HAWASSA UNIVERSITY

THE INSTITUTE OF TECHNOLOGY, FACULTY OF INFORMATICS DEPARTMENT OF COMPUTER SCIENCE

Bi-Directional Sidaamu Afoo - Amharic Statistical Machine Translation

A Thesis submitted to the School of Graduate Studies of Hawassa University in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science

By: Kebebush Kamiso

Advisor: Tesfaye Bayu (PhD)

Co-Advisor: Andargachew Mekonnen (PhD)

Hawassa, Ethiopia

December, 2023

APPROVAL SHEET-I

This is to certify that the thesis entitled, " **Bi-Directional Sidaamu Afoo - Amharic Statistical Machine Translation**" submitted in partial fulfillment of the requirements for the degree of Master's with specialization in Computer Science, the Graduate Program of the Faculty of Informatics, and has been carried out by **Kebebush Kamiso**. Therefore we recommend that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

____________________________	__________________	__________________
Name of major advisor	Signature	Date

____________________________	__________________	__________________
Name of co-advisor	Signature	Date

EXAMINERS' APPROVAL SHEET

SCHOOL OF GRADUATE STUDIES

HAWASSA UNIVERSITY EXAMINERS' APPROVAL SHEET

(Submission Sheet-2)

We, the undersigned, members of the Board of Examiners of the final open defense by **Kebebush Kamiso** have read and evaluated his/her thesis entitled "**Bi-Directional Sidaamu Afoo - Amharic Statistical Machine Translation**", and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirements for the degree.

Name	Signature	Date
____________________	______________	__________
Name of Major Advisor	Signature	Date
____________________	______________	__________
Name of Internal Examiner-I	Signature	Date
____________________	______________	__________
Name of Internal Examiner-II	Signature	Date
____________________	______________	__________
Name of External Examiner	Signature	Date
____________________	______________	__________
SGS Approval	Signature	Date

Final approval and acceptance of the thesis is contingent upon the submission of the final copy of the thesis to the School of Graduate Studies (SGS) through the Department/School Graduate Committee (DGC/SGC) of the candidate's department.

Stamp of SGS Date: ____________

Declaration

I declare that this research is my original work and has not been presented for a degree in any University and that all sources of material used for the research have been properly acknowledged.

Declared by:

Name	Signature	Date
Kebebush Kamiso	________________	________________

This research has been submitted for Examination with my approval as University advisor.

Name Advisor	Signature	Date
Tesfaye Bayu (PHD	________________	________________
Co-Advisor	Signature	Date
Andargachew Mekonnen (PHD)	________________	________________

Hawassa, Ethiopia

December, 2023

Abstract

Machine translation (MT) is the area of Natural Language Processing (NLP) that focuses on obtaining a target language text from a source language text using automatic techniques. It is a multidisciplinary field and the challenge has been approached from various points of view including linguistics and statistics.

MT usually involves one or more approaches. Our preference for this study is to develop the bi-directional Sidaamu Afoo - Amharic machine translation system, make use of a statistical machine translation (SMT) approach.

To conduct the experiment, a parallel corpus was collected from all possible available sources. These include mostly the Old and New Testaments of the Holy Bible for both languages. We used the monolingual Contemporary Amharic Corpus and the Sidama Afoo corpus compiled by a research team in the Informatics Faculty of Hawassa University. Different preprocessing tasks such as tokenization, cleaning, and normalization have been done to make the corpus suitable for the system.

To accomplish the objective of this thesis work, we conducted four experiments using word and morpheme-based translation units with SMT for Sidaamu Afoo - Amharic language pairs.

The first two experiments focus on word-based SMT and the next two on morpheme-based translation using unsupervised morphological segmentation tool; Morfessor. For each experiment, we used 30,100 parallel sentences. Out of the total parallel sentences, we used 80% (24,100) of randomly selected parallel sentences for training, 10% (3,000) for tuning and another 10% (3,000) for testing.

The basic tools used for accomplishing the machine translation are Moses for the translation process which is MGIZA ++ for word and morpheme alignment and KenLM for language modeling; Morfessor for morphological segmentation. For evaluation SacreBLEU package which are BLEU, ChrF and TER metrics.

According to the experimental findings, the differences between Amharic to Sidaamu Afoo and Sidaamu Afoo to Amharic in the Word-based alignment translation were 6.2, 16, and 1.9 for BLUE, ChrF2, and TER, respectively. In the Morpheme-based alignment, the differences between Amharic to Sidaamu Afoo and Sidaamu Afoo to Amharic translation were 7.5, 20.4, and 5.1, for BLUE, ChrF2, and TER respectively.

In conclusion, the results show that morpheme-based alignment performance is better than word-based alignment, for Amharic to Sidaamu Afoo than Sidaamu Afoo to Amharic.

Keywords: SMT, morpheme level alignment, word level alignment, morfessor

Acknowledgment

First and foremost, I would like to express my heartfelt thanks to Lord JESUS, who gave me the opportunity, endurance, strength, and wisdom to bring this thesis to completion.

Secondly, I am highly grateful to my research advisors; Dr. Tesfaye Bayu and Dr. Andargachew Mekonnen for their expertise, generous consideration, and giving me vital assistance in finishing this research. Dr. Andargachew`s spport was unforgateable ever.

Thirdly, I would like to express my gratitude to Ato Kitawu Ayele for his kindness in supporting me by shaping the structure of my document.

Fourthly, I would like to thank all those who helped me through the progress of this study, especially Dr. Degif, PG coordinator and Ato Wogene, Head department also to staff members W/ro Sisay, W/t Ruhama, Ato Solomon and Ato Adane for their cooperative responses during the time I was in need of their help.

I also like to express my gratitude to the research team at Hawassa University's Informatics Faculty for their cooperation throughout the collection of the monolingual corpus.

Lastly, I am also thankful to my family especially my husband Techane Roda, my brother Berhanu Bekele and my children for their patience, love, encouragement, and moral support in completing this thesis.

Finally, I am also grateful to my friends and colleagues, especially W/ro Misgana Elias and Ato Biniam Bekele for their support and encouragement.

Table of contents

List of Figures

List of Tables

List of Acronyms

BLEU - Bilingual Evaluation Understudy

ALPAC - Automatic Language Processing Advisory Committee

EBMT - Example-Based Machine Translation

MT - Machine Translation

NLP - Natural Language Processing

NMT - Neural machine translation

RBMT - Rule-Based Machine Translation

SOV - Subject-Object-Verb

SNRS - Sidaama National Regional State

SMT- Statistical Machine Translation

SL - Source Language

TL - Target Language

LT - Language Technology

AI - Artificial Intelligence

ATG - Analysis-Transfer-Generation

LK - Linguistic Knowledge

TER - Term Error Rate

IBM - Individual Based Model

WMT - Wechsler Memory Test

DMT - Direct Machine Translation

CV - Cross-Validation

CHAPTER ONE

1. Introduction

1.1 Background

A field of computer science which is concerned with the interactions between human (natural) languages and computers is referred to as Natural language processing (NLP). Further, it is defined as ways for computers to analyze, understand, and derive meaning from human language in a smart and useful way. To put it in other way, NLP is the process of teaching computers to understand the way humans learn and use languages. Among other technological attempts undergoing in NLP is Machine translation (MT) which is one area of language technology (LT). It is the field of automatically translating some text or speech from a source language (SL) to a target language (TL) (Brandt, 2011). I.e. it is the attempt to automate all or part of the process of translating from one human language to another. Thus, “Machine translation is the application of computers to the task of translating text or speech from one natural (human) language to another natural (human) language” (Brandt, 2011).

Machine translation has different advantages; the first one is Confidentiality, since people use machine translation systems to translate their private information, people communicate only with the system (MT) rather than other individuals, as a result, the privacy of the individuals is protected. The second advantage is fast translation. By using a machine translation system it is possible to save time while translating large texts like paragraphs and even documents in a short period of time. The third one is universality. Usually, a human translator translates the meaning of the text in their own context. This may bias the meaning of the text; but, in the case of machine translation a text will be translated with the same meaning anywhere and everywhere, this makes machine translation universal (Teshome, E, 2013).

Machine translation systems are based on their core methodology can be classified into two paradigms: the rule-based approach and the corpus-based approach (Okpor, 2014). In the rule-based approach, human experts specify a set of rules to describe the translation process, so an enormous amount of input from human experts is required. On the other hand, under the corpus-based approach, the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Combining the features of the two major classifications of machine translation systems gave birth to the hybrid machine translation approach, which has better efficiency in MT systems (Okpor, 2014). One of these or Hybrid approaches can be used for the purpose in need depending on the resource availability like (data, money, time, expertise, etc…) available. However, because of the above resource limitations, we were obliged to choose only **statistical machine translation approach** for this study. So, a bi-directional Sidaamu Afoo-Amharic Statistical machine translation system was chosen to be designed.

1.2 Motivation

Ethiopia is a country that entails different nations and nationalities with over 80 distinct languages. Of all the languages being spoken, Amharic is one of a widely spoken language all over the country.

It is a Semitic language mostly spoken in North Central Ethiopia. It is also used as the "official" language of the Federal Democratic Republic of Ethiopia (FDRE) and thus, has official status nationwide. It is also official working language for some of regional states within the federal system; including the Amhara region the place where the language come from and multi-ethnic Southern regional states and may be some others (Gambela, Beneshangul etc.) of the country (Teshome, E, 2013). Sidaama language (also known as sidaamu afoo) is one of the major languages spoken in Ethiopia. It is a Highland East Cushitic language spoken in the south-central part of the country. It is also sometimes called Sidamigna or Sidamic (in English, using the ic ending of "Amharic") (Kawachi, 2007). Currently, it is the working language of the Sidaama Regional State. More recent estimates the number of speakers of Sidaamu afoo is over 5 million (Kawachi, 2007). However, there has been relatively little research on the language.

The lack of natural language processing tools that understand Sidaamu Afoo text-Amharic text and vice versa in online machine translation is a major problem for many people who use languages as their means of information processing and usage. Further, this problem extends to government and non-government organizations, schooling systems, Judiciary services, etc. Thus, there exists an outstanding issue of standardized and automatic translation need that requires tecnology to solve ambiguity among the users. Therefore; I am initiated to contribute my effort and hoping this start could motivate other researchers in the field for further work in this regard.

1.3 Statement of the problem

Sidaamu Afoo is one of the major languages spoken in Ethiopia. It is a working language in government offices and courts for the Sidaama regional state that involves more than 5 million populations (HUMMEL, S., & PANINI, F, 2020). As well as it is also used as a medium of instruction in primary schools and it is also conducted as a language at secondary and tertiary levels at large in the region. Also day by day regional work performance reports, regional work performance reports, feedback, guidelines, and Court decisions should be conducted using Sidaamu Afoo; whenever and wherever depending on the situation at the federal government level or otherwise in different organistions, these works may be obviously required to be translated in Amharic language. Besides these, there is an acute shortage of reference and reading materials written in Sidaamu Afoo in the schools of the region, whereas the Amharic language is reach in this regard. So, due to the aforementioned multiple reasons, there is an outstanding bi-directional machine translation need in all parties (Educators, Governmental and non-governmental officials, advocacy workers, courts, and citizens who don’t understand both languages, etc.)

Therefore, the main issue factor that initiated this study to be carried out is the fact that an application that translates Sidaamu Afoo texts into Amharic and Amharic texts into Sidaamu Afoo is not adequately available as long as the researcher's knowledge is concerned, or there is no prior full-fledged study conducted a bi-directional Sidaamu Afoo to Amharic using Statistical machine translation system at hand.

For various reasons, the translation of bi-directional Sidaamu Afoo - Amharic texts become a necessity that is why this study was intiated. The bi-directional Sidaamu Afoo to Amharic Statistical machine translation system likely assists humans to accomplish an efficient translation with respect to time and resource.

To this end, this study attempted to answer the following research questions:

1. How to develop a parallel corpus for Sidaamu Afoo -Amharic machine translation system?
2. How to design the architecture of the system and test the performance of the system?

1.4 Objectives

1.4.1 General Objective

The general objective of this research is to design and develop a bi-directional Sidaamu Afoo - Amharic statistical machine translation system.

1.4.2 Specific Objectives

The specific objectives of this study are to:

- Review syntactic structure and relationship of the language pair;
- Collect Sidaamu Afoo - Amharic bilingual parallel corpus.
- Designe architecture for bi-directional Sidaamu Afoo – Amharic machine translation using statistical machine approach.
- Test the performance of the system both ways that is from Sidaamu Afoo to Amharic as well as from Amharic to Sidaamu Afoo and,
- To report the findings of the study and recommend for the upcoming research area.

1.5 Scope and limitation of the study

The bi-directional Sidaamu Afoo – Amharic machine translation is designed to translate text to text translation from the source language to the target language.

There are different limitations faced during the process of conducting this research. The first and the most challenge was the lack of a standardized multidisciplinary bilingual corpus for training and testing. This is due to the absence of a sufficient amount of digitally available documents and data in Sidaamu Afoo; other than the religious one. So we were obliged to test the performance of the system mostly using the religious domain. The other one is the study was so limited on word and morpheme alignment uints only, it does not include all types of units. Thus, the experimental result is presented as it is, i.e. result analysis was not done and this could be considered as a limitation.

1.6 Significance of the study

Machine translation has a great role in exchanging information among different languages around the world; the Machine translation rate is faster than a human translator (Solomon, Y, 2017).

The significance of this research is to develop machine translation software for Sidaamu Afoo to Amharic and vice versa. It is possible to address information and solve language barriers between individuals to read and understand different publications, assist medium of instruction in primary schools, secondary and tertiary level. So far it can be used to interpret work performance reports, feedback, guidelines and Court decisions. Also, it contributes to future research and development regarding Sidaamu Afoo–Amharic language pairs used as an additional component in the area of natural language processing; specifically in machine translation, Information retrieval, speech processing and text processing.

1.7 Methodology of the Study

1.7.1 Literature review

Research methodology is a way to systematically solve the research problem (Kothari C, 2004)**.** In order to achieve the general and specific objective of this study, different methodologies were employed. In this study, a detailed literature review has been done on machine translation on different language pairs. Published text documents, thesis, conference, books, and journal articles were also reviewed to explore the principles, methods, techniques, and tools employed.

Additionally, a conversation was under taken wisely with Amharic and Sidaamu Afoo language experts about the linguistic characteristics, grammatical structures and morphologies.

Furthermore, the different algorithms used in implementing them were studied carefully and the syntactic relationship between Sidaamu Afoo and Amharic languages has been reviewed.

1.7.2 Research design

To conduct the research, we followed experimental research design to explore words and morphemes level alignment on Statistical machine translation. Different experiments are - conducted for better performance of SMT. Experimental research is a scientific approach where one or more independent variables are manipulated and applied to one or more dependent variables to measure their effect on the latter. It includes a hypothesis, a variable that can be manipulated by the researcher, and variables that can be measured, calculated and compared (Reiter, E, 2018).

1.7.3 Data Source

To perform the experiments on corpus-based statistical machine translation, a parallel corpus was collected from all possible available sources. Such sources include the Old and New Testaments of the Holy bible for both languages and we used the monolingual Contemporary Amharic Corpus (Gezmu, et al, 2018) and Sidaamu Afoo corpus compiled by a research team in the Informatics Faculty of Hawassa University. The Size of the corpus for the experiment we collected 30,100 parallel sentences for each Sidaamu Afoo and Amharic languages training dataset, which were used for training the translation model. From 30,100 collected sentences, we used 3,000 for testing and 3,000 for tuning purposes. We used 24,100 parallel sentences for the language model for both the Sidaamu afoo and Amharic languages.

1.7.4 Tools

To develop statistical machine translation various tools are available. The basic tools used for accomplishing the machine translation are:

- Ubuntu 20.04: an operating system which is freely available and suitable for the Moses environment.
- Moses:
  - MGIZA ++: for word and morpheme alignment.
  - KenLM: for language modeling.
- Python programming language: is used as a tool for preprocessing in the Ubuntu operating system which is suitable for Moses environment.
- Unsupervised morpheme segmentation tool Morfessor 2.0: is used for morphological segmentation.
- SacréBLEU package: BLEU, ChrF and TER metrics: For evaluation method. It provides hassle-free computation of shareable, comparable, and reproducible BLEU scores.

1.7.5 Evaluation

Machine translation evaluation has been done by using manual or automatic evaluation methods. Manual evaluation gives a better result in order to measure the quality of translation and to analyze the errors within the system output. But the most challenging issues in conducting human evaluation are high costs and time consuming. Therefore we used automatic methods like Bilingual Evaluation Understudy (BLEU) score metrics to evaluate the performance of machine translation system.

The basic idea behind BLEU is, if the machine translation output is closer to human translation output; then it is considered as better translation (Post, 2018). BLEU was one of the metrics to achieve a high correlation with reference translation and remains one of the most popular automated and inexpensive metrics used in different researches for evaluation purpose. However, BLUE score metrics evaluation has major drawbacks, especially when applied to tasks that it was never intended to evaluate. To compensate for the limitations of BLEU (Reiter, E, 2018). We used SacreBLEU package metrics. SacreBLEU (Post, 2018) provides hassle-free computation of shareable, comparable, and reproducible BLEU scores. It supports Multi-reference Evaluation: BLEU, chrF and TER. Using Multi-reference Evaluation (Post, 2018) helps to regulate the BLUE score results.

1.8 Thesis organization

This thesis is organized into six chapters:

The first chapter gives a general overview of the whole thesis which describes the background of the research, the statement of the problem, the objectives of the research, the scope and limitation of the study, the methods used, and the organization of the thesis.

The second chapter reviews different literatures regarding Machine Translation together with its different approaches with a special focus on Statistical Machine Translation. The chapter covers the components of the SMT in detail and reviews related works.

The third chapter deals with an overview of the two languages, the relationship and difference between them and discusses the Morphology of Sidaamu Afoo and Amharic languages.

The fourth Chapter presents a detailed description of the design and development of the Sidaamu Afoo - Amharic language model including, an overview of the development of the MT system, Corpus Collection, Preparation and Architecture of the System.

The fifth Chapter deals with the experimentation of the study which includes different experiments, discussion of results, interpretation and findings.

The last, chapter presents conclusions and recommendations based on the findings of the study. This chapter also incites to future works. At the end references and Appendices are attached to the document.

CHAPTER TWO

2. Literature Review

2.1 Introduction

In this chapter, we are going to briefly discuss an overview of background information about Machine translation and its approaches such as Statistical Machine Translation (SMT), Rule-Based Machine Translation (RBMT), Example-Based Machine Translation (EBMT), Hybrid Machine Translation (HMT), and Neural machine translation (NMT) in detail. We also look at related works of some languages that have been done with machine translation using different approaches and methodologies.

2.2 Machine translation

Machine translation (MT) can be defined as the translation of information from one natural language (source language) to another language (target language) using computerized systems; automatic or semi-automatic (Chéragui, 2012). It is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. Due to the advent of computers and the internet, the world is becoming together to one (Okpor, 2014). Thus, the knowledge, culture, tradition, history, religious, and philosophy documents of one country language can be translated to another language and the rest of the world through Machine translation. To create a paperless working environment translation plays a great role and to make accessible the document of one language in another language. Sharing of Knowledge is also possible besides facilitating easy communication. No more being a language barrier for Communications in any way.

A language is used for conveying information or broadcasting the information. Stepping into the modern digital age, language as the information carrier has become the most significant means for a human to communicate. It has been considered as the barrier of communication between people from different countries and between people who speak different languages within the same country. The problem of converting one language into another quickly and efficiently has become a problem of common concern for humanity (Li, P, 2013). MT is an automatic translation of one language into one or more languages (in the case of multilingual) by means of a computer. High-quality translation requires a thorough understanding of the source text and its intended function as well as good knowledge of the target language. Translation itself is a challenging task for humans and is no less challenging for computers because it deals with natural languages (Mahata, S. K., Das, D., & Bandyopadhyay, S, 2019).

2.3 History of Machine Translation

The history of machine translation is traced from the pioneers/developer and early systems of the 1950s and 1960s, the impact of the Automatic Language Processing Advisory Committee (ALPAC) report in the mid-1960s, the revival in the 1970s, the appearance of commercial and operational systems in the 1980s, research during the 1980s, new developments in research in the 1990s, and the growing use of systems in the past decade (Slocum, J, 1985). These resulted in the birth of modern Machine translation.

2.4 Approaches of Machine Translation

Machine translation systems, based on their core methodology can be classified into two paradigms: the rule-based approach and the corpus-based approach (Okpor, 2014). In the rule-based approach, human experts specify a set of rules to describe the translation process, so an enormous amount of input from human experts is required. On the other hand, under the corpus-based approach, the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Combining the features of the two major classifications of machine translation systems gave birth to the hybrid machine translation approach (Okpor, 2014). Each of the machine translation approaches are explained below:

2.4.1 Rule-Based Machine Translation Approach (RBMT)

Rule-Based Machine Translation (RBMT), also known as Knowledge-Based Machine Translation, is a general term that describes machine translation systems based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively (Okpor, 2014). Having input sentences, an RBMT system generates them to output sentences on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a real translation task. RBMT methodology applies a set of linguistic rules in three different phases: (Okpor, 2014)analysis, transfer, and generation. Therefore, a rule-based system requires syntax analysis, semantic analysis, syntax generation, and semantic generation as shown in Figure 2.1 below:

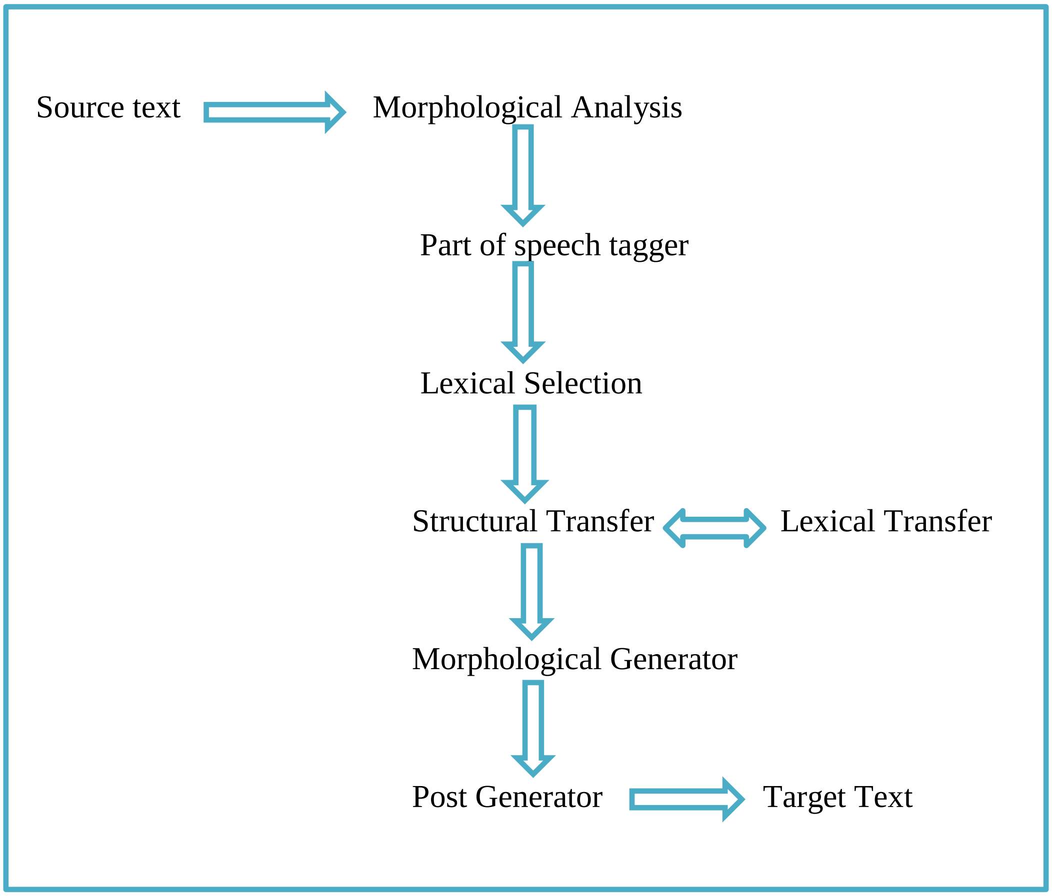


Figure 2-1: Architecture of RBMT Approaches

Based on the intermediate representation used this approach is further classified into the following approaches [19]. Direct approach, Interlingua approach and Transfer-Based MT approaches.

