ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

ENGLISH -TIGRIGNA FACTORED STATISTICAL MACHINE TRANSLATION

BY

TARIKU TSEGAYE

JUNE 2014

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

ENGLISH -TIGRIGNA FACTORED STATISTICAL MACHINE TRANSLATION

A thesis submitted to the school of graduate studies of Addis Ababa University in partial fulfillment of the requirements of the degree of Master of Science in Information Science

BY
TARIKU TSEGAYE

JUNE 2014

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

ENGLISH -TIGRIGNA FACTORED STATISTICAL MACHINE TRANSLATION

BY

TARIKU TSEGAYE

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
____________________________	Chairperson	____________________________	
____________________________	Advisor,	_____________	____________________
____________________________	Examiner,	________________________________	

Declaration

I declare that the thesis is my original work and has not been presented for a degree in any other university.

____

Date

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

____

Advisor

AKNOWLEDGMENTS

First and foremost, praise be to the Almighty God.

I am greatly indebted to a number of people who assisted me to successfully complete my graduate study and thesis. I would first of all like to express my heartfelt appreciation to my advisor Dr. Martha Yifiru. This study wouldn't have been fruitful without her consistent guidance and follow-up. Thank you for everything.

I would also like to acknowledge the support and encouragement I have received from my dear friends.

I would like to reserve my deepest gratitude to my family members; my uncle Tesfu, my father, my mother and my brothers. Dear Tesfu, Please note that you have always been inspirational in pursuing my graduate studies and advance my career as a whole.

Dear Faseye, Thank you for believing in me. I am so glad to have you by my side!

LIST OF ACRONYMS I

LIST OF TABLES II

LIST OF FIGURES III

LIST OF ALGORITHMS IV

LIST OF APPENDICES V

ABSTRACT VI

1. CHAPTER ONE: INTRODUCTION 1

1.1. Background 1
1.2. Statement Of The Problem 3
1.3. Objective 5
1.4. Methodology 5
1.5. Scope and limitation Of The Study 6
1.6. Organization Of The study 6

2. CHAPTER TWO : LITERATURE SURVEY 8

2.1 Machine Translation Approaches 9

2.1.1 Linguistic or Rule based Approaches 9

2.1.1.1 Direct Approach 9

2.1.1.2 Interlingua Approach 10

2.1.1.3 Transfer Approach 12

2.1.2 Non-Linguistic Approaches 12

2.1.2.1 Dictionary based Approach 12

2.1.2.2 Empirical or Corpus based Approach 13

2.1.2.3 Example based Approach 13

2.1.2.4 Statistical Approach 14

2.1.3 Hybrid Machine Translation System 15

2.3 Components of Statistical Machine Translation 16

2.4 Factored Statistical Machine Translation....................18
2.5 Related Studies....................19
2.5.1 Global Works....................19
2.5.2 Local Works....................20
3. CHAPTER THREE:TIGRIGNA LANGUAGE....................22
3.1. Introduction....................22
3.2. Morphology....................22
3.2.1.Derivation and Inflectional Morphology of Tigrigna....................24
3.2.1.1.Inflectional Morphology....................24
3.2.1.2.Derivational Morphology....................24
4. CHAPTER FOUR: METHODOLOGY....................33
4.1. Data Collection methods....................33
4.2. Preprocessing Techniques and Algorithms....................33
4.2.1.Sentence Level Segmentation....................34
4.2.2.Tokenization....................34
4.3. Morphological Segmentation....................35
4.4. POS Tagging....................36
4.5. Sentence Alignment....................37
4.6. Factored Corpus Preparation....................37
5. CHAPTER FIVE : EXPERIMENTATION AND DISCUSSION....................41
5.1. Experiment setup....................41
5.2. Experiment results....................46
5.3. Discussion....................48
6. CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS....................51
6.1. Conclusion....................51
6.2. Recommendation....................52
REFERENCES....................A
APPENDIX I: Geez Alphabet....................D
APPENDIX II: Baseline English-Tigrigna Parallel corpus excerpt....................E
APPENDIX III: Factored English-Tigrigna Parallel corpus excerpt....................I

LIST OF ACRONYMS

ALPAC – Automatic language processing Advisory Committee

CLIR – Cross Language Information Retrieval

EBMT – Example Based Machine Translation

FSMT- Factored statistical Machine Translation

MT- Machine Translation

NLP- Natural Language Processing

OANC – Open American National Corpus

POS- Part of Speech tagging

SMT - Statistical Machine Translation

LIST OF TABLES

Table 2.1: Example of direct approach 10
Table 3.1: Inflection of Imperfective tense 27
Table 3.2: Inflection in gerundive form of verb 27
Table 3.3: Inflection of imperative form of verb 28
Table 3.4: Inflection of nouns 28
Table 3.5: Inflection of Adjectives 29
Table 3.6: Nouns derived from other nouns 29
Table 5.1: System Environment for baseline and segmented experimentation 41
Table 5.2: System Environment for factored experimentation 42
Table 5.3: Prepared Corpus for segmented experimentation 46
Table 5.4: prepared corpus for segmented experimentation 47
Table 5.5: A BLEU score of the three systems...48
Table 5.6: Output translation of baseline system 49
Table 5.7: Sample Factored translation 50

LIST OF FIGURES

Figure 2.1: The vauquois triangle 11
Figure 2.2: Example of Factored Translation Model..18
Figure 4.1: Tigrigna Factored corpus preparation flow 38
Figure 4.2: English Factored Corpus preparation flow 39
Figure 5.1: Architecture of the proposed system..46
Figure 5.2: English sentence length graph..47
Figure 5.3: Tigrigna sentence length graph 48

LIST OF ALGORITHMS

Algorithm 4.1: Sentence level segmentation algorithm..34

Algorithm 4.2: Tokenization algorithm...35

Algorithm 4.3: Segmentation algorithm...36

Algorithm 4.4: Factored corpus preparation for both languages..40

LIST OF APPENDIX

Appendix I: Geez Alphabets

Appendix II: Baseline English-Tigrigna Parallel corpus excerpt

Appendix I: Factored English-Tigrigna Parallel corpus excerpt

ABSTRACT

In this paper, English to Tigrigna translation was conducted using Statistical machine translation approach. A total of 17,649 sentence pairs were used as a bilingual corpus to develop, train and test the translation system. Experiment was conducted using MOSES employing three types of corpus namely baseline, Segmented and finally factored corpus that integrates linguistic knowledge at word level. Some preliminary preprocessing task were performed namely sentence level segmentation and tokenization. These preprocessing tasks were done using a program codes written with python. In addition to that a lot of manual cleaning tasks were done when the preprocessing task required the researcher's judgment. After preprocessing, morphological segmentation, stemming and POS tagging were performed to prepare the factored corpora. The performance of the system was then tested using the BLEU metric. The result revealed that segmentation has contributed for the overall performance of the segmented system that has shown better performance compared to the baseline phrase-based system. When compared with the same segmented reference, the BLEU score for the segmented system is 22.65% which is a 1.61% increase from the baseline system that has a BLEU score 21.04. The factored corpus has shown a decrease of 6.15% from the segmented and 4.53% from the baseline system. The researcher believes that, the low performance of the factored system is accounted to the POS tags attached since the tagger was trained using a small manually tagged corpus prepared by the researcher.

INTRODUCTION

1.1 BACKGROUND

Natural Language Processing (NLP) is the field of computer science devoted to the development of models and technologies enabling computers to use human languages both as input and output (Jurafsky et al, 2005).

The ultimate goal of NLP is to build computational models that equal human performance in the task of reading, writing, learning, speaking and understanding. Computational models are useful to explore the nature of linguistic communication as well as for enabling effective human-machine interaction. (Jurafsky et al, 2005) describe Natural Language Processing as "computational techniques that process spoken and written human language as language". According to the Microsoft researchers(Martine et al, 2005), the goal of the Natural Language Processing (NLP) is "to design and build software that will analyze, understand and generate languages that humans use naturally, so that eventually one will be able to address their computer like addressing another person".

Machine Translation, one of the basic endpoints of NLP, is an automatic translation of one natural language text to another using computer. Numerous approaches like Rule based and linguistic based systems are used to develop a machine translation system. But currently, statistical methods are taking over the machine translation field. Statistical Machine Translation (SMT) approach draws knowledge from automata theory, artificial intelligence, data structure and statistics. SMT system treats translation as a machine learning problem. This means that a learning algorithm is applied to a large amount of parallel corpora. Parallel corpora are sentences in one language along with its translation. Learning algorithms create a model from parallel sentences and using this model, unseen sentences are translated. If parallel corpora are available for a language pair then it is easy to build a bilingual SMT system. The accuracy of the system is highly dependent on the quality and quantity of the parallel corpus and the domain. These parallel corpora are constantly growing. Parallel corpora are the fundamental resource for SMT system which can be obtained from any domain such as news, parliamentary documents where the corresponding translation is available.

In SMT system, statistical methods are used for mapping of source language phrases into target language phrases. Statistical model parameters are estimated from bi-lingual and mono-lingual corpora. There are two models in the SMT system. These are Translation model and Language model. The translation model takes parallel sentences and finds the translation hypothesis between the phrases. Language model is based on the statistical properties of n-grams. It uses the monolingual corpora. Several translation models are available in SMT system. Some important models are phrase based model, syntax based model and factored model. Phrase Based Statistical Machine Translation (PBSMT) is limited to the mapping of small text chunks. Factored translation model is an extension of phrase based model in that it integrates linguistic information at the word level.

Machine Translation is used for translating texts for assimilation purpose which aids bilingual or cross-lingual communication and also for searching, accessing and understanding foreign language information from databases and web-pages (Hutchins, 2001). MT highly contribute to the field of information retrieval particularly in Cross-Language Information Retrieval (CLIR), i.e. information retrieval systems capable of searching databases in many different languages (Aynalem et al., 2010).

1.2 STATEMENT OF THE PROBLEM

Translation plays a big role in closing the language barrier. Hutchins et al. (1992) point out that, in an increasingly global economy, the relatively limited number of professional translators cannot meet the growing demand for rapid translations.Human translators use a mixture of thought processes, abilities and resources to interpret the meaning of sentence and communicate it in a different language.

Though human translation provides accurate translation it is considered as expensive, time taking and usually unavailable when it is needed for communicating quickly and cheaply with people with whom we do not share a common language (Melby et al, 1995).Our country's short-coming of human translation are not different from the global perspective.

While humans remain the only truly reliable translators, machine translation provides the advantage of immediate turnaournd.Machine Translation is the application of computers to the task of translating texts from one natural language to another. Even though it was envisioned as a computer application in the 1950’s, it is still considered to be an open problem (Hutchins, 2001).Using MT system, the cost is minimized and speed is considerably maximized. Though the growing demand of speed and cost minimized translation is mitigated by MT, the problem still lays in coming up with close-to human translation. Therefore, we can consider human aided MT, in which the human editor/translator often pre-edits the text, or applies the criteria of controlled language, and works with special language domains. (Austermuhl, 2001).This will substantially reduce the limitation of MT.

The problem of Machine Translation on local languages has been explored using two MT approaches over the past 2 years. (Adugna et al, 2010), employed statistical approach to Oromo-English machine translation with promising result. Then Gasser (2012) has used rule based approach on Amharic and found an interesting result even though the work is still in progress. A study by (Mulu et al., 2012), employed Amharic Statistical approaches and found a BLEU result in the range of 35-36%. The study by Mulu has continued by integrating linguistic knowledge at word level.

Because of its large speakers and the fact that it's used as the official working language of the Federal government of Ethiopia, Amharic is the language mostly explored in many Natural Language studies. Tigrigna, the third most spoken language after Amharic and Oromo, is one of the least investigated languages in NLP. Though some researches in the area of stemming and text categorization have been conducted, machine translation on Tigrigna has not been explored before. Unlike other approaches such as rule based, which requires the manual development of linguistic rules, SMT is not tailored for any specific pair of languages. It is an approach to MT that is characterized by the use of machine learning methods. In less than two decades, SMT has dominated the academic MT research, and has gained a share of the commercial MT market as well (Lopez, 2008). This makes it ideal for a less explored language like Tigrigna. Factored translation, which is an extension of phrase based translation, is a statistical machine translation approach that integrates linguistic features at word level. This approach has shown to be effective for highly morphologically inflectional language.

Taking into account the above limitation and opportunities, the present study sought to employ Statistical Machine Translation for Tigrigna. The study will be performed on three types of corpus namely baseline, segmented and finally Factored SMT that integrates linguistic knowledge into the corpus hoping to bring better performance.

This study then, attempts to address the following research questions:

- What will be the performance of SMT approach to English-Tigrigna pair?
- Can segmented and factored corpus improve the performance of English-Tigrigna Baseline phrase based SMT?

1.3 OBJECTIVE

General Objective

The main objective of this study is to develop English to Tigrigna statistical machine translation system by integrating Linguistic features.

Specific objective

Specific objectives include:-

- ✓ Pre-processing both Tigrigna and English corpus; preprocessing includes cleaning and factoring words in to lemmas and other morphological features
- ✓ prepare parallel corpus by aligning Tigrigna and English sentences
- ✓ build and train the system using aligned documents
- ✓ Evaluate the performance of the model

1.4 METHODOLOGY

Literature Survey

Several related articles, books and literatures will be reviewed to achieve the objective of this thesis.

Data Collection

SMT requires the availability of Parallel corpus, which is the direct translation of sentences of both languages. Therefore documents such as Bible and news available both in Tigrigna and English will be collected. Due to the scope of this thesis and taking into consideration the availability of parallel corpus in both languages, an estimate of 15,000 to 20,000 sentences will be used for the experiment after preprocessing.

Data Preprocessing

Prior to experimentation the collected documents will run through a series of preprocessing such as case normalization (for English) and sentence segmentation. Since this research employs integrating linguistic features in to the parallel corpora, the documents will then be prepared with morphological features attached with each words for both languages. The data will also be set in such a way it is suitable for MOSES and other software tools used during the research.

Programming Languages and tools Used

Numerous tools are available for Statistical Machine Translation. HUNALIGN is a sentence level alignment tool which proves to be a robust tool in the area of SMT and will be used for this thesis. MGIZA++ will be used for word alignment whereas SRILM will be employed for building and managing the language model. MOSES, the leading open-source toolkit for statistical machine translation, will be used for automatic translation. Finally BLEU, an evaluation metric, will evaluate the translation accuracy of the system.

Morfessor, will be used to generate different morphological features of English and segment each word to its root word. The stemmer by (Yonas, 2011) will be used to segment Tigrigna words to its stem. PYTHON was selected as a programming language for this research for preprocessing activities. Python is an open source, interpreted, object-oriented, high level programming language. It enables the researcher to implement text preprocessing for the sought system without any hassle and allows writing programs that are clear and readable.

Experimentation and Evaluation

The preprocessed data will first be aligned at sentence level and word level. The aligned documents will then be used to generate and train the translation model. A substantial amount of corpus will also be used to train the language model. The system's performance will then be evaluated

1.5 SCOPE AND LIMITATION OF THE STUDY

The paper will limit its study to English-to-Tigrigna Statistical Machine Translation using baseline, segmented and with factored corpus. The study conducted is mono-directional. i.e., from English to Tigrigna. The study will employ around 17,000 sentences because of the scarcity of multiple domain English-Tigrigna parallel corpora.

1.6 ORGANIZATION OF THE STUDY

The thesis is presented in six chapters. The first chapter presents an explanation about the research background, problem description, objective of the study, methodology as well as scope and limitation of the study.

Chapter two presents the various concepts in Statistical Machine Translation and Factored Statistical machine translation. Chapter three discusses the morphology of Tigrigna in general. It also describes the overview of Tigrigna language, word formation in Tigrigna and derivational and inflectional morphology of verbs, nouns, and adjectives.

Chapter four presents Methodology which includes the Data preparation and implementation of the machine translation system. This chapter discuses about the document collection used in the research, the different preprocessing, segmentation, POS tagging and factored corpus preparation. The finding of the experiment result of the three translation systems are discussed in chapter five.

Finally, Chapter six presents summary, conclusion and recommendation.

CHAPTER TWO

LITERATURE SURVEY

Mechanizing translation processes can be traced back to seventeenth century but has been an interesting area of research since the 1930's (Hutchins et al., 1992)**.** In 1933 Artsrouni designed a storage device on paper tape which could be used to find the equivalent of any word in another language. In 1950, (Bar-Hillel, 1960), noted that fully automatic translation would not be achieved without long-term basic research, and human assistance was essential, either to prepare texts or to revise the output.

In the 1950s and 1960s research tended to polarize between empirical trial-and-error approaches, often adopting statistical methods with immediate working systems as the goal, and theoretical approaches involving fundamental linguistic research and aiming for long-term solutions (Locke et al, 1955).There were many predictions of imminent breakthroughs and of fully automatic systems operating within a few years. However, disillusion grew as the intricacy of the linguistic problems became more and more obvious. In a review of Machine Translation progress, Bar-Hillel (1960) criticized the prevailing assumption that the goal of MT research should be the creation of fully automatic high quality translation systems producing results identical to those of human translators (Booth,1967). In 1964, Automatic Language Processing Advisory Committee (ALPAC), which was formed to examine the prospects of Machine Translation, reported that MT was slower, less accurate and twice expensive as human translation.

The main framework of MT research until the end of the 1980s was based on essentially linguistic rules of various kinds: rules for syntactic analysis, lexical rules, and rules for lexical transfer, rules for syntactic generation and the like. Since 1989, however, the dominance of the rule-based approach has been broken by the emergence of new methods and strategies which are now called corpus-based methods. At the very same time certain Japanese groups began to publish preliminary results using methods based on corpora of translation examples, i.e. using the approach now generally called `example-based' translation.

The most dramatic development has been the revival of the statistics-based approach to MT in the candied project at IBM with acceptable results either matching exactly the translations in the corpus, or expressed the same sense in slightly different words. The Statistical approaches existed in the early periods of Machine translation but the result was disappointing (Arnold et al, 1994).

Since MT requires large corpus for training and translation, the wide availability of high end computers with great processing speed and memory capacity has made an immense contribution to the increase in performance of MT systems.

2.1 Machine Translation Approaches

Since the emergence of ideas of using machines for language translation processes, different approaches have been proposed and put into practice (Hutchins, 1986). There are two major divisions of machine translation. These are linguistic/rule based and non-linguistic approaches. The combination of the two approaches is called hybrid approach. The linguistic approach comprises direct, Interlingua and transfer approaches which require some sort of linguistic knowledge to perform translations whereas Non-linguistic approaches consists of Dictionary, corpus based, example based and statistical approaches don't require any linguistic knowledge to translate the sentences (Arnold et al,1994).

2.1.1 Linguistic or Rule based Approaches

Rule based approaches need a lot of linguistic knowledge during the translation and so it utilizes grammar rules and computer programs which will be helpful in analyzing the text for determining grammatical information and features for each and every word in the source language, translating it by replacing each word by lexicon or word that have the same context in the target language. Rule based approach is the main methodology that was developed in machine translation. Linguistic knowledge will be required in order to write the rules for this type of approaches. These rules will play a fundamental role during the different levels of translation (Arnold et al, 1994).

Rule based approach are further divided into direct approach, Interlingua approach and transfer approach.

2.1.1.1 Direct Approaches

Direct translation approach can be taken as the first approach to machine translation. In this type of approach, the machine translation system is designed more specifically for one particular pair of language. There is no need of identifying the schematic roles and universal concepts in this approach. It involves the process of analyzing morphological information, identify the elements and reorder the words in the source language according to the word order pattern of the target language and then replace the words in the source language by the target language words using a lexical dictionary of that particular language pair and as a last step, inflect the words appropriately to produce translations. This approach as it is seen, looks like a lot of work has to be done in order to produce translations, but all those work which has to be employed will be simple and can be accomplished very easily, in a short span of time.

Table 2.1 describes the example, how the sentence “he went to school” will be translated from English to Tigrigna using the direct approach.

Input sentence in English		He went to school.
After	**Morphological Analysis**	He go PAST to school.
	Constituent Identification	<he><go><PAST><to><school>.
	Word Reordering	<he><to><school><go><PAST>
	Dictionary Lookup	ንሱ ናብ ቤት ትምህርቲ ከይዱ< PAST >
	Inflection (the final translated language)	ናብ ቤት ትምህርቲ ከይዱ

Table 2.1 example of direct approach

2.1.1.2 Interlingua Approach

Interlingua approach to machine translation mainly aims at transforming the texts in the source language to a common representation which is applicable to many languages. Using this representation, the translation of text to the target language is performed and it should be possible to translate to every language from the same Interlingua representation with the right rules.

Interlingua approach sees machine translation as a two stage process:

1. Analyzing and transforming the source language texts into a common language independent representation.

2. From the common language independent form generate the text in the target language.

The first stage is particular to source language and doesn't require any knowledge about the target language whereas the second stage is particular to the target language and doesn't require any knowledge from the source language. The main advantage of interlingua approach is that it creates an economical multilingual environment that requires 2n translation systems to translate among n languages where in the other case, the direct approach requires n(n-1) translation systems.

