Addis Ababa University

College of Natural and Computational Science

Bidirectional Amharic-Afaan Oromo Machine Translation Using Hybrid Approach

Gelan Tulu Heyi

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

Addis Ababa, Ethiopia

March 2020

Addis Ababa University

College of Natural Sciences

Gelan Tulu Heyi

Advisor: *Dida Midekso (Dr.)*

This is to certify that the thesis prepared by *Gelan Tulu*, titled: *Bidirectional Amharic-Afaan Oromo Machine Translation Using Hybrid Approach* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:			
Examiner:			
Examiner:			

Abstract

Machine translation is the area of Natural Language Processing (NLP) that focuses on obtaining a target language text from a source language text by means of automatic techniques. Machine translation is a multidisciplinary field and the challenge has been approached from various points of view including linguistic and statistics. Hybrid methods focus on combining the best properties of two or more machine translation approaches. Nowadays, it has become very popular to include rules in statistical machine translation approaches.

In this study, a bidirectional Amharic-Afaan Oromo machine translation system using hybrid approach has been developed. 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. Since there are no publicly available POS tagger tools for both Amharic and Afaan Oromo languages, tagged corpus is prepared manually. The linguistic background and nature of the two languages have been studied in order to design the reordering rules for different types of Amharic/Afaan Oromo phrases and sentences. Language models 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.

To check the accuracy of the system, two experiments were conducted 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.

Keywords: Machine Translation, Statistical Machine Translation, Hybrid Machine Translation, Reordering Rule.

Acknowledgments

First and foremost, I am very thankful to the Almighty God for entitling me to this opportunity. I would like to express my deepest gratitude to my advisor, Dida Midekso (PhD), without his constructive comments and inspiring suggestions this research work wouldn't have been possible.

My deepest gratitude goes to my family, especially my mother Tsige Demissie, for their unconditional love and endless motivation during the course of my study. Finally, all my friends deserve special thanks.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Acronyms and Abbreviations

BCE	Before Common Era
BLEU	Bilingual Evaluation Understudy
CBMT	Corpus Based Machine Translation
CLIR	Cross Language Information Retrieval
CW	Compound Word
EBMT	Example Based Machine Translation
LM	Language Model
LSTM	Long-Short Term Memory
MLE	Maximum Likelihood Estimate
NLP	Natural Language Processing
NMT	Neural Machine Translation
POS	Part of Speech Tagging
RBMT	Rule Based Machine Translation
RNN	Recuurent Neural Network
SMT	Statistical Machine Translation
TM	Translation Model

CHAPTER ONE: INTRODUCTION

1.1 Background

Machine translation is a branch of computational linguistics and is defined as an automatic process by computerized system that converts a piece of text (written or spoken) from one natural language referred to as a source language to another natural language called the target language with human intervention or not and with the objective of restoring the meaning of the original text in the translated text [1].

Machine translation systems can be designed either specifically for two particular languages, called a bilingual system, or for more than a single pair of languages, called a multilingual system. A bilingual system may be either unidirectional, from one source language into one target language, or may be bidirectional. Multilingual systems are usually designed to be bidirectional, but most bilingual systems are unidirectional.

Different approaches for machine translation were defined and gained maturity for practical use today. The main approaches to building a machine translation tools are: knowledge driven approach also known as Rule Based Machine Translation (RBMT), data driven machine translation approach which is also known as Corpus Based Machine Translation (CBMT), hybrid machine translation approach which combines the advantages of the RBMT and CBMT approaches and Neural Machine Translation (NMT) which emerged as a successor of corpus based machine translation.

RBMT generates output based on linguistic rules and language order, morphological, syntactic and semantic analysis of both the source and the target language. The RBMT systems follow various approaches for translation namely; direct approach, transfer approach and interlingua approach [2]. However, RBMT techniques are less accurate due to the difficulty in incorporating rule interaction in big systems, ambiguity and idiomatic expressions. The complexity of creating RBMT system paved way for developing other machine translation approaches like corpus based machine translation and hybrid machine translation.

CBMT requires huge amount of parallel corpus to ensure translation of the source language sentences to the target language sentences. The two major categories of CBMT are Statistical Machine Translation (SMT) and Example Based Machine Translation (EBMT). SMT uses parallel corpus to calculate the order of words in both the source and target languages using mathematical

statistical probability. EBMT systems use the sample sentences stored in the database for translation of new sentences.

The Hybrid approach of machine translation utilizes properties of RBMT and SMT. Some Hybrid systems use a rule based approach followed by correction of output using statistical information. On the other hand, in some Hybrid systems statistical preprocessing is done followed by correction using transfer rules.

CBMT systems fail to provide accurate translations between language pairs with significant grammatical differences. Thus, the emerging research in machine translation has turned towards Neural Machine Translation (NMT). Neural machine translation is a new architecture that aims at building a single neural network that can be jointly turned to maximize the translation performance. This neural network is trained using deep learining techiniques. NMT requires a very large number parallel corpus to train the network. This requirement hinders the applicability of NMT for language pairs that lack huge parallel corpus.

1.2 Motivation

Ethiopia has more than 80 languages spoken within the country. Amharic and Afaan Oromo are the two principal languages spoken in the country [3]. Due to a large number of speakers of Amharic and Afaan Oromo, need of translations from Amharic to Afaan Oromo and vice versa is highly increasing from time to time. This motivated us to study and investigate the development of bidirectional Amharic – Afaan Oromo machine translation system.

1.3 Statement of the Problem

Amharic is an Afro-Asiatic language of the Semitic group which is widely spoken in Ethiopia. Of the Kushitic languages spoken in Ethiopia, Afaan Oromo is the language with the largest number of speakers. Currently 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.

Bidirectional machine translation systems for different language pairs have been developed over the years. Most of the studies have been done on language pairs of English and the other languages. For instance, Filipino-English [4], Myanmar-English [5], English-Amharic [6], English-Afaan Oromo [7, 8]. Amharic-Tigrigna [9] is the only translation done on Ethiopian language pairs. To

the best of the researcher's knowledge there is no machine translation 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. These data would be valuable if they can be used by both language speakers. This calls for the development of bidirectional Amharic-Afaan Oromo translation system.

This study was attempted to answer the research question: What is the possible machine translation approach to overcome linguistic barriers and to address the knowledge among Amharic language and Afaan Oromo language speakers and users?

1.4 Objective of the Study

General Objective

The general objective of this research work is to design and develop a bidirectional Amharic – Afaan Oromo machine translation system using hybrid approach.

Specific Objectives

To fulfill the general objective, some specific objectives are identified. The specific objectives are:

- To review techniques and methodologies used for machine translation.
- To study syntactic structure and relationship of the language pair: Amharic and Afaan Oromo.
- To collect Amharic – Afaan Oromo bilingual parallel corpus.
- To develop a general architecture for bidirectional Amharic – Afaan Oromo machine translation using hybrid approach.
- To develop a prototype for the bidirectional Amharic – Afaan Oromo translation.
- To test and evaluate the performance of the prototype.

1.5 Methods of the Study

To achieve the objectives of the research, the following methods will be followed.

Literature Review

Systems and applications that are related to bidirectional machine translation in different language pair was reviewed. This consists of thesis, conference and journal articles, white papers and bidirectional systems developed for other languages. In addition, a discussion was made with

Amharic and Afaan Oromo language experts regarding the linguistic nature of the languages, like the grammatical structure and morphology of the languages.

Data Collection

Amharic-Afaan Oromo parallel corpus was collected from Fana Broadcasting Corporate News[1], some chapters of the Holy Bible and other simple sentences are used to perform the experiment. A total of 1402 parallel sentences were collected, out of which 1301 are used for traning and the rest parallel sentences i.e., 101 are used for testing.

Software Tools

For the development of bidirectional Amharic-Afaan Oromo machine translation prototype, the following tools will be used:

- Ubuntu 16.04: a complete desktop Linux operating system which is freely available and suitable for the Moses environment.
- Moses: a statistical machine translation system that allows to automatically train translation models for any language pair.
- Giza++: a toolkit to train word alignment models.
- MKCLS: a tool to train word classes by using a maximum-likelihood-criterion.
- IRSTLM: a language modeling toolkit.
- BLEU Score: to evaluate the performance of the system.
- Notepad: to make the corpus in system understandable format.
- Microsoft Office 2013: software for the documentation of the study.

Evaluation

Machine translation evaluation could be done by using manual or automatic evaluation methods. Manual evaluatin gives a better result in order to measure the quality of machine translation and to analyze the errors within the system output. The most challenging issues in conducting human evaluation of machine translation output are high costs and time consumption. Therefore automatic methods like Bilingual Evaluation Understudy (BLEU) were proposed to measure the performance of machine translation. We used BLEU score metrics to evaluate the performance of the prototype.

[1] http://www.fanabc.com

1.6 Application of Results

The following are the main applications of this research work:

- The parallel corpus which is used for training and testing purpose in this work can be used in other NLP applications such as, named entity recognition, cross language information retrieval (CLIR) for Amharic – Afaan Oromo language pair.
- The translation of different reading materials can easily be accomplished for Amharic – Afaan Oromo language pair.
- The translation system can be used as a tool in teaching and learning process of the languages.

1.7 Scope and Limitation of the Study

The bidirectional Amharic – Afaan Oromo machine translation using hybrid approach is designed to translate simple sentence written in Amharic text into Afaan Oromo text and vice versa. Compound and complex sentences are not included in the study.

1.8 Organization of the Thesis

This section describes the organization of the rest of the research work. The next chapter presents literature review which briefly discusses about an overview of the Amharic and Afaan Oromo languages and different machine translation approaches. The Third Chapter presents the related works on machine translation done on different language pairs. Chapter Four presents the design of bidirectional Amharic-Afaan Oromo machine translation using hybrid approach. The experiments and results are discussed in Chapter Five and Chapter Six presents conclusion and future works.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

In this chapter, a brief overview of Amharic and Afaan Oromo languages and different machine translation approaches are discussed. The major Amharic and Afaan Oromo word classes, which are nouns, verbs, adjectives and adverbs are also discussed in this chapter.

2.2 A Brief Overview of Amharic Language

Amharic is the second most widely spoken Semitic language after Arabic and the first most widely spoken language in Ethiopia. Semitic languages were introduced into Ethiopia by migrants from Yemen who crossed the Red Sea in the first millennium Before Common Era (BCE) where they entered in contact with Cushitic speakers [10]. The earliest records of Ethiopic Semitic are in Ge'ez, the classical language of Ethiopia spoken once in the Christian kingdom of Aksum until medieval times. While Ge'ez was preserved for written communication and as a liturgical language, one of its descendants, Amharic, developed as a lingua franca for trade and everyday communication since the 17th century. Amharic has been greatly influenced by Cushitic languages, such as Afaan Oromo, not only in its lexicon, but also in syntax and typology [10].

Amharic is written with a version of the Ge'ez script known as Fidel [11]. Amharic has seven vowels [10], as shown in Table 2.1.

Table 2.1: Amharic Vowels

Vowel	ኧ/ *ä*/	ኡ/ u/	ኢ/ i/	ኣ/ *a*/	ኤ/ e/	እ/ ï/	ኦ/ o/
Sounds like	**a**gain	m**oo**n	f**ee**t	f**a**ther	W**a**y	p**i**n	W**a**r

The Amharic script contains thirty four basic symbols. Each of the thirty four basic symbols has seven symbols, one for each of the seven vowels of Amharic [11]. The Amharic syllabary is presented in Figure 2.1.

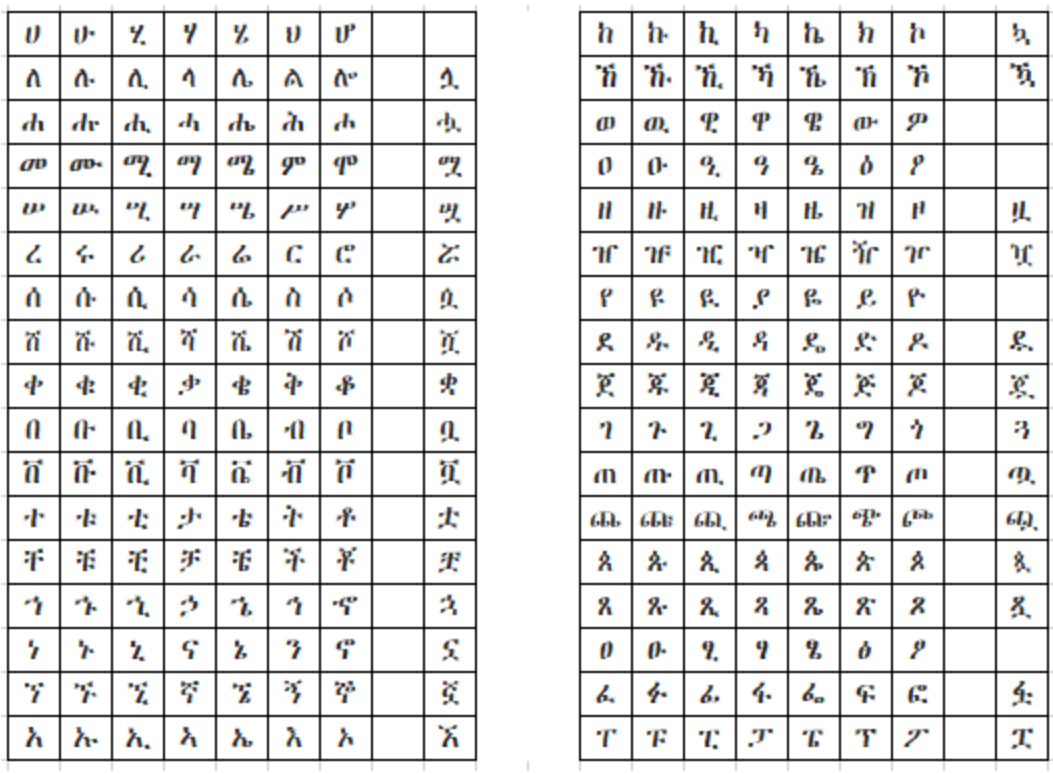

Figure 2.1: Amharic Alphabet

Amharic Punctuation Marks

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 end of a sentence.
- ፣ is used to separate comparative and sequential list of names, phrases, or numbers as well as to separate parts of a sentence that are not complete by themselves.
- ፤ 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.

2.2.1 Word Categories of Amharic

Baye Yimam [12] classified Amharic word classes into five types i.e., nouns, verbs, adjectives, adverbs and prepositions. This section discusses each Amharic word classes.

2.2.1.1 Nouns

Nouns are words that are used to identify names, things and places. A word is grouped under noun if it inflects for the Amharic plural marker '-ኦች' /'-och'/ or '-ዎች' [-woch], if it can be used as a subject or an object in a sentence, is modified by adjectives and comes after demonstrative pronouns [13].

Amharic plural nouns are mainly formed by adding suffixes: '-ኦች' /'*-och'*/ or '-ዎች' /'*-woch'*/. Table 2.2 shows suffixes used to form plural nouns in Amharic.

Table 2.2: Amharic plural noun formation using suffix

Singular Noun	**Plural marker**	**Plural Noun**
በሬ */bäre/* [ox]	-ዎች	በሬዎች / *bärewoch*/ [oxen]
አስተማሪ */xästämari/* [teacher]	-ዎች	አስተማሪዎች */xästämariwochï* / [teachers]
ላም */lam/* [cattle]	-ኦች	ላሞች /*lamoch* / [cattle]
ቤት */bet/* [house]	-ኦች	ቤቶች /*betoch* / [houses]
ሥራ */Sra/* [work]	-ዎች	ሥራዎች /*Srawoch*/ [works]
ጅብ /*jb*/ [hyena]	-ኦች	ጅቦች /*jboch* / [hyenas]

Amharic nouns can be either primary or derived. They are derived if they are related in their root consonants and/or meaning to verbs, adjectives or other nouns. Otherwise, they are primary [13, 14]. For example, the noun መንገድ */mängäd*/ [street] is primary but, 'መንገድ-ኛ' → መንገደኛ

/mängädäNa/ [traveler] is derived from the nominal base መንገድ by adding the morpheme '-ኛኛ'. Nouns can be derived from other nouns, adjectives, roots, stems and the infinitive form of a verb by affixation and intercalation. The morphemes '-ነት', '-ኛኛ', '-ኝት', '-ኣዊ', '-ተኛ', '-ኛ' and the prefix 'ባለ-' are used to derive nouns from other nouns. Table 2.3 shows examples of nouns derived from other base nouns.

Table 2.3: Nouns derived from other nouns

Base noun	**Derived noun**
ሰው */säw/* [Person]	ሰው-ነት → ሰውነት */säwnät/* [Body]
ዘብ */zäb/* [Sentinel]	ዘብ-ኛኛ → ዘበኛ */zäbäNa/* [Security guard]
ዋና */wana/* [Swimming]	ዋና-ተኛ → ዋናተኛ */wanatäNa/* [Swimmer]
ሃብት */habt/* [Wealth]	ባለ-ሃብት → ባለሃብት */balähabt/* [Wealthy]

A word that can be used in place of a noun is called a pronoun. Pronouns can be categorized based on their functions and meanings in the sentence. Amharic pronouns are categorized into personal pronouns, reflexive pronouns, demonstrative pronouns and possessive pronouns [12].

Personal Pronouns

A personal pronoun is a word that is used as a simple substitute for the proper name of a person. Amharic personal pronouns with equivalent English personal pronouns are shown in Table 2.3.

Table 2.4: Amharic personal pronouns

	1st Person	**2nd person**	**3rd person**
Singular	እኔ [I]	አንተ/አንቺ [you]	እሱ [he], እሷ [she]
Plural	እኛ [we]	እናንተ [you, plural]	እነሱ [they]

Within second-person and third-person singular, there are two additional polite independent pronouns, for reference to people to whom the speaker wishes to show respect. The polite personal pronouns in Amharic are እርስዎ [you, singular, polite] and እሳቸው [he/she, singular, polite].

Reflexive Pronouns

Reflexive pronouns are words that are used when the subject and the object of a sentence are the same. For example: እኔ በራሴ እተማመናለሁ */xne bärase etämamänalähu/* [I believe in myself]. The subject እኔ (I) and the object ራሴ (myself) indicate the same person.

A reflexive pronoun can also play the indirect object role in a sentence [12, 15]. For example: አልማዝ ሁልጊዜ ጠዋት ጠዋት ለራሷ ሻይ ትቀዳለች. */xälmaz hulgize Täwat Täwat läraswa śay tqädaläc/* [Almaz pours a cup of tea for herself every morning].

Amharic reflexive pronouns with equivalent English reflexive pronouns are as follows: እኔ ራሴ */xne rase/* [myself], እሱ ራሱ */xsu rasu/* [himself], እሷ ራሷ */xswa raswa/* [herself], አንተ ራስህ /x*äntä rash*/ [yourself, masculine, singular], አንቺ ራስሽ */xänchi räsś/* [yourself, feminine, singular], አንድ ራሱ /x*änd rasu* / (oneself), እሱ ራሱ */xsu rasu/* [itself], እኛ ራሳችን */xNa rasachn/* [ourselves], እናንተ ራሳችሁ */xnantä rasacu/* [yourselves], እነሱ ራሳቸው */xnesu rasacäw/* [themselves].

Demonstrative Pronouns

A demonstrative pronoun is a pronoun that is used to point to something specific within a sentence. Amharic makes a two way distinction between near ይህ/ይቺ */yh/yci/* [this], እነዚህ */xnäzih/* [these] and far ያ */ya/* [that], ያቺ */yacï/* [that], እነዚያ */xnäziya/* [those] demonstrative expressions (pronouns, adjectives, adverbs) and they can be either singular or plural. Amharic also distinguishes masculine gender ይህ */yh/* [this], ያ */ya/* [that]/ and feminine gender ይቺ */ycï* [this], ያቺ */yacï/* [that] in the singular.

Possessive Pronouns

Possessive pronouns show possession or ownership in a sentence [12, 15]. In Amharic there are two ways in which possession can be expressed. The first is through possessive suffixes. Amharic has a set of morphemes that are suffixed to nouns, signaling possession. For example: ቤት (house)

ቤት-ኤ → ቤቴ */bete/* [my house).

ቤት-ኣችን → ቤታችን */betäcn/* [our house].

ቤት-ህ → ቤትህ */beth/* [your house, masculine].

ቤት-ሽ → ቤትሽ */betś/* [your house, feminine].

ቤት-ኡ → ቤቱ */betu/* [his house].

ቤት-ኋ → ቤቷ */betwa/* [her house].

ቤት-ኣቹ → ቤታቹ */betacu/* [your house].

ቤት-ኣቸው → ቤታቸው */betacew/* [their house].

Morphemes -ኤ, -ኣችን, -ህ, -ሽ, -ኡ, -ኋ, -ኣቹ and -ኣቸው are affixed to the noun ቤት to indicate possession my, our, your (masculine, singular), your (feminine, singular), his, her, your (plural) and their respectively. The second way to express possession is through attaching prefix 'የ-' to the Amharic personal pronouns. For example:

ያ መኪና የአንተ (ያንተ) ነው */ya mäkina yantä näw/* [That car is yours]. The possessive pronoun የአንተ (ያንተ) is used as an object.

የአንተ (ያንተ) መኪና እየመጣች ነው */yantä mäkina eyämäTac näw/* [Your car is coming]. The possessive pronoun የአንተ (ያንተ) is used as a subject.

Amharic possessive pronouns with their equivalent English possessive pronouns is shown in Table 2.5.

Table 2.5: Amharic possessive personal pronouns.

	1st Person	**2nd person**	**3rd person**
Singular	የእኔ [my/mine].	የአንተ/የአንቺ [your/yours].	የእሱ [his/his, masculine], የእሷ [her/hers, feminine]
Plural	የእኛ [our/ours].	የእናንተ [your/yours].	የእነሱ [their/theirs].