2.4.1.1 Direct approach

Direct machine translation approach (DMT): Starting with the shallowest level at the bottom of the pyramid is the Direct Machine Translation Approach (Okpor, 2014) (Daba, J. & Assabie, Y, 2014). DMT approach is the oldest and less popular approach. Direct translation is made at the word level and MT systems that use this approach are capable of translating a language, called source language (SL) directly to another language, called target language (TL). Words of the SL are translated without passing through an additional/intermediary representation. The analysis of SL texts is oriented to only one TL. Direct translation systems are basically bilingual and unidirectional. The direct translation approach needs only a little syntactic and semantic analysis. SL analysis is oriented specifically to the production of representations appropriate for one particular TL. DMT is a word-by-word translation approach with some simple grammatical adjustments. Figure 2.2 shows the architecture of direct approach.

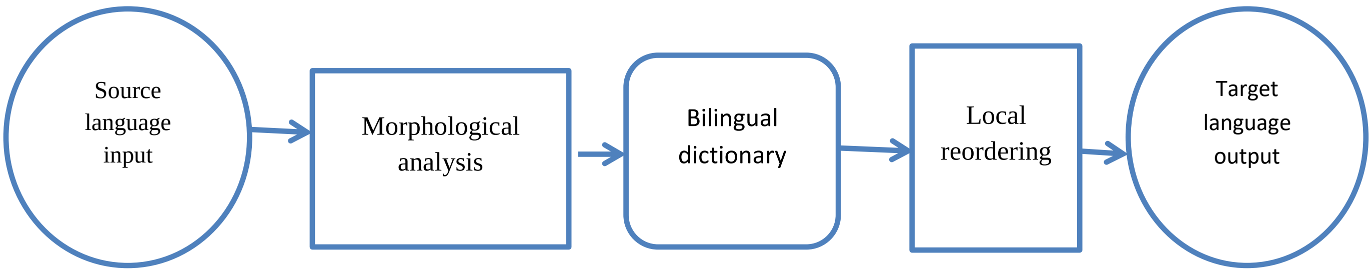


Figure 2.2: Direct machine translation (Akubazgi, G, 2017)

2.4.1.2 Interlingua-based Machine Translation

The translation is based on representing the source language text in an intermediary form, Interlingua (Mulugeta, D, 2015). The idea is to represent all sentences that tell the same thing in the same way, independent of any language. This approach translates by performing deep semantic analysis on the X input language into the Interlingua representation and providing translation to target language Y from that intermediate representation. It involves analysis and generation: analysis helps to derive an Interlingua representation.

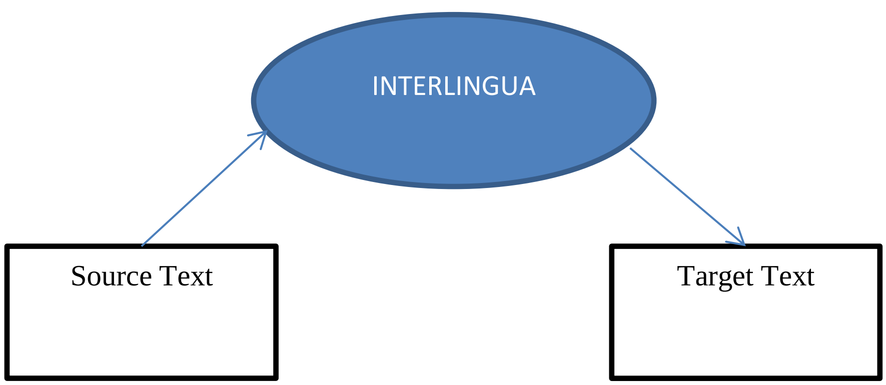


Figure 2.3: Interlingua Machine translation

The source language or the sentence to be translated is transformed into an Interlingua which is an abstract language-independent representation. Then the target language text is generated from that internal representation. This representation allows analyzers and generators to be written by monolingual system developers and handles very different languages from each other but it's applicable for a specific domain not for a wider domain (Syahrina, A, 2011) (Antony P, 2013).

Interlingua-based MT approach is the most attractive form of the rule-based approach since it works fine regardless of any language and it permits translation from and into the same language. Its drawbacks: hard to define Interlingua and it fails to take the similarities between languages.

2.4.1.3 Transfer based Machine Translation

The Transfer approach is also called the indirect Approach or the Linguistic Knowledge (LK) translation] (Dubey, P, 2013). It is the second generation of machine translation. In this approach, the source language is transformed into a less language-specific representation, and an equivalent representation is generated for the target language using bilingual dictionaries and grammar rules. The system uses an intermediate representation that captures the structure of the original text to generate the correct translation. Transfer transfer-based machine translation process involves three Phases: analysis, transfer, and generation (Irfan, M, 2017). Transfer-based approaches need rules for - syntactic transfer, Semantic transfer, and lexical transfer Syntactic transfer rules – will tell us how to modify the source parse tree to resemble the target parse tree.

- Semantic transfer – using semantic role labeling.
- Lexical transfer rules – based on a bilingual dictionary – The dictionary can be used to deal with lexical ambiguity.

2.4.2 Corpus-Based Machine Translation Approach (CBMT)

This is one of the main methods of machine translation is Corpus-Based Machine Translation. The corpus-based approach for machine translation has emerged as one of the most widely explored areas in machine translation since 1989 (Tripathi, S., & Sarkhel, J. K, 2010). Corpus-based machine translation overcomes the problem of the knowledge acquisition problem of rule-based machine translation. This approach as its name points uses a large amount of raw data in the form of parallel corpora. Because of the high level of accuracy achieved during the translation, this method has dominated over other approaches. The Corpus-based approach is further classified into Example-Based Machine Translation and Statistical Machine Translation (Tripathi, S., & Sarkhel, J. K, 2010).

2.4.2.1 Example-Based MT (Memory based translation)

EBMT is a translation method that retrieves similar examples (pairs of source phrases, sentences, or texts and their translations) from a database of examples adapting the examples to translate new input (Vani K, 2019). The system maintains an example-base (EB) consisting of translation examples. When a source language sentence is given to the system, the system retrieves a similar source language sentence from the EB with its translation. Then it adapts the example to generate the target language sentence for the input sentence. The basic premise is that, if a previously translated phrase occurs again, the same translation is likely to be correct again. Thus, the EBMT system rests on the idea that similar sentences will have similar translations. The system has two main modules 1) retrieval and 2) adaption (Jaganadh G, 2010). There are three tasks in EBMT: Matching fragments against existing examples, transferring (identifying the corresponding translation fragments), and recombining the fragments to give the target text (Vani K, 2019).

2.4.2.2 Statistical Machine Translation

Statistical machine translation (SMT) is generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora (Brunning, J. J. J , 2010). The initial model of SMT based on the Bayes Theorem, proposed by Brown et al. (Brown, P. et al, 1990) takes the view that every sentence in one language is a possible translation of a sentence in the other and the most appropriate is the translation that is assigned the highest probability by the system. SMT is a machine translation approach that uses human-produced translations known as parallel corpus (Daba, J. & Assabie, Y, 2014) According to (Lopez, 2008) the translation process by using SMT is considered a machine learning problem. After examining the parallel corpus, SMT algorithms automatically 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 models (Jurafsky, D & Martin, J, 2002). SMT focuses on the result of translation rather than the process. So, a true translation which is both, faithful to the channel equation shows that two components are needed. These are a translation model P(S|T), and a language model P(T). SMT works based on the Bayesian model which translates source language S (Sidaamu Afoo) to target language T (Amharic) or source language and the best translation is selected depending on the highest value of the translation model (P (T|S)) (Brown, P. et al, 1990). Therefore the noisy channel via Bayesian rule is given as shown below.

$$\text{Best-translation } \check{T} = argmax\ P(T|S)$$

$$\check{T} = argmax\ \frac{P(S|T)P(T)}{P(S)}$$

$$\check{T} = argmax\ P(S|T)P(T)$$

Applying the noisy channel model to machine translation requires thinking of things backward (Lopez, 2008). It needs to pretend that the Sidaamu Afoo source language input S must be translated into a corrupted version of some target (e.g. Amharic) sentence T, and that the task is to discover the hidden (target language) sentence T that generates the observation sentence S. The noisy channel model of statistical MT thus requires three components to translate from a Sidaamu Afoo/Source sentence S to an Amharic/Target sentence T (Okpor, 2014) (Daba, J. & Assabie, Y, 2014). These are the language model to compute P(T), the translation model to compute P(S |T) and the decoder, which is given S and produces the most probable T. Figure 2-4 below shows the components of the approach:

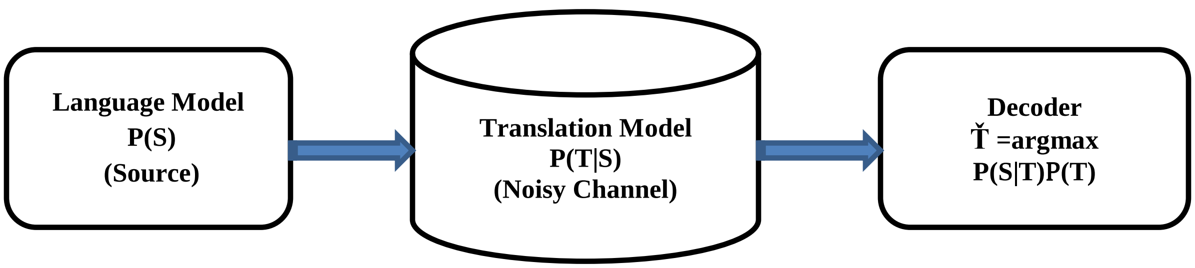


Figure 2-4 Components of Statistical Machine translation

2.4.2.2.1 Language model:

A statistical language model is a probability distribution over sequences of words. Given such a sequence with length $\boldsymbol{m}$, it assigns a probability, $(w1, w2, w3 \ldots \ldots . . wm)$ to the whole sequence. Having a way to estimate the relative likelihood of different phrases is useful in many natural language processing applications, especially ones that generate text as an output (Papineni, K. et al, 2002). The intuition of the N-gram model is that instead of computing the probability of a word given its entire history, we approximate the history by just the last few words (Martin, J. H., & Jurafski, D, 2000). To achieve this, we apply the Markov assumptions which say that the probability of a word depends only on the previous words. Markov models are the class of probabilistic models that assume that we predict the probability of some future unit without looking too far into the past. Based on it different kinds of N-gram probability exist such as unigram, bigram (looks one word into the past), trigram (looks two words into the past), and, in general, N-gram (looks N-1 words into the past) (Martin, J. H., & Jurafski, D, 2000). The N-gram model performs well, for the corpus with simple sentences with the unigram, bigram and trigram models since the words in the sentence are not that long. Yet a problem exists if the sentences are too long, and the solution would be smoothing which is avoiding zero probability. This means by avoiding zero probability no matter how long the decimal gets, it shouldn't be approximated to zero. Based on this method language model calculates the probabilities of N-grams which is used by the decoder (Martin, J. H., & Jurafski, D, 2000).

2.4.2.2.1 Decoding:

Decoding is the process of determining the most probable translation among all possible translations based on a searching algorithm. The search space is so huge because of different possible translation for each word (phrase) with different ordering in sentences. Different decoding algorithms were proposed for SMT. Most of these decoding algorithms are based on partial sentence evaluation as it is not possible to find the best translation. In order to solve decoding problem, most decoding algorithms are finding optimum solution instead of best solution. The Beam search algorithm, Greedy decoder and stack decoding algorithm are some examples. Most decoders in the SMT are based on the best-first search (Mulugeta, D, 2015). The A* was the first of the best-first search that was proposed by IBM group and implemented on word to word SMT where the search hypotheses are managed in a priority queue (stack) ordered by their scores (MARA M, 2018). The beam search is the other best-first search that is implemented on the phrase based decoding in the Moses system.

2.4.2.2.2 Translation Model:

To build a translation model as mentioned earlier, we should have a source language sentence S (E.g. Sidaamu Afoo (S)) and target language sentence T (E.g. Amharic (A)) of parallel corpus. Therefore, the job of the translation model is to assign a probability that S generates to T. As mentioned above, for a given source and target sentences S and T, it is the way sentences in S get converted to sentences in T which is denoted by (Martin, J. H., & Jurafski, D, 2000) (Koehn, 2009 &2010):

$$P\ (T|S) = \frac{\boldsymbol{Count}\ (T, S)}{\boldsymbol{Count}(S)}$$

The above equation may be difficult to achieve, if the sentences are too long. To overcome this problem the sentence is decomposed into words and sub-words called morpheme, as in language modeling (Koehn, 2009 &2010).

$$P\ (S|T) = \sum_{X} P(S, X\ |T)$$

The variable X represents alignments between the individual chunks in the sentence pair where the chunks in the sentence pair can be morphemes or words or phrases. In morpheme-based translation, the fundamental unit of translation is a morpheme. Phrase-based translations, most commonly used, translates whole sequences of words, where the lengths may differ in which blocks are not linguistic phrases but, phrases found using statistical methods from corpus (Girma E, 2021).

Alignment

Alignment is the arrangement of something in an orderly manner with something else (Koehn, 2009 &2010). It can be performed at different levels, from paragraphs, sentences, segments, words and characters.

Word alignment: This determines the translational correspondences at the word level given a parallel corpus. It is a mapping between the words in the source sentence and the words in the target sentence. The initial statistical models for machine translation are based on words as atomic units that may be translated, inserted, dropped, and reordered (Koehn, 2009 &2010).

Word-based statistical machine translation ignores possible morphological relatedness of the words. This is more of a problem for inflectional languages the richer their morphology (Sidaamu Afoo and Amharic), the larger training corpus has to be to cover most of the possible word forms (Gamback, B. et al, 2005). Currently, word alignment models for statistical machine translation do not address morphology beyond merely splitting words. A morpheme alignment is a function mapping a set of morpheme positions in a source language sentence to a set of morpheme positions in a target language sentence (Girma E, 2021).

Word alignment commonly has done using IBM Models 1-5. IBM Model 1 is weak in terms of conducting reordering or adding and dropping words. The IBM Model 2 has an additional model for an alignment that is not present in Model 1. Some source words may be translated into multiple target words (fertility of the words) this problem is addressed in IBM Model 3. In IBM Model 4, each word is dependent on the previously aligned word and the word classes of the surrounding words. IBM Model 5 reformulates IBM Model 4 by enhancing the alignment model with more training parameters to overcome model deficiency (Koehn, 2009 &2010) (Girma E, 2021).

2.4.2.2.3 Evaluation system

Evaluating the quality of a translation is an extremely subjective task, and disagreements about evaluation methodology are widespread. Nevertheless, evaluating MT results is important to know how good a Machine Translation system is and identify new development areas for improvement in translation quality (White et al, 1994). suggested three aspects of evaluating a translation how well the translation represents the source text (adequacy), the extent to which the translation is a well-formed and correct sentence (fluency), and comprehensiveness of the information for readers (In formativeness) (O'connell, M. J, 1994). Generally, MT evaluation can be performed through a human or automated system (Mulugeta, D, 2015). The human and automatic MT evaluations are briefly described below.

2.4.2.2.3.1 Human Evaluation

Human evaluation of machine translation means that the assessment of translation quality is done by human professional translators. This is the most effective option when it comes to determining the quality of machine translations down to the level of sentences. But human evaluation, as with human translation, is by nature more costly and time consuming (Lopez, M. et al, 2007).

Automatic evaluations of Machine Translation are based on evaluation metrics like precision. It compares system translation output with reference translations from the parallel corpus. The automatic evaluations are important as they run frequently and are cost efficient. There are different types of heuristic methods, such as BLEU, NIST, TER, Precision and Recall, and METEOR (Jurafsky, D & Martin, J, 2002). All heuristic methods except Bleu require human translation and time consuming. In BLEU each MT output is evaluated by a weighted average of the number of N-gram overlaps with the human translation.

2.4.2.2.3.2 BLEU (Bilingual Evaluation Understudy)

BLEU is one of the famous evaluation methods that can be used to have a comparison among different Machine Translation systems. BLEU scoring tool is used for the evaluation of the quality of the translation system based on the familiarity to the researcher and applicability with Moses. The BLUE evaluates the quality of text which has been machine-translated from one natural language to another based on the degree of correspondence between a machine's output and that of a human professional human translation (Fosler-Lussier, E. et al, 2012).

2.4.3 Hybrid Approaches

The approach makes use the strong side of the statistical a rule-based translation approach (Daba, J. & Assabie, Y, 2014). It has a better efficiency from all the machine translation approaches and used in different ways. One way is to perform the translation at the first stage using a rule-based approach followed by adjusting the output using statistical information. Moreover, in some cases rules used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. The later way is more suitable since it has more power, flexibility, and control in translation (Antony P, 2013).

2.4.4 Neural Machine Translation Approaches

Neural machine translation (NMT) is a new breed of corpus-based machine translation it is similar to the statistical machine translation technology that was the state of the art until very recently, but uses a completely different computational approach neural networks (Bahdanau, D et al, 2014). These systems are also known as sequence-to-sequence models or encoder-decoder networks and were initially fairly simple neural network models made out of two recurrent parts (Forcada, M. L, 2017) is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, which consists of many small subcomponents (words) that are tuned separately, neural machine translation attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation. Most of the proposed neural machine translation models belong to a family of encoder– decoders (Bahdanau, D et al, 2014) (Forcada, M. L, 2017). Two recurrent neural networks (RNN) Encoder, is used by the neural network to encode a source sentence into a fixed vector and decoder, used to predict words in the target language.

2.5 Related Works

Machine translation research has been conducted in many different ways utilizing various methodologies and approaches for many languages. Following are some of the studies with a particular emphasis on the SMT approach.

2.5.1 Bi-directional Amharic – Afaan Oromo Machine Translation Using Statistical Approach

The research was done on Bi-directional Amharic – Afaan Oromo Machine Translation system Using Statistical Approach, conducted by **Emebet Girma in 2021**. The main objective of the research is to design and develop a bi-directional Amharic - Afaan Oromo machine translation system using a statistical approach **(**Girma E, 2021**)**. The research work was done using statistical machine translation and compared Experimental performance results between Unsupervised Morpheme-level translation and word-level translation; applying 14600 parallel sentences in the entire parallel corpus that was used in the experiment. All corpuses used for the study were collected from holly bible and religious documents.

For this experiment tools used were: Moses for the translation process, MGIZA++ for word and morpheme alignment, IRSTLM for language modeling, and Morfessor for morphological segmentation.

The result showed for the unsupervised morpheme segmentation-based level alignment using the BLEU score has an average of 19.77 % accuracy for the Amharic to Afaan Oromo and 16.14 % for the Afaan Oromo to Amharic. For word-based alignment, the BLEU Score was 13.84 % for Amharic to Afaan Oromo and 9.72% for Afaan Oromo to Amharic.

As the results showed morpheme-level alignment translation performs better than word-level alignment translation according to emebet.

The researcher recommended, further study for better results using a large set of corpora that includes different disciplines. Also indicated to under go rule-based morphological segmentation or machine learning algorithms for designing an optimal model for segmentation taking in to account that both languages are morphological rich. (Girma E, 2021).

2.5.2 Morpheme-Based Bi-directional Ge’ez-Amharic Machine Translation

The research was done on a morpheme-based bi-directional Ge’ez-Amharic machine translation system using a statistical approach. Conducted by **Tadesse Kassa in 2018** the main objective of the research is to design morpheme-based bi-directional machine translation for Ge’ez-Amharic textual documents (Kassa, T, 2018).

The size of the parallel corpus used for the experiment contains a total of 13,833 simple and complex sentences and tools used for this experiment are: Moses for the translation process, MGIZA++ for alignment of word and morpheme, IRSTLM for language modeling and for morphological segmentation, Morfessor were used. The research work was implemented using statistical machine translation and compared Experimental performance results between Morpheme-level translation and word-level translation. The results showed a better performance of 15.14% and 16.15% BLEU scores using morpheme-based from Geez to Amharic and from Amharic to Geez translation, respectively as compared to word-level translation, there is on average 6.77% and 7.73% improvement from Geez-Amharic and Amharic-Ge’ez respectively (Kassa, T, 2018). The research shows an additional experiment using unsupervised and rule-based morpheme segmentation approaches. The BLEU score is 0.6% and 1.27% for Ge’ez to Amharic and Amharic to Ge’ez respectively. From the result, rule-based morpheme segmentation approaches are better than the unsupervised approach when Amharic is used as a source language and Ge’ez is used as a target language. Finally, the researcher recommended, Alignment of Ge’ez-Amharic text is a challenging task because of the many-to-many correspondence between words/morphemes of the two languages. Hence, there is a need to identify optimal alignment for Ge’ez-Amharic Machine translation and the researcher used prefixes and suffixes for rule-based morphological segmentation. In addition, the researcher recommended; since both languages are morphologically rich, there is a need to apply machine learning algorithms to designing an optimal model for segmentation (Kassa, T, 2018).

2.5.3 Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation

The research has been conducted by **Yitayew Solomonin 2017** (Solomon, Y, 2017). The main objective of this study is to explore the effect of word level, phrase level, and sentence level alignment on the Bi-Directional Afaan Oromo-English was conducted by using a statistical machine translation approach. Corpus was collected from different sources like the criminal code, FDRE constitution, Megleta Oromia and Holly Bible; and prepared 6400 simple and complex sentences and made the corpus suitable for use to train and test the system and used the 9:1 ratio respectively. The system is bi-directional; two language models are developed, for English 1900 monolingual sentences and Afaan Oromo monolingual 12200 sentences used. To develop the system different tools were used, such as Moses for decoding purposes, MGIZA++for alignment, Anymalign and hunalign, and IRSTLM for language modeling.

The researcher run six tests using aligned phrases of various lengths in both directions (from English to Afaan Oromo and Afaan Oromo to English). Utilizing 4 phrases with maximum lengths and 1 phrase with minimum lengths, the first and second experiments are conducted. The experiment's findings had BLEU scores of 21% for translating from English into Afaan Oromo and 42% for translating from Afaan Oromo into English. The results of the third and fourth studies, which used 16 maximum and 4 minimum phrase lengths, yielded BLEU scores of 27% for English to Afaan Oromo translation and 47% for Afaan Oromo to English translation. The fifth and sixth experiments use phrases with 30 maximum and 20 minimum characters to align sentences at the sentence level. BLEU scores range between 18% and 35%.

Conclusions drawn from the experimental results are: a better translation is obtained when Afaan Oromo is used as a source language and English is used as the target language. The experiment using 16 Maximum and 4 Minimum lengths of phrases shows better performance with BLEU scores of 27% for English to Afaan Oromo and 47% for Afaan Oromo to English translation. Finally, the researcher suggested using a hybrid technique to manage different alignments and provide superior SMT results.

2.5.4 Bi-directional English-Amharic Machine Translation: An Experiment using Constrained Corpus

The research was conducted by **Eleni Teshome in 2013** (Teshome, E, 2013). The objective of this study was to design and develop a bi-directional English-Amharic machine translation system using a constrained corpus. Sample text corpora used from relevant data sources with parallel text. For the simple sentences, 1020 sentences were manually prepared and for the complex sentences 1951 were collected. Out of it; 414 from the Public Procurement Directive and 1537 sentences from the Bible.

