ADDIS ABABA UNIVERSITY
College of Humanities, Language Studies, Journalism and Communication

A Morphosyntactic tagset for the Annotation of Texts in Tigrinya

By

Tsegay Woldemariam

June 2013
Addis Ababa

A Morphosyntactic Tagset for the Annotation of Texts in Tigrinya

By

Tsegay Woldemariam

A Thesis Submitted to the School of Graduate Studies, Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computational Linguistics

June 2013
Addis Ababa

ADDIS ABABA UNIVERSITY

College of Humanities, Language Studies, Journalism and Communication
Department of Linguistics

A Morphosyntactic Tagset for the Annotation of Texts in Tigrinya

By

Tsegay Woldemariam

	Approved by	Signature
Advisor	Tesfaalem Kiros	
Advisor	ERMIAS ABEBE	
Examiner	Solomon Teferra	
Examiner	Baye Yimam	

Acknowledgement

Before all I render my greater thanks with reverence to my God the almighty, who gave me more years to live and do such a work, and who allowed this work to be a part in my life.

I wish to express my profound sense of gratitude to my advisors Dr. Zelalem Liyew and Ato Ermias Abebe for introducing me to this research topic and providing their valuable guidance and unfailing encouragement throughout the course of the work. I am immensely grateful to them for their constant advice and support for the successful completion of this work. I would like to extend my due respect and gratitude to Ato Binyam Ephrem, coordinator of the Computational Linguistics program, for his immense contribution from the selection of the title to the compilation of the thesis.

I am very much thankful to all the staff members and research scholars of the Department of Linguistics, Information Science and Academy of Ethiopian Languages and Cultures for their direct or indirect help in various forms during my research work. I thank Dr. Hirut W.Mariam, Ato Mesfin Messele and Dr. Moges Yigezu for their encouragement in my study.

I pass my appreciation and thanks to Dr. Girma Awgichew Demeke, Kefyalew G.Egziabher, Eric Atwell, and Michael Gasser who helped me by giving their unreserved advice and materials which are necessary for my research.

My thanks also go to Dr. Ahmed Hussein, the owner and head of HiLCoE, for his unreserved encouragement throughout my MSc. learning at HiLCoE and AAU.

I am very grateful to my family, my wife Haregeweyn Teferra, my children - Amanuel, Zecharias, Yohannes and Betsue, for their abiding support throughout my career and life. I am also very thankful for my mother Birzaf Birru, who always prays for me unceasingly so that I may be well and successful. I also thank my brother Asmelash who was at my side in time of need.

I thank all my well-wishers who directly or indirectly contributed for the completion of this thesis.

Abstract

The major purpose of this thesis is to identify and develop a morphosyntactic tagset for morphosyntactic annotation of texts in Tigrinya, the Ethio-Semitic language having about seven to nine million speakers in Ethiopia and Eritrea (CSA, 2007; CIA 2012; http://en.wikipedia.org/wiki/Tigrinya_language#cite_ref-2). In relation to what is researched, there is almost no Natural Language Processing (NLP) resource for Tigrinya. The researcher thinks that Tigrinya is lucky to start with a comprehensive morphosyntactic tagset development; because morphosyntactic tagset is the foundation for many NLP applications. We have examined the Morphosyntactic features of Tigrinya words and assign a tag that can be applicable for these words in Tigrinya texts.

The thesis focuses only on the development of morphosyntactic tagset based on the morphological and morphosyntactic features of Tigrinya. As a result the developed morphosyntactic tagset for Tigrinya has 18 coarse-grained tags at the higher level, 105 fine-grained tags at the lower level, and even we can extend to more fine-grained features and we get 139 tags. We recommend for researchers to use the 105 tags for their applications, unless and otherwise they have a different purpose which needs the coarse-grained major category 18 tags or the very fine-grained 139 tags, even beyond.

The uses and applications of morphosyntactic tagsets provide an important level of linguistic information to a document. It is useful as a preprocessing step of parsing and most of all it is useful to develop a POS tagger, which is the basis for many higher NLP applications. Students, researchers and professionals like computational linguists/computer scientists who are engaged in Natural Language Processing applications like speech recognition, text to speech, natural language parsing, information retrieval, lexicography and machine translation are the beneficiaries of this research.

