ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLAGE OF NATURAL SCIENCEs

DEPARTMENT OF COMPUTER SCIENCE

A HYBRID APPROACH TO AMHARIC BASE PHRASE CHUNKING AND PARSING

BY

ABEBA IBRAHIM

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

MARCH 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLAGE OF NATURAL SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

A HYBRID APPROACH TO AMHARIC BASE PHRASE CHUNKING AND PARSING

BY

ABEBA IBRAHIM

Advisor: Yaregal Assabie (PHD)

Signature of the Board of Examiners for Approval

Name	Signature
1. Dr. Yaregal Assabie	_____
2. _____	_____
3. _____	_____

Declaration

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university and that all sources of material used for the thesis have been duly acknowledged.

ABEBA IBRAHM

MARCH 2013

The thesis has been submitted for examination with our approval as university advisor.

_____

Dr. Yaregal Assabie

Acknowledgements

Above all, my truthful thank and praise goes to the Almighty God for making this study successful. All the days what you have done for me is really beyond imagination. Indeed, thanks for all I have received and all that is yet to come.

I would like to express my deepest gratitude to my advisor, Dr. Yaregal Assabie for his close supervision, and constructive suggestion during my research work. He has been devoting his time and providing all necessary relevant information to carry out the research.

My special thanks go to my parents and my fiancé. They have been very patient with me, while I was working on this thesis and thus did not have enough time for them. Besides that, they have always supported me during my study and I am very grateful for that. They were the ones who encouraged me to study, and they have contributed to this thesis more than they might think. Finally, I would like to give my gratitude to people who are not mentioned in name but whose effort helped me much all along.

List of Figures

List of Tables

Abbreviations

NLP- natural language processing

WIC- Walta Information Center

HMM- Hidden Markov Model

IOB- Inside Outside Beginning

XML- Extensible Mark-up Language

NP – Noun Phrase

VP- Verb Phrase

PP- Prepositional Phrase

AdjP- Adjectival Phrase

AdvP- Adverbial Phrase

S- Sentence

WSJ- Wall Street Journal

NV- Verbal Nouns

NC- Noun with Conjunction

N- Noun in all forms

NP- preposition not separated from a noun

NB- Noun prefixed with

V- Verb

AUX- Auxiliary verbs

NPrep- a noun with preposition

VPrep- a verb preposition

J, ADJ- Adjective

C, CONJ- Conjunction

JC, ADJC- Adjective with Conjunction

REL, VREL - relative clause

ITJ, INT- interjection

ORD, NUMOR- ordinal number

CRD, NUMCR- cardinal number

Abstract

Nowadays, Natural Language Processing (NLP) concerns with the interaction between computers and human natural languages. The most difficult task in NLP is to learn natural languages for the computer. Enabling computers to understand natural language involves assigning of words with their Part Of Speech, extraction of phrases, extraction of meaning, etc from natural language sentences. Text chunking and sentence parsing are among the tasks of NLP. Text chunking or shallow parsing is one of the tasks of NLP which divides a text in syntactically correlated words from a stream of text. It is an intermediate step of full parsing. As well as, text chunking could be used as a precursor for many natural language processing tasks, such as information retrieval, named entity extraction, text summarization and so on.

The objective of this research is to extract different types of Amharic phrases by grouping syntactically correlated words which are found at different level of the parser using Hidden Markov Model (HMM) model and to transform the chunker to parser. Some rules are also used in this study to correct some outputs of HMM based chunker. Bottom-up approach with transformation algorithm is used to transform the chunker to the parser. For the identification of the boundary of the phrases IOB2 chunk specification is selected and used in this study.

In this study different sentences are collected from Amharic grammar books and news of Walta Information Center (WIC) for the training and testing datasets. Unlike the data collected from WIC, the data collected from Amharic grammar books are not tagged at all. Thus, these data sets were analyzed and tagged manually and used as a corpus for chunking. But the entire data sets were chunk tagged manually for the training data set and approved by linguistic professionals.

Experiments have been conducting using the training and testing data sets. The training and testing datasets are prepared using the 10 fold cross validation. The experiments on Amharic sentence chunking showed an average accuracy of 85.31% testing set before applying the rule for correction and an average accuracy of 93.75% on the test set after applying rules. And also the experiment on Amharic sentence parsing showed an average accuracy of 93.75%.

Keywords: Amharic Text chunking, Amharic partial parsing, Amharic shallow parsing, Amharic Parsing

Table of Contents

CHAPTER ONE

INTRODUCTION

1.1 Background

In computing, Natural language is a language that is spoken or written by humans for general purpose communication. Teaching computers to understand and realize the way how humans use and learn natural language is one of the most challenges in Artificial intelligence [1].

Natural Language Processing (NLP) is a very attractive method of human–computer interaction. NLP describes the function of computer system to analyze or synthesize spoken or written natural language [2]. It deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages [1].

Natural Language processing and understanding requires the general language structures of text at phonological, at morphological, at syntactical, at semantical, at discourse and pragmatic levels [3] to increase the capability of NLP applications that have been and are being done at different level. The syntactic level concerns how can be words put together to form correct sentences and determines what structural role each word plays in the sentence. Generally, analyzing a sentence is done in the syntactic level of NLP tasks.

Broadly speaking, the syntactic level deals with analyzing a sentence that generally consists of segmenting a sentence into words, grouping these words into a certain syntactic structural units such as noun phrase; recognize syntactic elements and their relationships within a structure. Syntactic level also indicates how the words are grouped together into phrases, what words modify other words, and what words are of central importance in the sentence [4]. Determining these phrases within a sentence is very useful for a variety of natural language processing NLP applications, and the process which directly labels these phrases is called chunking (Shallow parsing or light parsing).

Chunking is a Natural Language Processing (NLP) task that consists in dividing a text into syntactically correlated parts of words. These words are non-overlapping, i.e., a word can only be a member of one chunk and non exhaustive, i.e., not all words are in chunks [5]. Chunking helps to identify non-overlapping phrases from a stream of data, which are further used for the development of different NLP applications such as information retrieval (IR), information extraction (IE), named entity recognition (NER) and so on. These NLP tasks consist of recognizing some type of structure which represents linguistic elements of the analysis and their relations. In text chunking the main problem is to divide text into syntactically related non overlapping groups of words (chunks) [6].

Example:- *ትንሹ ልጅ ትንሽ እንጀራ በላ "The little boy ate little Injera"*

Here the sentence can be segmented into three basic syntactic non overlapping phrases: a noun phrase ትንሹ ልጅ "the little boy", a verb phrase በላ "ate", and another noun phrase ትንሽ እንጀራ "little Injera".

Chunking has become an interesting alternative to full parsing. As described above the main goal of chunking is to divide a text into segments which correspond to certain syntactic units. Abney [6] introduced the concept of chunk as an intermediate step providing input to further full parsing stages. Thus, Chunking can be seen as the basic task in full Parsing. Because it is a technique used to convert a sentence in to simple non over-lapping syntactic structures (phrases) and these are NP (Noun Phrase), VP (Verb Phrase), PP (prepositional Phrase), etc to parse a sentence.

Although the detailed information from a full parse is lost, chunking is a valuable process in its own right when the entire grammatical structure produced by a full parse is not required. For example, various studies indicate that the information that is obtained by chunking or partial parsing is sufficient for IE technologies rather than full parsing [7]. Alongside, partial syntactical information can help to solve many NLP tasks, such as text summarization, machine translation and spoken language understanding [8]. For example, Kutlu [9] stated that finding noun phrases and verb phrases is enough for IR technologies. Phrases that give us information about agent, time, places, objects, etc. are more significant than the complete configurational syntactic analyses of a sentence for question-answering, information extraction, text mining and automatic summarization.

Chunkers do not necessarily assign every word in the sentence like full parses to a higher-level constituent. They identify simple phrases but do not require that the sentence be represented by a single structure. By contrast Full parsers attempt to discover a single structure which incorporates every word in the sentence. Abney [6] proposed to divide sentences into labeled, non-overlapping sequences of words based on superficial analysis and local information. In general, many of NLP applications often require syntactic analysis at various NLP levels including full parsing and chunking. The chunking level identify all possible phrases and the full parsing analyzes the phrase structure of a sentence .The choice of which syntactic analysis level should be used depends on the specific speed or accuracy of an application. The chunking level is Efficient and fast in terms of processing than full parsing [10].

Chunkers can identify syntactic chunks at a different level of the parser so group of chunkers can build a complete parser [6]. Most of researches for other language like English and Germany use chunkers in parsers. Ejerhed and Church [11] use noun phrase chunk rules to express a grammar for Swedish. Abney [6] uses base chunker to find base phrases and then attaches them with the other attachment process to build a full chunk parser. Brants [12] used a cascade of Markov model chunkers for obtaining parsing results for the German NEGRA corpus.

Parsing is the use of syntax to determine the functions of words in the input sentence in order to generate a data structure that can be used to get at the meaning of the sentence [13, 14]. It is a process of converting words to their syntactic structure. Therefore, It is the method used to analyze the various parts of a string to differentiate whether the string is a sentence in the language or not. In addition to this, parsing deals with a number of sub problems such as identifying constituents that can fit together. Generally, parsing assists to understand how words put together to form the correct phrase or sentence, each and every structural role of the words, to understand which phrases are sub parts of other phrases i.e. word modifies the other word and to understand what words are important in the sentence.

Like chunking, parsing also plays an important role in NLP applications. Syntactic parsing is used as an input for semantic analysis. In linguistic, languages are sets of valid sentences. Thus these sentences are required to be grammatically correct and meaningful. The syntactic analysis (parsing) is only concerned on the grammatical correction of the sentence whereas the semantic

analysis is concerned on the meaning of the sentence. So one sentence to be correct, first it must be correct grammatically then it makes meaning. Hence, semantic analysis uses the syntactic one as an input to check whether the given sentence is grammatically correct or not. When comparing to semantic analysis, syntactic analysis is less expensive. Therefore, it plays a significant role in reducing overall system complexity [15].

Today there are a lot of chunking and parsing systems developed for various languages of the world such as Swedish [11], Chinese [16], Urdu [17], Germen [12], etc. As far as the researchers' knowledge, there is no such kind of chunking system developed so far for the Amharic language which is very important for many NLP applications. But Atelach [15] and Daniel [18] have worked on Amharic word parser for simple sentences and complex sentences, respectively. These works are reviewed in this research work. Hence, this paper deals with the development of text chunker for Amharic language and using of the output of the chunker to the parser.

1.2 Motivation

Many of the NLP applications have been conducted for different languages for different purposes such as Named Entity Recognition [19], Noun Phrase (NP) Chunking [20], Clause Identification [21] etc for English language. In recent times, NLP applications for local languages such as Amharic have become an area of research interest.

Amharic, a language that is mainly spoken in Ethiopia, is one of an Africa language. The number of speakers of the language is on the rise for two reasons. First, it is the working language of the Federal Democratic Republic of Ethiopia, a country with more than 85 million people. Second, unlike most other African languages, Amharic is a written language with its own alphabet and written materials, actively being used every day in newspapers and other media outlets. Its alphabet is known as *fidel (*"alphabet", "letter", or "character"*)* `which grew out of the Ge'ez *abugida*. And also this language has its distinct way of grammatical phrase construction, character (fidel) representation and statement formation [22].

There have been some NLP applications conducted for Amharic and other local language such as Oromo language and Wolita language. These include, for instance, 'An Amharic Word Parser'

by Abiyot Bayou [23], 'A Part-of-Speech Tagger' by Mesfin Getachew [24], 'An Amharic Sentence Parser' by Atelach Alemu [15] and 'Automatic Complex Sentence Parsing' by Daniel Gochel Agonafer [18], 'Statistical Morphological Analysis for Amharic' by Tesfaye Bayu [25] ,'An Automatic Sentence Parser For Oromo Language' by Diriba Megersa [27] and 'A Morphological Synthesizer for Amharic Perfectives' by Kibur Lisanu [26]. To solve many problems of NLP, there is a lot to be done.

As illustrated earlier, not all NLP application use the full information of parser, so text chunking is a technique to reduce cost of full parsing, it also trim down the search space. Abney [6] realized the need of chunking in the design of parsing and also in different NLP applications. Grover [28] considers chunking useful for Named Entity Recognition. Considering all these, text chunker has already been developed for different languages such as English [20, 29, and 30], Hindi [31] and so on. But, to the best knowledge of the researcher, text chunker has not been developed for Amharic yet.

Not only chunking but also parsing is useful for various NLP applications such as semantic analysis. The absence of syntactic parsing bound the application of higher NLP such as semantic analysis, discourse level, spelling and grammar checking, etc [15]. To the best of the researchers' knowledge, researches like the automatic simple sentence parser for Amharic developed by Atelach [15] and the automatic complex sentence parsing for Amharic by Daniel [18] were conducted in this regards.

The parser which was developed by Atelach had 85% accuracy. However, it had the following limitations:

- The automatic sentence parser accepts and parses only 4 words length declarative sentences.
- The text used in the study is transcribed
- Only 100 simple declarative sentences were used both for training and testing purposes

Whereas, a parser developed by Daniel had an accuracy of 89.6% on the training set and 81% on the test set. Nevertheless, the complex sentences that are included in the sample do not include

complex noun phrases. The parser parses only simple noun phrases and complex verb phrases. But this study improves the limitations of Atelache's parser.

Therefore, the major concern of this study is developing text chunker (shallow or partial parser) and then transforming this chunker to parser. The approach followed in this study is hybrid. And the sentences that are used for training and testing sets are simple sentences specifically declarative sentences and complex sentences which are composed of simple NP and complex VP, complex NP and simple VP and both complex NPs and complex VPs [1].

1.3 Statement of the problem

The text chunker or base phrase recognizer for different languages recognize the types of phrases such as noun phrase, adjectival phrase, adverbial phrase, prepositional phrase and verb phrase with their correct phrase construction. However, there has not been such text chunking developed for Amharic yet. Therefore, text chunker that considers the special characteristics of the language and that fulfill the stated requirements needs to be developed for Amharic. In this study, we will investigate the problems and limitations of Amharic text chunking, the effect of developing text chunking, and try to develop the design of Amharic text chunker. The problem is "Given an Amharic language sentence along with POS tags of tokens to identify base phrases. And use this base phrases as an input for parsing".

1.4 Objective of the Study

1.4.1 General Objective

The general objective of this research work is to analyze and design a text chunker for Amharic text and use the output of the chunker to a parser a text.

1.4.2 Specific Objective

The specific objectives of this research work are to:-

- Review word categories, types of phrases and sentence formation of Amharic language

[1]For more information see chapter 3

- Study the phrase construction rules appropriate for the language to recognize phrase types
- Collect sample simple and complex sentences that are used for the preparation of the corpus and serve for the experiment
- Review techniques of text chunking adopted for other language
- Review techniques and methods to transfer chunking algorithm to parsing algorithm
- Build the full system
- Evaluate the performance of the system

1.5 Methods

In the development of NLP applications, exploring of the property of the target language is needed. The following are methods that have been followed out to achieve the above mentioned general and specific objectives of the research work.

1.5.1 Literature review

Literature review on text chunking done in other languages has been conducted to understand techniques how text chunker works. Various related literature resources such as research papers, books, manuals, journal articles, previous related research work as well as electronic materials on the Web have been reviewed to have better knowledge of detecting types of phrase, to understand the phrase structure of Amharic language and to be aware of the strategies, techniques and approaches (Rule based, statistical based or corpus based and hybrid approach) how to chunk the text and how to transfer the chunker to a parser.

For this research the tagged text is used as an input so a brief review was made to understand and identify the tag set types. This all consideration of the review helped the researcher to put into practice the feature of the Amharic language.

1.5.2 Discussion

Continuous discussion with linguistic professionals and experts has been conducted for the linguistic knowledge acquiring like the correct and wrong phrase structure of the language and types of phrases with their differences.

1.5.3 Data collection

A number of Amharic corpora have been collected from Walta Information Center (WIC) [2]. And some sentences were collected from book entitled የአማርኛ ሰዋሰው "The Amharic Grammar" by Baye [32, 33]. And also a number of simple and complex sentences were collected from the research paper of Atelach and Daniel. Daniel has collected 350 sentences from books entitled"Amharic for Beginners" [34], and የአማርኛ ሰዋሰው "The Amharic Grammar" [32] and the sentences had been manually analyzed and tagged. Among these 350 sentences only some of the complex sentences have been used for this research.

Two types of data sets were used during the experimentation that are training data set and test data set. So the sentences had been broke up into two for the training set and test set. For the training set the sentences had been manually chunked by the researchers. It was then given to the linguistic experts in order to get feedback on the correctness of the manual chunk.

1.5.4 Design and Implementation

For this research the tagged document that is collected from news of Walta and research papers has been used as an input. In the experimentation IOB chunk tag set has been used to identify the chunk boundaries. And to label the chunks with their syntactic categories Hidden Markov Model has been used. The Viterbi algorithm has been adopted to find the optimal sequence of chunk tags by using the parameters of HMM. Plus some rules have been employed to group the phrases.
Finally, Bottom-Up parsing algorithm has been used to transfer the chunking algorithm to parsing algorithm.

Python 3.1 was used for the development of the prototype and implementation of the chunking algorithm plus transforming chunking to parsing algorithm. This programming language is selected for the various features it provides. One of the reasons is that the language can easily be manipulated to code the algorithm.

[2]Walta Information Center (WIC) is a government information center that produces and distributes news for broadcast over television and radio (Saba, 2001)

1.5.5 Testing

In the experimentation of this study, training and testing data sets are used to train and test the system, respectively. All the data sets are collected from different Amharic Grammar books, researches and news[3]. In this study, 320 sentences are collected from different aources and the corpus is prepared using the 10 fold cross validation for the training and testing datasets. Using the 10 fold cross validation the experiment in this study was conducted ten times. For each ten phases 288 sentences were used to train the system while 32 sentences were used to test the system. Finally the output has been crosschecked with the hand chunk tagged sets.

1.6 Application of Results

As outlined in the above sections, text chunking plays a significant role in many areas of NLP applications for the target language. Text chunking can be used in development of following NLP applications [17, 31, 19, 16, 35, 36, 28, 51].

A. Text summarization
B. Named Entity Recognition (NER)
C. Information extraction(IE)
D. Information Retrieval (IR)
E. Question Answer Applications (QA)
F. Machine Translation (MT)
G. Speech Synthesis and Recognition
H. Index Term Generation
I. Parsing
J. Text Mining
K. Searching

Even though text chunking is useful for NLP applications that do not require a complete syntactic analysis, parsing is also useful for many NLP tasks such as

A. semantic parser

[3]See the corpus preparation for this study in section 4.2

B. Spell checker
C. Grammar checker
D. Automatic abstracting

Linguists and students in the area of Amharic language could also apply the output of this research to chunk and to parse simple and complex sentences automatically. The output can also be used in language teaching for recognition of phrasal categories, and to see the relationship between words in a sentence. Moreover, those who are interested in generating the syntactic structure can use it.

1.7 Scope of the Study

The scope of this study is to chunk the given free text with POS tagged which is to find different types of phrases and indicate the way how this chunker transforms to parser. These phrases are useful in various NLP applications such as information extraction and they are useful to provide better platform for fully parsing of sentences. In this research, all types of declarative sentences are considers such as simple sentences and all combinations of complex sentences which are simple NP and Complex VP, complex NP and simple VP or both complex NP and complex VP. All these types of sentences will be discussed in the next chapters.

1.8 Limitations of the Study

This study has the following limitations:

- All kinds of Amharic sentences are not included in this study. The sentences that are included in the training and testing dataset do not contain interrogative and imperative sentences.
- The size of the corpus is very small. The corpus is prepared manually for the purpose of the work.
- The tree of the parser do not include in this study. Only the way how to convert the chunker to parser using bottom up algorithm is discussed.

1.9 Organization of the Paper

This section describes the composition of this paper. The paper is divided into six chapters. Chapter one describes what NLP, chunking and parsing are. This chapter also presents the motivation, statement of the problem, objective, scope and methodology used in this study. Chapter two also introduces Amharic language, word classes of Amharic language and phrase construction of the language with examples. Different types of sentences including simple and complex sentences also explained in this chapter. Different types of approaches and techniques for chunking and also parsing are discussed in chapter three. Different types of representation of chunk tags also presented in this chapter. The core of this study is discussed in chapter four. In this chapter the approach and technique used for the system chunking and algorithm to transform chunking to parsing is explained. Finally, the conclusions and recommendations made based on the findings of the study are presented in chapter six.

CHAPTER TWO

STRUCTURE OF AMHARIC

2.1 Introduction

A natural language is used as a tool for communication and people use it for communication by combining phonologies to form words, by combining words to form phrases and by combining phrases to form sentences. This chapter discusses the structure of Amharic word classes, phrases types and sentences formation with their types.

In the early Amharic Grammar books, word classes are classified in to eight different types i.e. nouns, pronouns, verbs, adjectives, prepositions, interjections, adverbs and conjunctions. But Baye [33] classified these word class into five types i.e. nouns, verbs, adverbs, adjectives and prepositions. Among the word classes nouns, verbs, adjectives, adverbs, prepositions and conjunctions are discussed in this chapter. Phrase structures of the Amharic language such as noun phrases, verb phrases, adjectival phrases, adverbial phrases and prepositional phrases classified by Getahun [22] are all briefly discussed. And sentence formalisms of the language, particularly the simple sentences and complex sentences, are all discussed in this chapter.
The discussions made in this chapter are based on the information took out from Baye [33], Getahun [22], Atelach [15] and Daniel [18].

