chart Parser for Amharic Sentences”, Master’s Thesis, Addis Ababa University, 2015
- [36] Gelan Tulu, Bidirectional Amharic-Afaan Oromo MT Using Hybrid Approach, Master’s Thesis, Addis Ababa University, 2020
- [37] Arfaso Birhanu, Bi-Directional English-Afan Oromo MT Using Convolutional Neural Network, Master’s Thesis, Addis Ababa University, 2019
- [38] Shivkaran Singh, M. Anand Kumar and K.P. Soman, Attention based English to Punjabi neural MT, Journal of Intelligent & Fuzzy Systems, 2018
- [39] Hirschberg, J. and Manning, C., 2020. Advances In Natural Language Processing. [online] Science.sciencemag.org. Available at: <<http://science.sciencemag.org/>> [Accessed 9 October 2020].
- [40] Thi-Vinh Ngo, Thanh-Le Ha, and Phuong-Thai Nguyen, “Combining Advanced Methods in Japanese-Vietnamese Neural MT”, presented at the 2018 10th international conference on knowledge and systems engineering (KSE).
- [41] Bender, L., Bowen, D., Cooper, R., Ferguson, C., “Language in Ethiopia”, Oxford University Press, London (1976)
- [42] Ethiopia Online Visa. 2020. Amharic: The Ethiopian Language. [online] Available at: <<https://www.ethiopiaonlinevisa.com/amharic-the-ethiopian-language/>> [Accessed 9 October 2020].
- [43] Samuel Eyassu, et.al, “Classifying Amharic News Text Using Self Organizing Maps”, Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 71–78, 2005.

- [44] Atelach Alemu Argaw, et.al, “Amharic Stemmer: Reducing Words to their Citation Forms” Proceedings of the 5th Workshop on Important Unresolved Matters, pages 104–110, Prague, Czech Republic, 2007
- [45] Bethelhem Mengistu, “N-gram-Based Automatic Indexing for Amharic Text”, Msc thesis, Addis Ababa University, Ethiopia, 2002
- [46] Nega Alemayehu and Peter Willett, “Stemming of Amharic Words for Information Retrieval”, Literary and Linguistic computing, 17(1): 1-17, 2002.
- [47] Jurafsky Daniel, H., James. Speech and Language Processing: “An introduction to natural language processing, computational linguistics and speech recognition”, June 25, 2007.
- [48] Martha Yifru, “Morphology-Based Language Modeling for Amharic”, 2010
- [49] Abeba Ibrahim, “A hybrid approach to Amharic base phrase chunking and parsing”, Master’s thesis, Addis Ababa University, 2013
- [50] ተባባሪ ፕሮፌሰር ኔታሁን አማራ, “የአማርኛ ስዎስውስ ቀላል አቀራረብ”, 1989.
- [51] Baye Yimam, “Root reduction and extensions in Amharic”, Ethiopian Journal of Language and Literature, 1999
- [52] Tigist Tensou Tessema, “Word Sequence Prediction for Amharic Language”, Master’s thesis, Addis Ababa University, Ethiopia, 2014.
- [53] Amare Getahun.: “ዘመናዊ የአማርኛ ስዎስውስ ቀላል አቀራረብ (Modern Amharic Grammar in a Simple Approach)”, Addis Ababa, Ethiopia, 2010
- [54] Medium. 2020. Neural MT. [online] Available at: <<https://towardsdatascience.com/neural-machine-translation-15ecf6b0b>> [Accessed 9 October 2020].
- [55] Africa.upenn.edu. 2020. Afaan Oromo. [online] Available at: <http://www.africa.upenn.edu/Hornet/Afaan_Oromo_19777.html> [Accessed 9 October 2020].
- [56] En.wikipedia.org. 2020. Regions Of Ethiopia. [online] Available at: <https://en.wikipedia.org/wiki/Regions_of_Ethiopia> [Accessed 9 October 2020].

- [57] E. Note, M. Bulcha, B. Deressa, T. Gamta, and B. Reviews, “Oromo Studies”, vol. I, no. 1, 1993.
- [58] Smartling.com. 2020. History OfMT. [online] Available at: <<https://www.smartling.com/resources/101/a-brief-history-of-machine-translation/>> [Accessed 9 October 2020].
- [59] Juran KrishnaSarkhel, “Approaches to MT”, 2014.
- [60] 2020. Stemming Algorithms. [online] Available at: <<http://citeseer.nj.nec.com/hull96stemming.html>> [Accessed 9 October 2020].
- [61] Komishinii Aadaaf Turizmii Oromiyaa, Gumii Qormaata Afaan Oromoo, “Caasluga Afaan Oromoo, Jildi I”, Finfinnee, Ethiopia, 1995.
- [62] Kula Kekeba Tune and Vasudeva Varma, “Oromo-English Information Retrieval Experiments at CLEF”, Language Technologies Research Centre, 2007
- [63] Debela Tesfaye “Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach”, Master’s thesis, Addis Ababa University, 2010
- [64] Wakweya Olani “Inflectional Morphology in Oromo”, Master’s Thesis, Addis Ababa University, 2014
- [65] Catherine Griefenow-Mewis, “A Grammatical sketch of Written Oromo”, Germany: Koln, 2001.
- [66] Gragg, G. Oromo of Wallaga. In M.L Bender (ed.) The Non-Semitic Languages of Ethiopia. Michigan: African Studies Center, 166-195, 1976
- [67] Mohammed Ali and A. Zaborski, “Handbook of the Oromo Language”, Stuttgart: Steiner, 1990
- [68] Nordfeldt, M., “A [Oromo] Grammar”, Lund: Le Monde Orientale, 1947

[69] Ishetu Kebede, “The Verb To Be In Oromo”, MA Thesis, Department of Linguistics, Addis Ababa University, 1981.