The Vauquois triangle was used in the linguistic rule-based era of machine translation to describe the complexity or sophistication of approaches to machine translation, and also the evolution of those approaches. The first approach used was a direct lexical conversion between languages. Later efforts moved up the pyramid and introduced more complex processing, and also a modularization of the process into steps, beginning with analysis of the source language, transfer of information between the languages, and then generation of target language output. According to the model, each step up the triangle required greater effort in source language analysis and target language generation, but reduced the effort involved in conversion between languages. The pinnacle and ideal of the was a complete analysis of each sentence into an "interlingua" - a schema capable of representing all meaning expressible in any language in language-independent form. Fig. 2.1 shows the Vauquois triangle

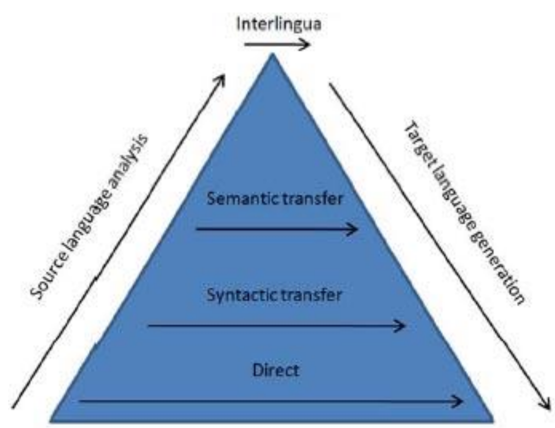

Fig. 2.1 The Vauquois triangle

2.1.1.3 Transfer Approach

Both transfer-based and interlingua-based machine translation have the same idea: to make a translation it is necessary to have an intermediate representation that captures the "meaning" of the original sentence in order to generate the correct translation. In interlingua-based MT this intermediate representation must be independent of the languages in question, whereas in transfer-based MT, it has some dependence on the language pair involved.

The way in which transfer-based machine translation systems work varies substantially, but in general they follow the same pattern: they apply sets of linguistic rules which are defined as correspondences between the structure of the source language and that of the target language. The first stage involves analyzing the input text for morphology and syntax (and sometimes semantics) to create an internal representation. The translation is generated from this representation using both bilingual dictionaries and grammatical rules.

2.1.2 Non-Linguistic Approaches

The non-linguistic approaches are those which don't require any linguistic knowledge to translate texts in the source language to target language. The only resource required by this type of approaches is data either the dictionaries for the dictionary based approach or bilingual and monolingual corpus for the empirical or corpus based approaches.

2.1.2.1 Dictionary based Approach

The dictionary based approach to machine translation uses dictionary for the language pair to translate the texts in word level from the source language to target language. In this approach words will be translated as a dictionary does, word by word, usually without much correlation of meaning between them. Dictionary lookups may be done with or without morphological analysis or lemmatisation.This approach, can either be preceded by some pre-processing stages to analyze the morphological information and lemmatize the word to be retrieved from the dictionary. This kind of approach can be used to translate the phrases in a sentence and is not efficient in translating a full sentence. This approach will be very useful in speeding up the human translation, by offering meaningful word translations and limiting the work of humans to correcting the syntax and grammar of the sentence.

2.1.2.2 Empirical or Corpus based Approach

The corpus based approaches don't require any explicit linguistic knowledge to translate the sentence. But a bilingual corpus of the language pair and the monolingual corpus of the target language are required to train the system to translate a sentence. This approach has driven lots of interest in world-wide.

2.1.2.3 Example based Approach

This approach to machine translation is a technique that is mainly based on how human beings interpret and solve the problems. That is, normally the humans split the problem into sub problems, solve each of the sub problems with the idea of they have solved this type of similar problems in the past and integrate them to solve the problem in a whole. This approach needs a huge bilingual corpus of the language pair among which translation has to be performed.

The Example Based Machine Translation system functions like a translation memory. A translation memory is a computer aided translation tool that is able to reuse previous translations. If the sentence or a similar sentence has been translated previously, the previous translation is returned. In contrast, the EBMT system can translate novel sentences and not just reproduce previous sentence translations. EBMT translates in three steps; matching, alignment and recombination. In matching, the system looks in its database of previous examples and find the pieces of text that together give the best coverage of the input sentence. This matching is done using various heuristics from exact character match to matches using higher linguistic knowledge to calculate the similarity of words or identify generalized templates. The alignment step is then used to identify which target words these matching strings corresponds to. This identification can be done by using existing bilingual dictionaries or automatically deduced from the parallel data. Finally these correspondences are recombined and the rejoined sentences are judged using either heuristic or statistical information.

2.1.2.4 Statistical Approach

Statistical MT models take the view that every sentence in the target language is a translation of the source language sentence with some probability (Brown et al., 1990)**.**Statistical machine translation is typically formulated under the framework of the noisy channel model. If we want to translate a sentence e in the source language E to a sentence a in the target language A, the noisy channel model describes the situation in the following way.

We assume that the sentence a to be translated was at first conceived in language E as some sentence e. During communication e was corrupted by the channel to a. Now, let us suppose that each sentence in E is a translation of a with some probability, and the sentence that we choose as the translation ($\dot{e}$) is the one that has the highest probability (Brown et al., 1990). In mathematical terms:

$$\dot{e} = argmax\ P\ (e|a)$$

Intuitively, $P\ (e|a)$ depend on two factors:

1. The kind of sentences that are likely in the language E. This is known as the language model - P (*e*).

2. The way sentences in E get converted to sentences in A. This is called the translation model -P (*a* |*e*).

The advantages of statistical approach over other machine translation approaches are as follows:

- The improved usage of resources available for machine translation such as manually translated parallel and aligned texts of a language pair, books available in both languages and so on. That is large amount of machine readable natural language texts are accessible with which this approach can be applied.
- In general, statistical machine translation systems are language independent i.e., it is not designed specifically for a pair of language.

- Rule based machine translation systems are usually expensive as they utilize manual creation of linguistic rules and also these systems cannot be generalized for other languages, whereas statistical systems can be generalized for any pair of languages, if bilingual corpora for that particular language pair is available.
- Translations produced by statistical systems are more natural compared to that of other systems, as it is trained from the real time texts available from bilingual corpora and also the fluency of the sentence will be guided by a monolingual corpus of the target language.

Statistical parameters are analyzed and determined from Bi-lingual and Monolingual corpora. Using these statistical parameters translation and language models are generated. Designing a statistical system for a particular language pair is a rapid process because the work lies on creating bilingual corpora for that particular language pair. In order to obtain better translations from this approach, the system needs at least more than two million words for a particular domain. Moreover, Statistical Machine Translation requires an extensive hardware configuration to create translation models in order to reach average performance levels (Brown et al., 1990) the key problems in statistical MT are: estimating the probability of a translation, and efficiently finding the sentence with the highest probability.

2.1.3 Hybrid Machine Translation System

Hybrid machine translation approach makes use of the advantages of both statistical and rule-based translation methodologies. Machine translation systems such as Systran (Senellart et al., 2007) were implemented using this approach. Some of the approaches used to develop a hybrid machine translation system are multi-engine, statistical rule generation and multi-pass. Multi-engine approach to hybrid machine translation involves running multiple machine translation systems in parallel. The final output is generated by combining the output of all the sub-systems. Most commonly, these systems use statistical and rule-based translation subsystems (Hutchins, 2007). The statistical rule generation approach involves using statistical data to generate lexical and syntactic rules. The input is then processed with these rules as if it were a rule-based translator (Hutchins, 2007). This approach attempts to avoid the difficult and time-consuming task of creating a set of comprehensive, fine-grained linguistic rules by extracting those rules from the training corpus. This approach still suffers from many problems of normal statistical machine translation, namely that the accuracy of the translation will depend heavily on the similarity of the input text to the text of the training corpus. As a result, this technique has had the most success in domain-specific applications, and has the same difficulties with domain adaptation as many statistical machine translation systems. (Chang et.al., 1997) the multi pass approach involves serially processing the input multiple times. The most common technique used in multi-pass machine translation systems is to pre-process the input with a rule-based machine translation system. The output of the rule-based pre-processor is passed to a statistical machine translation system, which produces the final output. This technique is used to limit the amount of information a statistical system need consider, significantly reducing the processing power required. It also removes the need for the rule-based system to be a complete translation system for the language, significantly reducing the amount of human effort and labor necessary to build the system (Hovy, 1996).

2.2 COMPONENTS OF STATISTICAL MACHINE TRANSLATION

Statistical machine translation requires the following basic components. These are: Language Model, Translation Model and Search algorithm.

Language Modeling

Language modeling is the process of assigning probability to a unit of text, where in the case of SMT, a unit of text represents a sentence. Given a sentence a in a language A, the task is to calculate $P(a)$ (Jurafsky and Martin, 2005**).**

For a sentence containing the word sequence $w_1w_2 \ldots w_n$, we can write,

$$P(a) = P(w_1w_2 \ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2) \ldots P(w_n|w_1w_2 \ldots w_{n-1})$$

One of the problems that arise in language modeling is data sparsity. For instance, how do we calculate probabilities such as P (wn|w1w2 . . . wn−1)? In no corpus will we find instances of all possible sequences of n words; actually we will find only a small fraction of such sequences. This problem can be addressed by using N-gram modeling.

In an N-gram model, (Jurafsky and Martin, 2005), the probability of a word given all the previous words is approximated by the probability of the word given the previous N-1 words. When N=2 we call it Bi-gram model whereas when N=3 we have a trigram model.

N -gram probabilities can be computed in a straightforward manner from a corpus. For example, Bi-gram probabilities can be calculated as

$$P\ (w_n | w_{n-1}) = \frac{Count(w_{n-1} w_n)}{\sum_w Count(w_{n-1} w\ )}$$

When calculating N-gram, Even though a sentence in a given corpus is a proper sentence, but if the probability of a word in that sentence is 0, the calculation will give 0 for the whole sentence. So there should be a way of distributing some of the probability to unseen or zero-probability sequences. This is called smoothing (Jurafsky and Martin, 2005).

Translation Model

A translation model tries to remember as much as possible how likely a source sentence is translated into a target sentence in training data. Most of state-of-the-art translation models used for regular text translation can be grouped into three categories: word-based models, phrase-based models, and syntax-based models. Word-based models use words as translation units. IBM Model 1 is one of the simplest and most widely used word-based models. This model is also called a lexical translation model, where the order of the words in the source and target sentence is ignored.

The steps in the IBM Model 1 are

1. We first choose the length for the target sentence I, according to the distribution.
2. Then, for each position in the target sentence, we choose a position j in the source sentence from which to generate the I^{th} target word according to the distribution, and generate the target word by translating according to the distribution. We include in position zero of the source sentence an artificial "null word", denoted by <null>. The purpose of the null word is to insert additional target words.

The phrase-based models form the basis for most of state-of-the-art SMT systems. Like the word-based models, the phrase-based models are generative models that translate an input sentence in a source language E into a sentence in a target language A. Unlike the word-based models that translate single words in isolation, the phrase-based models translate sequences of words (i.e., phrases) in *E* into sequences of words in *A*. The use of phrases as translation units is motivated by the observation that sometimes one word in a source language translates into multiple words in a target language, or vice versa(Chiang, D., 2005).

Syntax-based models rely on parsing the sentence in the source or the target language, or in some cases in both languages. Exploring syntax information for SMT is a long-standing research topic(Brown et al.,1993).

2.3 FACTORED STATISTICAL MACHINE TRANSLATION

Factored translation model, first experimented by (Koehn et al. 2007), is an extension of phrase based model, which map small text chunks by adding linguistic information into the training data. Integrating linguistic information helps to come up with a better result by avoiding data sparseness problem caused by limited training data. Many attempts has been made to add a richer information to statistical machine translation. Most trials concentrate on either preprocessing of the input corpus or post processing the output.

One of the inadequacies of traditional surface word approach is the poor handling of morphology. In this method each word is treated as token in itself. This means that the translation model treats ገዛ and ገዛታት as two completely different words. If ገዛ is known by the translation model but ገዛታት is not, then it will not be translated... The figure below shows an example of factored translation model:

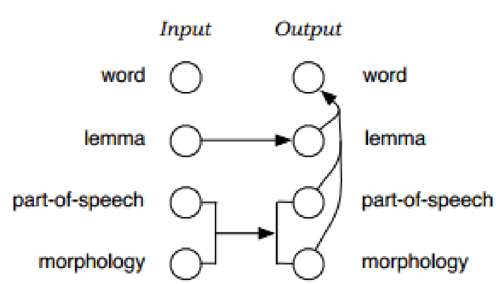

Figure 2.2: example of Factored Translation Model

2.5 RELATED STUDIES

There are considerable amount of studies, both published and unpublished, done to solve the problem of Machine translation. This research were made in different languages employing a variety of approaches and tools. This paper will review key works done globally on Factored Statistical Machine Translation and one work that make use of Rule-Based approach for Amharic-English MT. This section will also briefly examine available local works conducted on a local language particularly on machine translation.

2.4.1 Global Works

Factored SMT has become more and more popular these days as they often overcome the poor performance of phrase based SMT models employed on languages with reach morphological inflections. There are several researches made on Factored SMT using different factors (word features) into their corpus. e.g. (Bojar, 2007).In this study three word features were used and experimented on a series of steps. The experiment was first conducted with input (English) lowercase word forms directly translated to target (Czech) lowercase forms without attaching any word feature. This baseline scenario thus corresponds to a plain phrase-based SMT system. The experiment went on by sequentially adding more and more features such as lemma and POS to the training corpus. A BLEU score of the models show a significant increase in performance when additional factors were added to the words.

Another study by (Kumar, 2003) from English to Tamil used the same approach as (Bojar, 2007) but added automatic and rule based sentence reordering coupled with morphological feature at different stages of the experiment. For the baseline system a standard phrase based system is built using the surface form of the word without adding any additional linguistic knowledge with 4-gram language model. In the consecutive stages of the experimentation, root, part-of-speech and morphological information are included into the word in the form of *Surface|Root|Wordclass|Morphology* as an additional factors. The BLEU score shows a significant increase in performance when factored corpus is used. The average score for the baseline models with automatic and rule based approach is 0.2924. The use of factored corpus has made an increase of 0.39. The highest score is recorded when factored corpus is used with Rule Based Reordering. This system was also compared with different

MT models. The BLEU score shows 0.6 for this system and 0.3 for Google Translate demonstrating the impact of using linguistic features.

An Unfinished yet interesting research by (Gasser, 2012), employed Rule-based approach and has found a corner stone for a full Amharic-English translation system. He introduced L3model which is an evolving framework for Rule Based Machine Translation. This framework has features such as Bi-directional translation, capacity to handle structural divergence between typologically diverse language such as Amharic and English. This research is still on progress but the researcher is confident that he is on track on developing a complete Amharic to English machine translation.

2.4.2 Local Works

There are few research works conducted employing Statistical machine translation on Ethiopian languages.(Adugna et al, 2010), employed statistical approach to Oromo-English machine translation. They used a total of 62,300 sentences (1,024,156 words) of monolingual corpus train the language model system and bilingual corpus of 21,085 English sentences and 20,848 sentences of Oromo to build translation model of the system. From the total of the bilingual document 90 percent was used for training whereas the rest 10 percent was used for testing the system. The researcher experimented using up to 9-gram model and the N-gram score evaluated using BLEU sharply drops when the number of n increases. The researcher argues that the result drops because of the availability of only one reference translation and the diversity of the domain of the test data.

Another Statistical machine translation experiment, by (Mulu et al., 2012), was conducted on Amharic to English. Unlike (Adugna, 2010) who used multiple domain, the researcher used 19,115 Amharic and 25,730 English sentences obtained only from parliamentary documents. After a series of preprocessing, such as conversion of the Amharic corpus to a Unicode format, tokenization, sentence splitting and aligning, the system was trained and tested. The researcher developed two types of SMT called segmented and un-segmented. The un-segmented is the baseline system normally trained, developed and evaluated using the corpus. The segmented system has been developed by segmenting all the target texts. The BLEU score for the segmented system is 36.58%, which is a 0.92% increase from the baseline system that has a BLEU score of 35.66% .

Some of the challenges faced by (Mulu etal.,2012) were the incompatibility of the geez fonts with the system since they are developed with the intension to use them for the major European languages that have more or less similar writing system and punctuations, and readily unavailable Amharic corpus. Both researchers, though conducted on different languages, have similar recommendations. They mainly emphasized on efficient preprocessing of corpus and the integration of morphological and syntactic linguistic knowledge in to the corpus.

CHAPTER THREE

TIGRIGNA LANGUAGE

3.1 INTRODUCTION

Tigrigna is an Afro-Asiatic language, belonging to the family's Semitic branch. It is spoken by ethnic Tigray or Tigrigna in the Horn of Africa (Bauer, 2007). Tigrigna speakers primarily inhabit the Tigray Region in northern Ethiopia (65%), where its speakers are called Tigray, as well as the contiguous borders of southern Eritrea (35%), where speakers are known as the Tigrigna. Tigrigna is also spoken by groups of emigrants from these regions. In Tigrigna each symbol represents a consonant and vowel combination and the symbols are organized in groups of similar symbols on the basis of both the consonant and the vowel. For each consonant in each symbol, there is an unmarked symbol representing that consonant followed by a canonical or inherent vowel (Daniel, 2008).

3.2 MORPHOLOGY

Morphology is the branch of linguistics that deals with the internal structure of words and word formation, including affixation behavior, roots, and pattern properties (Spencer, 1991). Morphology is the main source of variation in natural language text, with suffixing and prefixing being the most common ways of creating a word variant.

Morphology can be classified as either inflectional or derivational. Inflection is variation or change of form that words undergo to mark distinctions of case, gender, number, tense, person, mood, voice, comparison. Inflectional morphology is applied to a given stem with predictable formation. It does not affect the word's grammatical category, such as noun, verb, etc. Case, gender, number, tense, person, mood, and voice are some examples of characteristics that might be affected by inflection. Derivational morphology, on the other hand, concatenates to a given word a set of morphemes that may affect the grammatical and syntactic category of the word.

A word can have several word forms, e.g., the word “write” can take the forms “writes”, “wrote” and “written”, usually called inflected forms. The root is the original form of the word before any transformation process, and it plays an important role in language studies. The root is the form of a word from which the other forms can be derived using the morphological rules of a language. A morpheme is the smallest unit of a language that has a meaning and cannot be broken down further into meaningful or recognizable parts and should impart a function or a meaning to the word which they are part of. An affix is a morpheme that can be added before (prefix) or after (suffix), or inserted inside (infix) a root or a stem to form new words or meanings (Gregory, 2001). Morphological information of a language is useful for several natural language applications such as stemming, morphological analysis, text generation, machine translation, document retrieval, etc.

85% of the words in Tigrigna are created from a root of three radicals (trilateral words) and to a lesser extent there are also quadliteral, pen-literal, or hexa-literal words (Kasa G., 2004). Each word group generates an increased verb forms and noun forms by the addition of derivational and inflectional affixes. Words in Tigrigna are built from the roots by means of a variety of morphological operations such as compounding, affixation, and reduplication (Amanuel, 1998).

An affix in Tigrigna is a morpheme that can be added before or after, or inserted inside, a root or a stem as a prefix, suffix or infix, respectively, to form new words or meanings. Tigrigna affixes have the feature of concatenating with each other in predefined linguistic rules. This feature increases the overall number of affixes (Kasa G., 2004). There are also some prefixes and suffixes which determine whether a word is a subject marker, pronoun, preposition, or a definite article. Tigrigna is highly productive, both derivationally and inflectionally. Definite articles, conjunctions, particles and other prefixes can attach to the beginning of a word, and large numbers of suffixes can attach to the end. A given headword can be found in huge number of different forms.

Tigrigna concatenative morphology regulates how a stem and affixes glue together, while non-concatenative one combines morphemes in more complex ways. Affixes in Tigrigna can be classified as four categories (Kasa G., 2004). Prefixes precede the base form, such as `እንተይ-, ኣይ-, ከምዘ-, ከተተ-, ስለ-, ዝተ- . Suffixes follow the base form, i.e. -ኩም, -ታት, -ታታት, -ነት, -ኣዊ and Infixes are inside the base form. Circumfixes are affixes attached before and after the base form at the same time. While circumfixes formally are combination of allowed prefixes and suffixes, they have to be treated as discontinuous units for semantic and grammatical reasons. Tigrigna non-concatenative morphology refers to reduplicated morpheme forms.

Reduplicated words based on morpheme regularity are grouped into full reduplication (e.g., the word ስትይስትይ is derived from the stem ስትይ) and partial reduplication of different kinds. The latter includes reduplicated stems with affixes (e.g. word ሰባበረ is derived from stem ሰበረ sebere, ተረጋገሙ 'teregagemu' is derived from stem ረገመ 'regeme', the word ገልጠምጠም 'gelTemTem' is derived from the stem ገልጠም 'gelTem') and there also various irregular reduplications.

3.2.1 DERIVATIONAL AND INFLECTIONAL MORPHOLOGY

There are five parts of speech in Tigrigna: adjectives, nouns, verbs, adverbs, and prepositions (Daniel, 2008). Prepositions and conjunctions are totally unproductive. Adverbs are few in number and are less productive. Therefore, the discussion of derivational and inflectional morphology concentrates on the remaining three parts of speech, namely verbs, nouns, and adjectives.

3.2.1.1 INFLECTIONAL MORPHOLOGY OF TIGRIGNA

As Tigrigna is a highly inflectional language definite articles, conjunctions, particles and other prefixes can attach to the beginning of a word, and large numbers of suffixes can attach to the end. A given root of word can be found in huge number of different forms.

3.2.1.1.1 INFLECTION OF VERBS

This section presents the inflection of verbs. It is compiled for the purpose of the study from Tigrigna grammar books by Daniel Teklu (2008) and Kasa G. (2004). A significantly large part of the vocabulary consists of verbs, which exhibit different morph syntactic properties based on the arrangement of the consonant-vowel patterns. For example, the root ሰበር, meaning 'to break' can have the perfect form ሰበረ with the pattern CVCVCV, imperfect form ትሰብር with the pattern CCVCC, gerund form ሰቢርካ with the pattern CVCVCCV, imperative form ሰበር with the pattern CCVC, causative form ኣስበረ with the pattern as-CVCV, passive form ተሰበረ with the pattern CVCVCV,etc. Subject, gender, number, etc. are also indicated as bound morphemes on the verb, as well as objects and possession markers, mood and tense, transitive, dative, negative, etc., producing complex verb morphology.