Contents

ACKNOWLEDGEMENT I

ABSTRACT II

ABBREVIATIONS VII

LIST OF TABLES X

CHAPTER ONE

INTRODUCTION 1

1.1. GENERAL BACKGROUND 1

1.2. BACKGROUND OF THE LANGUAGE 2

1.2.1. Tigrinya phonology and writing system 3

1.2.2. Morphology and syntax 5

1.2.3. Tigrinya Word Classes 5

1.3. STATEMENT OF THE PROBLEM 6

1.4. OBJECTIVE 7

1.5. SPECIFIC OBJECTIVES 7

1.6. SIGNIFICANCE OF THE RESEARCH 7

CHAPTER TWO

LITERATURE REVIEW 8

2.1. INTRODUCTION 8

2.2. CONCEPTUAL LITERATURE 9

2.2.1. Issues concerning POS tagset development 9

2.2.1.1. Tag, Tagset, and Tagging 9

2.2.1.2. Text 10

2.2.1.3. Morphemes and words 11

2.2.1.4. Morphosyntactic Tagsets 13
2.2.2. Morphological process 14
2.2.2.1. Inflection 14
2.2.2.2. Word formation 17
2.2.3. The morphosyntactic nature of morphosyntactic tagsets 19
2.2.4. Word Classes 20
2.2.5. Criteria and guidelines to develop a morphosyntactic tagset 21
2.2.5.1. Atwell's criteria for POS tagset development 21
2.2.5.2. Geoffrey Leech's standards for corpus annotation (Leech, 2005:29-30) 25
2.2.6. Grammatical input for morphosyntactic tagset development 26
2.2.6.1. Morphological information 26
2.2.6.2. Morphosyntactic information 26
2.2.6.3. Information about special elements in texts 27
2.2.7. Types of morphosyntactic tagsets 27
2.2.7.1. Flat/linear POS tagset 28
2.2.7.2. Hierarchical/fine-grained POS tagset 28
2.3. EMPIRICAL LITERATURE 31
2.3.1. International/global experience on POS tagset development 31
2.3.1.1. POS tagsets developed for English 31
2.3.1.2. Tagsets for Arabic 33
2.3.1.3. Tagsets for Hebrew 34
2.3.2. Local experience on POS tagset development 35
2.3.2.1. Amharic Tagsets 35
2.4. SUMMARY 38

CHAPTER THREE

METHODOLOGY 39

3.1. INTRODUCTION 39

3.2. THEORETICAL FRAMEWORK 39

3.3. DESIGN PRINCIPLES FOR TIGRINYA TAGSET 40

3.3.1. The tag name 40

3.3.2. Classification of words – by form and function 41

3.3.3. What counts as a word? 42

3.3.4. Multiword handling 42

3.3.5. Target users and/or application 42

3.3.6. Adherence to standards 43

3.3.7. Degree of coarseness 43

3.4. POS NAMES (TAG) 44

3.5. WHAT SHOULD BE INCLUDED IN THE TAGSET? 45

3.6. WHAT IS NOT INCLUDED IN TIGRINYA POS TAGSET 47

3.7. SIZE OF THE POS TAGSET 48

CHAPTER FOUR

TIGRINYA MORPHOSYNTACTIC TAGSET (TIGTAGS) DESIGN 49

4.1. INTRODUCTION 49

4.2. DESIGN OF THE TAGSET 49

4.2.1. Nouns <N> 50

4.2.2. Verbs <V> 53

4.2.3. Pronouns <Pn> 56

4.2.4. Adjectives <Aj> 59
4.2.5. Adverbs <Av> 60
4.2.6. Prepositions <Pp> 61
4.2.7. Conjunctions <Cj> 62
4.2.8. Determiners <Dt> 63
4.2.9. Interjections <Ij> 65
4.2.10. Punctuation <Pu> 66
4.2.11. Numerals <Nu> 67
4.2.12. Phrases <P> 69
4.2.13. Compound words <Cp> 75
4.2.14. Contractions <Cn> 76
4.2.15. Symbols <Sy> 77
4.2.16. Abbreviations <Ab> 77
4.2.17. Foreign words <Fn> 77
4.2.18. Residuals <Rs> 78
4.3. Summary of the morphosyntactic tagset 78

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS 84
5.1. CONCLUSION 84
5.2. RECOMMENDATIONS 85
REFERENCES 86
APPENDICES 92
APPENDIX I
TIGRIGNA CONSONANTS (ALPHABET) WITH THEIR PHONETIC DESCRIPTION. 92
APPENDIX II
TIGRINYA SAMPLE TEXT TAGGED WITH THE DEVELOPED MORPHOSYNTACTIC TAGSET 94

