
Addis Ababa University

School of Graduate Studies

College of Natural Sciences

Department of Computer Science

# Bidirectional English – Afaan Oromo Machine Translation Using Hybrid Approach

Jabesa Daba

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Science

Addis Ababa, Ethiopia

November 2013

Addis Ababa University

School of Graduate Studies

College of Natural Sciences

Department of Computer Science

# Bidirectional English – Afaan Oromo Machine Translation Using Hybrid Approach

Jabesa Daba

Signature of the Board of Examiners for Approval

| Name | Signature |
|---|---|
| 1. Dr.Yaregal Assabie, Advisor | _____ |
| 2. Dr. Dida Midekso, Examiner | _____ |
| 3. _____ | _____ |

**Dedicated to:**

1. **Zewuditu Terefe (Mother)**
2. **Lalise Daba (Sister)**

# Acknowledgements

Above all I would like to thank the almighty God, who gave me the opportunity and strength to achieve whatever I have achieved so far. I would like to express my gratitude to all the people who supported and accompanied me during the progress of this thesis.

First, I would like to express my deep-felt gratitude to my advisor, Dr Yaregal Assabie, whose excellent and enduring support shaped this work considerably and made the process of creating this thesis an invaluable learning experience.

I want to thank Ato. Nega Gerbaba who helped me in the process of data collection and evaluation of Afaan Oromo POS tagging.

I also want to thank W/t. Eleni Teshome for her kind support on installation of the tools we have used in this work.

Special thanks to my family; their endless motivation and unconditional love have been influential in whatever I have achieved so far.

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

# Table of Contents

## List of Figures

## List of Tables

## List of Algorithms

## Acronyms and Abbreviations

BLEU: Bilingual Evaluation Understudy

BP: Brevity Penalty

CLIR: Cross Language Information Retrieval

EASMT: English Amharic Statistical Machine Translation

EB: Example Based

EBMT: Example Based Machine Translation

FDRE: Federal Democratic Republic of Ethiopia

HMT: Hybrid Machine Translation

HMM: Hidden Markov Model

IBM: International Business Machines

LDC: Linguistic Data Consortium

LM: Language Model

MT: Machine Translation

NIST: National Institute of Standards and Technology

NLP: Natural Language Processing

POS: Part Of Speech

RBMT: Rule Based Machine Translation

SMT: Statistical Machine Translation

SOV: Subject – Object – Verb

SVO: Subject – Verb – Object

# Abstract

Machine translation is one of the applications of natural language processing that studies the use of computer programs and software to translate one natural language into another in the form of text or speech. Since there is a need for translation of documents between English and Afaan Oromo languages there needs to be a mechanism to do so. Thus, this study resulted in the development of a bidirectional English-Afaan Oromo machine translation system using a hybrid approach. The research work is implemented using a hybrid of rule based and statistical approaches. Since English and Afaan Oromo have different sentence structures, we implement syntactic reordering approach which makes the structure of source sentences to be similar to the structure of target sentences. So, reordering rules are developed for simple, interrogative and complex English and Afaan Oromo sentences. In order to achieve the objective of this research work, a corpus is collected from different domain and prepared in a format suitable for use in the development process and classified as training set and test set. The reordering rules are applied on both the training and test sets in a preprocessing step. Since the system is bidirectional, two language models are developed; one for English and the other for Afaan Oromo. Translation models which assign a probability that a given source language text generates a target language text are built and a decoder which searches for the shortest path is used. Two major experiments are conducted by using two different approaches and their results are recorded. The first experiment is carried out by using a statistical approach. The result obtained from the experiment has a BLEU score of 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation. The second experiment is carried out by using a hybrid approach and the result obtained has a BLEU score of 37.41% for English to Afaan Oromo translation and 52.02% for Afaan Oromo to English translation. From the result, we can see that the hybrid approach is better than the statistical approach for the language pair and a better translation is acquired when Afaan Oromo is used as a source language and English is used as a target language.



**Key words**: Machine Translation, Statistical Machine Translation, Hybrid Machine Translation, Reordering rule

# CHAPTER ONE

# Introduction

## 1.1 Background

Natural Language Processing (NLP) is an interdisciplinary research area at the border between linguistics and artificial intelligence aiming at developing computer programs capable of human-like activities related to understanding or producing texts or speech in a natural language [1]. It is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks [2].

Applications of NLP include a number of fields of studies, such as machine translation, morphology, syntax, named entity recognition, natural language text processing and summarization, multilingual and cross language information retrieval (CLIR), speech recognition, information retrieval and text clustering, and so on [2]. Among these applications, machine translation (MT) refers to computerized systems responsible for the production of translations with or without human assistance [3]. MT is aimed to enable a computer to transfer natural language expressions in either text or speech from one natural language into another while preserving the meaning and interpretation. Large number of languages exist in the world which reflects the linguistic diversity. It is difficult for an individual to know and understand all the languages of the world. Hence, the methodology of translation was adopted to communicate the messages from one language to another. Accessibility to other language documents and information has always been a concern for most persons. In order to effectively use these documents and information which is written in another language, there must be a mechanism of translating the documents and information into the language which is understandable by the user.

## 1.2 Statement of the Problem

Machine translation systems for different language pairs have been developed by using different methodologies and approaches used in the area of study. Most of the studies have been done on

language pairs of English and other foreign language such as Spanish [4], Chinese [5], Arabic [6], French [7] and most of major languages spoken in India [8] and some of the studies are done on foreign language pairs which are morphologically related to each other [9].

However, there has been a little work on machine translation for languages which are spoken in Ethiopia. Some of the studies are carried out on English - Amharic language pair [10, 11, 12] and there is one experiment on English - Afaan Oromo language pair [13]. The experiment which was conducted on English – Afaan Oromo language pair is done by using statistical methodology. The experiment was intended to translate English sentences into Afaan Oromo only in one direction which is still not a solution for the need of Afaan Oromo to English translation. The other thing is that the accuracy of the translation which has a BLEU score of 17.74% is not satisfactory. Therefore, this study is aimed to solve these problems by developing a machine translation system which can translate from English to Afaan Oromo and from Afaan Oromo to English in both directions by using hybrid approach.

## 1.3 Objectives of the Study

### 1.3.1 General Objective

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

### 1.3.2 Specific Objectives

The specific objectives of this research work are to:

- Review techniques and methodologies used for machine translation.
- Study syntactic structure and relationship of the language pair; English and Afaan Oromo.
- Collect English – Afaan Oromo bilingual parallel corpus.
- Develop a general architecture for bidirectional English – Afaan Oromo machine translation using hybrid approach.
- Develop an algorithm for the machine translation system.
- Develop a prototype for bidirectional English – Afaan Oromo machine translation using hybrid approach.
- Test and evaluate the performance of the system.

## 1.4 Methodologies

### 1.4.1 Literature Review

Literature review on machine translation systems done for other language pairs have been reviewed in order to understand how machine translation works for English to foreign language translation and foreign language to English translation. In addition to this, both English and Afaan Oromo grammar books have been studied for further understanding of both the languages‟ syntax for the linguistic knowledge acquiring.

### 1.4.2 Data Collection

We have used English – Afaan Oromo parallel corpus collected from different domain including the Holy Bible, the Constitution of FDRE, the Criminal Code of FDRE, international conventions, Megeleta Oromia and a bulletin from Oromia health bureau. We have also used a monolingual Afaan Oromo and English corpus collected from the web.

### 1.4.3 Development Tools

Since we have developed the bidirectional English – Afaan Oromo machine translation system by using hybrid approach which is a combination of rule based and statistical approach, we have used freely available software such as IRSTLM toolkit, GIZA++, and Moses for the statistical part and we have used Python programming language for the rule part.

### 1.4.4 Evaluation

Machine translation systems can be evaluated by using human evaluation method or automatic evaluation method. Since human evaluation is not accurate and not efficient with respect to automatic evaluation, we have used BLEU score to evaluate the performance of the system, which is an automatic evaluation technique.

## 1.5 Scope and Limitations of the Study

### 1.5.1 Scope of the study

The bidirectional English – Afaan Oromo machine translation using hybrid approach is designed to translate a sentence written in English text into Afaan Oromo text and vice versa. Therefore, speech translation is not included in the study.

### 1.5.2 Limitations of the study

The main limitation while conducting this research work is the absence of publically available automatic Afaan Oromo part-of-speech tagger. Therefore, Afaan Oromo sentences are tagged manually which took much time and effort. The other limitation is that there is a scarcity and impurity of English – Afaan Oromo parallel corpus.

## 1.6 Applications of the Study

The following are the main contributions 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 CLIR for English – Afaan Oromo language pair.
- The translation of different reading materials can easily be accomplished for English – Afaan Oromo language pair.
- The translation system can be used as a tool in teaching and learning process of the languages.
- The study can be used as additional component for studies in speech to text, text to speech and speech to speech translation regarding English – Afaan Oromo language pair.

## 1.7 Thesis Organization

This thesis paper is organized into Six Chapters including the current one. Chapter Two presents an overview of Afaan Oromo language and a literature review on machine translation. Chapter Three presents different related works on machine translation. Chapter Four presents design of bidirectional English – Afaan Oromo machine translation using hybrid approach. The experiments and results are discussed in Chapter Five and the conclusion and future works are presented in Chapter Six.

# CHAPTER TWO

# Literature Review

## 2.1 Introduction

In this Chapter, a brief overview on Afaan Oromo language and machine translation (MT) is provided. The Chapter briefly describes the state of the art of machine translation which includes statistical machine translation (SMT), rule based machine translation (RBMT), example based machine translation (EBMT) and hybrid machine translation (HMT).

## 2.2 Overview of Afaan Oromo

Afaan Oromo is one of the languages of the Low land East Cushitic within the Cushitic family of the Afro-Asiatic Phylum [14, 15]. It is also one of the major languages spoken in Ethiopia. According to [16] and [17], Afaan Oromo is the third most widely spoken language in Africa after Arabic and Hausa.

Like other African and Ethiopian languages, Afaan Oromo has a very rich morphology [18]. Latin based alphabet known as Qubee has been adopted and became an official script of Afaan Oromo since 1991 [19]. The language is widely used in Ethiopia and neighboring countries like Kenya and Somalia [20]. Currently, Afaan Oromo is an official language of Oromia Regional State and used as an instructional media for primary and junior secondary schools of the region. Even if the language is spoken by large number of the population, the number of literature works, newspapers, magazines, educational resources, official documents and religious writings written and published in this language are few in number.

### 2.2.1 Afaan Oromo Writing System

Afaan Oromo uses Latin based alphabet known as Qubee that consists of twenty-eight basic letters. From twenty eight basic letters, five of them are known as vowels, the other five letters are known as double consonants (Qubee dachaa) and the rest are known as consonants. Double consonant letters are derived from a combination of two consonant letters. Qubee is characterized by capital and small letters which is known in the English alphabet. Similar to

English language, vowels are sound makers and are sounds by themselves. Vowels in Afaan Oromo are characterized as short and long vowels.

The basic alphabet in Afaan Oromo does not contain „p", "v" and „z". This is because there are no native words in Afaan Oromo that are formed from these characters. However, in writing Afaan Oromo language they are used to refer to foreign words such as "Paappaayyaa" ("Papaya"). The complete list of Afaan Oromo alphabet is listed in Table 2.1.

Table 2.1. Afaan Oromo alphabet [21]

| Number | Capital | Small | Type | Long | Short |
|---|---|---|---|---|---|
| 1 | A | A | Vowel | Aa | A |
| 2 | B | B | Consonant | - | - |
| 3 | C | C | Consonant | - | - |
| 4 | D | D | Consonant | - | - |
| 5 | E | E | Vowel | Ee | E |
| 6 | F | F | Consonant | - | - |
| 7 | G | G | Consonant | - | - |
| 8 | H | H | Consonant | - | - |
| 9 | I | I | Vowel | Ii | I |
| 10 | J | J | Consonant | - | - |
| 11 | K | K | Consonant | - | - |
| 12 | L | L | Consonant | - | - |
| 13 | M | M | Consonant | - | - |
| 14 | N | N | Consonant | - | - |
| 15 | O | O | Vowel | Oo | O |
| 16 | P | P | Consonant | - | - |
| 17 | Q | Q | Consonant | - | - |
| 18 | R | R | Consonant | - | - |
| 19 | S | S | Consonant | - | - |
| 20 | T | T | Consonant | - | - |
| 21 | U | U | Vowel | Uu | U |
| 22 | V | V | Consonant | - | - |
| 23 | W | W | Consonant | - | - |
| 24 | X | X | Consonant | - | - |
| 25 | Y | Y | Consonant | - | - |
| 26 | Z | Z | Consonant | - | - |
| 27 | CH | Ch | Double consonant | - | - |
| 28 | DH | Dh | Double consonant | - | - |
| 29 | NY | Ny | Double consonant | - | - |
| 30 | PH | Ph | Double consonant | - | - |
| 31 | SH | Sh | Double consonant | - | - |

### 2.2.2 Afaan Oromo Sentence Structure

Afaan Oromo and English have differences in their syntactic structure. In Afaan Oromo, the sentence structure is subject-object-verb (SOV). Subject-object-verb (SOV) is a sentence structure where the subject comes first, then the object and the verb next to the object. For example, if we take Afaan Oromo sentence “Dagaagaan nyaata nyaate”, “Dagaagaan” is the subject, “nyaata” is the object and “nyaate” is the verb of the sentence.

In case of English, the sentence structure is subject-verb-object. For example, if the above Afaan Oromo sentence is translated into English it will be “Degaga ate food” where “Degaga” is the subject, “ate” is the verb and “food” is the object.

### 2.2.3 Articles

English language uses two types of articles known as definite article (the) and indefinite article (a, an, some, any). In case of Afaan Oromo, there are no articles that will be inserted before nouns unlike that of English rather the last vowel of the noun is dropped and suffixes (-icha, -ittii, -attii, -utti) are added to show definiteness instead of using definite articles.

### 2.2.4 Punctuation Marks

Other than apostrophe, punctuation marks used in both Afaan Oromo and English languages are the same and used for the same purpose. Apostrophe mark („) in English shows possession, but in Afaan Oromo it is used in writing to represent a glitch sound known as hudhaa. It plays an important role in Afaan Oromo reading and writing system. For example, it is used to write a word in which most of the time two vowels appear together like “kaa'uu”.

## 2.3 Machine Translation

Machine translation refers to computerized systems which are used for translation of natural languages (such as English and Afaan Oromo) with the help of human or without. Machine translation is aimed to enable a computer to transfer natural language expressions in either text or speech from one natural language (source language) into another (target language) while preserving the meaning and interpretation. A translation can be human-aided machine translation or it can be machine-aided human translation [22]. In case of machine-aided human translation,

the translation is performed by human translators with the help of computer-based translation tools. The translation tools help the human translator by providing access to resources which are used for translation such as on-line dictionaries and also by performing transmission and reception of texts and storing previously translated texts. In the case of human-aided machine translation, the translation process is performed by computer with the help of human. Humans are involved before the translation process which is called pre-editing or it can be after the translation process which is called post-editing. According to [23], MT can be viewed as a system that builds a representation of the same content in the form of different languages. It is based on the idea that the same content can be expressed by different languages. Ideally, machine translation is a batch process which is applied to a given text which produces a perfect translated text which then only needs to be printed out [24].

MT systems can be sub-language MT or it can be general purpose [22]. Sub-language MT systems are designed particularly for some specific domain for some specialized purposes. The specialized language is referred to as a sub-language. A sub-language is used by experts in certain fields of area for communication purpose. It contains words which are only known by those experts of that specific field of study or words which can be used in different ways. Sub-languages are also characterized by special grammatical patterns. The general purpose MT systems are designed for translation of texts and speech from the entire domain without any domain restriction.

MT systems can be bilingual systems or multilingual systems depending on the number of languages involved in the translation process [22]. Bilingual systems are designed specifically for two languages (single pair of languages) and multilingual systems are designed for more than two languages. The translation can be unidirectional or bidirectional [22]. In case of unidirectional, the system translates from the source language into the target language only in one direction. Bidirectional systems work in both directions in a way that one language can act as source language or a target language. Bilingual systems can be unidirectional or they can be bidirectional, but multilingual systems are usually designed to be bidirectional.

Machine translation has its own advantage in allowing communication between users who speak different languages which advances globalization of the information highway [24]. Since most

translation systems work in an online environment, a fast and speedy communication can be made between people who speak different languages, leave in different environments and locations which will contribute to the growth of information technology. In the increasing use of the Internet, international information traffic is assumed to be dominated by the English language which will increase the need for translation service between English and other languages.

## 2.4 Approaches to Machine Translation

Different methods of machine translation are being used by different researchers, and the basic approaches and methodologies according to [25] are: rule based machine translation (RBMT), statistical machine translation (SMT), example based machine translation (EBMT) and hybrid machine translation (HMT). The detailed descriptions of the methodologies are presented in the following sections.

## 2.5 Statistical Machine Translation

The idea of using statistical methods for machine translations started by a team of scientists at IBM 40 years after the first idea of machine translation by Waren W. [26]. Arnold D. etal [23] state that statistical machine translation (SMT) is a machine translation system which is based on the idea that there is a possibility of every target sentence to be a translation of the source sentence.

SMT is a machine translation approach that uses human produced translations known as parallel corpus. According to [27], the translation process by using SMT is considered as a machine learning problem. After examining the parallel corpus, SMT algorithms automatically learn how to translate new sentences. The algorithms are machine learning algorithms which learn how to translate new sentences from the parallel corpus which is a collection of previously translated texts. The translation accuracy of these systems mainly depends on the parallel corpus regarding its domain, quantity and quality. So, in order to have a good translation quality, the data must be preprocessed consistently.

SMT is a MT approach that builds probabilistic models of faithfulness and fluency, so that the most probable translation can be selected by combining the models [22]. SMT focuses on the result of translation rather than the process. So, true translation, which is both faithful to the

source language and natural as an expression in the target language, is sometimes impossible. If you are going to go ahead and produce a translation anyway, you have to compromise. This is exactly what translators do in practice: they produce translations that do tolerably well on both criteria.