2.2 Basic Word Class

The linguistic characteristics of Amharic language have been studied by different researchers in different time. Research works in the area of Amharic word categorization have been carried out by Mersi'hazen [37], Baye Yimam [33], Getahun Amhare [22] and etc. Among the researches, Mersi'hazen classified Amharic word classes into eight categories (as described above) i.e. nouns, pronouns, verbs, adjectives, prepositions, interjections, adverbs and conjunctions. However, in the recent works of Baye, the eight traditional grammars is summarized and divided into five categories in the language. These are noun, verb, adverb, preposition and adjective.

The reasons, that the three word classes stated in early Amharic grammar category (pronouns, interjections and conjunctions) are not included in the recent word category, are:

- Pronouns can be used as a noun for the reason that pronouns can replace nouns. For example, አበበ “Abebe”, a noun, can be replaced by እሱ “He”, a pronoun. So pronouns can be included in nouns word class.
- In view of the fact that conjunctions have the same property and function as of preposition, they have included in prepositions word class.
- Interjections are not considered as Grammatical Categories because they are words without syntactic functions

The current study adopts the classification scheme of the early scholars but pronouns and nouns are considered in the same category as of Baye. The fact, that this classification can be preferred, is the Walta document used as an input for this research took on this classification of part of speech.

The following sections of this chapter explores into the discussion of the grammatical categories of Amharic as a background to the tasks to be carried out in chapter four and five, which are the central and major contribution of this thesis in the area.

2.2.1 Noun word class

Amharic nouns are words that are used to identify any group of things, names, places, etc. For instance, they can identify or name the thing either that can be seen by our eye or not like ወንበር “chair” (that can be seen by eyes) and ጨለማ “darkness” (that cannot be seen by eyes). These word classes have the following common structures and properties.

- They may use -ኦች “-och” morpheme as a plural marker
- They can be used as a subject in the sentence
- They can be used as an object in the sentence
- They can take modifiers and quantifiers

For this study, the Amharic noun categories consist of nouns and pronouns. Unlike nouns, pronouns take እነ- “ene-” morpheme as its plural marker.

Amharic nouns can be either primitive or derived. They are said to be in their primitive forms if they exist in their original form whereas they are referred to as derived if they originated (derived) into their present state from a different, and possibly completely different categories (Daniel).

Amharic nouns can be derived from:

- Verbal Roots by infixing vowels. The vowels can be 'እ', 'ኧ', 'ኣ'.

  See the following table how to derive nouns from verbal roots

Verbal root	Derived noun
ጥ-ቅ-ም	ጥእቅእም (ጥቅም) "benefit"
ም-ል-ስ	ምኧልስ (መልስ) "answer"
ቅ-ል-ም	ቅኧልኧም (ቀለም)"Paint"
ደ-ክ-ም	ድእክኣም (ድካም) "tirednesss"

- Adjectives and nouns by suffixing bound morphemes[4]

  Example: - the derived nouns ደግነት "kindness" and ልጅነት "childhood" derived from an adjective ደግ "kind" and ልጅ "child" respectively by suffixing the morpheme -ነት "-ness or hood".

- Stems by prefixing or suffixing bound morphemes (ኧት፣ኤ፣እና፣ኣት፣አሽ፣ኡታ፣ኤት፤ኣ፤ኦት፣ኢት፣ኢ፣ኢያ፣ኤታ፤መ፣ኛ፣)

  Example: - derived noun እርጅና is derived from the stem እርጅ by suffixing the morpheme እና and the derived noun መሄድ is derived by prefixing the stem ሄድ

- Compound Words (sometimes by affixing the vowels ኧ and ኦ)

  Example: - the nouns ብረት and ምጣድ compounded to form the compound word ብረት ምጣድ

  ቤት መንግስት ኧ ቤተ መንግስት "palace"

  ሰርት አደር ኦ ሰርቶ አደር "laborer"

[4]A morpheme is minimal meaning-bearing unit in a language

2.2.2 Adjective word class

Amharic adjectives modify nouns or a pronouns by describing, identifying, or quantifying words. Adjectives always come before nouns or pronouns which they modify. But all the words that come before nouns cannot always be an adjective.

For example: - *ይህ በግ "This sheep"*

In this example ይህ "This" precedes the noun በግ "sheep" but this doesn't mean ይህ "this" is an adjective, it is a pronoun.

Like nouns adjectives also either primitive or derived. They can be derived from

- Verbal Roots by infixing vowels ኧ፣ኢ፣ኡ፣ ኣ between consonants as shown below
  ስ-ን-ፍ ስኧንኧፍ (ሰነፍ) "lazy"
- Nouns by suffixing bound morphemes (ኧኛ፣ ኣማ፣ኣም፣ አዊ)
- Stems by suffixing bound morphemes (ኣ፣ ኡ፣ ኢታ)
- Compound Words of nouns and adjectives by affixing the vowel -ኧ

2.2.3 Verb word class

Verbs are words which indicate action and they take place in the end of clause positions. The other property of Amharic verbs is, they take subject markers as a suffix like -ሁ /-hu/ for subject 'I', -ህ /-h/ for subject 'You', ች /-c/ for subject 'She' and so on, to agree with the subject of the sentence. Some properties of verbs are

- Among the seven Amharic writing symbols order, majority of the verb words use the first order Amharic writing system
- Verbs can use 'እየ-' prefix morpheme
- Verbs can change their last symbol to the Amharic seven order writing system. Finally this changed verbs may take '-አል' suffix morpheme

Similar to nouns and adjectives verbs also derived from Verbal Roots by affixing the vowel ኧ, Verbal Stems by affixing morphemes and compound Words of stems with verbs.

2.2.4 Adverb word class

In Amharic, adverbs are used to modify the coming verbs. Adverbs always come before the modified verb but it should be noted that any words come before verbs cannot be always considered as an adverb. In their nature, adverbs can be found either in their primitive form or compound form as grouping of preposition and other word categories. Adverbs indicate manner, time, place, cause, or degree and answers questions such as እንዴት "how", መቼ "when", የት "where". The primitive adverbs are very few in number and these are: ገና “yet”, ክፉኛ “severely”, ቶሎ “quickly”, ጅልኛ “foolish”, etc.

2.2.5 Preposition word class

Prepositions give meanings only if they combine with other words such as noun, adjective, verb, etc, unless they have no meaning. Prepositions link nouns, pronouns and phrases to other words in a sentence. The main properties of prepositions are: they never use affixes and they don’t assist to form other words. Like adverbs, prepositions also few in number and these are: ስለ “for”, እንደ “like”, ከ”from”, ወደ “to”, አጠገብ “near to”, ማዶ “there”, ወዲህ “here”. As described above, Prepositions will have meaning only when they combine with other word class therefore they are used as affixes by coming before and after words. Prepositions are consisting of two parts which are prepositions coming before words and after words. The one which come before words and after words are: ስለ “for”, እንደ “like”, ወደ “to”, እስከ “up to”, ከ “form” and አጠገብ “near to”, ማዶ “here”, respectively. See the following examples

Preposition	Example
ሰለ “for”	ስለ ገንዘብ “for money”
እንደ “like”	እንደ ሰው “like a man”
ወደ “to”	ወደ ደሴ “to Dessie”

In this example all the stated prepositions comes always before the words they combine and these prepositions can stand alone as separate words

Preposition	Example
ከ-አጠገብ	ከተማሪ አጠገብ “next to a
ከ-ማዶ	ከወንዝ ማዶ “beyond the river”

From the above example the stated prepositions comes always together by placing other words in between.

To sum up, prepositions can appear as an affix for other word, as simple preposition that stand alone as separate word or they can appear as a combined prepositions [33].

2.2.6 Conjunction word class

Conjunctions are one of part of speeches that are used to connect words, phrases, clauses and sentences. In the sentence, አበበ እና ከበደ ጎበዝ ተማሪዎች ናቸው "Abebe and Kebede are clever students" the conjunction እና "and" connects the two nouns አበበ "Abebe" and ከበደ "Kebede". This word class comprises of coordinate conjunctions and subordinate conjunctions. Coordinate conjunctions are used to join two equal words, phrases, clauses and sentences in contrast subordinating conjunctions connect two parts of a words that are not equal (dependent and independent clauses). Subordinate conjunctions introduce dependent clause (s) and indicates the nature of the relationship among the independent clause(s) and the dependent clause(s). Like adverbs and prepositions conjunctions are very few in number and these are: እና "and", ከ "from", ወይም "or", ስለዚህ "so", እስከ "upto", ግን "however", ስለ "for".

2.3 Phrasal category

Phrases are syntactic structures that consist of one or more than one word but lack the subject-predicate organization of a clause. These phrases are composed of either only head word or other words or phrases with the head combination. The other words or phrases that are combined with the head in phrase construction can be specifiers, modifiers and complements. The definitions are explained below:

Modifiers are used to specifically point out the amount, time, place, type, etc of the head word or phrase in the phrase construction. Modifiers can be adjectival phrase, noun phrase, prepositional phrase or sentences.

Example: - *ጥቁር በግ "Black sheep"*

Here ጥቁር "black" is adjectival phrase, as described above one word can be a phrase. This adjectival phrase ጥቁር "black" specifically point out the type or color of በግ "sheep".

To realize more, see the next example when the modifier is sentence:

ትናንት ከጎጃም የመጣው መኪና "the car which came from Gojam yesterday"

Here ትናንት ከጎጃም የመጣው "which came from Gojam yesterday" is a sentence that point out the head word መኪና "car" form time and place.

This all does not mean that only one modifier appears in phrase construction rather in phrase construction more than one modifier may come. Example: - ዛሬ የመጣው ጎበዙ ተማሪ "the clever student that came today"

In this example two modifiers come together and these are ዛሬ የመጣው "that came today" sentence and ጎበዙ "clever" adjectival phrase.

In phrase construction **specifiers** are used to specify the identity, location, number, and possession etc of the head. They can be genitive, deictic or quantifier. They may be either primitive or derived. They can be derived from

- Noun by prefixing 'የ' morpheme to indicate possession
  Example: - *የካሳ መጽሀፍ "Kassa's book"*
  Here the specifier is የካሳ "Kassa's" that shows the owner of the book.
- From modifiers (specially prepositional phrase) by combining specifiers

Complements are used to make ideas complete. See the following example:

ዳቦ በላሁ "I ate bread"

የስንዴ ዳቦ በላሁ "I ate wheat bread"

Here the first sentence does not give full information about the bread but in the second sentence የስንዴ "wheat" is complement that indicates from what the bread is made to get full information about the bread.

In this section ways of combination of words how to form phrases and types of phrases discussed briefly. The type of the phrase is determined by the words which it is formed from. All the word classes that are discussed in the above section used as a head in the phrases structure discussed below.

In Amharic, phrases are categorized into five categories, namely noun phrase (NP), verb phrase (VP), adjectival phrase (AdjP), adverbial phrase (AdvP) and prepositional phrase (PP) [33].

2.3.1 Noun phrase

A noun phrase (NP) is a phrase that has a noun as its head. In this phrase construction, the head of the phrase is always found at the end of the phrase. This type of phrase can be made from a single noun or combination of noun with either other word classes including noun word class or phrases. That means one noun can be a noun phrase. See the following example: *አንበሳው ሁለት ላሞች ገደለ "the lion killed two cows"* in this sentence there are two parts: the subject አንበሳው" the loin"and the object with the verb ሁለት ላሞች ገደለ"killed two cows". Thus, the first part (the subject) is a noun phrase and the second one is a verb phrase. Therefore, the noun phrase in the above example is only the noun አንበሳው"the lion".

The other words or phrases that are constituents in NP can be specifiers, modifiers and complements. In the noun phrase construction specifiers, modifiers and complement always come before the head (the noun).

Example: - *የወርቅ ቀለበት "Golden Ring"*
Here the constituent የወርቅ "Golden" is a complement that comes before the head word ቀለበት "ring".
Specifiers that are to be used in noun phrase construction can be noun phrases, adjectival phrases and sentences, whereas modifiers can be genitive, deictic or quantifier.

A NP can be simple or complex. The simplest NP construction consists of a single noun or pronoun for instance በግ "sheep", መኪና "car", እሱ "he", እሷ "she" are the simplest NP and deos not consist subordinate clauses in it. A complex NP can consist of a noun with other constituents (specifiers, modifiers and complements) but the phrase must contain at least one sentence (Baye).
Example: -In the NP *ካሳ የገዛው የዝናብ ኮት (the rain coat that Kassa bought)* here ካሳ የገዛው (that Kassa bought) is a sentence which is a modifier, whereas የዝናብ (the rain) is the single word that is a complement.

2.3.2 Verb phrase

A verb phrase (VP) is composed of a verb as a head, which is found at the end of the phrase, and other constituents such as complements modifiers and specifiers. But not all the verbs take the

same category of complement. Based on this, verbs can be dividing into two. These are transitive and intransitive. Transitive verbs do take transitive noun phrases as their complement and intransitive verbs do not.

Example 1: - *ካሳ [ምሳውን] በላ (Kassa ate his lunch)*

አስቴር [ብርጭቆ] ሰበረች (Aster broke glass)

ካሳ ትናንት [መኪና] ገዛ (Kassa bought a car yesterday)

Here all the words in the bracket are transitive NPs that are complements of the head verb next to them. These verbs (በላ (he ate), ሰበረች (she broke), ገዛ (he bought)) take one complements and also there are verbs that take two complements. For example, in the following VPs አስቴር [ለካሳ] [መጽሀፍ] ሰጠችው (Aster gave a book to Kassa) ካሳ [ለውንድሙ] [ምስጢር] ነገረው (Kassa told to his brother a secret), there are two complements in each sentence which are in the brackets.

As described above intransitive verbs do not take NPs as their complements rather they take prepositional phrase (PP).

Example 2:- *ካሳ [ወደ ቤተ ክርስቲያን] ሄደ "Kassa went to church"*

ካሳ [ወደ ቤቱ] ገባ "Kassa entered to his house"

Here the entire phrases which are in brackets are PP.

In VPs modifiers are found in the form of prepositional phrase, noun phrase, adverbial phrase and sentences. Like other phrases in VP construction, all the constituents listed above may come together.

Example 3: - *ሁለት ጊዜ በባንክ ለአስቴር ገንዘብ ልኮላታል (he has sent money to Aster two times through bank)*. All the constituents are found in this example and theses are specifier ሁለት ጊዜ"two times", modifier በባንክ"by bank" and complement ለአስቴር ገንዘብ "money to Aster". In Amharic VPs can be simple or complex. All examples stated above are examples of simple VPs. VPs are said to be complex if they contain more than one verb or embedded sentence.

Example 4: - *[ካሳ መኪና እንደገዛ] ሰማች "She heard that Kassa has bought a car"*

Here the VP is complex because it consist an embedded sentence which is found in the bracket ካሳ መኪና እንደገዛ "that Kassa has bought a car" and also in this phrase construction there exist two verbs እንደገዛ "that he has bought" and ሰማች "she heard".

2.3.3 Prepositional phrase

Amharic prepositional phrase (PP) is made up of a preposition (Prep) head and other constituents such as nouns, noun phrases, prepositional phrase, etc. unlike other phrase constructions, prepositions cannot be taken as a phrase instead they should have combine with other constituents and the constituents may come either previous to and subsequent to the preposition which is the head of the phrase. Broadly speaking, if the complements are nouns or NPs, the position of prepositions are in front of the complements whereas if the complements are PPs, the position will shift to the end of the phrase.

Example: - *head (preposition) + complement (nouns or NPs)* *እንደ ትልቅ ልጅ"like a big child"*

Complement (PP) + head (preposition) *ከወንዙ አጠገብ "Next to the river"*

Not only complements but also PP consists prepositional modifiers.

2.3.4 Adverbial phrase

Amharic Adverbial phrases (AdvP) are made upof one adverb as head word and one or more other lexical categories including adverbs itself as modifiers. The head of the AdvP also found at the end. Unlike other phrases, AdvPs do not take complements. Most of the time, the modifiers of AdvPs are PPs that comes always before adverbs.

Example: - *ካሳ [እንደ አባቱ ክፉኛ] ታመመ (Kassa is severely sick like his father)*

Here the phrase in the bracket is adverbial phrase and the head word is ክፉኛ (severely). The modifier that is found in the AdvP is እንደ አባቱ (like his father) which is comparative PP.

2.3.5 Adjectival phrase

As explained above, most of adjective words are derived from nouns or verbs. In this section the phrase construction of these words is discussed briefly. Adjectival phrase (AdjP) can be composed of an adjective (head) that is found at the end, and other constituents such as complements, modifiers and specifiers.

Example: - *ካሳ [በጣም እንደ እህቱ ሰው ፈሪ] ነው “Kassa is very shy like his sister”*

Here the phrase in the bracket is adjectival phrase. In this phrase construction there are modifier and specifier which እንደ እህቱ“like his sister” and በጣም “very” are respectively. When specifiers, complements and modifiers come together in the adjective phrase construction, the complements always come next to modifiers or specifiers and modifiers come next to specifiers. See the above example the specifier በጣም “very” comes before the modifier እንደ እህቱ “like his sister” which is PP and the complement ሰው comes after all the constituents that are modifier and specifier.

Like VPs and NPs, AdjP can be simple or complex. Similar to complex NPs and complex VPs, AdjP also can be simple or complex. The above example is simple AdjP. Complex AdjP are phrases that contain embedded sentence. For example in the phrases [አልማዝ ከአስቴር የበለጠች] ቆንጆ, “Almaz is more beautiful than Aster” the word construction in the bracket is an embedded sentence. Thus, the phrase is complex.

2.4 Sentence Formation

In Amharic grammar, the groups of phrases that together express ideas are called a sentence. Sentences are constructed from simple or complex NP and simple or complex VP but NP always comes first as a subject.

Example: - *a. ሁለት ትልልቅ ልጆች “two big children”*

b. ትናንት በመኪና ወደ ጎጃም ሄዱ “went to Gojam by car yesterday”

c. ሁለት ትልልቅ ልጆች ትናንት በመኪና ወደ ጎጃም ሄዱ “two big children went to Gojam by car yesterday”

The first two constructions do not express full idea but the last one do. Because the last one expresses full information such as who did go to Gojam? Where did the two children go? And etc. all this questions have been answered by the last word construction. In the last construction there are NP and VP which build the sentence and these are NP ሁለት ትልልቅ ልጆች “two big children and VP ትናንት በመኪና ወደ ጎጃም ሄዱ”went to Gojam by car yesterday”

The remaining phrases (other than NP and VP) are being constructed in NPs or VPs that are found in a sentence Based on this construction, sentences can be simple or complex.

2.4.1 Simple sentence

Simple sentences are sentences, which contain only one verb. Simple sentence can be constructed from NP followed by VP which only contain single verb.

Example 1: -*A. አስቴር ብርጭቆውን ሰበረችው "Aster broke the glass"*

B. ሰበር- ኧች-ው "broke"

Here the sentence contains only one verb ሰበረችው "she broke". And in B the pronoun morpheme -ኧች indicates the subject of the sentence and the morpheme -ው indicates the object of the sentence. This sentence contains transitive verb ሰበረችው "broke" that takes only one object ብርጭቆውን "the glass".

Example 2: - *አስቴር ለካሳ መጽሀፍ ሰጠችው "Aster gave Kassa a book"*

Here also the sentence contains only one verb ሰጠችው "gave" so it is simple sentence. The difference in this example from the previous one is the sentence here contains transitive verb ሰጠችው "gave" with two objects ለካሳ "Kassa" and መጽሀፍ "book".

Example 3: - *ካሳ ወደ ትምህርት ቤት ሄደ "Kassa went to school"*

Like the above examples, sentence of example 3 also consist one verb that is ሄደ "(he) went". This sentence contains intransitive verb ሄደ "(he) went" that do not take objects.
Generally, all the above stated examples are simple sentences that contain different types of verbs. Simple sentences may contain Intransitive verbs, transitive verbs with one object and transitive verbs with two objects

Simple sentences can be declarative sentences, interrogative sentence and imperative sentences. All these types of sentences are discussed below.

2.4.1.1 Declarative sentences

In contrast to command, question or exclamation, if the sentence is statement it is declarative sentence. In Amharic, Declarative sentences always ends with the Amharic punctuation mark "።"(See Appendix 2) which is equivalent of period (.) in English. They are used to convey

information. Declarative sentences can be positive or negative sentences. Negative sentences simply negate a declarative statement made about something.

Example 1: - *አስቴር ትምህርት ቤት ውስጥ ነች "Aster is at school"*
Here the sentence is declarative because it describes where aster is.

Example 2: - *ካሳ በግ አልገዛም "Kassa didn't buy a sheep"*
In this example the sentence is negative declarative sentence. The verb አልገዛም "did not buy" is negated by the prefix አል-"not"

2.4.1.2 Interrogative sentences

In Amharic Interrogative sentences are sentences that can form a question. The question can be the one that ask the known thing to be sure or the one that asks the unknown one. These types of sentences always end with question mark punctuation which is symboled as "?"(See Appendix 2).