Interrogative sentences are sentences that can form a question. According to Getahun Amare [15], the main interrogative pronouns used in Amharic are: ማን /*man*/ (who), ምን /*mn*/ [what], የት /*yät*/ [where], ስንት /*snt*/ [how much/ how many], መቼ /*mäce*/ [when], እንዴት /*xndet*/ [how], የትኛው /*yätNaw*/ [which]. When the interrogative pronouns are combined with preposition, we can get interrogative prepositional phrases ከማን /*kämän*/ [from who], ለምን /*lamn*/ [why], በምን /*bamn*/ [by what], ከየት /*käyät*/ [from where], የማነው /*yämanäw*/ [whose], etc.

2.2.1.2 Adjectives

Amharic adjectives modify nouns or pronouns by describing, identifying or quantifying words [12, 15]. Amharic adjectives always come before nouns or pronouns which they modify, but all the words that come before nouns cannot always be adjectives [13]. As it is true for nouns, adjectives can also be primary (such as ደግ */däg/* [kind], ፈጣን */fäTan/* [fast]) or derived. Adjectives are derived from nouns, stems or verbal roots by adding a suffix or a prefix and by intercalation. For example, it is possible to derive ድንጋይ-ኣማ → ድንጋያማ */dngayama/* [stony] from the noun ድንጋይ */dngay/* [stone]; ሀይል-ኛኛ → ሀይለኛ */hayläNa/* [powerful] from the noun ሀይል */hayl/* [power]; ስኝንኝፍ → ሰነፍ */sänäf/* [lazy] from the root ስንፍ */snf/*; ክብኡር → ክቡር */kbur/* [respectful] from the root ክብር */kbr/* [respect] by suffixation and intercalation.

2.2.1.3 Verbs

A verb is a word that expresses action, state of being in or relationship between two things [16]. Amharic verbs take subject markers as a suffix like '**-ሁ'** for subject 'I' as in መጣሁ */mäTahu/* [I came], '**-ህ'** for subject 'you' as in መጣህ */mäTah/* (you came), **'-ች'** for subject 'she' as in መጣች */mäTac/* [She came], and so on, to agree with subject of the sentence. Amharic verbs often have additional morphology that indicate the person, number and (second person and third person singular) gender of the object of the verb. For example: አንቺን አየ**ሁሽ** */xäncin xäyähuś/* [I saw you], '-ሁሽ' indicates second person, singular, feminine, and in the sentence አልማዝን አየ**ኋት** */xälmazn xäyäwat/* [I saw Almaz] '-ኋት' indicates third person, singular, feminine.

2.2.1.4 Adverbs

In Amharic, adverbs are used to modify the coming verbs. Adverbs always come before the modified verb. Adverbs can be found either in their primitive form or compound form as grouping of preposition and other word categories [13].

For example: in the adverbial phrase, መምጣት አለመምጣቷን ገና አልወሰነችም */mämTat xälämämTatwan gäna xälwäsänäcm/* [She hasn't yet decided if she wants to come or not], ገና */gäna/* [yet] is the only adverb that formed the adverbial phrase.

2.2.1.5 Prepositions

Prepositions and postpositions together are called adpositions. A preposition or a postposition typically combines with a noun or a pronoun or more generally a noun phrase, this being called its complement. A preposition comes before its complement; a postposition comes after its complement. In Amharic, adposition link one word with another word [12]. Amharic adpositions are very few in number, these are: ስለ */slä/,* እንደ */xnde/,* ከወደ */käwädä/,* አጠገብ */xäTägäb/,* ማዶ */mado/,* ባሻገር */bashagär/,* ወዲህ */wädih/*. In Amharic, adpositions give meaning when they come with other words. Consider the following phrases:

ስለ ገንዘብ */slä gänzäb/*

እንደ ሰው */xndä säw/*

ከወንዝ ማዶ */käwänz mado/*

እስከ ጎጃም ድረስ */xskä gojam dräs/*

Prepositions ‘ስለ’, ‘እንደ’, ‘ከ’ and ‘እስከ’ comes before the nouns ‘ገንዘብ’, ‘ሰው’, ‘ወንዝ’ and ‘ጎጃም’ and postpositions ‘ማዶ’ and ‘ድረስ’ comes after the nouns ‘ወንዝ’ and ‘ጎጃም’.

2.2.2 Amharic Phrasal Categories

A phrase is a small group of words that adds meaning to a sentence. In a phrase, the main word, or the word that is what the phrase is about, is called the head. According to Eleni Teshome [6], Amharic phrases are categorized into noun phrases, verb phrases, adjectival phrases, adverbial phrases and prepositional phrases, and researcher Baye Yimam [12] adds one more category called conjunction phrases. The descriptions of each Amharic phrases is presented in the following section.

Noun Phrase: A noun phrase is a phrase that has a noun as its head. In Amharic noun phrase, one or more words work together to give more information about the noun. For example: in the following Amharic noun phrase ሁሉም የምወዳቸው ልጆቼ */hulum yämwädacäw ljoce/* [all my dear children], ሁሉም (all) is a specifier, የምወዳቸው (my dear) is an adverbial modifier and ልጆቼ (children) is a noun.

Verb Phrase: Amharic verb phrase is constructed with a verb as a head and other constituents such as complements, modifiers and specifiers. For example: in the following Amharic verb

phrase, ከትምህርት ቤት መጣሁ */kätmhrt bet mäTahu/* [I came from school], ከትምህርት ቤት (from school) is prepositional phrase modifying the verb መጣሁ (came).

Adjectival phrase: In Amharic adjective phrase, one or more words work together to give more information about an adjective. For example: in the sentence, ወንድሜ በስራው በጣም ደስተኛ ነው */wändme bäsraw bätam dästäNa näw/* [My brother is very happy with his work], ደስተኛ */dästäNa/* [happy] modifies the prepositional phrase በስራው */bäsraw/* [with his work].

Prepositional Phrase: Amharic prepositional phrase is made up of a preposition head and other constituents such as nouns, noun phrases, etc. unlike other phrase constructions, a preposition cannot be taken as a phrase, instead it should be combined with other constituents. Prepositions link nouns, pronouns and phrases to other words in a sentence. Prepositions give meanings only if they combine with other words such as noun, adjective, verb. For example: in the prepositional phrase በወንበር ላይ */bäwänbär lay/* (on the chair) በ */bä/* and ላይ /*lay*/ are prepositions which are combined with the noun ወንበር /*wänbär/* [chair].

Adverbial Phrases: Amharic adverbial phrases are made up of an adverb as head word and one or more other lexical categories including adverbs themselves as modifiers, the head of the adverbial phrase is placed at the end [13]. Unlike other phrases, adverbial phrases do not take complements. Most of the time, the modifiers of the adverbial phrases are prepositional phrases that come always before adverbs. Examples: ክፉኛ */kfuNa/* [severely], በጣም ክፉኛ */bäTam kfuNa*/ [very severely], እንደ ወንድሙ በጣም ክፉኛ /x*ndä wändmu bäTam kfuNa/* [very severely like his brother].

Conjunction Phrases: A conjunction is a part of speech that connects words with words, phrases with phrases and sentences with sentences [12]. The primary types of conjunctions in Amharic are coordinating conjunctions and subordinating conjunctions.

Coordinating conjunctions connect words, phrases, and clauses. A coordinating conjunctions give equal emphasis or importance to clauses, phrases, and words. For example, consider the following Amharic sentences.

አራት በጎች **እና** ሶስት ፍየሎች */xärat bägoc xna sost fyäloc/* [Four sheep and three goats].

ሁለት እንጀራ **ወይም** ሶስት ጠርሙስ ቢራ /*hulät xnjära wäym sost Tärmus bira*/ [Two Injera or three bottles of beer]

'እና' /*xna*/ [and] and 'ወይም' /*wäym*/ [or] are coordinating conjunctions used to connect related phrases.

Subordinating conjunctions connect two clauses together, but in doing so, they make one clause dependent (or subordinate) upon the other clause (or main clause). For example, consider the following simple sentences.

ጠራሁት **ሆኖም** አልመጣም */Tärahut honom xälmätam/* [I called him though he did not come].

በልቻለሁ **ግን** አልጠገብኩም /*bälcaläw gn xältägäbkum*/ [I ate but I didn't satisfied].

ጠራሁት (I called him) and በልቻለሁ /*bälcaläw/* (I ate) are the main clauses, አልመጣም /*xälmätam/* (he did not come) and አልጠገብኩም */xältägäbkum/* (I didn't satisfied) are the dependent clauses and ሆኖም */honom/* [though] and ግን */gn/* [but] are subordinate conjunctions.

2.2.3 Amharic Morphology

Amharic is a consonant root-based language with vowels added on to the consonants. Morphemes can be added as articles, prepositions, personal pronouns, numbers, conjunctions and adjectives [17, 18]. The roots of verbs and most nouns in the Amharic are characterized as a sequence of consonants known as radicals.

For example:

ውስድ is the root form for ወሰደ */wäsädä/* [take] and ተወሰደ /*täwäsädä*/ [taken].

Subject-Verb agreement

Amharic verbs agree with their subjects that is, the person, number and gender of the subject of the verb (in the second and third person singular) are marked by suffixes or prefixes on the verb. The affixes on the verb that signal subject agreement vary greatly with the particular verb tense, aspect or mood. In Amharic sentence, the verb goes at the end of the sentence and the order is Subject – Object – Verb (SOV) [18].

For example:

እሱ ተማሪ ነው */xsu tämari näw/* [He is a student].

እሱ /*xsu*/ [He] is a subject, ‘ተማሪ’ /*tämari*/ [student] is an object and ’ነው’ /*näw*/ [is] is a verb.

Amharic Articles

Indefinite articles are generally unmarked in Amharic, but definite articles are always marked by a suffix called the definite marker [19]. For singular, a distinction is made between a noun treated as masculine form, for example: ቤቱ /bet-u/ [his house] or as feminine form, ቤቷ /*bet-wa*/ [her house], definite-female.

2.2.4 Amharic Sentence Structure

The usual word order of a sentence in Amharic is Subject-Object-Verb (SOV) [10]. For example, in the sentence:

አበበ ትምህርት ቤት ሄደ /*xäbäbä tmhrt bet hedä*/ [Abebe went to school], ‘አበበ’ /*xäbäbä*/ [Abebe] is the subject, ‘ትምህርት ቤት’ /*tmhrt bet*/ [school] is the object and ‘ሄደ’ /*hedä*/ [went] is the verb.

Simple Amharic sentences can also be constructed using a subject and a predicate.

For example: ‘ውሻው ሮጠ’ /*wśaw roTä*/ [the dog ran], ‘ውሻው’ /*wśaw*/ [the dog] is the subject of the sentence, because the sentence is telling something about the dog. And what is it telling? It says ውሻው ሮጠ /*wśaw roTä*/ [the dog ran], so the predicate is ሮጠ /*roTä*/ [ran].

Amharic sentences can also be constructed from simple or complex noun phrases and simple or complex verb phrases. Simple sentences are constructed from simple noun phrase followed by simple verb phrase which contains only a single verb. The following examples show the various structures of simple sentences.

- አበበ ሄደ /*xäbäbä hedä*/ [Abebe went.]
- አበበ መኪና ገዛ /*abäbä mäkina gäza*/ [Abebe bought a car.]
- ማን መኪና ገዛልህ? /*man mäkina gäzalh?*/ [Who did buy a car for you?]
- ሁለት ትልልቅ ልጆች በመኪና ወደ ጎጃም ሄዱ /*hulät tllq ljocï bämäkina wädä gojam hedu*/ [Two big children went to Gojjam by car.]

2.3 A Brief Overview of Afaan Oromo

Afaan Oromo is one of the major indigenous African languages that is widely spoken and used in most part of Ethiopia and some parts of the neighboring countries [20]. Besides, Afaan Oromo has long history of and well developed oral tradition. Despite of this and the size of its speakers as well as its value as widely spoken language in the Horn of Africa, it remained as unwritten language for a long period of time. The writing system of Afaan Oromo is called Qubee, a latin alphabet [21, 22].

Afaan Oromo has five vowels, five double consonants and twenty consonant phonemes, i.e., sounds that make a difference in word meaning. Afaan Oromo vowels are represented by the letters, a, e, o, u and i, or long vowels: aa, ee, oo, uu and ii. The length of the vowel makes a difference in word meaning [22]. For example:

Laga [river] and Laagaa [roof of the mouth].

Lafa [ground] and Laafaa [soft]

Afaan Oromo double consonants are represented by the letters: Ch, Dh, Ny, Ph and Sh and the rest consonants are represented by the letters: B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y and Z [23].

Afaan Oromo words do not have the consonants ‘p’, ‘v’ and ‘z’, because there are no native Afaan Oromo words that are formed from these characters [7, 23]. However, in writing Afaan Oromo they are used to refer to foreign words such as police (poolisii) and virus (vaayirasii).

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.
- **?** Question mark is used to indicate a direct question when placed at the end of a sentence. For example: Na wajjin dhufitta? (Can you come with me?)
- **!** Exclamation mark is used at the end of command and exclamatory sentences.
- **,** Comma is used to show a separation of ideas or elements within the structure of a sentence. For example: Koonsarticharratti Dooktar Artiist Alii Birraa, Heelan Mallas, Kaahsaay Barihaa fi Salamoon Haylee ni hirmaatu.

- **:** Colon is used to separate and introduce lists, clauses, and quotations, along with several conventional uses.
- **;** Semi colon is used to connect independent clauses. It shows a closer relationship between the clauses than a period would show.

An apostrophe mark (‘) in Afaan Oromo is used to represent a glitch called hudhaa sound. It is used to write the word in which most of the time two vowels appeared together like ba’e to mean “get out” with the exception of some words like ja’a ‘six’, hin danda’amu ‘impossible’, which are identified from the sound created. Sometimes apostrophe mark (‘) in Afaan Oromo interchangeable with the spelling “h”. For instance, “ba’e”, “ja’a” can be interchanged by the spelling “h” like “bahe”, “jaha” respectively still the senses of the words is not changed.

2.3.1 Word Categories of Afaan Oromo

Words are the basic unit of a language that has meaning and can be spoken or written. Afaan Oromo words are composed of two parts: the root (base morpheme), which generally consists of basic sound and provides the basic lexical meaning of the word, and the pattern, which consists of prefixes and/or suffixes and gives grammatical meaning to the word [20]. For example, the root ‘bar’ combines with the pattern ‘-e’ gives bare (learned), whereas the same root combines with the pattern ‘-te’ gives barte (she learned). The combination of words on the basis of the language resulted in phrases, clauses or sentences. Afaan Oromo words can be placed into five grammatical categories: nouns, verbs, adverbs, adjectives and adposition [24]. According to Abdi Sani [21] pronouns are included under the noun category, and conjunctions and interjections under adposition.

2.3.1.1 Noun

A noun is a part of speech that names a person, place, thing, idea, action or quality. In Afaan Oromo, a noun (maqaa) mainly occurs at the beginning of a sentence. For example: Tolaan hucuu adii bitate. (Tola bought white cloth). Tolaan (Tola) ‘name of person’ is a noun, comes at the beginning of the sentence.

A word that is categorized as a noun in a sentence can be a subject or an object [25]. In Afaan Oromo, a subject mostly comes at the beginning whereas an object mostly comes after subject and before verbs in a sentence. For example: in the sentence, Tolaan mana ijare. (Tola built a house). The noun ‘Tolaan’ (name of person) is the subject and the noun ‘mana’ (house) is the object.

Most Afaan Oromo nouns are marked for gender: masculine or feminine. Afaan Oromo nouns derived from verbs adds suffix '–aa' and '–tuu' to the verb root for the masculine and feminine gender respectively [26].

For example:

barsiisuu [to teach]: verb,

barsiisaa [teacher] masculine: noun;

barsiistuu [teacher] feminine: noun.

barachuu (to learn): verb,

barataa [student], masculine: noun;

barattuu [student], feminine: noun.

Afaan Oromo plural nouns are mainly formed by adding suffixes: '–oota', '–ota', '–wwan', '–een', '–lee' and '–yyi' [24, 27]. Table 2.6 shows suffixes used to form plural nouns in Afaan Oromo.

Table 2.6: Afaan Oromo plural noun formation using suffix

Singular Noun	**Plural marker**	**Plural Noun**
Sangaa [ox]	-ota	Sangota [oxen]
Barsiisaa [teacher]	-ota	Barsiisota [teachers]
Sa'a [cattle]	-wwan	Saawwan [cattle]
Mana [house]	-een	Manneen [houses]
Hojii [work]	-lee	Hojiilee [works]
Waraabeessa [hyena]	-yyi	Waraabeeyyi [hyenas]

Pronoun

A pronoun is a word that can be used in place of a noun. Afaan Oromo pronouns can be categorized based on their functions and meanings in the sentence [25]. These are personal pronouns, possessive pronouns, reflexive pronouns and demonstrative pronouns. The descriptions of each Afann Oromo pronouns is presented in this section.

Personal Pronouns

Afaan Oromo personal pronouns refer to the person speaking, the person spoken to or the person spoken about. For example, in the following sentences,

Isheen kitaaba dubbifte. (She read a book).

Inni ishee jaalata. (He likes her).

Nuti isa binna. (We buy it).

Isheen (she), inni (he) and nuti (we) are personal pronouns. Table 2.7 illustrates Afaan Oromo personal pronouns that can be used in the subject positions.

Table 2.7: Afaan Oromo personal pronouns

Number	**1st Person**	**2nd person**	**3rd person**
Singular	Ani (I)	Ati (you)	Inni/isa (he), Ishee/isii (she)
Plural	Nuti (we)	Isin (you)	Isaan/Jarii (they)

Possessive Pronouns: Possessive pronouns are pronouns that indicate ownership of something. For example:

Re'een suni tiyya. (That goat is mine).

Konkolaataan sun keessani. (That car is yours).

'tiyaa' (mine) and keessani (yours) are possessive pronoun. Table 2.8 below shows Afaan Oromo possessive pronouns that can be used in the subject positions.

Table 2.8: Afaan Oromo possessive personal pronouns

Number	**1st Person**	**2nd person**	**3rd person**
Singular	Kiyya/Kooti/tiyya (mine)	kee (yours)	Kan isaa (his), Kan ishee (hers)
Plural	Keenya (ours)	Keessan (yours)	Kan isaanii (theirs)

Afaan Oromo possessive case can also be formed by prefixing 'kan'. For example: kan koo (mine), kan keenya (ours), kan isaa (his), kan ishee (hers), kan kee (yours), kan isaanii (theirs).

Reflexive Pronouns

According to Getachew Mamo and Million Meshesha [25], Afaan Oromo has two ways of expressing reflexive pronouns (myself, ourselves, yourself, yourselves, himself, herself and themselves). One is to use the noun meaning 'self': of(i) or if(i). This noun is inflected for case but, unless it is being emphasized, not for person, number, or gender.

For example:

Isheen of laalti. (base form of of). (She looks at herself).

Isheen ofiif konkolaataa bitte. (dative of of). (She bought a car for herself).

The other possibility is to use 'mataa', with possessive suffixes. For example: mataa koo (myself), mataa kee (yourself , singular).

Afaan Oromo has a reciprocal pronoun wal (each other) that is used like of/if. It is inflected for case but not for person, number, or gender.

For example:

Wal jaalatu. (They like each other).

Kennaa walii bitan. (They bought gifts each other).

Demonstrative Pronouns

Afaan Oromo makes a two-way distinction between proximal ('this, these') and distal ('that, those') demonstrative pronouns and adjectives [25, 22]. Proximal pronouns have masculine and feminine gender whereas distal pronouns do not have. However, singular and plural demonstrative pronouns are not distinguished. Table 2.9 shows Afaan Oromo demonstrative pronouns.

Table 2.9: Afaan Oromo demonstrative pronouns

Case	**Proximal ('this, these')**	**Distal ('that, those')**
Base	Kana (Tana, feminine)	San
Nominative	Kuni (Tuni, feminine)	Suni

In Afaan Oromo interrogative sentences are used to form a question. According to Jabesa Daba and Yaregal Assabie [7], the main Afaan Oromo interrogative pronouns are: maal(i) (ምን, what), maaliif(i) (why), akkam(i) (how), yoom (መቼ, when), eessa (የት, where), eessaa (from where),

eenyu (ማን, who, what), kan eenyu (whose), meeqa (ስንት, ምን ያህል, how much, how many), kam(i) (which).

2.3.1.2 Verb

A verb (xumura) is a word that express action, state of being in or relationship between two things [16]. In Afaan Oromo verbs mostly appear at the end of a sentence [22]. For example: Turaan wayaa adii bitate. (Tura bought white cloth). Bitate (bought) is the verb of the sentence.

Like Amharic, Afaan Oromo verbs can be modified to indicate person, gender, tense and number [20, 25, 22]. The prefixes and suffixes for person, gender, tense and number are essentially identical in all forms. For example, root 'deem-' has the basic meaning of 'waking'. The root may be conjugated in simple past, present, continuous and perfect tense, in singular and plural forms as shown in Table 2.10.

Table 2.10: Different forms of root 'deem'[20]

Person	**Number**	**Past**	**Present**	**Continuous**	**Perfect**
1st person	Singular	Deem**e**	**Ni**deem**a**	Deem**aara**	Deem**eera**
	Plural	Deem**ne**	**Ni**deem**na**	Deem**aarra**	Deem**neerra**
2nd person	Singular	Deem**te**	**Ni**deem**ta**	Deem**aarta**	Deem**teerta**
	Plural	Deem**tan**	**Ni**deem**tu**	Deem**aartu**	Deem**taniirtu**
3rd person	Singular	Deem**e**	**Ni**deem**a/ti**	Deem**aara**/**arti**	Deem**eera**/**teerti**
	Plural	Deem**an**	**Ni**deem**u**	Deem**aaruu**	Deem**aniiru**

Most Afaan Oromo verbs are in their infinitive form, for example, beekuu (to know). The verb stem 'beek-' is the infinitive form 'beekuu' with the final '–uu' dropped. Afaan Oromo verbs can be categorized into main (transitive or intransitive) and auxiliary verbs [22].