Depending on this idea, the goal of translation can be modeled as the production of an output that maximizes some value function that represents the importance of both faithfulness and fluency. SMT is the name for a class of approaches that do just this, by building probabilistic models of faithfulness and fluency, and then combining these models to choose the most probable translation. If we chose the product of faithfulness and fluency as our quality metric, we could model the translation from a source language sentence S to a target language sentence T as [22]:

$$\text{best-translation T} = argmax_T \text{ faithfulness(T,S) fluency(T)} \quad (2.1)$$

Where $argmax_T$ = a function that maximizes the product of faithfulness and fluency

faithfulness (T,S) = a target language faithful to the source language

fluency (T) = similarity of the target language to the source language

The above equation is similar to the Bayesian noisy channel model. This equation can be formalized for SMT. So, in order to translate from a foreign language sentence F = f1, f2, ..., fm to English, in a probabilistic model, the best English sentence E = e1,e2, ...,el is the one whose probability P(E|F) is the highest. Therefore by rewriting the noisy channel model via Bayes rule another equation can be derived [22]:

$$\begin{aligned} E &= argmax_E \, P(E|F) \\ &= argmax_E \, \frac{P(F|E)P(E)}{P(F)} \\ E &= argmax_E \, P(F|E)P(E) \end{aligned} \quad (2.2)$$

Where *P (E|F)* = The translation model for foreign to English translation

*P(F|E)* = The translation model for English to foreign translation

*P(E)* = language model for English

The denominator *P(F)* can be ignored inside the argmax , because the aim is to choose the best English sentence for a fixed foreign sentence F, and hence *P(F)* is a constant. The resulting noisy

channel equation shows that two components are needed. These are a translation model *P(F|E),* and a language model *P(E).*

Applying the noisy channel model to machine translation requires thinking of things backwards. It needs to pretend that the foreign (source language) input F must be translated in a corrupted version of some English (target language) sentence E, and that the task is to discover the hidden (target language) sentence E that generates the observation sentence F. The noisy channel model of statistical MT thus requires three components to translate from a foreign sentence *F* to an English sentence E [22]:

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

### 2.5.1 Language Model: N-Gram Language Model

The task of language model is to focus exclusively on the form of the target language sentence, irrespective of the manner in which that hypothesis was constructed from the input. A good language model should be able to clearly discriminate between grammatical and ungrammatical words in a language [28].

SMT systems are based on the same N-gram language models as speech recognition and other applications [22]. The task of predicting the next word can be stated as attempting to estimate the probability function P [29]:

$$P(W_n|W_1, \dots, W_{n-1}) \quad (2.3)$$

In such a statistical problem, it is good to use a classification of previous words, the history to predict the next word. On the basis of having looked at a lot of text, it is easy to know which words tend to follow other words. For this task, each textual history cannot be considered separately: most of the time we will be listening to a sentence that we have never heard before, and so there is no previous identical textual history on which to base our predictions, and even if we had heard the beginning of the sentence before, it might end differently this time. And so we

need a method of grouping histories that are similar in some way to give reasonable predictions as to which words we can expect next.

Estimators like N-grams that assign a conditional probability to possible next words can be used to assign a joint probability to an entire sentence [22]. Whether estimating probabilities of next words or of whole sequences, the N-gram model is one of the most important tools in speech and language processing. N-grams are essential in any task in which we have to identify words in noisy, ambiguous input. N-gram models are also essential in SMT. An N-gram will choose the most fluent translation sentence, i.e. the one that has the highest probability. N-grams are also crucial in NLP tasks like part-of-speech tagging, natural language generation, and word similarity. The language model component is monolingual, and so acquiring training data is relatively easy.

### 2.5.2 Translation Model

Statistical methods are applied to generate translated version using bilingual corpora. This methodology uses different kinds of translation models [26]:

- Statistical word-based translation model
- Statistical phrase-based model
- Statistical syntax-based model

#### 2.5.2.1 Statistical Word-Based Translation Model

Statistical machine translation is based on the idea that every target sentence is a possible translation of every source sentence. A reasonable assumption is that the probability of a given target sentence being a good choice for translation relies heavily on which source sentence is under consideration for translation. It is therefore possible to condition the probability on the source sentence, yielding the posterior probability of the target sentence given the source sentence P(t|s) [26], where t=target; s=source sentence.

This model of translation correlates with an intuitive view of translation; given a source sentence that we want to translate, which is the best target sentence to choose as its translation. That is, the target sentence that leads to the highest value for P(t|s). Since the translation model is modelled as the reverse process due to the Bayesian inversion, the following description of the translation

model will appear as a description of how a target sentence is translated into the source language. Peter F. Brown etal [30] change the problem of modelling translation through the problem of determining all possible word alignments between two sentences:

$$P(s|t) = \sum_a P(s, a|t) \quad (2.4)$$

Where P(s|t) = probability of the source sentence given the target sentence

$P(s, a|t)$ = probability of word alignment of source sentence to the target sentence

#### 2.5.2.2 The Phrase-Based Translation Model

The job of the translation model, given an English sentence E and a foreign sentence F is to assign a probability that E generates F [22]. Modern SMT is based on the perception that a good way to compute these probabilities is by considering the behavior of phrases. The idea of phrase-based SMT is to use phrases (sequences of words) instead of single words as the fundamental units of translation.

The generative story of phrase-based translation has three steps. The first step is to group the English source words into phrases *e1, e2...eI*. The next step is to translate each English phrase *ei* into a foreign phrase fj. Finally, each of the foreign phrases is reordered. The probability model for phrase-based translation relies on a translation probability and a distortion probability. The factor $\phi(fj|ei)$ is the translation probability of generating foreign phrase *f j* from English phrase *ei* [22].

Reordering of foreign phrases is done by the distortion probability d. Distortion in SMT refers to a word having a different (distorted) position in the foreign sentence than it had in the English sentence; it is thus a measure of the distance between the positions of a phrase in the languages. The distortion probability in phrase-based MT means the probability of two consecutive English phrases being separated in foreign by a span (of foreign words) of a particular length. More formally, the distortion is parameterized by $d(a_i - b_{i-1})$, where $a_i$ is the start position of the foreign phrase generated by the *ith* English phrase ei, and $b_{i-1}$ is the end position of the foreign phrase generated by the *i−1th* English phrase *ei−1*. We can use a very simple distortion probability, in which we simply raise some small constant α to the distortion: $d(a_i - b_{i-1})$

$= \alpha^{|a_i - b_i - 1|}$. This distortion model penalizes large distortions by giving lower probability the larger the distortion.

The final translation model for phrase-based MT is [22]:

$$P(F|E) = \prod_{i=1}^{I} \phi(fi, ei) d(a_i - b_{i-1}) \quad (2.5)$$

Where $\phi(fi, ei)$ = the translation probability of generating foreign phrase *fi* from English phrase *ei*

$d(a_i - b_{i-1})$ = distortion probability

In order to use the phrase-based model, we need two more things. We need a model of decoding, so we can go from a surface foreign string to a hidden English string. And we need a model of training, so we can learn parameters.

The main set of parameters that needs to be trained is the set of phrase translation probabilities $\phi(fi|ei)$. These parameters, as well as the distortion constant $\alpha$, could be set if we had only a large bilingual training set, in which each foreign sentence was paired with an English sentence, and if furthermore we knew exactly which phrase in the foreign sentence was translated by which phrase in the English sentence. We call such a mapping a phrase alignment.

### 2.5.3 The Decoder

The job of the decoder is to take a foreign source sentence F and produce the best (English) translation E according to the product of the translation and language models. Finding the sentence which maximizes the translation and language model probabilities is a search problem, and decoding is thus a kind of search. Decoders in MT are based on best first search, a kind of heuristic or informed search [22]; these are search algorithms that are informed by knowledge from the problem domain. Best-first search algorithms select a node n in the search space to explore based on an evaluation function *f(n)*. MT decoders are variants of a specific kind of best-first search called **A*** search.

**A*** search was first implemented for machine translation by IBM [30]. The basic intuition is to maintain a priority queue which is traditionally referred to as a stack with all the partial translation hypotheses together with their scores. The job of a decoder is to find the highest scoring sentence in the target language (according to the translation model) corresponding to a

given source sentence. It is also possible for the decoder to output a ranked list of the translation candidates, and also to supply various types of information about how it came to its decision [32].

### 2.5.4 Alignment

Text alignment is not part of the translation process [29]; Instead text alignment is mostly used to create lexical resources such as bilingual dictionaries and parallel grammars, which then improve the quality of machine translation. All statistical translation models are based on the idea of a word alignment. A word alignment is a mapping between the source words and the target words in a set of parallel sentences [22]. For phrase-based SMT, the alignment algorithms are used just to find the best alignment for a sentence pair (F,E), in order to help extract a set of phrases. It is also possible to use these word alignment algorithms as a translation model *P(F,E)* as well. There are different alignment models used in SMT which are discussed in the following sections.

#### 2.5.4.1 IBM Model 1

IBM Model 1 is a word alignment model which is the first and simplest of five models proposed by IBM researchers [30]. The general IBM Model 1 generative story for how to generate a foreign sentence from an English sentence $E = e1,e2, ...,eI$ of length I is:

1. Choose a length K for the foreign sentence, henceforth $F = f_1, f_2, ..., f_K$.
2. Choose an alignment $A = a_1,a_2, ...,a_J$ between the English and foreign sentences.
3. For each position $j$ in the foreign sentence, choose a foreign word $f_j$ by translating the English word that is aligned to it.

An IBM Model 1 consists of finite set E of English words, a set F of foreign words, and integers M and L specifying the maximum length of Foreign and English sentences respectively. The parameters of the model are as follows [32]:

- t(f|e) for any f Ɛ F, e Ɛ E U {NULL}. The parameter t(f|e) can be interpreted as the conditional probability of generating foreign word f from English word e.

Given these definitions, for any English sentence e1 … el where each $e_j$ Ɛ E, for each length m, we define the conditional distribution over foreign sentences
f1 … fm and alignments a1 … am as [32]:

$$p(f_1 \dots \dots f_{m,}\, a_1 \dots \dots a_m | e_1 \dots \dots e_l, m) = \prod_{i=1}^{m} \frac{1}{(l+1)} \text{ x } t(fi|e_{ai}) = \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(fi|e_{ai}) \quad (2.6)$$

The parameters of IBM Model 1 can be estimated using the Expectation Maximization algorithm [22].

#### 2.5.4.2 IBM Model 2

IBM Model 2 model consists of finite set E of English words, a set F of foreign words, and integers M and L specifying the maximum length of foreign and English sentences respectively. The parameters of the model are as follows [32]:

- t(f|e) for any f Ɛ F, e Ɛ E U {NULL}. The parameter t(f|e) can be interpreted as the conditional probability of generating foreign word f from English word e.
- q(j|i,l,m) for any l Ɛ { 1….L}, m Ɛ {1….m}, iƐ{1…m}, j Ɛ {0…l}. The parameter q(j|i, l, m) can be interpreted as the probability of alignment variable ai taking the value j, conditioned on the lengths l and m of the English and foreign sentences.

Given these definitions, for any English sentence *e1 … el* where each ej Ɛ E, for each length m, we define the conditional distribution over foreign sentences *f1 … fm* and alignments a1 … am as:

$$p(f_1 \dots \dots f_{m,}\, a_1 \dots \dots a_m | e_1 \dots \dots e_l, m) = \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{ai}) \quad (2.7)$$

The alignment parameters, q(j|i, l, m) specify a different distribution (q(0|i, l, m); q(1|i, l, m),…., q(l|i, l, m)) for each possible value of the tuple i, l, m, where i is the position in the foreign sentence, l is the length of the English sentence, and m is the length of the Foreign sentence. This will allow us, for example, to capture the tendency for words close to the beginning of the foreign sentence to be translations of words close to the beginning of the English sentence.

#### 2.5.4.3 HMM Alignment

HMM alignment model is based on the familiar HMM model. As with IBM Model 1, the HMM alignment tries to compute *P(F,A|E)* [33]. The HMM model is based on a restructuring of this probability using the chain rule as follows:

$$p(f_1^J, a_1^J | e_1^I) = p(J|e_1^I) x \prod_{j=1}^{j} p(a_j, | f_1^{j-1}, a_1^{j-1}, e_1^I)\ x\ p(f_j, | f_1^{j-1}, a_1^{j-1}, e_1^I) \quad (2.8)$$

Where $p(J|e_1^I)$ = a length probability, $p(a_j, |f_1^{j-1}, a_1^{j-1}, e_1^I)$ = an alignment probability

$p(f_j, |f_1^{j-1}, a_1^{j-1}, e_1^I)$ = a lexicon probability

The main advantage of the HMM alignment model is that there are well-understood algorithms both for decoding and for training. For decoding, we can use the Viterbi algorithm [39] to find the best alignment for a sentence pair (F,E). For training, we can use the Baum-Welch algorithm [22].

## 2.6 Rule Based Machine Translation

Rule based machine translation has much to do with the morphological, syntactic and semantic information about the source and target language. Linguistic rules are built over this information. The methodology has several approaches known as direct approach, interlingua approach and transfer based approach, Figure 2.1.

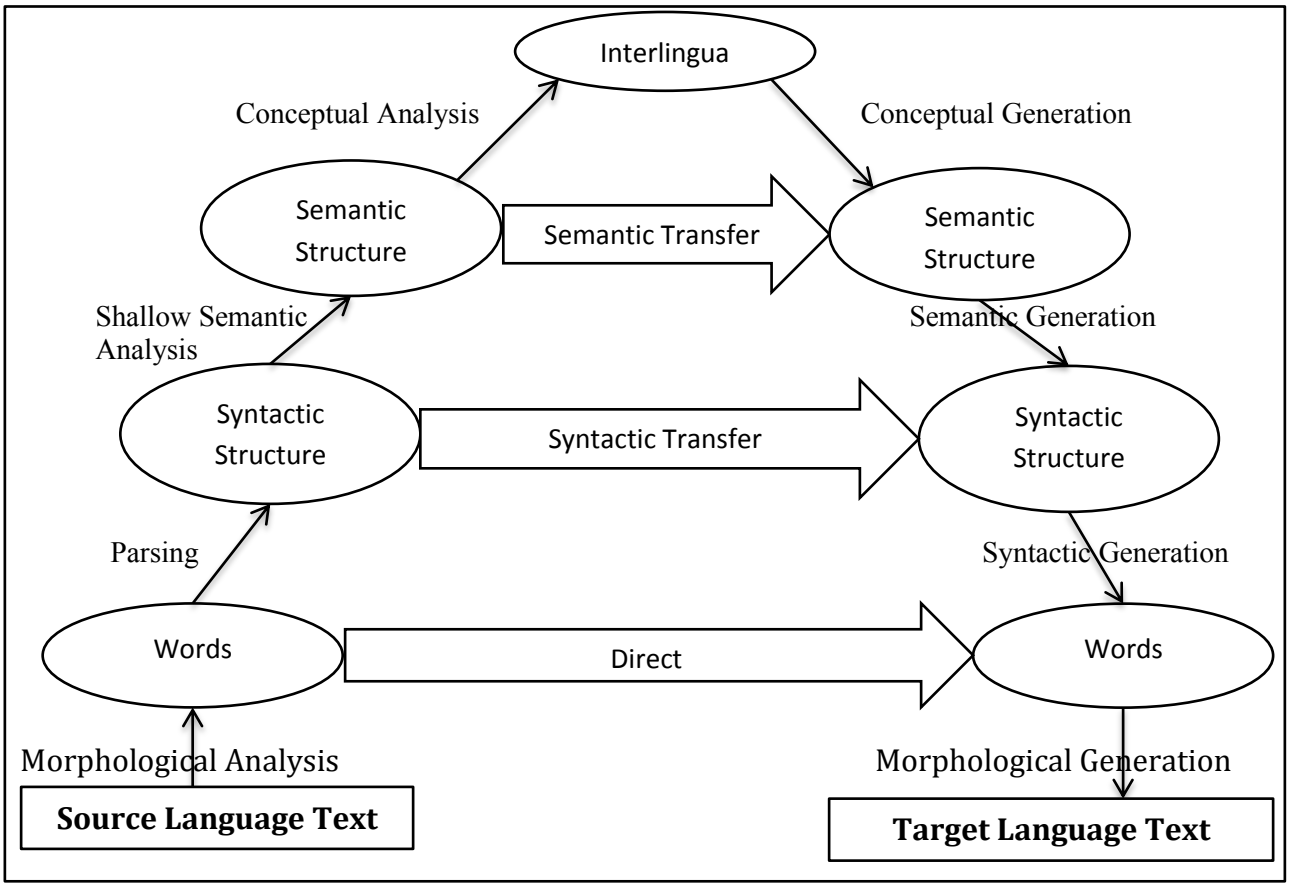



Figure 2.1. The Vauquois Triangle [22]

**Direct approach** is a MT approach which lacks any kinds of intermediate stages in translation processes [22]: the processing of the source language input text leads directly to the desired target language output text. In broad outline, first generation direct MT systems began with what we might call a morphological analysis phase, where there would be some identification of word endings and reduction of inflected forms to their uninflected basic forms, and the results would be input into a large bilingual dictionary look-up program. There would be no analysis of syntactic structure or of semantic relationships. In other words, lexical identification would depend on morphological analysis and would lead directly to bilingual dictionary look-up providing target language word equivalences. There would follow some local reordering rules to give more acceptable target language output, perhaps moving some adjectives or verb particles, and then the target language text would be produced. The direct approach is summarized in Figure 2.2.

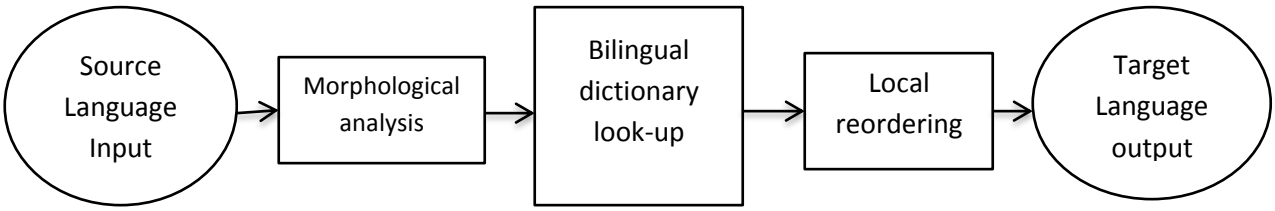



Figure 2.2. Direct machine translation system [22]

The approach can be characterized as 'word-for-word' translation with some local word-order adjustment. It gives the kind of translation quality that might be expected from someone with a very cheap bilingual dictionary and only the most rudimentary knowledge of the grammar of the target language: frequent mistranslations at the lexical level and largely inappropriate syntax structures which mirrored too closely those of the source language.