Example 1: - *ማን ጫማ ገዛልህ? "Who did buy the shoes for you?"*

In the above example, the question is for the unknown thing just to get full information about it. This type of question or interrogative sentences consist interrogative pronouns which are ማን "who", መቼ "when", ምን "what", ስንት "how many", የት "where", etc.

Example 2: - *ካሳ መጣ እንዴ? (Did kassa come?)*

In this example, the inquisitor knows that Kassa would come but does not know whether he came or not.

2.4.1.3 Imperative sentences

When someone wants to pass instruction or commands, imperative sentences can be used. Most of the time, the subject of imperative sentences are second person pronouns. But when the command is passed for the third person the subject of the sentence can be third person pronouns or nouns.
Example: - *ወጥ አምጪ (bring wat)*
Here the subject is (you) second person feminine singular

Example: - *ካሳ ልብስ ይጠብ (Kassa, wash clothes)*

Here the command is for the third person that does not exist at the time the order is transferred. So the subject is (he) third person singular masculine.

2.4.2 Complex Sentences

As described above, sentences are composed of one or more NPs and VPs. These phrases can be simple or complex. Complex sentences are sentences that contain at least one complex NP or complex VP or both complex NP and complex VP. Complex NPs are phrases that contain at least one embedded sentence in the phrase construction. The embedded sentence can be complements. See the following examples.

Example 1: - *[ካሳ የገባበት የሳር ቤት] በጣም ትልቅ ነው "the thatched house that Kassa has entered is so big"*

Here the head of the noun phrase [ካሳ የገባበት የሳር ቤት] "that"is ቤት "the thatched house that Kassa has entered". The head with the complement የሳር "thatched" form simple noun phrase የሳር ቤት "thatched house" and this noun phrase has been combined with the embedded sentence or clause ካሳ የገባበት "that Kassa has entered" to form complex noun phrase. But the clause that makes the complex phrase is dependent which is identified by the morpheme የ "that".

Example 2: - *[ካሳ የገዛው መጽሀፍ] ዛሬ ጠፉ"the book that Kassa has bought is lost today"*

Here the NP is found in the bracket and the head መጽሀፍ "the book"has combined with the dependent clause ካሳ የገዛው "that Kassa has bought" to form complex NP which is found in the bracket. The relativizer የ **"**that**"** in the dependent clause indicates that the clause is a subordinate clause and it cannot stand alone

Likewise, complex VPs is complex if they contain at least one sentence or more than one verb. Similar to complex NPs, complex VPs also contain dependent clauses or sentences. These clauses can be complement or modifier.

Example 3: - *አስቴር [[ካሳ መኪና እንደገዛ] ሰማች] "Aster heard that Kassa has bought a car"*
In this sentence, the phrase in the first bracket is VP and the embedded sentence or dependent clause is in the second bracket. In this example the dependent clause is used as a complement in the VP construction. The prefix እንደ- in the clause indicates that the clause is dependent

Example 4: - *አስቴር [[ካሳ ወደ ጎጃም ሰለሄደ] አለቀሰች] "Aster wept for the reason that Kassa went to Gojam."*
Here the word construction in the second bracket is dependent clause which is identified by the prefix ሰለ- that founds in the verb. This clause is a modifier of the phrase.
Generally, we can understand from all these examples that embedded sentences or dependent clauses in the VPs can be modifiers or complements.

Moreover, more than one embedded sentences or dependent clauses can be found in a complex VP. See the following example.

Example 5: - *ካሳ [ከጎጃም እንደመጣ] [አስቴር ወደ ናዝሬት እንደ ሄደች] ሰማ "when Kassa came from Gojam he heard that Aster went to Nazret"*
In this example the VP is ከጎጃም እንደመጣ አስቴር ወደ ናዝሬት እንደ ሄደች ሰማ "came from Gojam he heard that Aster went to Nazret" and there are two dependent sentences in it: ከጎጃም እንደመጣ "came from Gojam " and አስቴር ወደ ናዝሬት እንደ ሄደች "that Aster went to Nazret". These sentences are to be used in this phrase construction as a modifier and complement respectively to their order.

Unlike the above examples, Complex sentences also may contain both complex NP and complex VP. Some examples of complex sentences which contain both complex NP and complex VP are given below.

Example 6: - ከጎጃም የመጣችው ልጅ ካሳ እንደወደዳት አወቀች "The girl who came from Gojam knew well that Kassa is in love with her"
Here is a complex sentence that contains complex NP and VP which are ከጎጃም የመጣችው ልጅ "The girl who came from Gojam "and ካሳ እንደወደዳት አወቀች "knew well that Kassa is in love with her", respectively.

Generally speaking, simple sentences are composed of simple noun phrases and simple verb phrases whereas complex sentences can be composed of complex NP and simple VP, simple NP and complex VP or both complex NP and VP.

2.5 Conclusion

To conclude, this chapter broadly speaks about the categories of word classes based on early and recent scholars. However, scholars classified word classes into eight and five, for the purpose of this study words are classified into six. These are noun, verb, adjective, adverb, preposition and conjunction. In this study pronouns are treated as nouns. This categorization for pronouns is forwarded by the scholar Baye [33].

Five types of Amharic phrases, which are classified by Baye [33], have been discussed in this chapter. Such phrases are noun phrases, verb phrases, adjectival phrases, prepositional phrases and adverbial phrases. In the other hand, types of sentences as classified by Getahun [22] also were adopted for the purpose of this study which are simple and complex sentences. Different types of the simple sentences also were identified and discussed briefly in this chapter. According to Getahun simple sentences categorized into three, namely declarative sentences, interrogative sentences and imperative sentences. Finally ways of formation of complex sentences was indicated.

CHAPTER THREE

LITERATURE REVIEW

3.1 Introduction

As indicated in the first chapter, determining different types of phrases in a sentence is very useful task for varieties of NLP applications. The process of determining theses phrases is called chunking. Chunking is the process of first identifying proper chunks from a sequence of words, and then classifying these chunks into some grammatical classes. Broadly speaking, Tasks of chunking are extracting the non-overlapping segments from a stream of data and identifying them with non-recursive cores of various types of phrases. In such a way that syntactically related words are grouped in the same phrase. These identified phrases are the bottom part of the parser so a group of chunkers can form parse. So chunking is considered to be the preprocessing stage that may facilitate the full parsing of sentences of a certain language. This task has already been proven using different methods. These methods are discussed below.

3.2 Methods of chunking

3.2.1 Rule-based chunking

Rule-based chunking is the most uncomplicated and cheap. A set of static rules are defined to chunk the text. This method is an approach based on regular expression rules developed by a human, Often these rules are described in terms of a grammar or finite state automaton (FSA). The advantages of rule based chunking are extremely simple to implement and it does not require training corpus. Some examples of rule based chunking for other languages are rule based chunking using XML [28] and Text chunking using transformation-based learning [20].

Grover [28] indicates how to develop a chunker which is reusable and configurable to different chunking styles. The main tools that are being used in this study are LT-XML2 and LT-TTT2. The chunking pipeline in this study only recognizes noun and verb groups.

The data is being used for training this system is CoNLL[5] data. The result of the research is 89.1% precision and 88.57% recall for noun group and 88.10% precision and 91.86% recall for verb group for English .

Abney [4] introduce FSA for partial parsing. In this study FSA is expressed by using regular grammar. See the following example how FSA of Abney works[6].

$$S_0 \rightarrow dS_1 \mid aS_1 \mid nNP \mid pNP \mid dNP$$
$$S_1 \rightarrow aS_1 \mid nNP \mid pNP$$
$$NP \rightarrow rcNP \mid ppNP$$

Here the regular grammar is for noun phrase detection. S_0, S_1 and NP are non terminal symbols and transitions whereas the others are terminal such as d, a, n, p which are determinant, adjective, noun, pronoun respectively. As well as, rc and pp stand for relative clause and prepositional phrase. Speaking broadly, the determinant d and adjective a lead to the first transition S_1 and non terminals noun n, pronoun p and determinant d lead to transition NP.

While in transition S1 non terminal adjective a loop back and stay in S1 along with rc and pp show the way to NP. It continues in this fashion till the correct NP found. This technique which is used by Abney was able to achieve high accuracy.

The cons of rule based chunking are

- Rules are not sufficiently complete; they cannot predict the subtleties and exceptions that occur naturally in language.
- Very difficult to generate new rules, without the skills of an accomplished linguist.

3.2.2 Statistical chunking

The shortcomings of rule based chunking are solved by the statistical methods. Statistical or machine learning methods can use supervised or unsupervised learning approaches. The

[5]Data provided for Conference on Computational Natural Language Learning (CoNLL)

[6]The example is taken from the paper of St.Charles that is noun phrase extraction

supervised learning approach is the type of corpora which is annotated whereas; the unsupervised approach is the natural corpora as it found from news or somewhere else. For example the text with POS tagged is a sample of supervised approach. The major problem of using supervise approach is lack of manually chunked text (corpora). Some statistical chunking methods are discussed.

3.2.2.1 Maximum Entropy

Maximum entropy (ME) learners use a statistical method to predict which possibility is the most likely, that is, whether the current word begins, falls within, or falls outside a noun phrase. MEM is an exponential model that offers the flexibility of integrating multiple sources of knowledge into a model [42]. One of the main advantages of using MEM is the ability to incorporate various features into the conditional probability framework and as compared to other models is that it potentially tags words which have never been seen in the training data. The framework estimates probabilities based on the constraints derived from the training data and making least number of assumptions possible.

Koeling [20] used maximum entropy model for chunking task. The first step in this task is implementing classifiers to tag every word in a sentence with chunk tag by using local lexical information. Information sources which were be used for prediction chunk tag are current word, POS tag of current word, surrounding words and POS tags of surrounding words. In specifically for the surrounding word limitation only three words from left and two words from right side were being used. Using ME model the recall and precision were 91.86% and 92.08% respectively.

3.2.2.2 Hidden Markov Model

Hidden Markov Model (HMM) is a statistical structure with stochastic transitions and observations [39]. It can be used to solve classification problems involved in modeling sequential data. This statistical model can be applied as long as the system being modeled possesses the hidden Markov property. A system with the hidden Markov property must have a discrete number of possible states, and the probability distribution of future states must depend only on the present state and be completely independent of past states. These states are not directly observable. Li [38] proposed the Chinese chunking model based on conventional HMM.

Singh [31] proposed HMM for Hindi text chunking. This chunker took a text with POS tags as its input and gave marked chunk boundaries in its output. The main tasks of this word were: the first one was identifying the chunk boundaries and the second was labeling the chunks with their syntactic categories. The symbol of chunk tag sets which was used by this work is 2- tag scheme {STRT and CNT}, 3-tag scheme {STRT, CNT AND STP} and 4-tag scheme {STRT, CNT, STP and STRT_STP} where STRT stands for start, CNT stands for continuation, STP stands for stop and STRT_STP stands for start stop. STRT indicates that the token is the start of the chunk; CNT indicated that the token lies in the middle of a chunk; STP points toward the token lies at the end of a chunk and STRT_STP indicates the token lies in a chunk of its own. In this study, different types of input token were used which were only words, only POS tag, word_POS tag and Postag_word. The chunker was tested on 20,000 words of testing data and 92% precision with 100% recall achieved for chunk boundaries.

3.2.2.3 Conditional Random Field

Conditional Random Field (CRF) machine learners are actually quite similar to Hidden Markov learners. The primary advantage of CRF's over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMM's. The transition probabilities of the HMM have been transformed into feature functions that are conditional upon the input sequence.

Sha presented how to train CRF to achieve high performance on CoNLL task for noun phrase chunking. The result of his NP chunking has been reported on the modified CoNLL-2000 version of Tjong Kim Sang and Buchholz [5]. The test set he has used consists of Wall Street Journal (WSJ) section 21 tagged with the Brill POS tagger. The system achieves 94.38% F score.

3.2.2.4 Support Vector Machine

Support vector machines (SVM) are the most complex machine learning technique explored in this paper, however they are also the most accurate and computationally inexpensive. SVM's are binary classifiers. This means that they are best used to separate one class of items from another. In the NP extraction case, they can be used to separate noun phrases from non-noun phrases.

Kudo and Matsumoto [30] implemented SVM learning technique to identify English base phrases. They also applied weighted voting of 8 SVM based system to achieve higher accuracy. They derived weighting strategy from theoretical basis if the SVM for the weighted voting systems. They have used three annotated corpora for their experiments. The base NP standard dataset and base NP large data set which consists sections (15-18) and sections (02-21) of WSJ part of the Penn Treebank for the training data and section 20 and section 00 for the testing data, which are used for the noun phrase identification, respectively. The chunking data set that is used for all types of phrase identification also consists of sections (15-18) of the WSJ part of Penn Treebank for the training data and section 20 for the test data. As they reported their approach achieves 94.15% precision and 94.29% recall for baseNP-S data set, 95.62% precision and 95.93% recall for baseNP-L data set and 93.89% precision and 93.92% recall for the chunking dataset.

3.2.3 Hybrid chunking

Park et al [40] in their paper described a new approach of chunking for Korean language which is hybrid approach. Initially, the rule based chunking is done. Memory based learning technique is used for the correction of errors, which were exceptions to rules.

Ramshaw et al [20] have introduced transformation based learning for locating chunks in the tagged text as tagging problem as of Eric Brill's who used transformation based learning for tagging with high accuracy. For the chunking purpose they use IOB chunking specification. The learning process in this study is based on template rules. The first step is derivation of rules, second is scoring of rules, and third is selection of one rule with maximal positive effect. The process is iterative. This technique achieved precision and recall of 88% for complex chunks and 92% for baseNP.

3.3 Representation of the chunk structure and boundary

3.3.1 Chunking specification

There are many decisions to be made about where the boundaries of a group should lie and, as a consequence, there are many different 'styles' of chunking. There are different types of chunk tags and chunk boundary identification. The tag of each chunk type can be noun phrases, verb phrases, adjectival phrases, etc as of the target natural language construction rule. Nevertheless In order to

identify the boundaries of each chunk in sentences, there are five boundary types. These are IOB1, IOB2, IOE1, IOE2, IO, [and] [20].

The first four formats are complete chunk representation which can identify the beginning and ending of phrases. All use I tag for words that are inside a phrase and an O tag for words that are outside a phrase. They differ in their treatment of chunk-initial and chunk-final words

- IOB1- the first word inside a phrase immediately following another phrase receives a B tag
- IOB2- all phrases- initial words receive a B tag
- IOE1- the final word inside a phrase immediate preceding another same phrase type receives an E tag
- IOE2- all phrases- final words receive an E tag

Whereas, the last three are partial chunk representation

- IO- words inside a phrase receive an I tag, others receive an O tag
- [all phrase-initial words receive [tag other words receive . tag
-] all phrase-final words receive] tag and other words receive. Tag

See the following example with the chunk representation in table 2.1

ሁለቱ ልጆች በትልቅ መኪና ወደ ጎጃም ሄዱ “the two children went to Gojam by a big car”

Table 3.1: Representations of the chunk structure of the given sentence

	ሁለቱ	ልጆች	በትልቅ	መኪና	ወደ	ጎጃም	ሄዱ
IOB1	I-NP	I-NP	B-NP	I-NP	I-PP	I-PP	O
IOB2	B-NP	I-NP	B-NP	I-NP	B-PP	I-PP	O
IOE1	I-NP	E-NP	I-NP	I-NP	I-PP	I-PP	O
IOE2	I-NP	E-NP	I-NP	E-NP	I-PP	E-PP	O
IO	I-NP	I-NP	I-NP	I-NP	I-PP	I-PP	O
[ ]	[-NP	.	[	.	[	.	.
]	.	]	.	]	.	]	.

3.4 Strategies of parsing

Parsing is the process of assigning syntactic (and semantic) structures to input strings, according to a grammar. As indicated earlier parsing algorithm can be described as a procedure that searches through various ways of combining grammatical rules to find a combination that generates a tree that could be the structure of the input sentence [41].

3.4.1 Top down parsing

Top-down parsing starts with the symbol S and then searches through different ways to rewrite the symbols until the input sentence is generated. Top down parsing begins with the start symbol (usually a sentence S) and applies the grammar rules forward until the symbols at the terminals of the tree correspond to the components of the sentence being parsed.

In the top-down approach, a parser tries to derive the given string from the start symbol by rewriting non terminals one by one using productions. The non terminal on the left hand side of a production is replaced by it right hand side in the string being parsed.

3.4.2 Bottom up parsing

Bottom-up parsing starts with words in a sentence and uses production rules backward to reduce the sequence of symbols until it consists solely of S. Bottom-up parsing begins with the sentence to be parsed and applies the grammar rules backward until a single tree whose terminals are the words of the sentence and whose top node is the start symbol (usually S, for sentence) has been produced. In the bottom-up approach, a parser tries to reduce the given string to the start symbol step by step using productions. The right hand side of a production found in the string being parsed is replaced by its left hand side.

Atelach [15] implemented the Inside Outside algorithm with a bottom up chart parsing strategy. The probabilistic context free grammar has been used as a grammatical formalism to represent the phrase structure rules of the language. In this study only four word length sentences are considered and the others are recommended for further research.

CHAPTER FOUR

DESIGN OF CHUNKER AND PARSER FOR AMHARIC LANGUAGE

4.1 Introduction

In this chapter, data preparation and the model that is developed in this study are discussed. The syntactic property of Amharic discussed in chapter three is used in the design of the work.

The first section in this chapter begins by discussing the sample text that is being used in the study. The chunking specification used in this work with the chunk tag set is briefly discussed in the next section. Architecture of the study also is discussed in this chapter. Approaches of designing chunker and parser and also algorithm that is used to transform the chunker to parser are discussed in section five. The last section concludes and summarizes the whole chapter.

4.2 The sample corpus and their preparation for the system

For the current work, part of speech (POS) tagged corpus containing 320 sentences is used that were collected from two widely used Amharic grammar books, research papers and WIC. Still the size of the corpus is small due to lack of annotated text especially for the training set. The sentences in the training set must be chunk tagged sentences (See appendix 7). To the best knowledge of the researchers, still chunk tagged documents do not exist for Amharic. For this reason, sentences are chunk tagged manually.

Sentences that are collected from the research of Daniel and WIC were already tagged manually by the researchers and linguistic experts. Whereas, sentences that are gathered from the research of Atelach tagged using a POS tagger developed by Mesfin [24].

The main challenge in the collected corpus is the difference of the POS tag name. Daniel and Atelach use the same POS tag name but the WIC's is dissimilar from them (see Appendix 3 and Appendix 4). Unlike Atelach and Daniel, the tagging task of WIC corpus is word by word which didn't treat more than one word as one to assign towards the POS tag. For instance, in the case of Atelach and Daniel ጥቁርና ነጭ treated as one word and is assigned to JC tag name, which is an adjective not separated from a conjunction, but in the case of the sentences of WIC they are

considered as two words which are ጥቁርና AdjC ነጭ Adj. With the exception of this type of problem, the other tagging process is the same in both cases excluding the name. Adjective is symboled as ADJ tag name in the annotated sentences of WIC but J for sentences that are found from researches of Atelach and Daniel. Finally the tag names, which are used in WIC corpus, are used in this study except NP, VP and ADJP which are replaced by Nprep, Vprep and Adjprep (See Appendix 4). The reason for this replacement is NP, VP and ADJP are used to represent phrases in this study like NP for Noun Phrase, VP for Verb Phrase and ADJP for adjectival phrase. Moreover, the sentences collected from Atelach and Daniel are re tagged with the tag name of WIC by the researchers only by changing the tag name for the consistency of the document (See Appendix 6). See table 4.1. Beside this the data collected form WIC needed some modification to be compatible to the system. Thus, some texts that are not either POS of the word or normal text were removed from the corpus such as <title>, <document>, etc and only some sentences, which are taken randomly, were used.

Even if there are researches that are conducted in the area of POS tagger for Amharic language, for some reasons it was too difficult to get the system. Thus the rest of the sentences that are collected from Amharic grammar books, were hand tagged by the researchers and comments and suggestion are taken from linguistic experts.

Generally, for the tagged corpus 31 tags are used to tag the data and 13 chunk tags are used to identify the begging and ending of the chunk (see Appendix 4 and Appendix 5). These sample sentences were transcribed according to the Amharic ፊደል (Fidel or alphabet) Unicode standard (See Appendix 1). All types of simple and complex declarative sentences are included in the corpus. In this study all possible combinations of complex sentences are included.

Table 4.1: substitution of the old tag name with new

The old tag name	The new tag name
NV	N
NB	N
J	ADJ
VCO	V
C	CONJ
JC	ADJC
JP	ADJPREP
REL	VREL
ITJ	INT
ORD	NUMOR
CRD	NUMCR

The collected data is classified into two for the training and testing purposes. As illustrated above, the training dataset that are taken from the corpus were hand chunked by the researchers according to the phrase construction rule of Amharic language. Then, comments and suggestions were taken from Amharic language professionals.

4.3 Chunking specification

To make clear the specification, a chunk is a constituent whose children are pre terminals. In this study, there are six different kinds of chunk in Amharic languages namely, noun phrase (NP), verb phrase (VP), Adjective phrase (AdjP), Adverb phrase (AdvP), prepositional phrase (PP) and sentence (S). Table 4.2 shows the tag of each chunk type.