[70] Jacob Elming, “Syntactic Reordering In Statistical MT”, Copenhagen Business School, a PhD thesis, June 2008

[71] Yehenew Shiferaw, “Design and development of human-aided rule-based english-amharic MT”, MA Thesis, June 2004

[72] 2020. Man To Machine A Tutorial On The Art Of MT. [online] Available at: <<http://www.slideshare.net/jaganadhg/a-tutorialon-machine-translation>> [Accessed 10 October 2020].

[73] Dawit Mulugeta, “Geez to Amharic Automatic MT: A Statistical Approach”, Master’s thesis, Addis Ababa University, Ethiopia, 2015.

[74] 2007. Example-Based MT (EBMT). [online] Available at: <<http://www.cs.cmu.edu/afs/cs/user/alavie/11-731/731-cmt/www/ebmt2007.pdf>> [Accessed 10 October 2020].

[75] Hutchinsweb.me.uk. 2020. Example Based MT – A Review And Commentary. In Recent Advances In Example-Based MT.. [online] Available at: <<http://www.hutchinsweb.me.uk/MTJ->> [Accessed 10 October 2020].

[76] Biruk Abel, “Geez to Amharic MT”, Master’s thesis, Addis Ababa University, Ethiopia, 2018.

[77] 2003. Statistical Phrase Based MT. [online] Available at: <<http://www.aclweb.org/anthology/N/N03/N03-1017.pdf>> [Accessed 10 October 2020].

[78]. Zens R, Och F J and Ney H, “Phrase based MT”, Lecture Notes in Computer Science; Springer (2002), pp. 35-56.

[79] 2010. What Can Syntax-Based MT Learn From Phrase-Based MT?. [online] Available at: <<http://www.isi.edu/natural-language/mt/ats-vs-ghkm>> [Accessed 10 October 2020].

[80] 2010. A Syntax-Based Statistical Translation Model. [online] Available at: <<http://www.aclweb.org/anthology/P/P01/P01-1067>> [Accessed 10 October 2020].

- [81] V. Machines and H. Karlbom, “Hybrid MT Choosing the best translation with Support”, September, 2016.
- [82] Dzmitry Bahdanau, KyungHyun Cho, “Neural MT by jointly learning to align and translate”, ICLR, 2015
- [83]2020. Recurrent Continuous Translation Models. [online] Available at: <<http://www.aclweb.org/anthology/D13-1176>> [Accessed 10 October 2020].
- [84]K. Cho et al., “Learning Phrase Representations using RNN Encoder-Decoder for SMT”, In Proceedings of the Conf. on EMNLP, October 25-29, Doha, Qatar, 2014, pp. 1724-1734.
- [85]Sutskever, I., Vinyals, O., & Le, Q. V., “Sequence to sequence learning with neuralnetworks”, In Advances in neural information processing systems, 2014.
- [86]A. Lamb and M. Xie, “Convolutional Encoders for Neural MT”, pp. 1–8, 2016
- [87] Y. Wu et al., “Google’s Neural MT System : Bridging the Gap between Human and MT” pp. 1–23, 2016.
- [88] D. Bahdanau, K. Cho, and Y. Bengio, “Neural MT”, pp. 1– 15, 2015.
- [89] G. Neubig, “Neural MT and Sequence-to-sequence Models : A Tutorial” pp. 1–65, 2017
- [90] Y. Liu, L. Ji, R. Huang, T. Ming, and C. Gao, “An attention-gated convolutional neural network for sentence classification,” pp. 1–19, 2018
- [91] Niessen, S., F.J Och, G. Leusch, and H. Ney, “An Evaluation Tool for MT: Fast Evaluation for MT Research”, in Proceesing of the 2nd InternationalConference on Language Resources and Evaluation, Athens, Greece. 2000.
- [92] En.wikipedia.org. 2020. Evaluation OfMT. [online] Available at: <https://en.wikipedia.org/wiki/Evaluation_of_machine_translation> [Accessed 10 October 2020].
- [93] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of MT”, ACL, 2002

- [94] Somers, H. "Round-Trip Translation: What Is It Good For?", Proceedings of the Australasian Language Technology Workshop, 2005
- [95] Somers, H., Gaspari, F. and Ana Niño, "Detecting Inappropriate Use of Free Online MT by Language Students - A Special Case of Plagiarism Detection". Proceedings of the 11th Annual Conference of the European Association of MT, Oslo University (Norway) pp. 41–48, 2006
- [96]ALPAC, "Languages and machines: computers in translation and linguistics", National Academy of Sciences, National Research Council, Washington, D.C., 1966.
- [97] Turian, J., Shen, L. and Melamed, I. D., "Evaluation of MT and its Evaluation", Proceedings of the MT Summit IX, 2003
- [98] White, J., O'Connell, T. and O'Mara, F. "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches", Proceedings of the 1st Conference of the Association for MT in the Americas. Columbia, 1994
- [99] White, J., "Approaches to Black Box MT Evaluation", Proceedings of MT Summit, 1995
- [100] Han, A.L.F., Wong, D.F., and Chao, L.S., "LEPOR: A Robust Evaluation Metric for MT with Augmented Factors", in Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), 2012
- [101] Nithya B and Shibily Joseph, "A Hybrid Approach to English to MalayalamMT" International Journal of Computer Applications, 2013
- [102] Unnikrishnan P , Antony P J and Dr. Soman K P, "A Novel Approach for English to South Dravidian Language Statistical MT System", International Journal on Computer Science and Engineering, 2010
- [103] Charu Verma, Aarti Singh, Swagata Seal, Varsha Singh, Iti Mathur, "Hindi-English Neural MT Using Attention Model", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 11, 2019
- [104]D.Do, M. Utiyama, and E. Sumita, "MT from Japanese and French to Vietnamese, the difference among language families", October 2015.