Statistical machine translation approach was used for the study. Researcher discovered and supposed that as the size of the corpora increases, the accuracy of the translation improves.

A two methodology testing has been used. The first one is the BLEU score and the second one preparing a questionnaire manually. BLEU score is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. The second testing methodology is the manual questionnaire technic. Questionnaires developed were different, thus the results obtained from Amharic to English and from English to Amharic were not similar.

From the experimental results for the simple sentence using BLEU score had on an average of 82.22% and 90.59% from English to Amharic and Amharic to English respectively and, using the manual questionnaire method the accuracy is 91% and 97% form English to Amharic and Amharic to English respectively. For the complex sentences, the result acquired from the BLEU Score was approximately 73.38% for English to Amharic, 84.12% for Amharic to English, and from the questionnaire method from English to Amharic 87% and from Amharic to English was 89%. In this regard, manual questionnaire method has better performance for a simple sentence. Finally, the author recommended that; further research in machine translation of Amharic to other languages, in Ethiopia such as Tigrigna, Afaan Oromo or so could be performed while preparing a large corpus.

2.5.5 English-Afaan Oromo machine translation: An experiment using statistical approach

The study was conducted by **Sisay Adugna in 2009**. The main objective of this study is to develop a prototype English-Afaan Oromo machine translation system using a statistical approach, i.e., without explicit formulation of linguistic rules (Adugna, S, 2009).

The researcher 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 purposes. The data is organized into training and testing data in the proportion of 9:1 (90% for training and 10% for testing).

The researcher used different tools for the implementation of the system, SRILM toolkit was used for language modeling, GIZA++ for word alignment, which implements the word alignment methods IBM1 to IBM5, Decoding is done using Moses, and the documents were preprocessed using different scripts written for this purpose like the apostrophe. Sentence aligning, tokenization, lowercasing, and truncating long sentences that take the alignment to be out of optimality were done by those scripts.

The researcher performs different experiments by a varying number of N-grams, the n-gram score for values of n equals 1, 2, 3, 4, 5, 6, 7, 8 and 9 is observed to be 43.96%, 21.57%, 14.42%, 10.72%, 8.04%, 5.52%, 3.76%, 2.23% and 1.30% respectively.

Finally, the author strongly recommends the addition of more bilingual data for further experimentation and the development of a spell checker for Afaan Oromo that will help facilitate the document preparation.

2.5.6 Geez to Amharic Automatic Machine Translation: A Statistical Approach

The research has been conducted by **Dawit Mulugeta in 2015** (Mulugeta, D, 2015). The general objective of this research is to investigate the application of Statistical Machine learning technique to Machine Translation from Geez to Amharic. This researcher uses quantitative experimental methodology.

Researchers gathered 12860 parallel bilingual corpora from the Bible and a few other holy texts to undertake corpus-based machine translation (Wedase Mariam and Arganon). Over 90% of the whole data set was used as a training set after being separated into training and testing sets. Moses for translation, IRSTLM and SRILM for language modelling, GIZA++ for word alignment toolset used to train IBM Model 1 to Model 5, the Hidden Markov Model, and the BLEU score were all employed by the researcher to construct the system.

The researcher conducted experiments based on sentence-level alignment the performance of the result and the BLEU score obtained were 8.14%. Further investigation has been done to crosscheck and improve the performance score using the 10-fold cross-validation (CV) method. The BLEU score result obtained on the trials was 9.11%, 7.44%, 7.61%, 6.36%, 10.26%, 9.39%, 8.01%, 8.54%, and 7.72%. The result verifies that the performance is highly dependent on the training and testing data domain. The researcher also checks the performance of the system after splitting each book of the Bible into the training and testing set. The trials have been done three times to see the result and average accuracy is calculated to compare with the 10-fold CV test result and get better performance than the 10-fold CV result.

The author concluded by suggesting that this study be expanded using various morphological segmentation and synthesizing mechanisms, using a larger corpus size and various domains of contents other than the religious one, and should be carried out using the Example-based Machine Translation approach.

2.5.7 Bi-directional English-Afaan Oromo Machine Translation Using Hybrid Approach

The research was conducted by **Jabesa Daba** in 2013 with the objective of developing a bi-directional English-Afaan Oromo machine translation system using hybrid approach (Daba, J, 2013). The researcher used corpus for the experiment collected from different sources including the Holy Bible, the Constitution of FDRE, the Criminal Code of FDRE, International conventions, Megeleta Oromia, a bulletin from the Oromia Health Bureau, and different websites. The documents were preprocessed using different scripts like: Tokenization, True-casing, and cleaning to make them ready for an experiment. The size of the parallel corpus used for the experiment contains a total of 3000 corpora from this 90% were used for training and the remaining 10 percent for testing for both approaches. 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 by using a statistical approach.

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 by using a rule-based approach. From the result of both experiments, the Author concludes that the result recorded from a BLEU score shows that the hybrid approach is better than the statistical approach for the English-Afaan Oromo language pair.

Finally, the author recommends that 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.

Summary

In this section, a few works related to machine translation have been discussed for different language pairs.

Conducted by	Title	Objective	Types of experiment		BLEU score		Research gap
Emebet Girma	Bi-directional Amharic- Afaan Oromo Machine Translation Using Statistical Approach	To design and develop a bi-directional Amharic - Afaan Oromo machine translation system using a statistical approach	Word-based	AM - AO	13.84 %		Lack of standardized corpus and used Small data size
				AO- AM	9.72%		
			Unsupervised morpheme based	AM - AO	19.77 %		
				AO- AM	16.14%		
Tadesse Kassa	Morpheme-Based Bi-directional Ge'ez - Amharic Machine Translation	Morpheme-Based Bi-directional Ge'ez -Amharic Machine Translation	Word-based	Geez- AM	8.37%		Limited in religious documents only
				AM- Geez	8.42%		
			Morpheme Based Translation	Using Morfessor	Geez- AM	14.54%	
				Using Rule-based	AM- Geez	14.88%	
					Geez- AM	15.14%	
					AM- Geez	16.15%	
Eleni Teshome	Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus	To explore the effects of simple and complex sentence	Simple sentence	Eng.-AM	BLEU	Manually	Used a Small data size
					82.22% 9	91%	
				AM - Eng.	90.59%	97%	
Yitayew Solomon	Optimal Alignment for Bidirectional Afaan Oromo-English Statistical Machine Translation	To explore the effect of word, phrase and sentence level alignment	4 max and 1 min phrase Complex sentence	Eng.- Orom	21%		The absence of a standardized corpus
				Orom- Eng.	42%		
			16 max and 4 min phrase	Eng.- Orom	27%		
				Orom- Eng	. 47%		
			30 max and 20 min phrase	Eng.- Orom	18%		
				Orom- Eng.	35%		
Sisay adugna	English – Afaan Oromo Machine Translation: An Experiment Using Statistical Approach.	To develop a prototype English-Amharic SMT without explicit formulation of linguistic rule	experiments by a varying number of Ngrams, the ngram score	Eng.-AM n equals 1, 2, 3, 4, 5, 6, 7, 8 and 9	43.96%, 21.57%, 14.42%, 10.72%, 8.04%, 5.52%, 3.76%, 2.23% and 1.30% respectively		The absence of a sufficient amount of machine-readable documents
Dawit Mulugeta	Geez to Amharic Automatic Machine Translation: A Statistical Approach	To investigate the application of Statistical Machine learning technique to Machine Translation from Geez to Amharic.	Sentence level	Geez –Am	8.14%		Lack of standardized corpus and used Small data size
			10-fold cross validation (CV)	Geez –Am	9.11%, 7.44%, 7.61%, 6.36%, 10.26%, 9.39%, 8.01%, 8.54% and 7.72%.		

Table 2-1 related works summary

Whatever the case, we attempted to understand from the associated literature review above that, a result varies based on the approaches utilized by each researcher.

Those who used a large set of corpora, all kinds of translation units (word, morpheme, phrase, sentences, etc.) and hybrid approach likely got better results than those who used a limited number of corpora and constrained with a single approach.

CHAPTER THREE

3. Amharic and Sidaamu Afoo languages

3.1 Introduction

This chapter briefly discusses an overview of the Amharic and Sidaamu Afoo languages the relationship and difference between them and the nature of the Language pair.

3.2 Ethiopian Languages

Ethiopia consists of more than 86 different languages with over 200 dialects spoken. The Largest ethnic and linguistic groups are the Oromo, Amhara, etc. Ge'ez is Ethiopia's original, ancient language and was introduced as the official written language during the first Aksumite kingdom. The unique script is derived from the Sabean alphabet and is still used today by the Ethiopian Orthodox Tewahedo Church. Amharic is derived from Ge'ez. It is the official national language of Ethiopia. The Ethiopian languages are divided into four major language groups. These are Semitic, Cushitic, Omotic, and Nilo-Saharan (Ethiopian Treasures, 2013).

3.3 A Brief Overview of Amharic Language

Amharic is one of the Semitic languages spoken in Ethiopia. Next to Arabic, it is the second most spoken Semitic language in the world and it is the official working language of the Federal Democratic Republic of Ethiopia. It is also the first largest spoken language in Ethiopia and possibly one of the five largest languages on the African continent. The Amharic alphabet is called Fidel 'ፊደል', which grew out of the Ge'ez abugida-called in the Ethiopian Semitic language. The usual word order of Amharic is Subject-Object-Verb (SOV) (Teshome, E, 2013). Modern-written Amharic uses a unique script called hohiyat (ሆሂያት) which is conveniently written in a tabular format of seven columns (Abel, B, 2018). The first order is the basic form; the other orders are derived from it by more or less regular modifications indicating the different vowels (Argaw, A. A., & Asker, L, 2007). The alphabet is written from left to right in contrast to some other Semitic languages such as Arabic and Hebrew. It consists of 34 consonants, giving 7*34=238 syllable patterns, or fidels 'ፊደል' (Eyassu, S., & Gambäck, B, 2005). In addition to the 238 characters, there are other non-standard alphabets which contain special features such as "ቋ ሟ ሎ ቧ ኋ…" (Teshome, E, 2013) each alphabet represents a consonant together with its vowel. The vowels are fused to the consonant form in the form of diacritic markings. The diacritic markings are strokes attached to the base characters to change their order (Bethelhem, M, 2002).

3.4 A Brief Overview of Sidaamu Afoo

Sidaamu Afoo is one of the languages in Ethiopia, widely used in Sidaama National Regional State (SNRS). It is also the official working language of Sidaama National Regional State (SNRS), next to Afaan Oromo and Somale, Sidaamu Afoo is the biggest language under the Cushitic category with more than 5 million speakers of the language (HUMMEL, S., & PANINI, F, 2020). Sidaamu Afoo is also serving as the instructional medium for primary and secondary school; until 1990 Ge'ez script for writing. However, since 1991 its official writing system is the Latin alphabet system (Kawachi, 2007). The Sidaamu Afoo script which is called 'Fidalla' has 33 symbols. These are: 25 of them are single-digit letters found in Latin script as found in English, except v (Abdella, K. M, 2010), the 7 two-digit letters; and the symbol (') is called a glottal stop. It is a phoneme in Sidaamu Afoo. The two-digit letters are ch, dh, ny, ph, sh, ts, and zh (Sidama Information and Culture Department , 2007) (Mohammed, A. K., & Pandey, A, 2018) (Kawachi, 2007). According to (Demeke, G. A, 2006), in Fidalla, like in English, both capital letters and small letters are used. The small letters are called 'Shimmaada Fidalla' and the capital letters are 'Jajjaba' Fidalla.

The usage of the apostrophe (') and double apostrophe (''), when there is a break between vowels or between vowels and consonants is another crucial aspect of the Sidaamu Afoo. According to (M. W/Giorgis, 2019), (') is considered one of the alphabets, and the number of alphabets in Sidaamu Afoo became 33. The apostrophe (') and double apostrophe ('') indicate that the vowels are pronounced independently from each other.

For example:

- Ce'a /bird, Cee'ma /to be lazy/, sa'a /to pass/
- Ka''u /She rised/ Ka''i /He rised./, ko''ee /this one/

Sidaamu Afoo has five short vowel phonemes (/i/, /e/, /a/, /o/, /u/) and their long counterparts (/ii/, /ee/, /aa/, /oo/, /uu/) (Kawachi, 2007). The length of the vowel makes a difference in word meaning.

For example:

- i/sinna 'branches'
- [=ii/ siinna 'coffee cups.
- e/ tenne 'at that time, then'
- ee/- teenne 'flies'.

3.5 Nature of the Language pair

Amharic and Sidaamu Afoo languages have differences and similarities in linguistic structure discussed below:

3.5.1 Writing system

Amharic is written using Fidel, ፊደል, which grew out of the Geez. Modern Amharic has inherited its system of writing by way of the language of the old kingdom of Axum, Geez, which is still the classical and ecclesiastical language of Ethiopia. The roots, then, of Amharic orthography; like those of the language itself, are Semitic. Both Geez and the related languages of Ethiopia are written and read from left to right, in contrast to the other Semitic languages like Arabic and Hebrew (Eyassu, S., & Gambäck, B, 2005).

Amharic has its own writing system, a semi-syllabic system. It possesses 34 primary characters, each representing a consonant and each having 7 variations in form to indicate the vowel which follows the consonant. These 34 sets of 7 forms are the "ordinary characters"; but besides them there are also a number of "diphthong characters", each representing a consonant and a following vowel with a w sound (or, in one case, a y sound) interposed between them. In writing, none of them is indispensable because the same sounds can always be represented by combinations of the ordinary characters, but many of them are in common use and, on the whole, they cannot be ignored (Bethelhem, M, 2002).

Until the mid-1970's Sidaama was a vernacular language. It was not confined and as such was not used for teaching or other purposes. However, the Military Regime of Ethiopia (MRE) permitted Sidaama language to be used for the ongoing literacy campaign. During this period Sidaama was written using the Ethiopian script. However, since localization was done inaccurately, reading and understanding the Sidaama documents and books was very difficult. This emanated from the discrepancy in the number of vowels in Amharic and Sidaama and inappropriate decision regarding the representation of vowels (Teferra, A, 2000). However, the selection of Sidaama as a language of literacy enabled the publication of several educational primers. Another fundamental change with regards to the status of Sidaama language took place in 1992. The Transitional Government of Ethiopia (TGE) at the time declared that each and every ethnic group can use its own language for primary education and also for official and administrative purposes. In addition permission was given to use any type of script in order to write the vernacular languages. Thus, beginning from 1992, Sidaama is being used as a language of primary education, and for administrative and judicial matters. In addition the Latin script was adopted as the writing system (Kawachi, 2007).

3.5.2 Word order (Sentence Structure of Amharic and Sidaamu Afoo)

Amhraic and Sidaamu Afoo are the same in sentence structure order. Like Amhraic, the sentence structure of Sidaamu Afoo is Subject-Object-Verb (SOV). For example, in the sentence,
አበበ ትምህርት ቤት ሄደ:: /Ababbi rosu minira hadhi. / [Abebe went to school.], ‘አበበ’ / Ababbi / [Abebe] is the subject, ‘ትምህርት ቤት’ / rosu minira / [school] is the object and ‘ሄደ’ / hadhi / [went] is the verb.

3.5.3 Nouns

Nouns in Amharic are words that are used to identify names, of things and places. A word is grouped under noun if it inflects for the Amharic plural marker ‘-ኦች’ /‘-och’/ or ‘-ዎች’ [-woch], if it can be used as a subject or an object in a sentence, is modified by adjectives and comes after demonstrative pronouns (Ibrahim, A., & Assabie, Y, 2013) Amharic plural nouns are mainly formed by adding suffixes: ‘-ኦች’ /‘-och’/ or ‘-ዎች’ /‘-woch’/. Table 3.1 shows suffixes used to form plural nouns in Amharic.

Singular Noun	**Plural marker**	**Plural Noun**
በሬ /bäre/ [ox] -	-ዎች	በሬዎች / bärewoch/ [oxen]
አስተማሪ /xästämari/ [teacher]	-ዎች	አስተማሪዎች /xästämariwochï / [teachers]
ላም /lam/ [cattle]	-ኦች	ላሞች /lamoch / [cattle]
ቤት /bet/ [house]	-ኦች	ቤቶች /betoch / [houses]

Table 3.1: Amharic plural noun formation using suffix

Amharic nouns can be either primary or derived. They are derived if they are related in their root consonants and/or meaning to verbs, adjectives or other nouns. Otherwise, they are primary (Ibrahim, A., & Assabie, Y, 2013) (Teklewold, A. A, 2013). For example, the noun መንገድ /mängäd/ [street] is primary but, ‘መንገድ-ኧኛ’ ⟶ መንገደኛ /mängädäNa/ [traveler] is derived from the nominal base መንገድ by adding the morpheme ‘-ኧኛ’. Nouns can be derived from other nouns, adjectives, roots, stems and the infinitive form of a verb by affixation and intercalation. The morphemes ‘-ነት’, ‘-ኧኛ’, ‘-ኧት’, ‘-ኣዊ’, ‘-ተኛ’, ‘-ኛ’ and the prefix ‘ባለ-‘ are used to derive nouns from other nouns. Table 3.2 shows examples of nouns derived from other base nouns.

Base noun	**Derived noun**
ሰው /säw/ [Person]	ሰው-ነት → ሰውነት /säwnät/ [Body]
ሃብት /habt/ [Wealth]	ባለ-ሃብት → ባለሃብት /balähabt/ [Wealthy]

Table 3.2: Nouns derived from other nouns

Nouns in Sidaamu Afoo are any word that can be used to name or identify place, object or ideas. Two types of grammatical genders exist in Sidaamu Afoo nouns. These are masculine and feminine, and the entire nouns of the language belong to one of these gender categories. Similarly, there are two numbers (singular and plural) which can be identified by the morpheme it adds. Plural form of a given noun can be formed by adding suffix to the root noun. Various types of suffixes can be added to transform a singular noun to its plural form. All of these suffixes change singular noun to plural without variation in meaning. The last vowel of the singular noun is dropped before the suffix is added. There are several forms of suffixes that turn verbs into nouns, namely, -a, -o, -e, -anso, - atto, -ano, -ille, -imma, and -aancho. Because the citation form of any word in Sidaamu Afoo ends in - a, -o, or -e, it is sometimes difficult to decide whether nouns ending in one of these vowels derive from verbs by adding a nominalizing suffix to the verb stem, or the verbs derive from the nouns. Examples are shown blow in Table 3.3 (Kawachi, 2007).

e, o, and, i	Verb	Noun
-a	gurd- ‘to knot’	gurd-a ‘knot’
	wi’l- ‘to cry’	wi’l-a ‘condolensces, cryin
	giir- ‘to burn’	giir-a ‘flame’
-o	egenn- ‘to know’	egenn-o ‘knowledge’
	gorr- ‘to slaughter’	gorr-o ‘slaughtering’
	hanq- ‘to get angry’	hanq’-o ‘anger’
-e	fool- ‘to breathe’	fool-e ‘breath’
	godo’l- ‘to play’	godo’l-e ‘play’
	mund - ‘to bleed’	mund-ee ‘blood’

Table 3.3: Example to justify the above discussion

There are five other suffix forms that more clearly derive nouns from verbs, -ansho, -atto, -ano, - ille, and -imma. Examples are shown in the next discussion in Table 3.4. -ille and -imma can also be used to derive abstract nouns from adjectives and other nouns (Addisu, B, 2016).

Suffix	**Verb**	**Noun**
-ansho	giw- ‘to hate sb’	giw-ansho ‘hatred, disagreement’
	k’olch- ‘to outdistance sb’	k’olch-anso ‘race’
	gan- ‘to hit, beat, fight’	gan-ansho ‘fighting’
-atto	ag- ‘to drink’	ag-atto ‘drinking’’
	ha’r- ‘to go’	ha’r-atto ‘going’
	huucc’- ‘to pray’	huucc’-atto ‘prayer
-ano	mug- ‘to become sleepy’	mug-ano ‘sleepiness’
	huucc- ‘to pray, ask’	huucc’-ano ‘advice’
-imma	gedh- ‘to become old’	gedh-imma ‘old age’
	hala’l- ‘to become wide’	hala’l-imma ‘width’

Table 3.4: Examples to justify the behavior suffix **(Addisu, B, 2016)**

3.5.4 Personal Pronouns

Personal pronouns **in Amharic** are pronouns that are associated primarily with a particular grammatical person, first-person (as I (እኔ)), second-person (as you (አንተ/አንቺ)), and third-person (as he, as she (እሱ/እሷ). The subject form is the nominative case which occurs' as the subject of a simple verb, a verbal phrase, and a nominal sentence (Ryan Cotterell, et al, 2016).

In Sidaam Afoo Pronouns are words that can be used in place of nouns. Similar to that of nouns, pronouns have number and gender. For example, ise/isi which means ‘she’ is feminine (singular) whereas isi which means 'he' is masculine (singular) and ‘insaa’ which means 'they' is plural can be masculine or feminine. Pronouns can also be categorized based on their functions and meanings in the sentence. These are personal pronoun, possessive pronoun, demonstrative pronoun, relative pronoun or reciprocal pronoun (Lalego, D.L, 2020). Table 3.4.2.1 illustrates personal pronouns that can be in the subject positions.

In Amharic and Sidaamu Afoo personal pronouns can be classified as singular and plural. These are shown in the Table 3.5 below:

	1 st Person	2 nd person	3 rd person
Singular	እኔ/Ani [I].	አንተ/አንቺ Ati[you].	እሱ/እሷ Isii/Ise[He/She]
Plural	እኛ/Ninke[we]	እናንተ/Ki'ne[you]	እነሱ/Insaa[they]

Table 3.5 Personal pronoun in Amharic and Sidaamu Afoo language

In both Amharic and Sidaamu Afoo languages, there are some unique features; such as within second-person and third-person singular, there are two additional polite independent pronouns, for reference to people to whom the speaker wishes to show respect. The polite personal pronouns are እርስዎ/ Ki'ne [You, singular, polite] and እሳቸው/Inisa [He/She, singular, polite].

3.5.5 Adjective

Amharic adjectives modify nouns or pronouns by describing, identifying, or quantifying words (ይማም, 2010), (አማረ, 2010). Amharic adjectives always come before nouns or pronouns which they modify, but all the words that come before nouns cannot always be adjectives (Ibrahim, A., & Assabie, Y, 2013). As is true for nouns, adjectives can also be primary (such as ደግ /däg/ [kind], ፈጣን /fäTan/ [fast]) or derived. Adjectives are derived from nouns, stems, or verbal roots by adding a suffix or a prefix and by intercalation. For example, it is possible to derive ድንጋይ-ኣማ ➔ ድንጋያማ /dngayama/ [stony] from the noun ድንጋይ /dngay/ [stone]; ሀይል-ኧኛ ➔ሀይለኛ /hayläNa/ [powerful] from the noun ሀይል /hayl/ [power]; ስኧንኧፍ ➔ ሰነፍ /sänäf/ [lazy] from the root ስንፍ /snf/; ክብኡር ➔ ክቡር /kbur/ [respectful] from the root ክብር /kbr/ [respect] by suffixation and intercalation (Heyi, G.T , 2020).

In Sidaamu Afoo an adjective could be a word that describes nouns and pronouns. It specifies which one, how much, what kind, and more. An adjective enables listeners and readers to visualize things more clearly by using their senses.

It can be simple or derived. Derived adjectives are usually formed from verb stems by the derivational suffixes -a, -čo or -ado. According to (Teferra, A, 2000) the derivational morpheme –a is more frequent than the other two. Derived and simple adjectives agree with their head noun in number, gender and case. This means that adjectives can inflect, like nouns, to nominal grammatical categories. However, the inflectional forms of adjectives may not be identical to that of nouns (Teferra, A, 2000).