Transitive verbs are main verbs which transfer message to complements or objects. Consider the following examples:

Tolaan bishaan waraabe. [Tola fetch water].

Tolaan ulee cabse. [Tola broke a stick].

Each of the verbs, waraabe [fetch] and cabse [broke] in these sentences have objects that complete the verbs' actions.

Intransitive verbs are main verbs which do not take object or complement in a sentence. For example: in the sentence, Ijoolleen rafan (Children slept), it is impossible for an object to follow the verb rafan (slept).

Auxiliary verbs support the main verbs used in a sentence, add functional or grammatical meaning to the clauses in which they appear. For example:

Tolaan kaleessa ganama fiigaa ture. [Tola was running yesterday morning.]

Yeroo obboleessi koo naaf bilbilu, ani rafeen ture. [I was sleeping when my brother called me.]

Taphni ijoolleef faayidaa baay'ee qaba. [playing has many advantages for childrens.]

In the above sentences the words 'ture' and 'qaba' are auxiliary verbs. The following are Afaan Oromo auxiliary verbs 'dhaa', 'ta`e', 'qaba', 'ture', 'jira', etc.

Like Amharic, Afaan Oromo verbs take subject markers such as '-e', '-ine', '-ite' and '-ani' for subjects I, we, she and they respectively to agree with the subject of the sentences, as shown in the following examples:

Ani isa gors**e.** [I advised him.]

Nu`i isa gors**ine.** [We advised him.]

Isheen isa gors**ite.** [She advised him.]

Isaan isa gors**ani.** [They advised him.]

2.3.1.3 Adverb

Adverbs are words which modify verbs and adjectives. Adverbs could be categorized as adverbial time, adverbial place and adverbial condition [25]. In Afaan Oromo adverbs precede verbs they modify. For example:

- Isheen baayee furdaada. [She is very fat.], baayee [very] indicates the degree how fat she is.
- Isheen amma dufte. [She came now.], 'amma'/ [now] is a time adverb.
- Toolaan baayee deeraada. [Tola is very tall.], the adverb baayee modifies the verb deeraa.

Some common Afaan Oromo adverbs are: amma (now), kaleessa (yesterday), harr'a (today), edana (tonight), bor (tomorrow), dhiyootti (soon), dafee (quickly), suuta (slowly), walii wajjin (together), baayee (very), yeroo hunda (always), yeroo baayyee (usually), gaaffii gaaf (sometimes), darbee (rarely), matuma (never).

2.3.1.4 Adjective

In Afaan Oromo adjectives (addeessa) come after the nouns they qualify. For example: in the following adjectival phrases, uffata adii (white cloth) and muka gabaabaa (short stick), adii (white) and gabaabaa (short) are adjectives that qualifies the nouns uffata and muka respectively.

Afaan Oromo adjectives can be marked for gender, by the presence of gender markers '-cca', '-aa', etc for masculine and '-ttii', '-tuu', '-oo', etc for feminine [22]. Table 2.10 presents inflection of adjectives for gender.

Table 2.11: Adjectives inflection for gender

Adjective	**Masculine**	**Feminine**
gurraacca (black)	gurraacca (by affixing –cca)	gurraattii (by affixing –ttii)
deeraa (tall)	deeraa (by affixing –aa)	deertuu (by affixing –tuu)
furdaa (fat)	furdaa (by affixing -aa)	furdoo (by affixing -oo)

2.3.1.5 Adposition

Prepositions and postpositions together are called adpositions. Adpositions are class of words used to express spatial or temporal relations [24]. A preposition comes before its complement; a postposition comes after its complement. Consider the following examples.

- Toolaan waaye ofisaa dubbaccuu jaalata (Tola likes to talk about himself). 'waaye' (about) occurs preceding the nominal 'ofisaa'.
- Toolaan abbaasaa wajjin dhufe. (Tola came with his father). 'wajjin' (with) occurs after the nominal 'abbaasaa'.

Some common prepositions are: gara (towards), eega, erga (since, from, after), haga, hanga (until), hamma (upto, as much as), akka (like as), waa'ee (about, in regard to).

Some common postpositions are: ala (out, outside), bira (beside, with, around), booda (after), cinaa (beside, near, next to), dur, dura (before), duuba (behind, back of), irra (on), irraa (from), itti (to,

at, in), jala (under, beneath), jidduu (middle, between), keessa (in, inside), malee (without, except), wajjin (with, together), gubbaa (on, above), fuuldura (in front of), gad(i) (down, below), ol(i) (up, above).

Afaan Oromo Conjunction

A conjunction is a word that can be used to connect two phrases, clauses and sentences. Conjunctions can be divided into coordinating and subordinating conjunctions. Coordinating conjunctions are used to connect two independent clauses [28], whereas, subordinating conjunctions are used to connect main clauses with subordinate clauses [25]. Consider the following examples:

- Ittoo shiroon jaaladha garuu ittoo misira caalaa jaaladha (I like shiro watt, but I like lentil watt more). 'Garuu' is used to connect the two independent sentences "Ittoo shiroon jaaladha" and "ittoo misira caalaa jaaladha".
- Nyaatan barbaada sababiinsa nan beela'e. (I want food because I am hungry). 'Sababiinsa' is used as a subordinating conjunction. It connects the independent clause "Nyaatan barbaada (I want food)" and the subordinating clause "nan beela'e (I am hungry)".

Some common Afaan Oromo conjunctions are: fi (and), garuu/immoo (but), yookin-for declaratives, moo-for questions (or), haa ta'u malee (however), etc.

Afaan Oromo Subordinating conjunctions are yoo (if), akka waan (as if), sababiin isaa, sababiinsa (because), kanaafuu (so, therefore), akka (so that, in order to), ta'us (though), tu'ullee (even though), wanta/yeenna (when), hamma (until), erga (after), dursa (before), etc.

2.3.2 Afaan Oromo Phrasal Categories

In Afaan Oromo there are five different kinds of phrases, namely noun phrase, verb phrase, prepositional phrase, adjectival phrase and adverbial phrase. This section discusses Afaan Oromo phrasal categories.

Noun Phrases

A noun phrase is a phrase that has a noun or indefinite pronoun as its head. For example: in the sentence, Manni Toolaan sun jige. [That Tola's house has damaged], "Manni Toolaan" is a noun phrase, and the head (noun) of the noun phrase is "Manni".

Verb Phrases

In a verb phrase the word that the phrase about is the verb. For example: in the sentence, Caaltun **biddeena xaafii tolchite**. [Chaltu made teff injera], 'tolchite' is the head of the verb phrase "biddeena xaafii tolchite". The verb phrase tells what Chaltu 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. [When the rain stops raining we went to the city], "gara magaalaa" is a prepositional phrase and the head of the prepositional phrase is 'gara' [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 cimtuudha.** [Chaltu is very cleaver in her education.], the phrase "barnoota ishiitiin daran cimtuudha" is adjectival phrase.

Adverbial Phrases

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

- Mucaan *suutaan* deema [The boy went slowly]; *suutaan* indicates the manner of an action.
- Abbaan isaa *darbannii darbanii* mana dhufu. [His father come to home *seldom.*]; *darbannii darbanii* indicates the time of an action.
- Dabtara keessan *bakka kana* ka'aadha deemaa. [Put your exercise book *here* and go.]; *bakka kana* indicates location.
- Obbo Caalaan lafa *ballinaa* qotan [Mr. Chala is farming a *large* land]; *ballinaa* indicates degree.
- Inni hojii *suutaan* hojjechu filata. [He prefers to do his work *quickly*.], *suutaan* indicates the manner of an action.

2.3.3 Afaan Oromo Morphology

Morphology is the study of morphemes and their arrangements in forming words [29]. Morphemes are the minimal meaningful units which may constitute words or parts of words. In Afaan Oromo, words can be formed from morphemes in two ways: inflectional morphology and derivational

morphology [30]. In inflectional morphology, words are formed by the combination of stem with a grammatical morpheme, usually results in a word of the same class as the original stem. Inflectional morphemes modify a word's tense, number, aspect and so on. For example: barsiisaa (teacher): singular, barsiis**ota** (teachers): plural. Derivational morphology deals with word formation from stem and grammatical morphemes, usually results in a word of different lexical class [30, 31]. For example: hojjete [work]: verb is derived from hojjet**oota** [workers]: noun.

Subject-Verb agreement

Like Amharic, Afaan Oromo verbs agree with their subjects. The person, number and gender of the subject of the verb are marked by suffixes or prefixes on the verb [32].

For example:

- Isheen Ameerikaa irraa dhufte. (She came from America).
- Inni Ameerikaa irraa dhufe. (He came from America).
- Isaan Ameerikaa irraa dhufan. (They came from America).

The verbs dhufte, dhufe and dhufan agree with the subject pronouns Isheen (she), Inni (he) and Isaan (they) respectively.

Definiteness

Afaan Oromo has no indefinite articles but it indicates definiteness with suffixes '-(t)icha' for masculine nouns and '-(t)ittii' for feminine nouns and the last vowel of the noun is dropped before suffixes (-icha, -ittii, -attii, -utti) are added [23, 26].

For example:

karaa 'road', karaa + icha (karicha) (the road),

nama 'man', nama + (t)icha (namicha /namticha/) (the man).

For animated nouns that take either male or female gender, the definite suffix may indicate the intended gender. For example: qaalluu (priest), qaalicha (the priest, masculine), qallittii (the priest, feminine).

2.3.4 Afaan Oromo Sentence Structure

Afaan Oromo and Amhraic are the same in sentence structure order. Like Amhraic, the sentence structure of Afaan Oromo is Subject-Object-Verb (SOV) [33]. For example, in the sentences,

- Inni bishaan fide (He brought water), Inni [he] is the subject, bishaan (water) is the object and fide (brought) is the verb.
- Isheen hoolaa bitte (She bought sheep), Isheen (she) is the subject, hoolaa (sheep) is the object and bitte (bought) is the verb.

2.4 Machine Translation

Machine translation is a subfield of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another [34]. Machine translation systems, based on their core methodology can be classified into two paradigms: the rule-based approach and the corpus-based approach. In the rule-based 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. On the other hand, under the corpus-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Combining the features of the two major classifications of machine translation systems gave birth to the hybrid machine translation approach [35]. Each of the machine translation approach is explained below.

2.4.1 Rule Based Machine Translation Approach

Rule Based Machine Translation (RBMT), also known as knowledge based machine translation uses linguistic rules and language order for its conversion to target language [36]. Having input sentences in some source language, a rule based system generates output sentences in some target language, based on the morphological, syntactic and semantic analysis of both the source and the target languages.

The various approaches that RBMT systems follow are direct, transfer and interlingua. They differ in the depth of analysis of the source language and the level to form language-independent representation of meaning between source language and target language [37]. The Vauquois triangle in Figure 2.2 shows an increasing depth of analysis required on both the analysis and generation end as we move from the direct approach through transfer approach to interlingua approach.

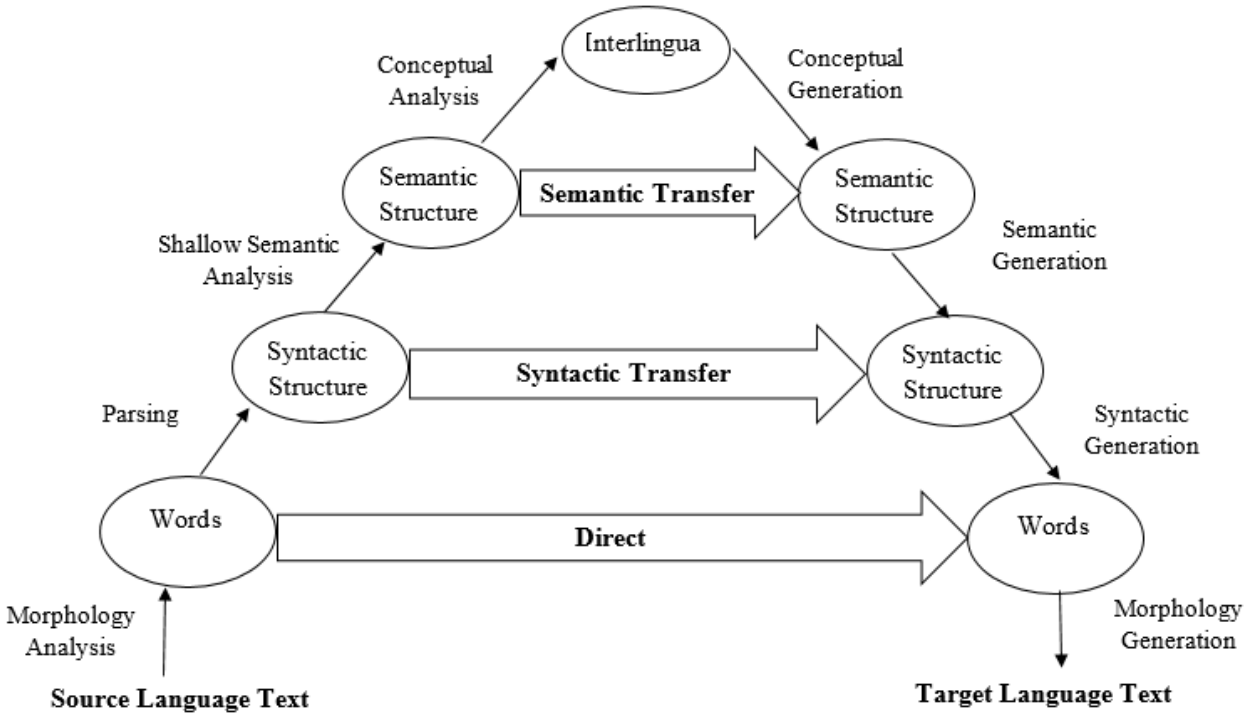

Figure 2.2: Vauquois Triangle [38]

The most intuitive form of translation is simply translating every word, one by one, looking up the word in a bilingual lexicon. This is also the basis of the so called direct translation approach, found at the bottom of the Vauquois tringle [38]. One level above the direct approach is the transfer approach. In syntactic transfer, the syntax structure of the source sentence is analyzed, and the resulting syntactic structure is mapped, by rules, to a new syntactic structure in the target language. Semantic transfer, is similar to syntactic transfer, but attempts to analyse the semantic structure of the source sentence, and uses rules to map these to a semantic structure in the target language. The interlingua model is found at the top of the Vauquois triangle. Direct and transfer approaches rely extensively on various sets of rules that map words, syntax, or semantic roles from the source language to the target language. This is a limitation, when there are multiple languages to relate to each other, because it requires to reconstruct the rule sets for each language pair. The interlingua approach is a solution to the limitation of direct and transfer approach. The basic idea behind interlingua approach is, instead of translating from all languages to all others, translation goes from the source languages to one interlingua representation and from that representation to the target languages. Each of RBMT approaches are discussed below.

2.4.1.1 Direct Machine Translation Approach

Direct machine translation (DMT) approach is the oldest and less popular approach. Machine translation systems that use this approach are capable of translating a source language directly to a target language. Words of the source language are translated into target language with the same word-for-word arrangement with the help of bilingual dictionary, without passing through an intermediary representation [37]. Source language analysis is oriented specifically to only one target language. Direct machine translation systems are basically unidirectional and bilingual.

As depicted in Figure 2.3, DMT approach requires the following stages for the generation of a sentence in the target language.

- The morphological inflections are removed from the words of the source text according to the different grammar rules of the word.
- The target language equivalent is found in a bilingual dictionary.
- Necessary syntactical arrangements are performed.
- Lastly, the output is generated in the target language.

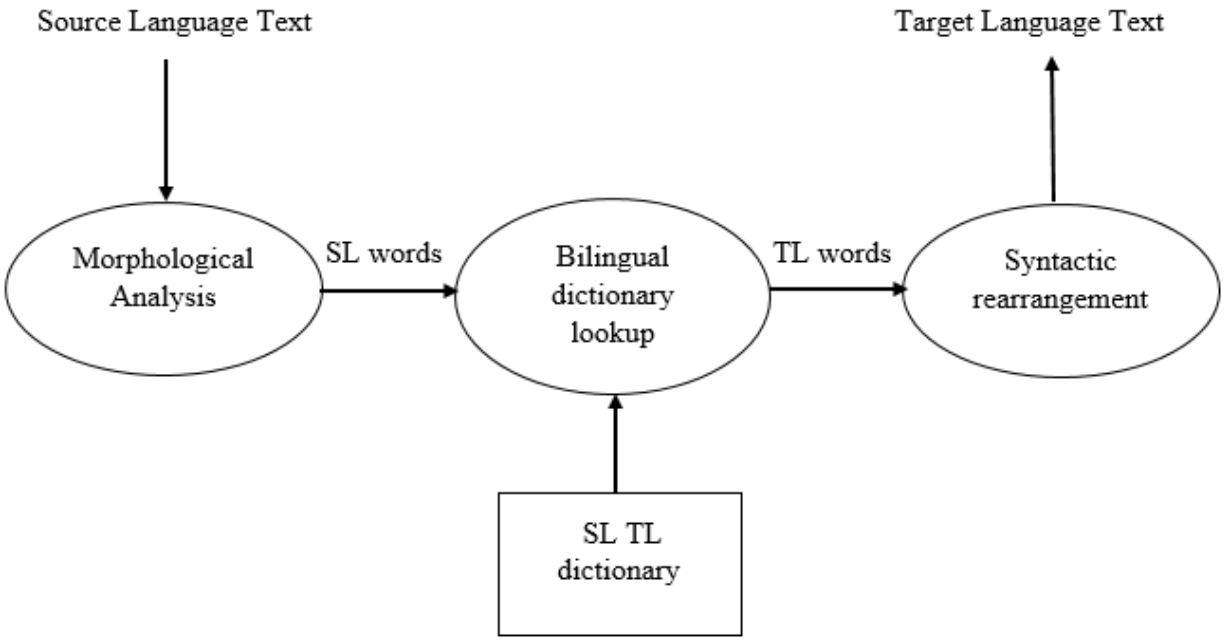

Figure 2.3: Direct machine translation approach [39]

DMT approach requires little syntactic and semantic analysis and its performance depends on morphological analysis, text processing software and word-by-word translation with minor

adjustments on word order and morphology. DMT involves only lexical analysis, i.e., it does not consider structure and relationships between words and also it is developed for a specific language pair are among the limitations [35].

2.4.1.2 Transfer Based Machine Translation Approach

In transfer based machine translation, the source language is transformed into an abstract, less language-specific representation. An equivalent representation, with the same level of abstraction, is generated for the target language using bilingual dictionaries and grammar rules. On the basis of the structural differences between the source and target language, a transfer based system can be broken down into three different stages: analysis, transfer and generation [35]. In the first stage, analysis of the source text is done based on linguistic information such as morphology, part-of-speech, syntax and semantics. Algorithms are applied to parse the source language and derive the syntactic or the semantic structure of the text to be translated. During transfer stage, the syntactic or semantic structure of the source language is then transferred into the syntactic or semantic structure of the target language. In generation stage, the necessary morphological inflections for the sentences are added. Accuracy of output can be enhanced if the translation is limited to a particular domain. The quality of translation can be further increased by pre-processing the input sentence.

2.4.1.3 Interlingua Machine Translation Approach

In interlingua machine translation approach, source language is transformed into an interlingua representation which is independent of any of the languages involved in the translation. The interlingua representation is then translated to the target language in order to produce meaningful translation, as shown in Figure 2.4.

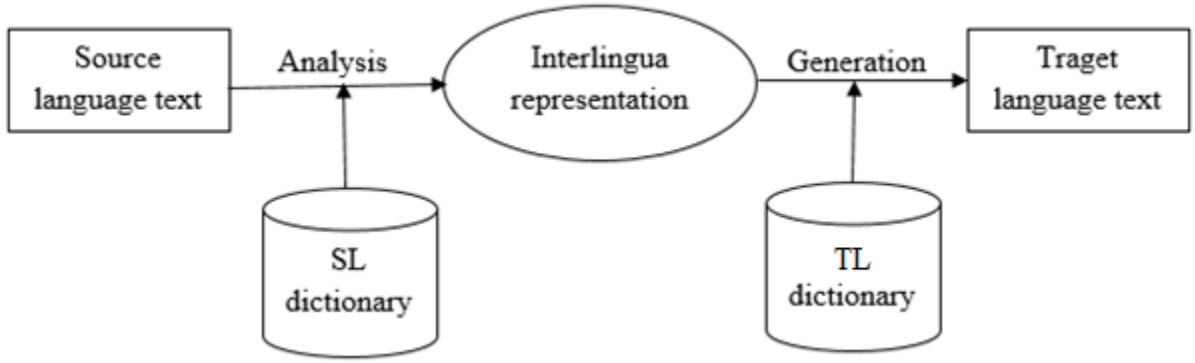

Figure 2.4: Interlingua-based RBMT [40]

Because of its independency on the language pair for translation, this approach is useful for multilingual machine translation system.

2.4.2 Corpus Based Machine Translation Approach

Corpus-based machine translation (CBMT), also referred as data driven machine translation, is an alternative approach for machine translation to overcome knowledge acquisition problem of rule-based machine translation (RBMT). CBMT automatically acquires the translation models from bilingual corpora that may not be there for under resourced languages [41]. CBMT approach is further classified into two major approaches: Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT) approach.