The simplest form of the verb is the third person masculine singular of the perfect tense. In most Tigrigna dictionaries, all the words derived from a triliteral root are entered under the third person masculine singular form of the verb. Each three-consonant (or "triliteral") root belongs to one of three conjugation classes, conventionally known as A, B, and C. This division is a basic feature of Ethiopian Semitic languages.

Most three-consonant roots are in the A class. In the citation form (perfect), these have no germination but the vowel ‘e’ appears between both pairs of consonants. Examples are: ደረፈ derefe “he sung”, ደየበ deyebe “he climbed”, ሰተየ seteye “he drank”. The B class is distinguished by the gemination of the second consonant in all forms. Some Examples are: ደቀሰ deqqese ‘sleep’ ወሰኸ wesseKe ‘add’. The relatively few members of the C class take the vowel a between the first and second consonants. Examples are ባረኸ bareKe 'bless' and ናፈቐ nafeqe 'long for, miss'.

Tigrigna also has a significant number of four-consonant (or "quadriliteral") roots. These fall into a single conjugation class. Examples are መስከረ meskere ‘testify’ and ቀልጠፈ qelTefe 'hurry’. The language also has five-consonant (or "quinquiliteral") roots. Most, if not all, of these are "defective" in the sense that their simplest form takes the te- prefix.

Examples are ተንቀጥቀጠ te-nqetkeTe 'tremble' and ተምበርከኸ te-mberkeKe 'kneel'. Tigrigna verbs have two tenses: perfect and imperfect. Perfect tense denotes actions completed, while imperfect denotes uncompleted actions. The imperfect tense has four moods: indicative, subjective, jussive, and imperative. Tigrigna verbs in perfect tense consist of a stem and a subject marker. The subject marker indicates the person, gender, and number of the subject. The form of a verb in perfect tense can have subject marker and pronoun suffix. The form of a subject-marker is determined together by the person, gender, and number of the subject. Other elements like negative markers also inflect verbs in Tigrigna.

As in other Semitic languages, Tigrigna verbs are very complex consisting of a stem and up to four prefixes and four suffixes. The stem in turn is composed of a root, representing the purely lexical component of the verb, and a template, consisting of slots for the root segments and for the vowels (and sometimes consonants) that are inserted around and between these segments. The template represents tense, aspect, and mood.

Each lexeme can appear in four different tense-aspect-mood (TAM) categories, conventionally referred to as perfective, imperfective, jussive/imperative, and gerund. For example, the verb ኣይፍተውን ayftewn ‘he is not liked’ has the lemma ተፈተወ tefetewe ’he was liked’, which is derived from the verb root ፈተወ ftw. Every Tigrigna verb must agree with its subject. As in other Semitic languages, subject agreement is expressed by suffixes alone in some TAM categories (perfective and gerundive) and by a combination of prefixes and suffixes.

Tigrigna verbs may also have a suffix representing the person, number, and gender of a direct object or an indirect object that is definite. Tigrigna verbs are inflected for person, gender, number, and time with basic verb form being the third person masculine singular. Tigrigna verbs have two tenses: perfect and imperfect. Perfect tense denotes actions completed, while imperfect denotes uncompleted actions. The imperfect tense has four moods: indicative, subjective, jussive, and imperative. Tigrigna verbs are conjugated in perfective, imperfective, indicative, subjective, jussive, and imperative. In conjugating the verbs affixes are attached to the verbs.

INFLECTION OF PERFECTIVE TENSE

The perfect tense which is the basic form normally expresses the past tense and consist of a stem and a subject marker. The form of a verb in perfect tense can have subject marker and pronoun suffix. The subject marker indicates the person, gender, and number of the subject. The form of a subject-marker is determined together by the person, gender, and number of the subject (Daniel T., 2008).

INFLECTION OF IMPERFECTIVE TENSE

The imperfect tense has four moods: indicative, subjective, jussive, and imperative and is inflected by prefixing and suffixing gender, person, and number morphemes to the imperfective verb stem. The table below shows how suffixes and prefixes are added for the root verb “ገበረ”.

Person Singular Plural	Person Singular Plural	Person Singular Plural
1st person	እገብር(`I-gebr)	ንገብር(n-gebr)
2nd person-masculine	ትገብር(t-gebr)	ትገብሩ(t-gebr-u)
2nd person- feminine	ትገብሪ(t-gebr-i)	ትገብራ(t-gebr-a)
3rd person-masculine	ይገብር(y-gebr)	ይገብሩ(y-gebr-u)
3rd person-feminine	ትገብር(t-gebr)	ይገብራ(y-gebr-a)

Table 3.1: Inflection of Imperfective tense

In imperative tense prefixes እ(`I-), ት(t), ይ(y),ን(n) and the suffixes attached are ኡ(u), ኢ(i), ኣ(a). To indicate negative verbs the morphemes ኣይ/ay/, ኣይት/ayt/, ኣይን/ayn/ are added as prefixes and ን/n/, ኣን/an/, ኡን/un/ are added as suffixes.

INFLECTION OF GERUNDIVE FORM OF VERB

Person Singular Plural	Person Singular Plural	Person Singular Plural
1st person	ነቢረ(nebir-e)	ነበርና(neber-na)
2nd person-masculine	ነበርካ(neber-ka)	ነበርኩም(neber-kum)
2nd person- feminine	ነበርኪ(neber-ki)	ነበርክን(neber-kn)
3rd person-masculine	ነቢሩ(neber-u)	ነበሩ(neber-u)
3rd person-feminine	ነቢራ(nebir-a)	ነቢረን(nebir-en)

Table 3.2: Inflection in Gerundive form of verb

To make the gerundive form of the verb አ/e/, ካ/ka/, ኪ/ki/, ኡ/u/, ኣ/a/, ና/na/, ኩም/kum/, ክን/kn/, አን/en/ morphemes are attached to the root word.

INFLECTION OF IMPERATIVE FORM OF VERB

Person Singular Plural	Person Singular Plural	Person Singular Plural
1st person	ክፅሕፍ(kSHf)	ንፅሕፍ(nSHf)
2nd person-masculine	ትፅሕፍ(tSHf)	ትፅሓፉ(tSHafu)
2nd person- feminine	ትፅሕፊ(tSHfi)	ትፅሕፋ(tSHfa)
3rd person-masculine	ይፅሓፍ(ySHaf)	ይፅሓፉ(ySHfu)
3rd person-feminine	ትፅሓፍ(tSHaf)	ይፅሓፋ(ySHafa)

Table 3.3: Inflection of Imperative form of verb

3.2.1.1.2 INFLECTION OF NOUNS

Tigrigna nouns inflect for case, number, definiteness, and gender (Daniel Teklu, 20068). A noun has the nominative case when it is a subject; accusative when it is the object of a verb; and genitive when it is the object of a preposition. The form of Tigrigna noun is determined by its gender, number, and grammatical case. Most plural nouns are formed by adding a plural marker affix (-ታት or -ኣት) to the singular form. Although when referring to groups belonging to a certain tribe or country –yan is affixed. There are a set of affixes that are used to make plural nouns and are attached as prefix or suffixes to the nouns. The affixes -ታት/-tat/, ኣት/-at/, ኣን/-a/, ኦት/-ot/, ውቲ/-wti/, ቲ/-ti/ are used as suffixes to inflect nouns. Here are some examples to show the inflection of nouns.

Noun Noun	suffix After suffixation	Noun Noun
ባሕሪ	ባሕሪ-ታት	ባሕሪታት
እምባ	እምባ-ታት	እምባታት
ስእሊ	ስእሊ-ታት	ስእልታት
ሃገር	ሃገር-ኣት	ሃገራት
ሰማይ	ሰማይ-ኣት	ሰማያት
ምእመን	ምእመን-ኣን	ምእመናን
መምህር	መምህር-ኣን	መምህራን

Table 3.4: Inflection of Imperative form of verb

Another form of inflection for nouns is created by attaching the morpheme ኣ/a-/ as prefix. For example in the words ገረብ(gereb, forest), ፈረስ(feres, horse), ከረን(keren, mountain), ኣግራብ(agrab, forests), ኣፍራስ(afras, horses), ኣክራን(Akran, mountains). Tigrigna has two grammatical genders: masculine and feminine, and all nouns belong to either one or the other and inanimate objects are take one of the genders. Some noun pairs for people distinguish masculine and feminine by their endings, with the feminine signaled by ኢት/it/ and the masculine signaled by ኢ/i/. These include agent nouns derived from verbs — ከፈተ kefete 'open', ከፋቲ kefati 'opener (m.)', ከፋቲት kefatit 'opener (f.)' — and nouns for nationalities or natives of particular regions - ትግራዋይ tgraway 'Tigrean (m.)', ትግራወይቲ tgraweyti 'Tigrean (f.)'.

3.2.1.1.2 INFLECTION OF ADJECTIVES

Tigrigna adjectives inflect for number and gender. Tigrigna adjectives may have separate masculine singular, feminine singular and plural forms, and adjectives usually agree in gender and number with the nouns they modify (Daniel T., 2008). The plural forms follow the same patterns as noun plurals; that is, they may be formed by suffixes or internal changes or a combination of the two. The affixes that are used for the inflections of the adjectives are ኦ/o/, ቲ/ti/, ኣት/at/, ኣን/an/ and ኦት/ot/. Table 3.5 shows inflection adjectives for number.

Singular Adjectives	**Plural Adjectives**	**Affix attached**
SheKali	SheKalo	-o
Kedani	Kedano	-o
Mehazi	Mehazti	-ti
Qetal	Qetelti	-ti
Kbur	Kburat	-at
Senef	Senefat	-at
Harestay	Harestot	-ot
Zebenay	Zebenot	-ot

Table 3.5: Inflection of Adjectives

Adjectives are also inflected for gender by adding the infix –a- and suffix –ti. For example, words such as qeyaH(fm) , qeyaHti and qeyh(m), qetan(fm) ,qetenti and qetin(m) ,belaH(fm), belaHti and beliH (fm) are inflected forms of the qyH,qtn,blH respectively.

3.2.1.1 DERIVATIONAL MORPHOLOGY OF TIGRIGNA

Derivational morphology describes how affixes combine with word stems to derive new words. Derivational affixes may affect the part-of-speech and meaning of a word.

3.2.1.1.1 DERIVATION OF VERBS

Unlike the other word categories such as nouns and adjectives, the derivation of verbs from other parts of speech is not common. Almost all Tigrigna verbs are derived from root consonants, as indicated by Daniel (2008). Traditionally a distinction is made between simple and derived verbs.

Simple verbs are those verbs derived from roots by intercalating vowel patterns whereas derived verbs are considered as derivatives of simple verbs. The derivation process can be an internal one in which consonant-vowel patterns are changed, an external one where derivational affixes are attached to the simple derived verbs or a combination of the internal and external derivational processes. The derivation of causative, passive, repetitive and reciprocal verbs is presented below.

1. **Causative**: Causative verbs are derived by adding the derivational morphemes ‘a- and to the verb stem as in the examples በፅሐ /beSHe-/ ‘arrive’ - ኣብፅሐ- /‘abSHe/ ’cause to arrive’ and ወሰደ /wesed/ ’take’ ኣውሰደ /awsede-/ ’cause to take’. In most cases the ‘a- morpheme is used to form causative of intransitive verbs, transitive ones and verbs of state. Some exceptions are the verbs that begin with ‘a, always take the morpheme ‘a but add the morpheme I after the morpheme ‘a to form causative e.g. ኣሰረ /‘asere/, ኣእሰረ/‘aIsere /.
2. **Passive/Reflexive**: The passive verbs are derived using the derivational morpheme ተ/te/. This derivational morpheme is realized as ተ-/te-/ before consonants and as ት-/t-/before vowels. Moreover, in the imperfect, jussive and in derived nominal like verbal noun, the derivational morpheme ት/t-/ is used. In this case, it assimilates to the first consonant of the verb stem, and as a result, the first radical of the verb geminates. Some exceptions are intransitive verbs like ፈሊሑ /faliHu/ ’it boiled’ that form their passive forms using the prefix ተ- /te-/ as in ተፈሊሑ/tafeliHu/ ’it was boiled’. Such kind of verbs can derive their passive from their causative form (ኣፍሊሑ/afliHu/’he boiled’).

3. **Reduplicative/repetitive:** Reduplicative stems indicate an action which is performed repeatedly. For tri-radical verbs, such stems are formed by duplicating the second consonant of the root and using the ኣ-/a-/ after the duplicated consonant as in ሰባበረ/seba-bere/ ’he broke repeatedly’ derived from the root ስብር/sbr/ break. All verb types, Type A, B and C have the same reduplicative forms.
4. **Reciprocal**: Reciprocal verbs are derived by prefixing the derivational morpheme ተ- /te-/ either to the derived type C forms (that use the vowel a after the first radical) or to the reduplicative stem. For example, reciprocal forms of ተቓተሉ /teqatelu/ ’killed each other’ and ተቐታተሉ/teqetatelu/ ’killed one another’ are derived from the derived type C stem qatelu- and reduplicative stem qetatelu-, respectively. The causative of reciprocal verbs are formed by adding the causative prefix ’a- to the reciprocal verb forms. However, the reciprocal verb prefix t- or ‘a- assimilates to the stem-initial consonant (thus causes the first radical of the stem to geminate) and does not show up in the surface form of the reciprocal causative.

3.2.1.1.2 DERIVATION OF NOUNS

Tigrigna nouns can be either primary or derived. They are derived if they are related in their root consonants and/or meaning to verbs, adjectives, or other nouns. Otherwise, they are primary. For example, a noun እግሪ /Igri/ ’foot, leg’ is primary but, እግረኛ/IgreNa/ ’pedestrian’ is derived from the nominal base Igri by adding the morpheme –‘eNa. Nouns are derived from other nouns, adjectives, roots, stems, and the infinitive form of a verb by affixation and intercalation. The morphemes-ነት/-net/, -ኢት/-`it/, -ኣት/-`at/,-ኡት/ut/, -ትኦ/-to/, -ኦ/o/, -ኢ/i/, -ኣ/a/, -ኣን/an/, -ኣኛ/-‘eNa/, -ኛ/-Na/, -ኣት/-et/, ኣዊ/-awi/, -ተኛ/-teNa/, -ና/-na/ and the prefix መ-/me-/ are used to derive nouns from other nouns. From the adjectives, nouns can be derived using the suffixes /net/ and /-et/ as in the examples /merzamanet/ ’generosity’ which is derived from the adjective /merzam/ ’poison’ and flTet ’knowledge’ from the adjective fluT ’known’. Nouns can also be derived from verbal roots by intercalation and affixation. Table 3.6 shows some examples of derived nouns from other nouns.

Base form	Bound morpheme	Derived noun
Mskr	-net	mskrnet
Selam	-awi	selamawi
Areb	-Na	ArebNa

Xlm	-at	Xlmat
bSh	-it	bSHit
Whb	-to	whbto
Dfn	-o	dfno
Hlm	-i	Hlmi
Lmn	-a	lmena

Table 3.6: Nouns derived from other nouns

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

3.2.1.1.3 DERIVATION OF ADJECTIVE

Adjectives in Tigrigna include all the words that modify nouns and can be modified by the word ብጣዕሚ /btaimi/ 'very, greatly' (Daniel T., 2008). As it is true for nouns, adjectives can also be primary (such as ለዋህ /lewah/ 'kind') or derived, although the number of primary adjectives is very small. Adjectives are derived from nouns, stems or verbal roots by adding a suffix and by intercalation. The suffixes ኣም/-am/, -ዊ/-wi/, -ኣዊ/-awi/, -ኣይ/-ay/, -ኣታይ/-atay/,-ታይ/-tay/, -አኛ/-eNa/ and -ኣዋይ/-away/ are used in the derivation of adjectives from nouns. For example it is possible to derive ሃፍታም/haftam/'rich, wealthy', ተንኮለኛ/tenkoleNA/, ዘበናዊ /zebenawi/'modern 'and ማእኸላይ/maIKelay/ 'central' from the nouns ሃፍቲ/hafti/ 'wealth', ተንኮል/tenkol/ '', ዘበን/zeben/ 'period' and ማእኸል/maIKel/ 'center', respectively. Adjectives can also be derived either from roots by intercalation of vocalic elements or attaching a suffix to bound stems.

CHAPTER FOUR

METHODOLOGY

This study aims to model English to Tigrigna machine translator using three types of corpus. Towards that, various data collection and preprocessing tasks have been performed and different tools were selected to be used for the experiment.

4.1Data Collection methods

A statistical machine translation requires two basic corpora namely monolingual and bilingual corpus. The monolingual corpus is used to train the language model of the target language, in this case Tigrigna. Since there is no well-prepared linguistic resource for Tigrigna, the researcher collected documents from the international news website http://www.voanews.com/ and the Bible, obtained from www.geezexperience.com, as a monolingual corpus for the language model. But, since the Bible is available in both English and Tigrigna and contains more sentences, it has become the first candidate to be used as a parallel corpus.

The corpus associates probabilities with translations empirically by counting co-occurrences in the data. Estimates of probabilities get more accurate as the size of data increases. With these regards, the researcher has used 31,256 English sentences and 31,234 Tigrigna sentences before preprocessing.

4.2 Preprocessing techniques and algorithms

In order to prepare the raw data collected for the machine translation tools (IRSTLM, GIZA++, MOSES), preliminary preprocessing task which are sentence level segmentation and tokenization were performed. These preprocessing tasks were implemented in python. In addition to that a lot of manual cleaning tasks were done when the preprocessing task required the researcher's judgment.

The bible both in English and Tigrigna were obtained in PDF format. Each corpus contains information related to chapter of each verse , page numbers, content description, publisher, date and place of publication. Furthermore, there is also pagination, header and footer content attached to all pages. This contents were removed manually by going through each page to keep the consistencies in both documents.

4.2.1 Sentence level segmentation

Before any automated alignment was made, the researcher has to put each corpus one sentence per line. For this purpose a sentence segmenter code was written using python. Since no abbreviation acronyms exists in the bible the program finds "**.**" on the English corpus and "***::***" on the Tigrigna corpus and puts each sentence per line in a new file. Algorithm 4.1 shows the sentence segmentation

```
open corpus for reading
while not end of corpus do
        read words
        for each word in the file
                if space end of line is reached //depending on the language
                        write into new file with new_line character
                end if
        end for
close file
```

Algorithm 4.1. Sentence level segmentation Algorithm

Although parallel religious data is mostly sentence by sentence aligned but after sentence segmentation, because of some unknown reasons, the researcher found some misalignments in the data. Due to only 2 to 3 unaligned sentences the researcher had to manually analyze the entire corpora and find the proper locations in the corpora with mismatch sentences. Output of this phase is the sentence by sentence aligned corpora ready for the cleaning process.

4.2.2 Word Tokenization

Tokenization is the process of splitting a sentence into words so as to make it ready for subsequent processes namely, POS tagging and stemming. The main tasks of tokenization in this research is removing of punctuation marks and putting one word per line. Tokenization was a challenge in cases where spaces were missed and two words were found to be attached together.

In such situations, the researcher was required to do some corrections manually. Algorithm 4.2 shows the algorithm for tokenization in the context of this research.

```
open corpus for reading
while not end of corpus do
        read word
        for each word in the file
                if space is found
                        write into new file with new_line character
                end if
        end for
close file
```

Algorithm 4.2. Tokenization Algorithm

4.2.3 Morphological segmentation

Segmentation is a technique used to separate morphological variants of a word into its stem word, prefixes, affixes and suffixes. This task is an important task for Tigrigna texts preprocessing due to the morphological richness of the language. As discussed in the previous chapter, Tigrigna makes use of prefixes suffixes and infixes to create different word form.

In machine translation the word *በለዐ* and *ስለዘይተበለዐ* are considered as two different words not as an extension of the other. For this research segmentation is performed on the target side, which is on Tigrigna words. Different tools where considered for the segmentation of the corpus. Primarily segmentation was conducted using morfessor. After general look at the output, most of the words are not segmented well, the researcher believes morfessor is not suitable for Tigrigna language. Therefore the researcher was forced to use the stemming algorithm developed by (Yonas, 2011). He implemented a rule based stemmer algorithm where he primarily collected the list of prefixes and affixes and used sequential checking and removing of affixes. The researcher claims that the stemmer has an accuracy of 85.8% and 86.3% using two different corpus. In this research the pseudo code obtained was re-written using python and modified to segment the prefix and suffix rather than removing them. Algorithm 4.3 shows the algorithm used for segmentation.

```
open corpus for reading
while not end of corpus do
    read word
    for each word in the file
        for each prefix in list_of_prefixes
            if word begins with prefix
                segment word using '_'
            end if
        end for
        for each suffix in list_of_suffixes
            if word ends with suffix
                segment word using '_'
            end if
        end for
        write segmented word to new file
    end for
close file
```

Algorithm 4.3. Segmentation Algorithm

Morphological segmentation was also performed for English using Morfessor to obtain the lemma of the preparation of factored corpus. The Open American National Corpus (OANC) was used to train the segmenter. The output of the segmenting algorithm is in the same order of the original file. This is important to keep the document from losing its sentence structure.