3.5.6 Verb

A verb is a word that expresses action, state of being in or relationship between two things (Heyi, G.T , 2020). Amharic verbs take subject markers as a suffix like '-ሁ' for subject 'I' as in መጣሁ /mäTahu/ [I came], '-ህ' for subject 'you' as in መጣህ /mäTah/ (you came), '-ች' for subject 'she' as in መጣች /mäTac/ [She came], and so on, to agree with subject of the sentence. Amharic verbs often have additional morphology that indicate the person, number and (second person and third person singular) gender of the object of the verb. For example: አንቺን አየሁሽ /xäncin xäyähuś/ [I saw you], '-ሁሽ' indicates second person, singular, feminine, and in the sentence አልማዝን አየኋት /xälmazn xäyäwat/ [I saw Almaz] '-ኋት' indicates third person, singular, feminine (Heyi, G.T , 2020).

In **Sidaamu Afoo** language verbs are words that are used to indicate some action or event occurrence within time boundaries. Sidaamu Afoo verbs are either intransitive or transitive. There are almost no ambitransitive verbs like the English verb break. Sidaamu Afoo has a limited number of intransitive-transitive pairs of morphologically unrelated lexical verbs that seem to show a semantic contrast, but it is not clear whether they are contrastive only with respect to causativity (Kawachi, 2007).

Intransitive	**Transitive**
re- 'to die'	shi&- 'to kill'
ub- 'to fall'	tuγ- 'to drop (sth light)'
'ba'- 'to disappear, get lot; to become destroyed/spoiled'	kaar- 'to drop (sth heavy)
	hooγ- 'to lose sth' hun- 'to destroy, spoil'

Table 3.6 Show intransitive transitive verbs

Sidaamu Afoo verb forms showing a contrast in transitivity follow one of the three patterns. First, there are a large number of intransitive verb roots whose causative forms are transitive verbs. Second, the passive form of a transitive verb root acts as an intransitive verb that expresses a state-change. Third, there are many pairs of idiomatic expressions with ass- 'to do' and y- 'to say' that show a transitivity contrast.

3.5.7 Adverb

In Amharic, adverbs are used to modify the coming verbs. Adverbs always come before the modified verb. Adverbs can be found either in their primitive form or compound form as grouping of preposition and other word categories (Ibrahim, A., & Assabie, Y, 2013). For example: in the adverbial phrase,

መምጣት አለመምጣቷን ገና አልወሰነችም /mämTat xälämämTatwan gäna xälwäsänäcm/ [She hasn't yet decided if she wants to come or not], ገና /gäna/ [yet] is the only adverb that formed the adverbial phrase. There are only few adverbs in Sidaama. Adverbial functions are often expressed by postpositional phrases and subordinate clauses. Although few in number there are simple, derived and compound adverbs. The latter outnumbers the former two. Unlike the other word classes, adverbs do not inflect (Teferra, A, 2000).

Amharic coordinating conjunction	SidaamuAfooCoordinating conjunction	English coordinating conjunction
እና /ɨnä/	Naa	And
ወይም /wɛyɨmɨ/	Woyimmi	Or
ግን /gɨnɨ/	Kaayin	But
ለ /lɛ/	-raa	For
ስለዚህ /sɨlɛzihɨ/	Konnidaafira	So, therefore, As a result Consequently
ቢሆንም/bihonm	Ikkollana	'however/so'

Table 3.7 Amharic and Sidaamu Afoo coordinating conjunction

3.5.8 Conjunction

A conjunction is a word that can be used to connect two phrases, clauses and sentences. Conjunctions can be divided into coordinating and subordinating conjunctions. Coordinating conjunctions are used to connect two independent clauses (Gizaw, 2013), whereas, subordinating conjunctions are used to connect main clauses with subordinate clauses (Meshesha, G. Mamo and Million, 2011). Table 3.7-shows list of Amharic Coordinating conjunctions and their Sidaamu Afoo equivalent below.

3.5.9 Punctuation mark

Punctuation marks are symbols used in sentences and phrases to clarify the meaning. The most commonly used punctuation marks in Amharic and Sidaamu Afoo are presented as follows:

Amharic Symbol	Sidaamu Afoo symbol	English name	Purpose
(**:**) (hulet netib)	White space	Space	To separate individual word
(**::**) (arat netib)	**(.)** (cooyifoollisho)	**(.)** Period	To separate sentences
(**፣**) (netela serez)	(**,**) Taxxeessu malaate	(**,**) Comma	To separate text in a list
(**፤**) (dirb serez)	(**;**) Ittisu malaate	(**;**)Semicolon	To indicate a pause to independent clauses
(**?**) (tiyakemilket)	(**?**) (xamoote malatee)	(**?**)Question mark	To ask a question, placed at the end of the sentence
(" ")(timerte tikes)	(**" "**) or **(' ')** Maqishshu malaate	(**" "**) or **(' ')** Quotation Mark	Used around direct speeches, quotations or to emphasize a word or phrase
(**!**) (kal agano)	(**!**) Darbu malaate	(**!**) Exclamation mark	To symbolize the anger, surprise or excitement of that particular sentence.
(**-**) (serez)	(**-**) Harancho haawiittote bicamme (-)	(**-**) Hyphen	To link the parts of a compound word or phrase
((**)**) (knef)	((**)**)Gombote malaate	((**)**) Bracket	To enclose an additional inserted word

Table 3- 8 Amharic and Sidaamu Afoo punctuation mark

The Amharic language uses its own script; most of the punctuation mark used in Amharic is different from Sidaamu Afoo. Sidaamu Afoo's punctuation is the same as in English except in the case of Apostrophe marks (') and double apostrophe (''). An Apostrophe mark (') in Sidaamu Afoo, is used in writing to represent a Glottal (Lalego, D.L, 2020) sound known as 'Bicaamme' appearing between vowels or between vowels and consonants in Sidaamu Afoo. However, the Amharic language is not used it.

3.5.10 Article

The Amharic language has no articles before nouns instead of that suffixes are added to show definiteness instead of using the definite article. For Example; the girl ልጅቷ, "Girl" refers to ልጅ and the definite article "the" is replaced by the suffix "ቷ" to show definiteness. In Sidaamu Afoo, like the Amharic language, there are no articles (HUMMEL, S., & PANINI, F, 2020) that are inserted before nouns. The last vowel of the noun is dropped and suffixes. For Example; the boy Beetuu "Boy" refers to beetoo and the definite article "the" is replaced by the suffix "uu" to show definiteness.

3.5.11 Morphology

Morphology describes how words are formed in the language and it tries to discover the rules that govern the formation of words in a language (MARA M, 2018). The study of the combining morphemes to form words and deriving the morphemes from words is called morphology. The meaning of a word is refined by placing other words morphology around it. The smallest meaningfull unit of a language is a morpheme. A morpheme may either be a word or part of a word (Girma E, 2021) (Adugna, S, 2009).

Amharic is a consonant root-based language with vowels added on to the consonants. Morphemes can be added as articles, prepositions, personal pronouns, numbers, conjunctions and adjectives (John H , 2017), (Russell, 2009). The roots of verbs and most nouns in the Amharic are characterized as a sequence of consonants known as radicals.

Sidaamu Afoo is an agglutinative language, where almost all derivational and inflectional morphology involve affixation. Most affixes are suffixes, though there are case suprafixes (specifically, the accusative, oblique, and genitive case suprafixes). There are no prefixes in this language (though the negative proclitic **di** is often treated as a verb prefix rather than a clitic in the literature. Reduplication is not uncommon in Sidaamu Afoo, though not productive.

CHAPTER FOUR

4. Development of the MT system

4.1 Overview

In this chapter, corpus preparation, software tools used, and the system's overall architecture are discussed in detail.

4.2 Corpus Collection and Preparation

4.2.1 Corpus collection

A large amount of data, monolingual and bilingual, is needed to conduct statistical machine translation. The monolingual corpus is required to estimate the right word orders that the target language should look like and the bilingual, which is sentence-aligned, is used to build the translation model training and decoding purpose that determines the word (phrase) alignment between the two aligned sentences. As we mentioned above at literature review part, for this study, the corpus is collected from different online sources which contain parallel text (Amharic text and Sidaamu Afoo text).

The source includes: for **Parallel corpus:** Old and New Testaments of the Holy bible Holly Bible. And for **Monolingual corpus:** Contemporary Amharic Corpus (Gezmu, et al, 2018) and Sidaamu Afoo corpus compiled by a research team in the Informatics Faculty of Hawassa University.

The Size of the corpus for the experiment used 30,100 parallel sentences. Out of the total parallel sentences, we used 80% (24,100) of randomly selected parallel sentences for training, 10% (3,000) for tuning, and another 10% (3,000) for testing.

For word level alignment we have used the prepared corpus and MGIZA++ toolkit to align the corpus using IBM model 1-5. For morpheme level alignment we have used the prepared corpus and Morfessor.

4.2.2 Corpus preparation

Preprocessing

Before feeding the machine with data, we have to pre-process the data to facilitate the training by simplifying the repeated reading of data by changing data into the appropriate format. The task of data pre-processing includes several activities like Tokenization, which replaces sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security. (For example, Insert a glottal stop symbol ("ʔ") instead of an Apostrophe symbol (') in Sidaamu Afoo. An Apostrophe symbol (') in Sidaamu Afoo, is used in writing to represent a Glottal

(Lalego, D.L, 2020) sound known as ‘Bicaamme’ which appears between vowels or between vowels and consonants. It assumes a consonant sound that is used for those who wait and care for words in the language). The tokenized paragraph inserts space between words and punctuation marks. Tokenizing of corpus makes use of a Perl script. True-casing: is the problem in natural language processing (NLP) of determining the proper capitalization of words where such information is unavailable. This applies to the Sidaamu Afoo language. Cleaning: that cleans up a parallel corpus, so it works well with the training script. It performs removes empty lines and removes redundant space characters. The cleaning step is removing longer sentences with more than 80 words.

The data is finally detokenized after it has been tokenized. It was obtained when we compared the blue score values to sacreBLUE.

4.2.3 Morpheme-based Dataset preparation

Unsupervised morpheme segmentation

Morphological segmentation is an important sub-task in many natural language processing (NLP) applications, aiming to break words into meaning-bearing sub-word units called morphemes (Ryan Cotterell, et al, 2016). Many methods in NLP, information retrieval, and text mining make use of word-level information. However, since the number of word forms in a language is often infinite, morphological preprocessing may be vital for such methods to generalize to new forms. Morphological segmentation may allow us to break them down into more familiar units that have been observed before in the data.

In our case, morpheme-based translation was conducted using the same corpus. But dataset preparation for morpheme-based translation is different from that of word-based, for word segmentation, done using an unsupervised segmentation tool called Morfessor.

The steps below are what morfessor does when segmenting (Creutz M. and Lagus K, 2015). The first step is to create a model for both corpus using morfessor script using training and test data sets. Then, the model is used to segment an input corpus. Using the created model and morfessor - segment script, text corpus as an input for both languages redirects to a new file. The third step is to reassemble the segmented new file text using Python script. The fourth steps are to provide the morpheme-aligned sentences to the MGIZA++.

The following syntax is used to create the model for the Amharic and Sidaamu Afoo corpora using training and segment input corpora:

I. morfessor-train

- ➲ Morphessor -train corpus.am -s morphmodelam.bin (for amharic)
- ➲ Morphessor -train corpus.si -s morphmodelsi.bin (for sidaamu afoo)

II. morfessor-segment

➲ Morphessor-segment -l morphmodel.bin corpus.si-am.am > train.si-am.morf.am (for amharic)

➲ Morphessor-segment -l morphmodel.bin corpus.si-am.si > train.si-am.morf.si (for Sidaamu Afoo)

Table 4-1 shows in the below each word is represented with a morpheme including prefixes and suffixes. The list of words cannot be translated row by row before being concatenated at the sentence level.

Amharic	Sidaamu Afoo	Amharic	Sidaamu Afoo
በላይ በኩል ያለ _ውን የ _ግዮን _ን ምንጭ መውጫ ገድበ ውሃው _ን ወደ ምዕራብ , ወደ ዳዊት ከተማ , በ _ቦይ እንዲ _ወርድ ያደረገ ራሱ	Ki _ncho kurkur _e alee _nni noo _hu Gi _yooni Xa _shshi galchi _maa _nni hige lola _he Yerusaalame eanno gede assino _hu _no	እኔ ለምን ተስፋ እ _ቈ _ር _ጣለሁ ? ለምን _ስ እጨነቃለሁ . በ _እግዚአብሔር እተማመ _ና _ለሁ ስለሚ _ረዳኝ _ም እርሱ _ን አመሰግናለሁ . ሕዝቅያስ ነው ; ይህ ሕዝቅያስ ሊሠራ ባቀደ _ው ነገር ሁሉ የ _ተሳካለት ሰው ነበር ;	Lubbo _ʔya ballo , mayi _ra dadilla _tta ? Mayi _ra da _assatta _e? Xaa _no gala _gale Magano gatisa _ancho _ʔya galaxxe _emmo _na , ati _no iso hexxi . Hiziqiyaasi _iti . Isi _ra Assino coyi baalu qiniino _si .

Table 4-1 sample morpheme generated for Amharic and Sidaamu Afoo.

4.3 Architecture of the System

This section shows the architecture of the bi-directional Sidaamu Afoo – Amharic statistical machine translation. The architecture of the SMT system is shown in Figure 4.1. The system takes bilingual and monolingual corpora as inputs. The data are preprocessed with different preprocessing tools such as Tokenization, Normalization & Cleaning. The Translation Modeling component takes the part of the bilingual corpus as input and produces the translation model for the given language pairs. The Decoding component takes the language model, translation model, Tuning Set and the source text to search and produce the best translation of the given text. The Evaluation component of the system takes the system output and the reference translation of the input to produce the metric that compares the system output and the reference translation.

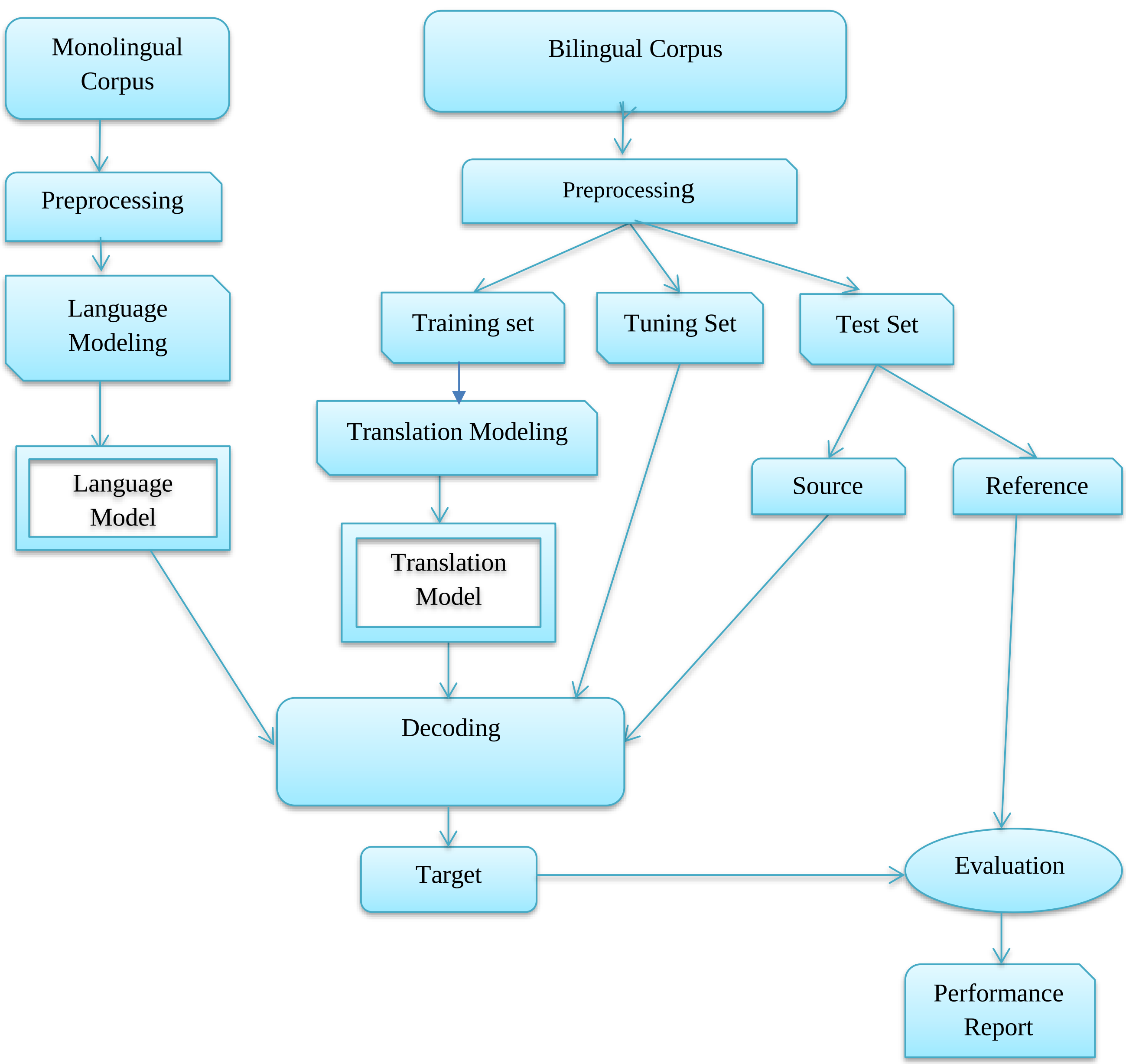


Figure 4.1 Architecture of the system

4.3.1 Language model

Language modeling (LM) is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. Language models analyze bodies of text data to provide a basis for their word predictions (Lorentzen, 2022).

Language models determine word probability by analyzing text data. They interpret this data by feeding it through an algorithm that establishes rules for context in natural language. Then, the model applies these rules in language tasks to accurately predict or produce new sentences. The model essentially learns the features and characteristics of basic language and uses those features to understand new phrases (Lorentzen, 2022). Language modeling is to estimate the probability distribution of various linguistic units, e.g., words, sentences, etc. A statistical language model is a probability distribution over sequences of words. There are various software packages available to build a statistical language model. The IRSTLM language modeling toolkit is one of them and is used to train the language model for this study. It uses for building and applying statistical language models. An appropriate five-gram language model was built. For the language model, we used monolingual corpora which is automatically generated by combining the train and tune set. 30,100 sentences are used for both Amharic and Sidaamu Afoo language modeling. It is the same amount used for both word-based and morpheme-based Machine Translation.

4.3.2 Translation Model

Translation models describe the mathematical relationship between two or more languages. We call them models of translational equivalence because the main thing that they aim to predict is whether expressions in different languages have equivalent meanings (Melamed, I. D., & Wang, W, 2004).

The translation model assigns the probability of a given source language which will generate the target language sentence. For the translation model we used the results of MGIZA++ & Morfessor for word level and morpheme level aligned corpus. For the translation model, we used a bilingual corpus.

4.3.2.1 Alignment tools

Word Alignment is the task of finding the correspondence between source and target words in a pair of sentences that are translations of each other (Cooper, 1989) MGIZA++ is software based on the famous word-alignment software, it is a freely available, widely used SMT toolkit that is used for both word and morpheme-level aligned corpus for the translation model by using IBM models (1-5). In this research, the MGIZA++ toolkit is used for word and morpheme alignment.

4.3.2.2 Decoder

Decoding is a searching problem that can be reformulated to search for the shortest path in an implicit graph. A decoder searches for the best sequence of transformations that translates the source sentence to the corresponding target sentence. It looks up all translations of every source word or phrase, using a word or phrase translation table, and recombines the target language phrases that maximize the translation model probability multiplied by the language model probability. For this study, the decoder, by following the above procedure performs the translation process from both directions.

4.3.2.3 Tuning

Tuning refers to the process of finding the optimal weights for this linear model, where optimal weights are those which maximize translation performance on a small set of parallel sentences (the tuning set) (Philipp K, 2022 & 2013). For this study, we used 3000 sentences for both Sidaamu Afoo and Amharic sentences. The bilingual corpora used for the tuning are preprocessed with tokenization and cleaning processes.

4.3.2.4 Evaluation

Evaluation is done using the sacer BLEU packege: BLEU, ChrF2 and TER metrics is an algorithm for evaluating the quality of text that has been machine-translated from one natural language to another (Teshome, E, 2013). To evaluate the performance of the prototype, using a reference translation corpus and the translation quality of the system output which was translated can be evaluated by using sacerblue packege: which supports Multi-reference Evaluation. Multi-reference Evaluation helps to regulate the BLUE score results.

CHAPTER FIVE

5. Experimentation

5.1 Introduction

This chapter discusses the experimental results of this thesis work by showing the experimental setups, and performance testing results of the experimental systems using BLEU, ChrF2, and TER metrics as implemented in SacreBLEU[1] package. The study is based on designing and implementing a bi-directional Sidaamu Afoo-Amharic machine translation system at a word-level and morpheme-level.

5.2 Datasets

In order to accomplish the objective of this thesis work, we conduct four experiments using word and morpheme based translation unit with statistical machine translation for Sidaamu Afoo - Amharic language pairs. The first two experiments focus on word based SMT and the next two experiments focuses on morpheme based translation using unsupervised morphological segmentation tool, Morfessor. For each experiment we used 30,100 parallel sentences. Out of the total parallel sentences, we used 80% (24,100) of randomly selected parallel sentences for training, 10% (3,000) for tuning, and another 10% (3,000) for testing. Additionally, for language models we used the monolingual Contemporary Amharic Corpus[2] (Gezmu, et al, 2018) and Sidaamu Afoo corpus[3].

5.2.1 Experimental setup

We used the Moses (Koehn, et al, 2007) toolkit to train the models. For word alignment, we used MGIZA++ (Och and Ney, 2003) and the grow-diag-final-and heuristic for summarization. We used the phrase-based reordering model (Koehn, et al, 2007) with three different orientations: monotone, swap, and discontinuous in both backward and forward directions, being conditioned on both the source and target languages. We removed sentence pairs with extreme length ratios and sentences longer than eighty tokens.

We used five-gram language models smoothed with the modified (Kneser, R., & Ney, H, 1995). KenLM (Heafield, 2011) language modeling toolkit was engaged for this purpose.

1 Available at: https://github.com/mjpost/sacrebleu

2 Available at: http://dx.doi.org/10.24352/ub.ovgu-2018-144

3 The Sidama Afoo corpus is compiled by a research team in the Informatics Faculty of Hawassa University.

5.2.2 Word-based bi-directional translation

The first two experiments are baseline experiments. We used word aligned corpus for the bi-directional translation process from Sidaamu Afoo to Amharic and Amharic to Sidaamu Afoo.

The first experiment is conducted to test word-based Sidaamu Afoo to Amharic machine translation. The source language is Sidaamu Afoo (input text for the translation processes) and the target language is Amharic (which is the output of the translation processes). The results are reported using BLEU, ChrF2, and TER. Experimental results acquired from BLEU were 11.3 ChrF2 was 26.4 and TER 83.4. Table 5-1 presents the sacreBleu output of the results.

System	Translation Direction	Result of experiment in sacreBLEU		
		BLUE	ChrF2	TER
Word-based	Sidaamu Afoo – Amharic	11.3	26.4	83.4

Table 5 -1 Results of Sidaamu Afoo to Amharic translation at the word-level.

The second experiment is on word-based Amharic to Sidaamu Afoo machine translation. The source language is Amharic (input text for the translation processes) and the target language is Sidaamu Afoo (which is the output of the translation processes). The results are reported using BLEU, ChrF2, and TER. Experimental results acquired from BLEU were 17.5, ChrF2 was 42.4, and TER was 81.5. Table 5-2 shows the results of Amharic to Sidaamu Afoo translation at the word-level

System	Translation Direction	Result of experiment in sacreBLEU		
		BLUE	ChrF2	TER
Word-based	Amharic – Sidaamu Afoo	17.5	42.4	81.5

Table 5-2 Results of Amharic to Sidaamu Afoo translation at the word-level.