Table 4.2 the tag of chunk

Chunk description Type	Chunk Tag Name	Example
Noun phrase chunk	NP	ትላንት የተገዛው ኮት “the coat that has bought yesterday”
Verb phrase chunk	VP	ስለታመመ በመኪና ወደ ሆስፒታል ሄደ “(he) went to hospital by car because he is sick”
Prepositional phrase chunk	PP	ከወንድሙ ጋር “with his brother”
Adverbial phrase chunk	AdvP	እንደ አባቱ ክፉኛ “like his father severley”
Adjectival phrase chunk	AdjP	እንደ እህቱ ሰው ፈሪ “(he is) shy like his sister”
Sentence chunk	S	ደራሲው የሳለውን “that the artist has painted”

To identify the chunks, it is necessary to find the positions where a chunk can end and a new chunk can begin. The POS tag assigned to every token is used to discover these positions. As discussed in chapter two, there are four kinds of complete chunk boundary representations namely IOB1, IOB2, IOE1 and IOE2. In this study, to identify the boundaries of each chunk in sentences the IOB2 tag set is used for chunk tagged annotated text. Here I is a token inside a chunk, o is a token outside a chunk and B tag is a token that exists at the beginning of the chunk. See the following example ካሳ ያመጣው ትንሽ ልጅ እንደ አባቱ በጣም ታመመ “the little boy that kassa has brought is very sick like his father” with the chunk representation in Table 4.3.

Table 4.3 Example of IOB2 chunk representation

	IOB2
ካሳ	B-S
ያመጣው	I-S
ትንሽ	B-NP
ልጅ	I-NP
እንደ	B-PP
አባቱ	I-PP
በጣም	B-VP
ታመመ	I-VP

It is same as IOB1, except that a B tag is given for every token, which exists at the beginning of the chunk. Using the chunk type and IOB2 tag set, 13 phrase tags were used. These are B-NP, I-NP, B-VP, I-VP, B-PP, I-PP, B-ADJP, I-ADJP, B-ADVP, I-ADVP, B-S, I-S and O (See Appendix 5).

The followings are some example sentences of chunk tagged. For more see Appendix 7.

Example 1:- ሁለቱ NUMCR O ትልልቅ ADJ B-NP ልጆች N I-NP በመኪና NPREP O ወደ Prep B- PP ጎጃም N I-NP ሄዱ V O

Example 2:- የኢትዮጵያ ADJ B-NP ጠላቶች N B-NP ሀገሪቱ N O በድርጅቱ NPREP B-PP ውስጥ Prep I-PP የተሰጣትን V O ቦታ N O ተቃወሙ V O

Example 3:- ንጉሱ N O ፋሺስቶች N B-S የተርበደበዱበትን V I-S ጀግና ADJ B-NP ሰው N I-NP ሰቀሉ V O

Example 4:- ኢትዮጵያ N O ፈንጣጣ N B-S የተወገደበትን V I-S እለት N O ትላንት ADV B-VP አከበረች <V> I-VP

4.4 Architecture of the System

This sub-section elaborates the overall architecture of the system. The system has two phases the training phase and the testing phase. In the training phase, the system first accepts words with POS tags and chunk tagged. Then, the hidden markov model is trained with this training set. Likewise in the test phase, the system accepts words with POS tags. Finally the system outputs appropriate chunk tag sequences against each POS tag using HMM model and the decoding algorithm that is Viterbi algorithm to select the best path. Figure 4.1 shows the architecture of the chunker.

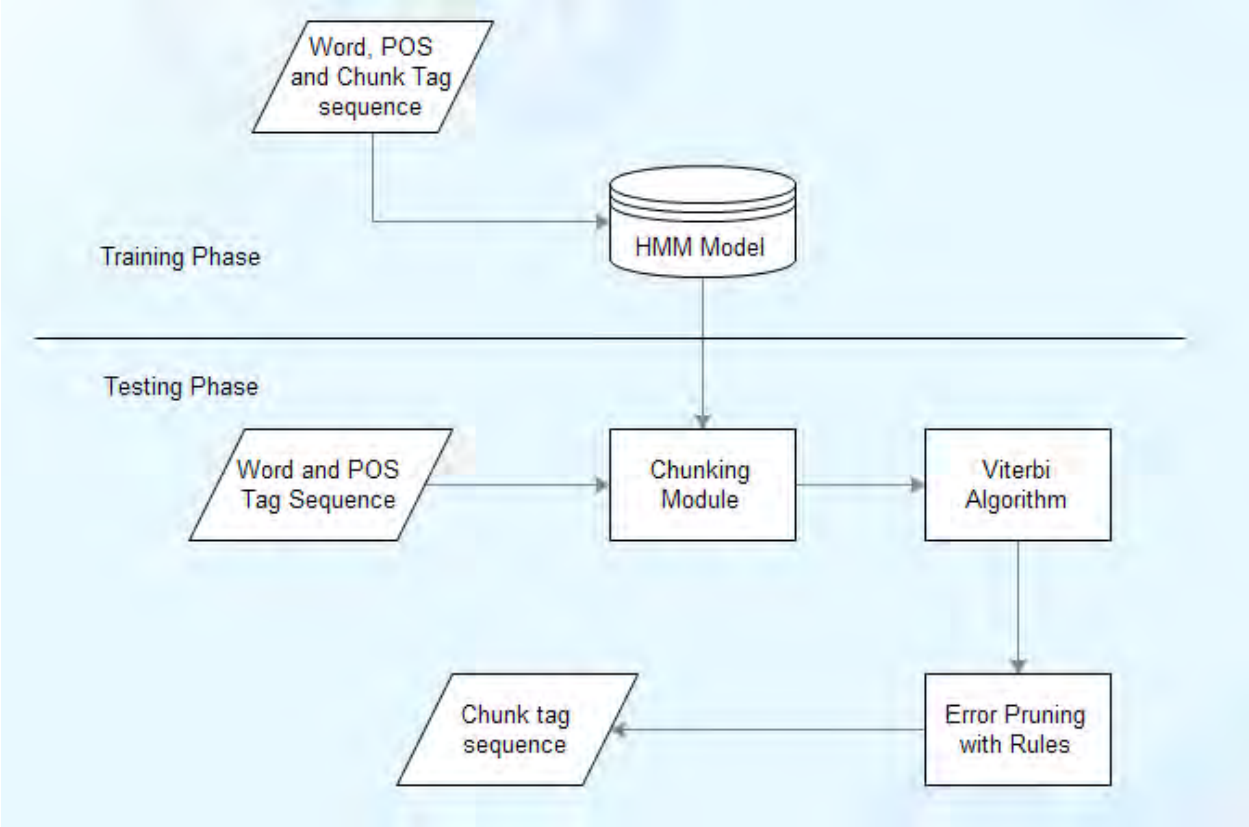

Figure 4.1 the Architecture of the chunker

Figure 4.2 elaborates the architecture of the chunker and the work flow of using the output of the chunker for parser. The output of the chunker is the sequence of words which are grouped syntactically correlated (phrase). In this sequence if base phrase is different from prepositional phrase and subordinate clause exists, only the head of the phrase will pass to the next step. Otherwise the PP ans S will take the word next to them to form the new phrase by taking the new word as a head of the newly formed phrase. The tagged head with other tagged words also are taken by the chunker for further chunking starting from finding base phrases from the new data stream. Then repeat the above mentioned process.

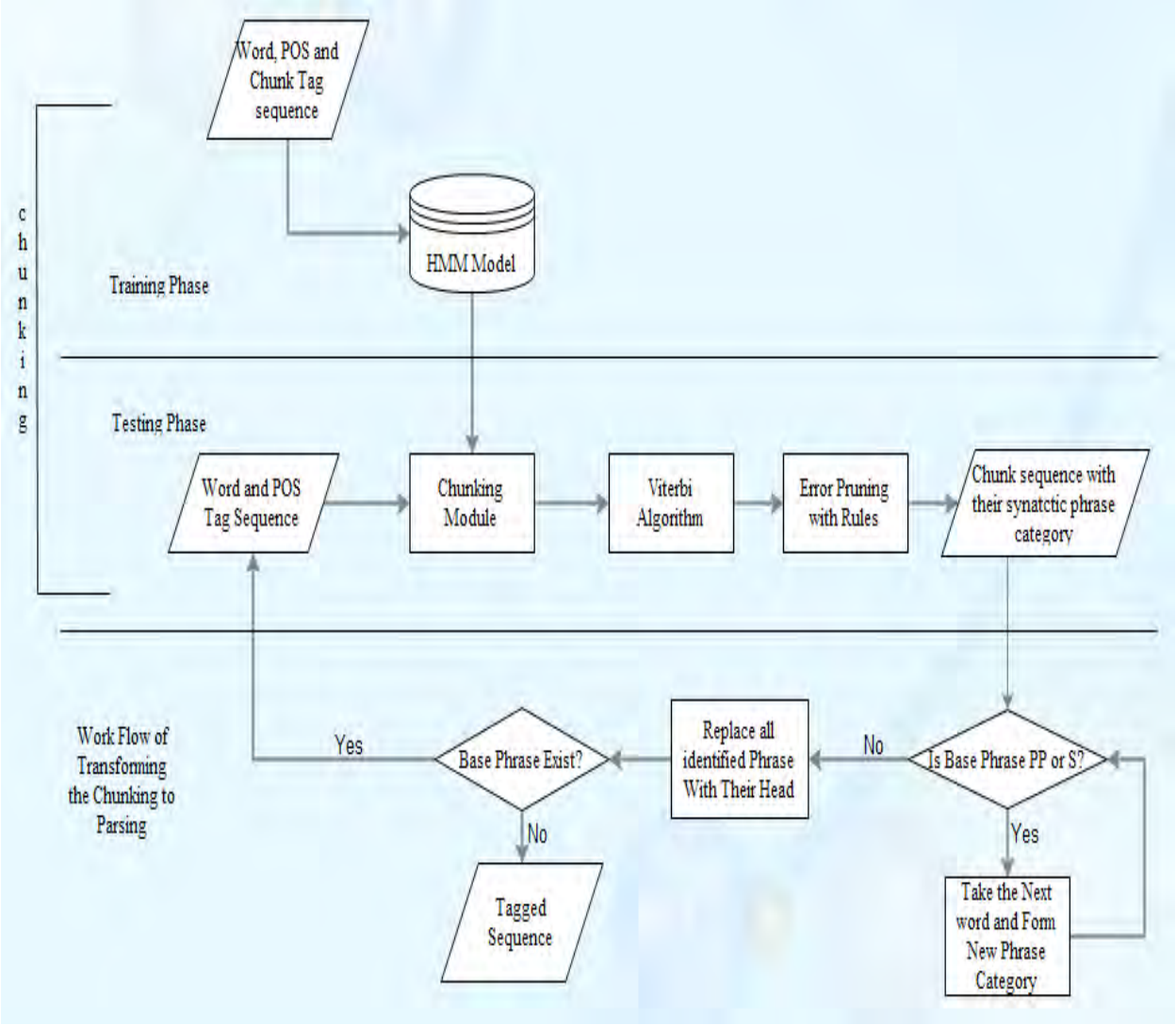

Figure 4.2 The overall Architecture of the system

4.5 Approach of the chunker and parser

4.5.1 Chunking module

In this work chunking treated as a tagging problem and statistical chunking is used then some rules are used to group chunking labels. First POS annotated corpus is prepared for statistical model (see Appendix 6). POS tags are Input of the system and IOB tags are the output.

Given the sequence of corresponding POS tags $T^n = (t_1, t_2, \ldots, t_n)$, ti $\in$ T where T is the POS tag set and C is a sequence of c_1 to c_n chunk tags. So, the problem is to get chunk tag sequence (C) given the sequence of POS tag sequence (T). The probabilistic model for this problem is as under:

$$\boldsymbol{C = \underset{c}{argmax}\, P(C \mid T)} \qquad (1)$$

Applying Bayes' rule :

$$\boldsymbol{C = \underset{C}{argmax}\, \frac{P(T|C)P(C)}{P(T)}} \qquad (2)$$

Simplify by dropping the denominator because it is constant for all sequences. So it can be

$$\boldsymbol{C = \underset{c}{argmax}\, P(T|C)\, P(C)} \qquad (3)$$

For each chunk tag sequence calculate the product of the probability of the tag sequence given the chunk tag sequence (likelihood) and the probability of the tag sequence (prior probability) is still difficult. Using Markov assumption, the whole Chunk tag sequence is estimated using Trigrams, and likelihood is also simplified such that a POS tag t_i depends only on corresponding chunk tag c_i. Hence, two simplifying assumptions make it possible to estimate the probability of chunk tag sequences given tag sequences.

$$\text{Emission Probabilities} = P(t_i \mid c_i) \qquad (4)$$

$$\text{Transition Probabilities} = P(c_i \mid c_{i-2}, c_{i-1}) \qquad (5)$$

The emission probability is the probability of a tag only dependent on its own chunk tag and transition probability is the probabilityof chunk tag dependent on the two previous chunk tags.

By Combining (4) and (5)

$$\arg\max_{c} \pi_{i=1}^{n} \mathbf{P}(\mathbf{t_i}|\mathbf{c_i}) \ \ \mathbf{P}(\mathbf{c_i}|\mathbf{c_{i-2}}\ \mathbf{c_{i-1}}) \quad (6)$$

For obtaining probability of $P(t_i \mid c_i)$ following equation is used:

$$\boldsymbol{P}(\mathbf{t_i}|\mathbf{c_i}) = \frac{\boldsymbol{count\ of}\ (\mathbf{t_i}, \mathbf{c_i})}{\boldsymbol{count\ of}\ \mathbf{c_i}} \quad (7)$$

For obtaining Trigram probability following equation is used:

$$\boldsymbol{P}(\mathbf{c_i}|\mathbf{c_{i-1}}\ \mathbf{c_{i-2}}) = \frac{\boldsymbol{count\ of}(\mathbf{c_{i-2}}, \mathbf{c_{i-1}}, \mathbf{c_i})}{\boldsymbol{count\ of}\ (\mathbf{c_{i-2}}, \mathbf{c_{i-1}})} \quad (8)$$

See the following example how HMM model works. First the system is trained with the chunk tagged training set i.e. to acquire frequency from the training set.

Suppose: - Chunk tags C={B-NP, I-NP, B-VP, I-VP, B-ADJP, I-ADJP,…}

Tags T={N, V, NPREP, VREL, VPREP,…}

Transition probability: P(B-S|O)=0.5, P(I-S| O B-S)=0.003, P(O| B-S I-S)=0.5, P(B-PP| I-S O)=0.05, P(I-PP|O B-PP)=0.1, P(O|B-PP I-PP)=0.3

Emission probability: P (N|O)=0.4, P(NPREP|B-S)=0.2, P(VREL|I-S)=0.02, P(N|O)=0.5, P(NPREP|B-PP)=0.005, P(PREP|I-PP)=0.5, P(V|O)=0.86

Initial probability: P_I (B-NP)=0.4, P_I (B-ADJP)=0.32, P_I (O)=0.5

Sentence 1 to chunk:- ወንበዴዎች N በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N ከጥቅም NPREP ውጭ PREP አደረጉት V

All possible chunk tagged sequence for sentences are calculated by multiplying all the number of the possible chunk tagged each tag has in a sentence. Example:- lets say in the sentence 1 the tag are 7 and N tag name has the possibility of 4 chunk tag name, NPREP has 3, VREL has 2, PREP has 2 and V has 3.The generated chunk tagged sequence for the sentence can be 1728. See the following sample chunk tagged sequences for sentence 1.

1. ወንበዴዎች (N, B-NP) በጎፈቃደኞች (NPREP, B-NP) የገነቡትን (VREL, I-S) ድርጅት (N, I-NP) ከጥቅም (NPREP, B-PP) ውጭ (PREP, I-PP) አደረጉት (V, O)
2. ወንበዴዎች (N, O) በጎፈቃደኞች (NPREP, B-PP) የገነቡትን (VREL, I-S) ድርጅት (N, B-NP) ከጥቅም (NPREP, B-S) ውጭ (PREP, B-NP) አደረጉት (V, O)
3. ወንበዴዎች (N, I-NP) በጎፈቃደኞች (NPREP, B-S) የገነቡትን (VREL, B-NP) ድርጅት (N, B-ADJP) ከጥቅም (NPREP, O) ውጭ (PREP, I-PP) አደረጉት (V, I-VP)
4. ወንበዴዎች (N, I-NP) በጎፈቃደኞች (NPREP, B-NP) የገነቡትን (VREL, B-NP) ድርጅት (N, I-NP) ከጥቅም (NPREP, B-NP) ውጭ (PREP, B-PP) አደረጉት (V, I-VP)
5. ወንበዴዎች (N, O) በጎፈቃደኞች (NPREP, B-S) የገነቡትን (VREL, O) ድርጅት (N, B-ADJP) ከጥቅም (NPREP, B-S) ውጭ (PREP, I-PP) አደረጉት (V, I-VP)
6. ወንበዴዎች (N, O) በጎፈቃደኞች (NPREP, B-S) የገነቡትን (VREL, I-S) ድርጅት (N, B-S) ከጥቅም (NPREP, I-NP) ውጭ (PREP, O) አደረጉት (V, B-VP)
7. ወንበዴዎች (N, O) በጎፈቃደኞች (NPREP, B-S) የገነቡትን (VREL, I-S) ድርጅት (N, I-NP) ከጥቅም (NPREP, O) ውጭ (PREP, I-PP) አደረጉት (V, O)
8. ወንበዴዎች (N, O) በጎፈቃደኞች (NPREP, B-S) የገነቡትን (VREL, I-S) ድርጅት (N, O) ከጥቅም (NPREP, B-PP) ውጭ (PREP, I-PP) አደረጉት (V, O)
9. ወንበዴዎች (N, B-NP) በጎፈቃደኞች (NPREP, B-NP) የገነቡትን (VREL, I-S) ድርጅት (N, I-NP) ከጥቅም (NPREP, B-PP) ውጭ (PREP, I-PP) አደረጉት (V, O)

The probability calculation for each chunk tagged sequence can be calculated as follows. Lets take the following chunk tagged sequence for probability calculation

ወንበዴዎች (N, O) በጎፈቃደኞች (NPREP, B-S) የገነቡትን (VREL, I-S) ድርጅት (N, O) ከጥቅም (NPREP, B-PP) ውጭ (PREP, I-PP) አደረጉት (V, O)

$\pi_{i=1}^{n}$ $P(t_i|c_i)$ $P(c_i|c_{i-2}\ c_{i-1})$ = (P (N|O)*P_I (O)) (P(NPREP|B-S)*P(B-S|O)) (P(VREL|I-S)*P(I-S| O B-S)) (P(N|O)*P(O| B-S I-S)) (P(NPREP|B-PP)*P(B-PP| I-S O)) (P(PREP|I-PP)*P(I-PP|O B-PP)) (P(V|O)* P(O|B-PP I-PP))

$\pi_{i=1}^{n}$ $P(t_i|c_i)$ $P(c_i|c_{i-2}\ c_{i-1})$=(0.4*0.5)(0.2*0.5)(0.02*0.003)(0.4*0.5)(0.005*0.05)(0.5*0.1)(0.86*0.3)

=0.000000000000504

Like this by calculating all the possible chunk tagged sequence choose the most probable tagging. The fundamental problem of HMM model is finding best path. Thus, the optimal sequence of chunk tags is found using Viterbi algorithm which uses parameters of HMM.

4.5.2 The Viterbi algorithm

The Viterbi algorithm is described as an algorithm which finds the most likely path, i.e. shortest path, given a set of hidden states (Viterbi, A.J, 1967).

Observed_tags = $tag_1 \ldots tag_T$

States = $q_0, q_1 \ldots q_N, q_F$

A = N x N matrix such that $a_{i,j}$ is the probability of the transition from q_i to q_j

B = lookup table such that $b_i(tag_t)$ is the probability that word t is assigned POS i

viterbi = (N+2) x T matrix # columns are states, rows are words

backpointer = (N+2) x T matrix # highest scoring previous cells for viterbi

for states q from 1 to N:

initialize viterbi[q,1] to a0,q * bq(tag1) # score transition 0→q given w1

initialize backpointer[q,1] to 0 (start state)

for tags tag from 2 to T:

for state q from 1 to N: # for T-1 x N (w,q) pairs

$$\text{Viterbi}[q, tag] \leftarrow \max_{Q' = 1}^{N} \text{Viterbi}[q', t-1] * a_{q',q} * b_q(tagt)$$ # score = maximum previous * prior * likelihood

$$\text{Backpointer}[q, tag] \leftarrow \operatorname{argmax}_{q' = 1}^{N} \text{Viterbi}[q', t-1] * a_{q',q}$$ # backpointer= maximum previous

$$\text{Viterbi}[qF, T] \leftarrow \max_{q = 1}^{N} \text{Viterbi}[q, t] * a_{q,qF}$$ # score=maximum previous* prior * likelihood

$$\text{Backpointer}[qF, T] \leftarrow \operatorname{argmax}_{q = 1}^{N} \text{Viterbi}[q, T] * a_{q,qF}$$ # backpointer = maximum previous

return (best_path) # derive by following backpointers from (qF,T) to q0.

The Viterbi algorithm is used to find the best path of the sequence.