The linguistic and computational naivety of this approach was quickly recognized. From the linguistic point of view what is missing is any analysis of the internal structure of the source text, particularly the grammatical relationships between the principal parts of the sentences.

**Interlingua approach** is a MT approach where source language is transformed into an intermediary language (representation) which is independent of any of the languages involved in the translation. The translated verse for the target language is then derived through this intermediary representation. The intermediate representation includes all information necessary

for the generation of the target text without looking back to the original text. The representation is thus a projection from the source text and at the same time acts as the basis for the generation of the target text; it is an abstract representation of the target text as well as a representation of the source text. The method is interlingua in the sense that the representation is neutral between two or more languages. In the past, the intention or hope was to develop an interlingua representation which was truly universal and could thus be intermediary between any natural languages. At present, interlingua systems are less ambitious.

The interlingua approach is clearly most attractive for multilingual systems. Each analysis module can be independent, both of all other analysis modules and of all generation modules. Target languages have no effect on any processes of analysis; the aim of analysis is the derivation of an interlingua representation.

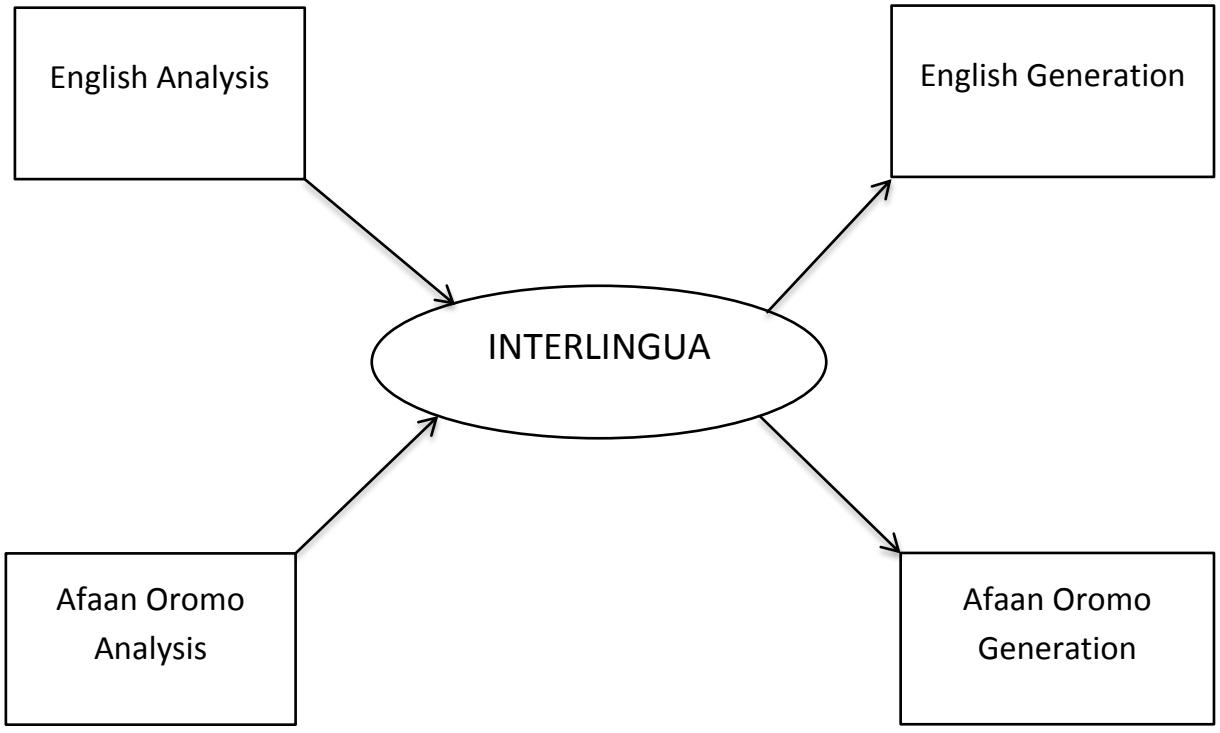



Figure 2.3. Interlingua model with two language models [22]

The advantage is that the addition of a new language to the system entails the creation of just two new modules: an analysis grammar and a generation grammar. By adding one analysts module in

Figure 2.3, for example, a French analysis grammar, the number of translation directions is increased from two (English to Afaan Oromo, and Afaan Oromo to French) to four (by the addition of French to Afaan Oromo and Afaan Oromo to English). The inclusion of another generation module, a French generation grammar, brings a further two pairs (English to French and French to English). Figure 2.4 shows interlingua model with three language models.

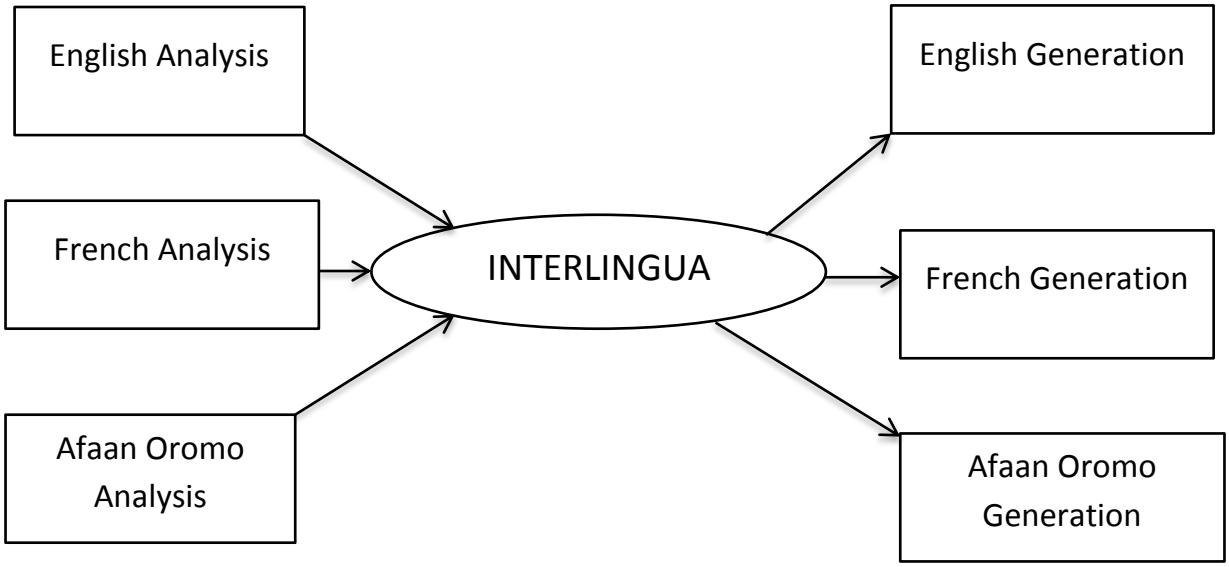



Figure 2.4. Interlingua model with three language models [22]

**Transfer based** approach is a MT approach where the source language is transformed into abstract, less language-specific representation [26]. An equivalent representation is then generated for the target language using bilingual dictionaries and grammar rules. First, the input text is parsed, and then rules are applied to transform the source language parse structure into a target language parse structure. Then, the target language sentence is generated from the parse structure.

In this approach, the translation proceeds in three stages, analyzing input sentences into a representation which still retains characteristics of the original, source language text [22]. This is then input to a special component (called a transfer component) which produces a representation which has characteristics of the target (output) language, and from which a target sentence can be synthesized as shown in Figure 2.5.

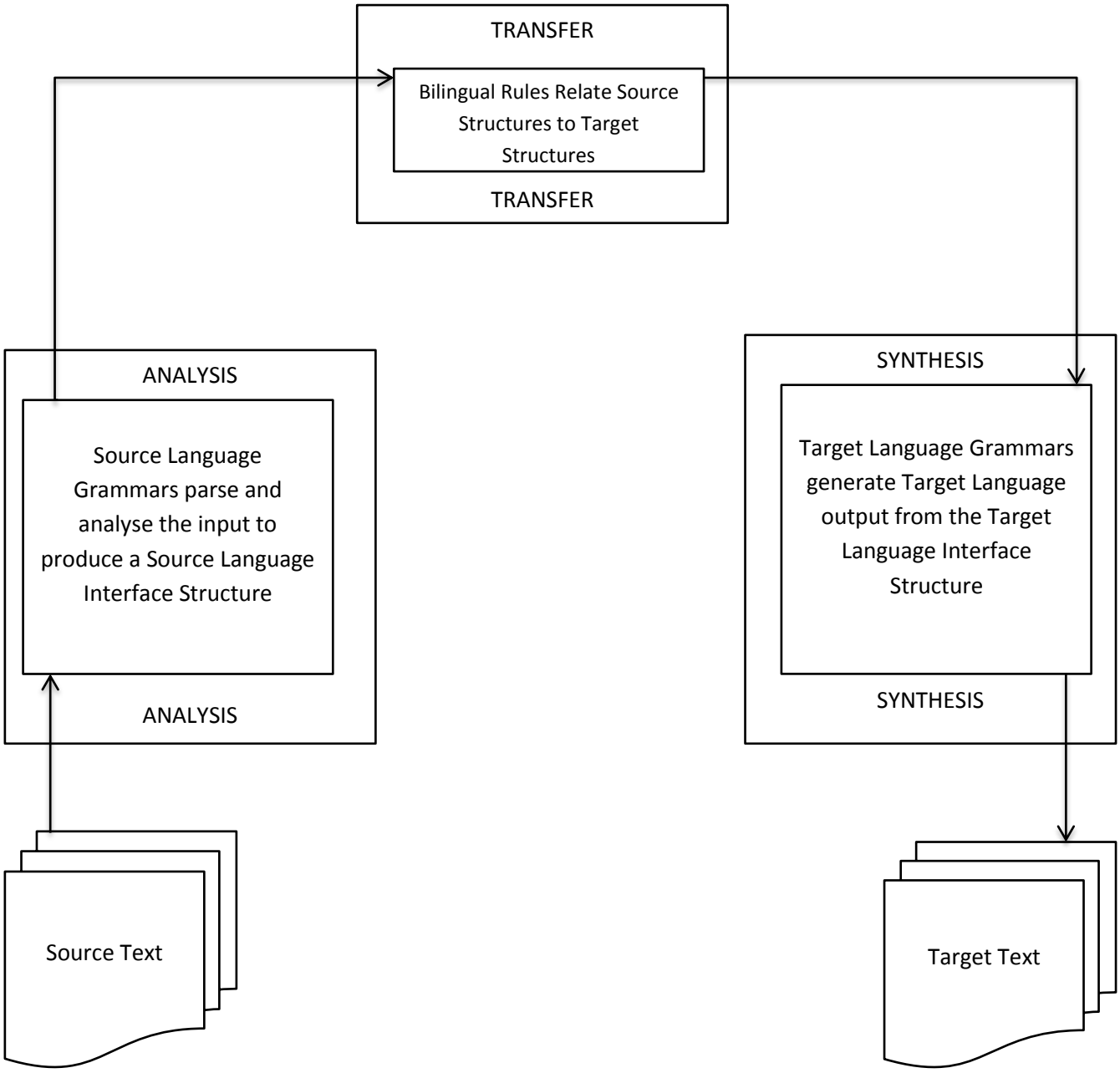



Figure 2.5. Components of a transfer system [23]

The triangle in Figure 2.1 shows the increasing depth of analysis required as we move from the direct approach through transfer approaches, to interlingua approaches. In addition, it shows the decreasing amount of transfer knowledge needed as we move up the triangle, from huge amounts of transfer at the direct level through transfer through interlingua.

## 2.7 Example Based Machine Translation

Example based machine translation is based on recalling/finding analogous examples of the language pairs. The examples are used to translate similar type of sentences of source language to the target language. The basic idea is that, if a previously translated sentence occurs again, the same translation is likely to be correct again [26].

The basic building block of EBMT is a large volume of translated texts (i.e., parallel bilingual texts), which have been translated by language professionals [34]. EBMT is about how to extract knowledge from bilingual texts.

In general, an example in context of EBMT is a pair of texts in two languages that are a translation of each other. There is no restriction on the size and linguistic level of the texts, the texts can be words, phrases, clauses, sentences or can be paragraphs. This implies that the example is not expected to have meaningful structure or constituent. Since there is a lower probability of longer sentences to show up in incoming texts, it is better to use short examples.

A critical issue that needs to be examined closely in this context is the number of examples over a large-scale bilingual corpus, which can be unlimited in practice. An example can be further decomposed, in more than one possible way, into sub-structures or shorter examples, and that examples can overlap with each other. Therefore, the example number can be exponentially large in respect to the corpus size, if we collect all possible examples from a bilingual corpus. Consequently, the impracticality and improbability of EBMT might arise, because any fragment of a sentence can be an example. We know that a language is well-known for utilizing limited resources to produce an unlimited number of expressions. Thus, it is an interesting issue to examine the practicality of EBMT in terms of the correlation of example number and corpus size. In practice, how to control an EB to a reasonable size becomes vitally critical. For this purpose, we need to determine what examples should be filtered out and which ones should be maintained in the EB, not only for the matter of efficiency but, more importantly, for practicality.

The relation of bilingual dictionary and EB is also worth of careful examination. Conventionally, a bilingual dictionary is a collection of lexical entries in one language and gives many possible translations in another language for each word. We can think of a bilingual dictionary as a

restricted EB, containing a collection of examples restricted at the word level. In return, an EB can be regarded as an extended bilingual dictionary. One might point out the fact that translating a word into another language following a bilingual dictionary is so uncertain, but translating a multiple word fragment of expressions in terms of an example from the EB is, in contrast, more sure. But this does not necessarily mean the dictionary and the EB do not share an intrinsic property for translation, namely, they provide choice of translation for a fragment, either single or multiple words, in an expression.

All choices need to be collected in the EB, highly similar to that in a dictionary, although the choices are significantly fewer. Hence, an empirical MT approach like EBMT can be understood as to tackle the following problem: given some observed translation as a set of expression fragments with their possible choices of translation in another language, find a reasonably good, if not the best, translation for next input word.

## 2.8 Hybrid Machine Translation

This approach leverages the strengths of statistical and rule – based translation methodologies. The methodology has two approaches [26]:

- *Rules post-processed by statistics*: Translations are performed using rule based engine. Statistics is then used in an attempt to adjust/correct the output from the rules engine.
- *Statistics guided by rules*: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating.

## 2.9 Evaluation of Machine Translation

Being able to evaluate a machine translation system is crucial. It provides a goal to work for, without which the exercise would be pointless. However, evaluation of machine translation has proven to be quite difficult [22]. For the most part, bilingual language users are able to perform a spontaneous evaluation on whether a translation is good or bad, or whether one translation is better than another. However, these judgments are subject to a high level of variance due to

many aspects such as expected purpose of the translation, the evaluators‟ level of conscious linguistic knowledge, and whether the evaluator weighs content or grammar highest.

Attempts have therefore been made at increasing inter-subjectivity by casting human evaluation within the boundaries of well-defined judgment tasks. These human evaluation tasks provide the best insight into the performance of a MT system, but they come with some major drawbacks. It is an expensive and time consuming evaluation method, and therefore it is less suited for tasks like everyday assessment of system progress or testing out new ideas. To overcome some of these drawbacks, automatic evaluation metrics have been introduced. These are much faster and cheaper than human evaluation, and they are consistent in their evaluation, since they will always provide the same evaluation given the same data. This is opposed to human evaluation where the intra-annotator agreement is far from perfect [36].

The disadvantage of automatic evaluation metrics is that their judgments are often not as correct as those provided by a human. The optimal metric should mirror the judgments of humans.

In short, MT evaluation is the task of scoring a translation given its source. Here, the MT system evaluates a variety of hypothesis translations created from the input. It then chooses the best translation based on this evaluation. In general, you need to solve the problem of MT to solve the problem of automatic MT evaluation.

The evaluation process, however, has the advantage that it is not tied by the realistic background of translation. Most often, evaluation is performed on sentences where one or more gold standard reference translations already exist. So, where the SMT system needs to piece its evaluation together from relevant references in its vast experience, the automatic evaluation metric possesses a fixed set of gold standard translations for reference that are targeted at exactly this translation task.

Nevertheless, even a large amount of gold standards will in reality not be enough to fully cover the potential variation leading to acceptable translations. This means that even though automatic evaluation has better premises, perfect evaluation still faces the same barriers as SMT in that it is necessary to evaluate based on an inadequate data set. Automatic MT evaluation therefore settles for high statistical correlations with human judgments over large amounts of data. This evens out

the noise brought on by the imperfections of automatic evaluation, but it also rules out certainty in evaluation of single sentences. The importance of human evaluation cannot be exaggerated. It provides the most correct and detailed picture of the performance of a MT system.

Nevertheless, for some tasks, automatic evaluation is necessary. In situations like everyday system evaluation, human evaluation can be too expensive, slow, and inconsistent. Therefore, an automatic evaluation metric that is reliable is very important to the progress of the field.

### 2.9.1 BLEU

BLEU score is an automatic machine translation evaluation method that compares n-gram overlap between a translation and possibly multiple references [37]. This is done based on the modified n-gram precision, which is calculated by dividing the number of n-grams in the translation that matches an n-gram in a reference, by the total number of n-grams in the translation. This is called modified, since each reference n-gram is only allowed to match once.

The BLEU score is measured on document level, not sentence level. This means that for a given n, modified precision is the total number of matching n-grams in the document divided by the total number of n-grams in all translations. This is formalized by equation:

$$precision_n = \frac{\sum_{c\epsilon\{candidates\}} \sum_{n-gram\ \epsilon\ c} count_{clip}(n-gram)}{\sum_{c'\epsilon\{candidates\}} \sum_{n-gram\ \epsilon\ c'} count_{clip}(n-gram')} \quad (2.9)$$

Where C = candidate words found in a reference

C‟‟ = candidate words which are not found in a reference

While it makes perfect sense to compare how much of a translation is found in any reference (precision), it makes less sense to examine how much of all the reference translations are present in the translation (recall). If for example the same meaning is expressed in four different ways in four references, then the translation will get full credit for hitting one of these in precision, but since only one of four is present in the translation, recall will be bad. Recall therefore seems inappropriate with multiple references.