5.2.3 Morpheme-based bi-directional translation using unsupervised morphological segmentation

The last two experiments are conducted on the basis of morphemes as a translation unit. To conduct the last two experiments the words are segmented using an unsupervised morphological segmentation tool, Morfessor.

For the third experiment we used, Sidaamu Afoo text as the source language and Amharic as the target language. Experimental results acquired from BLEU were 11.0 ChrF2 was 30.8 and TER was 84.4. Table 5-3 shows the results of Sidaamu Afoo to Amharic translation at the morpheme-level.

System	Translation Direction	Result of experiment in sacreBLEU		
		BLUE	ChrF2	TER
Morpheme-based	Sidaamu Afoo – Amharic	11.0	30.8	84.4

Table 5-3, Results of Sidaamu Afoo to Amharic translation at the morpheme-level.

In the last experiment, morpheme-based translation is done using Amharic and Sidaamu Afoo as the source and the target languages respectively. Experimental results acquired from BLEU were 18.5, ChrF2 was 51.2, and TER was 79.3. Table 5-4 shows the results of Amharic to Sidaamu Afoo translation at the morpheme-level.

System	Translation Direction	Result of experiment in sacreBLEU		
		BLUE	ChrF2	TER
Morpheme-based	Amharic – Sidaamu Afoo	18.5	51.2	79.3

Table 5-4 Results of Amharic to Sidaamu Afoo translation at the morpheme-level.

5.3 Discussion of Results

The main purpose of this study is to conducted experiments on bi-directional Sidaamu Afoo-Amharic, statistical machine translation to explore the effects of the word-based and morpheme-based translations for better performance of statistical machine translation in both directions. We conducted four different experiments by using words and morphemes as translation units. We conducted the first and second experiments at word-level bi-directional (from Sidaamu Afoo to Amharic and from Amharic to Sidaamu Afoo), and we conducted the third and fourth experiments at morpheme-level using the unsupervised morphological segmentation tool, Morfessor.

To evaluate the output of the system, we used the SacreBLEU package: BLEU, ChrF2, and TER metrics. SacreBLEU (Post, 2018) provides hassle-free computation of shareable, comparable, and reproducible **BLEU** scores.

The experiment is made by using word-based and morpheme-based SMT models. The experiments used the same datasets for each system.

The first experiment was done for word-based alignment. Experimental results acquired from BLEU were 11.3 ChrF2, 26.4, and TER 86.4 for sidaamu Afoo – Amharic. The second comparison is done for word-based alignment. Experimental results acquired from BLEU were 17.5 ChrF2, 42.4, and TER 81.5 for Amharic – sidaamu Afoo. The third comparison was done for morpheme–based alignment. Experimental results acquired from BLEU were 11.0 ChrF2 30.8 and TER 84.4 for sidaamu Afoo – Amharic. And the fourth comparison is done for morpheme–based alignment. Experimental results acquired from BLEU were 18.5 ChrF2 51.2 and TER 79.3 for Amharic – sidaamu Afoo. The summaries of sacreBLEU experimental comparison results are presented in the Table 5-5 below.

System	Translation Direction	Result of experiment in sacreBLEU		
		BLUE	ChrF2	TER
Word-based	Sidaamu Afoo – Amharic	11.3	26.4	83.4
	Amharic – Sidaamu Afoo	17.5	42.4	81.5
Morpheme-based	Sidaamu Afoo – Amharic	11.0	30.8	84.4
	Amharic – Sidaamu Afoo	18.5	51.2	79.3

Table 5-5 Summary of the experiment results

According to the experimental findings, the differences between Amharic to Sidaamu Afoo and Sidaamu Afoo to Amharic in the Word-based alignment translation were 6.2, 16, and 1.9 for BLUE, ChrF2, and TER, respectively. In the Morpheme-based alignment, the differences between Amharic to Sidaamu Afoo and Sidaamu Afoo to Amharic translation were 7.5, 20.4, and 5.1, for BLUE, ChrF2, and TER respectively.

In conclusion, the results show that morpheme-based alignment performance is better than word-based alignment, for Amharic to Sidaamu Afoo than Sidaamu Afoo to Amharic.

CHAPTER SIX

6. Conclusion and Recommendation

6.1 Conclusion

This paper work was aimed at addressing the objective of the study that meets the problem in question by designing and developing a bi-directional Sidaam Afoo-Amharic machine translation using a statistical approach.

In the development process:

- Literatures were reviewed from local and international studies. Additionally, related works were also reviewed.
- The Ethiopian language background, nouns, pronouns, adjectives, conjunctions, verb and word orders of both languages were discussed.
- Bilingual and monolingual corpus collected from different sources,
- To make the corpus suitable for the system, different preprocessing tasks applied.
- Language models for both languages by SMT approach was created, eventually evaluated by sacreBLEU

To accomplish the objective of this thesis work, we conducted four experiments:

- The first two experiments focused on word-based SMT and
- The second two on morpheme-based translation using unsupervised morphological segmentation tool; Morfessor.
- It results that unsupervised morpheme-based alignment performed better than word-based alignment from Amharic to Sidaamu Afoo.
- Accordingly, through these results the problem in question tried to be answered.

6.2 Recommendation

It is known for everybody that; automation of translation problems of one language can not be resolved on fortnight base, because it is huge task that needs adequate resource and ample time. Nevertheless this is a privet effort done by a single person at least to contribute a little bit to this huge problem with no sponsor. However, it may hopefully incite other researchers in the field for further full-flagged work.

So back to this study; it explores word-based and morpheme-based bi-directional machine translation for Sidaamu Afoo and Amharic languages using a Statistical machine translation approach. Based on the findings in the study, the following areas are recommended for further continuation of this study.

- The corpus taken for this study might not be enough; to provide better results; further research should have to be conducted using a large set of corpora.
- In this study we focused only on morphemes and words as a translation unit, further research should have to be done on other units of translation like phrases, and sentences.
- All or Most of the corpus used for this study is collected from Holly's bible and religious documents. To undertake a comprehensive experiment there is a need to prepare a corpus from different disciplines. It is a challenging task to collect and prepare data for local languages special for Sidaamu Afoo. So there is an immediate need to initiate efforts to prepare a standard corpus for local languages that can be used as a test bed to evaluate the advancement in machine translation for local languages.
- In this study, we use Morfessor for unsupervised morphological segmentation for both Sidaamu Afoo and Amharic languages. Since both languages are morphologically rich; therefore, it needs to apply rule-based morphological segmentation or machine learning algorithms to design an optimal model for segmentation.

7. References

Abdella, K. M. (2010). Speaker Dependent Speech Recognition for Sidaama Language. *(Doctoral dissertation, Addis Ababa University).*

Abel, B. (2018). "Geez to Amharic Machine Translation using Hybrid approach",. *unpublished masters thesis, Addis Ababa University, Ethiopia,.*

Addisu, B. (2016). DEVELOPMENT OF PART OF SPEECH TAGGER FOR SIDAAMA LANGUAGE. *COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES SCHOOL OF INFORMATION SCIENCE(Doctoral dissertation, Addis Ababa University).*

Adugna, S. (2009). English-Afaan Oromo Machine Translation:. *An Experiment Using Statistical Approach (Doctoral dissertation, MSc. Thesis, School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia).*(. (2009). English-Afaan Oromo Machine Translation:).

Akubazgi, G. (2017). Amharic-Tigrigna machine translation using hybrid approach. Addis Ababa University,. *Addis Ababa University, Addis Ababa*.

Antony P. (2013). Machine Translation Approaches and Survey for Indian Languages. *Computational Linguistics and Chinese Language Processing,* 47-78.

Argaw, A. A., & Asker, L. (2007, June). An Amharic stemmer: Reducing words to their citation forms. . *In Proceedings of the 2007 workshop on computational approaches to semitic languages:Common issues and resources* , 104-110.

Bahdanau, D et al. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473.l0.*

Bethelhem, M. (2002). N-gram-Based Automatic Indexing for Amharic Text. *unpublished masters thesis, Addis Ababa University, Ethiopia.*

Brandt, M. D. (2011). Developing an Icelandic to English shallow transfer machine translation system. Language.

Brown, P. et al. (1990). A statistical approach to machine translation. *Computational linguistics,*, 79-85.

Brunning, J. J. J . (2010). Alignment models and algorithms for statistical machine translation. *(Doctoral dissertation, University of Cambridge).*

Chéragui, M. A. (2012). Theoretical Overview of Machine translation. *ICWIT*, 160-169.

Cooper, R. A. (1989). Racing wheelchair rear wheel alignment. *Journal of rehabilitation research and development*, 47-50.

Creutz M. and Lagus K. (2015). Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor. *Neural Networks Research Centre, Helsinki University of Technology*, 1-27.

Daba, J. (2013). Bi-directional English-Afaan oromo machine translation using hybrid approach Addis Ababa University.

Daba, J. & Assabie, Y. (2014, September). A Hybrid Approach to the Development of Bidirectional English-Oromiffa Machine Translation. *In International Conference on Natural Language Processing*, 228-235.

Demeke, G. A. (2006). Sketch of Sidaama grammar. *Journal of African Languages and Linguistics, 20,*.( ), 87-116.

Dubey, P. (2013, December). Machine Translatiion System for Hindi-Dogri Language Pair. *In Proceeding - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA2013*, 422-425.

Ethiopian Treasures. (2013). *Addis Ababa University, Addis Ababa, Ethiopia,* . Retrieved from http://www.ethiopiantreasures.co.uk/pages/language.htm Constrained Corpus,".

Eyassu, S., & Gambäck, B. (2005, June). Classifying Amharic news text using self-organizing maps. *In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 71-78.

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation spaces, [Online].*, 291-309.

Fosler-Lussier, E. et al. (2012, June). Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics:. *Human Language Technologies.*

Gamback, B. et al. (2005). Natural language processing at the school of information studies for Africa. *In ACL 43rd Annual Meeting of the Association for Computational Linguistics; Workshop on Effective Tools and Methodologie*.

Gezmu, et al. (2018, May). Portable spelling corrector for a less-resourced language: Amharic. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Girma E. (2021). Bi-directional Amharic–Afaan Oromo Machine Translation Using Statistical Approach. *(Doctoral dissertation, St. Mary's University).*

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

Heafield. (2011, July). KenLM: Faster and smaller language model queries. *In Proceedings of the sixth workshop on statistical machine translation*, 187-197.

Heyi, G.T . (2020). Bi-directional Amharic–Afaan Oromo Machine Translation Using Statistical Approach. *(Doctoral dissertation, Addis Ababa University).*

HUMMEL, S., & PANINI, F. (2020). Bibliography of Ethiopian Semitic, Cushitic, and Omotic Linguistics XXIII:. *2019. Aethiopica, 23,,* 188-200.

Ibrahim, A., & Assabie, Y. (2013). A hybrid approach to Amharic base phrase chunking and parsing. *Addis Abeba University.*

Irfan, M. (2017, May 9). Machine translation. *Retrieved.*

Jaganadh G. (2010). Man to Machine: A tutorial on the art of Machine Translation. *http://www.slideshare.net/jaganadhg/a-tutorialon-machine-translation*.

John H . (2017). Spencer Ethiopia at Bay. *A Personal Account of the Haile Sellassie Years, Online book uploaded*.

Jurafsky, D & Martin, J. (2002). Speech and Language Processing. *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*

Kassa, T. (2018). Morpheme-Based Bi-Directional Ge'ez-Amharic Machine Translation. *Addis Ababa University, Addis Ababa, Ethiopia*..

Kawachi. (2007). A grammar of Sidaama (Sidamo), a Cushitic language of Ethiopia. *State. University of New York at Buffalo*.

Kneser, R., & Ney, H. (1995, May)). Improved backing-off for m-gram language modeling. *In international conference on acoustics, speech, and signal processing*, 181-184) IEEE.

Koehn, et al. (2007, June). Moses: Open source toolkit for statistical machine translation. *In Proceedings of the 45th annual meeting of the association for computational*.

Koehn, P. (2009 &2010). Statistical machine translation. *Cambridge University Press.*

Kothari C. (2004). Research methodology, india:. *New age international (p) limited, publishers*.

Lalego, D.L. (2020). Development of Morphological Analyzer for Sidaamu Afoo. *Addis Ababa University, Addis Ababa Ethiopia.* .

Li, P. (2013). A survey of machine translation methods. *TELKOMNIKA Indonesian Journal of Electrical Engineering,*, 11(12), 7125-7130.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR),* , 1-49.

Lopez, M. et al. (2007). Sustainable development and corporate performance. *A study based on the Dow Jones sustainability index. Journal of business ethics,*, 285-300.

Lorentzen, B. (2022). Social Biases in Language Models. *Gender Stereotypes in GPT-3 Generated Stories.*

M. W/Giorgis. (2019). Sidaamu Afiinni: Reading and writing in sidaamu Afoo. . *Addis Abbaba: Mega Publishing and Distribution P.L.C.,* .

Mahata, S. K., Das, D., & Bandyopadhyay, S. (2019). Machine translation using recurrent neural network on statistical machine translation. *Journal of Intelligent Systems*, 28(3), 447-453.

MARA M. (2018). ENGLISH-WOLAYTTA MACHINE TRANSLATION USING STATISTICAL APPROACH. *ST. MARY'S UNIVERSITY SCHOOL OF GRADUATE STUDIES, ADDIS ABABA, ETHIOPIA,*.

Martin, J. H., & Jurafski, D. (2000). An Introduction to Natural Language Processing.

Melamed, I. D., & Wang, W. (2004). Statistical machine translation by generalized parsing. *arXiv preprint cs/0407005*.

Meshesha, G. Mamo and Million. (2011). Parts of Speech Tagging for Afaan Oromo”. *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*.

Mohammed, A. K., & Pandey, A. (2018, January). Sphinx based speech recognition application for Sidama language. (IEEE., Ed.) *In 2018 2nd International Conference on Inventive Systems and Control (ICISC}*, (pp. 197-202).

Mulugeta, D. (2015, July)). Geez to Amharic Automatic Machine Translation:. *In Proceedings of the 21st International Conference on A Statistical Approach SCHOOL OF GRADUATE STUDIES COLLEGE OF NATURAL SCIENCES SCHOOL OF INFORMATION SCIENCE GEEZ TO AMHARI.*

Och and Ney. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 19-51.

O'connell, M. J. (1994). Improving adjuvant therapy for rectal cancer by combining protracted-infusion fluorouracil with radiation therapy after curative.

Okpor. (2014). Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI),*, 11(5), 159.

Papineni, K. et al. (2002, July). Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311-318.

Philipp K. (2022 & 2013, April 18,). MOSES statistical machine translation system. *University of Edinburgh*.

Post, M. (2018). A call for clarity in reporting BLEU score. *arXiv preprint arXiv:1804.08771*.

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics,*, 44(3), 393-401.

Russell, S. (2009). Amharic ማዕዝ TESL 539:. *Language Group Report Spring retrievedfrom*.

Ryan Cotterell, et al. (2016). "Morphological Segmentation InsideOut,". *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,*, 2325–2330.

Sidama Information and Culture Department . (2007). Sidaama-Amharic-Englishdictionary.pdf. *Master Printing Pres,Addis Ababa,*.

Slocum, J. (1985). A survey of machine translation. *Its history, current status and future prospects. Computational linguistics,*, 11(1), 1-17.

Solomon, Y. (2017, Solomon, Y., Meshesha, M., & Endale, W. (2017). . vol, 3, 73-77). Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation. 73-77.

Syahrina, A. (2011). Online Machine Translator System and Result Comparison.

Teferra, A. (2000). A grammar of Sidaama. *Hebrew University of Jerusalem.*

Teklewold, A. A. (2013). Automatic summarization for Amharic text using open text summarizer . *(Doctoral dissertation, Addis Ababa University).*

Teshome, E. (2013). Bi-directional Amharic–Afaan Oromo Machine Translation Using Statistical Approach. *(Doctoral dissertation, St. Mary's University).*

Tripathi, S., & Sarkhel, J. K. (2010). Approaches to machine translation:. *https://www.researchgate.net/publication/228574546.*, .

Vani K. (2019). Example-Based Machine Translation. *Internet: http://dspace.cusat.ac.in/jspui/*.

White et al. (1994). Comparison of 20 exclusive reactions at large t. *Physical Review D, 49(1),*, 58.

አማረ, ጌ. (2010). የአማርኛ ሰዋስው በቀላል አቀራረብ,. *የተሻሻለ ሁለተኛ ዕትም, አዲስ አበባ ዩኒቨርሲቲ ቢ.ኢ. ማተሚያ ቤት, Addis Ababa.*

ይማም, ባ. (2010). አጭርና ቀላል የአማርኛ ሰዋስው. *አልፋ አታሚዎች, Addis Ababa.*

Appendix

Appendix I

The Sample Translation input (a) and output (b) from Sidaamu Afoo to Amharic for **word -based translation**

Input text is Sidaamu Afoo (a)

dev.si-am.tok.si dev.si-am.tok.am

1 Hakkunni gedensaanni moote Maganu manchira , " Amo ane ledo minira haʔneenna mitore afooho wodhattona ; afanshano eemmohena " yiisi .
2 Qullaawu Ayyaani widoonni qolte borojjichikki anninke Daawitira , ‹ Daga mayira hanqitu ? Gosano mayira beebbatenni gano gantu ?
3 Abirihaamino Efirooni yiinota macciishshe , mannu afanna kulino garinni Abirihaami 400 kinchu birra keene uyisi .
4 Albasiinni noo beleqinni , kaadenna giirate buwani fule , duʔmu gomichi giddoonni hige sai .
5 Laabano rodoosi beettiha Yaiqoobi duduwo macciishshe haaʔrasira rahe mari ; hanqafe sunqisihu gedensaanni , minisira haaʔre haʔrisi . Yaiqoobino baalanka coye Laabara hasaawi .
6 Sagaleno ite dhaabbati . Shiima barra Demaasiqote heedhanno rosaanowa taxxi yii .
7 Yesuusi qole , " Mittu manchi jawa jila jile , lowo manna woshshino .
8 Qoleno , reyattowanna faradi gaarekki uurrite gattannowa , qunqumu gede asse halaʔladu xawira wigidiide tugannohe ; ate ikkitonni kainohunni roorrichikki saalanno .
9 Roduuwaʔya ! Kiʔne Yihudu gobbara heedhannota Yesuusi Kiristoosinni Maganu beetekiristaane ikkitinota lawishsha assidhinoonni . Insara Ayihude qarra iillishshunsante gede , kiʔnerano gobbaʔne manni qarra iillishinoʔne .
10 Kuri kaphaano , buuxa dandiitannokki coye xontanno ; insa woshshaadamatenna baʔate kalaqantino ; ayyaani galte heedhannonna heda dandiitannokki saada gedeereeti ; insano reyitanno garinni baʔanno .
11 Simiooni Pheexiroosi tenne ikkito lae Yesuusi albaanni guluphe , " Mooticha , ani cubbaataamohona anewiinni hoʔli " yiisi .
12 Mitu mitu manni cubbi xawoho leellanni noo daafira , albillitte daggasira noo yoo leellishanno . Wolu manni cubbi kayinni keeshshe leellanno .
13 Hakkiicho heedhe bayira beettose iltu ; wosinu galanno dargi nookki daafira , hoccunni xaaxxe hoowete saadate hayisso worranni badiri aana goxissu .
14 Bunshe malannoha , xagaraamoho yinanni .
15 Kiʔnenna maateʔne , Leewaawootunna giddoʔne heeʔrannohu manni gobba manni Maganiʔne Kaaliiqi uyinoʔnehu danchu coyi baalunkunni jiʔlitinanni hagiidhe .
16 Iseno , Kaaliiqi Muse hajajino garinni , umoho xaaxiʔnanni koffiyira albaanni bulu xilatinni usurtu .
17 Insa mereero xonamanno gede assitinoonnihu qullaawu suʔmiʔya dagate mereero ayirranno gede asseemmo yannara , ani Kaaliiqa ikkoommota affanno . Ani Mootichu Kaaliiqi konne coye coyiʔroommo . Ani Qullaawa ikkoommota daga kiʔne widoonni higge affanno gede asseemmo .
18 Iyosiyaasi moohihunni tonaa settikki diro , daga magansidhanno magansiʔra hune , gobbanna Qullaawa Mine keereensihu gedensaanni , Qullaawa Mine gatamartanno gede Atsaaliya beetto Saafaani , Yerusaalame gashshe nooha Maiseyanna borreessaancho Iyoahaazi beetto Yoaahi soyi .
19 Yesuusi Galiilaho heeʔri yannara , kuni meenti Yesuusi hoode haʔrino . Qoleno towatinosiho . Ledosi Yerusaalame dayinohu wolu lowo meentino hakko no .
20 Ledeno , " Gordunni daannohu woʔmunkura aliidiho ; uullate aanihu uullaho ; uullare coyiʔranno ; gordunni daannohu kayinni woʔmunkura aliidiho .
21 Hattoo manni kokkensara calla ikkinnina Mootichinke Kiristoosira soqqamannoha diʔikkino . Owaatunna awishshu qaalinni beebbaataamo manna dogganno .
22 Shiima yanna gedensaanni , mootichisi galte Yooseefi laʔe halchite , " Ane ledo goxi " yitusi .
23 " Kaaliiqa Moottichaʔya , ati Magano ikkootto daafira , coyiʔrootto qaale wonshatto ; xaano ledde tenne dhaggete hexxo anera oottoe .
24 Mootichu hajajo uyi ; lowo meenti duduwosi gobba tuqisi .
25 Elaami sirchinni : Maataaniya , Zakkaariyaasi , Yihieeli , Abiidi , Yeremootiitinna Eeliyaati .
26 Ayino Dukkanunniwa shiqqi yee ane hanqise Israeelete manna gudisiisannokki gede , Leewaawootu dukkannansa Qullaawu Dukkani gaangaawira qasidhe agartona " yii .
27 Kaaliiqi Iyyaasura , " Insa waajjitooti ; ani woʔma qeelle oommohe . Insa giddonni mittuno ittisannohehu diheeʔranno " yiisi .
28 Ita hisanyoote , aga hagiirre uyitanno ; woʔmanka kayinni woxunni hidhinanni .
29 Shiima yanna gedensaanni Gezeeri heedhanno Filisxeemooti ledo olu kai . Hatte yannara Husha ilaminohu Sibekayi , Refayimi ilama giddonni Siipayi yinanniha mitto haxirigga mancho shii ; Filisxeemootuno qeelantu .