4.2.4 POS tagging

In corpus linguistics, part-of-speech tagging, also called grammatical tagging, is the process of marking up a word in a corpus as corresponding to a particular part of speech, based on both its definition, as well as its context. Relationship with adjacent and related words in a phrase, sentence, or paragraph. POS tagging is required in this research to prepare factored corpus for both English and Tigrigna. Tagging the English corpus was easy compared to Tigrigna. In this research supervised tagging using SVMTool was done. The researcher has used OANC corpus tagged with the pen treebank tag set to train the tagger. Tagging the Tigrigna corpus was the difficult and time consuming task. Since there is no manually annotated corpus to train the tagger, the researcher tagged a 1,018 words.

4.5 Sentence Alignment

Sentence alignment is the process of mapping a two sentences from two different corpus. The aligned corpus will then be used to train the system. The alignment at the sentence level has been done using a sentence aligner called Hunalign. Hun align aligns bilingual text at sentence level using sentence-length information. In the simplest case, its output is a sequence of bilingual sentence pairs. In the presence of a dictionary, it combines this information with sentence length information. A small English-Tigrigna bilingual dictionary, which is obtained from *www.memhr.org*, of word lists sized 8,212 have been used. The aligner was able to align 25,578 English sentences and 25,730 Tigrigna sentences.

Although parallel religious data is mostly sentence by sentence aligned but after alignment, the researcher found some misalignments in the data. The misalignments is the result of two sentence appearing as one sentence separated using semicolon in the English corpus where the sentence appears as two sentence in the Tigrigna corpus. The researcher had to manually analyze the entire corpora and find the proper locations in the corpora with mismatch sentences.

Those sentences that do not have matching translations (0-1 or 1-0) have been dropped. Those sentence pairs with more than 200 tokens in length have been dropped as well in order to get a better performance of the decoder. As a result, 17,649 English sentences aligned with Tigrigna sentences have been retained and used for this experiment.

4.5 Factored Corpus Preparation

The state-of-the-art approach to statistical machine translation, so-called phrase-based models, is limited to the mapping of small text chunks without any explicit use of linguistic information, may it be morphological, syntactic, or semantic (Koehn et al., 2007). Such additional information has been demonstrated to be valuable by integrating it in pre-processing or post processing steps.

To prepare the factored corpus a program was written using python. The Input corpus to the program are result of the above preprocessing task. Primarily both the English and Tigrigna corpus was tokenized in to one word per line and saved to new files each. During this process the sentence ending are preserved. The new files are then POS tagged segmented to form *word | POS_tag* format. Morfessor was used to segment the English words whereas the stemmer by (Yonas, 2011) was used for Tigrigna. On Tigrigna side the segmenter uses POS tagged file, extracts the word from the word-POS tag combination while retaining relationship with tag and segments it into the lemma, prefixes and suffixes. Whereas on the English side two files where used where the first file contains one not segmented tokens per line and the other one contains space separated segmented tokens per line. At the last step, the surface form (original word) is attached at the beginning to form the factored corpus. Fig. 4.1 shows the flow of the algorithm.

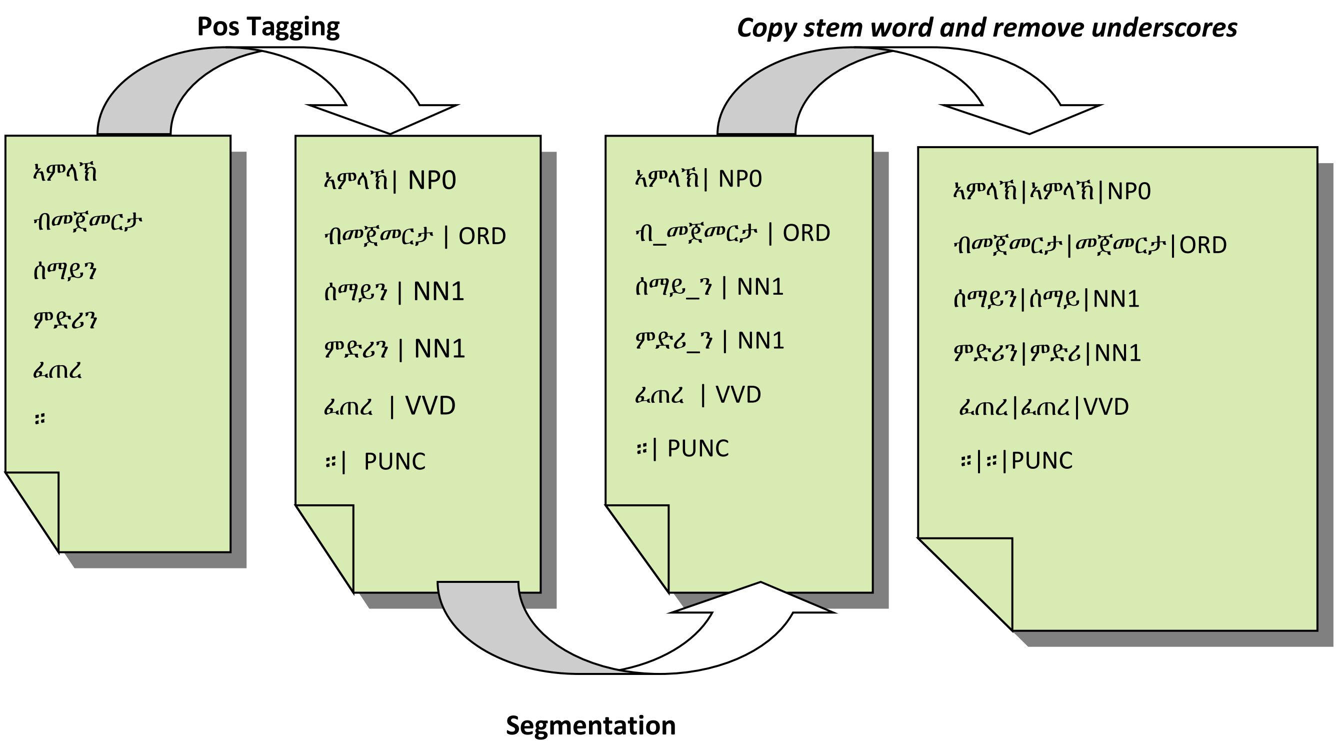

Fig. 4.1: Tigrigna Factored Corpus preparation flow

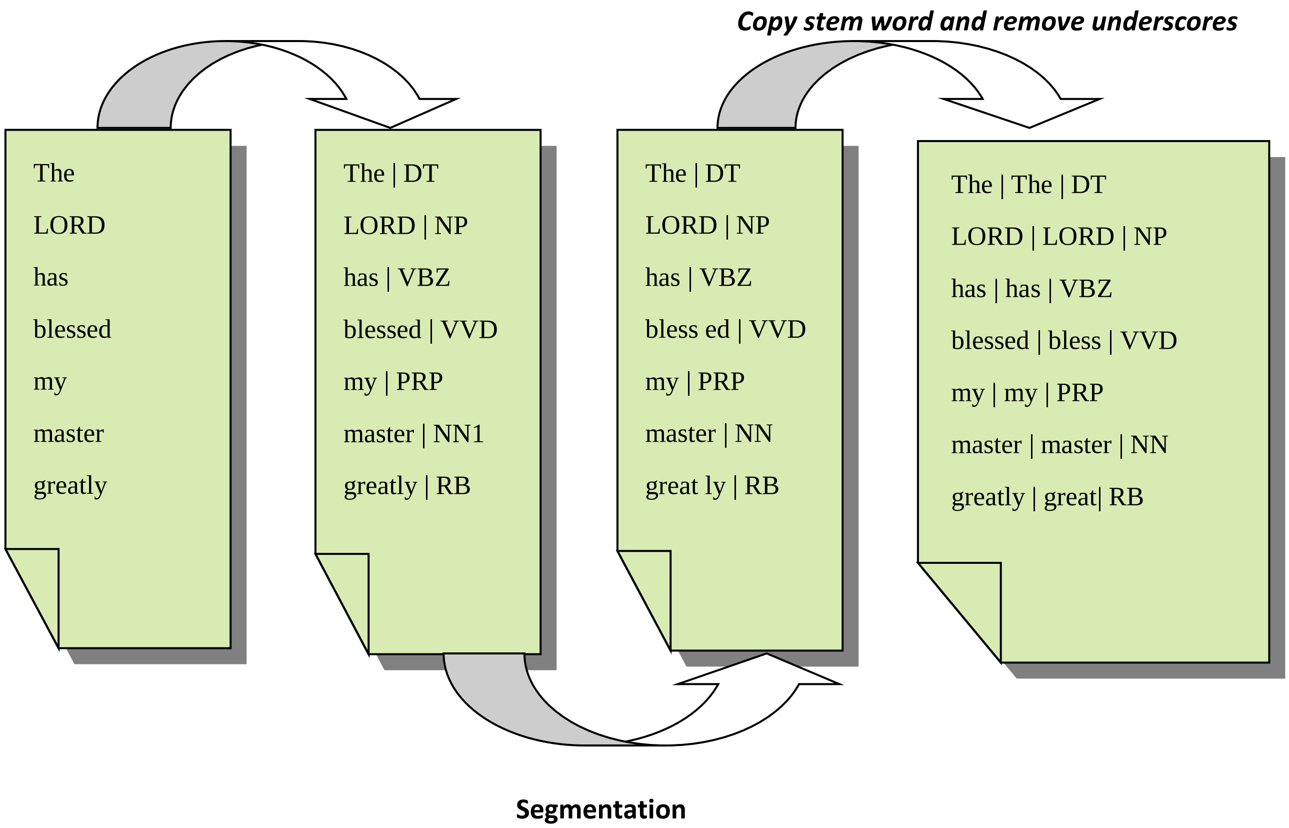

Fig. 4.2: English Factored Corpus preparation flow

Below, is the algorithm used to prepare the factored corpus for both languages.

```
open segmented corpus for reading
open tagged corpus for reading
while not end of tagged corpus do
        read word and save to t_ list
 while not end of segmented corpus do
        read word and save to list
for each_word in the t_list //t_list stands for Tagged corpus
                if each_word =='. ' // or ':: ' for Tigrigna
                        final_word=tlist[i] + '|' + stem + '|'+ tlist[i+1]+'/n'
                if each_word !='. ' // or ':: ' for Tigrigna
                        temp1=t_list[i]
```

```
            for  all_character in temp1
                    if all_character=="_"
                            temp2=temp1[start+1:]
            for all_character in temp2
                    if all_character=="_"
                            stem=temp2[:end-1]
                    end if
            end for
        final_word=tlist[i] + '|' + stem + '|'+ tlist[i+1]
        write final_word to file
close files
```

Algorithm 4.4 Factored corpus preparing algorithm for both languages

CHAPTER FIVE

EXPERIMENTATION AND DISCUSSION

This chapter presents the results of different set of experiments carried out for this study. The necessary detail of corpora that are used during the experiments is presented in the previous Chapter. The chapter starts with the experimental setup, followed by the description of the evaluation measure used to evaluate translation output. The main part of the chapter focuses on presenting and discussing the improvements in translation quality. The three major experiments conducted for this study are: baseline experiments, experiments with segmentation and experiments using factored based model. The chapter concludes by comparing the results and translation quality of the generated output using different experimental setups.

5.1 Experiment Setup

This section describe the toolkit used for building the language model. It also illustrate about the translation system used for conducting the experiments. It further discuss the translation procedure, together with the different parameter settings adopted for carrying out the experiment.

The experiment is conducted on a machine with good performance to run all tools smoothly.

System Environment	
Manufacturer	Dell
Model	OptiPlex 7010
Processor	Core i3-3220 CPU
Processor Speed	3.30 GHz(4 CPUs)
Memory	5GB
OS	Ubuntu 12.04

Table 5.1: System Environment for Baseline and segmented systems

System Environment	
Manufacturer	HP
Model	--
Processor	Intel Xeon CPU
Processor Speed	3.30 GHz(2 CPUs)
Memory	12 GB
OS	Ubuntu 12.04

Table 5.2: System Environment for Factored systems

The Statistical Language Modeling Toolkit

There are various software packages available to build Statistical Language Model. For example, the SRI Language Modeling toolkit. In this study, we use SRILM (Stolcke, 2002). SRILM toolkit is composed of set of tools for building and applying Statistical Language Models (LMs). The main purpose of SRILM is to support Language Model estimation and evaluation. Estimation means the creation of a model from training data; evaluation means computing the probability of a test corpus (Stolcke, 2002).

For this study, the researcher use the SRILM tool ngram-count to estimate two language models. One language model is built upon a text monolingual Tigrigna data. Second language model is comprised of part-of-speech tagged monolingual data.

Training and Translation System

The statistical phrase-based machine translation system, Moses (Koehn, et al., 2007), is used in this work to produce English-to-Tigrigna translation. According to (Koehn, et al., 2007) “The toolkit is a complete out-of-the box translation system for academic research. It consists of all the components needed to preprocess data, train the language models and the translation models. It also contains tools for tuning these models using minimum error rate training (MERT) (Och, 2003)”.

Moses automatically trains the translation models on the parallel corpora of the given language pair. It uses an efficient algorithm to find the maximum probability translation among the exponential number of candidate choices.

Translation Setup

The training process in Moses takes nine steps and all of them are executed using the script train-phrase-model.perl and train-factored-phrase-model.perl for baseline/segmented and factored corpus respectively. The training steps, external tools used for the training by Moses and also the parameters settings at each step are described below:

I. ***Prepare Data***: the selected corpus for the experiment is first cleaned using tok-dan.perl script. It removes the redundant space characters. It also removes the extra spaces on the start and end of the line. The English data is then converted to lowercase using the lowercase.perl script provided with the Moses implementation.

II. ***Word Alignment***: the researcher uses MGIZA++ toolkit which is freely available implementation of IBM models for extracting word alignments. Alignments are obtained by running the toolkit in both translation directions and then symmetrising the two alignments.

III. ***Extract Phrase***: Using the generated word alignment, Moses estimates the Maximum likelihood lexical translation table and extracts all those phrases in which words are aligned only to each other and not to any word outside the phrase.

IV. ***Score Phrases***: Phrases are scored from the stored phrase translation table. For each pair five different phrase translation scores are computed.

V. ***Reordering***: Moses builds the lexicalized reordering model that conditions the reordering on the actual phrases. It provides three different reordering models (i.e. different types of orientation of the phrases) together with number of variations of the lexicalized reordering model based on the orientation types.

VI. ***End of Training***: after creating reordering table, generation table is built using the target side of the training corpus.

After training the translation model, Moses standard MERT is executed on development set for tuning the weights of the individual models in our setup.

Evaluation Measures

One of the most difficult tasks in Machine Translation is to evaluate the output of the system. For this study the researcher have selected the BLEU (Bilingual Evaluation Understudy) as an evaluation metric. The Bleu metric is an IBM-developed metric and very well known for the machine evaluation for the machine translation. It checks how closer the candidate translation is to the reference translation based on the n-gram comparison between both translations. The Bleu score is based on the number of correct n-gram matches between candidate and reference translation, and these matches are position-independent.

The Bleu metric ranges from 0 to 1. If the candidate translation is identical to the reference translation it will attain the score 1 and 0 in case of no similarities. Bleu metric is based on the modified n-gram precision measure for comparing the candidate translation against multiple reference translations.

$$Precision = \frac{Number\ of\ words\ from\ the\ candidate\ that\ are\ found\ in\ the\ reference}{Total\ \ number\ of\ words\ in\ the\ candidate}$$

The metric modifies simple precision since MT system can over generate reasonable words, resulting in implausible, but high-precision, translations like the following example

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

All of the seven words in the candidate translation appear in both reference translations, thus the candidate text is given the unigram accuracy that :

$Unigram\ Precision = \frac{7}{7} = 1$

Now, for modified unigram precision calculation, for each word in the candidate translation, Bleu calculates its maximum total count in any of the reference translations. So in the previous example above, “the” appears twice in reference 1 and once in reference 2 so it’s MaxCount = 2. Now the total count of each word (Wc) in the candidate translation that is 7 for “the” in our example, is clipped to its MaxCount. Wc is then summed over all the words in the candidate translation.

$$\text{Modified } Unigram\ Precision = \frac{2}{7} = 0.28$$

Brevity penalty is introduced in the metric to penalize the shorter translations to receive too high score. Let, c be the length of the candidate translation and r be the effective reference corpus length. The brevity penalty (BP) is computed by,

$$\mathbf{BP}=\begin{cases} \mathbf{1} & \boldsymbol{if\ c > r} \\ \boldsymbol{e(1 - r/c)} & \boldsymbol{if\ c \leq r} \end{cases}$$

The final Bleu score is calculated by computing the geometric average of the modified n-gram precision, p_n using n-grams up to length N and positive weights w_n summing up to 1.

$$\boldsymbol{BLEU = BP \,.\, \exp(\textstyle\sum_{n=1}^{n} Wn\ log\ \mathrm{p}n)}$$

While it is better to use several independent reference translations (usually 4 if available), our English-Tigrigna parallel data contain only 1 reference translation per sentence.

Proposed System Architecture

Finally the system architecture is designed to represent the end to end process of the machine translation task. Fig. 5.1 shows the architecture of the proposed Phrase based and factored statistical Machine Translation. As can be seen from the diagram the system takes in parallel and monolingual corpus as input. In the case of factored translation the input parallel corpus contains factored word. These sentences are used to develop language and translation models respectively.

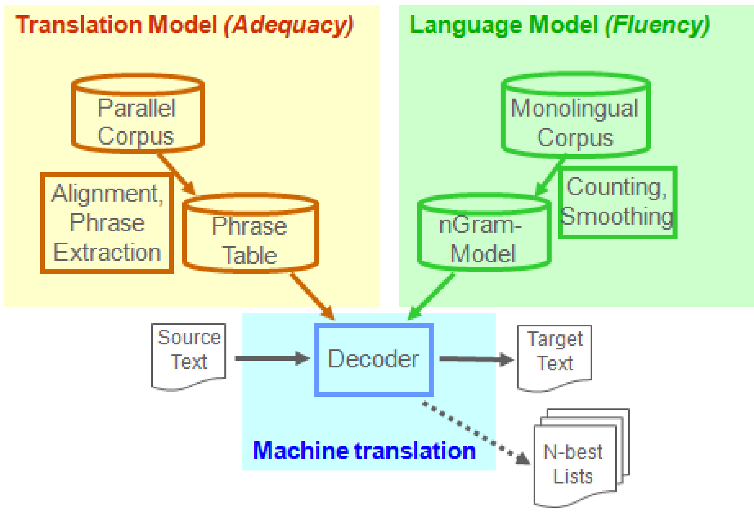

Fig: 5.1: Architecture of the proposed system

15.2 Experiment results

The baseline setup is a plain phrase-based translation model (i.e. single factored). The segmented translation system was trained and evaluated using segmented corpus whereas the factored translation model was trained and evaluated using factored corpus developed in a series of data preparation operations discussed in the previous chapter. Data is divided in training set, and test set.

Corpus	Sentence Pairs			English Tokens			Tigrigna Tokens		
	Training size	**Test Size**	**Total Sentence Pair**	**Training size**	**Test Size**	**Total Sentence Pair**	**Training size**	**Test Size**	**Total Sentence Pair**
Bible	17,000	649	**17,649**	527,830	19,367	**547,197**	541,958	20,936	**562,894**

Table 5.3: Prepared corpus for segmented experimentation

Corpus	Sentence Pairs			English Tokens			Tigrigna Tokens		
	Training size	**Test Size**	**Total Sentence Pair**	**Training size**	**Test Size**	**Total Sentence Pair**	**Training size**	**Test Size**	**Total Sentence Pair**
Bible	17,000	649	**17,649**	527,830	19,367	**547,197**	531,410	19,823	**551,233**

Table 5.4: Prepared corpus for Factored experimentation

The token size of the Tigrigna corpus is around 4036 words more than the token size of English corpus for the baseline and 15,697 for the segmented system. The same amount of sentence pair amounts are used for the baseline and factored systems.

As for average sentence length is between 100 to 131 words on average for English side of parallel corpus and 30 to 67 words on average for Tigrigna side of the parallel corpus. The English Bible corpus contains a few extraordinarily long sentences, with a size of even around 240 words. While, in Tigrigna Bible corpus the sentence length is roughly around 40 words and the maximum sentence length consists of around 170 to190 words. Fig. 5.2 and Fig. 5.3 shows the sentence length distribution of English and Tigrigna respectively after preprocessing.