2.4.2.1 Statistical Machine Translation Approach

SMT is a data driven approach which uses parallel aligned corpora and treat translation as a mathematical reasoning problem, in that every sentence in the target language is a translation with probability from the source language [40]. The higher the probability, then the higher the accuracy of translation and vice versa. SMT consists of language model, translation model and decoder. The basic sketch of SMT system is shown in Figure 2.5.

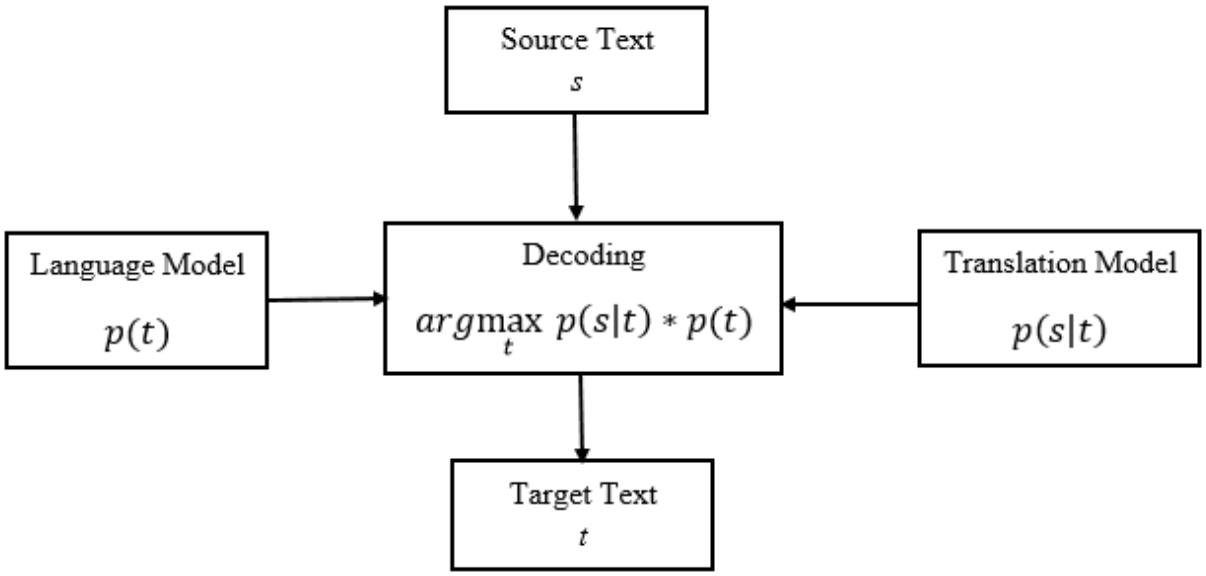

Figure 2.5: SMT Architecture [42]

The language model calculates the probability of the target language $p(t)$ and it models the fluency of the proposed target sentence.

Basically, an N-gram model predicts the occurrence of a word based on the occurrence of its N–1 previous words. For example, a bigram model (when N = 2) predicts the occurrence of a word, given only its previous word. Similarly, a trigram model (when N = 3) predicts the occurrence of a word based on its previous two words.

The Maximum Likelihood Estimate (MLE) of the unigram probability of a word w_i in a corpus is its count $c(w_{i)}$ normalized by the total number of word tokens N, as given by equation (1):

$$p(w_i) = \frac{C(w_i)}{N} \quad (1)$$

To compute a particular bigram probability of a word w_n, given a previous word w_{n-1}, we will compute the count of the bigram C($w_{n-1}w_n$) normalized by the sum of all the bigrams that share the same first word w_{n-1}:

$$p(w_n|w_{n-1}) = \frac{C(w_{n-1}w_{n)}}{\sum_w C(w_{n-1}w_n)} \quad (2)$$

We can simplify equation (2) into equation (3), since the sum of all bigram counts that start with a given word w_{n-1} must be the unigram count for that word w_{n-1}

$$p(w_n|w_{n-1}) = \frac{C(w_{n-1}w_{n)}}{C(w_{n-1})} \quad (3)$$

To compute some of the n-gram probabilities, consider the following mini-corpus of five Amharic sentences:

ሃና መፅሐፍ ገዛች

ሃና መፅሐፍ አነበበች

ሃና ሻይ አፈላች

አልማዝ ቡና ገዛች

አልማዝ ቡና ጠጣች

Some of the unigram probabilities from the corpus are:

$$p(\text{ሃና}) = \frac{C(\text{ሃና})}{N} = \frac{3}{15} = 0.20$$

$$p(\text{አልማዝ}) = \frac{C(\text{አልማዝ})}{N} = \frac{2}{15} = 0.13$$

where N is the total number of words seen in the corpus.

Some of the bigram probabilities from the corpus are:

$$p(\text{መፅሐፍ}|\text{ሃና}) = \frac{C(\text{ሃና መፅሐፍ})}{C(\text{ሃና})} = \frac{2}{3} = 0.67$$

$$p(\text{ሻይ}|\text{ሃና}) = \frac{C(\text{ሃና ሻይ})}{C(\text{ሃና})} = \frac{1}{3} = 0.33$$

Some of the trigram probabilities from the corpus are:

$$p(\text{አፈላች}|\text{ሃና ሻይ}) = \frac{C(\text{ሃና ሻይ አፈላች})}{C(\text{ሃና ሻይ})} = \frac{1}{1} = 1.0$$

$$p(\text{ጠጣች}|\text{አልማዝ ቡና}) = \frac{C(\text{አልማዝ ቡና ጠጣች})}{C(\text{አልማዝ ቡና})} = \frac{1}{2} = 0.50$$

The N-gram model performs well for unigram, bigram and trigram models for the corpus of simple sentences. Long sentences are difficult to observe in corpora and if any N-gram is missing, the language model will assign a probability of zero [43]. To keep a language model from assigning zero probability, smoothing techniques are used. Laplace smoothing adds one to all the counts, before we normalize them into probabilities. Since there are V words in a vocabulary and each one was incremented, we also need to adjust the denominator to take into account the extra V observations. Laplace smoothing to unigram probabilities is given by equation (4):

$$P_{Laplace}(w_i) = \frac{C(w_i)+1}{N+V} \quad (4)$$

where V is the size of the vocabulary, distinct words.

The formula of Laplace for smoothing of bigrams is given by equation (5)

$$P_{Laplace}(w_i|w_{i-1}) = \frac{C(w_{i-1},\ w_i)+1}{C(w_i-1)+V} \quad (5)$$

Using the above mini-corpus, the Laplace for smoothing of unigram:

$$P_{Laplace}(\text{ሃና}) = \frac{C(\text{ሃና})+1}{N+V} = \frac{3+1}{15+9} = \frac{4}{24} = \frac{1}{6}$$

Laplace for smoothing of bigram

$$P_{Laplace}(\text{ሙፅሐፍ}|\text{ሃና}) = \frac{C(\text{ሃና}, \text{ሙፅሐፍ}) + 1}{C(\text{ሃና}) + V} = \frac{2+1}{3+9} = \frac{3}{12} = \frac{1}{4} = 0.25$$

Translation model $p(s|t)$ is the probability that a sentence s in the source language S is the translation of a sentence t in the target language T. Performing the search efficiently is the work of a machine translation decoder that uses the foreign string, heuristics and other methods to limit the search space and at the same time keeping acceptable quality. The decoder gives the best translation possible $\hat{t}$ by maximizing the two probabilities $p(s|t)$ and $p(t)$ as given by equation (6) and make use of search algorithm.

$$\hat{t} = arg\max_{t}\ p(t|s) \qquad (6)$$

The Bayer Rule is applied to equation (6) to derive:

$$\hat{t} = arg\max_{t}\ p(s|t) * p(t) \qquad (7)$$

SMT system is not tailored to any specific pair of languages and it requires less virtual space than other models of machine translation, which makes it easier to operate and train on smaller systems. SMT does not work well for language pairs that have significantly different word orders and corpus development can be costly. SMT approach is subdivided into the following approaches namely: word based SMT, phrase based SMT, syntax based translation and hierarchical based SMT.

In word based SMT, sentences are broken down to the fundamental unit (word) and translation for source language to target language is done word by word. Once the target words are generated then they are arranged in a specific order by use of a reordering algorithm to generate the target sentence. However, compound words like idioms bring complexities [44].

Phrase based SMT proposed by Koehn [45] and mainly uses phrases as the fundamental unit of translation. The source and target language sentences contained in the parallel corpora are divided into phrases. Phrase based translation models are acquired from a word-aligned parallel corpus by extracting all phrase-pairs that are consistent with the word alignment based on Koehn [45] principle. The input and output phrases are aligned according to a specific order as suggested by Antony [44]. Though phrase based SMT may result in better performance, long phrases may degrade the performance.

Syntax based translation is based on the idea of translating syntactic units, rather than single words or strings of words (as in Phrase based SMT), i.e., parse trees of sentences/utterances.

Hierarchical phrase based SMT was proposed by Chiang [40] and combines the strengths of phrase-based and syntax-based translations. Phrase-based consists of the unit of block or segment of translation while the syntax based translation brings the rules of translation.

2.4.2.2 Example Based Machine Translation Approach

Example Based Machine Translation (EBMT) approach is introduced by Nagao [46] and can be defined as a data-driven approach that make use of analogy translation, similar in meaning and form from examples database. The database is made of parallel aligned bilingual corpora i.e., a set of sentences in the source language and the corresponding translations of each sentence in the target language with point to point mapping. The corpora is used to translate similar types of sentences of source language to the target language. The analogy translation uses three stages: matching, adaption and recombination.

In matching stage, the source language input text is fragmented depending on the granularity of the system followed by search for set of examples from database which matches or closely matches the input source language fragment string and the relevant fragments are picked. The target language fragments corresponding to the relevant fragments are extracted. In adaption stage, if the match is exact, the fragments are recombined to form target language output, else find the target language portion of the relevant match corresponds to specific portion in source language and align them. Finally, in recombination stage, combination of relevant target language fragments in order to form legal grammatical target text. The EBMT shown in Figure 2.6, shares similarities in structure with that shown in the Vauquois Triangle.

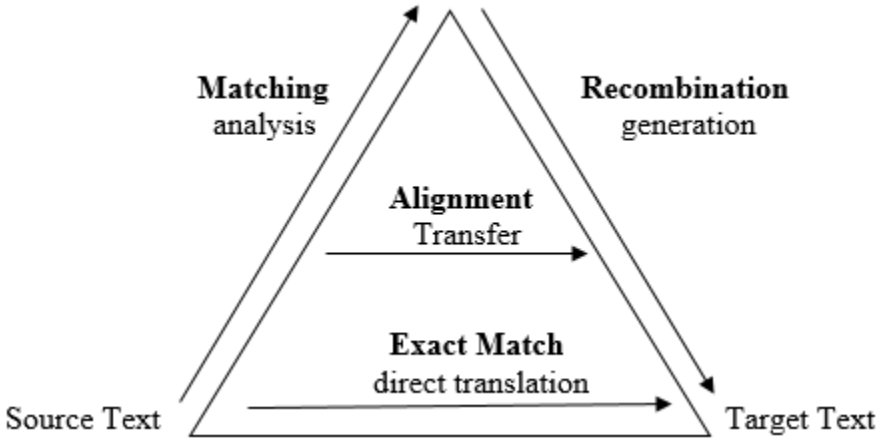

Figure 2.6: The Vauquois Triangle Modified for EBMT [47]

EBMT approach avoids the need for manually derived rules. However, it requires analysis and generation modules to produce the dependency trees needed for the examples database and for analyzing the sentence. EBMT is also computational efficiency, especially for large databases, although parallel computation techniques can be applied.

2.4.3 Hybrid Machine Translation Approach

Rule-based machine translation approach has high accuracy though it takes a lot of resource in terms of time and cost of development. On the other hand, data-driven machine translation approach has high coverage and cost of development is low as compared to rule-based machine translation approach. However, for data-driven machine translation approach, the need of corpora is a demerit especially for under resourced languages. By taking the advantage of both rule-based machine translation and statistical machine translation approaches, a new approach called hybrid machine translation was developed which has a better efficiency in the area of machine translation systems [7, 35]. The hybrid machine translation approach can be used in a number of different ways [35]. In some cases, translations are performed in the first stage using a rule based machine translation approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system.

2.4.4 Neural Machine Translation

Neural Machine Translation (NMT) is a recently proposed and an effective deep learning approach to machine translation that uses a large neural network based on vector representation of words

and that has shown an encouraging results [48]. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance [49]. The models proposed recently for neural machine translation often belong to a family of encoder-decoders. Recurrent Neural Network (RNN) encoder-decoder is proposed by K. Cho et al. [50]. The main idea is that the encoder encodes the source sentence with the RNN model and uses the last hidden state as input for the RNN decoder; this represents the output in the target sentence. The endoder and decoder are implemented using a RNN, especially Long-Short Term Memory (LSTM), convolutional neural networks, self-attention units or a combination of them [51]. In all these architectures, source and target sentences are handled separately as a one-dimensional sequence over time. One of the weaknesses of such models is that the encoder states are computed only once at the beginning and are left untouched with respect to the target histories.

If compared with SMT, there is no separate language model, translation model or reordering model, but just a single sequence model which predicts one word at a time. The prediction is conditioned on the source sentence and the already produced sequence in the target language. The prediction power of NMT is more promising than that of SMT, as neural networks share statistical evidence between similar words.

Although effective, the NMT systems still suffer some issues, such as scaling to larger vocabularies of words and the slow speed of training the models. In addition, large corpus is needed to train neural machine translation systems with performance comparable to statistical machine translation.

2.5 Evaluation of Machine Translation

Machine translation is the task to translate a text from a source language to a target language. As machine translation emerges as an important mode of translation, its quality is becoming more and more important. Evaluating machine translation results lacks an appropriate, consistent and easy to use criterion [52].

Machine translation evaluation could be done by using manual or automatic evaluation methods. Human evaluations gives a better result in order to measure the quality of machine translation. The most challenging issues in conducting human evaluation of machine translation output are high

costs and time consumption. Therefore, automatic metrics have been used in the evaluation of machine translated text. Some of automatic evaluation metrics are:

Word Error Rate

One of an automatic metrics used to evaluate machine translation systems is Word Error Rate (WER). WER is computed as the Levenshtein distance between the words of the system output and the words of the reference translation divided by the length of the reference translation. WER is the percentage of words, which are to be inserted, deleted or replaced in the translation inorder to obtain the sentence of reference [53].

The Levenshtein distance is computed using dynamic programming to find the optimal alignment between the machine translation output and the reference translation, with each word in the machine translation output aligning to either 1 or 0 words in the reference translation, and vice versa. Those cases where a reference word is aligned to nothing are labled as *deletions*, whereas the alignment of a word from the machine translation output to nothing is an *insertion.* If a reference word matches the machine translation output word it is aligned to, this is marked as a match, and otherwise is a *substitution.* The WER is then the sum of the number of substitutions (S), insertions (I) and deletions (D) divided by the number of words in the reference translation (N) as shown in Equation (8).

$$WER = \frac{\mathrm{S} + \mathrm{I} + \mathrm{D}}{\mathrm{N}} \qquad (8)$$

WER metric is computed efficiently and is reproducible. However, the main drawback is its dependency on the sentences of reference.

Sentence Error Rate

Sentence Error Rate (SER) indicates the percentage of sentences, whose translations have not matched in an exact mannar those of references. It shows similar advantages and shortcomings as WER. The Word Error Rate (WER) is based on the Levenshtein distance, the minimum number of substitutions, deletions and insertions that have to be performed to convert the generated text to the reference text [54]. The limitation of the WER is that it does not allow reordering of words, whereas the word order of the hypothesis can be different from the word order of the reference even though it is correct translation. In order to overcome this problem, the Position-Independent Word Error Rate (PER) compares the words in the two sentences without taking the word order

into account. The PER is always lower than or equal to the WER. The shortcoming of the PER is that the word order can be important in some cases. Therefore the best solution is to calculate both word error rates.

BLEU Score

BLEU is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another [55]. BLEU measures how many word sequences in the sentence under evaluation match the word sequences of some reference sentence. BLEU could be gamed by producing very short system outputs consisting only of highly confident n-grams, if it were not for the use of brevity penalty which penalized the BLEU score if the system output is shorter than the references.

$$p_n = \frac{\sum_{c\in\{Can\}} \sum_{n-gram\in c} Cnt_{clip}(n-gram)}{\sum_{c^r\in\{Can\}} \sum_{n-gram^r\in c^r} Cnt_{clip}(n-gram^r)} \quad (9)$$

Equation (9) shows the computation of the BLEU precision score for n-grams of length n, where *Can* are the sentences in the test-corpus, $Cnt(n-gram)$ is the number of times an n-gram occurs in a candidate, and $Cnt_{clip}(n-gram)$ is the minimum of the unclipped count and the maximum number of times it occurs in a reference translation.

$$BP = \begin{cases} 1, & c > r \\ e\left(1-\frac{r}{c}\right), & c \le r \end{cases} \quad (10)$$

Equation (10) shows the calculation of the BLEU brevity penalty, where *c* is the length of the candidate translation and r is the length of the reference translation. These terms are combined, as shown in Equation (11) to calculate the total BLEU score, wheren N is typically 4, and W_n is usually set to $\frac{1}{N}$.

$$BLEU = BP.\exp(\sum_{n=1}^{N} w_n \log p_n) \quad (11)$$

BLEU's score is always a number between 0 and 1. This value indicates how similar the candidate text is to the reference text, with values closer to 1 representing more similar texts. BLEU score alson includes a penalty for translations whose length differs significantly from that of the reference translation.

NIST

NIST metric is a method for evaluating the quality of text which has been translated using machine translation. Its name comes from the US National Institute of Standards and Technology. NIST is based on BLEU metric but introduced some modifications. BLEU calculates n-gram precision adding equal weight to each one but NIST gives information weight for each word, i.e. higher scores to more rate n-gram which are considered as more informative n-grams. NIST differs also from BLEU in brevity penality calculation, where small differences in translation length do not impact the overall score.

METEOR

Metric for Evaluation of Translation with Explicity Ordering (METEOR) is an automatic evaluation metric for machine translation output [56]. METEOR modifies BLEU in the way that it gives more emphasis to recall than to precision. METEOR was designed to fix some of the problems found in the more popular BLEU metric, and also produce good correlation with human judgement at the sentence or segment level. This differs from the BLEU metric in that BLEU seeks correlation at the corpus level.

Unlike BLEU which only calculates precision, METEOR calculates both precision and recall, and combine the two as shown in equation (12).

$$F_{mean} = \frac{P.R}{\alpha P + (1-\alpha)R} \qquad (12)$$

METEOR uses several stages of word matching between the system output and the reference translations in order to align the two strings. The matching stages are as follows:

a) Exact matching: strings which are identical in the reference and the hypothesis are aligned.
b) Stem matching: stemming is performed, so that words with the same morphological root are aligned.
c) Synonymy matching: words which are synonyms according to wordnet are aligned.

In each of these stages only words that were not matched in previous stages are allowed to be matched. Only unigrams, single words are compared for matches. Precision in METEOR is defined as the number of matches divided by the number of words in the system output and recall is defined as the number of matches divided by the number of words in the reference.

CHAPTER THREE: RELATED WORK

3.1 Overview

This chapter reviews the literature on machine translation done on different language pairs. This review covers machine translation system done for non-Ethiopian language pairs, for English and Ethiopian languages pairs and for Ethiopian language pairs. Finally a brief summary of this chapter is given.

3.2 Machine Translation Systems for Non-Ethiopian Language Pairs

The research by Joel Ilao *et al.* [4] was performed with the objective of translating Filipino-to-English Bidirectional Statistical Machine Translation (FEBSMT) using feedback. For this research, 22,061 instances of ASEANMT tourism parallel corpus were used for the initial training, development, and testing processes with the help of the Moses Toolkit. The output translations were evaluated with the use of evaluation metrics such as BLEU, NIST, METEOR, and TER. To further improve on the translation quality, user feedback was collected for statistical post-editing. The post-editing module is based on the Post-Edit Propagation (PEPr) system's concept of an Automatic Post-Editing (APE) system. For English-to-Filipino, FEBSMT showed a BLEU score of 0.34 after the five iterations of the automatic post-editing system. On the other hand, for Filipino-to-English, FEBSMT showed a 0.40 BLEU score with the same number of iterations.

The research by [5] deals with Myanmar – English bidirectional machine translation system with numerical particles identification. The system is implemented by applying rule based machine translation approach. Stanford and ML2KR parsers are used for preprocessing step. In English to Myanmar machine translation, the sentence is determined as the correct sentence when not only all words in a sentence were translated but also the translated sentence was meaningful sentence. Although all words in a sentence were translated, if the translated sentence was not acceptable, it is incorrect sentence. The performance of the system is tested by using the example sentences. Testing set of example sentences consist of over 1200 sentences. In addition, the system used Myanmar-English bilingual lexicon which contains 13,373 words to translate as Myanmar words. The evaluation measures for Myanmar-English machine translation system are defined in terms of success rate. Evaluation result of Myanmar to English machine translation with 1030 correct sentences and 224 incorrect sentences is 82.14%. Evaluation result of English to Myanmar machine translation with 971 correct sentences and 236 incorrect sentences is 80.45%.

3.3 Machine Translation Systems for English and Ethiopian language pairs

The research which was conducted by Jabesa Daba [7] mainly deals with English-Afaan Oromo machine translation system using a hybrid of rule-based and statistical approaches. Since English and Afaan Oromo have different sentence structures, the author implemented syntactic reordering with the purpose of making the structure of source sentences similar to the structure of target sentences. Accordingly, reordering rules are developed for simple, interrogative and complex English and Afaan Oromo sentences. Two groups of experiments are conducted by using purely statistical approach and hybrid approach. The Afaan Oromo-English SMT yields a BLEU score of 41.50% whereas English-Afaan Oromo SMT has a BLEU score of 32.39%. After applying local reordering rules, the system is improved to provide a BLEU score of 52.02% and 37.41% for Afaan Oromo-English and English-Afaan Oromo translations, respectively. The limitation of the study is that, the rules developed are used only for syntax reordering; morphological rules are not included.