Abbreviations

1	First person
2	Second person
3	Third person
Ab	Abbreviation
Ac	Accusative
Aj	Adjective
ARBTAGS	Arabic Tagset
As	Abstract
At	Active
Av	Adverb
Ax	Auxiliary
BNC	British National Corpus
C	Common
Ca	Cardinal
Cj	Conjunction
CLAWS	Constituent Likelihood Automatic Word-tagging System
Cn	Contraction
Co	Coordinating
Cp	Compound
Cr	Concrete
Ct	Countable
Df	Definite
Dg	Degree
Dm	Demonstrative
DM	Derivational Morpheme
Dr	Directional
Dt	Determiner
EAGLES	Expert Advisory Group on Language Engineering Standards
ELRC	Ethiopian Languages Research Center

F	Feminine
Fn	Foreign word
Fu	Future
Ge	Gerundive
Gn	Genitive
H	Honorific
Ic	Indicative
ICE	International Corpus of English
Id	Indefinite
Ig	Interrogative
Ij	Interjection
Im	Imperative
IM	Inflectional morpheme
Ip	Imperfect
It	Intransitive
Lc	Locative
LOB	Lancaster-Oslo-Bergen
M	Masculine
Mn	Main
Mr	Manner
Ms	Mass
MTW	Multi Token Word
N	Noun
Ng	Negative
NLP	Natural Language Processing
Nm	Nominative
NP	Noun phrase
Nt	Neuter
Nu	Numeral
Nv	Verbal noun

Or	Ordinal
P	Phrasal
Pe	Personal
Pf	Perfect
Pl	Plural
Pn	Pronoun
Po	Proper
POS	Part Of Speech
Pr	Present
Ps	Possessive
Pt	Past
Pu	Punctuation
Pv	Passive
Rc	Reciprocal
Rf	Reflexive
Rs	Residual
Sb	Subjunctive
SEC	Spoken English Corpus
Sg	Singular
SOV	Subject – Object – Verb
Su	Subordinating
Sy	Symbol
TAM	Tense, Aspect, and Mood
TIGTAGS	Tigrinya Tagset
Tm	Temporal
Tr	Transitive
Un	Unclassified
UPenn	University of Pennsylvania
V	Verb
WIC	Walta Information Center

List of Tables Page

Table 1: Consonants and Vowels of Tigrinya 4

Table 2: Punctuations of Tigrinya 4

Table 3: Tigrinya noun inflected to possessive noun 51

Table 4: Noun features of Tigrinya 52

Table 5: Verbal features of Tigrinya 56

Table 6: Prononominal features of Tigrinya 59

Table 7: Adjective features of Tigrinya 60

Table 8: Adverb features of Tigrinya 61

Table 9: Preposition features of Tigrinya 61

Table 10: Conjunction features of Tigrinya 63

Table 11: Determiner features of Tigrinya 64

Table 12: Interjection features of Tigrinya 66

Table 13: Punctuation features of Tigrinya 67

Table 14: Numeral features of Tigrinya 69

Table 15: Phrasal features of Tigrinya 75

Table 16: Compound features of Tigrinya 76

Table 17: Contraction features of Tigrinya 76

Table 18: Symbol features of Tigrinya **77**

Table 19: Abbreviation features of Tigrinya **77**

Table 20: Foreign features of Tigrinya **77**

Table 21: Features for the residuals of Tigrinya words **77**

Table 22: Tigrinya morphosyntactic tagset 78

Chapter One

Introduction

1.1. General Background

Through communication, we use different means to transfer our knowledge and information from one another. We people use language to communicate with each other in many ways. There is an ever-increasing availability of information and knowledge in many languages. To transfer this knowledge and information we use different types of technologies, including Natural Language Processing (NLP) technologies. Nowadays there is an ever-increasing NLP technology and interaction between languages and cultures than ever before. This technology of NLP is a multidisciplinary area, which is concerned with the design and building of softwares, which analyze, understand and generate languages that humans use naturally.

In the NLP tasks one of the main applications needed for the NLP application to develop is parts of speech (POS) tagger. This tagger allocates labels/tags to words in a text. The labels are the names given as their POS. This is called POS tagging, allocating tags to words in a written text. Here we understand that in order to construct a tagger, analyzing and developing POS tags is mandatory.

It is also essential to review the language before we pass on to the things concerning POS tagset development. This helps us to provide background for the discussion in the chapters to come. Concerning Tigrinya NLP resources, there are very few Tigrinya specialists and few grammar books on the language. Although Tigrinya is spoken by millions, it is not very well studied. The NLP research was much focusing on supplying resources for highly-resourced languages like English, French, German and other major European and Asian languages like Chinese, Japanese, Urdu, etc. African languages have, however, received too little attention. One of these is Tigrinya, a language which is spoken with about seven to nine million people.