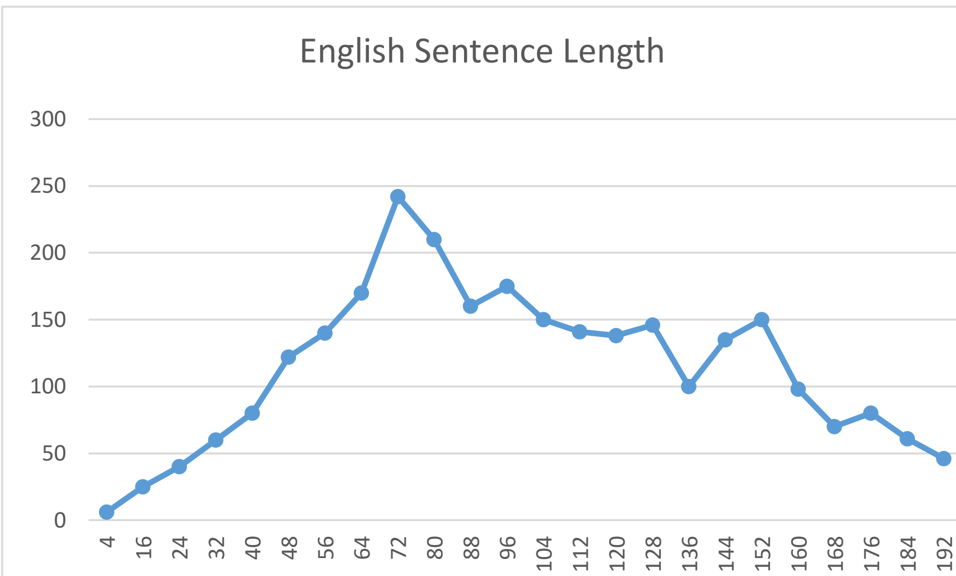

Fig: 5.2 English Sentence length

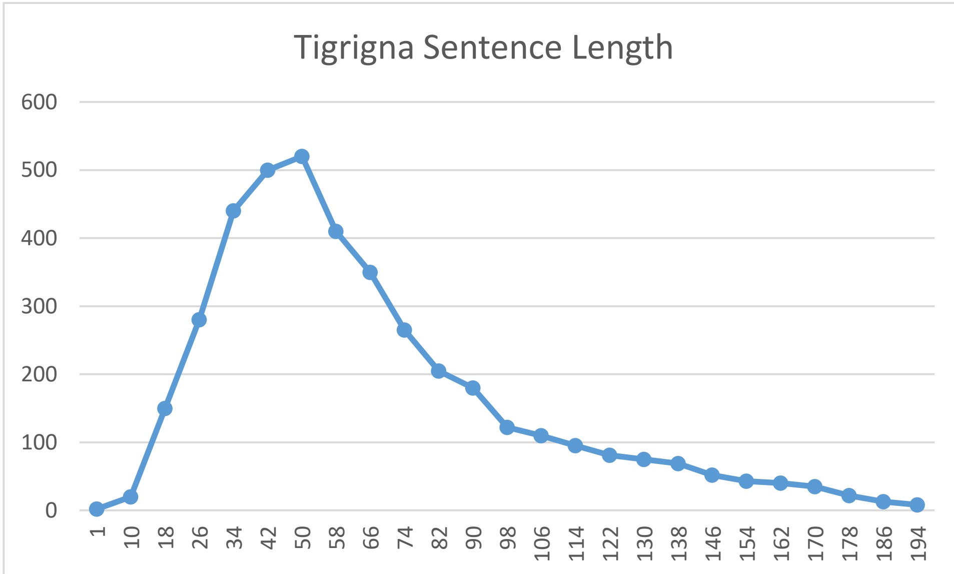

Fig: 5.3 Tigrigna Sentence length

In Table 5.5 presents the results that are achieved after performing the experiments using the three corpus.

		SYSTEM		
		Baseline	**Segmented**	**Factored**
REFERENCE	**Baseline**	**21.04%**	-	-
	Segmented	21.23%	**22.65%**	-
	Factored	-	-	**16.5%**

Table 5.5: A BLEU score of the three systems

5.3 Discussion

In Table 5.5, the outputs of the baseline, segmented and factored SMT have been scored with two types of reference: segmented and un-segmented. In the case of the segmented reference and the un-segmented SMT, the output of the system was re-segmented to be made consistent with the reference.

However, for the segmented system and un-segmented reference the researcher was unable to re-attach the segmented words system output to make it compatible with the un-segmented reference. The researcher see that the segmentation has contributed for the overall performance of the segmented system that has shown better performance compared to the baseline phrase-based system. When compared with the same segmented reference, the BLEU score for the segmented system is 22.65% which is a 1.61% increase from the baseline system that has a BLEU score 21.04.

The factored corpus has shown a decrease of 6.15% from the segmented and 4.53% from the baseline system. The researcher believes that, the low performance of the factored system is accounted to the POS tags attached since the tagger was trained using a small manually tagged corpus prepared by the researcher.

Table 5.6 shows input sentence from the Bible corpus, its reference translation and its respective output translation obtained using the first baseline experimental settings.

Input	And God divided the light from the darkness
Reference	ኣምላኽ ከኣ ነቲ ብርሃን ካብ ጸልማት ፈለዮ
Output	ኣምላኽ ከኣ divided ነቲ ብርሃን ካብ the darkness

Table 5.6 Output translation of baseline system

There are few issues associated with the translation generated by the baseline system. One of the major issue is the wrong syntactic ordering of phrases/words. We can see in table 5.6 that the baseline system is unable to model the translation between language-pair that have different word order structures. It also finds it difficult to translate verbs compared to other sentence parts such as nouns and conjunctions.

Input	Do|do|VDI not|not|XX0 seal|seal|VBD the|the|AT0 words|word|NN1 of|of|PRF the|the|AT0 prophecy|prophecy|NN1 of|of|PRF this|this|DT0 book|book|NN1 for|for|PRP the|the|AT0 time|time|NN1 is|is|VBD at|at|PRP hand|hand|NN1
Reference	ድማ|ድማ|CJC እቲ|እቲ|AT0 ጊዜ|ጊዜ|NN1 ቐረባ|ቐረባ|AJ0 እዩ|እዩ|VBZ እሞ|እሞ|AV0 ንቓል|ቓል|NN1

	ትንቢት|ትንቢት|NN1 እዚ|እዚ|AT0 መጽሓፍ|መጽሓፍ|NN1 ኣይትሕተሞ|ሕተሞ|NN1
Output	Do|do|VDI ኣይትሕተሞ|ሕተሞ|NN1 እቲ|እቲ|DT ንቓል|ቓል|NN1 of|of|PRF the|the|AT0 prophecy|prophecy|NN1 of|of|PRF እዚ|እዚ|AT0 መጽሓፍ|መጽሓፍ|NN1 for|for|PRF the|the|AT0 time|time|NN1 is|is|VBD at|at|PRP ኢድ|ኢድ|NN1

Table 5.7: sample Factored translation

There are few issues associated with the translation generated by the system. Firstly it failed to translate most of the words from the input sentence. Although the word "ኢድ" is also a correct translation of word "hand" it is wrong translation based on the context of the reference sentence. In other translation, it is also seen that, some words are translated with sequence of POS tags missing the surface words and lemma.

CHAPTER SIX
CONCLUSION AND RECOMMENDATIONS

In the preceding chapters we have seen the specific improvement techniques in the domain of statistical machine translation for English Tigrigna language pair. The general idea was to produce the grammatically coherent and human understandable translation given the input English sentence. In this final chapter, the approach and substantiate key results are summarized. we close this study work by drawing conclusions and giving recommendation for future research.

6.1Conclusion

The achievement of this research has been the development of Factored Statistical Machine Translation System for English to Tigrigna language by integrating linguistic features. Though Tigrigna is a linguistic resource scarce language, the researcher has managed to POS tag the corpus and segment it as part of this research work. Performing the above linguistic operations are challenging and demanding tasks especially for highly inflectional language like Tigrigna.

The performance of the statistical and machine learning methods mainly depends on the size and correctness of the corpus. If the corpus consists of all types of surface word forms, word categories and sentence structures, then it is possible for a learning algorithm to extract all required features. Preprocessing systems are automated for creating factored parallel and monolingual corpora. Though Factored corpora are an essential resource for developing a good Machine Translation system, for this research, it has shown a result that do not justify the costs of the added work.

Based on the experiments carried out for this study and the results obtained, the following conclusions are presented.

- The result obtained shows that the system translate the words with a maximum accuracy of 21.04% using baseline, 22.65% using Segmented and 16.5% using factored translation system using un-segmented and segmented references.

- The unavailability of manually tagged corpus and Tigrigna tag set has contributed for the low performance of the tagger.
- Segmentation has proven to increase the performance of the translation system where as the performance of factored translation system has significantly decreased.
- The accuracy of the stemmer by (Yonas, 2011) is not known on different corpus; it solely relies on the corpus used by the researcher making hard to assess the impact of the stemmer on the translation system.
- Based on an observation on the translation output of the three systems, the system was good in translating conjunctions such as 'ድማ' ፣ 'ግና' and nouns; whereas poor in translating Verbs.

6.2 **Recommendations**

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

- The Parallel corpus used in these research, the bible, contain long sentences and complex words. Further research should be conducted using a larger corpus from different domains.
- An efficient POS tagging must also be performed using a tagger that is trained with a large amount of tagged corpus. The involvement of linguists in such situation is vital.
- Due to the unavailability of a full morphological analyzer for Tigrigna, the segmentation performed is using a stemmer. A complete morphological analyzer and segmenter should be developed to obtain optimal result in segmented and factored translation systems.
- Sentence order mismatch is one of the drawbacks of the three translation systems. Therefore, the researcher recommends conducting further experiments by adding sentence reordering as a feature.
- A more thorough evaluation should be conducted on the segmented translation system with the outputs re-attached to make it compatible with the un-segmented reference. The researcher was unable to perform this task due to time limitation.
- Further research should also be conducted by integrating more linguistic features.

References

Daniel Jurafsky and James H. Martin, Speech and Language Processing, Pearson Education Inc, 2005.

Hutchins John, Machine translation and human translation: in competition or in complementation? International Journal of Translation, 2001

Aynalem Tesfaye, Kevin Scannel, Amharic – English Cross-lingual Information Retrieval: A Corpus Based Approach, Master's Thesis, Haramaya University

Hutchins John, Machine translation and human translation: in competition or in complementation? International Journal of Translation, 2001

Michael Gasser. Toward a Rule-Based Sys-tem for English-Amharic Translation. In LREC-SALTMIL - AfLaT Workshop on Language technology for normalization of less-resourced languages, 2012

Mulu Gebreegziabher Teshome, Laurent Besacier (Prof). Preliminary Experiment on English-Amharic Statistical Machine Translation (EASMT). In: Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU), 2012

Michael Gasser, HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrigna, Indiana University, USA

Hutchins, W. J. and Somers, H. L., An introduction to machine translation, Academic Press, London. 1992

Bar-Hillel, the present status of automatic translation of languages. Advances in Computers, 1960

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, andPaul S. Roossin, A Statistical Approach to Machine Translation, Computational Linguistics,1990

Ananthakrishnan Ramanathan, Statistical Machine Translation, PhD seminar Report, Indian Institute of Technology

Dugast, L., Senellart, J., Koehn, and P.: Statistical post-editing on SYSTRAN's rule based translation system. In: Proceedings of WMT07, Prague, Czech Republic, and Association for Computational Linguistics, June 2007

Chiang, D., A Hierarchical Phrase-Based Model for Statistical Machine Translation. ACL, 2005

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). the mathematics of statistical machine translation: Parameter estimation. Computational Linguistics

Philipp Koehn and Hieu Hoang. Factored Translation Models. In Proc. of EMNLP/CNLL, 2007

Locke W N, Booth AD, Machine translation of languages. MIT Press, Cambridge, Mass. 1955

Booth A.D (ed) Machine translation. North-Holland, Amsterdam. 1967

ALPAC Language and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee. National Academy of Sciences, Washington, DC. 1966

MT Summit 3: MT Summit III, July 1-4 1991, Washington D.C., USA.

Arnold D, Machine translation: an introductory guide. NCC/Blackwell, Manchester/Oxford. 1994

Hutchins W J. Machine translation: past, present, future. Ellis Horwood, Chichester, UK. (Halstead Press, New York), 1986

Bojar, O. English-to-Czech factored machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, ages 232–239, Prague, Czech Republic. Association for Computational Linguistics. 2007

Anand Kumar,Morphology Based Prototype Statistical Machine Translation System For English To Tamil Language, Phd Thesis, Amrita School Of Engineering, 2013

Adugna Sisay and Andreas Eisele, English – Oromo Machine Translation: An Experiment Using Statistical Approach, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 2010

Melby, Alan K. *The Possibility of Language: A discussion of the Nature of Language, with implications for Human and Machine Translation.* Amsterdam: John Benjamins Pub. Co, 1995

Austermuhl, Frank, *Electronic Tools for Translators*, Manchester: St. Jerome Pub. Co Baker, Mona 1998 Reexplorer la langue de la traduction: une approche par corpus', 2001

Lopez, A. Statistical machine translation. ACM Comput. Surv., 40, 3, Article 8 , 2008

DeRose, Steven J. "Grammatical category disambiguation by statistical optimization." Computational Linguistics 14(1): 31–39, 1988

Laurie Bauer , The Linguistics Student’s Handbook,Edinburgh, 2007

Daniel Teklu. “Zebenawi sewasw quanqua Tigrigna”, Mekelle: Mega printing, 2008

Gregory T.Stump (2001) Inflectional and Derivational morphology, Kentucky University,USA: Cambridge University Press.

Kassa G., Daniel G. Siwasw Tigrigna.Addis Ababa:Mega printing enterprise,2004.

Amanuel Sahle . sewasew Tigrigna bsefiḥu. Lawrencevill, NJ, USA: Red Sea Press, 2008

Chang, J. and Su, K. Corpus-based statistics-oriented (CBSO) machine translation researches in Taiwan. AMTA, 1997

Hovy, E. Deepening wisdom or compromised principles?, The hybridization of statistical and symbolic MT systems, 1996

ANNEX I: Geez Alphabets

ሀ	HA	ሁ	HU	ሂ	HI	ሃ	HA	ሄ	HE	ህ	H	ሆ	HO
ለ	LE	ሉ	LU	ሊ	LI	ላ	LA	ሌ	LE	ል	L	ሎ	LO
ሐ	HA	ሑ	HU	ሒ	HI	ሓ	HA	ሔ	HE	ሕ	H	ሖ	HO
መ	ME	ሙ	MU	ሚ	MI	ማ	MA	ሜ	ME	ም	M	ሞ	MO
ሠ	SE	ሡ	SU	ሢ	SI	ሣ	SA	ሤ	SE	ሥ	S	ሦ	SO
ረ	RE	ሩ	RU	ሪ	RI	ራ	RA	ሬ	RE	ር	R	ሮ	RO
ሰ	SE	ሱ	SU	ሲ	SI	ሳ	SA	ሴ	SE	ስ	S	ሶ	SO
ሸ	SHE	ሹ	SHU	ሺ	SHI	ሻ	SHA	ሼ	SHE	ሽ	SH	ሾ	SHO
ቀ	KE	ቁ	KU	ቂ	KI	ቃ	KA	ቄ	KE	ቅ	K	ቆ	KO
በ	BE	ቡ	BU	ቢ	BI	ባ	BA	ቤ	BE	ብ	B	ቦ	BO
ተ	TE	ቱ	TU	ቲ	TI	ታ	TA	ቴ	TE	ት	T	ቶ	TO
ቸ	CHE	ቹ	CHU	ቺ	CHI	ቻ	CHA	ቼ	CHE	ች	CH	ቾ	CHO
ኀ	HA	ኁ	HU	ኂ	HI	ኃ	HA	ኄ	HE	ኅ	H	ኆ	HO
ነ	NE	ኑ	NU	ኒ	NI	ና	NA	ኔ	NE	ን	N	ኖ	NO
ኘ	GNE	ኙ	GNU	ኚ	GNI	ኛ	GNA	ኜ	GNE	ኝ	GN	ኞ	GNO
አ	A	ኡ	U	ኢ	I	ኣ	A	ኤ	E	እ	I	ኦ	O
ከ	KE	ኩ	KU	ኪ	KI	ካ	KA	ኬ	KE	ክ	K	ኮ	KO
ኸ	HE	ኹ	HU	ኺ	HI	ኻ	HA	ኼ	HE	ኽ	H	ኾ	HO
ወ	WE	ዉ	WU	ዊ	WI	ዋ	WA	ዌ	WE	ው	W	ዎ	WO
ዐ	A	ዑ	U	ዒ	I	ዓ	A	ዔ	E	ዕ	I	ዖ	O
ዘ	ZE	ዙ	ZU	ዚ	ZI	ዛ	ZA	ዜ	ZE	ዝ	Z	ዞ	ZO
ዠ	ZHE	ዡ	ZHU	ዢ	ZHI	ዣ	ZHA	ዤ	ZHE	ዥ	ZH	ዦ	ZHO
የ	YE	ዩ	YU	ዪ	YI	ያ	YA	ዬ	YE	ይ	Y	ዮ	YO
ደ	DE	ዱ	DU	ዲ	DI	ዳ	DA	ዴ	DE	ድ	D	ዶ	DO
ጀ	JE	ጁ	JU	ጂ	JI	ጃ	JA	ጄ	GE	ጅ	J	ጆ	JO
ገ	GE	ጉ	GU	ጊ	GI	ጋ	GA	ጌ	TE	ግ	G	ጎ	GO
ጠ	TE	ጡ	TU	ጢ	TI	ጣ	TA	ጤ	CHE	ጥ	T	ጦ	TO
ጨ	CHE	ጩ	CHU	ጪ	CHI	ጫ	CHA	ጬ	PE	ጭ	CH	ጮ	CHO
ጰ	PE	ጱ	PU	ጲ	PI	ጳ	PA	ጴ	TSE	ጵ	P	ጶ	PO
ጸ	TSE	ጹ	TSU	ጺ	TSI	ጻ	TSA	ጼ	TSE	ጽ	TS	ጾ	TSO
ፀ	TSE	ፁ	TSU	ፂ	TSI	ፃ	TSA	ፄ	TSE	ፅ	TS	ፆ	TSO
ፈ	FE	ፉ	FU	ፊ	FI	ፋ	FA	ፌ	FE	ፍ	F	ፎ	FO
ፐ	PE	ፑ	PU	ፒ	PI	ፓ	PA	ፔ	PE	ፕ	P	ፖ	PO

APPENDIX II

Baseline English-Tigrigna Parallel Corpus Excerpt

At the first God made the heaven and the earth.	ኣምላኽ ብመጀመርታ ሰማይን ምድርን ፈጠረ ።
And the earth was waste and without form; and it was dark on the face of the deep: and the Spirit of God was moving on the face of the waters.	ምድሪ ድማ በረኻን ጥራያን ነበረት ፡ ጸልማት ከኣ ኣብ ልዕሊ መዓሙቕ ነበረ ። መንፈስ ኣምላኽ ድማ ኣብ ልዕሊ ማያት ይዝምቢ ነበረ ።
And God said, Let there be light: and there was light.	ኣምላኽ ከኣ ፥ ብርሃን ይኹን ፡ በለ ። ብርሃን ድማ ኾነ ።
And God, looking on the light, saw that it was good: and God made a division between the light and the dark,	ኣምላኽ ድማ እቲ ብርሃን ጽቡቕ ከም ዝኾነ ረኣየ። ኣምላኽ ከኣ ነቲ ብርሃን ካብ ጸልማት ፈለዮ።
Naming the light, Day, and the dark, Night. And there was evening and there was morning, the first day.	ኣምላኽ ነቲ ብርሃን መዓልቲ ኣውጽኣሉ። ነቲ ጸልማት ከኣ ለይቲ ኣውጽኣሉ። ምሸት ኰነ ብጊሓትውን ኰነ፡ ሓንቲ መዓልቲ።
And God said, Let there be a solid arch stretching over the waters, parting the waters from the waters.	ኣምላኽ ድማ፥ ንማያት ካብ ማያት ዚፈሊ ጠፈር ኣብ መንጎ ማያት ይኹን፡ በለ።
And God made the arch for a division between the waters which were under the arch and those which were over it: and it was so.	ኣምላኽ ነቲ ጠፈር ገበሮ። ነቲ ኣብ ትሕቲ ጠፈር ዘሎ ማያት ድማ ኻብቲ ኣብ ልዕሊ ጠፈር ዘሎ ማያት ፈለዮ፡ ከምኡውን ኰነ።
And God gave the arch the name of Heaven. And there was evening and there was morning, the second day.	ኣምላኽ ከኣ ነቲ ጠፈር ሰማይ ኣውጽኣሉ። ምሸት ኰነ ብጊሓትውን ኰነ፡ ካልኣይቲ መዓልቲ።
And God said, Let the waters under the heaven come together in one place, and let the dry land be seen: and it was so.	ኣምላኽ ድማ፥ እቲ ንቑጽ ምእንቲ ኺርኤስ፡ እቲ ኣብ ትሕቲ ሰማይ ዘሎ ማያት ናብ ሓንቲ ቦታ ይተኣከብ፡ በለ። ከምኡ ድማ ኾነ።
And God gave the dry land the name of Earth; and the waters together in their place were named Seas: and God saw that it was good.	ኣምላኽ ከኣ ነቲ ንቑጽ ምድሪ ኣውጽኣሉ፡ ነቲ እኩብ ማያት ድማ ባሕሪ ኣውጽኣሉ። ኣምላኽ ከኣ ጽቡቕ ከም ዝኾነ ረኣየ።
And God said, Let grass come up on the earth, and plants producing seed, and fruit-trees giving fruit, in which is their seed, after their sort: and it was so.	ኣምላኽ ድማ፥ እታ ምድሪ ሳዕርን ዘርኢ ዚህብ ብቘልን፡ ዘርኡ ኣብ ርእሱ ዘለዎ፡ ፍረ ከከም ዓይነቱ ኣብ ምድሪ ዚፈሪ ኦም ተውጽእ፡ በለ። ከምኡ ድማ ኾነ።
And grass came up on the earth, and every plant producing seed of its sort, and every tree producing fruit, in which is its seed, of its sort: and God saw that it was good.	እታ ምድሪ ሳዕርን ከከም ዓይነቱ ዘርኢ ዚህብ ብቘልን፡ ዘርኡ ኣብ ርእሱ ዘለዎ ፍረ ዚፈሪ ኣእዋም ከኣ ኣውጽኤት። ኣምላኽ ድማ ጽቡቕ ከም ዝኾነ ረኣየ።