The output text is Amharic (b)

dev.si-am.tok.si

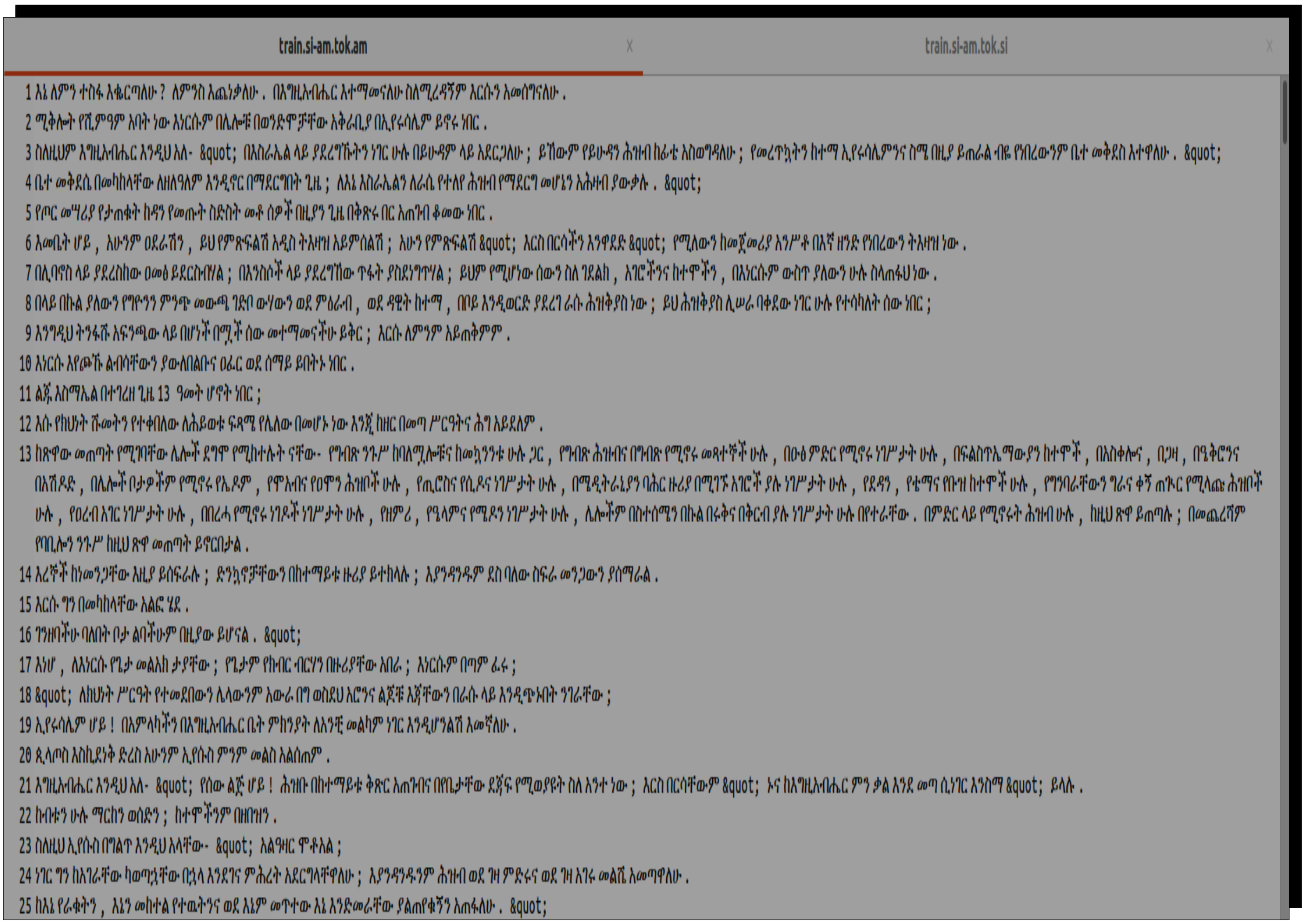
dev.si-am.tok.am

1 ከዚህም በኋላ ንጉሡ ነቢዩን " ና ከእኔ ጋር አብረህ ወደ ቤቱ ግባ ; እህልም ቅመስ ; ስጦታ እሰጥሃለሁ " አለው .
2 በመንፈስ ቅዱስ አማካይነት በአገልጋይህ በአባታችን በዳዊት አፍ እንዲህ ብለህ የተናገርከ አንተ ነህ ; " አሕዛብ ስለምን ተቈጡ ? ሕዝቦችስ ስለምን በከንቱ ዶለቱ ?
3 አብርሃምም በነገሩ ተስማምቶ ዔፍሮን በሕዝቡ ፊት በተናገረው መሠረት በነጋዴዎች ሚዛን ልክ አራት መቶ ብር መዝኖ ሰጠው .
4 በፊቱ ካለው ነጸብራቅ በረዶና የእሳት ብልጭታ ጥቅጥቅ ያለውን ደመና ሰንጥቆ ወጣ .
5 ላባም የእኅቱ ልጅ ያዕቆብ መምጣቱን በሰማ ጊዜ ሊቀበለው ፈጥኖ ሄደ ; ዐቅፎ ከሳመውም በኋላ ወደ ቤት አመጣው ; ያዕቆብ የሆነውን ሁሉ ለላባ ነገረው ;
6 መብልም በላና በረታ ; በደማስቆ ከነበሩት ደቀ መዛሙርት ጋር ለጥቂት ቀኖች ቈየ .
7 ኢየሱስ ግን እንዲህ አለው ; " አንድ ሰው ትልቅ የእራት ግብዣ አዘጋጅቶ , ብዙ ሰዎችን ጠራ ;
8 እርሱም ወደምትሞትበትና ሠረገሎችህም ቆመው ወደሚቀሩበት ቦታ እንደ ኳስ ጠቅልሎ ወደ ሰፊው ሜዳ ይወረውርሃል ; ለጌቶችህም ኀፍረት ትሆናለህ .
9 ወንድሞች ሆይ , እናንተም በይሁዳ ምድር የሚገኙትን በኢየሱስ ክርስቶስ የእግዚአብሔር አብያተ ክርስቲያን የሆኑትን አርአያ ተከትላችኋል ; እነርሱ በአይሁድ መከራ እንደ ደረሰባቸው ሁሉ እናንተም ከገዛ ወገኖቻችሁ መከራ ደርሶባችኋል .
10 እነዚህ ሐሰተኞች መምህራን ግን ለማስተዋል ያቃታቸውን ነገር ይሳደባሉ ; እነርሱ ለመጠመድና ለመገደል እንደ ተወለዱ , በተፈጥሮ ስሜት እንደሚኖሩና አእምሮ እንደሌላቸው እንስሶች ናቸው ; እንስሶች እንደሚጠፉ እነርሱም ፈጽመው ይጠፋሉ .
11 ስምዖን ጴጥሮስ ይህን ባየ ጊዜ በኢየሱስ ፊት ተንበርክኮ , " ጌታ ሆይ , እኔ ኃጢአተኛ ስለ ሆንኩ ወደ እኔ አትቅረብ ! " አለው .
12 የአንዳንድ ሰዎች ኃጢአት ግልጥ በመሆኑ አስቀድሞ ፍርዱን ያመለክታል ; የሌሎች ሰዎች ኃጢአት ግን የሚገለጠው ዘግየት ብሎ ነው .
13 በዚያም የበኵር ልጅዋን ወለደች ; በመታቀፊያ ጨርቅ ጠቀለለችው ; በእንግዶች ማደሪያ ስፍራ ስላላገኙ በበረት በከብቶች መመገቢያ ግርግም ውስጥ አስተኛችው .
14 ዘወትር ክፋትን ለማድረግ የሚያሤር ሁሉ " ተንኰለኛ " የሚል የቅጽል ስም ይወጣለታል ;
15 ለአንተና ለቤተሰብህም እግዚአብሔር ባደረገላችሁ መልካም ነገር ሁሉ ደስ ብሎአችሁ አመስግኑ ; ሌዋውያንና ከእናንተ ጋር የሚኖሩ መጻተኞች በዚህ የምስጋና ሥርዓት አፈጻጸም ላይ አብረው ይገኙ .
16 እርሱንም እግዚአብሔር ሙሴን ባዘዘው መሠረት ከመጠምጠሚያው ፊት በሰማያዊ ክር አሰሩት .
17 በሕዝቦች መካከል ተሰድቦ የነበረውን , ማለት እናንተ አሰደባችሁት የነበረውን የታላቁን ስሜን ቅድስና አሳያለሁ ; ከዚያ በኋላ በእናንተ አማካይነት በእነርሱ ፊት ቅድስናዬን በምገልጥበት ጊዜ ሕዝቦች እኔ እግዚአብሔር መሆኔን ያውቃሉ .
18 ኢዮስያስ በነገሠ በዐሥራ ስምንተኛው ዓመት የአሕዛብን አምልኮ በማጥፋት ምድሪቱንና ቤተ መቅደሱን ካነጻ በኋላ , የእግዚአብሔርን ቤተ መቅደስ ያድሱ ዘንድ የአጻልያን ልጅ ሳፋንን , የኢየሩሳሌም አስተዳዳሪ የሆነውን ማዕሤያንና የኢዮአሐዝን ልጅ ጸሐፊውን ዮአሕን ላከ ;
19 እነርሱ ኢየሱስ በገሊላ በነበረበት ጊዜ , ይከተሉትና ያገለግሉት ነበር . እንዲሁም ከኢየሱስ ጋር ወደ ኢየሩሳሌም የመጡ ብዙ ሌሎች ሴቶች ነበሩ .
20 ከላይ የሚመጣው ከሁሉ በላይ ነው ; ከምድር የሆነ ምድራዊ ነው ; የምድርንም ነገር ይናገራል ; ከሰማይ የሚመጣው ከሁሉ በላይ ነው .
21 እንደእነዚህ ዐይነቶቹ ሰዎች ራሳቸውን እንጂ ጌታችንን ክርስቶስን አያገለግሉም . በለዘቡና በሚያቈላምጡ ቃላት የየዋሆችን ልብ ያታልላሉ .
22 ከጊዜ በኋላ የጌታው ሚስት ዮሴፍን ተመልክታ ስለ አፈቀረችው , " ና ከእኔ ጋር አብረን እንተኛ " አለችው .
23 " አሁንም ልዑል እግዚአብሔር ሆይ , አንተ አምላክ ነህ ; ቃልህ ታማኝ ነው ; እነሆ , አሁንም ይህን መልካም የተስፋ ቃል ለአገልጋይህ ሰጥተሃል .
24 እግዚአብሔር ትእዛዝ ሰጠ ; ብዙ ሴቶችም ዜናውን አሠራጩ ;
25 ከዔላም ጕሣ- ማታንያ , ዘካርያስ የሒኤል , ዐብዲ , የሬሞትና ኤልያ ;
26 ነገር ግን ማንም ሰው ወደ ድንኳኑ በመቅረብ ቊጣዬን እንዳያስነሣ የእስራኤልን ማኅበር እንዳያስፈጅ ሌዋውያን በድንኳኑ ዙሪያ ሰፍረው ይጠብቁት . "
27 እግዚአብሔርም ኢያሱን " እነርሱን አትፍራቸው ; እኔ አንተን በእነርሱ ላይ ድልን አቀዳጅሃለሁ ; ከእነርሱ አንድም የሚቋቋምህ አይኖርም " አለው .
28 በምግብና በወይን ጠጅ ተድላ ደስታ ይገኛል ; ይሁን እንጂ ይህ ሁሉ ያለ ገንዘብ አይገኝም .
29 ዘግየት ብሎም በጌዘር የሚኖሩ ፍልስጥኤማውያን ላይ ጦርነት ተደረገ ; በዚያን ጊዜ የሑሻ ተወላጅ የሆነው ሲበካይ , ከራፋይም ወገን የነበረውን ሲፓይ ተብሎ የሚጠራውን ኀያል ሰው ገደለ ; ፍልስጥኤማውያንም ተሸነፉ .

Appendix II

The Sample Translation input (a) and output (b) from Amharic to Sidaamu Afoo for **word -based translation**

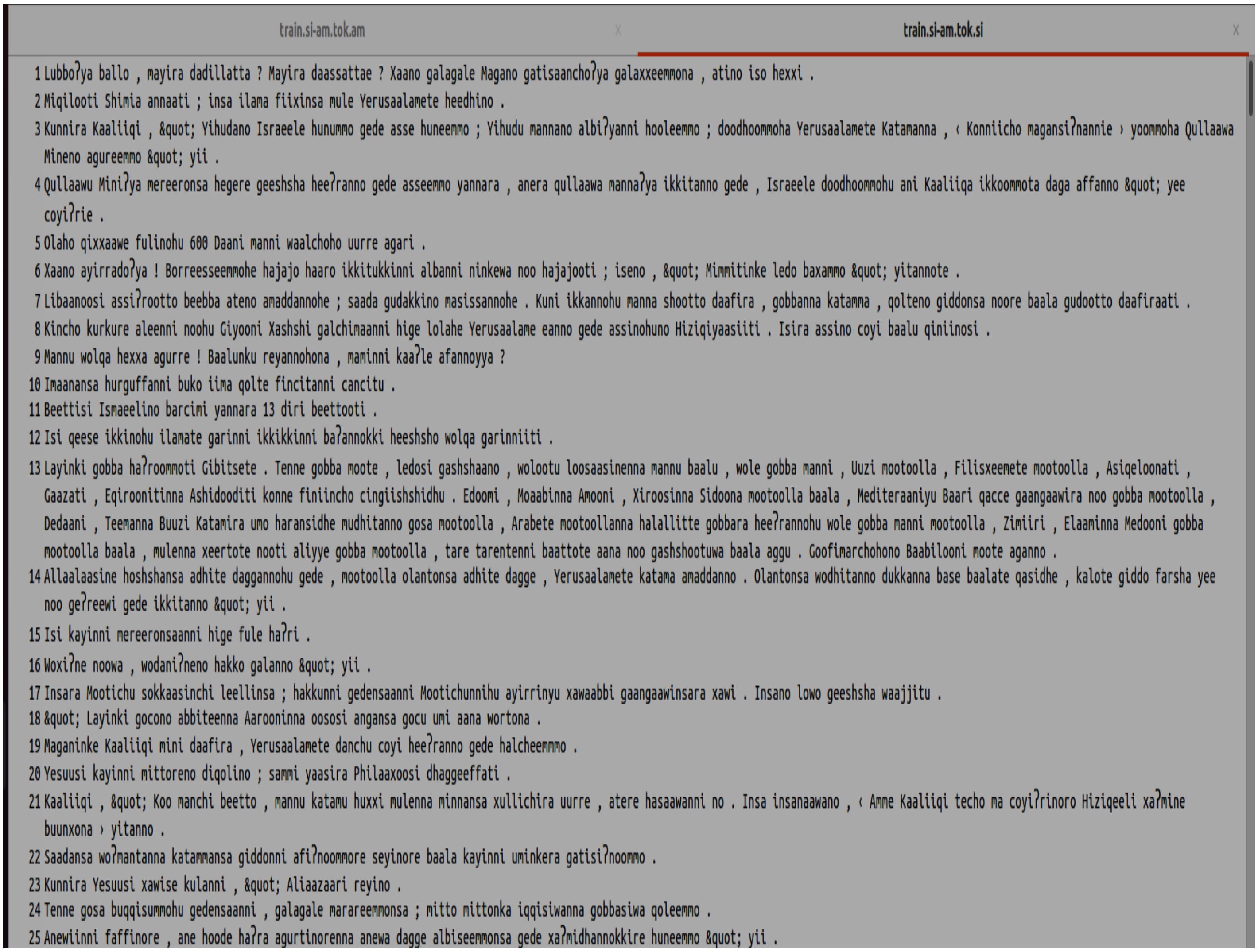
Input/source Amharic (a)



train.si-am.tok.am   train.si-am.tok.si

1 እኔ ለምን ተስፋ እቈርጣለሁ ? ለምንስ እጨነቃለሁ . በእግዚአብሔር እተማመናለሁ ስለሚረዳኝም እርሱን አመሰግናለሁ .
2 ሚቅሎት የሺምዓም አባት ነው እነርሱም በሌሎቹ በወንድሞቻቸው አቅራቢያ በኢየሩሳሌም ይኖሩ ነበር .
3 ስለዚህም እግዚአብሔር እንዲህ አለ፦ " በእስራኤል ላይ ያደረግኹትን ነገር ሁሉ በይሁዳም ላይ አደርጋለሁ ; ይኸውም የይሁዳን ሕዝብ ከፊቴ አስወግዳለሁ ; የመረጥኳትን ከተማ ኢየሩሳሌምንና ስሜ በዚያ ይጠራል ብዬ የነበረውንም ቤተ መቅደስ እተዋለሁ . "
4 ቤተ መቅደሴ በመካከላቸው ለዘለዓለም እንዲኖር በማደርግበት ጊዜ ; ለእኔ እስራኤልን ለራሴ የተለየ ሕዝብ የማደርግ መሆኔን አሕዛብ ያውቃሉ . "
5 የጦር መሣሪያ የታጠቁት ከዳን የመጡት ስድስት መቶ ሰዎች በዚያን ጊዜ በቅጽሩ በር አጠገብ ቆመው ነበር .
6 እመቤት ሆይ , አሁንም ዐደራሽን , ይህ የምጽፍልሽ አዲስ ትእዛዝ አይምሰልሽ ; አሁን የምጽፍልሽ " እርስ በርሳችን እንዋደድ " የሚለውን ከመጀመሪያ አንሥቶ በእኛ ዘንድ የነበረውን ትእዛዝ ነው .
7 በሊባኖስ ላይ ያደረስከው ዐመፅ ይደርስብሃል ; በእንስሶች ላይ ያደረግኸው ጥፋት ያስደነግጥሃል ; ይህም የሚሆነው ሰውን ስለ ገደልክ , አገሮችንና ከተሞችን , በእነርሱም ውስጥ ያለውን ሁሉ ስላጠፋህ ነው .
8 በላይ በኩል ያለውን የግዮንን ምንጭ መውጫ ገድቦ ውሃውን ወደ ምዕራብ , ወደ ዳዊት ከተማ , በቦይ እንዲወርድ ያደረገ ራሱ ሕዝቅያስ ነው ; ይህ ሕዝቅያስ ሊሠራ ባቀደው ነገር ሁሉ የተሳካለት ሰው ነበር ;
9 እንግዲህ ትንፋሹ አፍንጫው ላይ በሆነች በሚች ሰው መተማመናችሁ ይቅር ; እርሱ ለምንም አይጠቅምም .
10 እነርሱ እየጮኹ ልብሳቸውን ያውለበልቡና ዐፈር ወደ ሰማይ ይበትኑ ነበር .
11 ልጁ እስማኤል በተገረዘ ጊዜ 13 ዓመት ሆኖት ነበር ;
12 እሱ የክህነት ሹመትን የተቀበለው ለሕይወቱ ፍጻሜ የሌለው በመሆኑ ነው እንጂ ከዘር በመጣ ሥርዓትና ሕግ አይደለም .
13 ከጽዋው መጠጣት የሚገባቸው ሌሎች ደግሞ የሚከተሉት ናቸው፦ የግብጽ ንጉሥ ከባለሟሎቹና ከመኳንንቱ ሁሉ ጋር , የግብጽ ሕዝብና በግብጽ የሚኖሩ መጻተኞች ሁሉ , በዑፅ ምድር የሚኖሩ ነገሥታት ሁሉ , በፍልስጥኤማውያን ከተሞች , በአስቀሎና , በጋዛ , በዔቅሮንና በአሽዶድ , በሌሎች ቦታዎችም የሚኖሩ የኤዶም , የሞአብና የዐሞን ሕዝቦች ሁሉ , የጢሮስና የሲዶና ነገሥታት ሁሉ , በሜዲትራኒያን ባሕር ዙሪያ በሚገኙ አገሮች ያሉ ነገሥታት ሁሉ , የደዳን , የቴማና የቡዝ ከተሞች ሁሉ , የግንባራቸውን ግራና ቀኝ ጠጕር የሚላጩ ሕዝቦች ሁሉ , የዐረብ አገር ነገሥታት ሁሉ , በበረሓ የሚኖሩ ነገዶች ነገሥታት ሁሉ , የዘምሪ , የዔላምና የሜዶን ነገሥታት ሁሉ , ሌሎችም በስተሰሜን በኩል በሩቅና በቅርብ ያሉ ነገሥታት ሁሉ በየተራቸው . በምድር ላይ የሚኖሩት ሕዝብ ሁሉ , ከዚህ ጽዋ ይጠጣሉ ; በመጨረሻም የባቢሎን ንጉሥ ከዚህ ጽዋ መጠጣት ይኖርበታል .
14 እረኞች ከነመንጋቸው እዚያ ይሰፍራሉ ; ድንኳኖቻቸውን በከተማይቱ ዙሪያ ይተክላሉ ; እያንዳንዱም ደስ ባለው ስፍራ መንጋውን ያሰማራል .
15 እርሱ ግን በመካከላቸው አልፎ ሄደ .
16 ገንዘባችሁ ባለበት ቦታ ልባችሁም በዚያው ይሆናል . "
17 እነሆ , ለእነርሱ የጌታ መልአክ ታያቸው ; የጌታም የክብር ብርሃን በዙሪያቸው አበራ ; እነርሱም በጣም ፈሩ ;
18 " ለክህነት ሥርዓት የተመደበውን ሌላውንም አውራ በግ ወስደህ አሮንና ልጆቹ እጆቻቸውን በራሱ ላይ እንዲጭኑበት ንገራቸው ;
19 ኢየሩሳሌም ሆይ ! በአምላካችን በእግዚአብሔር ቤት ምክንያት ለአንቺ መልካም ነገር እንዲሆንልሽ እለምናለሁ .
20 ጲላጦስ እስኪደነቅ ድረስ አሁንም ኢየሱስ ምንም መልስ አልሰጠም .
21 እግዚአብሔር እንዲህ አለ፦ " የሰው ልጅ ሆይ ! ሕዝቡ በከተማይቱ ቅጽር አጠገብና በየቤታቸው ደጃፍ የሚወያዩት ስለ አንተ ነው ; እርስ በርሳቸውም " ኑና ከእግዚአብሔር ምን ቃል እንደ መጣ ሲነገር እንስማ " ይላሉ .
22 ከብቱን ሁሉ ማርከን ወሰድን ; ከተሞችንም በዘበዝን .
23 ስለዚህ ኢየሱስ በግልጥ እንዲህ አላቸው፦ " አልዓዛር ሞቶአል ;
24 ነገር ግን ከአገራቸው ካወጧቸው በኋላ እንደገና ምሕረት አደርግላቸዋለሁ ; እያንዳንዱንም ሕዝብ ወደ ገዛ ምድሩና ወደ ገዛ አገሩ መልሼ አመጣቸዋለሁ .
25 ከእኔ የራቁትን , እኔን መከተል የተዉትንና ወደ እኔም መጥተው እኔ እንድመራቸው ያልጠየቁኝን አጠፋለሁ . "

Output/target Sidaamu Afoo (b)



train.si-am.tok.am | train.si-am.tok.si

1 Lubboʔya ballo , mayira dadillatta ? Mayira daassattae ? Xaano galagale Magano gatisaanchoʔya galaxxeemmona , atino iso hexxi .
2 Miqilooti Shimia annaati ; insa ilama fiixinsa mule Yerusaalamete heedhino .
3 Kunnira Kaaliiqi , " Yihudano Israeele hunummo gede asse huneemmo ; Yihudu mannano albiʔyanni hooleemmo ; doodhoommoha Yerusaalamete Katamanna , ‹ Konniicho magansiʔnannie › yoommoha Qullaawa Mineno agureemmo " yii .
4 Qullaawu Miniʔya mereeronsa hegere geeshsha heeʔranno gede asseemmo yannara , anera qullaawa mannaʔya ikkitanno gede , Israeele doodhoommohu ani Kaaliiqa ikkoommota daga affanno " yee coyiʔrie .
5 Olaho qixxaawe fulinohu 600 Daani manni waalchoho uurre agari .
6 Xaano ayirradoʔya ! Borreesseemmohe hajajo haaro ikkitukkinni albanni ninkewa noo hajajooti ; iseno , " Mimmitinke ledo baxammo " yitannote .
7 Libaanoosi assiʔrootto beebba ateno amaddannohe ; saada gudakkino masissannohe . Kuni ikkannohu manna shootto daafira , gobbanna katamma , qolteno giddonsa noore baala gudootto daafiraati .
8 Kincho kurkure aleenni noohu Giyooni Xashshi galchimaanni hige lolahe Yerusaalame eanno gede assinohuno Hiziqiyaasiiti . Isira assino coyi baalu qiniinosi .
9 Mannu wolqa hexxa agurre ! Baalunku reyannohona , maminni kaaʔle afannoyya ?
10 Imaanansa hurguffanni buko iima qolte fincitanni cancitu .
11 Beettisi Ismaeelino barcimi yannara 13 diri beettooti .
12 Isi qeese ikkinohu ilamate garinni ikkikkinni baʔannokki heeshsho wolqa garinniiti .
13 Layinki gobba haʔroommoti Gibitsete . Tenne gobba moote , ledosi gashshaano , wolootu loosaasinenna mannu baalu , wole gobba manni , Uuzi mootoolla , Filisxeemete mootoolla , Asiqeloonati , Gaazati , Eqiroonitinna Ashidooditi konne finiincho cingiishshidhu . Edoomi , Moaabinna Amooni , Xiroosinna Sidoona mootoolla baala , Mediteraaniyu Baari qacce gaangaawira noo gobba mootoolla , Dedaani , Teemanna Buuzi Katamira umo haransidhe mudhitanno gosa mootoolla , Arabete mootoollanna halallitte gobbara heeʔrannohu wole gobba manni mootoolla , Zimiiri , Elaaminna Medooni gobba mootoolla baala , mulenna xeertote nooti aliyye gobba mootoolla , tare tarentenni baattote aana noo gashshootuwa baala aggu . Goofimarchohono Baabilooni moote aganno .
14 Allaalaasine hoshshansa adhite daggannohu gede , mootoolla olantonsa adhite dagge , Yerusaalamete katama amaddanno . Olantonsa wodhitanno dukkanna base baalate qasidhe , kalote giddo farsha yee noo geʔreewi gede ikkitanno " yii .
15 Isi kayinni mereeronsaanni hige fule haʔri .
16 Woxiʔne noowa , wodaniʔneno hakko galanno " yii .
17 Insara Mootichu sokkaasinchi leellinsa ; hakkunni gedensaanni Mootichunnihu ayirrinyu xawaabbi gaangaawinsara xawi . Insano lowo geeshsha waajjitu .
18 " Layinki gocono abbiteenna Aarooninna oososi angansa gocu umi aana wortona .
19 Maganinke Kaaliiqi mini daafira , Yerusaalamete danchu coyi heeʔranno gede halcheemmmo .
20 Yesuusi kayinni mittoreno diqolino ; sammi yaasira Philaaxoosi dhaggeeffati .
21 Kaaliiqi , " Koo manchi beetto , mannu katamu huxxi mulenna minnansa xullichira uurre , atere hasaawanni no . Insa insanaawano , ‹ Amme Kaaliiqi techo ma coyiʔrinoro Hiziqeeli xaʔmine buunxona › yitanno .
22 Saadansa woʔmantanna katammansa giddonni afiʔnoommore seyinore baala kayinni uminkera gatisiʔnoommo .
23 Kunnira Yesuusi xawise kulanni , " Aliaazaari reyino .
24 Tenne gosa buqqisummohu gedensaanni , galagale marareemmonsa ; mitto mittonka iqqisiwanna gobbasiwa qoleemmo .
25 Anewiinni faffinore , ane hoode haʔra agurtinorenna anewa dagge albiseemmonsa gede xaʔmidhannokkire huneemmo " yii .