To develop a POS tagset for a language, it is good to understand how other languages have done that, by reading materials that deal with the general trend of POS tagset development and related materials on the subject. Since POS tagset development is the basis for POS tagging, it is also basic to know the grammatical structure of the language under research. It is also compulsory to have some criteria and principle on how to develop a POS tagset. This helps the researcher to handle the process of POS tagset analysis and development to be fine and more accurate.

1.2. Background of the Language

Tigrinya is a Semitic language spoken in the Tigray region of Ethiopia and the highlands of Eritrea as a mother tongue. It is also spoken in other parts of Ethiopia, especially in the northern parts of Wollo and Gonder, as well as by Tigrinya speaking immigrants in different countries around the world. Tigrinya is classified under the north Ethio-Semitic sub-phylum and is grouped with Ge'ez, Tigre, and Amharic which share many linguistic features (Bender et al, 1976; Tesfay, 2002; Daniel, 1998:48).

Its name is Tigrinya (ትግርኛ, /tɨgrɨŋŋa/) or Tigray (ትግራይ /tɨgraj/according to many grammarians and lexicographers like Gallina (1894), Cimino (1904), Bassano (1918), Leonessa (1928), Caressa (1935), and others. It is also spelled Tigrinya, Tigrigna, and less commonly Tigrnia, Tigrina, and Tigriña. Tigrinya is the working language of Tigray region in Ethiopia, and Eritrea. Tigrinya should not be confused with the related Semitic Tigre language, which is spoken in the lowland regions of Eritrea to the north and west of the region where Tigrinya is spoken. Tigrinya speakers are called "Tigraway" for masculine singular, "Tigraweyti" for feminine singular, and when addressing in plural "Tigrawot" or "Tegaru" when speaking in Tigrinya. The Tigrinya speaking people are called "Tigre" for singular, and "Tigrewoch" or sometimes "Tigroch" for plural, by Amharic speakers.

Tigrinya is the second largest language among the Semitic languages of Ethiopia. It is also a major national language in Ethiopia. It is one of the prominent two working languages of Eritrea, which is the first language in number having 50% of the population (Chefena Hailemariam, Kroon, & Walters, (1999); CIA. (2012) *https://www.cia.gov/library/publications/the-world-factbook/fields/2075.html#er).* According to CIA World Fact Book estimation, in July 2012 the

major Semitic languages are: Hebrew 5,799,758; Arabic 221,002,544; Amharic 25,236,775 and Tigrinya 5,722,775 in Ethiopia and 3,347,572 in Eritrea, totally it is spoken by 9,070,347 people (CIA 2012; http://en.wikipedia.org/wiki/Tigrinya_language#cite_ref-2). Therefore Tigrinya is the third most widely spoken Semitic language after Arabic and Amharic.

1.2.1. Tigrinya phonology and writing system

This section introduces the number of phonemes and their orthographical representations relevant to my discussion. And mention in a senstence or two how POS tagsets description hings up on the discussion of phonemens and phonetic representations. Phonology is a branch of linguistics that studies the structure and systematic patterning of sounds in human language (Akmajian, et al., 2001). Since POS tagset is basic for tagging Tigrinya texts which are formed by the combination of different phonemes of the language, it is good to see the phonemes and some changes which happen in their process and how they are represented in the orthography.

Tigrinya is a written language with its own alphabet and writing synstem, lively being used every day. Tigrinya has a set of vowels, consonants and the usual seven-vowel system which is also used by some Ethio Semitic languages.

The writing system of Tigrinya is known as ፊደል /fidəl/, which means alphabet. Each *fidel* represents a consonant-vowel sequence. In other words, each *fidel* is a combination of a consonant phoneme with the seven different vowels indicated below. All graphemes from the first to the seventh order, except sometimes the sixth order has only a consonant, represent consonant-vowel combinations. Characters representing the same consonant followed by different vowels are similar but having slight difference in shape. This combination of consonant and vowel in one character makes the Tigrinya script syllabic[1]. The basic first order consonants 32 and the 5 labilized velars, totally there are 37 consonants. There are 7 vowels which change the basic phoneme of every consonant into different orders (Tesfay, 2002:25; Daniel, 2008:27-35). Tigrinya phonemes are represented by the following graphemic units.

[1] Syllabic script is a script (grapheme) containing a consonant and a vowel together in one grapheme.