The study which was conducted by Sisay Adugna [8] mainly deals with the translation of English documents to Afaan Oromo using statistical methods. The study was carried out with two main goals: the first one is to apply existing SMT system on English – Afaan Oromo language pair by using available parallel corpus and the second one is to identify the challenges that need a solution regarding the language pair. The author used parallel documents from different domains including spiritual, medical and legal documents. 20,000 bilingual sentences and 62,300 monolingual sentences were used for training and testing purpose. The BLEU score for the test data from legal, medical and religious domains are 13.69%, 1.97% and 21.72% respectively. Due to the spelling error of same Afaan Oromo words in the corpus, the system consider them as different. The limitation of the study is that it does not incorporate Afaan Oromo spell checker.

The study which was conducted by Eleni Teshome [6] mainly deals with translation of English documents to Amharic and Amharic documents to English. The research work implemented the statistical machine translation approach. Two language models were developed, one for Amharic and the other for English so as to ensure a bidirectional translation. Translation models were built which assign a probability that a given source language text generates a target language text. Two different corpora were prepared. Corpus I was made of about 1020 simple sentences that had been prepared manually. All sentences were used for the training set. For the test set, the sample text

that contains 102 simple sentences was prepared manually. Corpus II contains 1951 complex sentences out of which 40 sentences were used for the test set. Two methodologies were used to test the system. The first methodology is BLEU score and the second methodology used is preparing a questionnaire manually. The result on Corpus I recorded from the first methodology (BLEU Score) was 82.22% for the English to Amharic translation and 90.59% for the Amharic to English translation. The result recorded on Corpus I using the second methodology was 91% for the English-Amharic translation and 97% for the Amharic to English. The result on Corpus II recorded from the first methodology was 73.38% for English to Amharic translation and 84.12% for Amharic to English translation. The accuracy from the second methodology on Corpus II was 87% for English to Amharic translation and 89% for Amharic to English translation. The limitation of the study is that it does not handle larger set of complex sentences.

3.4 Machine Translation System for Ethiopian Language pair

The study which was conducted by the Akubazgi Gebremariam [9] mainly deals with the Amharic to Tigrigna machine translation using hybrid approach. Two major experiments are conducted using two different approaches and their results are recorded. 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 hybrid approach and the result obtained has a BLEU score of 17.47%. From the result, it can be concluded that the hybrid approach is better than the statistical approach for Amharic-to-Tigrigna machine translation system. The limitation of the study is that morphological rules are not developed, only rules for syntax reordering are developed.

3.5 Summary

In this section, we have discussed works related to machine translation for different language pairs. As to the researcher's knowledge there is no study conducted that deal with Amharic-Afaan Oromo machine translation. Since Amharic and Afaan Oromo are morphologically rich and less resourced languages and researches conducted on machine translation for different language pairs using different approaches cannot be directly applied for Amharic-Afaan Oromo or vice versa translation, this study deal to experiment bidirectional Amharic-Afaan Oromo machine translation using hybrid approach.

CHAPTER FOUR: BIDIRECTIONAL AMHARIC-AFAAN OROMO MACHINE TRANSLATION SYSTEM

4.1 Introduction

This chapter discusses bidirectional Amharic – Afaan Oromo machine translation system. The overall system architecture and its components are discussed in detail.

4.2 Architecture of the System

The architecture of the bidirectional Amharic – Afaan Oromo machine translation system shown in Figure 4.1 has four components which will be discussed next.

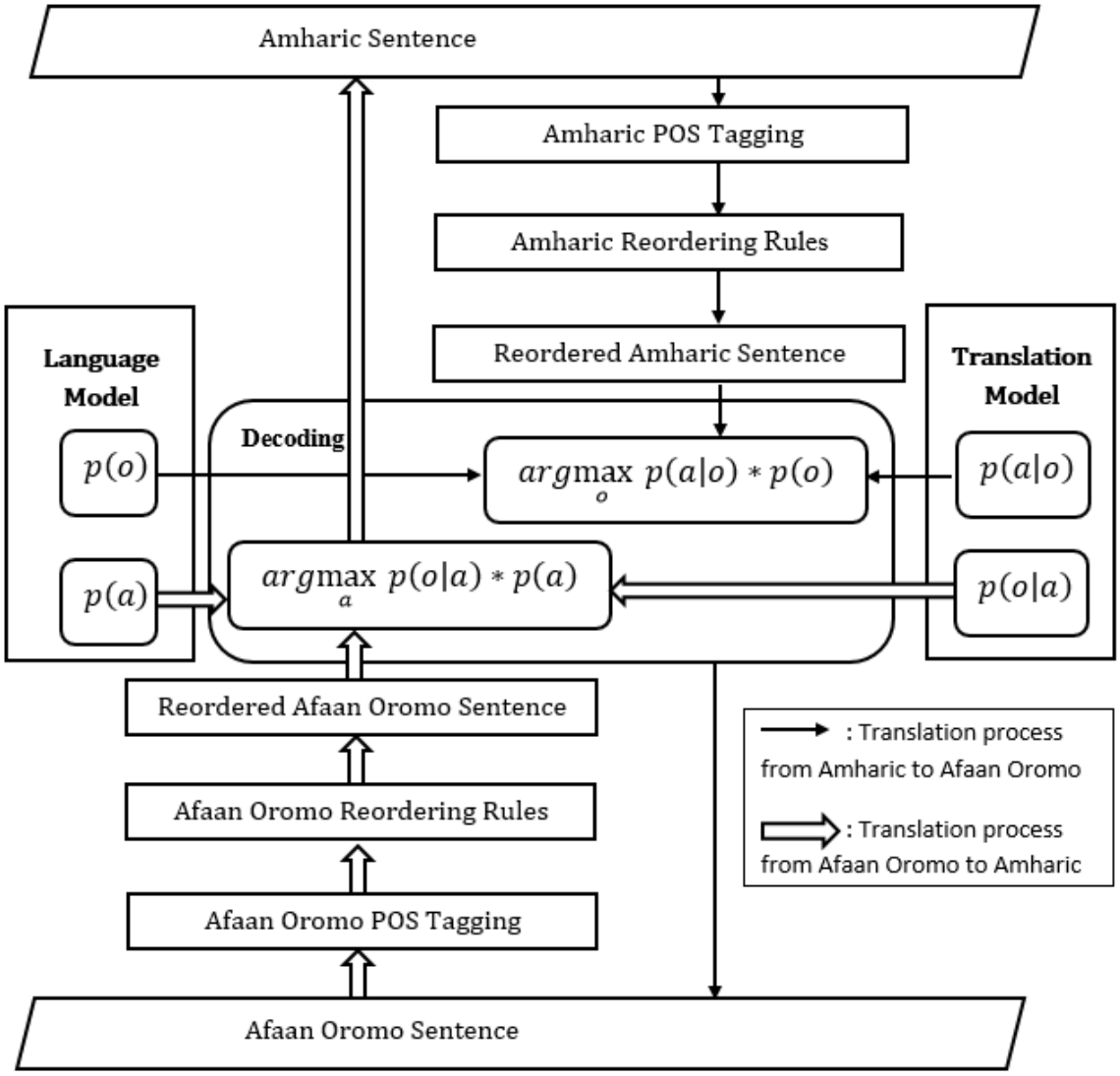

Figure 4.1: Architecture of the System

Amharic/Afaan Oromo sentences consist of lexical items called Part of Speech (POS). POS tagging is the process by which a specific tag is assigned to each word of an input sentence, to indicate the function of that word in the specific context. POS includes nouns, verbs, adjectives, adverbs, pronouns, conjunctions and their sub-categories.

Table 4.1 shows POS tag sets used in Amharic and Afaan Oromo in which most of them are adopted from the English Penn Treebank tag sets [7] and tag sets developed for Amharic-Tigrigna translation [9].

Table 4.1: Amharic and Afaan Oromo POS tag sets

No	**Tags**	**Description**
1	AXU	Auxiliary verb
2	CC	Conjunction and subordinate conjunction
3	CD	Cardinal number
4	CW	Any compound word (can be singular, plural and mass) which changes its order of words in the target language
5	ON	Ordinary number
6	IN	Preposition
7	JJ	Adjective
8	NN	Noun, singular or mass
9	NNP	Proper noun, singular
10	NNPS	Proper noun, plural
11	NNS	Noun, plural
12	NP	Noun phrase
13	PRP	Personal pronoun
14	PRP$	Possessive pronoun
15	PUN	Punctuation

16	RB	Adverb
17	SYM	Symbol
18	VB	Verb, base form
19	VBD	Verb, past tense
20	VBG	Verb, gerund or present participle
21	VBN	Verb, past participle
22	VBP	Verb, non-3rd person singular present
23	VBR	Relative verb
24	VBZ	Verb, 3rd person singular present
25	WP	Interrogative pronoun
26	WP$	Possessive wh-pronoun

Since there are no publicly available POS tagger tools for Amharic and Afaan Oromo, for this research, POS tagging is done manually. Input sentences in either languages (Amharic or Afaan Oromo) are POS tagged for sentence reordering. According to the sentence structure of Amharic or Afaan Oromo languages, the words and their tagged information are stored in a separate file.

4.2.1 Sentence Reordering

Word reordering is a preprocessing stage in machine translation system where the words of the source language sentence are reordered as per the grammatical structure of the target language to facilitate the training process. Amharic belongs to Semitic and Afaan Oromo belongs to Cushitic family of languages. The sentence structure of both languages is Subject-Object-Verb (SOV) that could be suitable for statistical machine translation.

There are some Amharic phrases and sentences that do not require reordering rules for translation to Afaan Oromo and vice versa.

Consider the following simple Amharic sentences and their translations in Afaan Oromo.

Amharic: ሃና መኪና ገዛች ። /*hana mäkina gäzac/* [Hana bought a car.]

Afaan Oromo: Haanaan konkoolaataa bitte.

Amharic POS tag: ሃና_NNP ሙኪና_NN ገዛች_VBD ።_PUN

Afaan Oromo POS tag: Haanaan_NNP konkoolaataa_NN bitte_VBD ._PUN

Amharic: ልጁ ወተት ጠጣ ። */lju wätät TäTa/* [The baby drank milk.]

Afaan Oromo: Mucaan aannan dhuge.

Amharic POS tag: ልጁ_NNP ወተት_NN ጠጣ_VBD ።_PUN

Afaan Oromo POS tag: Mucaan_NNP aannan_NN dhuge_VBD ._PUN

The structural order of words in both languages are similar, i.e., in Amharic and Afaan Oromo sentences, the subjects (‘ሃና’ , ‘Haanaan’, ‘ልጁ’, ‘Mucaan’) come before the objects (‘ሙኪና’, ‘konkoolaataa’, ‘ወተት’, ‘aannan’) and the verbs (‘ገዛች’, ‘bitte’, ‘ጠጣ’, ‘dhuge’) come after the objects. Therefore, for such kinds of simple sentences reordering rules are not required.

Consider the following Amharic sentences containing prepositions and their translations in Afaan Oromo.

Amharic: እሱ ወንበር ላይ ተቀመጠ ። */xsu wänbär lay täqämäTä/* [He sit on the chair).

Afaan Oromo: Inni barcuma irraa ta’e.

Amharic POS tagged: እሱ_PRP ወንበር_NN ላይ_IN ተቀመጠ_VBG ።_PUN

Afaan Oromo POS tagged: Inni_PRP barcuma_NN irraa_IN ta’e_VBG ._PUN

Amharic: እሷ ቤት ውስጥ ነች ። */xsWa bet wusT näc/* [She is in the house.]

Afaan Oromo: Isheen mana keessa jirti.

Amharic POS tagged: እሷ_PRP ቤት_NN ውስጥ_IN ነች_AUX ።_PUN

Afaan Oromo POS tagged: Isheen_PRP mana_NN keessa_IN jirti_AUX ._PUN

Amharic: እኛ ወደ ከተማ ሄድን ። */xNa wädä kätäma hedn/* [We went to the city).

Afaan Oromo: Nu’i gara magaalaa deemne.

Amharic POS tagged: እኛ_PRP ወደ_IN ከተማ_NN ሄድን_VBD ።_PUN

Afaan Oromo POS tagged: Nu’i_PRP gara_IN magaalaa_NN deemne_VBD ._PUN

When the prepositions ‘ላይ’ and ‘ውስጥ’ appear after the nouns ‘ወንበር’ and ‘ቤት’ in Amharic sentences then the equivalent prepositions ‘irraa’ and ‘keesaa’ in Afaan Oromo also appear after

the nouns ‘barcumaa’ and ‘mana’, and also when the preposition ‘ወደ’ appears before the noun ‘ከተማ’ in Amharic sentence then the equivalent preposition ‘gara’ in Afaan Oromo sentence appears before the noun ‘magaalaa’. Therefore, for such kinds of Amharic and Afaan Oromo sentences containing prepositions, reordering rules are not required.

Consider the following Amharic interrogative sentences and their translations in Afaan Oromo.

Amharic: ካሳ መቼ መጣህ? /*kasa mäce mäTah*?/ [Kassa when did you come?]

Afaan Oromo: Kaasaa yoom dhufte?

Amharic: ሃና ምን ፈለግሽ? /*hana mn fälägś* ?/ [Hana what do you want?]

Afaan Oromo: Haanaa maal barbaadee?

Amharic POS tagged: ካሳ_NNP መቼ_WP መጣህ_VBD ?

Afaan Oromo POS tagged: Kaasaa_NNP yoom_WP dhufte_VBD ?_PUN

Amharic POS tagged: ሃና_NNP ምን_WP ፈለግሽ_VBG ?_PUN

Afaan Oromo POS tagged: Haanaa_NNP maal_WP barbaadee_VBG ?_PUN

As shown in the Amharic interrogative sentences, when the interrogative pronouns (መቼ, ምን) come before the verbs (መጣህ, ፈለግሽ) then the interrogative pronouns (yoom, maal) in Afaan Oromo also come before the verbs (dhufte, barbaadee).

በግ የገዛው **ማን** ነው? /*bäg yägäzaw man näw* ?/ [Who bought a sheep?]

Hoolaa kan bitte **eenyu** dha?

When the interrogative pronoun ‘ማን’ comes after the verb ‘የገዛው’ in the Amharic interrogative sentence, then the interrogative pronoun ‘eenyu’ also comes after the verb ‘kan bitte’ in the Afaan Oromo interrogative sentence. Therefore, for such kinds of Amharic and Afaan Oromo interrogative sentences reordering rule is not required.

In Amharic to Afaan Oromo translation, Amharic reordering rules are used to make Amharic sentences in the corpus to have a similar sentence structure with that of Afaan Oromo and vice versa. In this section, Amharic/Afaan Oromo reordering rules are discussed, which are used to perform syntactic reordering on Amharic/Afaan Oromo words in the sentence.

Reordering Rule for compound word

A compound word (CW) is a combination of two words that can be treated as a single word in a sentence. Consider the following Amharic compound word and its Afaan Oromo translation:

Amharic: ትምህርት ቤት_CW /tmhrt bet/ [School]

Afaan Oromo: Mana barnootaa_CW

The compound word in Amharic has different word order compared to the compound word in Afaan Oromo. In order to have a similar structure in both Amharic and Afaan Oromo sentences, we apply the reordering rule defined by the Algorithm 4.1 to the Amharic/Afaan Oromo sentences.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.
Store all sentences in S.
Store all words in W.
For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.
For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.
If POS (w_j) is 'CW'.
$temp = w_j$
$w_j = w_{j-1}$
$w_{j-1} = temp$
Store s_i in S'
End if
End for
End for
Write S' to file

Algorithm 4.1: Algorithm for reordering compound words

Reordering Rule for noun phrases

A noun phrase is a phrase where the head word is a noun. A noun phrase can be a single word, just the noun or more than one word. Noun phrases can function in several different ways in a sentence. For instance, a noun phrase can be a subject, a direct object, the object of a preposition or an

indirect object. This section discusses noun phrases and their different functions in Amharic and Afaan Oromo sentences.

Consider the following Amharic noun phrases and their translation in Afaan Oromo.

Amharic: የስንዴ ዳቦ /*yäsnde dabo*/ [a wheat bread]

Afaan Oromo: Daabo kamadi

Amharic: የማር ጠጅ /*yämar Täj*/

Afaan Oromo: daadii damma

Amharic: የገብስ ጠላ /*yägäbs Täla*/

Afaan Oromo: Farso garbuu.

Amharic POS tagged: የስንዴ_NP ዳቦ_NN

Afaan Oromo POS tagged: Daabo_NN kamadi_NP

Amharic POS tagged: የማር_NP ጠጅ_NN

Afaan Oromo POS tagged: daadii_NN damma_NP

Amharic POS tagged: የገብስ_NP ጠላ_NN

Afaan Oromo POS tagged: Farso_NN garbuu_NP

The head words ዳቦ */dabo/*, ጠጅ */Täj/* and ጠላ /*Täla*/ are nouns. The noun phrases የስንዴ /*yäsnde*/ and የማር */yämar/* and የገብስ /*yägäbs/* indicate from what the head words ዳቦ */dabo/*, ጠጅ */Täj/* and ጠላ /T*äla*/ are made from respectively.

Consider the following noun phrases

Amharic: የበጋ ፀሀይ /*yäbäga Ťähäy*/ [the winter sun]

Amharic: የሰሜን ኮከብ /*yäsämen kokäb*/ [the North Star]

Amharic: የእርሻ መሬት */yäxrśa märet/* [agricultural land]

Amharic POS tag: የበጋ_NP ፀሀይ_NN

Amharic POS tag: የሰሜን_NP ኮከብ_NN

Amharic POS tag: የእርሻ_NP መሬት_NN

ፀሀይ */tsehay/*, ኮከብ /*kokeb*/ and መሬት / *märet/* are nouns which are used as the head word. The noun phrase የበጋ / *yäbäga*/ indicates the wheather condition of the sun, የሰሜን /*yäsämen*/ indicates the direction of the ኮከብ /*kokeb*/ and የእርሻ / *yäxrśa* / indicates the usage of the land.

Consider the following Amharic noun phrases and their translation in Afaan Oromo.

Amharic: የካሳ መፅሃፍ */yäkasa mäĬhaf/* [Kasa's book]

Afaan Oromo: Kitaaba Kaasaa

Amharic: የአስቴር ቤት /*yäxäster bet*/ [Aster's house]

Afaan Oromo: Mana Aster.

The noun phrases የካሳ */yäkasa/ and* የአስቴር /*yäxäster*/ indicate the owner of the book /መፅሃፍ/ and house/ቤት/ respectively.

From the above Amharic and Afaan Oromo noun phrases discussion, the structure for the above Amharic noun phrases is NP => NP NN and the structure for the above Afaan Oromo noun phrases is NP = > NN NP. From the above discussion, Amharic noun phrases have different structure from Afaan Oromo noun phrases. In order to have a similar structure in both Amharic and Afaan Oromo sentences, we apply the reordering rule defined by the Algorithm 4.2 to the Amharic/Afaan Oromo sentences.

Now consider the following example of Amharic sentence and its translation in Afaan Oromo where the noun phrase is used as a *direct object*.

Amharic: ገመቹ የድንጋይ ቤት ሰራ ። */gämäcu yädngay bet sera/* [Gemechu made a house from stone.]

Afaan Oromo: Gammachuun mana dhagaa ijaare.

Amharic POS tagged: ገመቹ_NNP የድንጋይ_NP ቤት_NN ሰራ_VBD ።_PUN

Afaan Oromo POS tagged: Gammachuun_NNP mana_NN dhagaa_NP ijaare_VBD ._PUN

The noun phrases 'የድንጋይ ቤት' and 'mana dhagaa' are used as a direct object in the Amharic and Afaan Oromo sentences respectively. They have different order. In order to have a similar structure in both Amharic and Afaan Oromo sentences, we apply the reordering rule defined by Algorithm 4.2 to the Amharic/Afaan Oromo sentences that have a noun phrase used as a direct object.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'NP' and POS (w_{j+1}) is 'NN' or 'NNS' or 'NNP' or 'NNPS') OR

If (POS (w_j) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+1}) is 'NP')

$temp = w_j$

$w_j = w_{j+1}$

$w_{j+1} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.2: Algorithm for reordering noun phrases

Reordering Rule for adjective words

A noun phrase can be a single word modified by an adjective. Consider the following example of Amharic sentence and its translation in Afaan Oromo where the noun phrase is used as a subject.

Amharic: ቢጫው መኪና የሚሸጥ ነው ። */biCaw mäkina yämiśät näw/* [The yellow car is for sale.]

Afaan Oromo: Konkoolaataa booran kan gurguramuu dha.

Amharic POS tagged: ቢጫው_JJ መኪና_NN የሚሸጥ ነው ።_PUN

Afaan Oromo POS tagged: Konkoolaataa_NN booran_JJ kan gurguramuu dha ._PUN

In the Amharic sentence, the noun phrase 'ቢጫው መኪና' is used as a subject of the sentence and the noun adjective 'ቢጫው' appears before the noun 'መኪና' whereas in Afaan Oromo sentence, the noun adjective 'booran' appears after the noun 'konkoolaataa'. In order to have a similar

structure in both Amharic and Afaan Oromo sentences, we apply the reordering rule defined by the Algorithm 4.3 to the Amharic/Afaan Oromo sentences that contain adjectives.