Appendix III

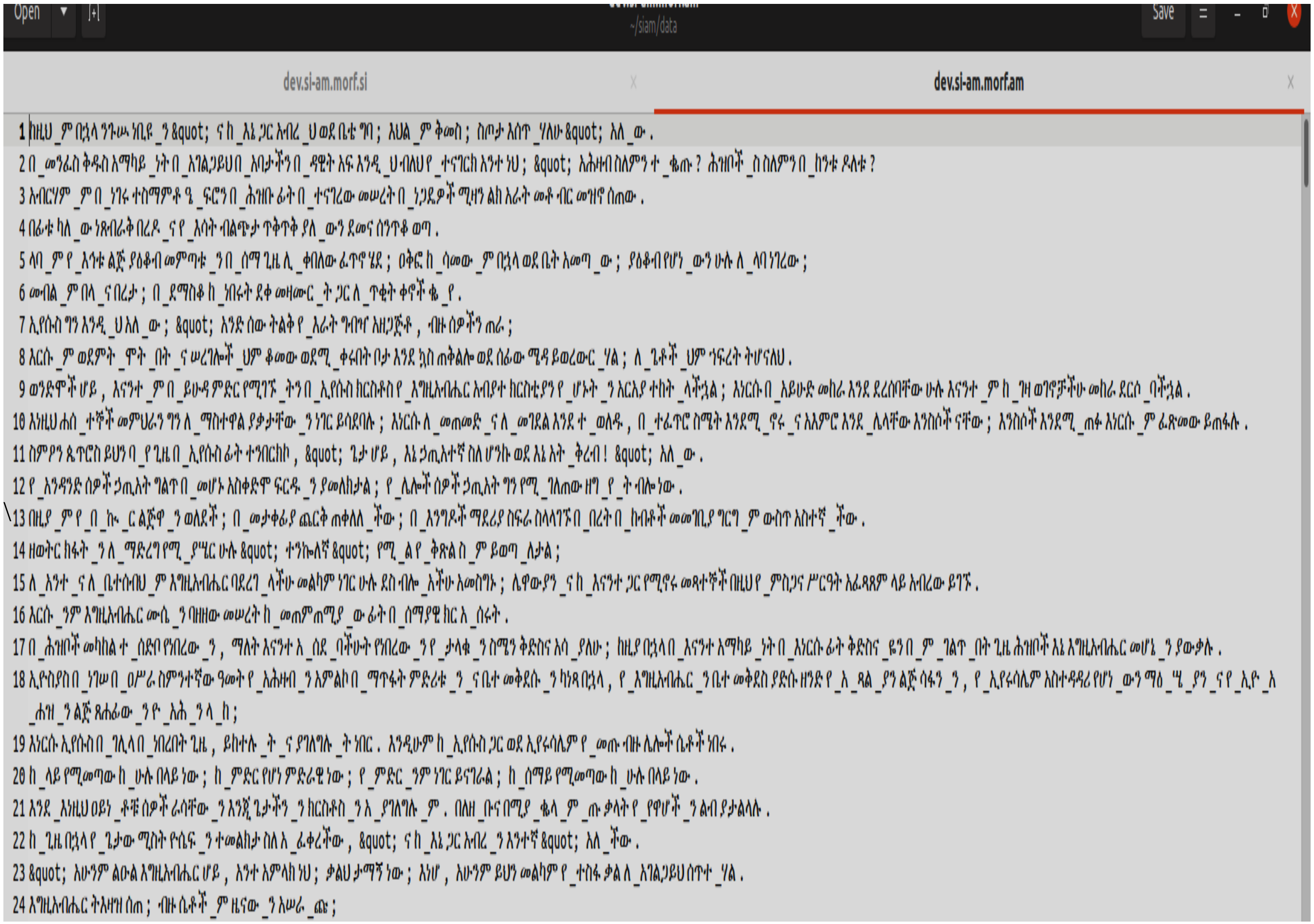
The Sample Translation input (a) and output (b) from Sidaamu Afoo to Amharic for **morpheme level alignment.**

The input text is Sidaamu Afoo (a)

dev.si-am.morf.si | dev.si-am.morf.am

```
1 Hakku _nni gedensaanni moote Maganu manchi _ra , &quot; A _mo ane ledo mini _ra ha _ʔne _enna mito _re afoo _ho wodha _tto _na ; afa _nsha _no eemmohe _na &quot; yii _si .
2 Qulla _awu Ayyaani widoonni qol _te borojjichi _kki anninke Daawiti _ra , ‹ Daga mayi _ra hanqi _tu ? Go _sa _no mayi _ra beebba _tenni ga _no gantu ?
3 Abirihaami _no Efi _ro _oni yiino _ta macciishshe , mannu afa _nna kulino gari _nni Abirihaami 400 kinchu bir _ra kee _ne uyi _si .
4 Alba _sii _nni noo be _le _qi _nni , kaade _nna giira _te buwani fule , duʔm _u gomi _chi giddoo _nni hige sai .
5 Laa _ba _no rodoo _si beetti _ha Yaiqoobi duduwo macciishshe haaʔra _sira ra _he mari ; hanqafe sunq _isi _hu gedensaanni , mini _sira haaʔr _e haʔri _si . Yaiqoobi _no baalanka coye Laa _ba _ra
  hasaawi .
6 Sagale _no ite dhaabb _ati . Shiima barra Demaasiqo _te heedhanno rosaano _wa ta _xxi yii .
7 Yesuusi qole , &quot; Mittu manchi jawa jila jil _e , lowo ma _nna woshshi _no .
8 Qol _e _no , reya _tto _wa _nna fara _di gaare _kki uurri _te gattanno _wa , qunqum _u gede asse halaʔla _du xawi _ra wigidii _de tuganno _he ; ate ikkito _nni kaino _hunni roorrichi _kki saala
  _nno .
9 Roduuwa _ʔya ! Kiʔne Yihudu gobba _ra heedhanno _ta Yesuusi Kiristoosi _nni Maganu beetekiristaane ikkitino _ta lawishsha assidhino _onni . Insa _ra Ayihude qarra iilli _shshu _nsa _nte gede , kiʔne
  _rano gobba _ʔne manni qarra iillishi _no _ʔne .
10 Kuri kapha _ano , buuxa dandiitanno _kki coye xo _ntanno ; insa woshshaada _ma _tenna baʔa _te kalaqantino ; ayyaani galte heedhanno _nna heda dandiitanno _kki saada gedee _reeti ; insa _no re
  _yitanno gari _nni baʔanno .
11 Simiooni Pheexiroosi tenne ikkito la _e Yesuusi albaa _nni gulu _phe , &quot; Mooticha , ani cubba _ataamo _hona ane _wiinni hoʔli &quot; yii _si .
12 Mi _tu mi _tu manni cubbi xawo _ho leella _nni noo daafi _ra , albillitte dagga _sira noo yoo leellishanno . Wolu manni cubbi kayi _nni keeshsh _e leellanno .
13 Hakkii _cho heedhe bayira beetto _se iltu ; wo _si _nu gala _nno dargi nookki daafi _ra , hoccu _nni xaa _xxe hoowe _te saada _te hayisso worranni ba _diri aana goxi _ssu .
14 Bunshe malanno _ha , xagara _amo _ho yinanni .
15 Kiʔne _nna maate _ʔne , Leewaawootu _nna giddo _ʔne heeʔranno _hu manni gobba manni Magani _ʔne Kaaliiqi uyino _ʔne _hu danchu coyi baalunku _nni ji _ʔli _tinanni hagiidhe .
16 Ise _no , Kaaliiqi Muse hajajino gari _nni , umo _ho xaaxi _ʔnanni ko _ffi _yi _ra albaa _nni bu _lu xi _la _ti _nni usur _tu .
17 Insa mereero xonama _nno gede assitino _onnihu qu _lla _awu suʔmi _ʔya daga _te mereero ayirranno gede asseemmo yanna _ra , ani Kaaliiqa ikkoommo _ta affanno . Ani Mootichu Kaaliiqi konne coye coyi
  _ʔroommo . Ani Qullaawa ikkoommo _ta daga kiʔne widoonni higge affanno gede asseemmo .
18 Iyosiyaasi moohi _hunni tonaa se _tti _kki di _ro , daga magansidhanno magansiʔra hu _ne , gobba _nna Qullaawa Mine keereensi _hu gedensaanni , Qullaawa Mine gatamar _tanno gede A _t _saali _ya
  beetto Saa _faa _ni , Yerusaalame ga _shshe noo _ha Ma _ise _yanna borreessa _ancho Iyoahaazi beetto Yo _aa _hi soyi .
19 Yesuusi Galiila _ho heeʔri yanna _ra , ku _ni meenti Yesuusi hoode haʔrino . Qol _e _no towat _ino _si _ho . Ledo _si Yerusaalame dayino _hu wolu lowo meenti _no hakko no .
20 Le _de _no , &quot; Gordu _nni daanno _hu woʔmunku _ra aliidi _ho ; uulla _te aani _hu uulla _ho ; uulla _re coyiʔranno ; gordu _nni daanno _hu kayi _nni woʔmunku _ra aliidi _ho .
21 Hattoo manni kokke _nsara calla ikki _nni _na Mootichi _nke Kiristoosi _ra soqqamanno _ha diʔikkino . Owaa _tu _nna awi _shshu qaali _nni beebba _ataamo ma _nna dogganno .
22 Shiima yanna gedensaanni , mootichi _si galte Yooseefi laʔ _e halchi _te , &quot; Ane ledo goxi &quot; yitu _si .
23 &quot; Kaaliiqa Mooticha _ʔya , ati Magano ikkootto daafi _ra , coyi _ʔrootto qaale wonsha _tto ; xaa _no led _de tenne dhagge _te hexxo ane _ra ootto _e .
24 Mootichu hajajo uyi ; lowo meenti duduwo _si gobba tuqi _si .
```

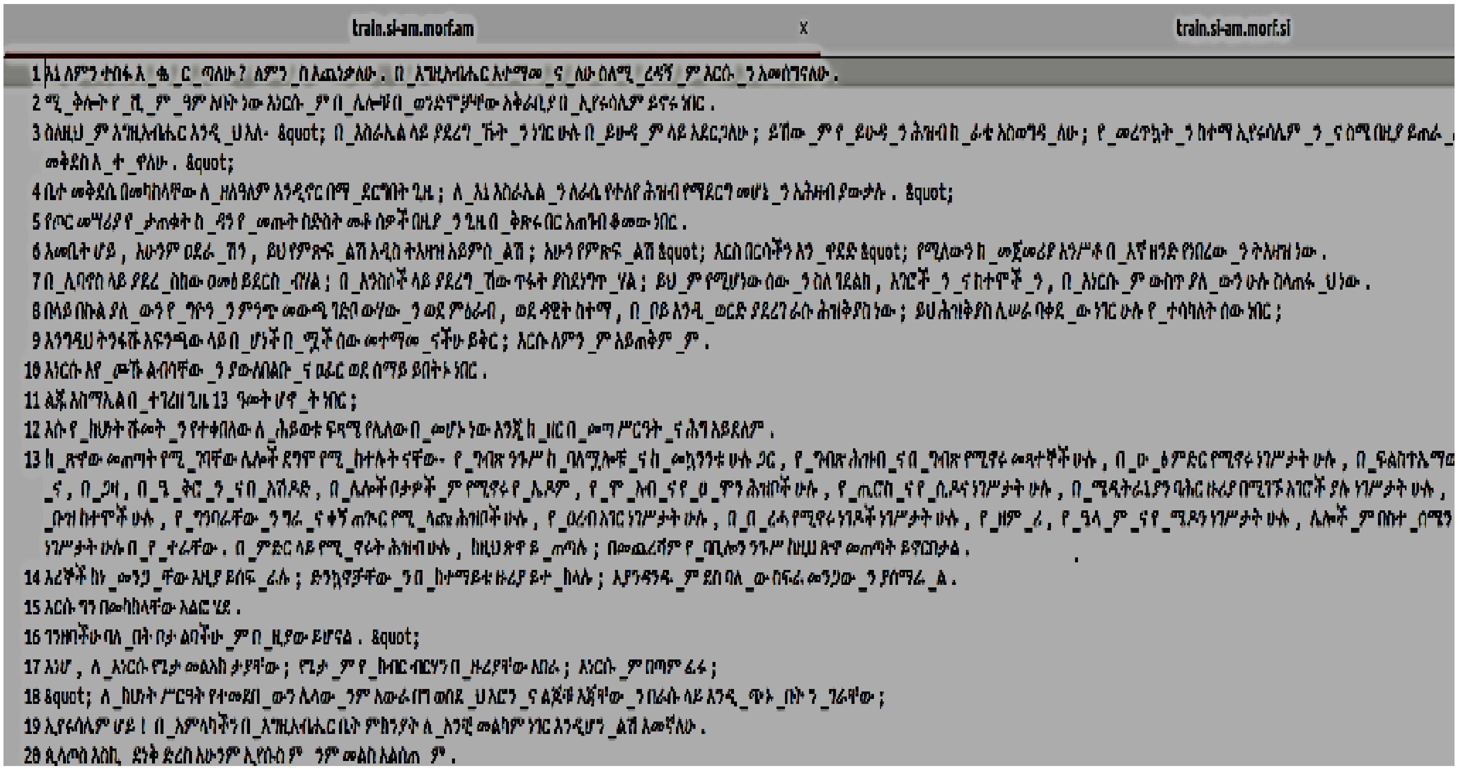
The output text is Amharic (b)



Appendix IV

The Sample Translation input (a) and output (b) from Amharic to Sidaamu Afoo for **morpheme level alignment.**

The input text is Amharic (a)



The output text is Sidaamu Afoo (b)

train.si-am.tok.am train.si-am.tok.si

1 Lubboʔya ballo , mayira dadillatta ? Mayira daassattae ? Xaano galagale Magano gatisaanchoʔya galaxxeemmona , atino iso hexxi .
2 Miqilooti Shimia annaati ; insa ilama fiixinsa mule Yerusaalamete heedhino .
3 Kunnira Kaaliiqi , " Yihudano Israeele hunummo gede asse huneemmo ; Yihudu mannano albiʔyanni hooleemmo ; doodhoommoha Yerusaalamete Katamanna , ‹ Konniicho magansiʔnannie › yoommoha Qullaawa Mineno agureemmo " yii .
4 Qullaawu Miniʔya mereeronsa hegere geeshsha heeʔranno gede asseemmo yannara , anera qullaawa mannaʔya ikkitanno gede , Israeele doodhoommohu ani Kaaliiqa ikkoommota daga affanno " yee coyiʔrie .
5 Olaho qixxaawe fulinohu 600 Daani manni waalchoho uurre agari .
6 Xaano ayirradoʔya ! Borreesseemmohe hajajo haaro ikkitukkinni albanni ninkewa noo hajajooti ; iseno , " Mimmitinke ledo baxammo " yitannote .
7 Libaanoosi assiʔrootto beebba ateno amaddannohe ; saada gudakkino masissannohe . Kuni ikkannohu manna shootto daafira , gobbanna katamma , qolteno giddonsa noore baala gudootto daafiraati .
8 Kincho kurkure aleenni noohu Giyooni Xashshi galchimaanni hige lolahe Yerusaalame eanno gede assinohuno Hiziqiyaasiiti . Isira assino coyi baalu qiniinosi .
9 Mannu wolqa hexxa agurre ! Baalunku reyannohona , maminni kaaʔle afannoyya ?
10 Imaanansa hurguffanni buko iima qolte fincitanni cancitu .
11 Beettisi Ismaeelino barcimi yannara 13 diri beettooti .
12 Isi qeese ikkinohu ilamate garinni ikkikkinni baʔannokki heeshsho wolqa garinniiti .
13 Layinki gobba haʔroommoti Gibitsete . Tenne gobba moote , ledosi gashshaano , wolootu loosaasinenna mannu baalu , wole gobba manni , Uuzi mootoolla , Filisxeemete mootoolla , Asiqeloonati , Gaazati , Eqiroonitinna Ashidooditi konne finiincho cingiishshidhu . Edoomi , Moaabinna Amooni , Xiroosinna Sidoona mootoolla baala , Mediteraaniyu Baari qacce gaangaawira noo gobba mootoolla , Dedaani , Teemanna Buuzi Katamira umo haransidhe mudhitanno gosa mootoolla , Arabete mootoollanna halallitte gobbara heeʔrannohu wole gobba manni mootoolla , Zimiiri , Elaaminna Medooni gobba mootoolla baala , mulenna xeertote nooti aliyye gobba mootoolla , tare tarentenni baattote aana noo gashshootuwa baala aggu . Goofimarchohono Baabilooni moote aganno .
14 Allaalaasine hoshshansa adhite daggannohu gede , mootoolla olantonsa adhite dagge , Yerusaalamete katama amaddanno . Olantonsa wodhitanno dukkanna base baalate qasidhe , kalote giddo farsha yee noo geʔreewi gede ikkitanno " yii .
15 Isi kayinni mereeronsaanni hige fule haʔri .
16 Woxiʔne noowa , wodaniʔneno hakko galanno " yii .
17 Insara Mootichu sokkaasinchi leellinsa ; hakkunni gedensaanni Mootichunnihu ayirrinyu xawaabbi gaangaawinsara xawi . Insano lowo geeshsha waajjitu .
18 " Layinki gocono abbiteenna Aarooninna oososi angansa gocu umi aana wortona .
19 Maganinke Kaaliiqi mini daafira , Yerusaalamete danchu coyi heeʔranno gede halcheemmmo .
20 Yesuusi kayinni mittoreno diqolino ; sammi yaasira Philaaxoosi dhaggeeffati .
21 Kaaliiqi , " Koo manchi beetto , mannu katamu huxxi mulenna minnansa xullichira uurre , atere hasaawanni no . Insa insanaawano , ‹ Amme Kaaliiqi techo ma coyiʔrinoro Hiziqeeli xaʔmine buunxona › yitanno .
22 Saadansa woʔmantanna katammansa giddonni afiʔnoommore seyinore baala kayinni uminkera gatisiʔnoommo .
23 Kunnira Yesuusi xawise kulanni , " Aliaazaari reyino .
24 Tenne gosa buqqisummohu gedensaanni , galagale marareemmonsa ; mitto mittonka iqqisiwanna gobbasiwa qoleemmo .
25 Anewiinni faffinore , ane hoode haʔra agurtinorenna anewa dagge albiseemmonsa gede xaʔmidhannokkire huneemmo " yii .

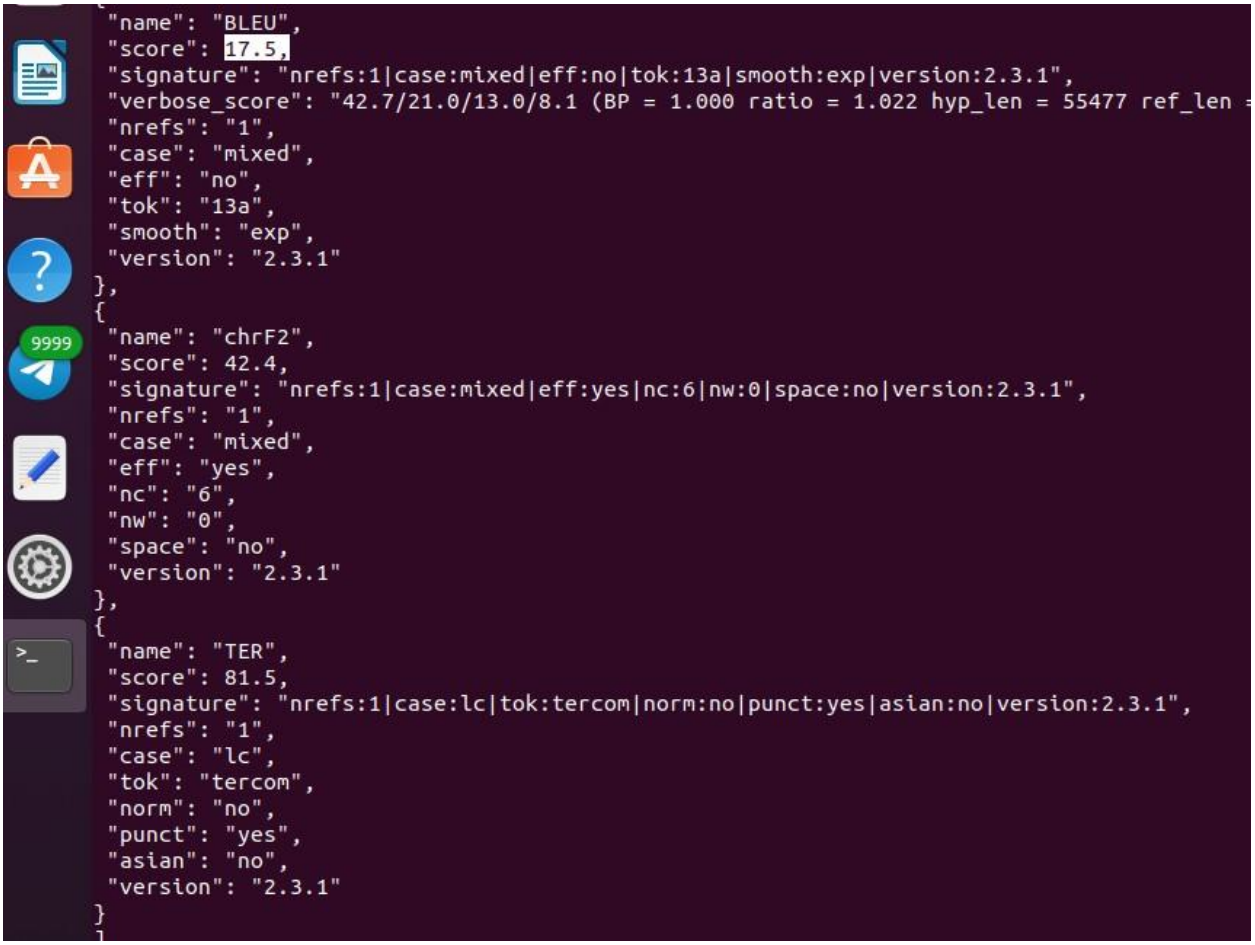
Appendix V

Sample comparison for input Sidaamu Afoo and for output Amharic word level alignment.


```
ubuntu@ubuntu-Lenovo-V330-15IKB:~/SMT$ sacrebleu /home/ubuntu/siam/data/test.si-am.dtok.am -i trans
[
{
 "name": "BLEU",
 "score": 11.3,
 "signature": "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1",
 "verbose_score": "39.4/15.3/7.5/4.0 (BP = 0.981 ratio = 0.981 hyp_len = 57068 ref_len = 58165)",
 "nrefs": "1",
 "case": "mixed",
 "eff": "no",
 "tok": "13a",
 "smooth": "exp",
 "version": "2.3.1"
},
{
 "name": "chrF2",
 "score": 26.4,
 "signature": "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1",
 "nrefs": "1",
 "case": "mixed",
 "eff": "yes",
 "nc": "6",
 "nw": "0",
 "space": "no",
 "version": "2.3.1"
},
{
 "name": "TER",
 "score": 83.4,
 "signature": "nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.3.1",
 "nrefs": "1",
 "case": "lc",
 "tok": "tercom",
 "norm": "no",
 "punct": "yes",
 "asian": "no",
 "version": "2.3.1"
}
]
```

Appendix VI

Sample comparison for input Amharic and for output Sidaamu Afoo word level alignment.