Graphemes/letters							
ሀ /hə/	ለ /lə/	ሐ /ħə/	መ /mə/	ረ /rə/	ሰ /sə/	ሸ /ʃə/	ቀ /k’ə/ *[2]
ቐ /χ’ə/*	በ /bə/	ቨ /və/	ተ /tə/	ቸ /tʃə/	ነ /nə/	ኘ /ɲə/	አ /ʔə/
ከ /kə/	ኸ /χə/	ወ /wə/	ዐ /ʕə/	ዘ /zə/	ዠ/ʒə/	የ /jə/	ደ /də/
ጀ /dʒə/	ገ /gə/	ጠ /t’ə/ *	ጨ /tʃə/*	ጰ /p’ə/ *	ፀ /s’ə/ *	ፈ /fə/	ፐ /pə/
labialized consonants							
ቈ /k’ʷə/ *	ቘ /χ’ʷə/*	ኰ /kʷə/	ኸዂ /χʷə/	ጐ /gʷə/			
Vowels							
ኧ /ə/	ኡ /u/	ኢ /i/	ኣ /a/	ኤ /e/	እ /ɨ/	ኦ /o/	

Table 1: Consonants and Vowels of Tigrinya

The above consonants, except the labialized velar and pharyngeal sounds having only their first, third, fourth, fifth, and sixth order, have seven orders. For example, here are the characters representing the seven orders of the first grapheme /hə/: ሀ /hə/, ሁ /hu/, ሂ /hi/, ሃ /ha/, ሄ /he/, ህ /h,hɨ/ and ሆ /ho/, but the labialized ones have five orders without the second and the seventh order, for instance ቈ /k’ʷə/ has, ቈ /k’ʷə/, ቍ /k’ʷi/, ቌ /k’ʷa/, ቈ /k’ʷe/, ቍ /k’ʷɨ/. This is how the letters are represented and Tigrinya texts are written.

Tigrinya has a set of punctuation symbols as listed below. They are similar to those used in Amharic, but the usage may not be identical. We can find the following punctuation marks and many more in Tigrinya texts which are indicated in chapter four under 4.2.10.

፡	space (not in modern use)	።	full stop
፤	semi colon	፣/፧	comma
፡-	preface colon	?	question mark
፧	question mark (not in modern use)	፨	paragraph separator (not in modern use)
,	for numbers to separate every three digit		

Table 2: Punctuations of Tigrinya

[2] Those graphemes with * are ejectives.

The traditional set of numerals used in Tigrinya texts is as shown below.

፩ 1, ፪ 2, ፫ 3, ፬ 4, ፭ 5, ፮ 6, ፯ 7, ፰ 8, ፱ 9, ፲ 10, ፳ 20, ፴ 30, ፵ 40, ፶ 50, ፷ 60, ፸ 70, ፹ 80, ፺ 90, ፻ 100, ፼ 10,000

These numerals have been replaced by the "Arabic" numerals, that are, the same ones used in English. But we may get them in different writings of Tigrinya.

1.2.2. Morphology and syntax

Tigrinya displays a wide variety of derivational phenomena. Tigrinya is one of the morphologically rich languages of Ethiopia. The order of the words in sentences is mostly SOV.

Let us take the following sentence

እቲ	ከልቢ	ነታ	ድሙ	ኣሳገጉዋ
ʔɨti	kəlbi	nəta	dɨmmu	ʔassag^{w}iguwwa
The	dog	the	cat	chased

'The dog chased the cat'

We see here the subject /kəlbi/ (dog) comes first, then comes the object /dɨmmu/ (cat) and at last comes the head verb /ʔassag^{w}iguwwa/ (chased). The morpheme /-u/ in /ʔassag^{w}ig + u + wwa/ is a subject marker (in this case 'the dog'), which is a 3rd person masculine noun phrase (NP-3MSg). The morpheme /-wwa/ in /ʔassagwig + u + wwa/ is an object marker (in this case 'the cat'), which is 3rd person feminine noun phrase (NP). Tigrinya can also be inflected by adding some affixes which signify like tense and number.

1.2.3. Tigrinya Word Classes

Girma (2006:14) in his paper presented at the workshop 'issues on Lexicography' says what word classes are: "The different categories under which words of a certain language are grouped are called word classes, or, traditionally, parts of speech." Word classes (part of speech) can be divided into two broad main categories: closed and open types. Closed classes are those that have relatively fixed membership (Booij, 2007:51; Jurafsky and Martin, 2008:139). An open class is one whose membership is in principle indefinite or unlimited. New items are continually being added, as new ideas, inventions, etc., emerge. Nouns, verbs, adjectives and adverbs are open-

classes, whereas conjunctions, pronouns, etc., are closed ones. For closed classes new items are not regularly added, as is the case with 'open-class' items (Crystal, 2008).