Load Amharic/Afaan Oromo sentences from POS tagged corpus

Store all sentences in S

Store all words in W

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S

For each w_j in s_i do, where j=0, 1, 2, …, n-1, where n is the number of words in s_i

If (POS (w_j) is 'JJ' and POS (w_{j+1}) is 'NN' or 'NNS' or 'NNP' or 'NNPS')

OR (POS (w_j) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+1}) is 'JJ')

$temp = w_j$

$w_j = w_{j+1}$

$w_{j+1} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.3: Algorithm for reordering adjective words

Reordering Rule for sentences containing a noun phrase and a compound word

Consider the following Amharic sentence containing a noun phrase and a compound word and Afaan Oromo translation, where the noun phrases 'የቂሊንጦ' and 'Qilinxoon' are used with the compound words 'ማረሚያ ቤት' and 'Mani Sirressaa' respectively.

Amharic: የቂሊንጦ ማረሚያ ቤት ተቃጠለ ። /*yäqilinTo marämiya bet täqaTälä*/

Afaan Oromo: Mani Sirressaa Qilinxoon gubate.

Amharic POS tagged: የቂሊንጦ_NP ማረሚያ_NN ቤት_CW ተቃጠለ_VBD ።_PUN

Afaan Oromo POS tagged: Mani_NN Sirressaa_CW Qilinxoon_NP gubate_VBD ._PUN

The Amharic sentence that contains the noun phrase 'የቂሊንጦ' and the compound word 'ማረሚያ ቤት' has different word order compared to its equivalent Afaan Oromo translated sentence.

In order to have a similar sentence structure in both Amharic and Afaan Oromo sentences that contain a noun phrase and a compound word, we apply the reordering rule defined by the Algorithm 4.4 to the Amharic sentence and Algorithm 4.5 to the Afaan Oromo sentence respectively.

Load Amharic sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'NP' and POS (w_{j+1}) is 'NN' and POS (w_{j+2}) is 'CW')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.4: Algorithm for reordering Amharic sentences containing a noun phrase and a compound word

Load Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'NN' and POS (w_{j+1}) is 'CW' and POS (w_{j+2}) is 'NP')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.5: Algorithm for reordering Afaan Oromo sentences containing a noun phrase and a compound word

Reordering Rule for sentences containing an adjective and a compound word

Amharic noun phrase could be constructed from an adjective followed by a compound word and in Afaan Oromo a noun phrase could be a compound word followed by an adjective.

Consider the following Amharic sentence containing a noun phrase modified by the adjective and its equivalent Afaan Oromo translation.

Amharic: ከፍተኛ ፍርድ ቤት */käftäNa frd bet/* [higher court]

Afaan Oromo: Mana murtii olaanaa.

Amharic POS tagged: ከፍተኛ_JJ ፍርድ_NN ቤት_CW ።_PUN

Afaan Oromo POS tagged: Mana_NN murtii_CW olaanaa_JJ ._PUN

The Amharic compound word 'ፍርድ ቤት' that is modified by the adjective 'ከፍተኛ' has different word order compared to Afaan Oromo compound word 'Mana murtii' modified by the the adjective 'olaanaa'. In order to have a similar structure in both Amharic and Afaan Oromo noun phrases, we apply the reordering rule defined by the Algorithm 4.6 to Amharic and Algorithm 4.7 to Afaan Oromo noun phrases that contain a compound word modified by an adjective respectively.

Load Amharic sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'JJ' and POS (w_{j+1}) is 'NN' or 'NNP' or 'NNPS' or 'NNS' and POS (w_{j+2}) is 'CW')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.6: Algorithm for reordering Amahric sentences containing an adjective and a compound word

Load Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'NN' or 'NNP' or 'NNPS' or 'NNS' and POS (w_{j+1}) is 'CW' and POS (w_{j+2}) is 'JJ')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.7: Algorithm for reordering Afaan Oromo sentences containing an adjective and a compound word

Reordering Rule for possessive pronouns

Possessive pronouns are pronouns that show possession or ownership of something in a sentence. Consider the following Amharic sentence containing possessive pronoun and its translation in Afaan Oromo.

Amharic: የእሱ ላሞች ሳር እየጋጡ ነው ። /*yäxsu lamoc sar xyägaTu näw*/ [His cows are gr azing grass].

Afaan Oromo: Saawwan isa margaa nyacha jiran.

Amharic POS tagged: የእሱ_PRP$ ላሞች_NNS ሳር_NN እየጋጡ_VBG ነው_AUX ።_PUN

Afaan Oromo POS tagged: Saawwan_NNS isa_PRP$ margaa_NN nyacha_VBG jiran_AUX ._PUN

In the Amharic sentence, the plural noun 'ላሞች' comes after the possessive pronoun 'የእሱ' but in the Afaan Oromo sentence the possessive pronoun 'Taaddasa' comes after the plural noun 'saawwan'. In order to have a similar structure in both Amharic and Afaan Oromo sentences, we apply the reordering rule defined by the Algorithm 4.8 to the Amharic/Afaan Oromo sentences containing possessive pronouns.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.
Store all sentences in S.
Store all words in W.
For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.
For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.
If (POS (w_j) is 'PRP$' and POS ($w_{j+1}$) is 'NN' or 'NNS' or 'NNP' or 'NNPS') OR
If (POS (w_j) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+1}) is 'PRP$')
$temp = w_j$
$w_j = w_{j+1}$
$w_{j+1} = temp$
Store s_i in S'
End if
End for
End for
Write S' to file

Algorithm 4.8: Algorithm for reordering possessive pronouns

Reordering Rule for cardinal numbers

Cardinal numbers refer to the counting numbers, because they show quantity. Consider the following Amharic and Afaan Oromo sentences that contain cardinal numbers.

Amharic: እኔ ሶስት ቋንቋዎችን እናገራለው ። /*xne sost qWanqWawocn xnagäralähu*/

[I speak three languages].

Afaan Oromo: Ani Afaawwan sadi dhubadaa.

Amharic: የእኔ ልጅ ሁለት ድመቶች አሉት ።/*yäxne lj hulät dmätoc alut*/[My son has two cats].

Afaan Oromo: Mucaan koo adurreewwan lama qabaa.

Amharic POS tagged: እኔ_PRP ሶስት_CD ቋንቋዎችን_NNS እናገራለው_VBG ።_PUN

Afaan Oromo POS tagged: Ani_PRP Afaawwan_NNS sadi_CD dhubadaa_VBG ._PUN

Amharic POS tagged: የእኔ_PRP$ ልጅ_NN ሁለት_CD ድመቶች_NNS አሉት_AUX ።_PUN

Afaan Oromo POS tagged: Mucaan_NN koo_PRP$ adurreewwan_NNS lama_CD qabaa_AUX ._PUN

In the Amharic sentences, cardinal numbers ‘ሶስት’ and ‘ሁለት’ are placed before nouns ‘ቋንቋዎችን’ and ‘ድመቶች’ whereas in Afaan Oromo cardinal numbers ‘sadi’ and ‘lama’ are placed after nouns ‘Afaawwan’ and ‘adurreewwan’ respectively. The reordering of Amharic and Afaan Oromo sentences that contain cardinal numbers is done by Algorithm 4.9

Load Amharic/Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'CD' and POS (w_{j+1}) is 'NN' or 'NNS' or 'NNP' or 'NNPS') OR

If (POS (w_j) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+1}) is 'CD')

$temp = w_j$

$w_j = w_{j+1}$

$w_{j+1} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.9: Algorithm for reordering cardinal numbers

Reordering Rule for ordinary numbers

Ordinary numbers tell the order of things and their rank in a sentence. For example, consider the following Amharic sentence containing an ordinary number and its translation in Afaan Oromo.

Amharic: እኔ ሶስተኛውን መፅሐፍ አንብቤዋለው ። /*xne sostäNawn meĬhaf xänbbewalew*/

[I read the third book.]

Afaan Oromo: Ani kitaabicha saddaffaa dubbiseera.

Amharic POS tagged: እኔ_PRP ሶስተኛውን_ON መፅሐፍ_NN አንብቤዋለው_VBD ።_PUN

Afaan Oromo POS tagged: Ani_PRP kitaabicha_NN saddaffaa_ON dubbiseera_VBD ._PUN

In the Amharic sentence, the ordinary number 'ሶስተኛውን' is placed before the noun 'መፅሐፍ' whereas in the Afaan Oromo sentence, the ordinary number 'saddaffaa' is placed after the noun 'kitaabicha'. Algorithm 4.10 shows the reordering of ordinary numbers in both languages.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'ON' and POS (w_{j+1}) is 'NN' or 'NNS' or 'NNP' or 'NNPS') OR

If (POS (w_j) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+1}) is 'ON')

$temp = w_j$

$w_j = w_{j+1}$

$w_{j+1} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.10: Algorithm for reordering ordinary numbers

Reordering Rule for noun phrases modified by an adjective

Consider the following Amharic noun phrase modified by an adjective and its translation in Afaan Oromo.

Amharic: አዲስ የቤት መኪና */xädis yäbet mäkina/* [New house car]

Afaan Oromo: Konkoolaataa mana haaraa

Amharic POS tagged: አዲስ_JJ የቤት_NP መኪና_NN

Afaan Oromo POS tagged: Konkoolaataa_NN mana_NP haaraa_JJ

In the Amharic noun phrase, the adjective 'አዲስ' comes before the noun phrase 'የቤት መኪና' but in the Afaan Oromo sentence the adjective 'haaraa' comes after the noun phrase 'konkoolaataa mana'. In order to have a similar structure in both Amharic and Afaan Oromo noun phrases, we apply the reordering rule defined by the Algorithm 4.11 to the Amharic/Afaan Oromo noun phrases.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'JJ' and POS (w_{j+1}) is 'NP' and POS (w_{j+2}) is 'NN' or 'NNP' or 'NNPS' or 'NNS') OR

If (POS (w_j) is 'NN' or 'NNP' or 'NNPS' or 'NNS' and POS (w_{j+1}) is 'NP' and POS (w_{j+2}) is 'JJ')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.11: Reordering rule for noun phrases modified by adjectives

Reordering Rule for sentences containing possessive pronoun and a noun phrase

Consider the following Amharic noun phrase and its translation in Afaan Oromo.

Amharic: የአልማዝ የወርቅ ቀለበት */yäxälmaz yäwärq qäläbät/* [Almaz's gold ring.]

Afaan Oromo: Amartii waqee Almaaz

Amharic POS tag: የአልማዝ_PRP$ የወርቅ_NP ቀለበት_ NN

Afaan Oromo POS tag: Amartii_NN waqee_NP Almaaz_PRP$

In the Amharic noun phrase, 'የአልማዝ' is used as a possessive pronoun i.e., it is described as the owner of the propery described by the noun phrase 'የወርቅ ቀለበት'. 'የአልማዝ' comes before the noun phrase 'የወርቅ ቀለበት' but in the Afaan Oromo sentence the owner of the property 'Amartii waqee' comes after it. In order to have a similar structure in both Amharic and Afaan Oromo noun phrases, we apply the reordering rule defined by the Algorithm 4.12 to the Amharic/Afaan Oromo sentences containing a possessive pronoun and a noun phrase.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'PRP$' and POS ($w_{j+1}$) is 'NP' and POS ($w_{j+2}$) is 'NN' or 'NNS' or 'NNP' or 'NNPS') OR

If (POS (w_j) is NN or NNS or NNP or NNPS and POS (w_{j+1}) is 'NP' and POS (w_{j+2}) is 'PRP$')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.12: Reordering rule for sentences containing a possessive pronoun and a noun phrase

Reordering Rule for sentences containing a cardinal number and a noun phrase

Consider the following Amharic sentence containing a cardinal number and a noun phrase and its translation in Afaan Oromo.

Amharic: 5 የጃፓን መኪናዎች */5 yäjapan mäkinawoc/* [5 Japan's cars].

Afaan Oromo: Konkoolaataawan Jaapaan 5

Amharic POS tagged: 5_CD የጃፓን_NP መኪናዎች_NNS

Afaan Oromo POS tagged: Konkoolaataawan_NNS Jaapaan_NP 5_CD

In the Amharic phrase, the cardinal number '5' comes before the noun phrase 'የጃፓን መኪናዎች' and in the Afaan Oromo, the noun phrase 'Konkoolaataawan Jaapaan' comes before the cardinal number '5'.

In order to have a similar structure in both Amharic and Afaan Oromo sentences, we apply the reordering rule defined by the Algorithm 4.13 to the Amharic/Afaan Oromo sentences containing a cardinal number and a noun phrase.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'CD' and POS (w_{j+1}) is 'PRP$' or 'NP' and POS ($w_{j+2}$) is 'NN' or 'NNP' or 'NNPS' or 'NNS') OR

If (POS (w_j) is 'NN' or 'NNP' or 'NNPS' or 'NNS' and POS (w_{j+1}) is 'PRP$' or 'NP' and POS ($w_{j+2}$) is 'CD')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.13: Reordering rule for sentences containing a cardinal number and a noun phrase

Reordering Rule for sentences containing an ordinary number and a noun combination

Consider the following Amharic sentence containing an ordinary number and a noun phrase and its translation in Afaan Oromo.

Amharic: 2ተኛው ዙር ውድድር /2*täNaw zur wddr*/ [The 2nd round tournament].

Afaan Oromo: Waldorgommin marsaan 2ffaan

Amharic POS tagged: 2ተኛው_ON ዙር_NN ውድድር_NN

Afaan Oromo POS tagged: Waldorgommin_NN marsaan_NN 2ffaan_ON

In the Amharic the noun phrase '2ተኛው ዙር' containing the noun phrase '2ተኛው' comes before the noun 'ውድድር' and in the Afaan Oromo, the noun phrase 'marsaan 2ffaan' containing the ordinary number '2ffaan' comes after the noun 'Waldorgommin'.

Load Amharic/Afaan Oromo sentences from POS tagged corpus.

Store all sentences in S.

Store all words in W.

For each $s_i \in$ S do, where i=0, 1, 2, …, m-1, where m is the number of sentences in S.

For each w_j in s_i do, where j=0, 1, 2, …, n-1. where n is the number of words in s_i.

If (POS (w_j) is 'ON' and POS (w_{j+1}) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+2}) is 'NN' or 'NNS' or 'NNP' or 'NNPS') OR

If (POS (w_j) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+1}) is 'NN' or 'NNS' or 'NNP' or 'NNPS' and POS (w_{j+2}) is 'ON')

$temp = w_j$

$w_j = w_{j+2}$

$w_{j+2} = temp$

Store s_i in S'

End if

End for

End for

Write S' to file

Algorithm 4.14: Reordering rule for sentences containing an ordinary number and noun combination

In order to have a similar structure in both Amharic and Afaan Oromo phrases, we apply the reordering rule defined by the Algorithm 4.14 to the Amharic/Afaan Oromo sentences containing an ordinary number and a noun combination.

4.2.2 Language Model

Since the system is bidirectional, a language model has been developed for both Amharic and Afaan Oromo. For Amharic to Afaan Oromo translation, the language model $p(o)$ should be trained on a small amount of monolingual corpus in Afaan Oromo compared to the parallel corpus used for the translation model. The language model $p(o)$ estimates how likely a string is in a given taget language (Afaan Oromo) i.e., it prefers fluent sentences. For example: it prefers "Inni gara *mana deeme*" than "Inni gara *deeme mana*", i.e., probability (Inni gara mana deeme) > probability (Inni gara deeme mana). Similarly, for Afaan Oromo to Amharic translation, the language model $p(a)$ should be trained on a small amount of monolingual corpus in Amharic compared to the parallel corpus used for the translation model.

4.2.3 Translation Model

For a given sentence pair (e, f) the translation model is used to indicate the probability that f is the translation of e. Since the system is bidirectional, two translation models have been developed for both Amharic and Afaan Oromo. When the translation is from Amharic to Afaan Oromo, the translation model probability $p(a|o)$ is used to measure the quality of the translation of the source Amharic sentence *a* to the given target Afaan Oromo sentence *o*. The translation model $p(a|o)$ encodes the faithfulness of o as a translation of a. For example:

probability (እሱ ወደ ቤት ሄደ ። | Inni gara mana deeme.) >

probability (እሱ ወደ ቤት ሄደ ። | Inni gara mana deemte.) >

probability (እሱ ወደ ቤት ሄደ ። | Inni gara magaalaa deeme.)

Similarly, when the translation is from Afaan Oromo to Amharic, the translation model probability $p(o|a)$ is used to measure the quality of the translation of source Afaan Oromo sentence o to the given target Amharic sentence a. The translation model finds out the correspondence between the source sentence and the target sentence in the source/target parallel corpus, which is called word-alignment. The basic unit of the correspondence is word. The alignment between the source word

and the target word could be one-to-zero, one-to-one or one-to-many. The translation system can produce multiple words from a single word, but not vice versa and this is a limitation of word-based model. One of the ways to overcome this limitation is to use phrase-based translation. The basis of phrase-based translation is to fragment the input sentence into phrases (sequence of consecutive words), translate and reorder these phrases into the target language. The phrase-based translation process is broken up into the following three mapping steps as shown in Figure 4.2.

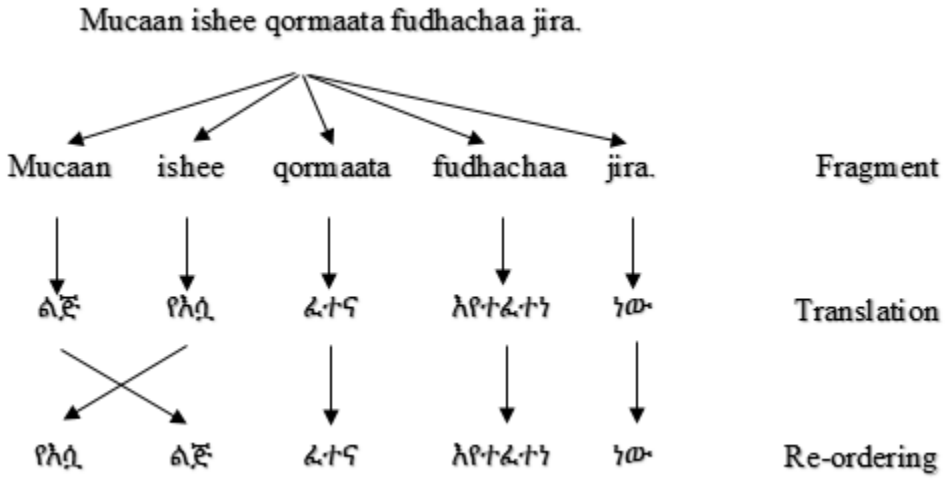

Figure 4.2: An example of phrase-based translation

4.2.4 Decoding

The decoder's task is aimed to find the best translation in the target language for a given input sentence by the statistical methods that count on the translation model and the language model. When translation is from Amharic to Afaan Oromo, the best translation is the one that maximizes the product of the probabilities $p(a|o)\ and\ p(o)$, i.e., $arg\max_{o}\ p(a|o) * p(o)$.

Similarly, when translation is from Afaan Oromo to Amharic, the best translation is the one that maximizes the product of the probabilities $p(o|a)\ and\ p(a)$, i.e., $arg\max_{a}\ p(o|a) * p(a)$.

CHAPTER FIVE: EXPERIMENT AND DISCUSSION

5.1 Introduction

Based on the design of Chapter Four, Amharic-Afaan Oromo bidirectional machine translation is experimented using a hybrid approach. This Chapter evaluates its performance by conducting two experiments by using a statistical approach and a hybrid approach.

5.2 Corpus Preparation

Hybrid approach requires bilingual parallel corpus. For this research work, parallel documents of Amharic and Afaan Oromo that are collected from Fana Broadcasting Corporate News[2], some chapters of the Holy Bible and other simple sentences are used.

The parallel corpus contains texts translated in Amharic and Afaan Oromo languages which are aligned at sentence level. After tokenizing, true casing and cleaning the collected Amharic-Afaan Oromo parallel corpus, we obtained exactly 1402 Amharic-Afaan Oromo parallel sentences. In this experiment, we have randomly selected around 7.2% of the total parallel sentences, i.e., 101 for testing the performance of the system and the rest around 93.8%, i.e., 1301 parallel sentences are used for training the system.

Four experiments were conducted using statistical and hybrid approaches. To conduct all the experiments, similar 1402 parallel sentences were used. The next section discusses each of these experiments.

5.3 Experiment I

The first two experiments, i.e., Amharic to Afaan Oromo translation and vice versa, were conducted by using a statistical approach.

5.3.1 Training the system

Moses which is freely available software is used to train the system in both directions, Amharic to Afaan Oromo and vice versa, by using similar and the same number of Amharic-Afaan Oromo parallel sentences. The training process includes the following procedures.

[2] http://www.fanabc.com

Language Model Training

The language model is used to ensure fluent output. Since the translation is bidirectional, the language model was built with Amharic as a target language for Afaan Oromo to Amharic translation and Afaan Oromo as a target language for Amharic to Afaan Oromo translation. IRSTLM toolkit was used to perform language modeling task. An appropriate 3-gram language model was built. First, the training was performed for Amharic to Afaan Oromo and then for Afaan Oromo to Amharic.

Training the Translation System

To train the translation model, we run word-alignment using GIZA++, phrase extraction and scoring, create lexicalized reordering tables and create Moses configuration file. The model specified by moses.ini file is used to decode/translate sentences from Amharic to Afaan Oromo and vice versa. The phrase table and reodering table were binarised, i.e., compiling them into a format that can be loaded quickly.

Tuning

Weights used by Moses to weight the different models against each other are not optimized. To find better weights we need to tune the translation system. Tuning requires a small amount of parallel data separate from the training data. Therefore, the parallel data was passed through tokenization and truecasing processes. The end result of tuning is an ".ini" file with trained weights.

5.3.2 Result of Test Set on Experiment I

We have used 101 Amharic and Afaan Oromo parallel sentences in order to measure and test the performance of the system in terms of the translation accuracy of translating a simple Amharic sentence to Afaan Oromo sentence and vice versa.