- [105] Sutskever, I., Vinyals, O., & Le, Q. V.. “Sequence to sequence learning with neural networks”, In Advances in neural information processing systems (pp. 3104-3112), 2014
- [106]2011. [online] Available at: <http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/> [Accessed 10 October 2020].
- [107] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, “Google’s Neural MT System: Bridging the Gap between Human and Machine Translation”, 2016
- [108]Samrawit Zewgneh, “English-Amharic Document Translation Using Hybrid Approach”, Master’s Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2017
- [109] Arnold D., Lorna B., Siety M., R. Lee H., Louisa S., “MT: an introduction guide”, NCC Blackwell, London, 1994
- [110] Mulubrhan Hailegebreal, “A Bidirectional Tigrigna – English Statistical MT”, Masters Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2017
- [111] Mohammed Hussen, “Part-of-speech tagging for Afaan Oromo language using Transformational Error driven Learning (TEL) approach”, Unpublished Master’s Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2010.
- [112] Dabala Goshu & Ronny Meyer, “Focus phenomena in Wellega Oromo”, Afrika und Übersee86: 161-197, 2003
- [113] Baye Yimam, “Oromo Substantive: Some Aspects of Morphology and Syntax”, MA. Thesis, Addis Ababa University, 1981
- [114]1-StopAsia. 2020. Neural Machine Translation - The Newest Technology Explained. [online] Available at: <<https://www.1stopasia.com/neural-machine-translation-the-newest-technology-explained/>> [Accessed 13 October 2020].
- [115] Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratn Shah, and Ponnurangam Kumaraguru, “Neural Machine Translation for English-Tamil” Jan 2019.
- [116] Brownlee, J., 2020.Encoder-Decoder Long Short-Term Memory Networks. [online] Machine Learning Mastery. Available at: <<https://machinelearningmastery.com/Encoder-Decoder-long-short-term-memory-networks/>> [Accessed 13 October 2020].
- [117] FloydHub Blog. 2020. Attention Mechanism. [online] Available at: <<https://blog.floydhub.com/attention-mechanism>> [Accessed 13 October 2020].

[118] Och, F. J., Tillmann, C., & Ney, H. (1999). Improved alignment models for statistical machine translation. In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[119] Gruber, N. and Jockisch, A., 2020. Are GRU Cells More Specific And LSTM Cells More Sensitive In Motive Classification Of Text?. [online] Available at: <<https://www.frontiersin.org/articles/10.3389/frai.2020.00040/full>> [Accessed 16 October 2020].

[120] Shuoheng Yang, Yuxin Wang, Xiaowen Chu. A Survey of Deep Learning Techniques for Neural Machine Translation. Feb 2020.

[121] Lilian Weng, Updated on 2020-04-07:]< Attention? Attention!><https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html> [Accessed 27 October 2020].

[122] Michael Gasser, Hornmorph User's Guide, 2012.

Appendix I: SAMPLE OF PARALLEL CORPUS

| | |
|---|--|
| ኢየሱስ ከስድስት ቀን በኋላ ጴጥሮስን፣ ያዕቆብንና ወንድሙን ዮሐንስን ብቻ ይዞ ወደ አንድ ረጅም ተራራ ወጣ። | Yesuus guyyaa ja'a booda Pheexiros, Yaaqoobii fi obboleessa Yaaqoob |
|---|--|