For these reasons, the BLEU metric avoids recall and instead introduces a heuristic brevity penalty (BP). This penalty punishes too short translations, which will otherwise have a tendency

to get higher precision due to fewer total n-grams. That is, the BP acts as a counterweight to the modified precision. The BP is based on the reference that is closest in length for each translation. Then summing over the entire corpus, a total length of all references (r) and all translations (t) is found. BP is then calculated by equation:

$$\text{BP} = \begin{cases} 1, & c > r \\ e\left(1 - \frac{r}{c}\right) & c \leq r \end{cases} \qquad (2.10)$$

Where BP = brevity penalty

r = references

c = corpus

The BLEU score is calculated by taking the geometric mean of the modified precision for N‟s up to a maximum n-gram length and multiplying it by the BP as in equation (2.11). The maximum n-gram length is usually set to N = 4, and n-grams are usually weighted equally $W_n = 1/N$.

$$\text{BLEU} = \text{BP} * \exp(\textstyle\sum_{n=1}^{N} \log p_n) \qquad (2.11)$$

The BLEU metric has been one of the most influential additions to the field of MT. This is both for good and for bad. BLEU was the first automatic evaluation metric proven to display a high level of correlation with human evaluations [37]. Now people had a subjective, fast and cheap evaluation option that it was feasible to optimize system parameters against. This made a lot more research possible, and less time was used on evaluation.

It is, however, uncertain to what extent this metric has controlled the direction of machine translation. If everybody is using the same metric, and this metric bias in favor of certain directions, then these directions will get more attention. To a certain extent, BLEU seems to contain such biases.

The main problem with BLEU may not as much be the metric itself, as it is the way people utilize it. BLEU stands for Bilingual Evaluation Understudy, and as the word ‟understudy‟ signals, the metric was meant as a supplement to human judges.

# CHAPTER THREE

# Related Work

## 3.1 Introduction

In this Chapter, we have critically reviewed previous works done on statistical machine translation (SMT) and hybrid machine translation (HMT). Studies on English – Afaan Oromo, English – Amharic machine translations and some studies conducted on English and other foreign languages are all reviewed.

## 3.2 English – Oromo Machine Translation: An Experiment Using a Statistical Approach

The study which was conducted by Sisay mainly deals with translation of English documents to Afaan Oromo using statistical methods [13]. The study was carried out with two main goals: the first one is to apply existing SMT systems 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 documents and legal documents. 20,000 bilingual sentences and 62, 300 monolingual sentences were used for training and testing purpose. The documents were preprocessed by using different scripts which are customized to handle some special behaviors of Afaan Oromo such as apostrophe. Sentence aligning, tokenization, lowercasing and truncating long sentences that take the alignment to be out of optimality were also done by those scripts. He used SRILM toolkit and GIZA++ for language modeling and word alignment respectively. The Moses decoder was used for decoding purpose. By using these resources, an average BLEU score of 17.74% was achieved based on the experimentation.

## 3.3 Bidirectional English-Amharic Machine Translation: An Experiment using constrained corpus

The study was done by Eleni with the objective of developing a bidirectional English-Amharic machine translation system using constrained corpus [12]. The research work was implemented by using statistical machine translation approach. In the study, she collected and prepared two

different corpora; the first corpus was made of 1,020 simple sentences and the other one is made of 1951 complex sentences. Since the translation is bidirectional, two language models were developed, one for Amharic and the other for English and translation models were also built. A decoder which searches for the shortest path was used and expectation maximization algorithm was used for aligning words in the accurate order. Two different experiments were conducted and the evaluation was performed by using two different methodologies. The first experiment was performed using simple sentences and evaluated by using manual questionnaire and automatic evaluation method. The result obtained for simple sentences using BLEU score has an average score of 82.22% accuracy for English to Amharic translation and 90.59% for Amharic to English translation and using the manual questionnaire method, accuracy of English to Amharic translation is 91% and accuracy of Amharic to English translation is 97%. For complex sentences, the result acquired from the BLEU score is 73.38% for English to Amharic translation and 84.12% for Amharic to English translation and from questionnaire method, accuracy of English to Amharic translation is 87% and Amharic to English translation is 89%.

The study shows Amharic to English translation has a better accuracy than English to Amharic translation.

## 3.4 Preliminary Experiments on English-Amharic Statistical Machine Translation (EASMT)

The experiment was performed with the objective of translating English text to Amharic text using the statistical machine translation approach [11]. The experiment on the EASMT system is conducted using available parallel documents. A total of 18,432 English-Amharic parallel sentences were used in the experiment. A pre-processing task is performed on the parallel documents in order to retain and convert the full content into a valid format suitable for the system. Some of these pre-processes include text conversion, trimming, sentence splitting, sentence aligning and tokenization. The process of trimming was performed before and after aligning at document level. Sentence splitting was done before starting aligning at sentence level while tokenization is performed after aligning at the sentence level. The alignment at the sentence level has been done using a sentence aligner called Hunalign. Out of the total collected data, 16,432 randomly selected sentence pairs were used for training while the remaining 2,000 sentence pairs were used for tuning and testing. The researcher used SRILM toolkit and GIZA++

for language modeling and word alignment respectively. He used Moses decoder for decoding purpose. Accordingly, he achieved a BLEU score of 35.32%. The researcher achieved a 0.34% increase in BLEU by applying morpheme segmentation to the tokens of the Amharic output result and the reference of the baseline system.

## 3.5 English Syntactic Reordering for English-Thai Phrase-Based Statistical Machine Translation

The study was carried out to show the way of improving a phrase based SMT in language pairs which have different word orders by using a syntactic reordering approach with SMT approach [38]. The study proposes reordering rules for English-Thai phrase based SMT system. The reordering rules transform both training and test English sentences in a preprocessing step. First, source language sentences were parsed by a parser then the parse trees of source language sentences were transformed by reordering rules for making word orders of the source sentences more similar to word-orders of the target sentences. The experiment was conducted on a hand-made English-Thai parallel corpus in the Moses phrase-based system. The parallel corpus they have used consists of 4,621 English-Thai sentence pairs. Thai sentences in the parallel corpus were translated from English sentences and segmented manually. English sentences in the training set were classified into different categories, such as affirmative sentences, interrogative sentences begin with "verb to be", interrogative sentences begin with "What", etc, and were parsed by the Stanford Parser. Reordering rules are extracted from the classified parse trees of the training set. Then reordering rules are applied for transforming English parse trees to make word orders of English sentences more similar to word orders of Thai sentences. After this, the phrase based SMT system is trained by the training set, which is reordered. After that, the reordered test set is translated by the reordered phrase based SMT system.

The reorder approach they have used improves accuracy of English-Thai translation in the Moses phrase based SMT system by increasing the BLEU score from 40.05% to 57.45%.

## 3.6 Chinese Syntactic Reordering for Statistical Machine Translation

The study introduces a reordering approach for translation of Chinese to English [5]. The study describes a set of syntactic reordering rules that exploit systematic differences between Chinese and English word order. The system is used as a preprocessor for both training and test

sentences, transforming Chinese sentences to be much closer to English in terms of their word order. The researchers used the Penn Chinese Treebank guideline in searching for a suitable set of reordering rules. They ruled out several phrase types as not requiring reordering rules. Three categories that are considered to be the most prominent candidates for reordering are identified. These phrases include VPs (verb phrases), NPs (noun phrases), and LCPs (localizer phrases, which frequently map to prepositional phrases in English). The baseline of the experiment was a phrase-based MT system trained using the Moses toolkit. The training data consists of nearly 637K pairs of sentences from various parallel news corpora distributed by the Linguistic Data Consortium (LDC). For tuning and testing, the official NIST MT evaluation data for Chinese from 2002 to 2006, which have four human generated English reference translations for each Chinese input was used. The evaluation data from 2002 to 2005 were split into two sets of roughly equal sizes: a tuning set of 2347 sentences was used for optimizing various parameters using minimum error training, and a development set of 2320 sentences was used for various analysis experiments. A series of processing steps were performed before the reordering rules were applied, which include segmentation, part-of-speech tagging, and parsing. They trained a Chinese Treebank-style tokenizer and part-of-speech tagger, both using a tagging model based on a perceptron learning algorithm. The reordering rules were then applied to the parse tree of each input. The reordered sentence is then re-tokenized to be consistent with the baseline system.

The reordering approach they have used improved the BLEU score for the Moses system from 28.52% to 30.86% on the NIST 2006 evaluation data.

## 3.7 Summary

Generally, the studies have been done by using statistical machine translation approach and hybrid machine translation approach. In most of the studies, English is used as a source language and statistical approach is used as a base translation system. For example, in English-Thai Phrase Based Statistical Machine Translation, they have used syntactic reordering approach in which rules are written in order to make the structure of the source sentence to be similar to the target language and then the training and the translation is done by using statistical approach. In most of the studies, different NLP tools such as SRILM, IRSTLM, GIZA++ and Moses are used for training and decoding purpose and BLEU score is used for the evaluation purpose. In this

research work, we use parallel corpus collected from different domains, syntactic reordering approach as a preprocessing step and statistical approach for training and translation purpose.

# CHAPTER FOUR

# Design of Bidirectional English – Afaan Oromo Machine Translation

## 4.1 Introduction

This Chapter discusses the bidirectional English – Afaan Oromo machine translation system. The overall system architecture and its components are all discussed in detail.

## 4.2 System Design

The architecture of the system which is shown in Figure 4.1 has different components which will be discussed in detail in the next sections.

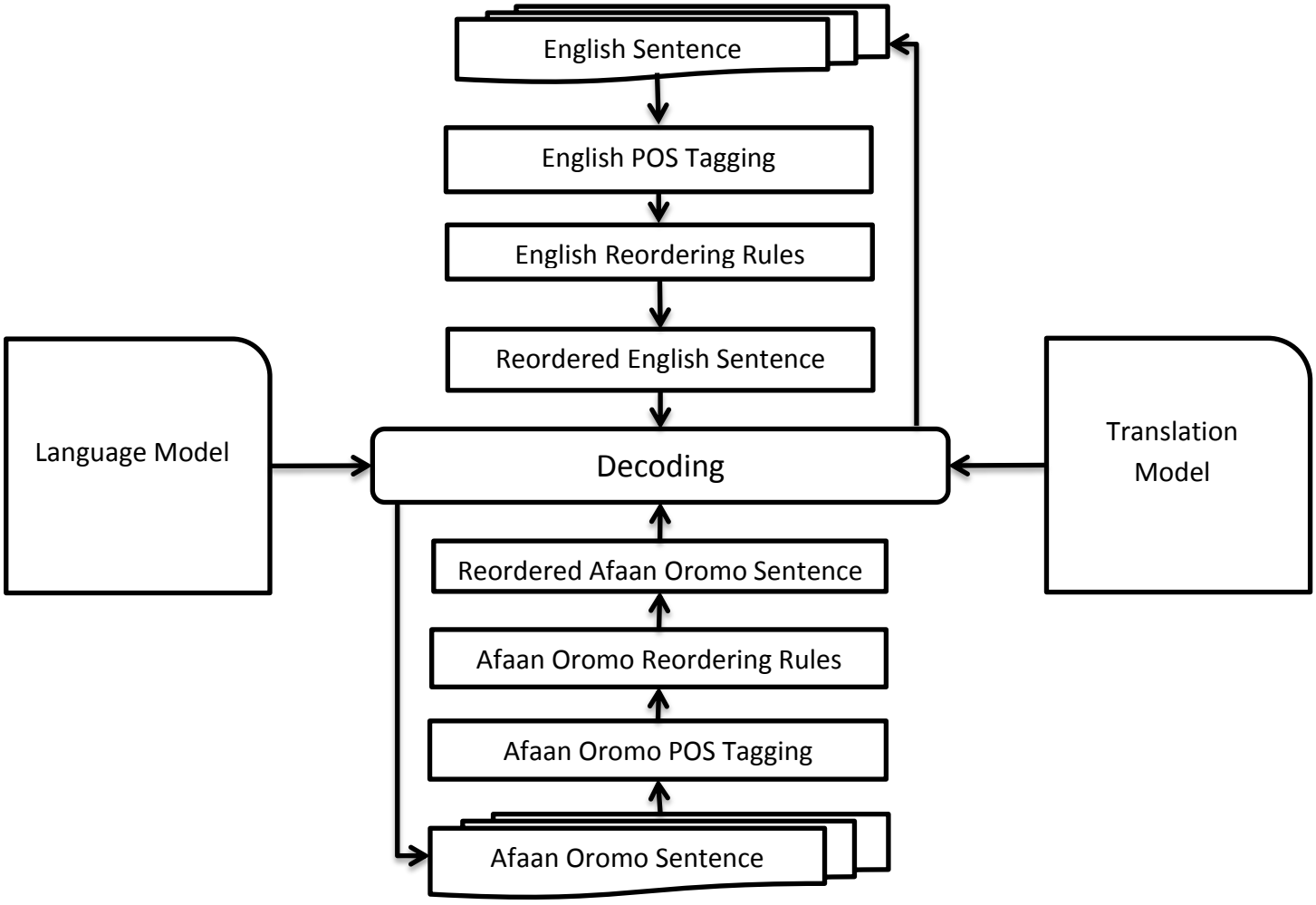



Figure 4.1. Architecture of the System

### 4.2.1 English POS Tagging

This component is the first step in rule part which assigns parts of speech to each word (and other token) for English sentences, so that the tagged sentences will be used as an input in English reordering rules. We have used a Stanford POS tagger which is a publically available POS tagger for English, German, Chinese and Arabic languages.The English tagger uses the Penn Treebank POS tag set as indicated in Table 4.1.

Table 4.1. The Penn Treebank POS tag set

| Number | Tag | Description |
|---|---|---|
| 1 | CC | Coordinating conjunction |
| 2 | CD | Cardinal number |
| 3 | DT | Determiner |
| 4 | EX | Existential *there* |
| 5 | FW | Foreign word |
| 6 | IN | Preposition or subordinating conjunction |
| 7 | JJ | Adjective |
| 8 | JJR | Adjective, comparative |
| 9 | JJS | Adjective, superlative |
| 10 | LS | List item marker |
| 11 | MD | Modal |
| 12 | NN | Noun, singular or mass |
| 13 | NNS | Noun, plural |
| 14 | NNP | Proper noun, singular |
| 15 | NNPS | Proper noun, plural |
| 16 | PDT | Predeterminer |
| 17 | POS | Possessive ending |
| 18 | PRP | Personal pronoun |
| 19 | PRP$ | Possessive pronoun |
| 20 | RB | Adverb |
| 21 | RBR | Adverb, comparative |
| 22 | RBS | Adverb, superlative |
| 23 | RP | Particle |
| 24 | SYM | Symbol |
| 25 | TO | *to* |
| 26 | UH | Interjection |
| 27 | VB | Verb, base form |
| 28 | VBD | Verb, past tense |
| 29 | VBG | Verb, gerund or present participle |
| 30 | VBN | Verb, past participle |
| 31 | VBP | Verb, non-3rd person singular present |
| 32 | VBZ | Verb, 3rd person singular present |
| 33 | WDT | Wh-determiner |
| 34 | WP | Wh-pronoun |
| 35 | WP$ | Possessive wh-pronoun |

For example, if the input sentence is “she is my sister” then the POS tagger will assign part of speech to each word in the sentence and it displays it as “she_PRP is_VBZ my_PRP$ sister_NN”.

The main problem of this POS tagger is that it tags some non-English name as a foreign word rather than tagging it with a proper tag label. For example, if the input is “Lalise” then it will tag this word as “Lalise_FW” rather than tagging it as “Lalise_NN”. In order to handle such problems we have corrected the tagged corpus manually by changing FW tag labels into a proper tag label such as NN, NNP, NNS and so on.

### 4.2.2 English Reordering Rules

English and Afaan Oromo have a different sentence structure which is a problem for SMT. As it was described in Section 2.2.2, English language has SVO sentence structure whereas Afaan Oromo has SOV sentence structure. Because of the difference in word order, words in the source sentence are aligned to target words which have a different position.

This is the main part of the rule in English to Afaan Oromo translation which makes English sentences in the corpus to have a more similar sentence structure with that of Afaan Oromo. The advantage of applying this reordering technique is to minimize the problem of syntactic reordering in case of SMT.

We classify reordering rules into three main categories: reordering rules for simple sentences, interrogative sentences and complex sentences.

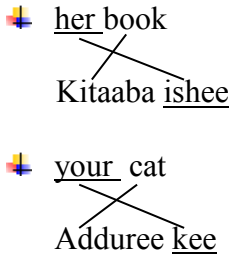
#### 4.2.2.1 Reordering rules for simple English Sentences

Reordering rules for simple sentences are farther classified into different sub-sections:

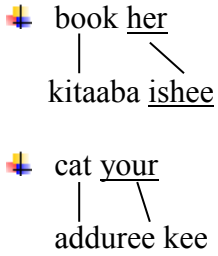
**1. Reordering rules for possessive pronouns**

A pronoun is a word that replaces a noun in a sentence, making the subject a person or a thing. Possessive pronouns are pronouns that demonstrate ownerships. In English language, possessive pronouns include my, mine, our, ours, its, his, her, hers, their, theirs, your, yours, whose, and ones which are all words that demonstrate ownership.

For example, in the following phrases, the underlined words are possessive pronouns:




As it is shown in the above two English phrases and their translations in Afaan Oromo, the pronouns appear before the nouns in English whereas the pronouns appear after the nouns in Afaan Oromo. Thus, in order to avoid such reordering problems we have reordered the words in the English phrases so that they can have similar structure with Afaan Oromo. After applying the reordering rules to the above two phrases, we can get the following reordered phrases:




The above reordering is done by using Algorithm 4.1.

1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i \, \mathcal{E} \, S$ do, where i=1,2,3,...k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,....l
6. if PRP$ in $w_j$ and NN or NNP or NNS or NNPS in $w_{j+1}$
7. tmp = $w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1}$ = tmp
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

Algorithm 4.1. Algorithm for reordering possessive pronouns in English sentences

**2. Reordering rules for prepositional phrases:**

A prepositional phrase is a phrase that begins with a preposition and ends with a noun, pronoun, gerund, or clause, the "object" of the preposition. The object of the preposition will often have one or more modifiers to describe it.