And there was evening and there was morning, the third day.	ምሸት ኮነ ብጊሓትውን ኮነ፡ ሳልሰይቲ መዓልቲ።
And God said, Let there be lights in the arch of heaven, for a division between the day and the night, and let them be for signs, and for marking the changes of the year, and for days and for years:	ኣምላኽ ድማ፤ ንመዓልቲ ኻብ ለይቲ ዚፈልዩ ብርሃናት ኣብ ጠፈር ሰማይ ይኹኑ፡ ንዘበናትን ንመዓልትታትን ዓመታትን ከኣ ንመፈለጥታ ይኹኑ።
And let them be for lights in the arch of heaven to give light on the earth: and it was so.	ኣብ ምድሪ ንምብራህ ኣብ ጠፈር ሰማይ ብርሃናት ይኹን፡ በለ። ከምኡ ድማ ኾነ።
And God made the two great lights: the greater light to be the ruler of the day, and the smaller light to be the ruler of the night: and he made the stars.	ኣምላኽ ከኣ ኽልተ ዓበይቲ ብርሃናት ገበረ፥ እቲ ዓብዩ ብርሃን ብመዓልቲ ኺስልጥን፡ እቲ ንእሽቶ ብርሃን ድማ ብለይቲ ኺስልጥን፡ ከዋኽብቲውን ገበረ።
And there was evening and there was morning, the fourth day.	ምሸት ኮነ ብጊሓትውን ኮነ፡ ራብዐይቲ መዓልቲ።
And God said, Let the waters be full of living things, and let birds be in flight over the earth under the arch of heaven.	ኣምላኽ ድማ፥ ማያት ህያው ነፍሲ ዘለዎ ውንጅርጅር ዚብል እንስሳ የውጽእ፡ ኣብ ልዕሊ ምድሪ ኣብ ትሕቲ ጠፈር ሰማይ ከኣ ኣዕዋፍ ይንፈራ፡ በለ።
And God made great sea-beasts, and every sort of living and moving thing with which the waters were full, and every sort of winged bird: and God saw that it was good.	ኣምላኽ ድማ ነቶም ዓበይቲ እንስሳ ባሕርን ነቲ ማያት ዘውጽኦ በብዓይነቱ ህያው ነፍሲ ዘለዎ ውንጅርጅር ዚብል ኵሉን፡ ክንፊ ዘለወን በብዓይነተን ኵለን ኣዕዋፍን ፈጠረ። ኣምላኽ ከኣ ጽቡቕ ከም ዝኾነ ረኣየ።
And God gave them his blessing, saying, Be fertile and have increase, making all the waters of the seas full, and let the birds be increased in the earth.	ኣምላኽ ድማ፥ ፍረዩን ተባዝሑን ንማያት ባሕሪ ምልእዎ፡ ኣዕዋፍ ከኣ ኣብ ምድሪ ይብዝሓ፡ ኢሉ ባረኾም።
And there was evening and there was morning, the fifth day.	ምሸት ኮነ ብጊሓትውን ኮነ፡ ሓምሰይቲ መዓልቲ።
And God said, Let the earth give birth to all sorts of living things, cattle and all things moving on the earth, and beasts of the earth after their sort: and it was so.	ኣምላኽ ድማ፥ ምድሪ ህያው ነፍሲ ዘለዎ በብዓይነቱ እንስሳን ለመምታን ኣራዊት ምድርን በብዓይነቱ ተውጽእ፡ በለ። ከምኡውን ኮነ።
And God made the beast of the earth after its sort, and the cattle after their sort, and everything moving on the face of the earth after its sort: and God saw that it was good.	ኣምላኽ ከኣ ኣራዊት ምድሪ በብዓይነቱን እንስሳ በብዓይነቱን ኵሉ ለመምታ ምድሪ ኸኣ በብዓይነቱ ገበረ። ኣምላኽ ድማ ጽቡቕ ከም ዝኾነ ረኣየ።
And God said, Let us make man in our image, like us: and let him have rule over the fish of the sea and over the birds of the air and over the cattle and over	ኣምላኽ ከኣ፥ ብመልክዕና ኸም ምስልና ሰብ ንግበር፡ ንዓሳ ባሕርን ነዕዋፍ ሰማይን ንእንስሳን ንብዘላ ምድርን ኣብ ምድሪ ለመም ንዚብል ኵሉ ለመምታን ይግዝኡ፡ በለ።

all the earth and over every living thing which goes flat on the earth.

And God made man in his image, in the image of God he made him: male and female he made them.

And God gave them his blessing and said to them, Be fertile and have increase, and make the earth full and be masters of it; be rulers over the fish of the sea and over the birds of the air and over every living thing moving on the earth.

And to every beast of the earth and to every bird of the air and every living thing moving on the face of the earth I have given every green plant for food: and it was so.

And God saw everything which he had made and it was very good. And there was evening and there was morning, the sixth day.

And the heaven and the earth and all things in them were complete.

And on the seventh day God came to the end of all his work; and on the seventh day he took his rest from all the work which he had done.

And God gave his blessing to the seventh day and made it holy: because on that day he took his rest from all the work which he had made and done.

These are the generations of the heaven and the earth when they were made.

In the day when the Lord God made earth and heaven there were no plants of the field on the earth, and no grass had come up: for the Lord God had not sent rain on the earth and there was no man to do work on the land.

But a mist went up from the earth, watering all the face of the land.

And the Lord God made man from the dust of the earth, breathing into him the breath of life: and man became a living soul.

ኣምላኽ ድማ ብመልክዑ ሰብ ፈጠረ። ብመልክዕ ኣምላኽ ፈጠሮ። ተባዕታይን ኣንስተይትን ገይሩ ፈጠሮም።

ኣምላኽ ከኣ ባረኾም። ኣምላኽ ድማ፡ ፍረዩን ተባዝሑን ንምድሪ ኽኣ ምልእዋን ምለኽዋን፡ ንዓሳ ባሕርን ነዕዋፍ ስማይን ኣብ ምድሪ ለመም ንዝብል ኩሉ እንስሳን ከኣ ግዝኡ፡ በሎም።

ንኹሉ ኣራዊት ምድርን ንኹለን ኣዕዋፍ ሰማይን ህያው ነፍሲ ንዘለዎ ኣብ ምድሪ ለመም ንዚብል ኵሉ ኽኣ ኵሉ ለምለም ሳዕሪ ንምግቦም ሂበዮም አሎኹ፡ በለ። ከምኡ ድማ ኾነ።

ኣምላኽ ከኣ ዝገበሮ ዘበለ ዅሉ ረኣየ፡ እንሆ፡ ብዙሕ ጽቡቕ ኰነ። ምሸት ኰነ ብጊሓትውን ኰነ፡ ሳድሰይቲ መዓልቲ።

ከምኡ ሰማይን ምድርን ኵሉ ሰራዊቶምን ተፈጸሙ።

ኣምላኽ ከኣ ነቲ ዝገበሮ ግብሩ በታ ሳብዐይቲ መዓልቲ ፈጸሞ፡ ብሳብዓይቲ መዓልቲ ድማ ኻብቲ ዝገበሮ ዅሉ ግብሩ ዐረፈ።

ኣምላኽ ከኣ ኻብቲ ዝፈጠሮን ዝገበሮን ኵሉ ግብሩ ብእኣ ስለ ዝዐረፈ፡ ነታ ሳብዐይቲ መዓልቲ ባረኻን ቀደሳን።

በታ እግዚኣብሄር ኣምላኽ ምድርን ሰማይን ዝፈጠረላ መዓልቲ ምስ ተፈጥሩ፡ ወለዶ ሰማይን ምድርን እዚ እዩ።

እግዚኣብሄር ኣምላኽ ኣብ ምድሪ ገና ኣየዝነመን ነበረ፡ ንምድሪ ዚዐዩ ሰብውን ኣይነበረን እሞ፡ ገና ገለ ኦም መሮር ኣይነበረን፡ ሳዕሪ መሮር ከኣ ሓንቲኳ ኣይበቘለትን ነበረት።

ግናኸ ንዅሉ ዝባን ምድሪ ዜስቲ ግመ ኻብ ምድሪ ይወጽእ ነበረ።

እግዚኣብሄር ኣምላኽ ከኣ ንሰብ ካብ ሓመድ ምድሪ ገበሮ፡ ኣብ ኣፍንጫኡ ድማ ትንፋስ ህይወት ኡፍ በለሉ እሞ እቲ ሰብ ህያው ነፍሲ ኾነ።

And the Lord God made a garden in the east, in Eden; and there he put the man whom he had made.

And out of the earth the Lord made every tree to come, delighting the eye and good for food; and in the middle of the garden, the tree of life and the tree of the knowledge of good and evil.

And a river went out of Eden giving water to the garden; and from there it was parted and became four streams.

እግዚኣብሄር ኣምላኽ ከኣ ኣብ ኤድን ብሸነኽ ምብራቕ ገነት ተኸለ። ነቲ ዝገበሮ ሰብ ድማ ኣብኣ ኣንበሮ።

እግዚኣብሄር ኣምላኽ ከኣ ምርኣዩ ዜብህግ፡ ምብላዑ ዝጥዑም ኵሉ ኦም ኣብ ምድሪ ኣብቈለ፡ ኣብ ማእከል ገነት ከኣ ኦም ህይወት፡ እታ ጽቡቕን ክፉእን እተፍልጥ ኦም ድማ።

ንገነት ዜስቲ ርባ ከኣ ካብ ኤድን ይውሕዝ ነበረ። ካብኡ ድማ ተፈላልዩ ኣርባዕተ ጨንፈር ኰነ።

APPENDIX III

Factored English-Tigrigna Parallel Corpus Excerpt

At|At|PRP the|the|AT0 first|first|ORD God|God|NP0 made|make|VVD the|the|AT0 heaven|heaven|NN1 and|and|CJC the|the|AT0 earth|earth|NN1 .|.|PUN

And|And|CJC the|the|AT0 earth|earth|NN1 was|was|VBD waste|waste|AJ0 and|and|CJC without|without|PRP form|form|NN1 ;|;|PUN and|and|CJC it|it|PNP was|was|VBD dark|dark|AJ0 on|on|PRP the|the|AT0 face|face|NN1 of|of|PRF the|the|AT0 deep|deep|AJ0 :|:|PUN and|and|CJC the|the|AT0 Spirit|Spirit|NN1 of|of|PRF God|God|NP0 was|was|VBD moving|move|VVG on|on|PRP the|the|AT0 face|face|NN1 of|of|PRF the|the|AT0 waters|water|NN1 .|.|PUN

And|And|CJC God|God|NP0 said|said|VVD ,|,|PUN Let|Let|VVB there|there|EX0 be|be|VBI light|light|AJ0 :|:|PUN and|and|CJC there|there|EX0 was|was|VBD light|light|AJ0 .|.|PUN

And|And|CJC God|God|NP0 ,|,|PUN looking|look|VVG on|on|PRP the|the|AT0 light|light|NN1 ,|,|PUN saw|saw|VVD that|that|CJT it|it|PNP was|was|VBD good|good|AJ0 :|:|PUN and|and|CJC God|God|NP0 made|made|VVD a|a|AT0 division|division|NN1 between|between|PRP the|the|AT0 light|light|NN1 and|and|CJC the|the|AT0 dark|dark|NN1 ,|,|PUN

Naming|Name|VVG the|the|AT0 light|light|NN1 ,|,|PUN Day|Day|NP0 ,|,|PUN and|and|CJC the|the|AT0 dark|dark|NN1 ,|,|PUN Night|Night|NN1 .|.|PUNC And|And|CJC there|there|EX0 was|was|VBD evening|evening|NN1 and|and|CJC there|there|EX0 was|was|VBD morning|morning|NN1 ,|,|PUN the|the|AT0 first|first|ORD day|day|NN1 .|.|PUN

And|And|CJC God|God|NP0 said|say|VVD ,|,|PUN Let|Let|VVB there|there|EX0 be|be|VBI a|a|AT0 solid|solid|NN1 arch|arch|NN1 stretching|stretch|NN1 over|over|AJ0 the|the|AT0

ኣምላኽ|ኣምላኽ|NP0 ብመጀመርታ|መጀመርታ|ORD ሰማይን|ሰማይ|NN1 ምድሪን|ምድሪ|NN1 ፈጠረ|ፈጠረ|VVD ።|።|PUNC

ምድሪ|ምድሪ|NN1 ድማ|ድማ|CJC በረኻን|በረኻ|NN1 ጥራያን|ጥራይ|VVD ነበረት|ነበረ|VBD ፡|፡|PUNCPUNC ጸልማት|ጸልማት|AJ0 ከኣ|ከኣ|CJC ኣብ|ኣብ|CJC ልዕሊ|ልዕሊ|PRP መዓሙቕ|መዓሙቕ|AJ0 ነበረ|ነበረ|VBD ።|።|PUNC መንፈስ|መንፈስ|NN1 ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ኣብ|ኣብ|CJC ልዕሊ|ልዕሊ|PRP ማያት|ማይ|NN2 ይዝምቢ|ይዝምቢ|VVG ነበረ|ነበረ|VBD ።|።|PUNCPUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ፥|፥|PUNC ብርሃን|ብርሃን|AJ0 ይኹን|ኹነ|VBI ፡|፡|PUNCPUNC በለ|በለ|VVD ።|።|PUNC ብርሃን|ብርሃን|AJ0 ድማ|ድማ|CJC ኾነ|ኾነ|VVD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC እቲ|እቲ|AT0 ብርሃን|ብርሃን|AJ0 ጽቡቕ|ጽቡቕ|AJ0 ከም|ከም|PNP ዝኾነ|ኾነ|VBI ረኣየ|ረኣየ|VVG ።|።|PUNC ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ነቲ|ነቲ|AT0 ብርሃን|ብርሃን|AJ0 ካብ|ካብ|PRP ጸልማት|ጸልማት|AJ0 ፈለዮ|ፈለየ|NN1 ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ነቲ|ነቲ|AT0 ብርሃን|ብርሃን|AJ0 መዓልቲ|መዓልቲ|NP0 ኣውጽኣሉ|ኣውጸዓ|VVG ።|።|PUNC ነቲ|ነቲ|AT0 ጸልማት|ጸልማት|AJ0 ከኣ|ከኣ|CJC ለይቲ|ለይቲ|NN1 ኣውጽኣሉ|ኣውጸዓ|VVG ።|።|PUNC ምሸት|ምሸት|NN1 ኰነ|ኰነ|VBD ብጊሓትውን|ብጊሓት|NN1 ኰነ|ኰነ|VBD ፡|፡|PUNC ሓንቲ|ሓንቲ|ORD መዓልቲ|መዓልቲ|NP0 ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፥|፥|PUNC ንማያት|ማይ|NN2 ካብ|ካብ|PRP ማያት|ማይ|NN2 ዚፈሊ|ፈለየ|NN1 ጠፈር|ጠፈር|NN1 ኣብ|ኣብ|CJC መንጎ|መንጎ|PRP ማያት|ማይ|NN2 ይኹን|ኹነ|VBI ፡|፡ |PUNC በለ|በለ|VVD ።|።|PUNC waters|water|NN2 ,|,|PUNC parting|part|AJ0 the|the|AT0 waters|water|NN2 from|from|PRP the|the|AT0 waters|water|NN2 .|.|PUNC

And|And|CJC God|God|NP0 made|make|VVD the|the|AT0 arch|arch|NN1 for|for|PRP a|a|AT0 division|division|NN1 between|between|PRP the|the|AT0 waters|water|NN2 which|which|DTQ were|were|VBD under|under|PRP the|the|AT0 arch|arch|NN1 and|and|CJC those|those|DT0 which|which|DTQ were|were|VBD over|over|PRP it|it|PNP :|:|PUN and|and|CJC it|it|PNP was|was|VBD so|so|AV0 .|.|PUN

And|And|CJC God|God|NP0 gave|give|VVD the|the|AT0 arch|arch|NN1 the|the|AT0 name|name|NN1 of|of|PRF Heaven|Heaven|NN1 .|.|PUNC And|And|CJC there|there|EX0 was|was|VBD evening|evening|NN1 and|and|CJC there|there|EX0 was|was|VBD morning|morning|NN1 ,|,|PUN the|the|AT0 second|second|ORD day|day|NN1 .|.|PUN

And|And|CJC God|God|NP0 said|say|VVD ,|,|PUN Let|Let|VVB the|the|AT0 waters|water|NN2 under|under|PRP the|the|AT0 heaven|heaven|NN1 come|come|VVB together|together|AV0 in|in|PRP one|one|CRD place|place|NN1 ,|,|PUN and|and|CJC let|let|VVB the|the|AT0 dry|dry|AJ0 land|land|NN1 be|be|VBI seen|see|VVN :|:|PUN and|and|CJC it|it|PNP was|was|VBD so|so|AV0 .|.|PUN

And|And|CJC God|God|NP0 gave|give|VVD the|the|AT0 dry|dry|AJ0 land|land|NN1 the|the|AT0 name|name|NN1 of|of|PRF Earth|Earth|NN1 ;|;|PUN and|and|CJC the|the|AT0 waters|water|NN2 together|together|AV0 in|in|PRP their|their|DPS place|place|NN1 were|were|VBD named|name|VVN Seas|Sea|NN2 :|:|PUN and|and|CJC God|God|NP0 saw|saw|VVD that|that|CJT it|it|PNP was|was|VBD good|so|AJ0 .|.|PUN

ኣምላኽ|ኣምላኽ|NP0 ነቲ|ነቲ|AT0 ጠፈር|ጠፈር|NN1 ገበሮ|ገበረ|VVD ።|።|PUNC ነቲ|ነቲ|AT0 ኣብ|ኣብ|CJC ትሕቲ|ትሕቲ|PRP ጠፈር|ጠፈር|NN1 ዘሎ|ዘሎ|VBD ማያት|ማይ|NN2 ድማ|ድማ|CJC ኻብቲ|ኻብቲ|PRP ኣብ|ኣብ|CJC ልዕሊ|ልዕሊ|PRP ጠፈር|ጠፈር|NN1 ዘሎ|ዘሎ|VBD ማያት|ማይ|NN2 ፈለዮ|ፈለየ|NN1 ፡|፡|PUNC ከምኡውን|ከምኡ|AV0 ኰነ|ኰነ|VBD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ነቲ|ነቲ|AT0 ጠፈር|ጠፈር|NN1 ሰማይ|ሰማይ|NN1 ኣውጽኣሉ|ኣውጻዓ|VVG ።|።|PUNC ምሸት|ምሸት|NN1 ኰነ|ኰነ|VBD ብጊሓትውን|ብጊሓት|NN1 ኰነ|ኰነ|VBD ፡|፡|PUNC ካልኣይቲ|ካልኣይቲ|ORD መዓልቲ|መዓልቲ|NP0 ።|። |PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፥|፥|PUNC እቲ|እቲ|AT0 ንቑጽ|ንቑጽ|AJ0 ምእንቲ|ምእንቲ|AV0 ኺርኤስ|ኺርኤስ|VVN ፡|፡|PUNC እቲ|እቲ|AT0 ኣብ|ኣብ|CJC ትሕቲ|ትሕቲ|PRP ሰማይ|ሰማይ|NN1 ዘሎ|ዘሎ|VBD ማያት|ማይ|NN2 ናብ|ናብ|TO0 ሓንቲ|ሓንቲ|ORD ቦታ|ቦታ|NN1 ይተኣከብ|ተኣከበ|NN1 ፡|፡|PUNC በለ|በለ|VVD ።|። |PUNC ከምኡ|ከምኡ|AV0 ድማ|ድማ|CJC ኾነ|ኾነ|VBD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ነቲ|ነቲ|AT0 ንቑጽ|ንቑጽ|AJ0 ምድሪ|ምድሪ|NN1 ኣውጽኣሉ|ኣውጻዓ|VVG ፡|፡|PUNC ነቲ|ነቲ|AT0 እኩብ|እኩብ|NN1 ማያት|ማይ|NN2 ድማ|ድማ|CJC ባሕሪ|ባሕሪ|NN1 ኣውጽኣሉ|ኣውጻዓ|VVG ።|።|PUNC ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ጽቡቕ|ጽቡቕ|AJ0 ከም|ከም|PNP ዝኾነ|ኾነ|VBI ረኣየ|ረኣየ|VVD ።|። |PUNC

And|And|CJC God|God|NP0 said|say|VVD ,|,|PUN Let|Let|VVB grass|grass|NN1 come|come|VVI up|up|AVP on|on|PRP the|the|AT0 earth|earth|NN1 ,|,|PUN and|and|CJC plants|plant|NN2 producing|produce|VVG seed|seed|NN1 ,|,|PUN and|and|CJC fruit-trees|fruit-tree|NN2 giving|give|VVG fruit|fruit|NN0 ,|,|PUN in|in|PRP which|which|DTQ is|which|VBZ their|their|DPS seed|seed|NN1 ,|,|PUN after|after|CJS their|their|DPS sort|sort|NN1 :|:|PUN and|:|CJC it|:|PNP was|was|VBD so|so|AV0 .|.|PUN

And|And|CJC grass|grass|NN1 came|came|VVD up|up|AVP on|on|PRP the|the|AT0 earth|earth|NN1 ,|,|PUN and|and|CJC every|every|AT0 plant|plant|NN1 producing|produce|VVG seed|seed|NN1 of|of|PRF its|its|DPS sort|sort|NN1 ,|,|PUN and|and|CJC every|every|AT0 tree|tree|NN1 producing|produce|VVG fruit|fruit|NN0 ,|,|PUN in|in|PRP which|which|DTQ is|is|VBZ its|its|DPS seed|seed|NN1 ,|,|PUN of|of|PRF its|its|DPS sort|sort|NN1 :|:|PUN and|and|CJC God|God|NP0 saw|saw|VVD that|that|CJT it|it|PNP was|was|VBD good|so|AJ0