| | |
|--|--|
| | Yohaannisiin kophaa isaanii fudhatee gaara dheeraa tokkotti ol ba'e. |
| በፊታቸውም ተለወጠ፤ ፊቱም እንደ ፀሐይ አበራ፤ ልብሱም እንደ ብርሃን አንጸባረቀ። | Innis isaan duratti ni jijjiirame; fuulli isaa akka aduu ife; uffanni isaas akka ifaa calaaqqise. |
| ከዚያም ሙሴና ኤልያስ ከእሱ ጋር ሲነጋገሩ ታዩአቸው። | Yeroo kanatti, Musee fi Eliyaas Yesuusii wajjin utuu dubbatanii isaanitti mul'atan. |
| በዚህ ጊዜ ጴጥሮስ ኢየሱስን “ጌታ ሆይ፣ እዚህ መሆን ለእኛ መልካም ነው። ከፈለግክ በዚህ ስፍራ አንድ ለአንተ፣ አንድ ለሙሴና አንድ ለኤልያስ ሦስት ድንኳኖች እተክላለሁ” አለው። | Achiis Pheexiros Yesuusii, “Yaa Gooftaa, iddoo kana ta'uun keenya nuuf gaarii dha. Fedha kee yoo ta'e, asitti dunkaana sadii, tokko siif, tokko Museedhaaf, tokko immoo Eliyaasiif nan dhaaba” jedheen. |
| ገና እየተናገረ ሳለም ብሩህ ደመና ጋረዳቸው፤ ከደመናውም “በጣም የምደሰትበት፣ የምወደው ልጄ ይህ ነው። እሱን ስሙት” የሚል ድምፅ ተሰማ። | Utuma inni kana dubbachaa jiruu, duumessi ifa of keessaa qabu tokko dhufee isaan haguuge; duumessicha keessaas sagaleen tokko, “Kun Ilma koo isa jaallatamaa, isa ani itti gammadu dha. Isa dhaga'aa” jedhe. |
| ደቀ መዛሙርቱ ይህን ሲሰሙ በፍርሃት ተውጠው በግንባራቸው ተደፉ። | Barattoonnis yommuu kana dhaga'an baay'ee sodaatanii adda isaaniitiin gombifaman. |
| በዚህ ጊዜ ኢየሱስ ወደ እነሱ ቀርቦ ዳሰሳቸውና “ተነሱ። አትፍሩ” አላቸው። | Achiis Yesuus isaanitti siqee isaan qaqqabuudhaan, “Ka'aa. Hin sodaatinaa” isaaniin jedhe. |
| ቀና ብለው ሲመለከቱ ከኢየሱስ በቀር ሌላ ማንንም አላዩም። | Yommuu ol jedhanii ilaalanis, Yesuusii malee nama tokko illee hin argine. |
| ከተራራው እየወረዱ ሳሉ ኢየሱስ “የሰው ልጅ ከሞት እስኪነሳ ድረስ ራእዩን ለማንም እንዳትናገሩ” ሲል አዘዛቸው። | Gaaricha irraa utuu bu'aa jiranii Yesuus, “Hanga Ilmi namaa du'aa kaafamutti mul'ata kana eenyutti iyyuu hin himinaa” jedhee isaan ajaje. |
| ይሁን እንጂ ደቀ መዛሙርቱ “ታዲያ ጸሐፍት ኤልያስ በመጀመሪያ መምጣት አለበት የሚሉት ለምንድን ነው?” ሲሉ ጠየቁት። | Haa ta'u malee barattoonni isaa, “Maarree barsiisonni seeraa, Eliyaas dursee dhufuu qaba kan jedhan maaliifi?” jedhanii isa gaafatan. |
| እሱም መልሶ እንዲህ አለ:- “በእርግጥ ኤልያስ መጥቶ ሁሉንም ነገር ወደ ቀድሞ ሁኔታው ይመልሳል። | Innis deebisee akkana jedheen: “Dhugaa dha, Eliyaas dhufee waan hundumaa iddootti deebisa. |
| እኔ ግን እላችኋለሁ፣ ኤልያስ መምጣቱን መጥቷል፤ እነሱ ግን የፈለጉትን ነገር አደረጉበት እንጂ አላወቁትም። የሰው ልጅም እንደዚሁ በእነሱ እጅ ይሠቃያል።” | Haa ta'u malee, ani isinitti nan hima, Eliyaas dhufee; isaan garuu wanta barbaadan isa irratti raawwatan malee isa hin beekne. Ilmi namaas akkasuma harka isaaniitti dhiphachuuf jira.” |
| በዚህ ጊዜ ደቀ መዛሙርቱ ኢየሱስ የነገራቸው ስለ መጥምቁ ዮሐንስ እንደሆነ ገባቸው። | Yeroo kanatti barattoonni isaa inni waa'ee Yohaannis Cuuphaa akka isaanitti dubbate hubatan. |
| ወደ ሕዝቡ በመጡ ጊዜ አንድ ሰው ወደ እሱ ቀረበና ተንበርክኮ እንዲህ አለው:- | Yommuu gara namootaa dhufanis, namichi tokko gara isaa dhufee isa duratti jilbeenfatee akkana jedheen: |
| “ጌታ ሆይ፣ ለልጄ ምሕረት አድርግለት፤ የሚጥል በሽታ ስላለበት በጠና ታሟል። አንዴ እሳት ውስጥ አንዴ ደግሞ ውኃ ውስጥ ይወድቃል። | “Yaa Gooftaa, ilma kootiif garaa naa laafi; inni dhukkuba gaggabdootiin qabamee rakkachaa |