For example, in the following prepositional phrases, the underlined words are prepositions:

- in the car
  konkolaataa keessa
- with Gadise
  Gaaddisee waliin

As it is shown in the above two English phrases and their translations in Afaan Oromo, the prepositions appear before the nouns in English whereas the prepositions appear after the nouns in Afaan Oromo. So, in order to avoid such reordering problems we have reordered English phrases so that they can have similar structure with Afaan Oromo. After applying the reordering rules to the above two phrases we can get the following reordered phrases:

- the car in
  konkolaataa keessa
- Gadise with
  Gaaddisee waliin

The above reordering is done by using Algorithm 4.2.

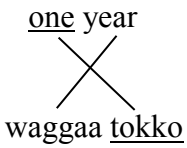
1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i \in S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if IN in $w_j$ and NN or NNP or NNS or NNPS in $w_{j+1}$
7. $tmp = w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1} = tmp$
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

Algorithm 4.2. Algorithm for reordering prepositional phrases in English sentences

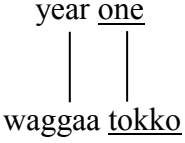
### 3. Reordering rules for cardinal number

Cardinal numbers are words that indicate a number such as one, two, ten, and hundred and so on. So, the cardinal numbers are structured in such a way that the cardinal number appears before nouns in English sentence whereas it appears after the noun in Afaan Oromo.

For example, in the following English and Afaan Oromo phrases, the underlined words are cardinal numbers whereas the other words are not.




Algorithm 4.3 solves such reordering problems in such a way that it puts the noun before the cardinals and the cardinals next to the noun. After applying this rule, the above English phrase is changed into the following phrase.

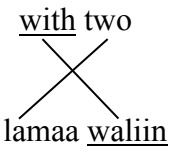


1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if CD in $w_j$ and NN or NNP or NNS or NNPS in $w_{j+1}$
7. tmp = $w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1}$ = tmp
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

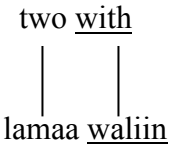
Algorithm 4.3. Algorithm for reordering cardinal numbers in English sentences

**4. Reordering rules for prepositions with cardinal number**

In English phrases and sentences prepositions are placed before cardinal numbers whereas they are placed after the cardinals in case of Afaan Oromo. For example, in the following English phrase and its Afaan Oromo translation the underlined words are prepositions and the others are cardinals.




Algorithm 4.4 solves such reordering problems in such a way that it puts the cardinals before the prepositions and the prepositions next to the cardinals. After applying this rule, the above English phrase is changed into the following phrase.

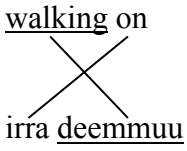


1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if IN in $w_j$ and CD in $w_{j+1}$
7. $tmp = w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1} = tmp$
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

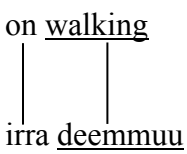
Algorithm 4.4. Algorithm for reordering preposition with cardinal number in English sentences

**5. Reordering rules for present participle verbs with prepositions**

In English phrases having present participle verbs and prepositions, the words are arranged in such a manner that the present participle verbs come before the preposition. If we take Afaan Oromo phrases, present participle verbs come next to the preposition. For example, in the following English and Afaan Oromo phrases the underlined words are present participle verbs and the others are prepositions.




Algorithm 4.5 rearranges the words of English in such phrases so that the phrase will have similar arrangement with that of Afaan Oromo phrase. So, if we apply this rule on the above phrase, we can get the following reordered phrase.

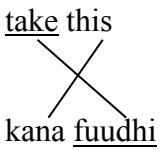


1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,2,3,….l
6. if VBG in $w_j$ and IN in $w_{j+1}$
7. tmp = wj
8. $w_j = w_{j+1}$
9. $w_{j+1}$ = tmp
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

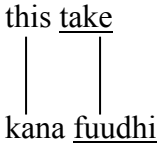
Algorithm 4.5. Algorithm for reordering present participle verbs with prepositions in English sentences

**6. Reordering rules for Verbs**

As it has been discussed in Section 2.2.2, in English the verb comes before the object and it comes next to the object in case of Afaan Oromo. The purpose of this rule is to make the subject, verb and object arrangement of English to be similar with that of Afaan Oromo. For example, if we take the following English and Afaan Oromo phrases the underlined words are verbs and the rest are words of other part of speech.



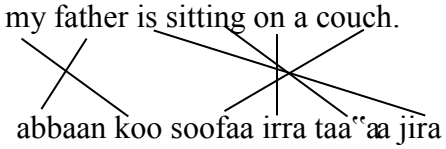

Algorithm 4.6 changes the place of verbs and puts them at the end so that the following result will be achieved.



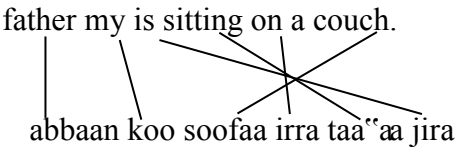
1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if VB in $w_j$
7. tmp = $w_j$
8. $w_j = w_{(length\ (s)-1)}$
9. $w_{(length\ (s)-1)}$ = tmp
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

Algorithm 4.6. Algorithm for reordering verbs in English sentences

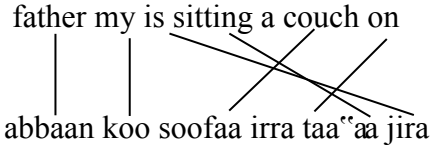
The rules which are mentioned in the above sub-sections are all combined together in order to reorder a complete English simple sentences. For example, if we have the following sentence:



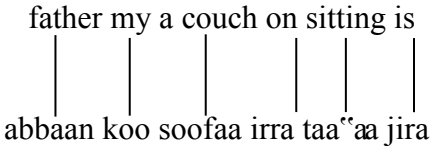

After applying the first rule it will be reordered into the following sentence




The next step is to apply the second rule and the sentence is changed into the following order



The last step is to apply rule number six which is changing the place of verbs in the sentence so that we can get the following reordered sentence

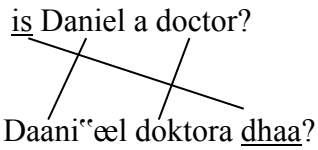



#### 4.2.2.2 Reordering rules for interrogative English sentences

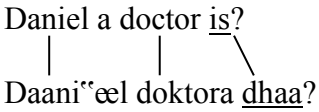
Reordering rules for interrogative English sentences are further sub classified by the beginning words of interrogative English sentences. The beginning words need to be reordered to new positions. The rules are discussed in detail in the following sections.

1. **Reordering rules for interrogative sentences beginning with "verb to be", "verb to do", "verb to have", and "auxiliary verb":**

All sentences starting by "is", "am", "are", "was", "were", "do", "does", "did", "can", "may", "will" and "would" are all categorized under this rule. The "verb to be", "verb to do", "verb to have", and "auxiliary verb" are reordered after the object of interrogative sentences. This makes the structure of the interrogative sentences to become like simple sentences. After that, we use reordering rules for simple sentences given in the previous subsection. For example, in the following sentence the underlined word is verb to be.




After applying the rule, the words are reordered and the sentence will have the same structure with Afaan Oromo



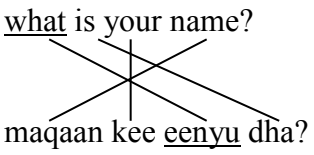
1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if QVB in $w_j$
7. tmp = $w_j$
8. $w_j = w_{(length\ (s)-1)}$
9. $w_{(length\ (s)-1)}$ = tmp
10. apply reordering rules for simple sentences
11. store $s_i$ in $S'$
12. End if
13. End for
14. End for
15. Write $S'$ to file

Algorithm 4.7. Algorithm for reordering interrogative sentences beginning with "verb to be", "verb to do", "verb to have", and "auxiliary verb in English sentences

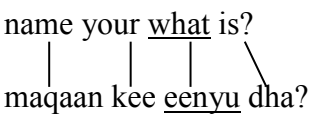
**2. Reordering rules for interrogative sentences beginning with "WHADVP"**

This rule is for interrogative sentences beginning with "WH" words. For example "Where", "What", "When", "How" and "Why" are all WHADVPs. The WHADVPs are repositioned to the end of the sentences. Finally, reordering rules for simple sentences are used for reordering the other words. The "WHADVP" is reordered after the object of interrogative sentences. This makes the structure of the interrogative sentence to become like the simple sentences. After that, we use reordering rules for simple sentences given in the previous subsection.

For example, in the following sentence the underlined word is "WHADVP"




After applying the rule, the words are reordered and the sentence will have the same structure with Afaan Oromo.



1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,...k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,....l
6. if WP or WRB in $w_j$ and NN or NNS or NNP or NNPS in $w_{j+1}$
7. tmp = $w_j$
8. tmptwo = $w_{j+1}$
9. $w_j = w_{(length\ (s)-2)}$
10. $w_{j+1} = w_{(length\ (s)-1)}$
11. $w_{(length\ (s)-2)}$ = tmp
12. $w_{(length\ (s)-1)}$ = tmptwo
13. apply reordering rules for simple sentences
15. store $s_i$ in $S'$
14. End if
15. End for
16. End for
17. Write $S'$ to file

Algorithm 4.8. Algorithm for reordering interrogative sentences beginning with ***"WHADVP"***

#### 4.2.2.3 Reordering rules for complex English Sentences

This reordering rule, Algorithm 4.9, is used to make the structure of complex English sentences more similar to the structure of Afaan Oromo. First, the complex sentences are divided into simple sentences by the rule, then the rules which are mentioned in Section 4.2.2.1, are applied on it. After this the reordered simple sentences are combined so that they can form a reordered complex sentence. For example, if we apply the rule on the following sentence:

Jack is eating dinner, but his brother is sleeping on the couch.

The rule splits the sentence into two different simple sentences depending on the conjunction word "*but*" and comma.

S1: Jack is eating dinner
S2: his brother is sleeping on the couch

After this the reordering rules for the simple sentences will be applied on S1 and S2 so that it gives the following two sentences:

S1: Jack dinner eating is
S2: brother his the couch on sleeping is

The last step is to combine the two sentences in order to get the following result:

1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=1,2,3,….l
6. if ',' in $w_j$ and CC in $w_{j+1}$
7. divide s into simple sentences as s1,s2,….sn
8. apply reordering rules for simple sentences
9. concatenate s1,s2...sn into s
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

Algorithm 4.9. Algorithm for reordering complex sentences

### 4.2.3 Language Model

The role of the language model is to give an estimate of how probable a sequence of words is to appear in the target language. The language model helps the translation system with selecting words or phrases appropriate for the local context and with combining them in a sequence with better word order. The most common approach to language modeling is to estimate the probability of a word conditioned on a window of preceding words called the history. These types of language models are called N-gram language models. In order to have a reliable estimate for the language model probabilities, the context of the language model is usually restricted to a few words. Another restriction on the context of language models is that the probability of words is computed within the boundaries of a sentence. The probabilities obtained from the N-gram model could be unigram, bigram, trigram or higher order N-grams. For example, if we have the following Afaan Oromo sentences,

Daggaaggaan konkolaataa bite.
Daggaaggaan konkolaataa oofe.
Ililliin konkolaaticha bitte.
Ililliin buna bitte.

Ililliin buna dhugdde.
Ililliin buna danfistte.

The unigram probability can be computed as

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

$$p(\text{Ililliin}) = \frac{4}{18} = 0.222$$

The bigram probability can be computed as

$$p(w2|w1) = \frac{\text{count (w1w2)}}{\text{count (w1)}}$$

$$p(\text{buna}|\text{Ililliin}) = \frac{3}{4} = 0.75$$

The trigram probability can be computed as

$$p(w3|w1w2) = \frac{\text{count (w1w2w3)}}{\text{count (w1w2)}}$$

$$p(\text{bite}|\text{Ililliin buna}) = \frac{1}{3} = 0.33$$

Since the system is bidirectional a language model has been developed for both English and Afaan Oromo by using N-gram language model with the help of IRSTLM tool. For English to Afaan Oromo translation, a language model has been developed for Afaan Oromo language, and for Afaan Oromo to English translation, a language model has been developed for English language.

**Smoothing**

To avoid assigning zero or low probabilities to unseen events, the maximum likelihood estimates for the conditional probabilities are adjusted to make the distribution more uniform. By doing so, zero and low probabilities are increased and high probabilities are decreased. By applying this type of technique called smoothing, the estimations are improved and unseen N-grams receive some probability.

### 4.2.4 Translation Model

In general, translations require many-to-many alignments between words. This means that a group of words in the source language should be translated by a group of words in the target language and there might not be a word-level correspondence between these groups. In phrase-

based models, the smallest unit used for translation is called a phrase. By translating phrases instead of words, the problem of word ambiguity is partially solved.

The translation model considers the sentences to be split into phrases, and all ways to split the sentences are equally likely. The likelihood of generating string O given string E is then factored in terms of probabilities of generating a phrase $O_i$ given a phrase $E_i$.

### 4.2.5 Decoding

Decoding is the main part of the translation system which takes source sentence E and produces the best translation O according to the product of the translation and language models. As it has been discussed in Section 2.5.3, decoding is a search problem which finds the sentence which maximizes the translation and language model probabilities. Since the system is phrased based translation, the decoder uses the phrase translation table to search for all translations of every source words or phrases, and then it recombines the target language phrases that maximizes the language model probability by the translation model probability multiplied. Therefore, when the translation is from English to Afaan Oromo, the best translation is the probability that maximizes the product of the probability of English – Afaan Oromo translation model p(e|o) and the language model of Afaan Oromo p(o), i.e.:

$$\underset{o}{\textbf{argmax}}\ \textbf{(p(e|o)*p(o))} \qquad (4.1)$$

When the translation is from Afaan Oromo to English, the best translation is the probability that maximizes the product of the probability of Afaan Oromo – English translation model p(o|e) and the language model of English p(e), i.e.:

$$\underset{e}{\textbf{argmax}}\ \textbf{(p(o|e)*p(e))} \qquad (4.2)$$

### 4.2.6 Afaan Oromo POS Tagging

The purpose of this component is to assign parts of speech to each word (and other token) for Afaan Oromo sentences, so that the tagged sentences will be used as an input in Afaan Oromo reordering rules. Since there is no publically available automatic POS tagger for Afaan Oromo

we have tagged the sentences manually by the tag set shown in Table 4.2 and the tag is examined by linguistic expert.

Table 4.2. The POS tag set used for Afaan Oromo POS tagging

| No. | Tags | Description |
|---|---|---|
| 1 | CC | Coordinating conjunction |
| 2 | CD | Cardinal numbers |
| 3 | IN | Preposition or subordinating conjunction |
| 4 | JJ | Adjective |
| 5 | NN | Noun, singular or mass |
| 6 | NNS | Noun, plural |
| 7 | NNP | Proper noun, singular |
| 8 | NNPS | Proper noun, plural |
| 9 | IN | Preposition |
| 10 | PRP$ | Possessive pronoun |
| 9 | RB | Adverb |
| 10 | VB | Verb, base form |
| 11 | VBD | Verb, past tense |
| 12 | VBG | Verb, gerund or present participle |
| 13 | VBN | Verb, past participle |
| 14 | VBP | Verb, non-3rd person singular present |
| 15 | VBZ | Verb, 3rd person singular present |
| 16 | PN | A tag for all punctuations in the language. |
| 17 | WH | A tag for all Wh question words in the language. |

For example, if we have untagged Afaan Oromo sentence “Daraaraan kolfaa jira”, then it will be tagged as “Daraaraan_NNP kolfaa_VBG jira_VBZ” because the word “Daraaraan” belongs to a proper noun, the word “Kolfaa” is a present participle verb and the word “jira” is a third person singular verb.

### 4.2.7 Afaan Oromo Reordering Rules

This is the other part of the rule which works in Afaan Oromo to English translation. This rule will make each Afaan Oromo sentence in the corpus to have a more similar sentence structure

with that of English. Therefore, the rule changes SOV sentence structure into SVO sentence structure, so that it will minimize the problem of syntactic reordering in case of SMT.

We classify reordering rules into three main categories: reordering rules for simple sentences, interrogative sentences and complex sentences:

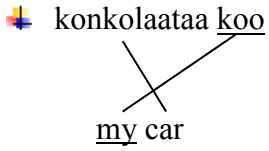
#### 4.2.7.1 Reordering rules for simple Afaan Oromo sentences

The reordering rules for simple sentences are farther classified into different sub-sections:

**1. Reordering rules for possessive pronouns**

A pronoun is a word that replaces a noun in a sentence, making the subject a person or a thing. Possessive pronouns are pronouns that demonstrate ownerships. In Afaan Oromo possessive pronouns include koo, kiyya, keenya, isaa, ishee, isaani, and kee which are all words that demonstrate ownership.

For example, in the following phrases, the underlined words are possessive pronouns:





As it is shown in the above two Afaan Oromo phrases and their translations in English, the possessive pronouns appear after the nouns in Afaan Oromo whereas the pronouns appear before the nouns in English. So, in order to avoid such reordering problems we have reordered Afaan Oromo phrases so that they can have similar structure with English. After applying the reordering rules to the above two phrases, we can get the following reordered phrases:

- keenya mana

  our home

The above reordering is done by using Algorithm 4.10.

1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i \, \mathcal{E} \, S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if NN or NNS or NNP or NNPS in $w_j$ and PRP$ in $w_{j+1}$
7. $tmp = w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1} = tmp$
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

Algorithm 4.10. Algorithm for reordering possessive pronouns in Afaan Oromo sentences

**2. Reordering rules for prepositional phrases:**

In Afaan Oromo a prepositional phrase will begin with a noun, pronoun, gerund, or clause, the "object" of the preposition and end with a preposition. For example, in the following prepositional phrases the underlined words are prepositions:

- kuticha keessa

  in the room

- kitaabicha irra

  on the book

As it is shown in the above two Afaan Oromo phrases and their translations in English the prepositions appear after the nouns in Afaan Oromo and they appear before the nouns in English.