And|And|CJC there|there|EX0 was|was|VBD evening|evening|NN1 and|and|CJC there|there|EX0 was|was|VBD morning|morning|NN1 ,|,|PUN the|,|AT0 third|third|ORD day|day|NN1 .|.|PUN

And|And|CJC God|God|NP0 said|say|VVD ,|,|PUN Let|Let|VVB there|there|EX0 be|be|VBI lights|light|NN2 in|in|PRP the|the|AT0 arch|arch|NN1 of|of|PRF heaven|heaven|NN1 ,|,|PUN for|for|PRP a|a|AT0 division|division|NN1 between|between|PRP the|the|AT0 day|day|NN1 and|and|CJC the|the|AT0 night|night|NN1 ,|,|PUN and|and|CJC let|let|VVB them|them|PNP be|be|VBI for|for|PRP signs|sign|NN2 ,|,|PUN and|and|CJC for|for|PRP marking|mark|VVG the|the|AT0 changes|change|NN2 of|of|PRF the|the|AT0 year|year|NN1 ,|,|PUN and|and|CJC for|for|PRP days|day|NN2 and|and|CJC for|for|PRP years|year|NN2 :|:|PUN

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፤|፤|PUNC እታ|እታ|AT0 ምድሪ|ምድሪ|NN1 ሳዕርን|ሳዕሪ|NN1 ዘርኢ|ዘርኢ|NN1 ዚህብ|ሀበ|VVG ብቝልን|ብቝል|NN1 ፡|፡|PUNC ዘርኡ|ዘርኢ|NN1 ኣብ|ኣብ|CJC ርእሱ|ርእሲ|NN2 ዘለዎ|ዘለዎ|VHB ፡|፡|PUNC ፍረ|ፍረ|NN1 ከከም|ከከም|AJ0 ዓይነቱ|ዓይነት|NN1 ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ዚፈሪ|ፈረየ|VVG ኦም|ኦም|NN1 ተውጽእ|ኣውጻዓ|VVG ፡|፡|PUNC በለ|በለ|VVD ።|።|PUNC ከምኡ|ከምኡ|AV0 ድማ|ድማ|CJC ኾነ|ኾነ|VBD ።|።|PUNC

እታ|እታ|AT0 ምድሪ|ምድሪ|NN1 ሳዕርን|ሳዕሪ|NN1 ከከም|ከከም|AJ0 ዓይነቱ|ዓይነት|NN1 ዘርኢ|ዘርኢ|NN1 ዚህብ|ሀበ|VVG ብቝልን|ብቝል|NN1 ፡|፡|PUNC ዘርኡ|ዘርኡ|NN1 ኣብ|ኣብ|CJC ርእሱ|ርእሲ|NN2 ዘለዎ|ዘለዎ|VHB ፍረ|ፍረ|NN1 ዚፈሪ|ፈረየ|VVG ኣእዋም|ኦም|NN1 ከኣ|ከኣ|CJC ኣውጽኤት|ኣውጻዓ|VVG ።|።|PUNC ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ጽቡቕ|ጽቡቕ|AJ0 ከም|ከም|PNP ዝኾነ|ኾነ|VBD ረኣየ|ረኣየ|VVD ።|። |PUNC

ምሸት|ምሸት|NN1 ኰነ|ኰነ|VBD ብጊሓትውን|ብጊሓት|NN1 ኰነ|ኰነ|VBD ፡|፡|PUNC ሳልሰይቲ|ሳልሳይ|ORD መዓልቲ|መዓልቲ|NP0 ።|። |PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፦|፦|PUNC ንመዓልቲ|መዓልቲ|NP0 ኻብ|ኻብ|PRP ለይቲ|ለይቲ|NN1 ዚፈልዩ|ፈለየ|NN1 ብርሃናት|ብርሃን|AJ0 ኣብ|ኣብ|CJC ጠፈር|ጠፈር|NN1 ሰማይ|ሰማይ|NN1 ይኹኑ|ኾነ|VBI ፡|፡|PUNC ንዘበናትን|ዘበን|NN1 መዓልቲ|መዓልቲ|NP0 ዓመታትን|ዓመት|NN1 ከኣ|ከኣ|CJC ንመፈለጥታ|ፈለጠ|VBD ይኹኑ|ኾነ|VBI ።|።|PUNC

And|And|CJC let|let|VVB them|them|PNP be|be|VBI for|for|PRP lights|light|NN2 in|in|PRP the|the|AT0 arch|arch|NN1 of|of|PRF heaven|heaven|NN1 to|to|TO0 give|give|VVI light|light|NN1 on|on|PRP the|on|AT0 earth|earth|NN1 :|:|PUN and|and|CJC it|it|PNP was|was|VBD so|so|AV0 .|.|PUNC

And|And|CJC God|God|NP0 made|make|VVD the|the|AT0 two|two|CRD great|great|AJ0 lights|light|NN2 :|:|PUN the|the|AT0 greater|great|AJC light|light|NN1 to|to|TO0 be|be|VBI the|the|AT0 ruler|rule|NN1 of|of|PRF the|the|AT0 day|day|NN1 ,|,|PUN and|and|CJC the|the|AT0 smaller|small|AJC light|light|NN1 to|to|TO0 be|be|VBI the|the|AT0 ruler|rule|NN1 of|of|PRF the|the|AT0 night|night|NN1 :|:|PUN and|and|CJC he|he|PNP made|make|VVD the|the|AT0 stars|star|NN2 .|.|PUNC

And|And|CJC there|there|EX0 was|was|VBD evening|evening|NN1 and|and|CJC there|there|EX0 was|was|VBD morning|morning|NN1 ,|,|PUN the|the|AT0 fourth|fourth|ORD day|day|NN1 .|.|PUNC

And|And|CJC God|God|NP0 said|say|VVD ,|,|PUN Let|Let|VVB the|the|AT0 waters|water|NN2 be|be|VBI full|full|AJ0 of|of|PRF living|live|AJ0 things|thing|NN2 ,|,|PUN and|and|CJC let|let|VVB birds|bird|NN2 be|be|VBI in|in|PRP flight|flight|NN1 over|over|PRP the|the|AT0 earth|earth|NN1 under|under|PRP the|the|AT0 arch|arch|NN1 of|of|PRF heaven|heaven|NN1 .|.|PUNC

And|And|CJC God|God|NP0 made|make|VVD great|great|AJ0 sea-beasts|sea-beast|NN2 ,|,|PUN and|and|CJC every|every|AT0 sort|sort|NN1 of|of|PRF living|live|NN1 and|and|CJC moving|move|VVG thing|thing|NN1 with|with|PRP which|which|DTQ the|the|AT0 waters|water|NN2 were|were|VBD full|full|AJ0 ,|,|PUN and|and|CJC every|every|AT0 sort|sort|NN1 of|of|PRF winged|wing|AJ0 bird|bird|NN1 :|:|PUN

ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ንምብራህ|በረኸ| ኣብ|ኣብ|CJC ጠፈር|ጠፈር|NN1 ሰማይ|ሰማይ|NN1 ብርሃናት|ብርሃን|AJ0 ይኹን|ኾነ|VVD ፡|፡|PUNC በለ|በለ|VVD ።|።|PUNC ከምኡ|ከምኡ|AV0 ድማ|ድማ|CJC ኾነ|ኾነ|VVD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ኽልተኽልተ|ኽልተ|CRD ዓበይቲ|ዓበይቲ|AJ0 ብርሃናት|ብርሃን|NN2 ገበረ|ገበረ|VVD ፥|፥|PUNC እቲ|እቲ|AT0 ዓብዩ|ዓብዩ|AJC ብርሃን|ብርሃን|NN1 ብመዓልቲ|መዓልቲ|NP0 ኺስልጥን|ሰልጠነ|NN1 ፡|፡ |PUNC እቲ|እቲ|AT0 ንእሽቶ|ንእሽቶ|AJC ብርሃን|ብርሃን|AJ0 ድማ|ድማ|CJC ብለይቲ|ለይቲ|NN1 ኺስልጥን|ሰልጠነ|NN1 ፡|፡|PUNC ከዋኽብቲውን|ከዋኽብት| ገበረ|ገበረ|VVD ።|።|PUNC

ምሸት|ምሸት|NN1 ኰነ|ኰነ|VBD ብጊሓትውን|ብጊሓት|NN1 ኰነ|ኰነ|VBD ፡|፡|PUNC ራብዐይቲ|ራብዓይ|ORD መዓልቲ|መዓልቲ|NP0 ።|። |PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፥|፥|PUNC ማያት|ማይ|NN2 ህያው|ህያው|AJ0 ነፍሲ|ነፍሲ|NN2 ዘለዎ|ዘለዎ|VHB እንስሳ|እንስሳ|NN2 የውጽእ|ኣውጻዓ|VVG ፡|፡|PUNC ኣብ|ኣብ|CJC ልዕሊ|ልዕሊ|PRP ምድሪ|ምድሪ|NN1 ኣብ|ኣብ|CJC ትሕቲ|ትሕቲ|PRP ጠፈር|ጠፈር|NN1 ሰማይ|ሰማይ|NN1 ከኣ|ከኣ|CJC ኣዕዋፍ|ኢፍ|NN2 ይንፈራ|ነፈረ|NN1 ፡|፡ |PUNC በለ|በለ|VVD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ነቶም|ነቶም|DT0 ዓበይቲ|ዓባይ|AJ0 እንስሳ|እንስሳ|NN2 ባሕርን|ባሕሪ|NN2 ነቲ|ነቲ|AT0 ማያት|ማይ|NN2 ዘውጽኦ|ኣውጻዓ|VVG በብዓይነቱ|ዓይነት|NN1 ህያው|ህያው|AJ0 ነፍሲ|ነፍሲ|NN1 ዘለዎ|ዘለዎ|VHB ውንጅርጅር|ውንጅርጅር|VBG ዚብል|ዚብል|AJ0 ኵሉን|ኵሉን|DT0 ፡|፡|PUNC ክንፊክንፊ|ክንፊ|AJ0 ዘለወን|ዘለወን|CJC በብዓይነተን|ዓይነት|NN1 ኵለን|ኲሉ|DT0 ኣዕዋፍን|ኢፍ|NN2 ፈጠረ|ፈጠረ|VVD ። |።|PUNC ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ጽቡቕ|ጽቡቕ|AJ0 ከም|ከም|PNP ዝኾነ|ኾነ|VBD ረአየ|ረአየ|VVD ።|።|PUNC and|and|CJC God|God|NP0 saw|saw|VVD that|that|CJT it|it|PNP was|was|VBD good|good|AJ0 .|.|PUNC

And|And|CJC God|God|NP0 gave|give|VVD them|them|PNP his|his|DPS blessing|bless|NN1 ,|,|PUN saying|say|VVG ,|,|PUN Be|Be|VBI fertile|fertile|AJ0 and|and|CJC have|have|VHB increase|increase|NN1 ,|,|PUN making|make|VVG all|all|DT0 the|the|AT0 waters|water|NN2 of|of|PRF the|the|AT0 seas|sea|NN2 full|full|AJ0 ,|,|PUN and|and|CJC let|let|VVB the|the|AT0 birds|bird|NN2 be|be|VBI increased|increase|VVN in|in|PRP the|the|AT0 earth|earth|NN1 .|.|PUNC

And|And|CJC there|there|EX0 was|was|VBD evening|evening|NN1 and|and|CJC there|there|EX0 was|was|VBD morning|morning|NN1 ,|,|PUN the|the|AT0 fifth|fifth|ORD day|day|NN1 .|.|PUNC

And|And|CJC God|God|NP0 said|say|VVD ,|,|PUN Let|Let|VVB the|the|AT0 earth|earth|NN1 give|give|VVI birth|birth|NN1 to|to|PRP all|all|DT0 sorts|sort|NN2 of|of|PRF living|live|AJ0 things|thing|NN2 ,|,|PUN cattle|cattle|NN2 and|and|CJC all|all|DT0 things|thing|NN2 moving|move|VVG on|on|PRP the|the|AT0 earth|earth|NN1 ,|,|PUN and|and|CJC beasts|beast|NN2 of|of|PRF the|the|AT0 earth|earth|NN1 after|after|PRP their|their|DPS sort|sort|NN1 :|:|PUN and|and|CJC it|it|PNP was|was|VBD so|so|AV0 .|.|PUNC

And|And|CJC God|God|NP0 made|make|VVD the|the|AT0 beast|beast|NN1 of|of|PRF the|the|AT0 earth|earth|NN1 after|after|PRP its|its|DPS sort|sort|NN1 ,|,|PUN and|and|CJC the|the|AT0 cattle|cattle|NN2 after|after|PRP their|their|DPS sort|sort|NN1 ,|,|PUN and|and|CJC everything|everything|PNI moving|move|VVG on|on|PRP the|the|AT0 face|face|NN1 of|of|PRF the|the|AT0 earth|earth|NN1 after|after|PRP its|its|DPS sort|sort|NN1 :|:|PUN and|and|CJC God|God|NP0 saw|see|VVD that|that|CJT it|it|PNP was|was|VBD good|good|AJ0 .|.|PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፥|፥|PUNC ፍረዩን|ፍረ|NN1 ተባዝሑን|ተባዝሑን|AJ0 ንማያት|ማይ|NN2 ባሕሪ|ባሕሪ|NN2 ምልእዎ|ምልእዎ|AJ0 ፡|፡|PUNC ኣዕዋፍ|ኢፍ|NN2 ከኣ|ከኣ|CJC ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ይብዝሓ|ይብዝሓ|NN2 ፡|፡|PUNC ኢሉ|ኢሉ|VVG ባረኾም|ባረኾም|NN1 ።|።|PUNC

ምሸት|ምሸት|NN1 ኰነ|ኰነ|VBD ብጊሓትውን|ብጊሓት|NN1 ኰነ|ኰነ|VBD ፡|፡|PUNC ሓምሰይቲ|ሓምሳይ|ORD መዓልቲ|መዓልቲ|NP0 ።|። |PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፥|፥|PUNC ምድሪ|ምድሪ|NN1 ህያው|ህያው|AJ0 ነፍሲ|ነፍሲ| ዘለዎ|ዘለዎ|VHB በብዓይነቱ|ዓይነት|NN1 እንስሳን|እንስሳ|NN2 ለመምታን|ለመም|VVG ኣራዊት|ኣራዊት|NN1 ምድርን|ምድሪ|NN1 በብዓይነቱ|ዓይነት|NN1 ተውጽእ|ኣውጻዓ|VVG ፡|፡ |PUNC በለ|በለ|VVD ።|።|PUNC ከምኡውን|ከምኡ|AV0 ኰነ|ኰነ|VBD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ኣራዊት|ኣራዊት|NN1 ምድሪ|ምድሪ|NN1 በብዓይነቱን|ዓይነት|NN1 እንስሳ|እንስሳ|NN2 በብዓይነቱን|ዓይነት|NN1 ኵሉ|ኵሉ|DT0 ለመምታ|ለመም|VVG ምድሪ|ምድሪ|NN1 ኸኣ|ኸኣ|NP0 በብዓይነቱ|ዓይነት|NN1 ገበረ|ገበረ|VVD ። |።|PUNC ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ጽቡቕ|ጽቡቕ |AJo ከም|ከም|PNP ዝኾነ|ኾነ|VVD ረኣየ|ረኣየ|VVD ።|።|PUNC

And|And|CJC God|God|NP0 said|say|VVD ,|,|PUN Let|Let|VVB us|us|PNP make|make|VVI man|man|NN1 in|in|PRP our|our|DPS image|image|NN1 ,|,|PUN like|like|PRP us|us|PNP :|:|PUN and|and|CJC let|let|VVB him|him|PNP have|have|VHI rule|rule|NN1 over|over|PRP the|the|AT0 fish|fish|NN0 of|of|PRF the|the|AT0 sea|sea|NN1 and|and|CJC over|over|PRP the|the|AT0 birds|bird|NN2 of|of|PRF the|the|AT0 air|air|NN1 and|and|CJC over|over|PRP the|the|AT0 cattle|cattle|NN2 and|and|CJC over|over|PRP all|all|DT0 the|the|AT0 earth|earth|NN1 and|and|CJC over|over|PRP every|every|AT0 living|live|AJ0 thing|thing|NN1 which|which|DTQ goes|go|VVZ flat|flat|AJ0 on|on|PRP the|the|AT0 earth|earth|NN1 .|.|PUNC

And|And|CJC God|God|NP0 made|made|VVD man|man|NN1 in|in|PRP his|his|DPS image|image|NN1 ,|,|PUN in|in|PRP the|the|AT0 image|image|NN1 of|of|PRF God|God|NP0 he|he|PNP made|make|VVD him|him|PNP :|:|PUN male|male|AJ0 and|and|CJC female|female|NN1 he|he|PNP made|make|VVD them|them|PNP .|.|PUNC

And|And|CJC God|God|NP0 gave|gave|VVD them|them|PNP his|his|DPS blessing|bless|NN1 and|and|CJC said|say|VVD to|to|PRP them|them|PNP ,|,|PUN Be|Be|VBI fertile|fertile|AJ0 and|and|CJC have|have|VHB increase|increase|NN1 ,|,|PUN and|and|CJC make|make|VVB the|the|AT0 earth|earth|NN1 full|full|AJ0 and|and|CJC be|be|VBI masters|master|NN2 of|of|PRF it|it|PNP ;|;|PUN be|be|VBI rulers|ruler|NN2 over|over|PRP the|the|AT0 fish|fish|NN0 of|of|PRF the|the|AT0 sea|sea|NN1 and|and|CJC over|over|PRP the|the|AT0 birds|bird|NN2 of|of|PRF the|the|AT0 air|air|NN1 and|and|CJC over|over|PRP every|every|AT0 living|live|AJ0 thing|thing|NN1 moving|move|VVG on|on|PRP the|the|AT0 earth|earth|NN1 .|.|PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ፥|፥|PUNC ብመልክዕና|መልክዕ|NN1 ኸም|ኸም|PRP ምስልና|ምስል|NN1 ሰብ|ሰብ|NN1 ንግበር|ገበረ|VVD ፡|፡|PUNC ንዓሳ|ዓሳ|NN0 ባሕርን|ባሕር|NN1 ነዕዋፍ|ኢፍ|NN1 ሰማይን|ሰማይ|NN1 ንእንስሳን|እንስሳ|NN2 ንብዘላ|ንብዘላ|NN1 ምድርን|ምድሪ|NN1 ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ለመም|ለመም|VVG ንዚብል|በለ|VVD ኵሉ|ኵሉ|DT0 ለመምታን|ለመም|VVG ይግዝኡ|ገዘዓ| ፡|፡|PUNC በለ|በለ|VVD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ብመልክዑ|መልክዕ|NN1 ሰብ|ሰብ|NN1 ፈጠረ|ፈጠረ|VVD ።|።|PUNC ብመልክዕ|መልክዕ|NN1 ኣምላኽ|ኣምላኽ|NP0 ፈጠሮ|ፈጠረ|VVD ።|።|PUNC ተባዕታይን|ተባዕታይ|AJ0 ኣንስተይትን|ኣንስተይ|AJ0 ገይሩ|ገይሩ|VVB ፈጠሮም|ፈጠሮም|VVD ።|።|PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ባረኾም|ባረኸ|VVD ።|።|PUNC ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ፡|፡|PUNC ፍረዩን|ፍረዩ|AJ0 ተባዝሑን|ተባዝሑ|NN1 ንምድሪ|ምድሪ|NN1 ኸኣ|ኸኣ|NP0 ምልእዋን|መለዓ|AJ0 ምለኽዋን|መለኽ|NN2 ፡|፡|PUNC ንዓሳ|ዓሳ|NN1 ባሕርን|ባሕሪ|NN2 ነዕዋፍ|ኢፍ|NN2 ሰማይን|ሰማይ|NN1 ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ለመም|ለመም|VVG ንዝብል|ዝብል| ኲሉ|ኲሉ|DT0 እንስሳን|እንስሳ|NN2 ከኣ|ከኣ|CJC ግዝኡ|ገዘዓ| ፡|፡|PUNC በሎም|በለ|VVD

And|And|CJC to|to|PRP every|every|AT0 beast|beast|NN1 of|of|PRF the|the|AT0 earth|earth|NN1 and|and|CJC to|to|PRP every|every|AT0 bird|bird|NN1 of|of|PRF the|the|AT0 air|air|NN1 and|and|CJC every|every|AT0 living|live|AJ0 thing|thing|NN1 moving|move|VVG on|on|PRP the|the|AT0 face|face|NN1 of|of|PRF the|the|AT0 earth|earth|NN1 I|I|PNP have|have|VHB given|give|VVN every|every|AT0 green|green|AJ0 plant|plant|NN1 for|for|PRP food|food|NN1 :|:|PUN and|and|CJC it|it|PNP was|was|VBD so|so|AV0 .|.|PUNC

And|And|CJC God|God|NP0 saw|see|VVD everything|everything|PNI which|which|DTQ he|he|PNP had|have|VHD made|make|VVN and|and|CJC it|it|PNP was|was|VBD very|very|AV0 good|good|AJ0 .|.|PUNC And|And|CJC there|there|EX0 was|was|VBD evening|evening|NN1 and|and|CJC there|there|EX0 was|was|VBD morning|morning|NN1 ,|,|PUN the|the|AT0 sixth|sixth|ORD day|day|NN1 .|.|PUNC

And|And|CJC the|the|AT0 heaven|heaven|NN1 and|and|CJC the|the|AT0 earth|earth|NN1 and|and|CJC all|all|DT0 things|thing|NN2 in|in|PRP them|them|PNP were|were|VBD complete|complete|AJ0 .|.|PUNC