BLEU score methodology is used to see the result of the translation process. The result recorded from the BLEU score methodology shows 89.39% for Amharic to Afaan Oromo translation and 80.33% for Afaan Oromo to Amharic translation.

5.4 Experiment II

Two experiments were conducted on Amharic-Afaan Oromo language pair by using a hybrid approach.

First, the sentence reordering rules mentioned in Chapter Four are applied on the training and test sets, then a statistical approach is applied on the reordered corpus.

5.4.1 Training the system

Moses which is freely available software is used to train the system in both directions, Amharic to Afaan Oromo and vice versa, by using similar and the same number of Amharic-Afaan Oromo parallel sentences. The training process includes the same procedures described in Section 5.3.1.

5.4.2 Result of Test Set on Experiment II

We have used 101 Amharic and Afaan Oromo parallel sentences in order to measure and test the performance of the system in terms of the translation accuracy and the time it takes to translate a single Amharic simple sentence to Afaan Oromo sentence and vice versa.

BLEU score methodology is used to see the result of the translation process in both directions. The result recorded from the BLEU score methodology shows 91.56% for Amharic to Afaan Oromo translation and 82.24% for Afaan Oromo to Amharic translation.

5.5 Discussion

When translating from Amharic sentences to Afaan Oromo, for example, “የአንተ ስም ማነው?” is translated as “Maqaan kee eenyu?” but when translating Afaan Oromo sentence “Maqaan kee eenyu?” to Amharic, it can be translated as “የአንተ ስም ማነው?” or “የአንቺ ስም ማነው?”. Similarly, “እሱ ሻይ መጠጣት አይወድም” can be translated as “Inni shaayii dhugu hin jaalatu” but “Inni shaayii dhugu hin jaalatu” can be translated as “እሱ ሻይ መጠጣት አይወድም” or “እሱ ሻይ መጠጣት አትወድም”. These indicate Afaan Oromo words like “kee” and “hin jaalatu” can be translated in Amharic as “የአንተ” or “የአንቺ” and “አይወድም” or “አትወድም” respectively. But both Amharic words “አይወድም” and “አትወድም” are translated as “hin jaalatu” in Afaan Oromo. This means an Amharic word can have more than one meaning/equivalent in Afaan Oromo. This might be the reason behind the difference between the performances in Amharic to Afaan Oromo and Afaan Oromo to Amharic in both the experiments.

The experiments are conducted by using two different approaches. From the results of the experiments we can see that the result recorded from a BLEU score shows that the hybrid approach is better than the statistical approach for Amharic-Afaan Oromo bidirectional machine translation.

CHAPTER SIX: CONCLUSION AND FUTURE WORK

6.1 Introduction

This chapter concludes the thesis and highlights the main contributions that were achieved based on the stated objective. Finally, some suggestions and recommendations are made for future work that could be done in similar area of research.

6.2 Conclusion

In this study, we have developed a bidirectional Amharic-Afaan Oromo machine translation prototype using, hybrid approach. 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 POS tagging and to better guide the statistical engine. We have prepared manually tagged corpus for both Amharic and Afaan Oromo languages since there are no publicly available POS tagger tools for both languages. The linguistic background and nature of the two languages have been studied in order to design the reordering rules for different types of Amharic/Afaan Oromo phrases and sentences. Language modeling, translation modeling and decoding are all components of the statistical approach which are freely available on the web and incorporated in the translation system. The language model estimates how likely a string is in a given target language, Afaan Oromo or Amharic. A language model has been developed for both Afaan Oromo and Amharic because the system is bidirectional. The translation model is used to measure the quality of the translation of the source language sentence given the target language sentence. Just like language models, two translation models were developed one for Amharic and the other for Afaan Oromo. The decoder is 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.

Amharic-Afaan Oromo hybrid bidirectional machine translation design involves collection of Amharic and Afaan Oromo parallel corpus, corpus preparation, POS tagging, implementing the reordering rules for Amharic and Afaan Oromo sentences using ASP.Net C# programming and SQL server 2014 as back end, language modeling by using IRSTLM tool, translation modeling by

using GIZA++ (for creating word alignment from the parallel corpus) and training the system by using Moses.

Finally, two experiments were conducted by using the collected data set to check the accuracy of the system 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. From the test results of the conducted experiments in this research, it can be concluded that the hybrid approach is better than the statistical approach.

6.3 Contribution

The contribution of this study is to confirm that hybrid machine translation approach is a better option to translate Amharic to Afaan Oromo and vice versa. This approach was capable of translating different Amharic and Afaan Oromo phrases and simple sentences containing compound words, adjectives, noun phrases, possessive pronounus, cardinal and ordinary numbers. Additionally, the parallel corpus used for this study can be used as input for other similar researches areas.

6.4 Future Work

This research work is developed in order to translate Amharic sentences into Afaan Oromo and vice versa. The system can further be enhanced with the following possible future works:

- Better results may be obtained by increasing the size of the parallel corpus used for training the system.
- Incorporating components like automatic POS tagger, morphological analyzer and generation may increase the performance of the translation system.
- Better results may be obtained by incorporating word sense disambiguation component.
- The sentence reordering rules can be expanded to handle complex sentences.

References

[1] Kituku, Benson, Lawrence Muchemi, and Wanjiku Nganga. "A Review on Machine Translation Approaches," *TELKOMNIKA Indonesian Journal of Electrical Engineering and Computer Science,* Vol. 1, No. 1, 2016, pp 182-190.

[2] Amal Ganesh and Aasha V.C., "Rule Based Machine Translation: English to Malayalam: A Servey" in *Proceedings of* 3rd *International Conference on Advanced Computing, Networking and Informatics,* India, Orissa, October 2015, Vol 43, pp 447-454.

[3] Benjamin Elisha Sawe, "What Languages Are Spoken In Ethiopia," retrieved from https://www.workdatlas.com/articles/what-languages-are-spoken-in-ethiopia.html, Last access on March 05, 2020.

[4] Joel Ilao, Jasmine Ang, Marc Randell Chan and Joyce Uy, "*Filipino-to-English Bidirectional Statistical Machine Translation Using Feedback*", retrieved from https://www.researchgate.net/publication/280561598_FEBSMT_Filipino-to-English_Bidirectional_Statistical_Machine_Translation_Using_Feedback, Last accessed on September 25, 2018.

[5] Yin Yin Win, Aye Thida, "*Myanmar-English Bidirectional Machine Translation System with Numerical Particles Identification*", retrieved from http://www.mecs-press.org/ijitcs/ijitcs-v8-n6/IJITCS-V8-N6-5.pdf, Last accessed on October 01, 2018.

[6] Eleni Teshome, "*Bidirectional English – Amharic Machine Translation: An Experiment using Constrained Corpus*", Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2013.

[7] Jabesa Daba and Yaregal Assabie, "A Hybrid Approach to the Development of Bidirectional English-Oromiffa Machine Translation" in *Proceedings of the 9th International Conference on NLP,* Warsaw, Poland, September 2014.

[8] Sisay Adugna, "*English-Oromo Machine Translation: An Experiment Using a Statistical Approach*", Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2009.

[9] Akubazgi Gebremariam, "*Amharic to Tigrigna Machine Translation using Hybrid Approach: An Experiment Using a Statistical Approach*", Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2017.

[10] The Language Gulper: *"An insatiable appetite for ancient and modern tongues"*, retrieved from, http://www.languagesgulper.com/eng/Amharic.html, Last accessed on February 04, 2020.

[11] Alicia and Miguel Grinberg, retrieved from, https://www.amharicmachine.com/default/alphabet, Last accessed on February 04, 2020.

[12] ባየ ይማም, *አጭርና ቀላል የአማርኛ ሰዋስው*, አልፋ አታሚዎች, Addis Ababa, 2010.

[13] Abeba Ibrahim, "*A Hybrid Approach to Amharic Base Phrase Chunking and Parsing"*, Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2013.

[14] Addis Ashagre, "*Automatic Summarization for Amharic Text using Open Text Summarizer"*, Unpublished Masters Thesis, School of Information Science, Addis Ababa University, Ethiopia, 2013.

[15] ጌታሁን አማረ, *የአማርኛ ሰዋስው በቀላል አቀራረብ, የተሻሻለ ሁለተኛ ዕትም*, አዲስ አበባ ዩኒቨርሲቲ ቢ.ኢ ማተሚያ ቤት, Addis Ababa, 2010.

[16] Mohammed Hussen, *"Part-of-speech tagging for Afaan Oromo language using Transformational Error driven Learning (TEL) approach",* Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2010.

[17] John H.Spencer, *Ethiopia at Bay: A Personal Account of the Haile Sellassie Years*, Online book uploaded 2017.

[18] Susan Russell, *Amharic ግዕዝ TESL 539: Language Group Report Spring 2009*, retrieved from, http://www.ritell.org/Resources/Documents/language%20project/Amharic%20.pdf, Last access on March 09, 2019.

[19] Ruth Kramer, "The Amharic Definite Marker and the Syntax-Morphology Interface", *Annual Meeting of the Linguistic Society of Americal,* University of California, Santa Cruz, 2008.

[20] Ibrahim Bedane, "The Origin of Afaan Oromo: Mother Language", *Global Journal of human-social science,* Vol. 15, Issue 12, 2015.

[21] Abdi Sani, "*Afaan Oromo Named Entity Recognition Using Hybrid Approach"*, Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2015.

[22] Abebe Mideksa, "*Statistical Afaan Oromo grammar checker"*, Unpublished Masters Thesis, School of Information Science, Addis Ababa University, Ethiopia, 2015.

[23] Gezehagn Gutema, "*Afaan Oromo Text Retrieval System"*, Unpublished Masters Thesis, School of Information Science, Addis Ababa University, Ethiopia, 2012.

[24] Wakshum Temesgen, "*Effect of morphological information in Afaan Oromo word sequence prediction"*, Unpublished Masters Thesis, School of Information Science, Addis Ababa University, Ethiopia, 2017.

[25] Getachew Mamo and Million Meshesha, "Parts of Speech Tagging for Afaan Oromo", *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence,* Vol. 1, No. 3, 2011.

[26] Kula Kekeba Tune, *"Development of Cross-Lingual Information Retrieval for Resource-Scarce African Languages",* Thesis submitted for the Degree of PhD in Computer Science and Engineering, International Institute of Information Technology, Hyderabad, Deemed University, India, 2015.

[27] Debela Tesfaye, "A rule-based Afaan Oromo Grammar Checker", *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 8, 2011, pp. 126 – 130.

[28] Abraham Gizaw, "*Improving Brill's tagger lexical and transformation rule for Afaan Oromo Language"*, Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, Ethiopia, 2013.

[29] Assefa W/Mariam, "*Development of Morphological Analyzer for Afaan Oromo*", Unpublished Masters Thesis, Department of Information Science, Addis Ababa University, Ethiopia, 2005.

[30] Debela Tesfaye, "*Designing a Stemmer for Afaan Oromo Text: A hybrid approach*", Unpublished Masters Thesis, Department of Information Science, Addis Ababa University, Ethiopia, 2010.

[31] Fiseha Berhanu, "*Afaan Oromo Automatic News Text Summarizer Based on Sentence Selection Function*", Unpublished Masters Thesis, Department of Information Science, Addis Ababa University, Ethiopia, 2013.

[32] Birhanu Demie, "*The Impact of Afaan Oromo dialectal variations on teaching-learning process of the language"*, Unpublished Masters Thesis, Department of Linguistics and Philology, Addis Ababa University, Ethiopia, 2010.

[33] Daniel Bekele, "*Afaan Oromo-English Cross-Lingual Information Retrieval (CLIR): A corpus based approach"*, Unpublished Masters Thesis, School of Information Science, Addis Ababa University, Ethiopia, 2011.

[34] Ekta Gupta, Shailendra Kumar Shrivastava. "A Result Analysis of Translation Techniques of English to Hindi Online Translation Systems," *International Journal of Computer Applications (0975 – 8887),* Vol. 156, No. 12, 2016.

[35] M. D. Okpor. "Machine Translation Approaches: Issues and Challenges," *IJCSI International Journal of Computer Science Issues,* Vol. 11, No. 2, 2014.

[36] V.C. Aasha and Amal Ganesh, "Rule Based Machine Translation: English to Malayalam: A Survey" in *Proceedings of the 3rd International Conference on Advanced Computing, Networking and Informatics,* India, January 2016.

[37] Shantanoo Dubey, "Survey of Machine Translation Techniques," *International Journal of Advanced Research in Computer Science and Management Studies,* Vol. 5, Issue 2, 2017, pp 39-51.

[38] Sadik Bessou and Mohamed Touahria, “Morphological Analysis and Generation for Machine Translation from and to Arabic”, *International Journal of Computer Applications,* Vol. 18, No. 2, 2011.

[39] Anand Balladb and Umesh Chandra Jaiswal, “A Study of Machine Translation Methods and Their Challenges”, *International Journal of Advance Research in Science and Engineering,* Vol. 4, No. 2, 2015.

[40] Benson Kituku, Lawrence Muchemi, Wanjiku Nganga, “A Review on Machine Translation Approaches: Issues and Challenges,” *Indonesian Journal of Electrical Engineering and Computer Science,* Vol. 1, No. 1, 2016, pp. 182-190.

[41] Zhou Dajun and Wang Yun, “Corpus-based Machine Translation: Its current development and perspectives”, *International forum of teaching and studies,* Vol. 11 No. 1-2, 2015.

[42] Thai Phuong Nguyen and Akira Shimazu, “Improving Phrase-Based SMT with Morpho-Syntactic Analysis and Transformation” in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* Cambridge, August 2006.

[43] Ahmed Fasis, Hisham Salam, “Smoothing Techniques evaluation of n-gram language model for Arabic OCR post-processing”, *Journal of Theoretical and Applied Information Technology,* Vol. 82, No. 3, 2015.

[44] Antony P J., “Machine Translation Approaches and Survey for Indian Languages,” *International Journal of Computational Linguistics and Chinese Language Processing,* Vol. 18, No. 1, 2013, pp. 47-78.

[45] Koehn P, Och J, Daniel Marcu, “Statistical Phrase-Based Translation” in *Proceedings of HLT NAACL,* Edmonton, May – June 2003, pp. 48-54.

[46] Nagao M., “A framework for mechanical translation between English and Japanese by Analogy principle.” *Artificial and Human Intelligence, North Holland,* 1984, pp 173-180.

[47] Harold Somers, “Review article: Example based machine translation”, *Machine Translation 14,* Vol. 4, pp 113-145, 1999.

[48] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems,* pp. 3104-3112, 2014.

[49] Dzmitry Bahdanau, KyugnHyun Cho, Yoshua Bengio, Neural Machine Translation By Jointly Learning to align and translate. Published as a conference paper at ICLR 2015.

[50] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *arXiv preprint arXiv:1406.1078*, 2014.

[51] Parnia Bahar, Christopher Brix and Hermann Ney, "Towards Two Dimensional Sequence Model in Neural Machine Translation", in *Proceedings of 2018 Conference on Emprical Methods in Natural Language Processing,* Brussels, Belgium, November 2018, pp 3009-3015.

[52] Niessen, S., F.J Och, G. Leusch, and H. Ney, "An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research", in *Proceesing of the 2nd International Conference on Language Resources and Evaluation,* Athens, Greece. 2000.

[53] Tillmann C., S. Vogel, H. Ney, H. Sawaf and A. Zubiaga, "Accelerated DP based Search for Statistical Translation", in *Proceedings of the 5th European Conference on Speech Communication and Technology,* Rhodes, Greece, 1997.

[54] Maja Popovic, Hermann Ney, "Word Error Rates: Decomposition over POS Classes and Applications for Error Analysis", in *Proceedings of the 2nd Workshop on Statistical Machine Translation,* Prague, 2007.

[55] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* Philadelphia, July 2002, pp 311-318.

[56] Laith S. Hadla, Taghreed M. Hailat and Mohammed N. Ak-Kabi, "Comparative Study Between METEOR and BLEU Methods of MT: Arabic into English Translation as a Case Study", *International Journal of Advanced Computer Science and Applications,* Vol. 6, No. 11, 2015.

Annex I: Sample Amharic and Afaan Oromo Tagged Sentences for Training

ሃና_NNP ምን_WP ፈለግሽ_VBG ?_PUN	Haanaa_NNP maal_WP barbaadee_VBG ?_PUN
ልጁ_NNP ወተት_NN ጠጣ_VBD ።_PUN	Mucaan_NN aannan_NN dhuge_VBD ._PUN
ቢጫው_JJ መኪና_NN የሚሸጥ_VBG ነው_AUX ።_PUN	Konkoolaataan_NN booran_JJ kan_UNK gurguramuu_VBG dha_AUX ._PUN
የታደሰ_PRP$ ላሞች_NNS ሳር_NN እየጋጡ_VBG ነው_AUX ።_PUN	Saawwan_NNS Taaddasa_PRP$ margaa_NN nyacha_VBG jiran_AUX ._PUN
1ኛው_ON ዙር_NN ውድድር_NN ።_PUN	Waldorgommin_NN marsaan_NN 1ffaan_ON ._PUN
የገመቹ_PRP$ መኪና_NN ነጭ_JJ ነው_AUX ።_PUN	Konkoolaataan_NN Gammachuu_PRP$ adii_JJ dha_AUX ._PUN
ሰላም_NNP ስልክ_NK እያወራች_VBG ነው_AUX ።_PUN	Salaam_NNP bilbilaan_NK haasa'a_VBG jirti_AUX ._PUN
የምርጫው_NP ቀን_NN ተራዘመ_VBD ።_PUN	Guyyaa_NN filannichaa_NP dheereffame_VBD ._PUN
ቦንቱ_NNP የኮምፒውተር_NP ጨዋታ_NN መጫወት_RB ትወዳለች_VBG ።_PUN	Boontun_NNP tapha_NN kompiitari_NP taphachuu_NN jaalatti_VBZ ._PUN
እነዚህ_PRP$ ላሞች_NNS የእሷ_PRP$ ናቸው_AUX ።_PUN	Saawwan_NNS kana_PRP$ kan_UNK ishee_PRP$ dha_AUX ._PUN
የእሱ_PRP$ ትልቁ_JJ ቤት_CW አዲስ_JJ ነው_AUX ።_PUN	Mani_NN guddaan_JJ isaa_PRP$ haaraa_JJ dha_AUX ._PUN
ፋጡማ_NNP ቆንጆ_JJ መኪና_NN አላት_AUX ።_PUN	Faaxumaan_NNP konkoolaataa_NN bareedaa_JJ qabdi_AUX ._PUN
የእኔ_PRP$ ልጅ_NN ሁለት_CD ድመቶች_NNS አሉት_AUX ።_PUN	Mucaan_NN koo_PRP$ adurreewwan_NNS lama_CD qaba_AUX ._PUN
እኛ_PRP የገብስ_NP ጠላ_NN አንጠጣም_VBG ።_PUN	Nu'i_PRP farsoo_NN garbuu_NP hindhugnuu_VBG ._PUN
እነዚህ_PRP$ በጎች_NNS የእሷ_PRP$ ናቸው_AUX ።_PUN	Hoolotta_NNS kana_PRP$ kan_UNK ishee_PRP$ dha_AUX ._PUN

እሱ_PRP ሞቀሌ_NNP ላይ_IN ትልቅ_JJ ሆቴል_NN ገዛ_VBD ።_PUN	Inni_PRP Maqalee_NNP irraa_IN hooteela_NN guddaa_JJ bitte_VBD ._PUN
ጫላ_NNP እያለቀሰ_VBG ነው_AUX ።_PUN	Caalaan_NNP boo'aa_VBG jira_AUX ._PUN
እኔ_PRP ሁለት_CD ቋንቋዎችን_NNS እናገራለው_VBG ።_PUN	Ani_PRP Afaawwan_NNS lama_CD dhubadaa_VBG ._PUN
የእሱ_PRP$ ቤት_NN ተቃጠለ_VBD ።_PUN	Mani_NN isaa_PRP$ gubate_VBD ._PUN
አብዲ_NNP ነገ_UNK ይሄዳል_NK ።_PUN	Abdiin_NNP bor_UNK nideema_VBG ._PUN
አብዲ_NNP ዶክተር_NN ነው_AUX ።_PUN	Abdiin_NNP dooktar_NN dha_AUX ._PUN
ሽማግሌው_NK ቤተሰብ_NK ይመርቃል_NK ።_PUN	Jaarsichi_NK maatii_NK eebbisa_NK ._PUN
ትልቅ_JJ ፈረስ_NN ነው_AUX ።_PUN	Fardi_NN guddaa_JJ dha_AUX ._PUN
ቦንሳ_NNP ቀይ_JJ እስክሪብቶ_NN አለው_AUX ።_PUN	Boonsaan_NNP biirii_NN diimaa_JJ qaba_AUX ._PUN
አብዲ_NNP ሱዳን_NNP ነበር_AUX ።_PUN	Abdiin_NNP Sudaan_NNP ture_AUX ._PUN
ቤቱ_NN አዲስ_JJ ነው_AUX ።_PUN	Mani_NN haaraa_JJ dha_AUX ._PUN
ሞስኮቱ_NN ተሰብሯል_VBD ።_PUN	Foddaan_NN cabame_VBD ._PUN
አብዲ_NNP ቤት_NN ገዛ_VBD ።_PUN	Abdiin_NNP mana_NN bite_VBD ._PUN
ጠቅላይ_NN ሚኒስትር_NN ዶክተር_NN አብይ_NNP በአስመራ_NP ከተማ_NN ችግኝ_NN ተከሉ_VBD ።_PUN	Muummeen_NN ministiraa_NN Dooktar_NN Abiy_NNP magaalaa_NN Asmaraatti_NP biqiltuu_NN dhaabaniiru_VBD ._PUN
ውድድሩ_NN ላይ_IN ለሞሳተፍ_NP 400_CD ተወዳዳሪዎች_NNS መቀሌ_NNP ገብተዋል_VBD ።_PUN	Waldorgommicha_NN irraa_IN hirmaachuuf_NP dorgomtoonni_NNS 400_CD Maqalee_NNP galaniiru_VBD _UNK ._PUN
በአፍሪካ_NP ዋንጫ_NN ናይጄሪያ_NNP 3ኛ_ON ደረጃ_NN በመያዝ_NP አጠናቀቀች_VBD ።_PUN	Waancaa_NN Afrikaan_NP Naayjeeriyaan_NNP sadarkaa_NN 3ffaa_ON qabachuun_NP xumurteetti_VBD ._PUN