So, in order to avoid such reordering problems the rule will reorder Afaan Oromo phrases so that they can have similar structure with English. After applying the reordering rules, Algorithm 4.11, to the above two phrases we can get the following reordered phrases:

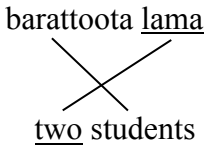
- keessa kuticha
  - in the room
- irra kitaabicha
  - on the book

1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i \, \mathcal{E} \, S$ do, where i=1,2,3,...k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,....l
6. if NN or NNS or NNP or NNPS in $w_j$ and IN in $w_{j+1}$
7. $tmp = w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1} = tmp$
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

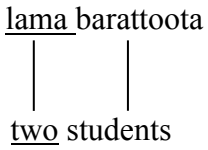
Algorithm 4.11. Algorithm for reordering prepositional phrases in Afaan Oromo sentences

### 3. Reordering rules for cardinal number

In Afaan Oromo, cardinal numbers are words that indicate a number such as tokko, lama, kudhan, kuma tokko and so on. So, these cardinal numbers are structured in such a way that the cardinal number appears after the noun in Afaan Oromo whereas it appears before nouns in English sentence. For example, in the following Afaan Oromo and their English translation phrases the underlined words are cardinal numbers.




The rule solves such reordering problems in such a way that it puts the noun after the cardinal numbers and the cardinal numbers before the noun. After applying this rule, Algorithm 4.12, the above Afaan Oromo phrase is changed into the following phrase.

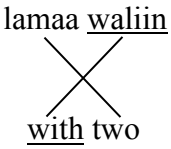



1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if NN or NNS or NNP or NNPS in $w_j$ and CD in $w_{j+1}$
7. tmp = $w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1}$ = tmp
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

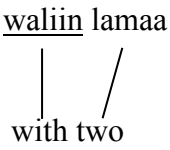
Algorithm 4.12. Algorithm for reordering cardinal number in Afaan Oromo sentences

**4. Reordering rules for prepositions with cardinal number**

In Afaan Oromo, phrases and sentences prepositions are placed after the cardinal numbers whereas they are placed before the cardinal numbers in English. For example, in the following Afaan Oromo and its English translation phrase the underlined words are prepositions and the other ones are cardinal numbers.




The rule solves such reordering problems in such a way that it puts the cardinals after the prepositions and the prepositions before the cardinals. After applying this rule, Algorithm 4.13, the above Afaan Oromo phrase is changed into the following phrase.




1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if CD in $w_j$ and IN in $w_{j+1}$
7. $tmp = w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1} = tmp$
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

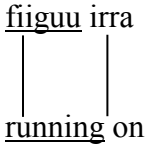
Algorithm 4.13. Algorithm for reordering prepositions with cardinal number in Afaan Oromo sentences

**5. Reordering rules for present participle verbs with prepositions**

In Afaan Oromo phrases having present participle verbs and prepositions, the words are arranged in such a manner that the present participle verbs come after the preposition. For example, in the

following Afaan Oromo and English phrases the underlined words are present participle verbs and the others are prepositions.


This rule rearranges the words of Afaan Oromo in such phrases so that the phrase will have the similar arrangement with that of English phrase. So, if we apply this rule, Algorithm 4.14, on the above phrase, we can get the following reordered phrase.

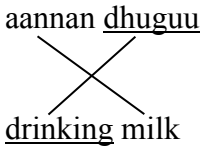



1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $S$ do, where i=1,2,3,...k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,....l
6. if IN in $w_j$ and VB in $w_{j+1}$
7. $tmp = w_j$
8. $w_j = w_{j+1}$
9. $w_{j+1} = tmp$
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

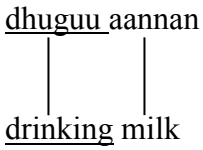
Algorithm 4.14. Algorithm for reordering present participle verbs with prepositions in Afaan Oromo sentences

## 6. Reordering rules for Verbs

As it has been discussed in Chapter Two in Afaan Oromo, the verb comes after the object and it comes before the object in case of English. The purpose of this rule is to make the subject, verb and object arrangement of Afaan Oromo to be similar with that of English. For example, if we take the following Afaan Oromo and English phrases the underlined words are verbs and the rest are words of other part of speech.



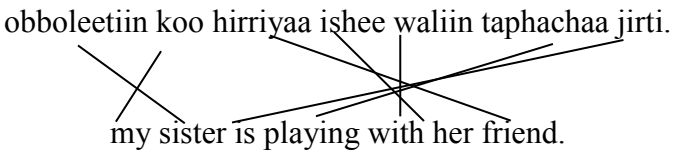

The rule, Algorithm 4.15, changes the place of verbs and puts them before the noun so that the following result will be achieved.



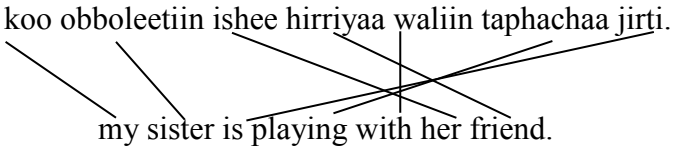

1. Load the sentences from POS Tagged corpus
2. Store all sentences in $\mathcal{S}$
3. Store all words in W
4. for each $s_i$ $\mathcal{E}$ $\mathcal{S}$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if VB in $w_{(length(s)-1)}$
7. tmp = $w_{(length\ (s)-1)}$
8. $w_{(length\ (s)-1)}$= $w_{(length\ (s)-2)}$
9. $w_{(length\ (s)-2)}$ = tmp
16. store $s_i$ in $\mathcal{S}'$
10. End if
11. End for
12. End for
13. Write $\mathcal{S}'$ to file

Algorithm 4.15. Algorithm for reordering verbs in Afaan Oromo sentences

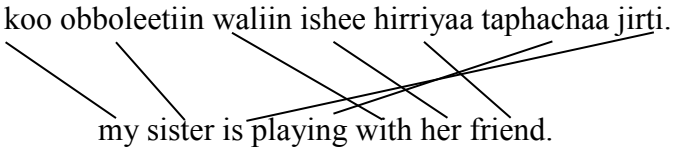
The rules which are mentioned in the above sub sections are all combined together in order to reorder a complete Afaan Oromo simple sentences. For example, if we have the following sentence:



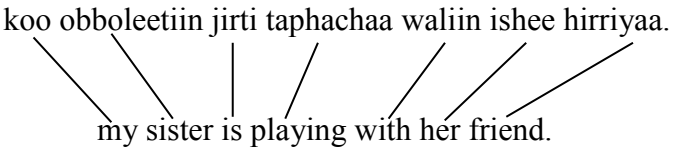

After applying the first rule it will be reordered into the following sentence




The next step is to apply the second rule and it is changed into the following order




The last step is to apply rule number six which is changing the place of verbs in the sentence so that we can get the following reordered sentence



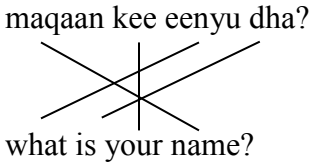

#### 4.2.7.2 Reordering rules for interrogative Afaan Oromo sentences

Reordering rules for interrogative Afaan Oromo sentences are different from one another depending on the interrogative pronouns of interrogative Afaan Oromo sentences. Interrogative pronouns are used in questions, and come before the verb and either before or after the subject. The main interrogative pronouns are:

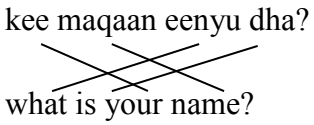
**Maal(i) -** What
**Maaliif(i) -** Why
**Akkam(i) -** How
**Yoom -** When
**Eessa -** Where

**Eessaa -** From where
**Eenyu –** Who, What
**Kan eenyu -** Whose
**Meeqa** - How much, How many
**Kam(i) -** Which

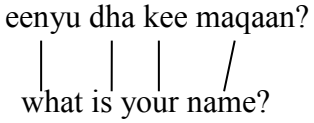
The rule, Algorithm 4.16, changes the structure of Afaan Oromo interrogative sentences so that it will be similar with that of English interrogative sentence structure. For example, if we have the following interrogative Afaan Oromo sentence:




First, the sentence will be considered as non-interrogative sentence and the rules which are mentioned in Section 4.2.7.1 will be applied on it. So, the above sentence will be changed into:




At last the reordering rule for interrogative sentences is applied on it and it will have the following sentence structure which is similar with that of English:



1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i \, \mathcal{E} \, S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if WH in $w_j$
7. $tmp = w_0$
8. $tmptwo = w_1$
9. $w_0 = w_j$
10. $w_1 = w_{j+1}$
11. $w_j = tmp$
12. $w_{j+1} = tmptwo$
13. apply reordering rules for simple sentences
14. store $s_i$ in $S'$
15. End if
16. End for
17. End for
18. Write $S'$ to file

Algorithm 4.16. Algorithm for reordering "WH" words in Afaan Oromo sentences

#### 4.2.7.3 Reordering rules for complex Afaan Oromo sentences

This reordering rule, Algorithm 4.17, is used to make the structure of complex Afaan Oromo sentences more similar to the structure of English sentences. First, the complex sentences are divided into simple sentences by the rule, then the rules which are mentioned in Section 4.2.7.1 are applied on it, after this the reordered simple sentences are combined so that they can form a reordered complex sentence. For example if we apply the rule on the following Afaan Oromo sentence:

inni mana keessa taa‟aa jira, haati isaa garuu deemteeti.

The rule splits the sentence into two different simple sentences depending on the comma.

S1: inni mana keessa taa‟aa jira
S2: haati isaa garuu deemteeti.

After this the reordering rules for the simple sentences will be applied on S1 and S2 so that it gives the following two sentences:

S1: inni jira taa‟aa keessa mana
S2: garuu isaa haati deemteeti

The last step is to combine the two sentences in order to get the following result:

inni jira taa‟a keessa mana garuu isaa haati deemteeti

1. Load the sentences from POS Tagged corpus
2. Store all sentences in $S$
3. Store all words in W
4. for each $s_i \; \mathcal{E} \; S$ do, where i=1,2,3,…k
5. for each $w_j$ in $s_i$ do, where j=0,1,2,3,….l
6. if ‘,’ in $w_j$ and CC in $w_{j+1}$
7. divide s into simple sentences as s1,s2,….sn
8. apply reordering rules for simple sentences
9. concatenate s1,s2...sn into s
10. store $s_i$ in $S'$
11. End if
12. End for
13. End for
14. Write $S'$ to file

Algorithm 4.17. Algorithm for reordering complex Afaan Oromo sentences

# CHAPTER FIVE

# Experiment

## 5.1 Introduction

Based on the design of Chapter Four, bidirectional English – Afaan Oromo machine translation is developed by using a hybrid of rule based and statistical approaches. This Chapter evaluates its performance by conducting two experiments. The first experiment focuses on statistical approach while the other experiment uses the hybrid approach.

## 5.2 Corpus Collection

Hybrid approach which is the combination of corpus based approach and rule based approach requires the availability of bilingual parallel corpus. In this research work, parallel documents of both English and Afaan Oromo languages that are publically available are used. This parallel corpus includes some chapters of the Holy Bible, the Constitution of FDRE, the Criminal code of FDRE, International conventions, Megeleta Oromia and a bulletin from Oromia health bureau. The parallel corpus is collected from different web sites and different offices. Most of the collected data is in a PDF file format and the others are collected in hard copy. The data which is collected in hard copy is all typed and changed into softcopy and saved in a text file format. The PDF files are also changed into text file so that it is ready for the next process.

## 5.3 Corpus Preparation

After the corpus has been collected it was prepared in a format that is suitable for the translation purpose. Therefore, the following three procedures have been applied on the collected corpus to make it ready for training and testing the translation system.

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

True-casing: This is the procedure that takes place after the tokenization step and makes the initial words in each sentence to be converted to their most probable casing. This helps reduce data being sparse.

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

We have used a total of 3000 English – Afaan Oromo parallel sentences for training and testing the system. From the total of 3000 parallel sentences, 2, 900 parallel sentences were used for training the system whereas the rest were used for testing the system.

Since the system is bidirectional, the training process was performed in both directions, that is, from English to Afaan Oromo and from Afaan Oromo to English. Similar parallel sentences and similar steps were conducted for both directions in order to train and test the system.

As it has been mentioned in Chapter One, showing the advantage of using hybrid approach rather than using statistical approach in machine translation for the language pair is one of the specific objectives of the research work. Therefore, two major experiments were conducted by using similar parallel sentences using two different approaches. The first experiment was conducted by using a statistical approach and the second experiment was conducted by using a hybrid approach. Similar 3000 parallel sentences were used to conduct both the experiments. Similar tools were used for training and testing in both the experiments. Both the experiments will be discussed briefly in the following sections as Experiment I and Experiment II.

## 5.4 Experiment I

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

### 5.4.1 Training the system

We have used the data set described in Section 5.3 to perform the training and testing procedures. From a total of 3000 English – Afaan Oromo parallel sentences 2,900 sentences were used to train the system.

Moses which is freely available software is used to train the system in both directions, English to Afaan Oromo and Afaan Oromo to English by using similar and the same number of English –

Afaan Oromo parallel sentences. The training process includes creating a language model, creating the translation model, tuning and decoding with the help of GIZA++ and IRSTLM tools. First the training was performed for English to Afaan Oromo translation and then it was performed for Afaan Oromo to English translation.

### 5.4.2 Result of Test set on Experiment I

We have used 100 English and Afaan Oromo parallel sentences in order to test the performance of the system in terms of translation accuracy and the time it takes to translate a single English sentence to Afaan Oromo sentence and vice versa. BLEU score methodology which was discussed in Chapter Two is used in order to see the result of the translation in both directions. The result recorded from the BLEU score methodology shows 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation.

The reason behind the difference between both the records was that there is a difference between feminine and masculine representation in English and Afaan Oromo languages. For Example, let‟s look at the following sentence:

"Kaku is running"

"*Kaku is running*" is translated as "*Kakuun fiigaa dha*" rather than translating it as "*Kakuun fiigaa jirti*" or "*Kakuun fiigaa jira*". In this translation process "*Kaku*" is translated as "*Kakuun*" and "*running*" is translated as "*fiigaa*" which are both correct translations. But the system is unable to translate the word "*is*" as "*jira*" or "*jirti*", rather it translates it as "*dha*" which is not a correct translation. The main reason behind this is that the system is unable to identify to which gender "*Kaku*" belongs to. But if we take the translation of the above two Afaan Oromo sentences "*Kakuun fiigaa jirti*" and "*Kakuun fiigaa jira*" they are both translated as "*Kaku is running*" because both "*jira*" and "*jirti*" are translated as "*is*". Because of this reason the translation accuracy of Afaan Oromo to English translation is better than English to Afaan Oromo translation. In addition to this the time it takes to translate Afaan Oromo sentence into English is less than the time it takes to translate English sentence into Afaan Oromo. Therefore, an efficient and accurate translation is acquired when Afaan is used as a source sentence and English is used as a target sentence.

## 5.5 Experiment II

This is the second and main experiment conducted on English – Afaan Oromo language pair by using a hybrid approach. As it was discussed in Chapter One the main objective of this research work is to develop a bidirectional English – Afaan Oromo machine translation using a hybrid approach.

First the rules mentioned in Chapter Four are applied on the training and test set then both the data sets were prepared for training and testing the system. Since the rules were applied before the training and testing step, the procedure of training and testing of the hybrid approach is similar with that of statistical approach. In case of English to Afaan Oromo translation first English reordering rules which are discussed in Section 4.2.2, are applied on English sentences so that the sentences will have similar syntactic structure with that of Afaan Oromo sentences. In case of Afaan Oromo to English translation, first Afaan Oromo reordering rules which are discussed in Section 4.2.7, are applied on Afaan Oromo sentences so that the sentences will have similar syntactic structure with that of English sentences. Finally, the statistical approach is applied on the prepared and reordered corpus.

### 5.5.1 Training the system

This experiment was conducted by using a reordered data set which is a total of 3000 English – Afaan Oromo parallel sentences. Out of 3000 parallel sentences we have used 2,900 parallel sentences for the training purpose and we have used 100 parallel sentences for the testing purpose.

Moses which is freely available software is used to train the system in both directions, English to Afaan Oromo and Afaan Oromo to English by using similar and the same number of English – Afaan Oromo parallel sentences. The training process includes creating a language model, creating the translation model, tuning and decoding with the help of GIZA++ and IRSTLM tools. First the training was performed for English to Afaan Oromo translation and then it was performed for Afaan Oromo to English translation.

### 5.5.2 Result of Test set on Experiment II

We have used 100 English and Afaan Oromo parallel sentences in order to test the performance of the system in terms of translation accuracy and the time it takes to translate a single English sentence to Afaan Oromo sentence and vice versa. In English to Afaan Oromo translation, reordered English sentences are used and in Afaan to English translation, Afaan Oromo reordered sentences are used. BLEU score methodology is used in order to see the result of the translation in both directions. The result recorded from the BLEU score methodology shows that 37.41% for English to Afaan Oromo translation and 52.02% for Afaan Oromo to English translation. The reason behind the difference between both the records was that there is a difference between feminine and masculine representation in English and Afaan Oromo. For Example, let"s look at the following sentence:

"Dame loves him"

"Dame loves him" is translated as "*Dammeen isa jaallata*" rather than translating it as "*Dammeen isa jaallati*" or "*Dameen isa jaallata*". The reason behind these is that the system is unable to identify to which gender "*Dame*" belongs to. Therefore, the system translates "*Dame*" as "*Dammeen*" which can also be translated as "*Dameen*". It also translates "*loves*" as "*jaallata*" which can also be translated as "*jaallati*". But, if we take the translation of the above Afaan Oromo sentence "*Dammeen isa jaallati*" it is translated as "*Dame loves him*" because "*Dammeen*" is translated as "*Dame*" and "*jaallati*" is translated as "*loves*". Because of this reason, the translation accuracy of Afaan Oromo to English translation is better than English to Afaan Oromo translation. In addition to this, the time it takes to translate Afaan Oromo sentence into English is less than the time it takes to translate English sentence into Afaan Oromo. Therefore, an efficient and accurate translation acquired when Afaan is used as a source language and English is used as a target language.

The experiments are conducted by using two different approaches. From the result of both 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 English – Afaan Oromo language pair. We can also see that better translation accuracy is acquired when Afaan Oromo is used as a source sentence and English is used as a target sentence.