And|And|CJC on|on|PRP the|the|AT0 seventh|seventh|ORD day|day|NN1 God|God|NP0 came|come|VVD to|to|PRP the|the|AT0 end|end|NN1 of|of|PRF all|all|DT0 his|his|DPS work|work|NN1 ;|;|PUN and|and|CJC on|on|PRP the|the|AT0 seventh|seventh|ORD day|day|NN1 he|he|PNP took|took|VVD his|his|DPS rest|rest|NN1 from|from|PRP all|all|DT0 the|the|AT0 work|work|NN1 which|which|DTQ he|he|PNP had|had|VHD done|do|VDN .|.|PUNC

ንኹሉ|ኹሉ|DT0 ኣራዊት|ኣራዊት|NN1 ምድርን|ምድሪ|NN1 ንኹለን|ኹሉ|DT0 ኣዕዋፍ|ኢፍ|NN2 ሰማይን|ሰማይ|NN1 ህያው|ህያው|AJ0 ነፍሲ|ነፍሲ|NN2 ንዘለዎ|ዘለዎ|VHB ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ለመም|ለመም|VVG ንዚብል|በለ|VVD ኵሉ|ኵሉ|DT0 ኸኣ|ኸኣ|NP0 ኵሉ|ኵሉ|DT0 ለምለም|ለምለም|AJ0 ሳዕሪ|ሳዕሪ|NN1 ንምግቦም|ምግቢ|NN1 ሂበዮም|ሃበ|VVN ኣሎኹ|ኣሎ|VVD ፡|፡|PUNC በለ|በለ|VVD ።|።|PUNC ከምኡ|ከምኡ|AV0 ድማ|ድማ|CJC ኾነ|ኾነ|VVD ።|። |PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ዝገበሮ|ገበረ|VVD ዘበለ|ዘበለ| ኹሉ|ኹሉ| ረኣየ|ረኣየ|VVD ፡|፡|PUNC እንሆ|እንሆ| ፡|፡|PUNC ብዙሕ|ብዙሕ| ጽቡቕ|ጽቡቕ|AJ0 ኰነ|ኰነ|VBD ።|ኰነ|VBD ምሸት|ምሸት|NN1 ኰነ|ኰነ|VBD ብጊሓትውን|ብጊሓት|NN1 ኰነ|ኰነ|VBD ፡ |፡|PUNC ሳድሰይቲ|ሳድሳይ| መዓልቲ|መዓልቲ|NP0 ።|። |PUNC

ከምኡ|ከምኡ|AV0 ሰማይን|ሰማይ|NN1 ምድርን|ምድሪ|NN1 ኵሉ|ኵሉ|DT0 ሰራዊቶምን|ሰራዊት|NN1 ተፈጸሙ|ተፈጸመ|AJ0 ።|። |PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ነቲ|ነቲ|AT0 ዝገበሮ|ገበረ|VVD ግብሩ|ግብሪ|NN1 በታ|በታ|PRP ሳብዐይቲ|ሳብዓይ|ORD መዓልቲ|መዓልቲ|NP0 ፈጸሞ|ፈጸሞ|NN1 ፡|፡|PUNC ብሳብዓይቲ|ሳብዓይ|ORD መዓልቲ|መዓልቲ|NP0 ድማ|ድማ|CJC ኻብቲ|ኻብቲ|PRP ዝገበሮ|ገበረ|VVD ኹሉ|ኹሉ|PRP ግብሩ|ግብሪ|NN1 ዐረፈ|ዐረፈ|NN1 ።|።|PUNC

And|And|CJC God|God|NP0 gave|give|VVD his|his|DPS blessing|bless|NN1 to|to|PRP the|the|AT0 seventh|seventh|ORD day|day|NN1 and|and|CJC made|make|VVD it|it|PNP holy|holy|AJ0 :|:|PUN because|because|CJS on|on|PRP that|that|DT0 day|day|NN1 he|he|PNP took|take|VVD his|his|DPS rest|rest|NN1 from|from|PRP all|all|DT0 the|the|AT0 work|work|NN1 which|which|DTQ he|he|PNP had|had|VHD made|make|VVN and|and|CJC done|do|VDN .|.|PUNC

These|These|DT0 are|are|VBB the|the|AT0 generations|generation|NN2 of|of|PRF the|the|AT0 heaven|heaven|NN1 and|and|CJC the|the|AT0 earth|earth|NN1 when|when|CJS they|they|PNP were|were|VBD made|made|VVN .|.|PUNC

In|In|PRP the|the|AT0 day|day|NN1 when|when|AVQ the|the|AT0 Lord|Lord|NP0 God|God|NP0 made|made|VVD earth|earth|NN1 and|and|CJC heaven|heaven|NN1 there|there|EX0 were|were|VBD no|no|AT0 plants|plant|NN2 of|of|PRF the|the|AT0 field|field|NN1 on|on|PRP the|the|AT0 earth|earth|NN1 ,|,|PUN and|and|CJC no|no|AT0 grass|grass|NN1 had|had|VHD come|come|VVN up|up|AVP :|:|PUN for|for|PRP the|the|AT0 Lord|Lord|NP0 God|God|NP0 had|had|VHD not|not|XX0 sent|send|VVN rain|rain|NN1 on|on|PRP the|the|AT0 earth|earth|NN1 and|and|CJC there|there|EX0 was|was|VBD no|no|AT0 man|man|NN1 to|to|TO0 do|do|VDI work|work|NN1 on|on|PRP the|the|AT0 land|land|NN1 .|.|PUNC

But|But|CJC a|a|AT0 mist|mist|NN1 went|go|VVD up|up|AVP from|from|PRP the|the|AT0 earth|earth|NN1 ,|,|PUN watering|water|VVG all|all|DT0 the|the|AT0 face|face|NN1 of|of|PRF the|the|AT0 land|land|NN1 .|.|PUNC

ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ኻብቲ|ኻብቲ|PRP ዝፈጠሮን|ፈጠረ|VVD ዝገበሮን|ገበረ|VVD ኵሉ|ኵሉ|DT0 ግብሩ|ግብሪ|NN1 ብእኣ|ብእኣ|DT0 ስለ|ስለ|CJC ዝዐረፈ|ዐረፈ|NN1 ፡|፡|PUNC ነታ|ነታ|PNP ሳብዐይቲ|ሳብዓይ|ORD መዓልቲ|መዓልቲ|NP0 ባረኻን|ባረኸ|VVD ቀደሳን|ቀደሰ|VVD ።|።|PUNC

በታ|በታ|DT0 እግዚኣብሄር|እግዚኣብሄር|NP0 ኣምላኽ|ኣምላኽ|NP0 ምድርን|ምድሪ|NN1 ሰማይን|ሰማይ|NN1 ዝፈጠረላ|ፈጠረ|VVD መዓልቲ|መዓልቲ|NP0 ምስ|ምስ|PRP ተፈጥሩ|ፈጠረ|VVD ፡|፡|PUNC ወለዶ|ወለዶ|NN2 ሰማይን|ሰማይ|NN1 ምድርን|ምድሪ|NN1 እዚ|እዚ|DT0 እዩ|እዩ|VBD ።|።|PUNC

እግዚኣብሄር|እግዚኣብሄር|NP0 ኣምላኽ|ኣምላኽ|NP0 ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ገና|ገና|AV0 ኣየዝነመን|ዘነመ|NN1 ነበረ|ነበረ|VBD ፡|፡|PUNC ንምድሪ|ምድሪ|NN1 ዚዐዪ|ዚዐዪ| ሰብውን|ሰብ|NN1 ኣይነበረን|ነበረ|VBD እሞ|እሞ| ፡|፡|PUNC ገና|ገና| ገለ|ገለ| ኦም|ኦም| መሮር|መሮር| ኣይነበረን|ነበረ|VBD ፡|፡ |PUNC ሳዕሪ|ሳዕሪ| መሮር|መሮር| ከኣ|ከኣ|CJC ሓንቲኳ|ሓንቲ|ORD ኣይበቘለትን|በቘለ| ነበረት|ነበረ|VBD ።|።|PUNC

ግናኸ|ግና|CJC ንኹሉ|ኹሉ|DT0 ዝባን|ዝባን|NN1 ምድሪ|ምድሪ|NN1 ዜስቲ|ዜስቲ|VVG ግመ|ግመ|NN1 ኻብ|ኻብ|PRP ምድሪ|ምድሪ|NN1 ይወጽእ|ወጸዐ|VVD ነበረ|ነበረ|VBD ።|።|PUNC

And|And|CJC the|the|AT0 Lord|Lord|NP0 God|God|NP0 made|make|VVD man|man|NN1 from|from|PRP the|the|AT0 dust|dust|NN1 of|of|PRF the|the|AT0 earth|earth|NN1 ,|,|PUN breathing|breath|VVG into|into|PRP him|him|PNP the|the|AT0 breath|breath|NN1 of|of|PRF life|life|NN1 :|:|PUN and|and|CJC man|man|NN1 became|become|VVD a|a|AT0 living|live|AJ0 soul|soul|NN1 .|.|PUNC

And|And|CJC the|the|AT0 Lord|Lord|NP0 God|God|NP0 made|make|VVD a|a|AT0 garden|garden|NN1 in|in|PRP the|the|AT0 east|east|NN1 ,|,|PUN in|in|PRP Eden|Eden|NP0 ;|;|PUN and|and|CJC there|there|AV0 he|he|PNP put|put|VVD the|the|AT0 man|man|NN1 whom|whom|PNQ he|he|PNP had|had|VHD made|make|VVN .|.|PUNC

And|And|CJC out|out|PRP of|of|PRP the|the|AT0 earth|earth|NN1 the|the|AT0 Lord|Lord|NP0 made|make|VVD every|every|AT0 tree|tree|NN1 to|to|TO0 come|come|VVI ,|,|PUN delighting|delight|VVG the|the|AT0 eye|eye|NN1 and|and|CJC good|good|AJ0 for|for|PRP food|food|NN1 ;|;|PUN and|and|CJC in|in|PRP the|the|AT0 middle|middle|NN1 of|of|PRF the|the|AT0 garden|garden|NN1 ,|,|PUN the|the|AT0 tree|tree|NN1 of|of|PRF life|life|NN1 and|and|CJC the|the|AT0 tree|tree|NN1 of|of|PRF the|the|AT0 knowledge|knowledge|NN1 of|of|PRF good|good|AJ0 and|and|CJC evil|evil|NN1 .|.|PUNC

And|And|CJC a|a|AT0 river|river|NN1 went|go|VVD out|out|PRP of|of|PRP Eden|Eden|NP0 giving|give|VVG water|water|NN1 to|to|PRP the|the|AT0 garden|garden|NN1 ;|;|PUN and|and|CJC from|from|PRP there|there|AV0 it|it|PNP was|was|VBD parted|part|VVN and|and|CJC became|become|VVN four|four|CRD streams|stream|NN2 .|.|PUNC

እግዚኣብሄር|እግዚኣብሄር|NP0 ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ንሰብ|ሰብ|NN1 ካብ|ካብ|PRP ሓመድ|ሓመድ|NN1 ምድሪ|ምድሪ|NN1 ገበሮ|ገበረ|VVD ፣|፣|PUNC ኣብ|ኣብ|CJC ኣፍንጫኡ|ኣፍንጫ|NN1 ድማ|ድማ|CJC ትንፋስ|ትንፋስ|NN1 ህይወት|ህይወት|NN1 ኡፍ|ኡፍ|VVG በለሉ|በለ|VVD እሞ|እሞ|AV0 እቲ|እቲ|AT0 ሰብ|ሰብ|NN1 ህያው|ህያው|AJ0 ነፍሲ|ነፍስ|NN1 ኾነ|ኾነ|VVD ።|።|PUNC

እግዚኣብሄር|እግዚኣብሄር|NP0 ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ኣብ|ኣብ|CJC ኤድን|ኤድን|NP0 ብምብራቕ|ምብራቕ|NN1 ገነት|ገነት|NN1 ተኸለ|ተኸለ|VVD ።|።|PUNC ነቲ|ነቲ|AT0 ዝገበሮ|ገበረ|VVD ሰብ|ሰብ|NN1 ድማ|ድማ|CJC ኣብኣ|ኣብኣ|PNP ኣንበሮ|ኣንበሮ|VVD ።|።|PUNC

እግዚኣብሄር|እግዚኣብሄር|NP0 ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ምርኣዩ|ረኣየ|VVD ዜብህግ|ዜብህግ|VVG ፣|፣|PUNC ምብላዑ|በለዐ|VVD ዝጥዑም|ጥዑም|AJ0 ኵሉ|ኵሉ|DT0 ኦም|ኦም|NN1 ኣብ|ኣብ|CJC ምድሪ|ምድሪ|NN1 ኣብቈለ|በቈለ|VVD ፣|፣|PUNC ኣብ|ኣብ|CJC ማእከል|ማእከል|NN1 ገነት|ገነት|NN1 ከኣ|ከኣ|CJC ኦም|ኦም|NN1 ህይወት|ህይወት|NN1 ፣|፣|PUNC እታ|እታ|DT0 ጽቡቕን|ጽቡቕ|AJ0 ክፉእን|ክፉእ|NN1 እተፍልጥ|ፈለጠ|NN1 ኦም|ኦም|NN1 ድማ|ድማ|CJC ።|።|PUNC

ንገነት |ገነት|NN1 ዜስቲ|ዜስቲ|VVG ርባ|ርባ|NN1 ከኣ|ከኣ|CJC ካብ|ካብ|PRP ኤድን|ኤድን|NN1 ይውሕዝ|ይውሕዝ|VVD ነበረ|ነበረ|VBD ።|።|PUNC ካብኡ|ካብኡ|PNP ድማ|ድማ|CJC ተፈላልዩ|ተፈላልዩ|VVN ኣርባዕተ|ኣርባዕተ|CRD ጨንፈር|ጨንፈር|NN2 ኰነ|ኰነ|VBD ።|።|PUNC

The|The|AT0 name|name|NN1 of|of|PRF the|the|AT0 first|first|ORD is|is|VBZ Pishon|Pishon|NP0 ,|,|PUN which|which|DTQ goes|go|VVZ round|round|AVP about|about|PRP all|all|DT0 the|the|AT0 land|land|NN1 of|of|PRF Havilah|Havilah|NP0 where|where|CJS there|there|EX0 is|is|VBZ gold|gold|NN1 .|.|PUNC

And|And|CJC the|the|AT0 gold|gold|NN1 of|of|PRF that|that|DT0 land|land|NN1 is|is|VBZ good|good|AJ0 :|:|PUN there|there|EX0 is|is|VBZ bdellium|bdellium|NN1 and|and|CJC the|the|AT0 onyx|onyx|NN1 stone|stone|NN1 .|.|PUNC

And|And|CJC the|the|AT0 name|name|NN1 of|of|PRF the|the|AT0 second|second|ORD river|river|NN1 is|is|VBZ Gihon|Gihon|NP0 :|:|PUN this|this|DT0 river|river|NN1 goes|go|VVZ round|round|AVP all|all|DT0 the|the|AT0 land|land|NN1 of|of|PRF Cush|Cush|NN1 .|.|PUNC

And|And|CJC the|the|AT0 name|name|NN1 of|of|PRF the|the|AT0 third|third|ORD river|river|NN1 is|is|VBZ Tigris|Tigris|NP0 ,|,|PUN which|which|DTQ goes|go|VVZ to|to|PRP the|the|AT0 east|east|NN1 of|of|PRF Assyria|Assyria|NP0 .|.|PUNC And|And|CJC the|the|AT0 fourth|fourth|ORD river|river|NN1 is|is|VBZ Euphrates|Euphrates|NP0 .|.|PUNC

And|And|CJC the|the|AT0 Lord|Lord|NP0 God|God|NP0 took|took|VVD the|the|AT0 man|man|NN1 and|and|CJC put|put|VVD him|him|PNP in|in|PRP the|the|AT0 garden|garden|NN1 of|of|PRF Eden|Eden|NP0 to|to|TO0 do|do|VDI work|work|NN1 in|in|PRP it|it|PNP and|and|CJC take|take|VVB care|care|NN1 of|of|PRF it|it|PNP .|.|PUNC

And|And|CJC the|the|AT0 Lord|Lord|NP0 God|God|NP0 gave|give|VVD the|the|AT0 man|man|NN1 orders|order|NN2 ,|,|PUN saying|say|VVG ,|,|PUN You|You|PNP may|may|VM0 freely|free|AV0 take|take|VVI

ስም|ስም|NN1 እቲ|እቲ|AT0 ሓደ|ሓደ|ORD ፒሶን|ፒሶን|NP0 እዩ|እዩ|VBZ ።|።|PUNC ነሱ|ነሱ|pnp ነታ|ነታ|PNP ወርቂ|ወርቂ|NN1 ዘለዋ|ዘለዋ|VHB ዅላ|ዅላ|DT0 ምድሪ|ምድሪ|NN1 ሓዊላ|ሓዊላ|NP0 ይኸብብ|ኸበበ|AVP ።|።|PUNC

ወርቂ|ወርቂ|NN1 እታ|እታ|AT0 ምድሪ|ምድሪ|NN1 እቲኣ|እቲኣ|AT0 ድማ|ድማ|CJC ጽቡቕ|ጽቡቕ|AJ0 እዩ|እዩ|VBZ ።|።|PUNC ኣብኡ|ኣብኡ|AT0 ብዶላህን|ዶላህ|NN1 ክቡር|ክቡር|NN1 ዕንቍን|ዕንቍ|NN1 አሎ|አሎ|VBD ።|።|PUNC

ስም|ስም|NN1 እቲ|እቲ|AT0 ኻልኣይ|ኻልኣይ|ORD ርባ|ርባ|NN1 ድማ|ድማ|CJC ጊሆን|ጊሆን|NP0 እዩ|እዩ|VBZ ፡|፡|PUNC ንሱ|ንሱ|AT0 ንዅላ|ዅሉ|PRP ምድሪ|ምድሪ|NN1 ኩሽ|ኩሽ|NN1 ይዞራ|ዞረ|VBD ።|። |PUNC

ስም|ስም|NN1 እቲ|እቲ|AT0 ሳልሳይ|ሳልሳይ|ORD ርባ|ርባ|NN1 ኸኣ|ኸኣ|CJC ሄዴቄል|ሄዴቄል|NP0 እዩ|እዩ|VBZ ፡|፡|PUNC ንሱ|ንሱ|AT0 ብቕድሚ|ቕድሚ|AJ0 ኣሶር|ኣሶር|NP0 ኣቢሉ|ኣቢሉ|PRP ዚኸይድ|ዚኸይድ|VVZ እዩ|እዩ|VBZ ።|።|PUNC እቲ|እቲ|AT0 ራብዓይ|ራብዓይ|ORD ርባ|ርባ|NN1 ድማ|ድማ|CJC ኤፍራጥስ|ኤፍራጥስ|NP0 እዩ|እዩ|VBD ። |።|PUNC

እግዚኣብሄር|እግዚኣብሄር|NP0 ኣምላኽ|ኣምላኽ|NP0 ከኣ|ከኣ|CJC ንሰብ|ሰብ|NN1 ወሲዱ|ወሰደ|VVD ፡|፡ |PUNC ኪዐይያን|ኪዐይያ|NN1 ኪሕልዋን|ኪሕልዋን|NN1 ኣብ|ኣብ|CJC ገነት|ገነት|NN1 ኤድን|ኤድን|NP0 አቐመጦ|አቐመጠ|VVD ።|።|PUNC

እግዚኣብሄር|እግዚኣብሄር|NP0 ኣምላኽ|ኣምላኽ|NP0 ድማ|ድማ|CJC ንሰብ|ሰብ|NN1 ፡|፡|PUNC ካብ|ካብ|PRP ኵሉ|ኵሉ|DT0 ኦም|ኦም|NN1 ገነት|ገነት|NN1 ከም|ከም|PNP ዝደሌኻ|ደለየ|AV0 ብላዕ|በለዐ|VVD ፡|፡|PUNC of|of|PRF the|the|AT0 fruit|fruit|NN0 of|of|PRF every|every|AT0 tree|tree|NN1 of|of|PRF the|the|AT0 garden|garden|NN1

But|But|CJC of|of|PRF the|the|AT0 fruit|fruit|NN0 of|of|PRF the|the|AT0 tree|tree|NN1 of|of|PRF the|the|AT0 knowledge|knowledge|NN1 of|of|PRF good|good|AJ0 and|and|CJC evil|evil|AJ0 you|you|PNP may|may|VM0 not|not|XX0 take|take|VVI ;|;|PUN for|for|PRP on|on|PRP the|the|AT0 day|day|NN1 when|when|AVQ you|you|PNP take|take|VVB of|of|PRF it|it|PNP ,|,|PUN death|death|NN1 will|will|VM0 certainly|certain|AV0 come|come|VVI to|to|PRP you|you|PNP .|.|PUNC

ካብታ|ካብታ|PRP ጽቡቕን|ጽቡቕ|AJ0 ክፉእን|ክፉእ|AJ0 እተፍልጥ|ፈለጠ|NN1 ኦም|ኦም|NN1 ግና|ግና|CJC ፡|፡ |PUNC ካብኣ|ካብኣ|PRP እትበልዕ|በለዐ|VVD መዓልትስ|መዓልቲ|NP0 ሞት|ሞት|VVD ክትመውት|ሞት|NN1 ኢኻ|ኢኻ|NN1 እሞ|እሞ|AJ0 ፡|፡ |PUNC ካብኣ|ካብኣ|PRP ኣይትብላዕ|በለዐ|VVD ፡|፡|PUNC ኢሉ|ኢሉ|AJ0 ኣዘዞ|ኣዘዘ|VVD ።|።|PUNC