1.3. Statement of the problem

In this era of information and knowledge, many languages are forced to become digitally accessible and develop to join the computational world through NLP. The well-resourced languages have long been integrated into the digital world and that has helped users or the public access their information easily and abundantly. These well-resourced languages construct different types of language technology applications to easily access and retrieve information from any corner of the world. The first task in NLP is developing part of speech (POS) tagger. Part of speech tagging is the process of labeling a part of speech tag like noun, verb, pronoun, adverb, or other word category markers to each word in a text. Part of speech tagging is a prerequisite for many NLP applications. Mainly POS tagsets are the backbones of part of speech tagging and we cannot develop a POS tagger without developing a POS tagset.

Many texts are known to the public in Tigrinya, through printed or the Internet. But it does not have computational linguistics resources. If this lack of resources is to be solved by developing NLP applications which are to be applied to languages, here the first question is how do we analyze and develop a morphosyntactic tagset for Tigrinya? Consequently we can say that there is a great need to develop many NLP application inputs for Tigrinya.

The main focus of this research is to analyze and develop morphosyntactic tagset for Tigrinya. Analyzing and developing a morphosyntactic tagset by itself is a research problem in computational linguistics, which requires its own careful analysis. Infact the main challenge in tagging is ambiguity which is caused by the complexity of morphology of the language and the morphosyntactic tagset. In order to solve the ambiguity, we need a very refined morphosyntactic tagset for every word occurrence in the language. This helps us to develop a more accurate POS tagger.

To my knowledge, there is no work done on this area in Tigrinya and it is invaluable to develop a morphosyntactic tagset for the language. This work will help those who develop POS tagger and

Tree bank for Tigrinya; in turn the tagger will be an input for many NLP works such as information retrieval, machine translation, lexicography, etc.

1.4. Objective

The major objective of this research is to identify features for morphosyntactic tagset for morphosyntactic annotation of texts in Tigrinya. Regarding Tigrinya, according to our knowledge, there is neither a tagger nor a tagset developed at this time. Therefore, it is mandatory to develop first a morphosyntactic tagset before doing any NLP application for Tigrinya.

1.5. Specific objectives

Deriven from the general objective, the specific objectives are:

Identify features for the major categories of Tigrinya words – coarse-grained,

Identify features of words at a middle level – fine-grained,

Identify features of words at a lower level – very fine-grained,

Identify features of phrases and other features found in Tigrinya texts.

1.6. Significance of the research

The uses and applications of POS tagsets provide important linguistic information about a document. It is useful as a preprocessing step of parsing and most of all, for developing a POS tagger, which is the basis for many higher NLP applications.

The beneficiaries of this research output are students, researchers and professionals like computational linguists/computer scientists who are engaged in NLP applications like speech recognition, text to speech, natural language parsing, information retrieval, lexicography and machine translation by using the morphosyntactic tagset as an input for their further work. As a result that will be beneficial to the owners of the language. The resources will make things easier and accessible.

Chapter two

Literature Review

2.1. Introduction

The literatures to be reviewed are of two types, namely: the conceptual literature and the empirical literature. The conceptual literature is concerned with the concepts and theories in the proposed research area. The empirical literature is the literature regarding the studies made earlier which are similar to the one proposed. When we see the technology of Natural Language Processing (NLP), many researches were done focusing on supplying resources for high-resourced languages.

Most of the well studied languages have developed a POS tagset, not for its own sake, but for the purpose of part of speech tagging and for other NLP applications, because without the development of a POS tagset no one can do these applications. Therefore, as mentioned above before doing any NLP application, it is a requirement for a person to develop a POS tagset. For instance, the English language has different types of tagsets for different purposes, like the Brown Corpus, British National Corpus (BNC) - the CLAWS series tagsets with different numbers of tags, the Penn Treebank, the EAGLES standard tagset for European languages including English and the like (Leech, 2005; Hardie, 2003; Atwell, 2008).

Even though there are some tagsets, which are constructed to develop a tagger, the low resourced Ethiopian languages do not have a well-developed tagset appicable for different NLP studies. According to the knowledge of the researcher, the first is Mesfin Getachew's (2001) tagset developed for Amharic POS tagger. There are also other tagsets developed for Amharic for the purpose of constructing a tagger, by Sisay (2005), and Girma and Mesfin (2006) to annotate Amharic news texts; see also the tagset developed for Afaan Oromo by Getachew Mamo (2009) for the same purpose. But Ermias' (forthcoming) tagset is different from the others, because the tagset is developed not to construct a tagger, but to develop a POS tagset for Amharic.

2.2. Conceptual Literature

In this part of the review we will see different concepts and issues dealt within the area of morphosyntactic tagset development, such as tag, tagset, morphosyntactic tagset, annotation and text.