| | |
|---|--|
| | jira. Yeroo baay'ee ibidda keessatti kufa; yeroo baay'ees bishaan keessatti kufa. |
| ወደ ደቀ መዛሙርትህ አመጣሁት፤ እነሱ ግን ሊፈውሱት አልቻሉም።” | Gara barattoota keetti isa fideen ture; isaan garuu isa fayyisuu hin dandeenye.” |
| ኢየሱስም መልሶ “እምነት የለሽና ጠማማ ትውልድ ሆይ፤ ከእናንተ ጋር እስከ መቼ መቆየት ሊኖርብኝ ነው? እስከ መቼስ እናንተን መታገሥ ሊኖርብኝ ነው? ልጁን ወደ እኔ አምጡት” አለ። | Yesuusiis deebisee, “Dhaloota amantiin hin qabnee fi micciiramaa nana, hanga yoomiittan isinii wajjin tura? Hanga yoomiittanis isiniif obsa? Mucicha as natti fidaa” jedhe. |
| ከዚያም ኢየሱስ ጋኔትን ገሠጸው፤ ጋኔትም ከልጁ ወጣ፤ ልጁም ከዚያች ሰዓት ጀምሮ ተፈወሰ። | Achiis Yesuus jinnii sana ifate; hafuurichis mucicha keessaa ba'e, mucichis sa'aatii sanaa jalqabee ni fayye. |
| ከዚህ በኋላ ደቀ መዛሙርቱ ብቻቸውን ወደ ኢየሱስ መጥተው “እኛ ልናስወጣው ያልቻልነው ለምንድን ነው?” አሉት። | Sana booda barattoonni isaa kophaa isaanii gara Yesuus dhufanii, “Nuti hafuuricha baasuu kan dadhabne maaliifi?” jedhaniin. |
| እሱም “እምነታችሁ ስላነሰ ነው። እውነት እላችኋለሁ፤ የሰናፍጭ ቅንጣት የምታክል እምነት ካላችሁ ይህን ተራራ ‘ከዚህ ተነስተህ ወደዚያ ሂድ’ ብትሉት ይሄዳል፤ የሚሳናችሁም ነገር አይኖርም” አላቸው። | Innis, “Amantiin keessan xinnoo waan ta'eefi. Dhuguman isiniin jedha, amantiin hanga firii sanaaficaa geessu utuu qabaattanii, gaara kanaan, ‘Asii ka'itii achi deemi’ yoo jettan ni deema; wanti isin hin dandeenyes hin jiraatu” isaaniin jedhe. |
| በገሊላ ተሰብስበው ሳሉ ኢየሱስ እንዲህ አላቸው፡- “የሰውን ልጅ ለሰዎች አሳልፈው ይሰጡታል፤ | Galiilaatti utuu walga'anii jiraniis, Yesuus akkana isaaniin jedhe: “Ilmi namaa harka namootaatti dabarfamee ni kennama; |
| እነሱም ይገድሉታል፤ እሱም በሦስተኛው ቀን ይነሳል።” ደቀ መዛሙርቱም በጣም አዘኑ። | isaanis isa ni ajjeesu; innis guyyaa sadaffaatti du'aa ni kaafama.” Yeroo kanatti barattoonni isaa baay'ee gaddan. |
| ቅፍርናሆም ከደረሱ በኋላ የቤተ መቅደሱን ግብር የሚሰበስቡት ሰዎች ወደ ጴጥሮስ ቀርበው “መምህራችሁ የቤተ መቅደሱን ግብር አይከፍልም?” አሉት። | Qifirnaahom erga ga'anii booda, namoonni gibira mana qulqullummaa sassaaban gara Pheexiros dhufanii, “Barsiisaan keessan gibira mana qulqullummaa hin kaffaluu?” jedhan. |
| እሱም “ይከፍላል” አላቸው። ይሁን እንጂ ወደ ቤት በገባ ጊዜ ኢየሱስ በቅድሚያ እንዲህ አለው፡- “ስምዎን ምን ይመስልሃል? የምድር ነገሥታት ቀረጥ ወይም ግብር የሚቀበሉት ከማን ነው? ከልጆቻቸው ወይስ ከሌሎች?” | Innis, “Ni kaffala” jedhe. Haa ta'u malee, yeroo Pheexiros mana seenu Yesuus dursee, “Simoon, maal sitti fakkaata? Mootonni lafaa gibira yookiin qaraxa eenyu irraa fudhatu? Ijoollee isaanii irraa moo keessummoota irraa ti?” jedheen. |
| እሱም “ከሌሎች” ብሎ ሲመልስለት ኢየሱስ እንዲህ አለው፡- “እንግዲያው ልጆቹ ግብር ከመክፈል ነፃ ናቸው። | Yommuu Pheexiros “Keessummoota irraa ti” jedhee deebisu, Yesuus akkana jedheen: “Yoos immoo ijoolleen gibira kaffaluu irraa walaba dha jechuu dha. |
| ሆኖም እንቅፋት እንዳንሆንባቸው ወደ ባሕሩ ሄደህ መንጠቆ ጣል፤ ከዚያም መጀመሪያ የምትይዘውን ዓሳ አፋን ስትከፍት አንድ የብር ሳንቲም ታገኛለህ። ሳንቲሙን ወስደህ ለእኔና ለአንተ ክፈል።” | Garuu nuti gufuu akka isaanitti hin taaneef, gara galaanaa dhaqitiin hokkoo ittiin qurxummii qabattu darbadhu; achiis qurxummii isa jalqaba qabattu yommuu afaan isaa bantu, saantima meetii tokko ni argatta. Isa fuudhiitii anaafis ofii keetiifis kaffali.” |

| | |
|---|--|
| ኢየሱስ ይህን ተናግሮ ከጨረሰ በኋላ ከገሊላ ወጥቶ ከዮርዳኖስ ማዶ ወዳሉት የይሁዳ ድንበሮች መጣ። | Yesuus wantootaa kana dubbatee yommuu xumuru, Galiilaadhaa ka'ee, gara daangaawwan Yihudaa warra Yordaanos gama jiranii dhufe. |
| እጅግ ብዙ ሰዎችም ተከተሉት፤ እሱም በዚያ ፈወሳቸው። | Yeroo kanatti namoonni hedduun isa duukaa kan bu'an yoo ta'u, innis achitti isaan fayyise. |
| ፈሪሳውያንም ወደ እሱ መጥተው ሊፈትኑት በማሰብ “አንድ ሰው በማንኛውም ምክንያት ሚስቱን እንዲፈታ ሕግ ይፈቅድለታል?” ሲሉ ጠየቁት። | Fariisannis isa qoruuf yaaduudhaan gara isaa dhufanii, “Namni tokko sababiidhuma ta'een haadha manaa isaa akka hiiku seerri isaaf ni heyyamaa?” jedhanii isa gaafatan. |
| እሱም መልሶ እንዲህ አለ፡- “ፈጣሪ ከመጀመሪያውም ወንድና ሴት አድርጎ እንደፈጠራቸው አላነበባችሁም?” | Innis deebisee akkana jedhe: “Waaqayyo jalqabumaa kaasee dhiiraa fi dubartii godhee akka isaan uume hin dubbisnee?” |