Example፡- ወንበዴዎች N በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N ከጥቅም NPREP ውጭ PREP አደረጉት V.

Words with their tags	Chunk tags
ወንበዴዎች N	B-NP I-NP O B-PP
በጎፈቃደኞች NPREP	B-NP B-S I-NP B-VP O
የገነቡትን VREL	B-S O B-VP
ድርጅት N	B-NP I-NP O B-PP
ከጥቅም NPREP	B-NP B-S I-NP B-VP O
ውጭ PREP	B-PP I-PP
አደረጉት V	B-VP I-VP O

Staring with B-NP (for ወንበዴዎች N) and then move to five different chunk tag states NPREP has, B-NP B-S I-NP B-VP O (for በጎፈቃደኞች NPREP) then these states move to different three states to form different new states and it goes in this fashion up to the end of the sentence which is discussed in HMM. But when the Viterbi algorithm is applied in each state formation the probability of the newly formed states calculated and the highest is selected as follows P(B-NP,B-NP) P(O,B-NP) P(B-NP,B-S)(O,B-S)

1.4292309796895281e-08: ('ወንበዴዎች', ('N', 'B-NP')), ('በጎፈቃደኞች', ('NPREP', 'I-NP'))

1.4534804928687978e-05: ('ወንበዴዎች', ('N', 'I-NP')), ('በጎፈቃደኞች', ('NPREP', 'B-VP'))

0.001142325814831578: ('ወንበዴዎች', ('N', 'O')), ('በጎፈቃደኞች', ('NPREP', 'B-VP'))

8.286573383526292e-06: ('ወንበዴዎች', ('N', 'I-NP')), ('በጎፈቃደኞች', ('NPREP', 'B-S'))

7.785687195273561e-06: ('ወንበዴዎች', ('N', 'O')), ('በጎፈቃደኞች', ('NPREP', 'O'))

1.3303986494700062e-05: ('ወንበዴዎች', ('N', 'O')), ('በጎፈቃደኞች', ('NPREP', 'B-NP'))

1.413370243859975e-07: ('ወንበዴዎች', ('N', 'B-NP')), ('በጎፈቃደኞች', ('NPREP', 'I-NP'))

4.062050225601187e-07:('ወንበዴዎች', ('N', 'I-NP')), ('በጎፈቃደኞች', ('NPREP', 'O'))

3.331311031577693e-08: ('ወንበዴዎች', ('N', 'I-NP')), ('በጎፈቃደኞች', ('NPREP', 'O'))

8.95007205882016e-10: ('ወንበዴዎች', ('N', 'B-PP')), ('በጎፈቃደኞች', ('NPREP', 'I-NP'))

0.0018860806571361231: ('ወንበዴዎች', ('N', 'O')), ('በጎፈቃደኞች', ('NPREP', 'B-S'))

2.0219443377496156e-06: ('ወንበዴዎች', ('N', 'O')), ('በጎፈቃደኞች', ('NPREP', 'O'))

The maximum probability is selected for the next step which is 0.0018860806571361231 that represent ('ወንበዴዎች', ('N', 'O')), ('በጎፈቃደኞች', and ('NPREP, B-S')). This new state takes the possible chunk tags of the next tag and form the new one and calculate the probability of the new states finally take the highest probability i.e. [('ወንበዴዎች', ('N', 'O')), ('በጎፈቃደኞች', ('NPREP', 'B-S')), ('የገነቡትን', ('VREL', 'I-S'))] . The final output is [('ወንበዴዎች', 'N'), ('በጎፈቃደኞች NPREP የገነቡትን VREL', 'S'), ('ድርጅት', 'N'), ('ከጥቅም NPREP ውጭ PREP', 'PP'), ('አደረጉት', 'V')]. For more outputs see Appendix 8

4.5.3 Approach Selected To Transform the Chunker to Parser

There are different types of approach for parsing [15, 18] but using chunking is the fastest and robust way. In this study, bottom up approach is employed by using the output of the chunker recursively as an input for the second round through using the chunker itself.

The algorithm that used for the transformation is

1. Take the tagged document
2. Use a chunker for identifying base phrases
3. if the base phrases are verb phrase, noun phrases, adjectival phrase and adverbial phrase

   Replace all identified phrases with their head

   Else

   The current phrase takes the word next to it and makes new phrase by taking the new word as a head

4. Find base phrases in the new data stream
5. If previous step discovered new phrases

   Repeat steps 3-5

   Else

   Stop

In this algorithm only step 2 and 4 require training phase. The others are fixed. The rule for identifying the head word of the phrase in step 3 is just taking of the last word of the phrase[8].
For more clarification, see the following example with description.

Example 1: -

Step 1: ወንበዴዎች N በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N ከጥቅም NPREP ውጭ PREP አደረጉት V

Step 2: [('ወንበዴዎች', 'N'), ('በጎፈቃደኞች NPREP የገነቡትን VREL', 'S'), ('ድርጅት', 'N'), ('ከጥቅም NPREP ውጭ PREP', 'PP'), ('አደረጉት', 'V')]

Step 3: ['ወንበዴዎች N', ('በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N', 'NP'), ('ከጥቅም NPREP ውጭ PREP አደረጉት V', 'VP')]

Step 4: [('ወንበዴዎች', 'N'), ('ድርጅት N አደረጉት V', 'VP')]

Step 5: ['ወንበዴዎች N', 'አደረጉት V']

Step 1 is taking the tagged sentence for all over the process to be chunked. That means it is the first step of the algorithm. Step 2 is the first output of the chunker which identifies possible base phrases. In this example ('በጎፈቃደኞች NPREP የገነቡትን VREL', 'S') and ('ከጥቅም NPREP ውጭ PREP', 'PP') are the base phrases which is the second step in the algorithm. When goes to step three, the prepositional phrase and the Subordinate clause or Sentence goes to the next word and converted to verb phrase and noun phrase, respectively. This is the behavior of Amharic phrase structure for prepositional phrases and subordinate clauses. All the above discussed in step three is the third step in the algorithm. The sentence in step 4 is reorganized for the recursive process, here only the head of

[8] See Section 2.3

the verb phrase is taken. This organization is similar to the sentence in step 1. Thus, the recursive procedure, stated in the fifth step of the algorithm, starts here. Here are some more examples and also see Appendix 10.

Example 2:

Step 1: አበራ <N> ወንድሙ <N> የጋበዘውን <V> ነበዝ <ADJ> የኮሌጅ <NPREP> ተማሪ < N> ተዋወቀ <V>

Step 2: [('አበራ', 'N'), ('ወንድሙ N የጋበዘውን V', 'S'), ('ነበዝ', 'ADJ'), ('የኮሌጅ NPREP ተማሪ N', 'NP'), ('ተዋወቀ', 'V')]

Step 3: ['አበራ N', ('ወንድሙ N የጋበዘውን V ነበዝ ADJ', 'ADJP'), 'ተማሪ N', 'ተዋወቀ V']

Step 4: ['አበራ N', 'ነበዝ ADJ', 'ተማሪ N', 'ተዋወቀ V')]

Step 5: [('አበራ', 'N'), ('ነበዝ ADJ ተማሪ N', 'NP'), ('ተዋወቀ', 'V')]

Step 6: ['አበራ N', 'ተማሪ N', 'ተዋወቀ V']

Step 7: [('አበራ', 'N'), ('ተማሪ N ተዋወቀ V', 'VP')]

Step 8: ['አበራ N', 'ተዋወቀ V']

Example 3:

Step 1: ልጅቷ <N> ትልቅ <ADJ> ወንድሟ <N> እንዲያስጠናት <V> ክፍል <N> ውስጥ <P> ነገረችው <V>

Step 2: [('ልጅቷ', 'N'), ('ትልቅ ADJ ወንድሟ N', 'NP'), ('እንዲያስጠናት', 'V'), ('ክፍል N ውስጥ P', 'PP'), ('ነገረችው', 'V')]

Step 3: ['ልጅቷ N', 'ወንድሟ N', 'እንዲያስጠናት V', ('ክፍል N ውስጥ P ነገረችው V', 'VP')]

Step 4: ['ልጅቷ N', 'ወንድሟ N', 'እንዲያስጠናት V', 'ነገረችው V']

Step 5: [('ልጅቷ', 'N'), ('ወንድሟ N እንዲያስጠናት V', 'S'), ('ነገረችው', 'V')]

Step 6: ['ልጅቷ N', ('ወንድሟ N እንዲያስጠናት V ነገረችው V', 'VP')]

Step 7: [('ልጅቷ', 'N'), ('ነገረችው', 'V')]

Here in example 2 and example 3, the prepositional phrase (PP) and the dependent sentence (S) take the next word to form the new phrase by taking the new word as ahead. For instance, the dependent sentence (‘ወንድሙ N የጋበዘውን V’, ‘S’) takes the next word ('ነበዝ', 'ADJ') and form the new phrase ('ወንድሙ N የጋበዘውን V ነበዝ ADJ', 'ADJP') where the new word ('ነበዝ', 'ADJ') is the head of the phrase.

To draw the tree of the parser use the above transformation algorithm with a little bit modification i.e. in Step 3 don't only replace the identified phrases with their head but also catch the identified phrases and pass their head. The tree is just for sample. It is not included in this study. Examples how to trace the tree of the parser. See the figures

Example 1:- ወንበዴዎች N በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N ከጥቅም NPREP ውጭ PREP አደረጉት V

[('ወንበዴዎች', 'N'), ('በጎፈቃደኞች NPREP የገነቡትን VREL', 'S'), ('ድርጅት', 'N'), ('ከጥቅም NPREP ውጭ PREP', 'PP'), ('አደረጉት', 'V')]

['ወንበዴዎች N', ('በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N', 'NP'), ('ከጥቅም NPREP ውጭ PREP አደረጉት V', 'VP')]

[('ወንበዴዎች', 'N'), ('ድርጅት N አደረጉት V', 'VP')]

['ወንበዴዎች N', 'አደረጉት V']

All the phases here are, the first step is finding base phrases from the sentence. The base phrases that are found in this step are ('በጎፈቃደኞች NPREP የገነቡትን VREL', 'S') and ('ከጥቅም NPREP ውጭ PREP', 'PP'). Both of them take the next word and form new phrases that are noun phrase (('በጎፈቃደኞች NPREP የገነቡትን VREL ', 'S') ድርጅት N, 'NP') and verb phrase (('ከጥቅም NPREP ውጭ PREP', 'PP') አደረጉት V, 'VP'), respectively. next by catching the whole phrase and pass only the head word all the entire phrases are captured and the head word with other words are tested to find other base phrase in it. If the head is passed the new sentence can become ወንበዴዎች N ድርጅት N አደረጉት V. Then in this sentence one base phrase is detected which is ድርጅት N አደረጉት V, 'VP') from the new sentence chunked [('ወንበዴዎች', 'N'), ('ድርጅት N አደረጉት V', 'VP')]. If there exist new base phrase combine the captured phrases with the new one. So the new chunked sentence can be modified like [('ወንበዴዎች', 'N'), (('በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N', 'NP'('ከጥቅም NPREP ውጭ PREP አደረጉት V', 'VP'), 'VP')]. It follows the same fashion till the head of the noun phrase and the verb phrase remains.

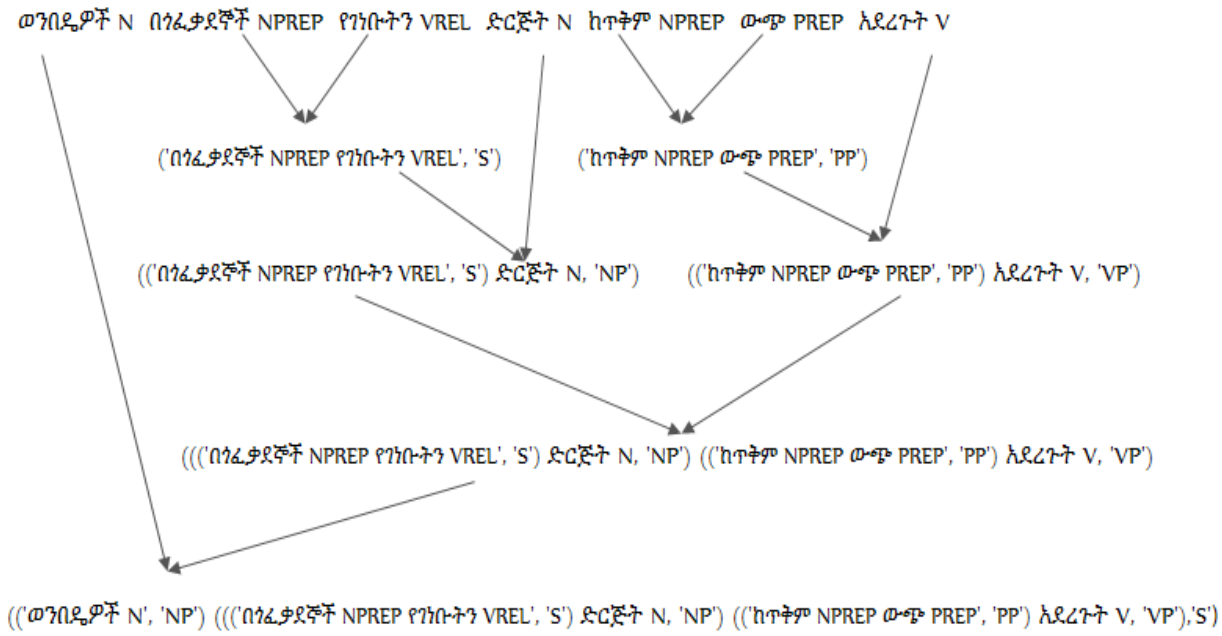

Figure 4.3 A bottom up parse structure for a sentence "ወንበዴዎች በጎፈቃደኞች የገነቡትን ድርጅት ከጥቅም ውጭ አደረጉት"

Example 2:- ኢማተልፈ N ያስገነባቸው VREL 17 NUMCR ትምህርትቤቶች N አገልግሎት N መስጠት V ጀመሩ V

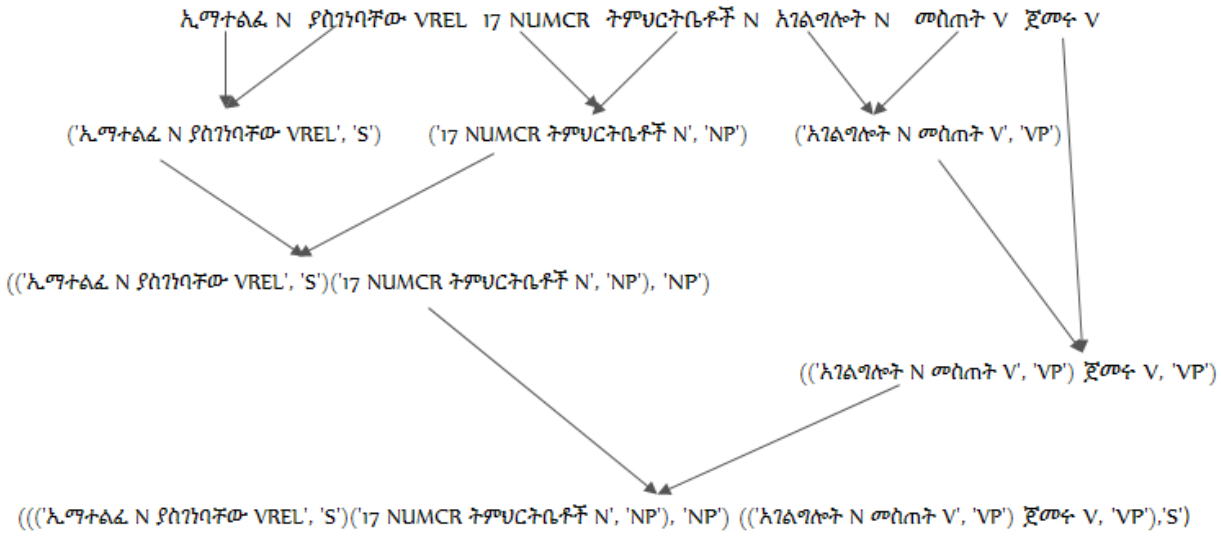

Figure 4.4 A bottom up parse structure for a sentence "ኢማተልፈ ያስገነባቸው 17 ትምህርትቤቶች አገልግሎት መስጠት ጀመሩ "

CHAPTER FIVE

EXPERIMENT AND RESULT

5.1 Experiment

The sample text selected, which was discussed in chapter four, was used for experimentation. Each sentence in the corpora had been tagged and hand chunked by the researcher, with comments and suggestions from linguists. The 320 sample sentences were selected randomly based on the distribution of the different phrase structures, using judgment sampling technique. The performance of Amharic chunking is measured in terms of the accuracy.

$$Accuracy = \frac{Number\ of\ correct\ \ chunk\ \ tagged\ \ sentences}{total\ \ number\ \ \ of\ sentences}$$

The sentences in the corpus are classified as training data set and testing data set using 10 fold cross validation. 10 fold cross validation (10FCV) is a system for testing trained classifiers. 10-Fold cross validation is a technique used to test how well a model adapts to fresh, previously unseen data. The procedure works like this:

1. Use a random sampling procedure to split the entire data set into 10 sub-samples which are used to test. Let's call these samples testing dataset 1, testing dataset 2, testing dataset 3 and so on, until we get to testing dataset 10.
2. As a first step, remove testing dataset 10 (T10) from the classified data set.
3. Train the machine learning algorithm using data from T1 to T9.
4. Once the machine has built a model based on data from T1 to T9, it sees how accurate the model predicts the unseen data of T10
5. Once the accuracy of predicting the values is tested in T10, T10 is putted back into the training set.
6. For the next step, remove testing dataset 9 (T9) from the training set.
7. Re-train the machine learning algorithm, this time using data from T1, T2, T3, T4, T5, T6, T7, T8, and T10 (i.e. leave out T9)

8. Once the machine has built a model based in the training set described in Step 7, it evaluates how accurately it can predict values in the new test set (i.e. T9)
9. Put T9 back into the training set.
10. Now, remove T8 from the training set, and repeat the testing procedure.

At the end of the sequence, the 10 results from the folds can be averaged to produce a single estimation of the model's predictive potential.

A big advantage of the 10-fold cross validation method is that all observations are used for both training and validation, and each observation is used for validation exactly once. This leads to a more accurate way to measure how efficiently the algorithm has "learned" a concept, based on training set data. Thus As outlined above the experiment for this study was carried out in ten phases.

5.2 Result

The following sections discuss the results on the obtained on the experiments carried on the Testing data Sets in different phases. It also reports the results of the test datasets.

5.2.1. Result on Test Dataset 1

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed to correct the output of the statistical

Table 5.1 Result of the chunker before applying rules on Test dataset 1

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 1	2	93.75%

Table 5.2 Result of the chunker after applying rules on Test dataset 1

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 1	1	96.875%

5.2.2. Result on Test Dataset 2

The result of the chunker for the testing dataset 2 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.3 Result of the chunker before applying rules on Test dataset 2

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	6	81.25%

Table 5.4 Result of the chunker after applying rules on Test dataset 2

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	4	87.5%

5.2.3. Result on Test Dataset 3

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.5 Result of the chunker before applying rules on Test dataset 3

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	4	87.5%

Table 5.6 Result of the chunker after applying rules on Test dataset 3

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	1	96.875%

5.2.4. Result on Test Dataset 4

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.7 Result of the chunker before applying rules on Test dataset 4

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	5	84.375%

Table 5.8 Result of the chunker after applying rules on Test dataset 4

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	1	96.875%

5.2.5. Result on Test Dataset 5

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.9 Result of the chunker after applying rules on Test dataset 5

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	7	78.125%

Table 5.10 Result of the chunker after applying rules on Test dataset 5

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	1	96.875%

5.2.6. Result on Test Dataset 6

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.11 Result of the chunker before applying rules on Test dataset 6

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	5	84.375%

Table 5.12 Result of the chunker after applying rules on Test dataset 6

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	3	90.625%

5.2.7. Result on Test Dataset 7

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.13 Result of the chunker before applying rules on Test dataset 7

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	4	87.5%

Table 5.14 Result of the chunker after applying rules on Test dataset 7

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	3	90.625%

5.2.8. Result on Test Dataset 8

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.15 Result of the chunker before applying rules on Test dataset 8

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	5	84.375%

Table 5.16 Result of the chunker after applying rules on Test dataset 8

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	2	93.75%

5.2.9. Result on Test Dataset 9

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.17 Result of the chunker before applying rules on Test dataset 9

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	4	87.5%

Table 5.18 Result of the chunker after applying rules on Test dataset 9

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	2	93.75%

5.2.10. Result on Test Dataset 10

The result of the chunker for the testing dataset 1 which is trained on the training dataset 1 is shown in the table 5.1 before hand crafted rules employed

Table 5.19 Result of the chunker before applying rules on Test dataset 10

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	5	84.375%

Table 5.20 Result of the chunker after applying rules on Test dataset 10

Data set	No of erroneously chunked sentences	Accuracy
Testing dataset 2	2	93.75%

The errors that can be pruned by the rule are the mis-chunk tag.