# CHAPTER SIX

# Conclusion and Recommendation

## 6.1 Conclusion

The purpose of this study is to develop a bidirectional English – Afaan Oromo machine translation system. Bidirectional English – Afaan Oromo machine translation is based on hybrid approach in which rules are used in preprocessing step and the translation is made by using a statistical approach.

The design process of bidirectional English – Afaan Oromo machine translation involves collecting English – Afaan Oromo parallel corpus, corpus preparation which also involves dividing the corpus as a training set and test set, designing reordering rules for both languages by using Python programming language, language modeling by using IRSTLM tool and training the system by using Moses. Language modeling, translation modeling and decoding are all components of the statistical approach.

Reordering rules are designed to make the structure of the source language to be more similar to the structure of the target language by using their part of speech. Stanford POS tagger is used in order to tag English sentences and Afaan Oromo sentences are tagged manually. In order to design reordering rules, syntactic structures of both the languages pair have been studied. The reordering rules are designed for simple sentences, interrogative sentences and complex sentences.

Finally two experiments were conducted by using the collected data set to check the accuracy and efficiency of the system by using two different approaches. The first experiment is conducted by using a statistical approach and it has a BLEU score of 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation. The second experiment is carried out by using a hybrid approach and it has a BLEU score of 37.41% for English to Afaan Oromo translation and 52.02% for Afaan Oromo to English translation. From the test result of the conducted experiment we have seen that the hybrid approach is better than the statistical approach for English – Afaan Oromo language pair.

## 6.2 Recommendation

Machine translation is an important research area of NLP application and the approach we have used has its own contribution to this, especially for English – Afaan Oromo language pair and other related languages such as English – Amharic and Afaan Oromo – Amharic language pairs. Finally, we would like to mention the following main points as a future work:

- Further results can be accomplished by increasing the size of the data set used for training the system. So, by increasing the size of the training data set one can develop a full-fledged bidirectional English – Afaan Oromo machine translation.
- The rules which are developed and used in the system are only used for syntax reordering. Therefore, additional results can be accomplished by further exploring the rules especially by developing morphological rules.
- The system is developed in order to translate English text into Afaan Oromo text and vice versa. Since the text translation is available, speech to text and text to speech translations can be developed for this language pair.
- The hybrid approach that we have used in this research work can be applied on other language pairs such as English – Amharic language pair and Afaan Oromo – Amharic language pair.
- Further results can also be achieved by adding a disambiguation component into the system.

## References


**[1]** Charniak Eugene. *Introduction to artificial intelligence*, Addison Wesley Publishing Company, Boston, 1984

**[2]** Karen Sparck Jones, *Natural Language Processing, a historical review*, University of Cambrige, Cambrige, October 2001

**[3]** W.John Hutchins, "Machine Translation: A Brief History", In: *proceedings of the Concise history of the language sciences: from the Sumerians to the cognitivists*, Oxford: Pergamon Press, Pages 431-445, 1995.

**[4]** Preslav Nakov, "Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing", *In: proceedings of the third workshop on statistical machine translation*, pages 147-150, Ohio, June 2008

**[5]** Chao Wang, Michael Collins and Philipp Koehn, "Chinese Syntactic Reordering for Statistical Machine Translation", In: *proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Pages 737–745, Prague, June 2007.

**[6]** Mouiad Alawneh, Nazlia Omar and Tengku Mohd Sembok *, "*Machine Translation From English To Arabic**",** In: *proceedings of the 2011 International Conference on Biomedical Engineering and Technology*, Singapore, 2011

**[7]** Holger Schwenk, Jean-Babtiste Fouet and Jean Senellart*, "*First Steps towards a general purpose French/English Statistical Machine Translation System*", In: proceedings of the third workshop on statistical machine translation*, pages 119-122, Ohio, June 2008

**[8]** Bushra Jawaid, Daniel Zeman, "Word-Order Issues in English-to-Urdu Statistical Machine Translation" *In: proceedings of the Prague Bulletin of Mathematical Linguistics*, pages 87–106, Prague, April 2011

**[9]** Nadir Durrani, Hassan Sajjad, Alexander Fraser, Helmut Schmid, "Hindi-to-Urdu machine translation through transliteration", *In: proceedings of the 48th annual meeting of the association for computational linguistics*, Pages 465-474, Stroudsburg, 2010

**[10]** Michael Gasser, "Toward a Rule-Based System for English-Amharic Translation", *In: proceedings of the 8th International Workshop of the ISCA Special Interest Group on Speech and Language Technology for Minority Languages and the 4th workshop on African Language Technology*, May 22, 2012

**[11]** Mulu Gebreegziabher Teshome, Laurent Besacier, *"*Preliminary Experiments on English-Amharic Statistical Machine Translation**",** *In: proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU)*, Cape town, May 9, 2012

**[12]** Eleni Teshome "*Bidirectional English – Amharic Machine Translation: An Experiment using constrained corpus*", MSc thesis, Addis Ababa University, Ethiopia, 2013

**[13]** Sisay Adugna, "*English-Oromo Machine Translation: An Experiment Using a Statistical Approach*", MSc thesis, Addis Ababa University, Ethiopia, 2009

**[14]** Gene B. Gragg, "Oromo Dictionary", *In: proceedings of journal of the American Oriental Society*, USA, 1982

**[15]** Gadaa Malbaa, *Oromia: an introduction to the history of the Oromo people*, Khartum, 1988

**[16]** Mahdi, Hamid M., *Oromo Dictionary, English-Oromo*, Sagalee Oromo Publishing co. Atlanta, Georgia 1995

**[17]** Ton Leus. Et al., *Borena Dictionary*, W.S.D. Grafisch Centram Schijndel, Holland, 1995

**[18]** Baye Yimam, "*The Phrase Structure of Ethiopian Oromo*". The Degree of Ph. D in Linguistics, School of Oriental and African University of London, London, 1986

**[19]** Hamiid M., *English-Oromo Dictionary*, Sagalee Oromoo Publishing Inc, Atlanta, 1996

**[20]** Tesfaye, Debela. "A rule base afan Oromo grammar checker", *In: proceedings of the International Journal of Advanced Computer Science and Applications*, USA, 2011

**[21]** Tilahun, G., "Qube Afaan Oromo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet", *In: proceedings of the Journal of Oromo Studies*, 1993

**[22]** Daniel Jurafsky. and James H. Martin, *Speech and Language Processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, 2006

**[23]** Arnold D., Lorna B., Siety M., R. Lee H., Louisa S., *Machine Translation: an introduction giude*, NCC Blackwell, London, 1994

**[24]** Jacob Elming, "*Syntactic Reordering In Statistical Machine Translation*", Copenhagen Business School, a PhD thesis, June 2008

**[25]** Sneha Tripathi and Juran Krishna Sarkhel, "*Approaches to machine translation*", December 2010.

**[26]** Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, John D. Lafferty, Robert L. Mercer & Paul S. Roossin, *A statistical approach to machine translation*, Computational Linguistics, 1990

**[27]** Adam Lopez, "Statistical Machine Translation", *In: proceedings of the ACM Computing Surveys*, University of Edinburgh, 2008

**[28]** Matthew J. Post, "*Syntax-based Language Models for Statistical Machine Translation*", a PhD thesis, University of Rochester, Rochester, New York, 2010

**[29]** Christopher D. Manning, Hinrich Schiitze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, May 1999

**[30]** Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra & Robert L. Mercer, *The mathematics of statistical machine translation: parameter estimation*, Computational Linguistics, 1993.

**[31]** Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lai, J., and Mercer, R. L., "*Method and system for natural language translation*", 1995

**[32]** Michael Collins, "*Statistical Machine Translation: IBM Models 1 and 2*",

**[33]** Stephan Vogel, Hermann Ney and Christoph Tillmann, "HMM-Based Alignment in Statistical Translation", *In: proceedings of the Second Workshop on Statistical Machine Translation*, pages 80–87, Germany, 2007

**[34]** Chunyu Kit, Haihua Pan and Jonathan J. Webster, "*Example-Based Machine Translation: A New Paradigm*", Department of Chinese Translation and Linguistics, City University of Hong Kong, Hong Kong, 1990

**[36]** Kishore Papineni, Salim Roukos, Todd Ward & Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation", *In: Proceedings of the 40th Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, 2007

**[37]** George Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics", *In: Proceeding of the ARPA Workshop on Human Language Technology*, Morgan Kaufmann Publishers Inc., USA, 2002

**[38]** Nawaphol Labutsri, Rapeeporn Chamchong, Richard Booth, Annupan Rodtook, "*English Syntactic Reordering for English-Thai Phrase-Based Statistical Machine Translation*" , Bangkok, Thailand, 2005

**[39]** M.S. Ryan and G.R.Nudd, *"The Viterbi Algorithm",* Department of Computer Science, University of Warwick, Coventry, England, 1993

# Appendices

## Appendix I: Sample Parallel Corpus for Training

| | |
|---|---|
| Husen is sawing a wood, but Fatuma is sitting idle. | Huseen muka soofaa jira, garuu faaxumaan calluma jettee taa'aa jirti. |
| There is no news about him, however I went on hoping. | Waa'een isaa homtuu hin dhaga'amne, haa ta'u malee anii abdii koo itti fufeen jira. |
| We went out for a picnic and visited some relatives on our way. | Nuyi qilleensa fudhachuuf baanee, achumaan firoota dubisne. |
| I haven't seen Bontu lately, and I won't see her again. | Anni Boontuu yeroo dhihooti hin argine kana boodees ishee hin argu. |
| Mr. Derara listens to the news every night, and then he goes to bed. | Obbo daraaraan galgala yeroo hundaa oduu dhageefatee achii gara sireeti deema. |
| Will you wait for me, or do you want to go ahead? | Anaan na eegda moo deemuu barbaada? |
| We had finished our homework, and I am tired. | Nutti Hojii manaa keenya xumureera aniis dadhabeera. |
| I had originally planned to attend the meeting however now I find I can't. | Anni jalqaba walgahicha hirmaachuu yaadeen ture ta'us amma akka hin dandeenye naaf gale. |
| Mrs. Derartu likes to drive she finds it relaxing. | Adde daraartuun ooffuu jaallati ishee bashanansiisa. |
| Bontu is afraid of the water consequently, she had trouble passing the swimming test. | Boontuun bishaan sodaati kanaafuu qormaata daakkaa darbuu hin dandeenye. |
| Should I eat this food, or do you want to eat it? | Anni Nyaata kana nyaadhu moo Atti nyaachu barbaada? |
| Helen was calm after the accident, but her sister was very nervous. | Balicha booda heelan tasgaboofteeti garuu obboleettiin ishee baay'ee naati. |
| The dog barked and it woke the baby. | Sarichi dutee mucaa hirribaa kaase. |
| I opened the door. | Anni balbalicha baneera. |
| Bontu is beautiful. | Bontuun bareedduu dha. |
| kalkidan is a student. | Qaalkidaan barattuu dha. |
| The childrens are taking an exam. | Ijoolleen qormaata fudhachaa jiru. |
| He loves enjera. | Inni buddeena jaalata. |
| Gelan is not crazy. | Galaan maraataa miti. |
| Hana is getting married. | Haanaan heerumuuf jetti. |
| Haile loves competition. | Hayileen dorgomii jaalatta. |
| Gifti wants to be a president. | Giiftiin pireezedantii ta'uu barbaaddi. |
| Lensa wants to be a doctor. | Leensaan doktara ta'uu barbaaddi. |
| She is a doctor. | Isheen doktara dha. |
| Gadise plays volleyball. | Gaaddiseen kubbaa saaphanaa taphatti. |
| Almaz likes playing volleyball. | Almaaz kubbaa saaphanaa taphachuu jaallatti. |
| The students are playing. | Barattootni taphachaa jiru. |
| Getnet has a ball. | Getnet kubbaa qaba. |
| He loves his mother. | Inni haadha isaa jaallata. |
| John is drinking tea. | Yohaannis shaayee dhugaa jira. |
| Simbo is his wife. | Simboon haadha manaa isaati. |

| | |
|---|---|
| Addisu has been married for two years. | Addisuun waggaa lama dura fuudhe. |
| Addisu is an employee. | Addisuun hojjetaa dha. |
| This penalty can only be carried out pursuant to a final judgment rendered by a competent court. | Adabbichi kan raawwatamu mana murtii seeraan dhaabbateen murtiin inni xumuraa yeroo itti kennamu qofa dha. |
| Amnesty, pardon or commutation of the sentence of death may be granted in all cases. | Haala kam keessattiyyuu, dhiifama taasisuun ykn adabbicha du'aa gara adabbii hidhaatti jijjiiruun ni danda'ama. |
| Sentence of death shall not be imposed for crimes committed by persons below eighteen years of age. | Adabni du'aa namoota umriin isaanii waggaa 18 gadi ta'e irratti hin murtaa'u. |
| In particular, no one shall be subjected without his free consent to medical or scientific experimentation. | Keessumatti ammoo nama isa kam irrattiyyuu fedhii isaatiin ala shaakallin wallaansaa fi saayinsawaa taasifamuu hin qabu. |
| Everyone has the right to liberty and security of person. | Namni kamiyyuu walabummaa fi nageenyi dhuunfaa isaa akka mirkanaa'uufiif mirga qaba. |
| No one shall be subjected to arbitrary arrest or detention. | Namni kamiyyuu seeraan ala hin qabamu ykn hin hidhamu. |
| Anyone who has been the victim of unlawful arrest or detention shall have an enforceable right to compensation. | Namni seeraan ala qabame ykn hidhame kamiyyuu beenyaa raaw'ii qabu argachuuf mirga qaba. |
| The right of men and women of marriageable age to marry and to found a family shall be recognized. | Dardaraa fi shamarran umriin gaa'elaaf qaqqabe wal fuudhanii bultii dhaabbachuuf mirga qabu. |
| No marriage shall be entered into without the free and full consent of the intending spouses. | Fedhii fi eyyama warra wal fuudhaniitiin ala gaa'elli hundeeffamuu hin qabu. |
| The committee shall consist of eighteen members. | Koreen kun miseensota 18 kan qabaatu ta'a. |
| The members of the Committee shall be elected and shall serve in their personal capacity. | Miseensonni Koree kanaa kan filatamanii fi mataa isaaniitiin kan tajaajilan ta'u. |
| What is your address? | Teessoon kee eessa dha. |
| Who is your teacher? | Barsiisaan kee eenyu dha. |
| What is your name? | Maqaan kee eenyu dha. |
| The book of the generation of Jesus Christ, the son of David, the son of Abraham. | Hiddi dhaloota Yesuus Kiristoosi isa sanyii Daawiti, sanyii Abraahami ta'e. |
| Abraham begat Isaac. | Abrahaam Yisaaqin dhalfatea. |
| Isaac begat Jacob. | Yisaaq Yaaqoobin dhalfate. |
| Jacob begat Judah and his brethren. | Yaaqoob Yihuudaa fi obboleeyyan isaa dhalfate. |
| Judah begat Pharez and Zerah of Tamar. | Yihuudaan Ti'imaar irraa Faaresii fi Zaaraa dhalfate. |
| Pharez begat Hezron. | Faares Hesronin dhalfate. |
| Hezron begat Ram. | Hesron Araamin dhalfate. |
| Ram begat Amminadab. | Araam Amiinaadaabin dhalfate. |
| Amminadab begat Nahshon. | Amiinaadaab Nahasoonin dhalfate. |
| Nahshon begat Salmon. | Nahasoon Salmoonin dhalfate. |
| Salmon begat Boaz of Rachab. | Salmoon Ra'aab irra Bo'eezin dhalfate. |
| Boaz begat Obed of Ruth. | Bo'eez Ruut irraa Yoobeedin dhalfate. |
| Obed begat Jesse. | Yoobeed Iseeyin dhalfate. |
| Jesse begat David the king. | Iseey Daawit Mooticha dhalfate. |

| | |
|---|---|
| David the king begat Solomon of her that had been the wife of Uriah. | Daawit niitii Oriyooni irraa Solomoonin dhalfate. |
| Solomon begat Rehoboam. | Solomoon Robi'aamin dhalfate. |
| Rehoboam begat Abijah. | Robi'aam Abiyaa dhalfate. |
| Abijah begat Asa. | Abiyaan Asaafin dhalfate. |
| Asa begat Jehoshaphat. | Asaaf Yosaafixin dhalfate. |
| Jehoshaphat begat Jehoram. | Yosaafix Yoraamin dhalfate. |
| Jehoram begat Uzziah. | Yoraam Oziyaa dhalfate. |
| Uzziah begat Jotham. | Oziyaan Yo'ataamin dhalfate. |
| Jotham begat Ahaz. | Yo'ataam Akaazin dhalfate. |
| Ahaz begat Hezekiah. | Akaaz Hisqiyaasin dhalfate. |
| Hezekiah begat Manasseh. | Hisqiyaas Minaasee dhalfate. |
| Manasseh begat Amon. | Minaaseen Amoonin dhalfate. |
| Amon begat Josiah. | Amoon Yoosyaasin dhalfate. |
| Josiah begat Jeconiah and his brethren, about the time they were carried away to Babylon. | Yoosiyaas bara boojuu Baabilooni keessa, Ikooniyaanii fi obboleeyyan isaa dhalfate. |
| after they were brought to Babylon, Jeconiah begat Shealtiel. | Boojuu Baabilooni booddee,Ikooniyaan Salaatiyaalin dhalfate. |
| Shealtiel begat Zerubbabel. | Salaatiyaal Zarubaabelin dhalfate. |
| Zerubbabel begat Abiud. | Zerubaabel Abiyuudin dhalfate. |
| Abiud begat Eliakim. | Abiyuud Eliyaaqeemin dhalfate. |
| Eliakim begat Azor. | Eliyaaqeem Aazaarin dhalfate. |
| Azor begat Zadok. | Aazaar Saadoqin dhalfate. |
| Zadok begat Achim. | Saadoq Akiimin dhalfate. |
| Achim begat Eliud. | Akiim Eliyuudin dhalfate. |
| Eliud begat Eleazar. | Eliyuud Alaazaarin dhalfate. |
| Eleazar begat Matthan. | Alaazaar Maataanin dhalfate. |
| Matthan begat Jacob. | Maataan Yaaqoobin dhalfate. |
| Jacob begat Josep | Yaaqoob Yooseefin dhalfate. |
| Joseph is the husband of Mary. | Yooseef dhiirsa Maariyaamiiti. |
| So all the generations from Abraham to David are fourteen generations. | Egaa walumaa galatti Abrahaamii hanga Daawitiitti dhaloota kudha afuri. |
| From David until the carrying away into Babylon are fourteen generations. | Daawitii hanga boojuu Baabilooniitti dhaloota kudha afuri. |
| From the carrying away into Babylon unto Christ are fourteen generations. | boojuu Baabilooniitii hanga Kiristoosiitti dhaloota kudhana afurtu ture. |
| The birth of Jesus Christ was on this wise. | Dhalachuuni Yesuus Kiristoosi akkana ture. |
| When as his mother Mary was espoused to Joseph. | Haati isaa Maariyaam kaadhima Yooseefi turte. |
| Before they came together, she was found with child of the Holy Ghost. | isheenis otoo Yooseef wajjiin walbira hin gayin hafuura Qulqulluun ulfooftee argamte. |
| Behold, a virgin shall be with child. | Kunoo, dubri tokko ni ulfoofti. |
| They shall call his name Immanuel. | maqaa isaas Amaanu'el jedhu. |
| In those days John the Baptist. | Bara sana Yohaannis Cuuphaani. |
| Blessed are the poor in spirit. | Warri hafuuraan hiyyeeyyii ta'an eebbifamoo dha. |
| Blessed are they that mourn. | Warri gaddanu eebbifamoo dha. |