Annex II: Sample Parallel Corpus for Testing

አብዲሳ ቤት መስራት ይፈልጋል ።	Abdiisaan mana hojjechu barbaada.
እኔ መፅሐፍ ገዛው ።	Ani kitaaba bite.
የአንተ ስም ማነው ?	Maqaan kee eenyu?
እኔ ኬንያ ነበርኩ ።	Ani Kenyaa ture.
በደረሰው አደጋ በ5 ሰዎች ላይ ከፍተኛ የአካል ጉዳት ደረሰ ።	Balaa ga'een namoota 5 irraa midhaa qaamaa cimaa qaqqabee.
አቶ ደመቀ ምክትል ጠቅላይ ሚኒስትር ናቸው ።	Obbo Dammaqaan Ittiaanaa Muummeen ministiraa dha.
አየለ ተማሪ ነው ነገርግን እሷ አስተማሪ ነች ።	Ayyalaan barataa dha garuu isheen barsiistuu dha.
ወንድሜ አዲስ መኪና እየነዳ ነው ።	Obboollessan koo konkoolaataa haaraa oofaa jira.
ቃልኪዳን መረብ ኳስ መጫወት ትወዳለች ።	Kaalkidaan kubbaa saaphana taphachuu jaalatti.
ጫላ አስተማሪ እና ዶክተር መሆን ይፈልጋል ።	Caalaan barsiisaa fi dooktar ta'uu barbaada.
እሷ መረብ ኳስ መጫወት ትወዳለች ።	Isheen kubbaa saaphana taphachuu jaalatti.
እሱ አልተማረም ነገርግን ብዙ ነገር ያውቃል ።	Inni hinbarannee garuu waan baay'ee beeka.
ቢጫው መኪና የሚሸጥ ነው ።	Konkoolaataan booran kan gurguramuu dha.
ገላን እግር ኳስ መጫወት አቆመ ።	Gallaan kubbaa miillaa taphachuu dhaabe.
የእኛ ትልቁ ቤት እየታደሰ ነው ።	Mani guddaan keenya ijaaramaa jira.
እናንተ ትላንት ወደ መቀሌ ሄዳቹ ።	Isin kaleessa gara Maqalee deemtan.
እነሱ ትላንት ወደ መቀሌ ሄዱ ።	Isaan kaleessa gara Maqalee deeman.
ቃልኪዳን ለጫላ መፅሐፍ ሰጠችው ።	Kaalkidaan Caalaaf kitaaba kenitte.
እኔ የቶክዮ ማራቶን እየሮጥኩ ነው ።	Ani maaraatoni Tokyoo fiigaa jira.
ሃና ትምህርት ቤት ውስጥ ነች ።	Haanaan mana barnoota keessa jirti.
እሷ ሻይ መጠጣት አትወድም ።	Isheen shaayii dhugu hin jaalatu.
ሃና መረብ ኳስ ትጫወታለች ።	Haanaan kubbaa saaphana taphatti.
እኛ እንጨት ቆረጥን ።	Nu'i muka murne.
የእንግሊዝ ጠቅላይ ሚኒስቴር ።	Muummeen ministiraa Briitaaniyaa.
እሷ ወንበር ላይ ተቀምጣለች ።	Isheen barcuma irraa ta'a jirti.
እሱ ትላንት ወደ ሐረር ሄደ ።	Inni kaleessa gara Harar deeme.
እኛ ነገ ወደ መቀሌ እንሄዳለን ።	Nu'i bor gara Maqalee nideemna.
እሷ ማራቶን እየሮጠች ነው ።	Isheen maaraatoni fiigaa jirti.
የድሬዳዋ ከተማ አስተዳደር ።	Bulchiinsa magaalaa Diredaawaa.

ቃልኪዳን እና ጫላ ተገናኙ ።	Kaalkidaan fi Caalaan walargan.
ዮሃንስ ሻይ እየጠጣ ነው ።	Yohaannis shaayee dhugaa jira.
እኔ የገዛውት ዶሮ ሞተ ።	Ani kan bitte handaaqqoo du'e.
እነሱ ማራቶን እየሮጡ ነው ።	Isaan maaraatoni fiigaa jiran.
እነሱ ትላንት መቀሌ ሄዱ ።	Isaan kaleessa Maqalee deeman.
እሱ ትልቅ የድንጋይ ቤት ሰራ ።	Inni mana dagaa guddaa ijaare.
እሱ የገብስ ጠላ አይጠጣም ።	Inni farsoo garbuu hindhuguu.
እኔ ነገ ወደ መቀሌ እሄዳለው ።	Ani bor gara Maqalee nideema.
እሷ ትላንትና ስትሮጥ ነበር ።	Isheen kaleessa fiigaa turte.
እሷ አዲስ አበባ ልትሄድ ነው ።	Isheen Addis Ababaa deemufi.
የእኔ ትንሹ ቤት አሮጌ ነው ።	Mani xiqqaan koo moofaa dha.
እሷ የገብስ ጠላ ጠጣች ።	Isheen farsoo garbuu dhugte.
አብዲ ትልቅ ቤት ሰራ ።	Abdiin mana guddaa ijaare.
እሱ ትልቅ ቤት ሊሰራ ነው ።	Inni mana guddaa ijaarufi.
እኔ ቡና መጠጣት አልወድም ።	Ani buna dhugu hinjaaladhu.
እናንተ የገዛቹት በግ ሞተ ።	Isin kanbitan hoolaa du'e.
እሷ ሶስት ላሞች አሏት ።	Isheen saawwan sadi qabdi.
እሱ ትላንት ሲሮጥ ነበር ።	Inni kaleessa fiigaa ture.
እኔ ሶስት ኪሎ ቡና ገዛው ።	Ani kiiloo sadi buna bite.
እኛ መኪና ልንገዛ ነው ።	Nu'i konkoolaataa bituufi.
እሷ አራት ላሞች ገዛች ።	Isheen saawwan afur bitte.
ይህ የጫልቱ ቤት ነው ።	Kuni mana Caaltuu dha.
እሷ አልጋ ላይ ተኛች ።	Isheen siree irraa rafte.
እነሱ አንድ ከብት አላቸው ።	Isaan sangaa tokko qaban.
እሱ መኪና ሊገዛ ነው ።	Inni konkoolaataa bitufi.
አስቴር ቡና እየጠጣች ነው ።	Asteer buna dhugaa jirti.
ቦንቱ ዶሮ ገዛች ።	Boontun handaaqqoo bitte.
እኔ መኪና የለኝም ።	Ani konkoolaataa hinqabu.
እነሱ አንድ በግ አላቸው ።	Isaan hoolaa tokko qaban.
እሷ ሐሙስ ትመረቃለች ።	Isheen kamisa eebbifamti.
እነሱ ቤት ውስጥ ናቸው ።	Isaan mana keessa jiran.
እሷ መፅሐፍ አነበበች ።	Isheen kitaaba dubbifte.
ሃይማኖት መኪና ትወዳለች ።	Haaymaanoot konkoolaataa jaalati.
እኛ ሁለት በጎች ገዛን ።	Nu'i hoolotta lama bine.
ስሟ ሜላት ነው ።	Maqaan ishee Melaat dha.
ጫልቱ አስተማሪ ነች ።	Caaltuun barsiistuu dha.
ሃና ኪስ ትጫወታለች ።	Haanaan kubbaa taphatti.
እሱ በፍጥነት እየነዳ ነው ።	Inni ariitin fiiga jira.

Annex III: Sample language model for Amharic

\data\

ngram 1= 1545

ngram 2= 3511

ngram 3= 650

\1-grams:

-3.72387 <s> -0.643701

-3.42284 የአላጌ -0.270106

-3.42284 ቴክኒክና -0.270106

-3.42284 ሙያ -0.270106

-3.42284 ኮሌጅ -0.166498

-3.54778 መምህራንና -0.193245

-3.32593 ሰራተኞች -0.153718

-3.54778 በኢትዮጵያ -0.11749

-3.72387 ተደርጎ -0.11749

-3.72387 በነበረው -0.11749

-3.42284 ሀገር -0.11749

-3.54778 አቀፍ -0.11749

-3.72387 ምርጫ -0.11749

-2.38144 ላይ -0.258859

-3.72387 ተሳትፈን -0.11749

-3.54778 አናውቅም -0.11749

-3.54778 አሉ -0.193245

-0.952646 ። -2.69646

-0.952646 </s> -2.86487

-3.1798 የውጭ -0.166498

-3.07065 ጉዳይ -0.234462

-3.54778 መመሪያ -0.11749

-3.72387 ድርሻ -0.11749

-3.32593 ኢትዮጵያ -0.11749

-3.32593 የዓለም -0.153718

-3.72387 ሁኔታ -0.11749

-3.72387 በመረዳት -0.11749

-3.72387 በሃሳብ -0.11749

-2.53353 እና -0.136435

-3.72387 በተግባር -0.11749

-3.72387 የተረጋገጠ -0.11749

-3.72387 ለሁሉም -0.11749

-3.72387 መፍጠር -0.11749

-3.72387 እንደሚያስፈልግ -0.11749

-3.24675 ተቀመጠ -0.491955

-3.72387 እውነት -0.11749

-3.72387 እላቹኃለሁ -0.11749

-3.72387 በዚህ -0.11749

-3.72387 ከቆሙት -0.11749

-3.12181 ሰዎች -0.11749

-3.72387 የእግዚአብሔርን-0.11749

-3.1798 መንግስት -0.166498

-3.72387 በጎይል -0.11749

-3.72387 ስትመጣ -0.11749

-3.72387 እስኪያዩት -0.11749

-3.54778 ድረስ -0.11749

-3.72387 ሞትን -0.11749

-3.72387 የማይቀመሱት -0.11749

-3.42284 ትናንት -0.11749

-3.72387 ምሽት -0.11749

-3.72387 በጣሊያን -0.11749

-3.72387 ሮም -0.11749

-2.8488 ከተማ -0.13782

-3.72387 በተካሄደ -0.11749

-3.54778 የዳይመንድ -0.193245

-3.54778 ሊግ -0.11749

-3.07065 ውድድር -0.135226

-3.42284 ኢትዮጵያውያን -0.166498

-3.42284 አትሌቶች -0.11749

-3.42284 በበላይነት -0.270106

-3.42284 አጠናቀዋል -0.270106

-3.54778 ሚሼል -0.193245

-3.54778 ፕላቲኒ -0.11749

-3.32593 ዋንጫ -0.153718

-2.74614 ኪስ -0.335255

-3.32593 የእግር -0.11749

-3.72387 የአለም -0.11749

-3.72387 ከ2022ቱ -0.11749

-2.72387 ጋር -0.213767

-3.72387 በተያያዘ -0.11749

-3.72387 ሙስና -0.11749

Annex IV: Sample language model for Afaan Oromo

\data\

ngram 1= 1539

ngram 2= 3569

ngram 3= 801

\1-grams:

-3.75259 <s> -0.693656

-3.20852 Barsiistooni -0.462829

-2.38152 fi -0.166393

-3.35465 hojetootnii -0.156635

-3.57649 koleejjii -0.188457

-3.45156 Teknikaa -0.356403

-3.45156 Ogummaa -0.356403

-3.45156 Allaagee -0.166986

-2.6734 , -0.132

-3.75259 Filannoo -0.126987

-3.45156 Biyyoolessaa -0.126987

-3.75259 Itoophiyaatti -0.126987

-3.75259 gaggeffamaa -0.126987

-2.93967 ture -0.763859

-2.35465 irraa -0.279177

-3.75259 hirmaannee -0.126987

-3.75259 hinbeeknu -0.126987

-3.57649 jedhu -0.188457

-0.981734 . -2.95079

-0.981366 </s> -2.95116

-3.57649 imaammata -0.126987

-3.57649 hariiroo -0.126987

-3.45156 dhimma -0.166986

-3.57649 alaa -0.126987

-2.97444 Ityoophiyaa -0.148347

-3.75259 haala -0.126987
-3.57649 addunyaa -0.126987
-3.75259 hubachuudhan -0.126987
-3.75259 yaadaa -0.126987
-3.75259 hojimaata -0.126987
-3.75259 qabatamaa -0.126987
-3.45156 irratti -0.126987
-3.75259 hundaa -0.126987
-1.65222 ' -0.775036
-2.62225 e -0.965745
-3.75259 uumuun -0.126987
-3.75259 barbaachisaa -0.126987
-2.60646 ta -1.31064
-3.57649 uun -0.126987
-3.27547 ka -0.578251
-3.75259 eera -0.126987
-3.75259 Dhuguma -0.126987
-3.75259 dhuguman -0.126987
-3.75259 isinitti -0.126987
-3.75259 hima -0.126987
-3.20852 namoota -0.146526
-3.57649 as -0.126987
-3.75259 dhaabatanii -0.126987
-2.87753 jiran -0.830806
-2.77486 keessa -0.30272
-3.75259 kaan -0.126987
-3.57649 utuu -0.126987
-2.22754 hin -0.325472
-2.79834 du -1.10973
-3.20852 in -0.126987
-3.27547 mootummaan -0.150541

-3.57649 Waaqayyoo -0.126987
-3.75259 humnaan -0.126987
-3.75259 dhufee -0.126987
-3.75259 arguug -0.126987
-3.05362 jiru -0.638921
-3.35465 dorgommii -0.156635
-3.57649 Diyaamand -0.188457
-3.57649 Liigii -0.126987
-2.26122 kaleessa -0.428132
-3.27547 galgala -0.249741
-3.45156 Xaaliyaan -0.126987
-2.90749 magaalaa -0.144991
-3.75259 Roomitti -0.126987
-3.75259 gaggeeffameen -0.126987
-3.45156 atleetonni -0.356403
-3.45156 olaantummaadhan -0.356403
-3.45156 xumuraniiru -0.356403
-3.57649 Miishal -0.188457
-3.57649 Plaatiiniin -0.126987
-3.35465 waancaa -0.126987
-2.97444 kubbaa -0.293032
-3.27547 miillaa -0.126987
-3.75259 2022n -0.126987
-3.75259 walqabatee -0.126987
-3.75259 malaammaltummaadhan -0.126987
-3.75259 shakkamaniiti -0.126987
-3.27547 to -0.578251
-3.75259 annaa -0.126987
-3.27547 jala -0.383648
-2.62225 kan -0.252327
-3.75259 oolfaman -0.126987

Annex V: Transliteration from Amharic alphabets to Latin characters

First Order		Second Order		Third Order		Fourth Order		Fifth Order		Sixth Order		Seventh Order				
ሀ	*hä*	ሁ	*hu*	ሂ	*hi*	ሃ	*ha*	ሄ	*he*	ህ	*h*	ሆ	*ho*			
ለ	*lä*	ሉ	*lu*	ሊ	*li*	ላ	*la*	ሌ	*le*	ል	*l*	ሎ	*lo*		ሏ	*lWa*
ሐ	*Hä*	ሑ	*Hu*	ሒ	*Hi*	ሓ	*Ha*	ሔ	*He*	ሕ	*H*	ሖ	*Ho*		ሗ	*HWa*
መ	*mä*	ሙ	*mu*	ሚ	*mi*	ማ	*ma*	ሜ	*me*	ም	*m*	ሞ	*mo*		ሟ	*mWa*
ሠ	*Sä*	ሡ	*Su*	ሢ	*Si*	ሣ	*Sa*	ሤ	*Se*	ሥ	*S*	ሦ	*So*		ሧ	*SWa*
ረ	*rä*	ሩ	*ru*	ሪ	*ri*	ራ	*ra*	ሬ	*re*	ር	*r*	ሮ	*ro*		ሯ	*rWa*
ሰ	*sä*	ሱ	*su*	ሲ	*si*	ሳ	*sa*	ሴ	*se*	ስ	*s*	ሶ	*so*		ሷ	*sWa*
ሸ	*śä*	ሹ	*śu*	ሺ	*śi*	ሻ	*śa*	ሼ	*śe*	ሽ	*ś*	ሾ	*śo*		ሿ	*śWa*
ቀ	*qä*	ቁ	*qu*	ቂ	*qi*	ቃ	*qa*	ቄ	*qe*	ቅ	*q*	ቆ	*qo*		ቋ	*qWa*
በ	*bä*	ቡ	*bu*	ቢ	*bi*	ባ	*ba*	ቤ	*be*	ብ	*b*	ቦ	*bo*		ቧ	*bWa*
ቨ	*vä*	ቩ	*vu*	ቪ	*vi*	ቫ	*va*	ቬ	*ve*	ቭ	*v*	ቮ	*vo*		ቯ	*vWa*
ተ	*tä*	ቱ	*tu*	ቲ	*ti*	ታ	*ta*	ቴ	*te*	ት	*t*	ቶ	*to*		ቷ	*tWa*
ቸ	*cä*	ቹ	*cu*	ቺ	*ci*	ቻ	*ca*	ቼ	*ce*	ች	*c*	ቾ	*co*		ቿ	*cWa*
ኀ	*Ĥä*	ኁ	*Ĥu*	ኂ	*Ĥi*	ኃ	*Ĥa*	ኄ	*Ĥe*	ኅ	*Ĥ*	ኆ	*Ĥo*		ኇ	*ĤWa*
ነ	*nä*	ኑ	*nu*	ኒ	*ni*	ና	*na*	ኔ	*ne*	ን	*n*	ኖ	*no*		ኗ	*nWa*
ኘ	*Nä*	ኙ	*Nu*	ኚ	*Ni*	ኛ	*Na*	ኜ	*Ne*	ኝ	*N*	ኞ	*No*		ኟ	*NWa*
አ	*xä*	ኡ	*xu*	ኢ	*xi*	ኣ	*xa*	ኤ	*xe*	እ	*x*	ኦ	*xo*		ኧ	*xWa*
ከ	*kä*	ኩ	*ku*	ኪ	*ki*	ካ	*ka*	ኬ	*ke*	ክ	*k*	ኮ	*ko*		ኳ	*kWa*
ኸ	*Kä*	ኹ	*Ku*	ኺ	*Ki*	ኻ	*Ka*	ኼ	*Ke*	ኽ	*K*	ኾ	*Ko*		ኵ	*KWa*
ወ	*wä*	ዉ	*wu*	ዊ	*wi*	ዋ	*wa*	ዌ	*we*	ው	*w*	ዎ	*wo*			
ዐ	*Xä*	ዑ	*Xu*	ዒ	*Xi*	ዓ	*Xa*	ዔ	*Xe*	ዕ	*X*	ዖ	*Xo*			
ዘ	*zä*	ዙ	*zu*	ዚ	*zi*	ዛ	*za*	ዜ	*ze*	ዝ	*z*	ዞ	*zo*		ዟ	*zWa*
ዠ	*Zä*	ዡ	*Zu*	ዢ	*Zi*	ዣ	*Za*	ዤ	*Ze*	ዥ	*Z*	ዦ	*Zo*		ዧ	*ZWa*
የ	*yä*	ዩ	*yu*	ዪ	*yi*	ያ	*ya*	ዬ	*ye*	ይ	*y*	ዮ	*yo*			
ደ	*dä*	ዱ	*du*	ዲ	*di*	ዳ	*da*	ዴ	*de*	ድ	*d*	ዶ	*do*		ዷ	*dWa*
ጀ	*jä*	ጁ	*ju*	ጂ	*ji*	ጃ	*ja*	ጄ	*je*	ጅ	*j*	ጆ	*jo*		ጇ	*jWa*
ገ	*gä*	ጉ	*gu*	ጊ	*gi*	ጋ	*ga*	ጌ	*ge*	ግ	*g*	ጎ	*go*		ጓ	*gWa*
ጠ	*Tä*	ጡ	*Tu*	ጢ	*Ti*	ጣ	*Ta*	ጤ	*Te*	ጥ	*T*	ጦ	*To*		ጧ	*TWa*
ጨ	*Cä*	ጩ	*Cu*	ጪ	*Ci*	ጫ	*Ca*	ጬ	*Ce*	ጭ	*C*	ጮ	*Co*		ጯ	*CWa*
ጰ	*Pä*	ጱ	*Pu*	ጲ	*Pi*	ጳ	*Pa*	ጴ	*Pe*	ጵ	*P*	ጶ	*Po*		ጷ	*PWa*
ጸ	*ťä*	ጹ	*ťu*	ጺ	*ťi*	ጻ	*ťa*	ጼ	*ťe*	ጽ	*ť*	ጾ	*ťo*		ጿ	*ťWa*
ፀ	*Ťä*	ፁ	*Ťu*	ፂ	*Ťi*	ፃ	*Ťa*	ፄ	*Ťe*	ፅ	*Ť*	ፆ	*Ťo*			
ፈ	*fä*	ፉ	*fu*	ፊ	*fi*	ፋ	*fa*	ፌ	*fe*	ፍ	*f*	ፎ	*fo*		ፏ	*fWa*
ፐ	*pä*	ፑ	*pu*	ፒ	*pi*	ፓ	*pa*	ፔ	*pe*	ፕ	*p*	ፖ	*po*		ፗ	*pWa*

Declaration

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

Declared by:

Name: _____.

Signature: _____.

Date: _____.

Confirmed by advisor:

Name: _____.

Signature: _____.

Date: _____.