Example:- ('ትናንት', (' ADV ', 'B-S')), ('የተገዛው', (' VREL ', 'I-S')), ('ትልቅ', (' ADJ ', 'B-NP')), ('የሀረር ', (' NPREP ', 'I-NP')), ('ሰንጋ', (' N ', 'O')), ('ከመኪና', (' NPREP ', 'B-VP')), ('እንደወረደ', (' VPREP ', 'I-VP')), ('ጠፉ', (' V ', 'O'))

Here the sentence chunking is corrected by using rule 1 in Appendix 9. So it can be corrected as follows

('ትናንት', (' ADV ', 'B-S')), ('የተገዛው', (' VREL ', 'I-S')), ('ትልቅ', (' ADJ ', 'O')), ('የሀረር ', (' NPREP ', 'B-NP')), ('ሰንጋ', (' N ', ' I-NP ')), ('ከመኪና', (' NPREP ', 'B-VP')), ('እንደወረደ', (' VPREP ', 'I-VP')), ('ጠፉ', (' V ', 'O'))

The major errors that are not pruned by rules are the tag sequence conflict.

Example:- [('አስቴር', 'N'), ('ለልጁ NPREP ገንዘብ N', 'NP'), ('ላከችለት', 'V')]

Here ‘ለልጁ NPREP ገንዘብ N’ which is chunked as noun phrase is not correct but the tag sequence NPREP N is correct for other sentence chunking. For example [('ካሳ', 'N'), ('የሱፍ NPREP ኮት N', 'NP'), ('ገዛ', 'V')]. Here 'የሱፍ NPREP ኮት N' is correct noun phrase.

By taking the average of all the ten results the overall accuracy of the result is 85.31% for the statistical chunker and 93.75% for the hybrid chunker. And the overall accuracy of the parser is the same that of the chunker i.e. 93.75% for the hybrid one because if the chunker chunks correctly the parser do the same.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

This thesis described the design and development Amharic text chunker according to the language grammar construction rules and approach to convert the chunker to parser by using the output of the chunker. In this study the way how to develop a text chunker for Amharic language using HMM model with the decoding algorithm i.e. the Viterbi algorithm and the way how to transform the chunker to parser using bottom up algorithm are tried to described. All simple and complex Amharic declarative sentences are included in this study.

The thesis began with a brief discussion on the concept and applications of NLP at different levels. In this discussion, it is indicated that NLP and natural langiage understanding (NLU) require the general language structure of the data at different level of the applications to increase its capability. To achieve these text chunking, which is a series of processes first identifying chunks from a sequence of tokens or words and second classify these chunks to some syntactically related classes, is used to achieve the above objective of NLP. And also parsing that is the syntactic level of NLP and its benefits for other NLP applications are discussed.

Following this, literatures in the area of different types of Amharic word class and phrasal categories are reviewed and discussed that are the main component in designing and developing the chunker and parser. Thus, features of the language that were considered in designing the various components of the chunker were made clear. Almost all lexical and phrasal categories, sentence formalisms, typical characteristics of simple sentences and complex sentences, and features of the language that were considered in designing the various components of the chunker were also discussed.

Next, literatures about approaches and models that are used to chunk texts according to the grammar construction rule of the target languages and approaches of parsing were reviewed. Various chunk tagging schemes for marking different chunk boundaries with brief example also

reviewed. Among the chunking tagging scheme IOB2 chunk boundary method is used in this study.

The sample corpus preparation that is used in this study and the problems that the researchers have phased during the preparation of the corpus and collecting of the data also presented. This sample corpus is used to generate the phrase construction rule of Amharic language. Due to lack of systems and time, the entire corpus chunk tagged manually with the help of linguistic experts. Finally, this chunk tagged data is used as an input by the chunker. The corpus which is used in this study is small in size for the reasons that lack of annotated large corpora with POS tags and chunk tags. But it has been solved using 10 fold cross validation.

The thesis also presented the Model and approach used to develop and design the Amharic text chunker and to transform the chunker to the parser. The algorithm which is used to find the shortest path also presented here. For this purpose the system is developed using Python 3.1.

Experiments were conducted in ten phases by using 10 fold cross validation. Each phase of the result of the expermint is stated. Evaluation of the chunker performance was made based on the evaluation procedures outlined in the thesis. In the study only one parameter, the percentage of correctly chunker sentences in the sampled text was used to measure the performance of the chunker. The results achieved based on the small sample were high, 93.5%.

6.2 Recommendation

There are many shortcomings in this research. These limitations are pointed out below and are active research areas, which should be addressed by interested individuals in the area. Thus, the following could be recommended as possible research areas.

- This work is done using Tri-gram model of HMM. It is considered that chunking task must be performed by Bi-gram, Uni-gram and Tetra-gram to have comparison that which n-gram suits best for the chunking task.
- Replicate this work using a large data and incorporating all types of sentences such as interrogative sentences and exclamatory sentences
- Other techniques like Support Vector Machines (SVM), Memory based Chunking, Decision Trees and Decision forests would be investigated in future work for accuracy

- Replicate this work for other language

References

[1] J. Truscott (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255-272.

[2] P. Jackson & I. Moulinier (2007). *Natural language processing for online applications: Text retrieval, extraction and categorization* (Vol. 5). John Benjamins Publishing Company.

[3] D. Jurafsky & J. H.Martin (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

[4] S. P. Abney (1991). Parsing by chunks. *Principle-Based Parsing*, Kluwer Academic Publishers.

[5] E. F. Tjong Kim Sang & S. Buchholz (2000, September). Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7* (pp. 127-132). Association for Computational Linguistics.

[6] S. P. Abney (1992). *Parsing by chunks* (pp. 257-278). Springer Netherlands.

[7] R. Yangarber & R. Grishman (1998). NYU: Description of the Proteus/PET system as used for MUC-7. *In Proceedings of the Seventh Message Understanding Conference*.

[8] A. Molina & F. Pla (2002). Shallow parsing using specialized hmms. *The Journal of Machine Learning Research*, *2*, 595-613.

[9] M. Kutlu (2010). *Noun phrase chunker for Turkish using dependency parser*. Doctoral dissertation, BILKENT UNIVERSITY.

[10] N. T. H. Thao, N. P. Thai, N. Le Minh, & H. Q. Thuy, (2009, October). Vietnamese noun phrase chunking based on conditional random fields. In *Knowledge and Systems Engineering. KSE'09. International Conference on* (pp. 172-178). IEEE.

[11] E. Ejerhed and K.W. Church (1983). Finite state parsing. *Papers from the Seventh Scandinavian Conference of Linguistics, University of Helsinki*, Finland.

[12] T. Brants (1999). Cascaded markov models. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* (pp. 118-125). Association for Computational Linguistics.

[13] P. H. Winston (1984). Artificial intelligence. *Reading, Mass.: Ad-dison-Wesley*.

[14] J. Harris (1992). Natural Language Understanding. *Reston, Virginia: Reston Publishing.*

[15] A. Argaw (2002). Automatic sentence parsing for Amharic text an Experiment using probabilistic context free grammars. Master's Thesis, *Computer Science Department, Addis Ababa University*. Addis Ababa, Ethiopia.

[16] F. Xu, C. Zong, & J. Zhao (2006). A Hybrid Approach to Chinese Base Noun Phrase Chunking. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing* (Vol. 22223, p. 87293). Sydney.

[17] W. Ali, & S. Hussain (2010, August). A hybrid approach to Urdu verb phrase chunking. In *Proceedings of the 8th Workshop on Asian Language Resources (ALR-8), COLING-2010. Beijing, China* (pp. 137-143).

[18] D. Gochel (2003)."An Integrated approach to automatic complex sentence parsing for Amharic text". Master's Thesis, *Computer Science Department, Addis Ababa University*. Addis Ababa, Ethiopia.

[19] G. Zhou & J. Su (2002, July). Named entity recognition using an HMM-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 473-480). Association for Computational Linguistics.

[20] L. A. Ramshaw & M. P. Marcus (1995, June). Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora* (pp. 82-94).

[21] E. F. Tjong, K. Sang, & H. Déjean (2001, July). Introduction to the CoNLL-2001 shared task: clause identification. In *Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7* (p. 8). Association for Computational Linguistics.

[22] G. Amare. የአማርኛ ሰዋሰው በቀላል አቀራረብ (1989). *Commercial Printing Press*, Addis Ababa.

[23] A. Bayou (2000). Developing Automatic Word Parser for Amharic Verbs and Their Derivation. Master's Thesis, *Computer Science Department, Addis Ababa University*. Addis Ababa, Ethiopia.

[24] M. Getachew (2001). Automatic Part of Speech Tagging for Amharic Language: An Experiment Using Stochastic Hidden Markov (HMM) Approach. Master's Thesis, *Computer Science Department, Addis Ababa University*. Addis Ababa, Ethiopia.

[25] T. Bayou (2002). Automatic Morphological Analyzer: An Experiment Using Unsupervised and Autosegmental Approach. Master's Thesis, *Computer Science Department, Addis Ababa University*. Addis Ababa, Ethiopia.

[26] K. Lisanu (2002). Design and development of authomatic morphological synthesizer for Amharic perfective verb forms. Master's Thesis, *Computer Science Department, Addis Ababa University*. Addis Ababa, Ethiopia.

[27] Magersa, Diriba (2002).A Automatic sentence parser for Oromo language using supervised learning technique. Master's Thesis, *Computer Science Department, Addis Ababa University*. Addis Ababa, Ethiopia.

[28] C. Grover & R. Tobin (2006). Rule-based chunking and reusability. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

[29] R. Koeling (2000). Chunking with maximum entropy models. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7* (pp. 139-141). Association for Computational Linguistics.

[30] T. Kudo & Y. Matsumoto (2001). Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

[31] A. Singh, S. Bendre & R. Sangal (2005). Hmm based chunker for Hindi. In *Proceedings of International Joint Conference on Natural Language Processing*.

[32] B. Yimam (1987). "የአማርኛ ሰዋሰው". *E.M.P.D.A, Addis Ababa.*

[33] B. Yimam (2002). "የአማርኛ ሰዋሰው". *E.M.P.D.A, Addis Ababa.*

[34] M. Frydenlund & K. Svensen (1967). *Amharic for beginners.* Norwegian Lutheran Mission Language School.

[35] A. Voutilainen (1995). NPtool, a detector of English noun phrases. *arXiv preprint cmp-lg/9502010*.

[36] J. Veenstra & S. Buchholz (1998). Fast NP chunking using memory-based learning techniques. *Proceedings of BENELEARN'98*, 71-78.

[37] M. woldeqirqos (1934). የአማርኛ ሰዋሰው. *Birhanena Selam Printing Press*, Addis Ababa.

[38] Li, S. J. (2003). Chunk parsing with maximum entropy principle. *CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION-*, *26*(12), 1722-1727.

[39] L. R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *In Proceedings of the IEEE*, *77*(2), 257-286.

[40] S. B. Park & B. T. Zhang (2003, July). Text chunking by combining hand-crafted rules and memory-based learning. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 497-504). Association for Computational Linguistics.

[41] J. Allen (1995). *Natural language understanding*. Addison-Wesley.

[42] A. L. Berger, V. J. D. Pietra & S. A. D. Pietra (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, *22*(1), 39-71.

Appendices

Appendix 1. The Amharic Alphabet, Unicode representation

	120	121	122	123	124	125	126	127	128	129	12A	12B
0	ሀ	ሐ	ሠ	ሰ	ቀ	ቐ	በ	ተ	ኀ	ነ	አ	ኰ
1	ሁ	ሑ	ሡ	ሱ	ቁ	ቑ	ቡ	ቱ	ኁ	ኑ	ኡ	
2	ሂ	ሒ	ሢ	ሲ	ቂ	ቒ	ቢ	ቲ	ኂ	ኒ	ኢ	኱
3	ሃ	ሓ	ሣ	ሳ	ቃ	ቓ	ባ	ታ	ኃ	ና	ኣ	ኳ
4	ሄ	ሔ	ሤ	ሴ	ቄ	ቔ	ቤ	ቴ	ኄ	ኔ	ኤ	ኴ
5	ህ	ሕ	ሥ	ስ	ቅ	ቕ	ብ	ት	ኅ	ን	እ	ኵ
6	ሆ	ሖ	ሦ	ሶ	ቆ	ቖ	ቦ	ቶ	ኆ	ኖ	ኦ	
7	ኋ	ሗ	ሧ	ሷ	ቇ		ቧ	ቷ	ኇ	ኗ	ኧ	
8	ለ	መ	ረ	ሸ	ቈ	ቘ	ቨ	ቸ	ኈ	ኘ	ከ	ኸ
9	ሉ	ሙ	ሩ	ሹ			ቩ	ቹ		ኙ	ኩ	ኹ
A	ሊ	ሚ	ሪ	ሺ	ቊ	ቚ	ቪ	ቺ	ኊ	ኚ	ኪ	ኺ
B	ላ	ማ	ራ	ሻ	ቋ	ቛ	ቫ	ቻ	ኋ	ኛ	ካ	ኻ
C	ሌ	ሜ	ሬ	ሼ	ቌ	ቜ	ቬ	ቼ	ኌ	ኜ	ኬ	ኼ
D	ል	ም	ር	ሽ	ቍ	ቝ	ቭ	ች	ኍ	ኝ	ክ	ኽ
E	ሎ	ሞ	ሮ	ሾ			ቮ	ቾ		ኞ	ኮ	ኾ
F	ሏ	ሟ	ሯ	ቿ			ቯ	ቿ		ኟ	ኯ	
	12c	12d	12e	12f	130	131	132	133	134	135	136	137
0	ዀ	ዐ	ዠ	ደ	ጀ	ጐ	ጠ	ጰ	ፀ	ፐ	※	፧

1												
2												
3												
4												
5												
6												
7												
8												
9												
A												
B												
C												
D												
E												
F												

Appendix 2. Amharic punctuation mark

NO	Punctuation mark	Symbol	Purpose
	The four dots (or double colon)	**::**	Marks end of a word and at the same time the end of a sentence
	Colon	**:**	Separate individual words in a sentence
	White space		Separate individual words in a sentence, the current practice
	Question mark	?	Marks the end of an interrogative sentence
	Exclamation	!	Used at the end of such sentences or interjections those express such emotions as…
	Semi-colon		Serves roughly the same function as comma. It separates related meanings
	Three dots	**…**	Marks deliberate omission of words, phrases or sentence
	Quotation marks	**“”** **‘’**	Used at the beginning and end of words that are being quoted
	Parenthesis	**()**	Encloses elaboration of Amharic meanings
	Stroke	**|**	Separates date, month, year on official let’ers of an organization

Appendix 3. POS tags by Mesfin

No	TAG	DESCRIPTION
1	N	Nouns including all pronouns, invariant for number, gender and case except for verbal nouns and such nouns formed using the prefix balä(e.g. ልጅ "child", ልጆች " Children", እሱ "He", ደግነት "kindness")
2	NV	Verbal nouns (e.g. መብላት "Eating", መጠጣት "Drinking."
3	NB	Noun formed by prefixing the prefix ባለ to nouns (e.g. ባለ በግ "The Sheep owner"
4	Nprep	A word with a preposition not separated from a noun (e.g. በመኪና "By car", ስለሀገር "about a country")
5	NC	A noun suffixed with a conjunction, i.e. a word with noun not separated from a conjunction (e.g. ሎሚና ብርቱካን "Lemon and orange", ዘይትስ "how
6	V	Verb in any form except auxiliary verbs, compound verbs and other forms of the auxiliary and compound verbs (e.g. ገደለ "He killed", ገድሎ "after he killed", ገደለች"še killed")
7	AUX	Auxiliary verbs and all their other forms. This does not include compounds of አለ "He, It present", አደረገ "He did" and all their other forms (e.g. አለ "He, It present" ,ነበረ "He, It was", ነው "It is", ነች "Še is",ናቸው "They are")
8	VCO	Compound verbs, i.e. compounds of አለ "He, It present", አደረገ "He did" and all their other forms (e.g ብቅ አለ "He appears", ብድግ አደረገ "He takes it up.")
9	Vprep	Any verb (main or auxiliary) headed by a preposition. The preposition not separated from the verb (e.g. ስለመጣ "Since he came", ከሄደ "If he went")
10	VC	Any verb suffixed or prefixed (i.e. headed) by a conjunction. That is, a word with the conjunction not separated from the verb (e.g. መጣና "He come and" , ሲመጣ "When he comes", እስክትጨርስ "Until you finišh")

11	J	An adjective which is preceded by neither prepositions nor conjunctions (e.g. ደግ “Kind”, ክፉኛ “Dangerously”, ትልቅ “Big”)
12	JC	An adjective not separated from a conjunction (e.g. ደግና የዋህ “Kind and Innocent”, ጥቁርና ነጭ “Black and white”)
13	JNU	A numeral that function as an adjective (e.g ሁለት ብርጭቆ “ two glasses”)
14	JPN	A preposition not separated from a noun but that function as an Adjective (e.g. የጠላ ብርጭቆ “A glass for “tella””, የቻይና ሳህን “A china made plate”)
15	JP	A word with a preposition not separated from the adjective. That is, the adjective is headed by a preposition (e.g. በደህና “In a fine way”)
16	PREP	A preposition that appear being not at’ached with other words (e.g. ከ “From”, ለ “To”, ስለ “For Sb/Sth.”, እንደ “Like”)
17	ADV	An adverb (e.g. ቶሎ “In a hurry”, ትናንትና “Yesterday”, ዛሬ “Today”, ሁል ጊዜ “Always”)
18	ADVC	An adverb which has a conjunction suffixed to it (e.g. አሁንም “Even now”)
19	C	Coordinating conjunctions that appear being not at’ached with other words (e.g. ነገር ግን “However”, ወይስ “Or”)
20	REL	A word which is a relative clause (e.g. የተሰረቀበት “one who is stolen”, የቆሙት “those that stand”)
21	ITJ	Interjections (e.g. ጎሽ! “Wonderful”, ዋ! “Take care! Be careful! Watch out!”)
22	ORD	Ordinal number (e.g. አምስተኛ “The fifth” ,አስረኛ “Tenth”)
23	CRD	Cardinal number (e.g. አምስት “Five”, አስር “Ten”)
24	PUNC	Punctuation (e.g. :, !,፩)

25	UNC	Unrecognized word, i.e. a word not found in the lexicon of the tagger

Appendix 4: Pos tags by WIC (used by this study)

NO	TAG	DESCRIPTION	EXAMPLE
1	ADJ	Adjective	በጣም()
2	N	Noun	ወንበር
3	VREL	Relative verb	የዛው
4	NUMCR	Cardinal Number	አንድ
5	V	Verb	ሄደ
6	ENDPUNC	Sentence end punc	
7	NPrep	Noun with preposition	በመኪና
8	VPrep	Verb with preposition	
9	NUMPrep	Number with preposition	በመቶ
10	PREP	Preposition	ጋር
11	VN	Verbal noun	
12	ADJPrep	Adjective with preposition	
13	NC	Noun with conjunction	ጥቁርና
14	ADV	Adverb	ዛሬ
15	PUNC	Punctuation	
16	NPC	Noun with preposition and conjunction	በመኪናና
17	AUX	Auxiliary verbs	ነው
18	PRONP	Pronoun with preposition	
19	CONJ	Conjunction	
20	NUMOR	Ordinal Number	
21	VPC	Verb with preposition and conjunction	
22	PRON	Pronoun	
23	PRONPC	Pronoun with preposition and conjunction	

24	ADJC	Adjective with conjunction	
25	VC	Verb with conjunction	
26	PRONC	Pronoun with conjunction	
27	UNC	Unclear	
28	ADJPC	Adjective with preposition and conjunction	
29	INT	Interjection	
30	NUMC	Number with conjunction	
31	NUMPC	Number with preposition and conjunction	

Appendix 5. Chunk tags

Chunk tags	Description
B-NP	Beginning of Noun Phrase
I-NP	Inside Noun Phrase
B-PP	Beginning of Prepositional Phrase
I-PP	Inside Prepositional Phrase
B-VP	Beginning of Verb Phrase
I-VP	Inside Verb Phrase
B-ADJP	Beginning of Adjectival Phrase
I-ADJP	Inside Adjectival Phrase
B-ADVP	Beginning of Adverbial Phrase
I-ADVP	Inside Adverbial Phrase
O	Outside from any phrase construction
B-S	Beginning of Sentence
I-S	Inside sentence