| | |
|---|---|
| Blessed are the meek. | Warri garraamiin eebbifamoo dha. |
| Blessed are they which do hunger and thirst after righteousness. | Warri qajeellummaa beela'anii fi dheebotan eebbifamoo dha. |
| Blessed are the merciful. | Araar-qabeeyyiin eebbifamoo dha. |
| Blessed are the pure in heart. | Warri garaan isaanii qulqulluu eebbifamoo dha. |
| Blessed are the peacemakers. | Warri nagaya buusani eebbifamoo dha. |
| Blessed are they which are persecuted for righteousness' sake. | Warri qajeelummaadhaaf jedhanii gaargalfaman eebbifamoo dha. |
| After this manner therefore pray. | Egaa akkana jedhaa kadhadhaa. |
| Our Father which art in heaven. | Yaa Abbaa keenya kan samii irra jirtu. |
| Hallowed be thy name. | Maqaan kee haa qulqullaa'uu. |
| Give us this day our daily bread. | Soora keenna kan guyyuma guyyaan nu barbaachisu hardha nu kenni. |
| The light of the body is the eye. | Ibsaani nafaa Ija. |
| Melchizedek was the king of Salem and priest of the most high God. | Malkiseedeq mootii Saaleemiitii fi luba Waaqa guddichaa ture. |
| I will put my laws into their mind. | Ani seera kiyya yaada isaanii keessa nan kaa'a. |
| I will be to them a God. | ani Waaqa isaanii nan ta'a. |
| They shall be to me a people. | isaanis saba kiyya ni ta'u. |
| I will be merciful to their unrighteousness. | Ani balleessaa isaanii nan dhiisaaf. |
| Their sins and their iniquities will I remember no more. | Cubbuu isaaniis deebi'ee hin yaadadhu. |
| Let brotherly love continue. | Jaalalli obbollumaa itti haa fufu. |
| To the saints and faithful brethren in Christ which are at Colossae. | Gara qulqullootaa fi obboleeyyan Kiristoositti amananii Qoloosaayis keessa jiraataniitti. |
| Grace be unto you, and peace, from God our Father and the Lord Jesus Christ. | Abbaa keenya Waaqa biraa ayyaannii fi nageenni isinii haa ta'uu. |
| Wives submit yourselves unto your own husbands. | Yaa niitotaa akka malutti dhiirsota keessaniif ajajamaa. |
| Everyone has the right to freedom of movement and residence within the borders of each State. | Namni kamiyyuu naannoo biyya isaa keessa bilisaan sossohuu fi jiraachuuf mirga ni qaba. |
| Everyone has the right to leave any country, including his own, and to return to his country. | Namni kamiyyuu biyya kamiyyuu keessaa bahu fi biyya isaatti deebi'uuf mirga qaba. |
| Everyone has the right to seek and to enjoy in other countries asylum from persecution. | Namni kamiyyuu miidhaa duraa baqachuu fi biyyoota biroo keessatti irkataa tahee jiraachuuf gaafachuuf mirga ni qaba. |
| Everyone has the right to a nationality. | Namni kamiyyuu mirga lammummaa argachuudhaa ni qaba. |
| No one shall be arbitrarily deprived of his nationality nor denied the right to change his nationality. | Namni kamiyyuu mirgi lammummaa isaa garmalee hin mulqamu akkasumas mirga lamummaa isaa geeddaruus hin dhorkamu. |
| Marriage shall be entered into only with the free and full consent of the intending spouses. | Gaa'elli kan raawwatamu fedhii bilisaa fi guutuu namoota walfuuchuuf fedhanii qofaani. |
| The family is the natural and fundamental group unit of society and is entitled to protection by society and the State. | Maatiin qaama uumamummaatii fi bu'ura hawaasaa waan ta'eef eegumsi hawaasaa fi mootummaa godhamuufii qaba. |
| Everyone has the right to own property alone as | Namni kamiyyuu dhuunfaatis tahee namoota biroo |

| | |
|---|---|
| well as in association with others. | waliin tahee mirga abbaa qabeenyaa tahu ni qaba. |
| No one shall be arbitrarily deprived of his property. | Namni kamiyyuu garmalee qabeenyaa isaa akka dhabu hin taasifamu. |
| The Court may not impose penalties or measures other than those prescribed by law. | Manni Murtichaa kanneen seeraan tumaman malee adabbiilee fi tarkaanfileen biroo murteessuu hin danda'u. |
| The above provisions shall not prevent the Court from interpreting the law. | Tumaaleen armaan olitti tuqaman Manna Murtiicha seericha hiikuu hin dhorkani. |
| I want to read Article three. | Anni Keewwata saddii dubbisuun barbaada. |
| The Ethiopian flag shall consist of green at the top, yellow in the middle and red at the bottom. | Alaabaan itiyoophiyaa gara irraa magariisa gidduun keelloo jalaan diimaa ta'ee ni qabaata. |
| The three colors shall be set horizontally in equal dimension. | Bifti sadanuu walqixa ta'anii dalgaan taa'u. |
| Members of the Federation may have their respective flags and emblems. | Miseensonni Federaalawaa alaabaa fi aasaxaa mata-mataa isaanii qabaachuu ni danda'u. |
| They shall determine the details thereof through their respective legislatures. | Tarreeffama isaanii Mana marii mata-mataa isaaniitiin murteeffatu. |
| City Court means a court established to decide on city-related cases. | Mana Murtii Dhimmoota Magaalaa jechuun qaama dhimmoota magaalaa ilaallatanirratti aangoo abbaa seerummaa qabaatee dhaabbate jechuu dha. |
| Constitution means the Constitution of Oromia National Regional State. | Heera jechuun Heera Mootummaa Naannoo Oromiyaa jechuu dha. |
| Mayor means a chief executive officer of a city. | Kantiibaa jechuun abbaa aangoo olaanaa qaama raawwachiisaa magaalaa jechuu dha. |
| City Manager means an official who executes municipal services in the city. | Hojii Adeemsisaa Magaalaa jechuun raawwachiisaa tajaajiloota mana qopheessaa magaalichaa ta'ee kan hojjetu jechuu dha. |
| Bureau means the Regional Office of industy and Urban Development. | Biiroo jechuun Biiroo Industrii fi Misooma Magaalaa Naannichaa jechuu dha. |
| This Proclamation shall apply to incorporated cities in Oromia Regional state | Labsiin kun Mootummaa Naannoo Oromiyaa keessatti magaalota qaamni seerummaa kennameef irratti raawwatamummaa ni qabaata. |
| An urban local government in Oromia shall have power over local issues. | Bulchiinsi naannoo magaalaa Naannoo Oromiyaa keessatti argamu dhimma magaalichaa irratti aangoo guutuu ni qaba. |
| The urban Government Model applicable in the Regional Government shall be the Council-Mayor systems. | Sirni gaggeessaa magaalaa, magaalota Mootummaa Naannichaa keessatti hojiirra oolu sirna Mana Marii-kantibaa jedhamu dha. |
| The speaker of the City Council shall be accountable to the City council. | Itti waamamni af-yaa'ii Mana Marii Magaalaa mana Marii magaalichaaf ta'a. |
| The speaker of the City council shall have the tenn of years of the city council. | Barri hojii af-yaa'ii mana marichaa bara hojii mana marichaa ta'a. |
| City Court means a court established to decide on city-related cases. | Mana Murtii Dhimmoota Magaalaa jechuun qaama dhimmoota magaalaa ilaallatanirratti aangoo abbaa seerummaa qabaatee dhaabbate jechuu dha. |

| | |
|---|---|
| Constitution means the Constitution of Oromia National Regional State. | Heera jechuun Heera Mootummaa Naannoo Oromiyaa jechuu dha. |
| Mayor means a chief executive officer of a city. | Kantiibaa jechuun abbaa aangoo olaanaa qaama raawwachiisaa magaalaa jechuu dha. |
| City Manager means an official who executes municipal services in the city; | Hojii Adeemsisaa Magaalaa jechuun raawwachiisaa tajaajiloota mana qopheessaa magaalichaa ta'ee kan hojjetu jechuu dha. |
| Bureau means of industy and Urban Development the Regional Office. | Biiroo jechuun Biiroo Industrii fi Misooma Magaalaa Naannichaa jechuudha. |
| This Proclamation shall apply to incorporated cities in Oromia Regional state | Labsiin kun Mootummaa Naannoo Oromiyaa keessatti magaalota qaamni seerummaa kennameef irratti raawwatamummaa ni qabaata. |
| An urban local government in Oromia shall have power over local issues. | Bulchiinsi naannoo magaalaa Naannoo Oromiyaa keessatti argamu dhimma magaalichaa irratti aangoo guutuu ni qaba. |
| The urban Government Model applicable in the Regional Government shall be the Council-Mayor systems. | Sirni gaggeessaa magaalaa, magaalota Mootummaa Naannichaa keessatti hojiirra oolu sirna Mana Marii-kantibaa jedhamu dha. |
| The speaker of the City Council shall be accountable to the City council. | Itti waamamni af-yaa'ii Mana Marii Magaalaa mana Marii magaalichaaf ta'a. |
| The speaker of the City council shall have the tenn of years of the city council. | Barri hojii af-yaa'ii mana marichaa bara hojii mana marichaa ta'a. |
| The baby woke up when the doorbell rang. | Mucaan bilbilli balbalaa yeroo bilbilame hirribaa ka'e. |
| I can't help you if you can't tell me what's wrong. | rakkoo jirru yoo natti himte malee anni si hin gargaaru. |
| I will pay you back as soon as I get the money. | akkuman qarshii argadheen deebisee siif kaffala. |
| Before he was a famous writer, Abdisa was a maintenance man. | uttuu bareessaa beekkamaa hin ta'iin dura, Abdiisaan hojjii suphaa hojjeta ture. |
| After she graduates this year, chaltu will work in her father’s company. | Erga barana eebifamtee booda, caaltuun dhaabbata abbaa ishee keessa hojjetti. |

## Appendix II: Sample Parallel Corpus for Testing

| | |
|---|---|
| He is sitting on the left but she is sitting on the right. | Inni Karaa bitaa taa'aa jira garuu isheen karaa mirgaa taa'aa jirti. |
| I am sitting behind her but she is sitting in front of TV. | Anni duuba ishee taa'aan jira garuu isheen fuula dura Tv taa'aa jirti. |
| Bontu wants to read a book. | Boontuun Kitaaba dubbisuu barbaadi. |
| We have Human and Democratic Rights. | Nuyyi Mirgoota Namummaa fi Dimookiraasummaa qabna. |
| Human rights and freedoms, emanating from the nature of mankind, are inviolable and inalienable. | Mirgoonni Namummaa fi Bilisummaa, uumama dhala namaa irraa kan maddan, kan hin cabnee fi hin mulqamne dha. |
| Human and democratic rights of citizens and peoples shall be respected. | Mirgoonni Namummaa fi dimookiraasummaa Lammiiwwanii fi Uummattootaa ni kabajamu. |
| Degaga is driking coffee, but chaltu is eating food. | Dagaagaan buna dhugaa jira garuu caaltuun nyaata nyaachaa jirti. |
| Derara opened the window. | Daraaraan foddicha baneera. |
| Bona is handsome. | Bonaan bareedaa dha. |
| Tola is a student but she is a teacher. | Tolaan barataa dha garuu isheen barattuu dha. |
| My son is taking an exam. | mucaan koo qormaata fudhachaa jira. |
| She has Ethiopian nationality. | isheen lammummaa Ittiyoophiyaa qabdi. |
| Chaltu is getting married. | Caaltuun heerumuuf jetti. |
| Haile wants to be a teacher. | Hayileen barsiisaa ta'uu barbaada. |
| My sister has a car. | Obboleetiin koo konkolaataa qabdi. |
| Kaku is going to school because she is a student. | Kakuun gara mana baruumsaa deemaa jirti sababiin isaa isheen baratuu dha. |
| Dagim is a doctor and he likes patients. | Daagim doktora dha kanaafuu inni dhukkubsataa jaallata. |
| She is sick she must see a doctor. | Isheen dhukkubsateeti doktora ilaaluu qabdi. |
| Everyone has a right to go to school. | Namni kamiyyuu gara mana baruumsaa deemuuf mirga qaba. |
| She likes Ethiopian flag. | isheen alaabaa Ittiyoophiyaa jaallati. |
| Her brother is an architect. | obboleesi ishee arkiteektii dha. |
| Degaga has the right to liberty and security of person. | Dagaagaan walabummaa fi nageenyi dhuunfaa isaa akka mirkanaa'uufiif mirga qaba. |
| kalkidan is sleeping because she is tired. | Qaalkidaan rafaa jirti sababiin isaa isheen dadhabdeeti. |
| Chala wants to be a teacher and a doctor. | Caalaan barsiisaa fi doktora ta'uu barbaada. |
| kenaf is not a teacher neither a doctor. | keenaaf barsiistuus doktoras mitti. |
| My father bought a pen. | abbaan koo peennaa bite. |
| Derara is crying because he lost his sister. | Daraaraan boo'aa jira sababiin isaa inni obboleetii isaa dhabee. |
| Dame went to school yesterday. | Dameen kaleesa gara mana baruumsaa deemte. |
| Bob is sweating because he was playing a football. | Boob dafqaa jira sababiin isaa inni kubbaa miilaa taphachaa ture. |
| He wants to write International Covenant On Civil and Political Rights. | Inni waadaa mirgoota siviilii fi siyaasaa idil-addunyaa bareesuu barbaada. |

## Appendix III: Sample language model for Afaan Oromo

iARPA

\data\

ngram 1= 2219

ngram 2= 4847

ngram 3= 3124

\1-grams:

-4.018929 <s> -0.532639

-0.964699 </s> 0.000000

-4.018929 Huseen -0.374390

-3.541808 muka -0.266669

-3.620989 soofaa -0.243450

-2.360918 jira -0.993573

-2.124060 , -0.301732

-3.365717 garuu -0.243450

-3.620989 faaxumaan -0.243450

-3.620989 calluma -0.243450

-3.620989 jettee -0.243450

-2.977537 taa' -0.826935

-2.603956 aa -0.471992

-2.564084 jirti -1.088548

-1.148818 . -2.738547

-3.620989 waa' -0.243450

-3.278567 een -0.190352

-2.360918 isaa -0.346727

-3.620989 homtuu -0.243450

-1.903319 hin -0.366666

-3.541808 dhaga' -0.266669

-3.620989 amne -0.243450

-3.143868 haa -0.441801

-2.493884 ta' -0.477830

-3.474861 u -0.419541

-3.173831 malee -0.223640

-3.240778 anii -0.204208

-3.416869 abdii -0.185458

-2.788480 koo -0.243400

-2.814809 itti -0.290364

-3.620989 fufeen -0.243450

-3.018929 Nuyi -0.238599

-3.620989 qilleensa -0.243450

-3.620989 fudhachuuf -0.243450

-3.620989 baanee -0.243450

-3.620989 achumaan -0.243450

-3.620989 firoota -0.243450

-3.620989 dubisne -0.243450

-1.893448 anni -0.733341

-3.620989 Boontuu -0.243450

-2.657201 yeroo -0.274751

-3.620989 dhihooti -0.243450

-3.620989 argine -0.243450

-3.018929 kana -0.224972

## Appendix IV: Sample Language model for English

iARPA

\data\

ngram 1= 1604

ngram 2= 4412

ngram 3= 3388

\1-grams:

-4.043951 <s> -0.681400

-0.989913 </s> 0.000000

-4.043951 Husen -0.394109

-1.688883 is -0.498126

-3.646011 sawing -0.278049

-1.896275 a -0.377372

-3.441891 wood -0.220057

-2.284284 , -0.413608

-3.390739 but -0.278049

-3.646011 Fatuma -0.278049

-3.043951 sitting -0.255773

-3.646011 idle -0.278049

-1.013352 . -3.007416

-2.930008 there -0.351527

-2.525437 no -0.727479

-3.231038 news -0.251193

-3.089709 about -0.249499

-2.983254 him -0.428837

-3.265800 however -0.286779

-1.830434 I -0.797983

-3.198853 went -0.285955

-2.376498 on -0.780395

-3.646011 hoping -0.278049

-2.839831 we -0.415138

-2.897823 out -0.351838

-2.447354 for -0.377874

-3.646011 picnic -0.278049

-1.977625 and -0.351196

-3.646011 visited -0.278049

-3.441891 some -0.245397

-3.646011 relatives -0.278049

-3.043951 our -0.265732

-3.646011 way -0.278049

-3.646011 haven -0.278049

-3.114532 't -0.237591

-3.499883 seen -0.313398

-3.066228 Bontu -0.554184

-3.646011 lately -0.278049

-3.265800 won -0.416352

-3.303589 see -0.232994

-2.742921 her -0.359330

-3.646011 again -0.278049

-3.646011 Mr. -0.278049

-3.303589 Derara -0.245387