APPENDIX II: THE LAST EPOCH RESULTS WITH LOSS LEVEL AND A TIME TO TAKEN FOR 163 BATCHES.

| | |
|--------------------------------|---------------------------------|
| Epoch 55 Batch 0 Loss 0.022612 | Epoch 55 Batch 39 Loss 0.026516 |
| Epoch 55 Batch 1 Loss 0.019935 | Epoch 55 Batch 40 Loss 0.023581 |

| | |
|---------------------------------|----------------------------------|
| Epoch 55 Batch 2 Loss 0.030306 | Epoch 55 Batch 41 Loss 0.028616 |
| Epoch 55 Batch 3 Loss 0.019593 | Epoch 55 Batch 42 Loss 0.018198 |
| Epoch 55 Batch 4 Loss 0.015300 | Epoch 55 Batch 43 Loss 0.022614 |
| Epoch 55 Batch 5 Loss 0.034801 | Epoch 55 Batch 44 Loss 0.024303 |
| Epoch 55 Batch 6 Loss 0.019256 | Epoch 55 Batch 45 Loss 0.021562 |
| Epoch 55 Batch 7 Loss 0.020439 | Epoch 55 Batch 46 Loss 0.018507 |
| Epoch 55 Batch 8 Loss 0.023125 | Epoch 55 Batch 47 Loss 0.022773 |
| Epoch 55 Batch 9 Loss 0.026950 | Epoch 55 Batch 48 Loss 0.024186 |
| Epoch 55 Batch 10 Loss 0.024589 | Epoch 55 Batch 49 Loss 0.016637 |
| Epoch 55 Batch 11 Loss 0.029814 | Epoch 55 Batch 50 Loss 0.026707 |
| Epoch 55 Batch 12 Loss 0.025414 | Epoch 55 Batch 51 Loss 0.017651 |
| Epoch 55 Batch 13 Loss 0.032700 | Epoch 55 Batch 52 Loss 0.025844 |
| Epoch 55 Batch 14 Loss 0.032776 | Epoch 55 Batch 53 Loss 0.024368 |
| Epoch 55 Batch 15 Loss 0.034641 | Epoch 55 Batch 54 Loss 0.030082 |
| Epoch 55 Batch 16 Loss 0.034385 | Epoch 55 Batch 55 Loss 0.025968 |
| Epoch 55 Batch 17 Loss 0.021293 | Epoch 55 Batch 56 Loss 0.020050 |
| Epoch 55 Batch 18 Loss 0.021591 | Epoch 55 Batch 57 Loss 0.016511 |
| Epoch 55 Batch 19 Loss 0.017303 | Epoch 55 Batch 58 Loss 0.018534 |
| Epoch 55 Batch 20 Loss 0.026813 | Epoch 55 Batch 59 Loss 0.034844 |
| Epoch 55 Batch 21 Loss 0.027987 | Epoch 55 Batch 60 Loss 0.023615 |
| Epoch 55 Batch 22 Loss 0.025438 | Epoch 55 Batch 61 Loss 0.027566 |
| Epoch 55 Batch 23 Loss 0.027117 | Epoch 55 Batch 62 Loss 0.018883 |
| Epoch 55 Batch 24 Loss 0.020781 | Epoch 55 Batch 63 Loss 0.018252 |
| Epoch 55 Batch 25 Loss 0.034669 | Epoch 55 Batch 64 Loss 0.030944 |
| Epoch 55 Batch 26 Loss 0.021480 | Epoch 55 Batch 65 Loss 0.024011 |
| Epoch 55 Batch 27 Loss 0.014888 | Epoch 55 Batch 66 Loss 0.021740 |
| Epoch 55 Batch 28 Loss 0.019768 | Epoch 55 Batch 67 Loss 0.022235 |
| Epoch 55 Batch 29 Loss 0.030496 | Epoch 55 Batch 68 Loss 0.024909 |
| Epoch 55 Batch 30 Loss 0.035808 | Epoch 55 Batch 69 Loss 0.020738 |
| Epoch 55 Batch 31 Loss 0.021373 | Epoch 55 Batch 70 Loss 0.029248 |
| Epoch 55 Batch 32 Loss 0.025391 | Epoch 55 Batch 71 Loss 0.029843 |
| Epoch 55 Batch 33 Loss 0.038369 | Epoch 55 Batch 72 Loss 0.028153 |
| Epoch 55 Batch 34 Loss 0.031353 | Epoch 55 Batch 73 Loss 0.025690 |
| Epoch 55 Batch 35 Loss 0.029892 | Epoch 55 Batch 74 Loss 0.037451 |
| Epoch 55 Batch 36 Loss 0.020342 | Epoch 55 Batch 75 Loss 0.026202 |
| Epoch 55 Batch 37 Loss 0.030650 | Epoch 55 Batch 76 Loss 0.027383 |
| Epoch 55 Batch 38 Loss 0.026123 | Epoch 55 Batch 77 Loss 0.019998 |
| Epoch 55 Batch 78 Loss 0.029825 | Epoch 55 Batch 113 Loss 0.023970 |
| Epoch 55 Batch 79 Loss 0.024828 | Epoch 55 Batch 114 Loss 0.031283 |
| Epoch 55 Batch 80 Loss 0.016894 | Epoch 55 Batch 115 Loss 0.021852 |
| Epoch 55 Batch 81 Loss 0.035585 | Epoch 55 Batch 116 Loss 0.015415 |
| Epoch 55 Batch 82 Loss 0.022113 | Epoch 55 Batch 117 Loss 0.028828 |