Appendix 6. Sample tagged document

1. ፕሬዚዳንቱ N አደጋው N የኢትዮጵያን NPREP ህዝብ N እንዳሳዘነ VPREP ገለጹ V
2. አመልድ N ለ300ሺ NUMPREP ወገኖች N ኢርዳታ N አከፋፈለ V
3. በቦረና NPREP የኦሀዴድ NPREP ተሃድሶ N ውይይት N ተጀመረ V
4. አለምአቀፍ ADJ የሩጫ NPREP ውድድር N በአዲስአበባ NPREP ሊካሄድ V ነው AUX
5. የጋምቤላ NPREP ፖርክን N ለማሻሻል NPREP 110ሺ NUMCR ብር N ተመደበ V
6. የኢሃዴግ NPREP 4ኛ NUMOR ድርጅታዊ ADJ ጉባኤ N ማምሻውን ADV ተከፈተ V
7. ከኤርትራ NPREP 205 NUMCR እትዮጵያውያን N ወደ PREP አገራቸው N ተመለሱ V
8. በእንግሊዝ NPREP የሚኖሩ VREL ኢትዮጵያውያን N የጋራ NPREP መድረክ N መሰረቱ V
9. የኢሃዴግ NPREP አራተኛ NUMOR ድርጅታዊ ADJ ጉባኤ N አጀንዳዎቹን N አጸደቀ V
10. ዲስትሪክቱ N በ36ሚሊየን NUMPREP ብር N የመንገድ NPREP ጥገና N አከናውኗል V
11. የኢትዮጵያ NPREP የትምህርት NPREP ሽፋን N በ6ነጥብ4በመቶ NUMPREP አደጉ V
12. ኢማተልፈ N ያስገነባቸው VREL 17 NUMCR ትምህርትቤቶች N አገልግሎት N መስጠት V ጀመሩ V
13. በደንቆሮ NPREP ደን N የቀይ ADJPREP ቀበሮ N ቁጥር N እየጨመረ V ነው AUX
14. የ14 NUMPREP የምርመራ NPREP ላብራቶሪዎች N ግንባታ N ተጠናቀቁ V
15. የካሳ NPREP ጓደኛ N እንደ PREP ካሳ N ንበዝ ADJ ተማሪ N ሆነ V
16. አንድ NUMCR የእርሻ NPREP በሬ N ከመኪና NPREP እንደወረደ VPREP ጠፋ V
17. ሁለት NUMCR ጣሳ N የገብስ NPRE ጠላ N ከገበያ NPREP ተገዛ V
18. ከአስቴር NPREP የበለጠች VREL ቆንጆ ADJ ልጅ N ከጎጃም NPREP መጣች V
19. ሁለት NUMCR ሊትር N ንጹህ ADJ የማር NPREP ጠጅ N ተሸጠ V
20. የአዲስአበባ NPREP ዩኒቨርሲቲ N ለዶክተር ADJPREP ብርሃነ N የፕሮፌሰርነት NPREP ማእረግ N ሰጠ V
21. አረጋውያን N ራሳቸውን N የሚያቋቁሙበትን VREL ማህበር N መሰረቱ V
22. አስቴር N ሊቀመንበሩን N ጠራችው V
23. በሶስት NUMCR ከተሞች N ችግረኛ ADJ ህጻናትን N ማቋቋም V ተጀመሬ V
24. ካሳ N ቤት N ውስጥ PREP አለ V
25. አስቴር N ሁለት ADJ ልጅ N አላት V
26. የትናንትናው ADVPREP ልጅ N ሰነፍ ADJ ነው AUX
27. አስቴር N አልጋው N ላይ PREP ወደቀች V
28. ልጁ N አባቱን N በሀይል NPREP ተሳደበ V
29. ሁለቱ NUMCR ልጆች N ወደ PREP ጎጃም N ሄዱ V
30. ፕሬዝዳንቱ N ሚኒስትሮችን N በማዕረግ NPREP ሾሙ V

31. ሚኒስትሩ N ወታደሮቹ N ሀገራቸውን N ከወራሪዎች NPREP ስለታደጉ VPREP በጣም ADJ አመሰገኑ V
32. ሰራተኞቹ N ድርጅቱ N የሸለመውን VREL ኮከብ ADJ ሰራተኛ N ሊቀመንበር N አደረጉት V
33. ግርማዊነታቸው N አዳዲስ ADJ ድርጅቶች N ጎበኙ V
34. ካሳ N የገዛው VREL የሱፍ NPREP ኮት N በጣም ADJ ያምራል V
35. ጎበዙ ADJ ተማሪ N ከትምህርትቤት NPREP እንደወጣ VPREP ወደ PREP ቤት N መጣ V
36. ተቃዋሚዎቹ N መምህር N በውድ NPREP ቀጠሩ V
37. ብዙ ADJ ሰዎች N ጃፓን N የሰራችውን VREL እቃ N መግዛት V ፈለጉ N
38. ካሳ N ከጎጃም NPREP የመጣበት VREL መኪና N ክፉኛ ADV ተጋጨ V
39. ካሳ N ስለዛገ V ተጣለ V
40. ካሳ N በመኪና NPREP ሄደ V
41. አስቴር N ጎበዝ ADJ ትመስላለች V
42. ካሳ N ወደ PREP ቤተክርስቲያን N ሄደ V
43. እነዚያ PRON ልጆች N ከትምህርትቤት NPREP እንደመጡ VPREP አጠኑ V
44. ያ PRON ትልቅ ADJ ቀለበት N እንደተገዛ VPREP ጠፋ V
45. እሷ PRON አባቷን N ትመስላለች V
46. የካሳ NPREP ላሞች N በጣም ADJ ታመሙ V
47. የኑግ NPREP ዘይት N በጣም ADJ ረከሷል V
48. አስቴር N ቤቷን N ሰርታ V የጨረሰች VPREP ይመስላል V
49. አለምአቀፍ ADJ የሩጫ NPREP ውድድር N በአዲስአበባ NPREP ሊካሄድ V ነው AUX
50. ከኤርትራ NPREP እትዮጵያውያን N ወደ PREP አገራቸው N ተመለሱ V
51. ኢማተልፈ N ያስገነባቸው VREL 17 NUMCR ትምህርትቤቶች N አገልግሎት N መስጠት V ጀመሩ V
52. የአውሮፓ NPREP አምባሳደሮች N በሻእቢያ NPREP አፈና N ላይ PREP ተቃወሙ V

Appendix 7. Sample chunk tagged document for the training data set

1. ፕሬዚዳንቱ N O አደጋው N O የኢትዮጵያን NPREP B-NP ህዝብ N I-NP እንዳሳዘነ VPREP O ገለጹ V O
2. አመልድ N O ለ300ሺ NUMPREP B-NP ወገኖች N I-NP ኢርዳታ N O አከፋፈለ V O
3. በቦረና NPREP O የኦህዴድ NPREP B-NP ተሃድሶ N I-NP ውይይት N O ተጀመረ V O
4. አለምአቀፍ ADJ O የሩጫ NPREP B-NP ውድድር N I-NP በአዲስአበባ NPREP B-VP ሊካሄድ V I-VP ነው AUX O
5. የጋምቤላ NPREP B-NP ፖርክን N I-NP ለማሻሻል NPREP O 110ሺ NUMCR B-NP ብር N I-NP ተመደበ V O
6. የኢሃዴግ NPREP O 4ኛ NUMOR O ድርጅታዊ ADJ B-NP ጉባኤ N I-NP ማምሻውን ADV B-VP ተከፈተ V I-VP
7. ከኤርትራ NPREP O 205 NUMCR B-NP እትዮጵያውያን N I-NP ወደ PREP B-PP አገራቸው N I-PP ተመለሱ V O
8. በእንግሊዝ NPREP B-S የሚኖሩ VREL I-S ኢትዮጵያውያን N O የጋራ NPREP B-NP መድረክ N I-NP መሰረቱ V O
9. የኢሃዴግ NPREP O አራተኛ NUMOR O ድርጅታዊ ADJ B-NP ጉባኤ N I-NP አጀንዳዎቹን N O አጸደቀ V O
10. ዲስትሪክቱ N O በ36ሚሊየን NUMPREP B-NP ብር N I-NP የመንገድ NPREP B-NP ጥገና N I-NP አከናውኗል V O
11. የኢትዮጵያ NPREP O የትምህርት NPREP B-NP ሽፋን N I-NP በ6ነጥብ4በመቶ NUMPREP B-VP አደጉ V I-VP
12. ኢማተልፈ N B-S ያስገነባቸው VREL I-S 17 NUMCR B-NP ትምህርትቤቶች N I-NP አገልግሎት N O መስጠት V O ጀመሩ V O
13. በደንቆሮ NPREP B-NP ደን N I-NP የቀይ ADJPREP B-NP ቀበሮ N I-NP ቁጥር N O እየጨመረ V O ነው AUX O
14. የ14 NUMPREP O የምርመራ NPREP B-NP ላብራቶሪዎች N I-NP ግንባታ N O ተጠናቀቁ V O
15. የካሳ NPREP B-NP ጓደኛ N I-NP እንደ PREP B-PP ካሳ N I-PP ጎበዝ ADJ B-NP ተማሪ N I-NP ሆነ V O
16. አንድ NUMCR O የእርሻ NPREP B-NP በሬ N I-NP ከመኪና NPREP B-VP እንደወረደ VPREP I-VP ጠፋ V O
17. ሁለት NUMCR B-NP ጣሳ N I-NP የገብስ NPRE B-NP ጠላ N I-NP ከገበያ NPREP O ተገዛ V O

18. ከአስቴር NPREP B-S የበለጠች VREL I-S ቆንጆ ADJ B-NP ልጅ N I-NP ከነጃም NPREP O መጣች V O
19. ሁለት NUMCR B-NP ሊትር N I-NP ንጹህ ADJ O የማር NPREP B-NP ጠጅ N I-NP ተሸጠ V O
20. የአዲስአበባ NPREP B-NP ዩኒቨርሲቲ N I-NP ለዶክተር ADJPREP B-NP ብርሃነ N I-NP የፕሮፌሰርነት NPREP B-NP ማእረግ N I-NP ሰጠ V O
21. ልማትፈንዱ N O 16ሚሊየን NUMCR B-NP ነዋሪዎችን N I-NP ተጠቃሚ N B-VP ማድረጉን V I-VP ገለጸ V O
22. ምክርቤቱ N O የሀገሪቱን NPREP B-NP ፕሬዚዳንት N I-NP መተዳደሪያ N B-NP አዋጅ N I-NP አጸደቁ V O
23. በሶስት NUMCR B-NP ከተሞች N I-NP ችግረኛ ADJ B-NP ህጻናትን N I-NP ማቋቋም V O ተጀመሬ V O
24. ድርጅቶቹ N O ከ57ሚሊየን NUMCR B-NP ብር N I-NP በላይ NPREP O የፕሮጀክቶች NPREP B-NP ግንባታ N I-NP አስጀመሩ V O
25. መስተዳድሩ N O ከፍተኛ ADJ B-NP ውጤት N I-NP ላስመዘገቡ VPREP O ተማሪዎች N O ሽልማት N O ሰጡ V O
26. በኬንያ NPREP B-S የሚኖሩ VREL I-S ኢትዮጵያውያን N O 35ሺ797 NUMCR B-NP ዶላር N I-NP ለገሱ V O
27. የሞሮኮ NPREP B-NP መንግስታት N I-NP የደስታ ADJPREP B-NP መግለጫ N I-NP አስተላለፉ V O
28. የሱማሊያ NPREP B-NP ስደተኞች N I-NP ወደ PREP B-PP ሀገራቸው N I-PP ተመለሱ V O
29. የጡረተኞቹ NPREP B-NP ማህበር N I-NP ስራዎችን N O እያከናወነ V O ነው AUX O
30. ካሳ N O ለአስቴር NPREP B-VP ሰጣት V I-VP
31. ካሳ N O ለሰውየው NPREP B-VP ነገረው V I-VP
32. ካሳ N O ቤት N B-PP ውስጥ PREP I-PP አለ V O

Appendix 8. Sample chunk output

1. [('ሁለቱ NUMCR ልጆች N', 'NP'), ('ወደ P እኛ PRON', 'PP'), ('መጡ', 'V')]
2. [('መንግስት', 'N'), ('ሀገሪቱ N የደረሰባትን VPREP', 'S'), ('ከፍተኛ', 'ADJ'), ('የኢኮኖሚ ADJPREP ኪሳራ N', 'NP'), ('አመነ', 'V')]
3. [('ወንበዴዎች', 'N'), ('በነፈቃደኞች NPREP የገነቡትን VPREP', 'VP'), ('ድርጅት', 'N'), ('ከጥቅም NPREP ውጭ PREP', 'PP'), ('አደረጉት', 'V')]
4. [('ቢሮው', 'N'), ('የንግድ NPREP ፈቃድ N', 'NP'), ('የተሰረዘባቸውን', 'VPREP'), ('ነጋዴዎች', 'N'), ('ዝርዝር N አወጣ V', 'VP')]
5. [('ግለሰቡ', 'N'), ('የስኳር NPREP ጨረታውን N', 'NP'), ('ስላጠናቀቀ', 'VPREP'), ('ምርቱን', 'N'), ('በመኪና NPREP ጫነ V', 'VP')]
6. [('እኛ', 'PRON'), ('አስተማሪው N የሰጠንን VPREP', 'S'), ('የቤት NPREP ስራ N', 'NP'), ('ክፍል N ውስጥ PREP', 'PP'), ('ሰራን', 'V')]
7. [('አንድ', 'NUMCR'), ('የነጃም', 'NPREP'), ('የእርሻ NPREP በሬ N', 'NP'), ('ከመኪና NPREP እንደወረደ VPREP', 'VP'), ('ጠፋ', 'V')]
8. [('የሰፈር NPREP ዱርዬዎች N', 'NP'), ('እናቱ', 'N'), ('የላከችውን', 'VPREP'), ('ትንሽ ADJ ልጅ N', 'NP'), ('ከፉኛ ADV ደበደቡት V', 'VP')]
9. [('ሁለቱ', 'NUMCR'), ('ትልልቅ ADJ ልጆች N', 'NP'), ('በመኪና', 'NPREP'), ('ወደ PREP ነጃም N', 'PP'), ('ሄዱ', 'V')]
10. [('ካሳ', 'N'), ('ከአገሩ NPREP እንደመጣ VPREP', 'VP'), ('አስቴር', 'N'), ('ወደ PREP ናዝሬት N', 'PP'), ('እንደሄደች', 'VPREP'), ('ሰማ', 'V')]
11. [('ከአስቴር NPREP የበለጠች VREL', 'S'), ('ቆንጆ ADJ ልጅ N', 'NP'), ('ከነጃም NPREP መጣች V', 'VP')]
12. [('ሁለት NUMCR ጣሳ N', 'NP'), ('የጉበስ NPRE ጠላ N', 'NP'), ('ከገበያ NPREP ተገዛ V', 'VP')]
13. [('አስቴር', 'N'), ('ነጃም N ሄደች V', 'VP')]
14. [('የካሳ NPREP ጓደኛ N', 'NP'), ('እንደ PREP ካሳ N', 'PP'), ('ነበዝ ADJ ተማሪ N', 'NP'), ('ለመሆን', 'VPREP'), ('ሞከረ', 'V')]
15. [('ፕሬዚዳንቱ', 'N'), ('አደጋው', 'N'), ('ህዝብ N እንዳሳዘነ VPREP', 'VP'), ('ገለጹ', 'V')]
16. [('በእንግሊዝ NPREP የሚኖሩ VREL', 'S'), ('ኢትዮጵያውያን', 'N'), ('የጋራ NPREP መድረክ N', 'NP'), ('መሰረቱ', 'V')]
17. [('ትምህርትቤቶች', 'N'), ('አገልግሎት N መስጠት V', 'VP'), ('ጀመሩ', 'V')]

18. [('አበራ', 'N'), ('ወንድሙ N የጋበዘውን VREL', 'S'), ('ነበዝ', 'ADJ'), ('የኮሌጅ NPREP ተማሪ N', 'NP'), ('ተዋወቀ', 'V')]
19. [('ቢሮው', 'N'), ('ፈቃድ N የተሰረዘባቸውን VREL', 'S'), ('ነጋዴዎች', 'N'), ('ዝርዝር N አወጣ V', 'VP')]
20. [('ሰራተኛዋ', 'N'), ('ዝናብ N ያረጠበውን VREL', 'S'), ('ልብስ', 'N'), ('ቤት N ውስጥ PREP', 'PP'), ('አሰጣች', 'V')]
21. [('ትናንት ADV የተገዛው VREL', 'S'), ('ትልቅ', 'ADJ'), ('የሀረር NPREP ሰንጋ N', 'NP'), ('ከመኪና NPREP እንደወረደ VPREP', 'VP'), ('ጠፉ', 'V')]
22. [('ልጅቷ', 'N'), ('ትልቅ ADJ ወንድሟ N', 'NP'), ('እንዲያስጠናት', 'V'), ('ክፍል N ውስጥ PREP', 'PP'), ('ነገረችው', 'V')]
23. [('አስቴር', 'N'), ('እህቷ', 'N'), ('በጣም ADJ ስለወፈረች V', 'VP'), ('ከመጠን PREP በላይ PREP', 'PP'), ('ተናደደች', 'V')]
24. [('ወልዴ', 'N'), ('ካሳ N የያዘውን VREL', 'S'), ('አዲስ ADJ ቦርሳ N', 'NP'), ('አደነቀ', 'V')]
25. [('ሽምግሌው', 'N'), ('ጥርሳቸው N ስላለቀ VPREP', 'VP'), ('ብዙ ADJ ጥሬ N', 'NP'), ('መብላት N አቆሙ V', 'VP')]
26. [('እነዚያ', 'PRON'), ('የእሱ PRONPREP በጎች N', 'NP'), ('ምሳ N እንደበሉ VPREP', 'VP'), ('ጠፉ', 'V')]
27. [('እንግሊዝ', 'N'), ('በኤርትራ NPREP የሚፈጸመውን VREL', 'S'), ('እስራት', 'N'), ('ተቃወመች', 'V')]
28. [('ተዘዋዋሪ ADJ ችሎቱ N', 'NP'), ('550 NUMCR እስረኞች N', 'NP'), ('እንዲፈቱ', 'VPREP'), ('በየነ', 'N')]
29. [('ኮኮቦች', 'N'), ('በንጋት', 'ADVPREP'), ('በጣም ADJ ያበራሉ V', 'VP')]
30. [('የሞሮኮ NPREP መንግስታት N', 'NP'), ('የደስታ ADJPREP መግለጫ N', 'NP'), ('አስተላለፉ', 'V')]
31. [('ሰውየው', 'N'), ('ቤት N ድረስ PREP', 'PP'), ('ሸኘን', 'V')]
32. [('ካሳ', 'N'), ('ሰው', 'N'), ('እጅግ ADJ ያከብራል V', 'VP')]
33. [('የካሳ NPREP መጽሀፍ N', 'NP'), ('ተቀደደ', 'V')]
34. [('ካሳ', 'N'), ('ትንሽ ADJ መጽሀፍ N', 'NP'), ('ሰጠ', 'V')]
35. [('አስቴር', 'N'), ('ቤቷን N ሰርታ V', 'VP'), ('የጨረሰች', 'VPREP'), ('ይመስላል', 'V')]
36. [('አለምአቀፍ', 'ADJ'), ('የሩጫ NPREP ውድድር N', 'NP'), ('በአዲስአበባ NPREP ሊካሄድ V', 'VP'), ('ነው', 'AUX')]
37. [('ከኤርትራ NPREP እትዮጵያውያን N', 'NP'), ('ወደ PREP አገራቸው N', 'PP'), ('ተመለሱ', 'V')]
38. [('ኢማተልፈ N ያስገነባቸው VREL', 'S'), ('17 NUMCR ትምህርትቤቶች N', 'NP'), ('አገልግሎት N መስጠት V', 'VP'), ('ጀመሩ', 'V')]

39. [('ሚኒስትሩ', 'N'), ('ወታደሮቹ N ሀገራቸውን N', 'PP'), ('ከወራሪዎች NPREP ስለታደጉ VPREP', 'VP'), ('በጣም ADJ አመሰገኑ V', 'VP')]
40. [('ሰራተኞቹ', 'N'), ('ድርጅቱ N የሸለመውን VPREP', 'S'), ('ኮከብ ADJ ሰራተኛ N', 'NP'), ('ሊቀመንበር N አደረጉት V', 'VP')]
41. [('ግርማዊነታቸው', 'N'), ('አዳዲስ ADJ ድርጅቶች N', 'NP'), ('ጎበኙ', 'V')]
42. [('ካሳ N የገዛው VPREP', 'S'), ('የሱፍ NPREP ኮት N', 'NP'), ('በጣም ADJ ያምራል V', 'VP')]
43. [('ጎበዙ ADJ ተማሪ N', 'NP'), ('ከትምህርትቤት NPREP እንደወጣ VPREP', 'VP'), ('ወደ PREP ቤት N', 'PP'), ('መጣ', 'V')]
44. [('ተቃዋሚዎቹ', 'N'), ('መምህር', 'N'), ('በውድ NPREP ቀጠሩ V', 'VP')]
45. [('በቦረና', 'NPREP'), ('የኦህዴድ NPREP ተሃድሶ N', 'NP'), ('ውይይት N ተጀመረ V', 'VP')]
46. [('ከኤርትራ', 'NPREP'), ('205 NUMCR እትዮጵያውያን N', 'NP'), ('ወደ PREP አገራቸው N', 'PP'), ('ተመለሱ', 'V')]
47. [('እነዚያ', 'PRON'), ('የአንተ PRONPREP ልጆች N', 'NP'), ('ከትምህርትቤት NPREP እንደመጡ VPREP', 'VP'), ('አጠኑ', 'V')]
48. [('ግርማዊነታቸው', 'N'), ('አውራጃው N የገነባውን VPREP', 'S'), ('አዳዲስ ADJ ድርጅቶች N', 'NP'), ('ጎበኙ', 'V')]
49. [('ትናንት', 'ADV'), ('ከጎጃም NPREP የመጣው VREL', 'S'), ('ልጅ', 'N'), ('እኛ PRON ቤት N', 'NP'), ('መጣ', 'V')]
50. [('መንግስት', 'N'), ('ደርግ N ላፈናቀላቸውን VPREP', 'S'), ('ደሃ ADJ ገበሬዎች N', 'NP'), ('ቀለብ N አከፋፈለ V', 'VP')]
51. [('ቦርሳው', 'N'), ('ብዙ ADJ እቃ N', 'NP'), ('በውስጡ NPREP ስለያዘ VPREP', 'VP'), ('በጣም ADV ከበደን V', 'VP')]
52. [('ትናንት', 'ADV'), ('ከጎጃም NPREP የመጣው VREL', 'S'), ('ልጅ', 'N'), ('አስቴር N እንደወደደቸው VPREP', 'VP'), ('አወቀ', 'V')]
53. [('ካሳ', 'N'), ('መጽሀፍ N የላከለት VREL', 'S'), ('ልጅ', 'N'), ('ከስራ NPREP ስትመጣ VPREP', 'VP'), ('መኪና N ገጫት V', 'VP')]
54. [('ወታደሮቹ', 'N'), ('ወደ PREP ግቢያቸው N', 'PP'), ('ገቡ', 'V')]
55. [('ተዘዋዋሪ ADJ ችሎቱ N', 'NP'), ('አራት NUMCR ተከሳሾችን N', 'NP'), ('በነጻ NPREP አሰናበተ V', 'VP')]
56. [('ካሳ', 'N'), ('ሚስጥር N ነገረው V', 'VP')]
57. [('አስቴር', 'N'), ('እንደ PREP እህቴ N', 'PP'), ('ወደቀች', 'V')]
58. [('አስቴር', 'N'), ('ወደ PREP ቤት N', 'PP'), ('ገባች', 'V')]
59. [('ልጆቹ', 'N'), ('ዛፍ N ላይ PREP', 'PP'), ('ወጡ', 'V')]
60. [('አስቴር', 'N'), ('ትንሽ ADJ እህት N', 'NP'), ('አላት', 'V')]
61. [('ፕሬዝዳንቱ', 'N'), ('አምስት NUMCR ሚኒስትሮችን N', 'NP'), ('በከፍተኛ ADJPREP ማዕረግ N', 'NP'), ('ሾሙ', 'V')]