Appendix VII

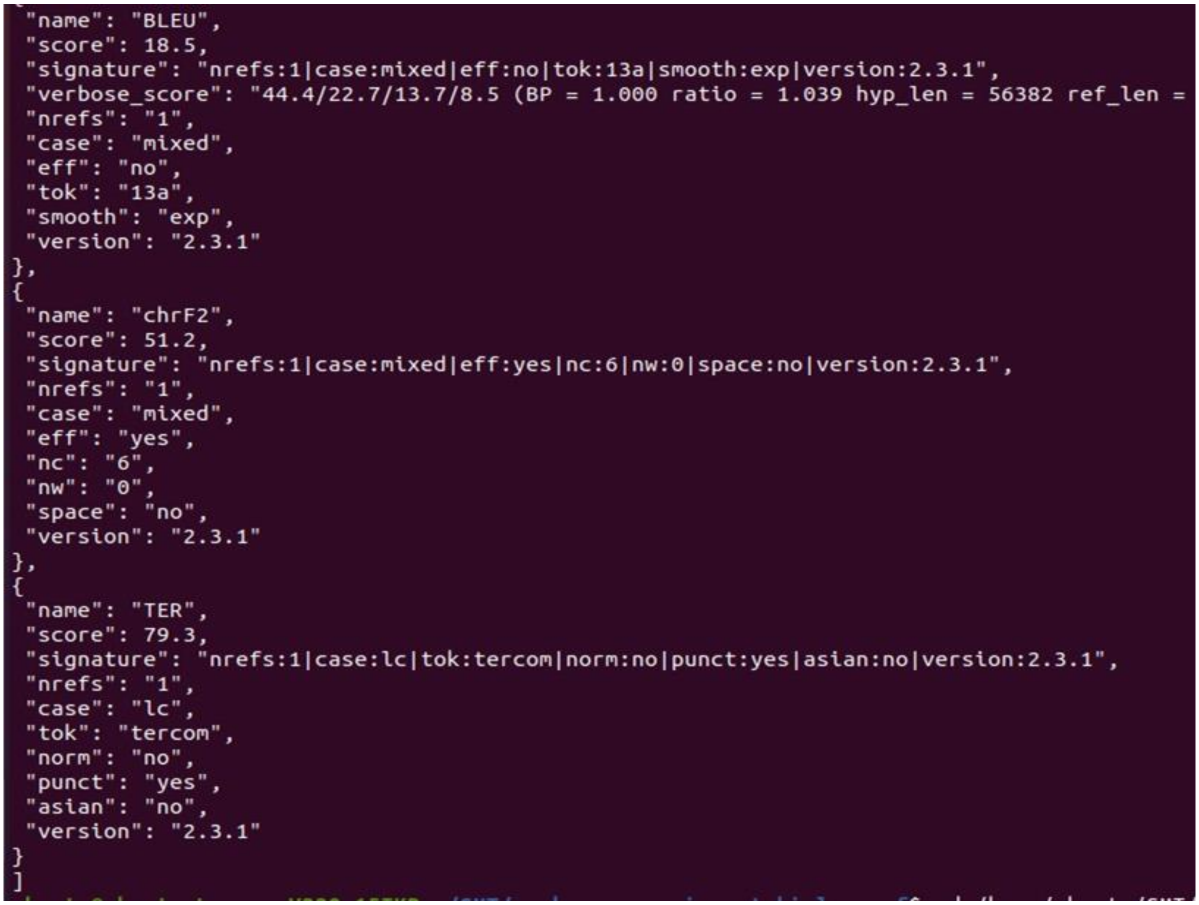
Sample comparison for input Sidaamu Afoo and for output Amharic morpheme level alignment.

```
ubuntu@ubuntu-Lenovo-V330-15IKB:~/SMT$ sacrebleu /home/ubuntu/siam/data/test.si-am.dtok.am -i translated_siam_dmorf.am -m bleu chrf ter
```


```
[
{
 "name": "BLEU",
 "score": 11.0,
 "signature": "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1",
 "verbose_score": "38.7/15.0/7.2/3.7 (BP = 0.988 ratio = 0.988 hyp_len = 57451 ref_len = 58165)",
 "nrefs": "1",
 "case": "mixed",
 "eff": "no",
 "tok": "13a",
 "smooth": "exp",
 "version": "2.3.1"
},
{
 "name": "chrF2",
 "score": 30.8,
 "signature": "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1",
 "nrefs": "1",
 "case": "mixed",
 "eff": "yes",
 "nc": "6",
 "nw": "0",
 "space": "no",
 "version": "2.3.1"
},
{
 "name": "TER",
 "score": 84.4,
 "signature": "nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.3.1",
 "nrefs": "1",
 "case": "lc",
 "tok": "tercom",
 "norm": "no",
 "punct": "yes",
 "asian": "no",
 "version": "2.3.1"
```

Appendix VIII

Sample comparison for input Amharic and for output Sidaamu Afoo morpheme level alignment.



```
 "name": "BLEU",
 "score": 18.5,
 "signature": "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1",
 "verbose_score": "44.4/22.7/13.7/8.5 (BP = 1.000 ratio = 1.039 hyp_len = 56382 ref_len =
 "nrefs": "1",
 "case": "mixed",
 "eff": "no",
 "tok": "13a",
 "smooth": "exp",
 "version": "2.3.1"
},
{
 "name": "chrF2",
 "score": 51.2,
 "signature": "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1",
 "nrefs": "1",
 "case": "mixed",
 "eff": "yes",
 "nc": "6",
 "nw": "0",
 "space": "no",
 "version": "2.3.1"
},
{
 "name": "TER",
 "score": 79.3,
 "signature": "nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.3.1",
 "nrefs": "1",
 "case": "lc",
 "tok": "tercom",
 "norm": "no",
 "punct": "yes",
 "asian": "no",
 "version": "2.3.1"
}
]
```

Appendix IX

Sample of morpheme level aligned corpus segmented using morfessor

Hakku _nni gedensaanni moote Maganu manchi _ra , " A _mo ane ledo mini _ra ha _ʔne _enna mito _re afoo _ho wodha _tto _na ; afa _nsha _no eemmohe _na " yii _si .	ከዚህ _ም በኋላ ንጉሡ ነቢዩ _ን " ና ከ _እኔ ጋር አብረ _ህ ወደ ቤቴ ግባ ; እህል _ም ቅመስ ; ስጦታ እሰጥ _ሃለሁ " አለ _ው .
Qulla _awu Ayyaani widoonni qol _te borojjichi _kki anninke Daawiti _ra , ‹ Daga mayi _ra hanqi _tu ? Go _sa _no mayi _ra beebba _tenni ga _no gantu ?	በ _መንፈስ ቅዱስ አማካይ _ነት በ _አገልጋይህ በ _አባታችን በ _ዳዊት አፍ እንዲ _ህ ብለህ የ _ተናገርከ አንተ ነህ ; " አሕዛብ ስለምን ተ _ቈጡ ? ሕዝቦች _ስ ስለምን በ _ከንቱ ዶለቱ ?
Abirihaami _no Efi _ro _oni yiino _ta macciishshe, mannu afa _nna kulino gari _nni Abirihaami 400 kinchu bir _ra kee _ne uyi _si .	አብርሃም _ም በ _ነገሩ ተስማምቶ ዔ _ፍሮን በ _ሕዝቡ ፊት በ _ተናገረው መሠረት በ _ነጋዴዎች ሚዛን ልክ አራት መቶ ብር መዝኖ ሰጠው .
Alba _sii _nni noo be _le _qi _nni , kaade _nna giira _te buwani fule , duʔm _u gomi _chi giddoo _nni hige sai .	በፊቱ ካለ _ው ነጸብራቅ በረዶ _ና የ _እሳት ብልጭታ ጥቅጥቅ ያለ _ውን ደመና ሰንጥቆ ወጣ .
Laa _ba _no rodoo _si beetti _ha Yaiqoobi duduwo macciishshe haaʔra _sira ra _he mari ; hanqafe sunq _isi _hu gedensaanni , mini _sira haaʔr _e haʔri _si . Yaiqoobi _no baalanka coye Laa _ba _ra hasaawi .	ላባ _ም የ _እኅቱ ልጅ ያዕቆብ መምጣቱ _ን በ _ሰማ ጊዜ ሊ _ቀበለው ፈጥኖ ሄደ ; ዐቅፎ ከ _ሳመው _ም በኋላ ወደ ቤት አመጣ _ው ; ያዕቆብ የሆነ _ውን ሁሉ ለ _ላባ ነገረው ;
Sagale _no ite dhaabb _ati . Shiima barra Demaasiqo _te heedhanno rosaano _wa ta _xxi yii .	ሙብል _ም በላ _ና በረታ ; በ _ደማስቆ ከ _ነበሩት ደቀ መዛሙር _ት ጋር ለ _ጥቂት ቀኖች ቈ _የ .
Yesuusi qole , " Mittu manchi jawa jila jil _e , lowo ma _nna woshshi _no .	ኢየሱስ ግን እንዲ _ህ አለ _ው ; " አንድ ሰው ትልቅ የ _እራት ግብዣ አዘጋጅቶ , ብዙ ሰዎችን ጠራ ;
Qol _e _no , reya _tto _wa _nna fara _di gaare _kki uurri _te gattanno _wa , qunqum _u gede asse halaʔla _du xawi _ra wigidii _de tuganno _he ; ate ikkito _nni kaino _hunni roorrichi _kki saala _nno .	እርሱ _ም ወደምት _ሞት _በት _ና ሠረገሎች _ህም ቆመው ወደሚ _ቀሩበት ቦታ እንደ ኳስ ጠቅልሎ ወደ ሰፊው ሜዳ ይወረውር _ሃል ; ለ _ጌቶች _ህም ኀፍረት ትሆናለህ .
Roduuwa _ʔya ! Kiʔne Yihudu gobba _ra heedhanno _ta Yesuusi Kiristoosi _nni Maganu beetekiristaane ikkitino _ta lawishsha assidhino _onni . Insa _ra Ayihude qarra iilli _shshu _nsa _nte gede , kiʔne _rano gobba _ʔne manni qarra iillishi _no _ʔne .	ወንድሞች ሆይ , እናንተ _ም በ _ይሁዳ ምድር የሚገኙ _ትን በ _ኢየሱስ ክርስቶስ የ _እግዚአብሔር አብያተ ክርስቲያን የ _ሆኑት _ን አርአያ ተከት _ላችኋል ; እነርሱ በ _አይሁድ መከራ እንደ ደረሰባቸው ሁሉ እናንተ _ም ከ _ገዛ ወገኖቻችሁ መከራ ደርሶ _ባችኋል .
Kuri kapha _ano , buuxa dandiitanno _kki coye xo _ntanno ; insa woshshaada _ma _tenna baʔa _te kalaqantino ; ayyaani galte heedhanno _nna heda dandiitanno _kki saada gedee _reeti ; insa _no re _yitanno gari _nni baʔanno .	እነዚህ ሐሰ _ተኞች መምህራን ግን ለ _ማስተዋል ያቃታቸው _ን ነገር ይሳደባሉ ; እነርሱ ለ _መጠመድ _ና ለ _መገደል እንደ ተ _ወለዱ , በ _ተፈጥሮ ስሜት እንደሚ _ኖሩ _ና አእምሮ እንደ _ሌላቸው እንስሶች ናቸው ; እንስሶች እንደሚ _ጠፉ እነርሱ _ም ፈጽመው ይጠፋሉ .
Simiooni Pheexiroosi tenne ikkito la _e Yesuusi albaa _nni gulu _phe , " Mooticha , ani cubba _ataamo _hona ane _wiinni hoʔli " yii _si .	ስምዖን ጴጥሮስ ይህን ባ _የ ጊዜ በ _ኢየሱስ ፊት ተንበርክኮ , " ጌታ ሆይ , እኔ ኃጢአተኛ ስለ ሆንኩ ወደ እኔ አት _ቅረብ ! " አለ _ው .

Appendix X

Sample sentences for Word-based translation from Amharic to Sidaamu Afoo

Hakkunni gedensaanni moote Maganu manchira , " Amo ane ledo minira haʔneenna mitore afooho wodhattona ; afanshano eemmohena " yiisi .	ከዚህም በኋላ ንጉሡ ነቢዩን " ና ከእኔ ጋር አብረህ ወደ ቤቴ ግባ ; እህልም ቅመስ ; ስጦታ እሰጥሃለሁ " አለው .
Qullaawu Ayyaani widoonni qolte borojjichikki anninke Daawitira , ‹ Daga mayira hanqitu ?	በመንፈስ ቅዱስ አማካይነት በአገልጋይህ በአባታችን በዳዊት አፍ እንዲህ ብለህ የተናገርከ አንተ ነህ ; " አሕዛብ ስለምን ተቈጡ ?
Gosano mayira beebbatenni gano gantu ?	ሕዝቦችስ ስለምን በከንቱ ዶለቱ ?
Abirihaamino Efirooni yiinota macciishshe , mannu afanna kulino garinni Abirihaami 400 kinchu birra keene uyisi .	አብርሃምም በነገሩ ተስማምቶ ዔፍሮን በሕዝቡ ፊት በተናገረው መሠረት በነጋዴዎች ሚዛን ልክ አራት መቶ ብር መዝኖ ሰጠው .
Albasiinni noo beleqinni , kaadenna giirate buwani fule , duʔmu gomichi giddoonni hige sai .	በፊቱ ካለው ነጸብራቅ በረዶና የእሳት ብልጭታ ጥቅጥቅ ያለውን ደመና ሰንጥቆ ወጣ .
Laabano rodoosi beettiha Yaiqoobi duduwo macciishshe haaʔrasira rahe mari ; hanqafe sunqisihu gedensaanni , minisira haaʔre haʔrisi . Yaiqoobino baalanka coye Laabara hasaawi .	ላባም የእኅቱ ልጅ ያዕቆብ መምጣቱን በሰማ ጊዜ ሊቀበለው ፈጥኖ ሄደ ; ዐቅፎ ከሳመውም በኋላ ወደ ቤት አመጣው ; ያዕቆብ የሆነውን ሁሉ ለላባ ነገረው ;
Sagaleno ite dhaabbati . Shiima barra Demaasiqote heedhanno rosaanowa taxxi yii .	መብልም በላና በረታ ; በደማስቆ ከነበሩት ደቀ መዛሙርት ጋር ለጥቂት ቀኖች ቈየ .
Yesuusi qole , " Mittu manchi jawa jila jile , lowo manna woshshino .	ኢየሱስ ግን እንዲህ አለው ; " አንድ ሰው ትልቅ የእራት ግብዣ አዘጋጅቶ , ብዙ ሰዎችን ጠራ ;
Qoleno , reyattowanna faradi gaarekki uurrite gattannowa , qunqumu gede asse halaʔladu xawira wigidiide tugannohe ; ate ikkitonni kainohunni roorrichikki saalanno .	እርሱም ወደምትሞትበትና ሠረገሎችህም ቆመው ወደሚቀሩበት ቦታ እንደ ኳስ ጠቅልሎ ወደ ሰፊው ሜዳ ይወረውርሃል ; ለጌቶችህም ኀፍረት ትሆናለህ .
Roduuwaʔya ! Kiʔne Yihudu gobbara heedhannota	ወንድሞች ሆይ , እናንተም በይሁዳ ምድር የሚገኙትን
Yesuusi Kiristoosinni Maganu beetekiristaane ikkitinota lawishsha assidhinoonni . Insara Ayihude qarra iillishshunsante gede , kiʔnerano gobbaʔne manni qarra iillishinoʔne .	በኢየሱስ ክርስቶስ የእግዚአብሔር አብያተ ክርስቲያን የሆኑትን አርአያ ተከትላችኋል ; እነርሱ በአይሁድ መከራ እንደ ደረሰባቸው ሁሉ እናንተም ከገዛ ወገኖቻችሁ መከራ ደርሶባችኋል .
Kuri kaphaano , buuxa dandiitannokki coye xontanno ; insa woshshaadamatenna baʔate kalaqantino ; ayyaani galte heedhannonna heda dandiitannokki saada gedeereeti ; insano reyitanno garinni baʔanno .	እነዚህ ሐሰተኞች መምህራን ግን ለማስተዋል ያቃታቸውን ነገር ይሳደባሉ ; እነርሱ ለመጠመድና ለመገደል እንደ ተወለዱ , በተፈጥሮ ስሜት እንደሚኖሩና አእምሮ እንደሌላቸው እንስሶች ናቸው ; እንስሶች እንደሚጠፉ እነርሱም ፈጽመው ይጠፋሉ .
Simiooni Pheexiroosi tenne ikkito lae Yesuusi albaanni guluphe , " Mooticha , ani cubbaataamohona anewiinni hoʔli " yiisi . Mitu mitu manni cubbi xawoho leellanni noo daafira , albillitte daggasira noo yoo leellishanno . Wolu manni cubbi kayinni keeshshe leellanno .	ስምዖን ጴጥሮስ ይህን ባየ ጊዜ በኢየሱስ ፊት ተንበርክኮ , " ጌታ ሆይ , እኔ ኃጢአተኛ ስለ ሆንኩ ወደ እኔ አትቅረብ ! " አለው . የአንዳንድ ሰዎች ኃጢአት ግልጥ በመሆኑ አስቀድሞ ፍርዱን ያመለክታል ; የሌሎች ሰዎች ኃጢአት ግን የሚገለጠው ዘግየት ብሎ ነው . በዚያም የበኵር ልጅዋን ወለደች ; በመታቀፊያ ጨርቅ ጠቀለለችው ; በእንግዶች ማደሪያ ስፍራ ስላላገኙ በበረት በከብቶች መመገቢያ ግርግም ውስጥ አስተኛችው .

Hakkiicho heedhe bayira beettose iltu ; wosinu galanno dargi nookki daafira , hoccunni xaaxxe hoowete saadate hayisso worranni badiri aana goxissu . Bunshe malannoha , xagaraamoho yinanni . Kiʔnenna maateʔne , Leewaawootunna giddoʔne	
heeʔrannohu manni gobba manni Maganiʔne Kaaliiqi uyinoʔnehu danchu coyi baalunkunni jiʔlitinanni hagiidhe . Iseno , Kaaliiqi Muse hajajino garinni , umoho xaaxiʔnanni koffiyira albaanni bulu xilatinni usurtu . Insa mereero xonamanno gede assitinoonnihu qullaawu suʔmiʔya dagate mereero ayirranno gede asseemmo yannara , ani Kaaliiqa ikkoommota affanno . Ani Mootichu Kaaliiqi konne coye coyiʔroommo . Ani Qullaawa ikkoommota daga kiʔne widoonni higge affanno gede asseemmo . Iyosiyaasi moohihunni tonaa settikki diro , daga magansidhanno magansiʔra hune , gobbanna Qullaawa Mine keereensihu gedensaanni , Qullaawa Mine gatamartanno gede Atsaaliya beetto Saafaani , Yerusaalame gashshe nooha Maiseyanna borreessaancho Iyoahaazi beetto Yoaahi soyi . Yesuusi Galiilaho heeʔri yannara , kuni meenti Yesuusi hoode haʔrino . Qoleno towatinosiho . Ledosi Yerusaalame dayinohu wolu lowo meentino hakko no .	ዘወትር ክፋትን ለማድረግ የሚያሤር ሁሉ " ተንኰለኛ " የሚል የቅጽል ስም ይወጣለታል ; ለአንተና ለቤተሰብህም እግዚአብሔር ባደረገላችሁ መልካም ነገር ሁሉ ደስ ብሎአችሁ አመስግኑ ; ሌዋውያንና ከእናንተ ጋር የሚኖሩ መጻተኞች በዚህ የምስጋና ሥርዓት አፈጻጸም ላይ አብረው ይገኙ . እርሱንም እግዚአብሔር ሙሴን ባዘዘው መሠረት ከመጠምጠሚያው ፊት በሰማያዊ ክር አሰሩት . በሕዝቦች መካከል ተሰድቦ የነበረውን , ማለት እናንተ አሰደባችሁት የነበረውን የታላቁን ስሜን ቅድስና አሳያለሁ ; ከዚያ በኋላ በእናንተ አማካይነት በእነርሱ ፊት ቅድስናዬን በምገልጥበት ጊዜ ሕዝቦች እኔ እግዚአብሔር መሆኔን ያውቃሉ . ኢዮስያስ በነገሠ በዐሥራ ስምንተኛው ዓመት የአሕዛብን አምልኮ በማጥፋት ምድሪቱንና ቤተ መቅደሱን ካነጻ በኋላ , የእግዚአብሔርን ቤተ መቅደስ ያድሱ ዘንድ የአጻልያን ልጅ ሳፋንን , የኢየሩሳሌም አስተዳዳሪ የሆነውን ማዕሤያንና የኢዮአሐዝን ልጅ ጸሐፊውን ዮአሕን ላከ ; እነርሱ ኢየሱስ በገሊላ በነበረበት ጊዜ , ይከተሉትና ያገለግሉት ነበር . እንዲሁም ከኢየሱስ ጋር ወደ ኢየሩሳሌም የመጡ ብዙ ሌሎች ሴቶች ነበሩ . ከላይ የሚመጣው ከሁሉ በላይ ነው ; ከምድር የሆነ ምድራዊ ነው ; የምድርንም ነገር ይናገራል ; ከሰማይ የሚመጣው ከሁሉ በላይ ነው .