| | |
|----------------------------------|----------------------------------|
| Epoch 55 Batch 83 Loss 0.026675 | Epoch 55 Batch 118 Loss 0.025532 |
| Epoch 55 Batch 84 Loss 0.025330 | Epoch 55 Batch 119 Loss 0.020534 |
| Epoch 55 Batch 85 Loss 0.015237 | Epoch 55 Batch 120 Loss 0.017933 |
| Epoch 55 Batch 86 Loss 0.032015 | Epoch 55 Batch 121 Loss 0.028519 |
| Epoch 55 Batch 87 Loss 0.029528 | Epoch 55 Batch 122 Loss 0.031279 |
| Epoch 55 Batch 88 Loss 0.030724 | Epoch 55 Batch 123 Loss 0.019246 |
| Epoch 55 Batch 89 Loss 0.025832 | Epoch 55 Batch 124 Loss 0.026116 |
| Epoch 55 Batch 90 Loss 0.031679 | Epoch 55 Batch 125 Loss 0.027826 |
| Epoch 55 Batch 91 Loss 0.022763 | Epoch 55 Batch 126 Loss 0.025162 |
| Epoch 55 Batch 92 Loss 0.026745 | Epoch 55 Batch 127 Loss 0.033733 |
| Epoch 55 Batch 93 Loss 0.028692 | Epoch 55 Batch 128 Loss 0.023237 |
| Epoch 55 Batch 94 Loss 0.025873 | Epoch 55 Batch 129 Loss 0.030318 |
| Epoch 55 Batch 95 Loss 0.026786 | Epoch 55 Batch 130 Loss 0.023102 |
| Epoch 55 Batch 96 Loss 0.020712 | Epoch 55 Batch 131 Loss 0.026452 |
| Epoch 55 Batch 97 Loss 0.026458 | Epoch 55 Batch 132 Loss 0.023335 |
| Epoch 55 Batch 98 Loss 0.024913 | Epoch 55 Batch 133 Loss 0.027048 |
| Epoch 55 Batch 99 Loss 0.034014 | Epoch 55 Batch 134 Loss 0.024235 |
| Epoch 55 Batch 100 Loss 0.023481 | Epoch 55 Batch 135 Loss 0.026792 |
| Epoch 55 Batch 101 Loss 0.019513 | Epoch 55 Batch 136 Loss 0.029464 |
| Epoch 55 Batch 102 Loss 0.025351 | Epoch 55 Batch 137 Loss 0.020033 |
| Epoch 55 Batch 103 Loss 0.022305 | Epoch 55 Batch 138 Loss 0.030387 |
| Epoch 55 Batch 104 Loss 0.025315 | Epoch 55 Batch 139 Loss 0.035456 |
| Epoch 55 Batch 105 Loss 0.023254 | Epoch 55 Batch 140 Loss 0.019982 |
| Epoch 55 Batch 106 Loss 0.026191 | Epoch 55 Batch 141 Loss 0.035147 |
| Epoch 55 Batch 107 Loss 0.024624 | Epoch 55 Batch 142 Loss 0.022067 |
| Epoch 55 Batch 108 Loss 0.029276 | Epoch 55 Batch 143 Loss 0.031032 |
| Epoch 55 Batch 109 Loss 0.019359 | Epoch 55 Batch 144 Loss 0.024076 |
| Epoch 55 Batch 110 Loss 0.028335 | Epoch 55 Batch 145 Loss 0.020744 |
| Epoch 55 Batch 111 Loss 0.020892 | Epoch 55 Batch 146 Loss 0.026939 |
| Epoch 55 Batch 112 Loss 0.022601 | Epoch 55 Batch 147 Loss 0.023274 |
| Epoch 55 Batch 148 Loss 0.023724 | Epoch 55 Batch 156 Loss 0.033226 |
| Epoch 55 Batch 149 Loss 0.025865 | Epoch 55 Batch 157 Loss 0.029772 |
| Epoch 55 Batch 150 Loss 0.031250 | Epoch 55 Batch 158 Loss 0.017353 |
| Epoch 55 Batch 151 Loss 0.018793 | Epoch 55 Batch 159 Loss 0.021683 |
| Epoch 55 Batch 152 Loss 0.020413 | Epoch 55 Batch 160 Loss 0.024364 |

| | |
|----------------------------------|--|
| Epoch 55 Batch 153 Loss 0.022950 | Epoch 55 Batch 161 Loss 0.028058 |
| Epoch 55 Batch 154 Loss 0.029428 | Epoch 55 Batch 162 Loss 0.018112 |
| Epoch 55 Batch 155 Loss 0.027573 | Epoch 55 Batch 163 Loss 0.023591 Time taken for this epoch 146.2265121936798 sec |

APPENDIX III: SAMPLE OUTPUT

```
translate(u'እጁንም ከጫነባቸው በኋላ ከዚያ ስፍራ ተነስቶ ሄደ')
```

Input in Amharic Sentence: <start> እጁንም ከጫነባቸው በኋላ ከዚያ ስፍራ ተነስቶ ሄደ <end>

Predicted Oromic equivalent translation: innis harka isaa erga isaan irra kaa4ee booda bakka sana dhiisee deeme <end>

BLEU score for this translation: 0.5847065326973129

/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:

Corpus/Sentence contains 0 counts of 2-gram overlaps.