62. [('ንጉሱ', 'N'), ('ፋሺስቶች N የተርበደበዱበትን VPREP', 'S'), ('ጀግና ADJ ሰው N', 'NP'), ('ሰቀሉ', 'V')]
63. [('ዞኑ N ያስገነባቸው VREL', 'S'), ('ፕሮጀክቶች', 'N'), ('ነዋሪዎችን', 'N'), ('ተጠቃሚ ADJ አደረጉ V', 'VP')]
64. [('ትላንት ADV የዘነበው VPREP', 'S'), ('የከረምት NPREP ዝናብ N', 'NP'), ('ሀይለኛ ADV ነው AUX', 'VP')]
65. [('ሰውየው', 'N'), ['በለጠ', 'N'], ('ወደ PREP ጎጃም N', 'PP'), ('እንደሄደ', 'V'), ('አሁን ADV ነገሩን V', 'VP')]
66. [('ደጓ ADJ ሴትዮ N', 'NP'), ('የጸሎት NPREP መጽሀፋቸው N', 'NP'), ('ስለጠፋ', 'VPREP'), ('በጣም ADJ አዘኑ V', 'VP')]
67. [('መንግስት', 'N'), ('ሀገሪቱ N የደረሰባትን VPREP', 'S'), ('ከፍተኛ', 'ADJ'), ('የኢኮኖሚ ADJPREP ኪሳራ N', 'NP'), ('አመነ', 'V')]
68. [('ልዩ', 'ADJ'), ('ዞኑ N ያስገነባቸው VREL', 'S'), ('ፕሮጀክቶች', 'N'), ('50ሺ NUMCR ነዋሪዎችን N', 'NP'), ('ተጠቃሚ ADJ አደረጉ V', 'VP')]
69. [('ካሳ', 'N'), ('ከአገሩ NPREP እንደመጣ VPREP', 'VP'), ('አስቴር', 'N'), ('ወደ PREP ናዝሬት N', 'PP'), ('እንደሄደች', 'VPREP'), ('ሰማ', 'V')]
70. [('ልጁ', 'N'), ('ቤቱ', 'N'), ('በቁጣ NPREP ስለተናገሩት V', 'VP'), ('ከቤት NPREP ወጣ V', 'VP')]
71. [('አሮጊቷ', 'N'), ('ሌቦች', 'N'), ('አምና ADV የገደሉትን VPREP', 'VP'), ('ብቸኛ ADJ ልጃቸውን N', 'NP'), ('አስታወሱ', 'V')]
72. [('እኔ', 'PRON'), ('ልብሴ', 'N'), ('ዛሬ ADV ስለታጠበ VPREP', 'VP'), ('አሮጌ ADJ ልብስ N', 'NP'), ('ለበኩ', 'V')]
73. [('ገበያተኞቹ', 'N'), ('እኛ PRON የቆምንበትን VREPE', 'S'), ('አውላላ ADJ ሜዳ N', 'NP'), ('ወድያው ADV ሞሉት V', 'VP')]
74. [('ብርጭቆ N የሰበረው VREL', 'S'), ('ልጅ', 'N'), ('ወተት N ጠጣ V', 'VP')]
75. [['እኔ', 'PRON'], ('አልጋ N ተዋስኩ V', 'VP')]
76. [('ነጋዴው', 'N'), ('ትንሽ', 'ADJ'), ('የደብረብርሀን NPREP በግ N', 'NP'), ('ገዛ', 'V')]
77. [('ካሳ', 'N'), ('የሱፍ NPREP ኮት N', 'NP'), ('ገዛ', 'V')]
78. [('አዛዡ', 'N'), ['ወታደሮቹ', 'N'], ('ወደ PREP ግቢያቸው N', 'PP'), ('እንደገቡ', 'VPREP'), ('በምስጋን NPREP ተቀበሏቸው V', 'VP')]
79. [('የኢሃዴግ', 'NPREP'), ('አራተኛ', 'NUMOR'), ('ድርጅታዊ ADJ ጉባኤ N', 'VP'), ('አጀንዳዎቹን N አጸደቀ V', 'VP')]
80. [('የ14', 'NUMPREP'), ('የምርመራ NPREP ላብራቶሪዎች N', 'NP'), ('ግንባታ N ተጠናቀቁ V', 'VP')]
81. [('የኢትዮጵያ NPREP ጠላቶች N', 'NP'), ('ሀገሪቱ', 'N'), ('በድርጅቱ NPREP ውስጥ P', 'PP'), ('የተሰጣትን', 'VPREP'), ('ቦታ', 'N'), ('ተቃወሙ', 'V')]

Appendix 9 Sample rules for correction

1. If POS(wi)=ADJ and POS(wi+1)=NPREP, NUMCR ,then chunk tag for wi is O
2. If POS(wi)=ADJ and POS(wi-1)!= ADJ and POS(wi+1)= AUX,V ,then chunk tag for wi is B-VP
3. If POS(wi)=NPREP and POS(wi+1)=N ,then chunk tag for wi is B-NP
4. If POS(wi)=NUMCR and POS(wi+1)=NPREP ,then chunk tag for wi is O
5. If POS(wi)=N and POS(wi+1)=VPREP and POS(wi-1)=N, ADJ, PRON,NPREP ,then chunk tag for wi is B-VP
6. If POS(wi)=ADJ and POS(wi+1)=ADJ ,then chunk tag for wi is B-ADJP

Appendix 10. Sample parsing

1. [('ልጅቷ', 'N'), ('በጣም ADJ አመለኛ ADJ', 'ADJP'), ('ነች', 'V')]

   ['ልጅቷ N', 'አመለኛ ADJ', 'ነች V']

   [('ልጅቷ', 'N'), ('አመለኛ ADJ ነች V', 'VP')]

   ['ልጅቷ N', 'ነች V']

   [('ልጅቷ', 'N'), ('ነች', 'V')]

2. [('ኮሚሽኑ', 'N'), ('የ1ነጥብ3ሚሊየን NUMPREP ብር N', 'NP'), ('የፕሮጀክት NPREP ስምምነት N', 'NP'), ('ተፈራረመ', 'V')]

   ['ኮሚሽኑ N', 'ብር N', 'ስምምነት N', 'ተፈራረመ V']

   [('ኮሚሽኑ', 'N'), ('ብር', 'N'), ('ስምምነት N ተፈራረመ V', 'VP')]

   ['ኮሚሽኑ N', 'ብር N', 'ተፈራረመ V']

   [('ኮሚሽኑ', 'N'), ('ብር N ተፈራረመ V', 'VP')]

   ['ኮሚሽኑ N', 'ተፈራረመ V']

   [('ኮሚሽኑ', 'N'), ('ተፈራረመ', 'V')]

3. [('አንባቢያን', 'N'), ('ደራሲው N የሳለውን VREL', 'S'), ('ጨካኝ ADJ ገጸ-ባህሪ N', 'NP'), ('ተአማኒነት N ተጠራጠሩ V', 'VP')]

   ['አንባቢያን N', 'ደራሲው N የሳለውን VREL', 'ገጸ-ባህሪ N', 'ተጠራጠሩ V']

   [('አንባቢያን', 'N'), ('ደራሲው N የሳለውን VREL', 'S'), ('ገጸ-ባህሪ', 'N'), ('ተጠራጠሩ', 'V')]

   ['አንባቢያን N', ('ደራሲው N የሳለውን VREL ገጸ-ባህሪ N', 'NP'), 'ተጠራጠሩ V']

   [('አንባቢያን', 'N'), ('ገጸ-ባህሪ N ተጠራጠሩ V', 'VP')]

   ['አንባቢያን N', 'ተጠራጠሩ V']

   [('አንባቢያን', 'N'), ('ተጠራጠሩ', 'V')]

4. [('ወንበዴዎች', 'N'), ('በጎፈቃደኞች NPREP የገነቡትን VREL', 'S'), ('ድርጅት', 'N'), ('ከጥቅም NPREP ውጭ PREP', 'PP'), ('አደረጉት', 'V')]

   ['ወንበዴዎች N', ('በጎፈቃደኞች NPREP የገነቡትን VREL ድርጅት N', 'NP'), ('ከጥቅም NPREP ውጭ PREP አደረጉት V', 'VP')]

   [('ወንበዴዎች', 'N'), ('ድርጅት N አደረጉት V', 'VP')]

   ['ወንበዴዎች N', 'አደረጉት V']

   [('ወንበዴዎች', 'N'), ('አደረጉት', 'V')]

5. [('ቢሮው', 'N'), ('የንግድ NPREP ፈቃድ N', 'NP'), ('የተሰረዘባቸውን', 'VREL'), ('ነጋዴዎች', 'N'), ('ዝርዝር N አወጣ V', 'VP')]

   ['ቢሮው N', 'ፈቃድ N', 'የተሰረዘባቸውን VREL', 'ነጋዴዎች N', 'አወጣ V']

   [('ቢሮው', 'N'), ('ፈቃድ N የተሰረዘባቸውን VREL', 'S'), ('ነጋዴዎች', 'N'), ('አወጣ', 'V')]

   ['ቢሮው N', ('ፈቃድ N የተሰረዘባቸውን VREL ነጋዴዎች N', 'NP'), 'አወጣ V']

   [('ቢሮው', 'N'), ('ነጋዴዎች N አወጣ V', 'VP')]

   ['ቢሮው N', 'አወጣ V']

   [('ቢሮው', 'N'), ('አወጣ', 'V')]

6. [('ነበዙ ADJ ተማሪ N', 'NP'), ('ወደ PREP ቤት N', 'PP'), ('መጣ', 'V')]

   ['ተማሪ N', ('ወደ PREP ቤት N መጣ V', 'VP')]

   [('ተማሪ', 'N'), ('መጣ', 'V')]

7. [('እኛ', 'PRON'), ('አስተማሪው N የሰጠንን VREL', 'S'), ('የቤት NPREP ስራ N', 'NP'), ('ክፍል N ውስጥ PREP', 'PP'), ('ሰራን', 'V')]

   ['እኛ PRON', 'አስተማሪው N የሰጠንን VREL', 'ስራ N', ('ክፍል N ውስጥ PREP ሰራን V', 'VP')]

   [('እኛ', 'PRON'), ('አስተማሪው N የሰጠንን VREL', 'S'), ('ስራ', 'N'), ('ሰራን', 'V')]

['እኛ PRON', ('አስተማሪው N የሰጠንን VREL ስራ N', 'NP'), 'ሰራን V']

[['እኛ', 'PRON'], ('ስራ N ሰራን V', 'VP')]

['እኛ PRON', 'ሰራን V']

[('እኛ', 'PRON'), ('ሰራን', 'V')]

8. [('ካሳ N የገዛው VREL', 'S'), ('የትናንቱ ADVPREP በግ N', 'NP'), ('ታረደ', 'V')]

   ['ካሳ N የገዛው VREL', 'በግ N', 'ታረደ V']

   [('ካሳ N የገዛው VREL', 'S'), ('በግ', 'N'), ('ታረደ', 'V')]

   [('ካሳ N የገዛው VREL በግ N', 'NP'), 'ታረደ V']

   [('በግ', 'N'), ('ታረደ', 'V')]

9. [('የአዲስአበባ NPREP ዩኒቨርሲቲ N', 'NP'), ('ለዶክተር ADJPREP ብርሃነ N', 'NP'), ('የፕሮፌሰርነት NPREP ማእረግ N', 'NP'), ('ሰጠ', 'V')]

   ['ዩኒቨርሲቲ N', 'ብርሃነ N', 'ማእረግ N', 'ሰጠ V']

   [('ዩኒቨርሲቲ', 'N'), ('ብርሃነ', 'N'), ('ማእረግ N ሰጠ V', 'VP')]

   ['ዩኒቨርሲቲ N', 'ብርሃነ N', 'ሰጠ V']

   [('ዩኒቨርሲቲ', 'N'), ('ብርሃነ N ሰጠ V', 'VP')]

   ['ዩኒቨርሲቲ N', 'ሰጠ V']

   [('ዩኒቨርሲቲ', 'N'), ('ሰጠ', 'V')]

10. [('ኢማተልፈ N ያስገነባቸው VREL', 'S'), ('ትምህርትቤቶች', 'N'), ('አገልግሎት N መስጠት V', 'VP'), ('ጀመሩ', 'V')]

    [('ኢማተልፈ N ያስገነባቸው VREL ትምህርትቤቶች N', 'NP'), 'መስጠት V', 'ጀመሩ V']

    [('ትምህርትቤቶች', 'N'), ('መስጠት V ጀመሩ V', 'VP')]

    ['ትምህርትቤቶች N', 'ጀመሩ V']

[('ትምህርትቤቶች', 'N'), ('ጀመሩ', 'V')]

11. ['የጋምቤላ NPREP', 'ፖርክን N', 'ለማሻሻል VPREP', 'ብር N', 'ተመደበ V']

    [('የጋምቤላ NPREP ፖርክን N', 'NP'), ('ለማሻሻል', 'VPREP'), ('ብር', 'N'), ('ተመደበ', 'V')]

    ['ፖርክን N', 'ለማሻሻል VPREP', 'ብር N', 'ተመደበ V']

    [('ፖርክን N ለማሻሻል VPREP', 'S'), ('ብር', 'N'), ('ተመደበ', 'V')]

    [('ፖርክን N ለማሻሻል VPREP ብር N', 'NP'), 'ተመደበ V']

    [('ብር', 'N'), ('ተመደበ', 'V')]

12. [('እኔ', 'PRON'), ('ካሳ N የገዛውን VREL', 'S'), ('ትንሽ ADJ በግ N', 'NP'), ('ትላንት ADV አየሁት V', 'VP')]

    ['እኔ PRON', 'ካሳ N የገዛውን VREL', 'በግ N', 'አየሁት V']

    [('እኔ', 'PRON'), ('ካሳ N የገዛውን VREL', 'S'), ('በግ', 'N'), ('አየሁት', 'V')]

    ['እኔ PRON', ('ካሳ N የገዛውን VREL በግ N', 'NP'), 'አየሁት V']

    [['እኔ', 'PRON'], ('በግ N አየሁት V', 'VP')]

    ['እኔ PRON', 'አየሁት V']

    [('እኔ', 'PRON'), ('አየሁት', 'V')]

13. [('አበራ', 'N'), ('ወንድሙ N የጋበዘውን VREL', 'S'), ('ነበዝ', 'ADJ'), ('የኮሌጅ NPREP ተማሪ N', 'NP'), ('ተዋወቀ', 'V')]

    ['አበራ N', ('ወንድሙ N የጋበዘውን VREL ነበዝ ADJ', 'ADJP'), 'ተማሪ N', 'ተዋወቀ V']

    [('አበራ', 'N'), ('ነበዝ ADJ ተማሪ N', 'NP'), ('ተዋወቀ', 'V')]

    ['አበራ N', 'ተማሪ N', 'ተዋወቀ V']

    [('አበራ', 'N'), ('ተማሪ N ተዋወቀ V', 'VP')]

    ['አበራ N', 'ተዋወቀ V']

[('አበራ', 'N'), ('ተዋወቀ', 'V')]

14. [('ታላቋ ADJ ኢትዮጵያ N', 'NP'), ('ፈንጣጣ', 'N'), ('የተወገደበትን', 'VREL'), ('እለት', 'N'), ('ትላንት ADV አከበረች V', 'VP')]

    ['ኢትዮጵያ N', 'ፈንጣጣ N', 'የተወገደበትን VREL', 'እለት N', 'አከበረች V']

    [('ኢትዮጵያ', 'N'), ('ፈንጣጣ N የተወገደበትን VREL', 'S'), ('እለት', 'N'), ('አከበረች', 'V')]

    ['ኢትዮጵያ N', ('ፈንጣጣ N የተወገደበትን VREL እለት N', 'NP'), 'አከበረች V']

    [('ኢትዮጵያ', 'N'), ('እለት N አከበረች V', 'VP')]

    ['ኢትዮጵያ N', 'አከበረች V']

    [('ኢትዮጵያ', 'N'), ('አከበረች', 'V')]

15. [('ማረት N ያስገነባቸው VREL', 'S'), ('146', 'NUMCR'), ('የውሃ NPREP ፕሮጀክቶች N', 'NP'), ('አገልግሎት N መስጠት V', 'VP'), ('ጀመሩ', 'V')]

    ['146 NUMCR', 'ፕሮጀክቶች N', 'መስጠት V', 'ጀመሩ V']

    [('146 NUMCR ፕሮጀክቶች N', 'NP'), ('መስጠት', 'V'), ('ጀመሩ', 'V')]

    ['ፕሮጀክቶች N', 'መስጠት V', 'ጀመሩ V']

    [('ፕሮጀክቶች', 'N'), ('መስጠት V ጀመሩ V', 'VP')]

    ['ፕሮጀክቶች N', 'ጀመሩ V']

    [('ፕሮጀክቶች', 'N'), ('ጀመሩ', 'V')]

16. [('ጋዜጠኛው', 'N'), ('ንጉሱ', 'N'), ('በአሜሪካ NPREP የተደረገላቸውን VREL', 'VP'), ('አቀባበል', 'N'), ('በደንብ ADV ዘገቡ V', 'VP')]

    ['ጋዜጠኛው N', 'ንጉሱ N', 'የተደረገላቸውን VREL', 'አቀባበል N', 'ዘገቡ V']

    [('ጋዜጠኛው', 'N'), ('ንጉሱ N የተደረገላቸውን VREL', 'S'), ('አቀባበል', 'N'), ('ዘገቡ', 'V')]

    ['ጋዜጠኛው N', ('ንጉሱ N የተደረገላቸውን VREL አቀባበል N', 'NP'), 'ዘገቡ V']

[('ጋዜጠኞው', 'N'), ('አቀባበል N ዘገቡ V', 'VP')]

['ጋዜጠኞው N', 'ዘገቡ V']

[('ጋዜጠኞው', 'N'), ('ዘገቡ', 'V')]

17. [('ትላንት ADV የዘነበው VREL', 'S'), ('ዝናብ', 'N'), ('ሀይለኛ ADV ነው AUX', 'VP')]

    [('ትላንት ADV የዘነበው VREL ዝናብ N', 'NP'), 'ነው AUX']

    [('ዝናብ', 'N'), ('ነው', 'AUX')]

    ['ዝናብ N', 'ነው AUX']

18. [('ኮሚሽኑ', 'N'), ('የ7ነጥብ7ሚሊየን NUMPREP ብር N', 'NP'), ('የእርዳታ NPREP ስምምነት N', 'NP'), ('ተፈራረመ', 'V')]

    ['ኮሚሽኑ N', 'ብር N', 'ስምምነት N', 'ተፈራረመ V']

    [('ኮሚሽኑ', 'N'), ('ብር', 'N'), ('ስምምነት N ተፈራረመ V', 'VP')]

    ['ኮሚሽኑ N', 'ብር N', 'ተፈራረመ V']

    [('ኮሚሽኑ', 'N'), ('ብር N ተፈራረመ V', 'VP')]

    ['ኮሚሽኑ N', 'ተፈራረመ V']

    [('ኮሚሽኑ', 'N'), ('ተፈራረመ', 'V')]