2.2.1. Issues concerning POS tagset development

It is essential to review about the language components related with POS tagset development before we pass on to other things concerning POS tagset. This helps us to provide background for the discussions to come.

2.2.1.1. Tag, Tagset, and Tagging

Crystal (2008:502) defines a ***tag*** as a grammatical label/symbol/mark attached to a word in a corpus to show its class. Tags may be added manually or automatically, by constructing a tagger, to a text. POS tags indicate the classes to which the words in a text belong. There are different types of tags like POS tags, phonetic tags, semantic tags, pragmatic tags, discourse tags, stylistic tags and lexical tags which we can annotate to texts accordingly (Leech, 2005:25-26).

According to Jurafsky (2008) and Leech (2005) a ***tagset*** is a list of symbols or parts of speech used for representing different categories of words. The set of all tags is called a ***tagset***. When words are considered in isolation, they can have one or more tags. But when these words are used in a certain context, the tags representing morphological and syntactic features may reduce to one tag. These tagsets not only differ with each other from language to language, but also vary within the language itself in different ways for different purposes and applications (Jurafsky, 2008:138; Leech, 2005:34).

The reasons for the variation of tags is that taggers give additional information like grammatical features such as number, gender, person, case markers for noun inflections; tense, mood and aspect markers for verbal inflections. The number of tags used in different systems varies depending on the information encoded in the tag. The tagset design plays a vital role when data is tagged according to it and consequently it affects the development of NLP application tools within that language.

The process of assigning part of speech for every word in a given text according to the context is called part of speech ***tagging***. It is the addition of tags, or labels, indicating the word class to which words in a text belong. The lists of all possible grammatical tags that are allowed in a corpus constitute a tagset, and the amount of detail of these tags to be assigned for every word in a text can vary across tagsets. ***Annotation*** is the process of adding some further information (grammatical features like word category, case markers, or other morph features) about the word labeled to each word of the text. It is the manual or automatic tagging of words in texts of the language under processes. According to Leech and Wilson (1996):

> Corpus annotation is the practice of adding interpretative, especially linguistic, information to a text corpus, by coding added to the electronic representation of the text itself. A typical case of corpus annotation is that of morphosyntactic annotation (also called grammatical tagging), whereby a label or a tag is associated with each word token in the text, to indicate its grammatical classification (Leech and Wilson, 1996:6).

Part of speech tagging has many uses in the areas of computational linguistics and natural language processing. It plays an important role in distinguishing words, for instance words which have the same spelling, but different meanings or pronunciation, in Speech and NLP such as Speech Recognition, Speech Synthesis, Information Retrieval, Word sense disambiguation and Machine translation (Leech, 2005:25; Jurafsky, 2008:4,138).

2.2.1.2. Text

There are different types of texts used for NLP purposes. There are spoken texts and written texts. The *spoken ones* are those that are recorded and processed for NLP purpose. The *written texts* are those that are written down, typed, or printed (Microsoft Encarta Dictionary, 2009). But to process them for NLP purpose they should be machine-readable. When compared to the spoken text, it is a relatively easy task to prepare a text for automated POS tagging using a tagger. But spoken language records involve many preprocessing before applying an automatic tagger (Leech, 2005:84).

2.2.1.3. Morphemes and words

Every human being expresses his/her thoughts with words, and every language has enough words to help its speakers express themselves and communicate with each other. These thoughts can be expressed either through speech or writing. In Natural Language Processing one of the central units of the process is the 'word'. Every word should have a group name to which it belongs. We start by developing a label or a tag or a name for each word in a language. Without words we cannot proceed to develop a tag for each word.

> Words play an integral role in the human ability to use language creatively. Far from being a static repository of memorized information, a human vocabulary is a dynamic system. We can add words at will. We can even expand their meanings into new domains (Akmajian, et al., 2001:11).

a. What is a morpheme?

Since this research is concerned with different word formation processes and developing morphosyntactic tags, it is good to deal with the field of linguistics, which has a major role in both word formation and development of morphosyntactic tags; that is morphology, which has the concept morpheme into its centerpiece. Aronoff and Rees-Miller (2003:214) define ***morpheme*** as "the smallest meaningful component of a word". Morphemes are classified into two classes: free and bound morphemes. A ***free morpheme*** is a morpheme that can stand by itself as a free morpheme in a phrase, (e.g., tree as in 'Here is a tree.'). A ***bound morpheme*** is a morpheme that cannot stand-alone, but must be attached to another morpheme:- it only exists as part of a complex word (e.g. the English plural morpheme -s is a bound morpheme and can only occur attached to nouns). Some bound morphemes are known as affixes (e.g., -s), and others as bound base morphemes (e.g., in English cran- in cranberry) (Akmajian, et al., 2001:18).