BLEU scores might be undesirable; use SmoothingFunction().

```
warnings.warn(_msg)
```

```
translate(u'እርግጥ ነው፣ ፍጹም የሆነ ሰው የለም።')
```

Input in Amharic Sentence: <start> እርግጥ ነው ፍጹም የሆነ ሰው የለም <end>

Predicted Oromic equivalent translation: namni mudaa hin qabne akka hin jirre beekamaa dha <end>

BLEU score for this translation: 0.6147881529512643

/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:

Corpus/Sentence contains 0 counts of 2-gram overlaps.

BLEU scores might be undesirable; use SmoothingFunction().

```
warnings.warn(_msg)
```

```
translate(u'እርግጥ እንዲህ ማድረግ ሁልጊዜ ቀላል ላይሆን ይችላል።')
```

Input in Amharic Sentence: <start> እርግጥ እንዲህ ማድረግ ሁልጊዜ ቀላል ላይሆን ይችላል <end>

Predicted Oromic equivalent translation: kana gochuun yeroo hunda salphaa ta4uu dhiisuu danda4a <end>

BLEU score for this translation: 0.6363082311973168

/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:

Corpus/Sentence contains 0 counts of 2-gram overlaps.

BLEU scores might be undesirable; use SmoothingFunction().

```
translate(u'እነሱም "ጌታ ሆይ፣ ይህን ምግብ ሁልጊዜ ስጠን" አሉት።')
```

```
Input in Amharic Sentence: <start> እነሱም ጌታ ሆይ ይህን ምግብ ሁልጊዜ ስጠን አሉት <end>
```

```
Predicted Oromic equivalent translation: isaanis yaa gooftaa yeroo hunda nyaata kana nuu kenni jedhaniin <end>
```

```
BLEU score for this translation: 0.6147881529512643
```

```
/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:
```

```
Corpus/Sentence contains 0 counts of 2-gram overlaps.
```

```
BLEU scores might be undesirable; use SmoothingFunction().
```

```
warnings.warn(_msg)
```

```
translate(u'ጋብቻ በታሪክ ዘመናት ሁሉ የነበረ ነገር ነው።')
```

```
Input in Amharic Sentence: <start> ጋብቻ በታሪክ ዘመናት ሁሉ የነበረ ነገር ነው <end>
```

```
Predicted Oromic equivalent translation: gaa4elli kutaa jireenyaa ti <end>
```

```
BLEU score for this translation: 0.7172835948406505
```

```
/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:
```

```
Corpus/Sentence contains 0 counts of 2-gram overlaps.
```

```
BLEU scores might be undesirable; use SmoothingFunction().
```

```
warnings.warn(_msg)
```

```
translate(u'በጭም እንዲህ አለደረገችም።')
```

```
Input in Amharic Sentence: <start> በጭም እንዲህ አለደረገችም <end>
```

```
Predicted Oromic equivalent translation: lakki hin deebine <end>
```

```
BLEU score for this translation: 0.7348889200874658
```

```
/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:
```

```
Corpus/Sentence contains 0 counts of 2-gram overlaps.
```

```
BLEU scores might be undesirable; use SmoothingFunction().
```

```
translate(u'ቶማስም መልሶ ጌታዬ አምላኬ!" አለው።')
```

```
Input in Amharic Sentence: <start> ቶማስም መልሶ ጌታዬ አምላኬ አለው <end>  
Predicted Oromic equivalent translation: toomaasis deebisee gooftaa ko waaqa ko jedheen <end>  
BLEU score for this translation: 0.6419358938757569  
/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:  
Corpus/Sentence contains 0 counts of 2-gram overlaps.  
BLEU scores might be undesirable; use SmoothingFunction().  
warnings.warn(_msg)
```

```
translate(u'ይህን የተናገረው በቅርብ ጊዜ ውስጥ ምን ዓይነት አሟሟት እንደሚሞት ለማመልከት ነው።')
```

```
Input in Amharic Sentence: <start> ይህን የተናገረው በቅርብ ጊዜ ውስጥ ምን ዓይነት አሟሟት እንደሚሞት ለማመልከት ነው <end>  
Predicted Oromic equivalent translation: inni kana erga dubbatee booda garuu addaam dubbate <end>  
BLEU score for this translation: 0.6303647413359293  
/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:  
Corpus/Sentence contains 0 counts of 2-gram overlaps.  
BLEU scores might be undesirable; use SmoothingFunction().  
warnings.warn(_msg)
```

```
translate(u'ይህም የሆነው "ከሰጠኝን ከእነዚህ መካከል አንዱም እንኳ አልጠፋብኝም" ብሎ የተናገረው ቃል ይፈጸም ዘንድ ነው።')
```

```
Input in Amharic Sentence: <start> ይህም የሆነው ከሰጠኝን ከእነዚህ መካከል አንዱም እንኳ አልጠፋብኝም ብሎ የተናገረው ቃል ይፈጸም ዘንድ ነው <end>  
Predicted Oromic equivalent translation: kunis kan ta4e wanti inni du4aaf ni baate jedhee dubbate akka raawwatuufi <end>  
BLEU score for this translation: 0.5909442849243819  
/usr/local/lib/python3.6/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:  
Corpus/Sentence contains 0 counts of 2-gram overlaps.  
BLEU scores might be undesirable; use SmoothingFunction().
```