A base morpheme or a stem[3] is a morpheme to which an affix can be attached. A base morpheme may be free (e.g. tree; therefore tree is both a free morpheme and a free base) or bound (like cran-) (Akmajian, et al., 2001:18-19). It is the stem which forms the basis for every word formation, but not for the whole word form. Stems can be either simplex or complex. A word

[3] A stem is a word-form minus inflectional affixes (Booij, 2007:28).

having only one morpheme which cannot be decomposed in to smaller meaningful units is a ***simple word*** (e.g. girl, house, go, mark). If they are simplex they are called roots[4]. Roots may be changed into stems by the addition of a morpheme, i.e. an affix (Booij, 2007:28-29). A ***complex word*** is a word having two or more morphemes which can be decomposed in to simple constituents (e.g. girl-s, go-ing, re-mark-able) (Booij, 2007:22).

When affixes are attached to the beginning of another morpheme they are called ***prefixes*** (e.g. in English re- in words such as rebuild, rewrite, rethink), and when affixes are attached to the ending of another morpheme they are called ***suffixes*** (e.g. -ize in words such as socialize, actualize, centralize). When affixes are inserted into another morpheme they are called ***infixes*** (e.g., In Bonto Igorot, a language of the Philippines, one can insert the infix -in- to the word *kayu*, meaning ‘‘wood,’’ immediately after the first consonant *k* to form the word *kinayu*, meaning ‘‘gathered wood’’ (Akmajian, et al., 2001:18). When affixes are attached as a prefix and a suffix to a stem/root they are called ***circumfix*** (e.g. in Amharic to the word በላ /bəlla/ meaning ‘he ate’, one can add a negative marker /ʔal-m / as a circumfix (አልበላም /ʔal + bəll + a + m/ ‘he didn’t eat’).

b. What is a word?

The word ‘word’ is the most fundamental unit of linguistic structure (Akmajian, et al., 2001:11). The name ‘word’ may have different meanings looking at it from different perspectives. The definition differs according to the way we use the notion about the word ‘word’. We can isolate the most frequently implied meanings of ‘word’. According to Todd (1987:25-26) the most frequently implied meanings of ‘word’ are the following four. (1) An *Orthographic word* is one which has a space on either side of it. This applies only to the written text. (2) A *Morphological word* is a unique form which considers form only not meaning. For example in English the words ‘bank’ and ‘banks’ are two morphological words, because they are not identical in form. (3) A *Lexical word* is a word having various forms which are closely related to meanings, not by form e.g., ‘bank’, ‘banking’, and ‘banks’ are three morphological words, but one lexical word. (4) A *Semantic word* distinguishes words which may be morphologically identical but differ in

[4] A root is a set of consonants which contains the basic meaning of a lexical item. A root is the base form of a word which cannot be further analysed without total loss of identity (Crystal, 2008:445).

meaning e.g., the polysemous word ‘bank’ is one lexical word, but are two words which are not closely related in meaning, one referring to the edge of a river and the second a financial institution.

The common perception is the orthographic word, the word as delimited by spaces or some punctuation on the page. There are also spoken words, which are distinguished by pauses among them. ***For our purpose of developing morphosyntactic tagset, we take ‘word’ as the text which is seen on a printed page clearly separated by white spaces or punctuation***. But here lies a problem with the compound words (hyphenated words and contractions) which are separated orthographically by white spaces or other means (e.g., In English the two words ‘income’ and ‘tax’ are taken as one compound word ‘income tax’; In Tigrinya the compound word ልቢ ወለድ /lɨbbi wəlləd/, ልቢ-ወለድ /lɨbbi-wəlləd/, ልቦወለድ /lɨbbəwəlləd/, ልቦለድ /lɨbbolləd/ meaning ‘novel’ can be found in texts written in these ways). Such cases found in texts should be addressed very carefully in the development of POS tagset according to the purpose of the Tigrinya morphosyntactic tagset development.

2.2.1.4. Morphosyntactic Tagsets

There are many types of tagsets. Among them are those which are directly related to word classes are POS tagsets and Morphosyntactic tagsets. POS tagsets are those, which show only the main categories or parts of speech of words not the detailed features of words at a lower level; for instance the eight parts of speech (noun, pronoun, verb, adjective, preposition, conjunction, adverb, interjection) which are mostly indicated in the traditional parts of speech classification.

The term *morphosyntax* is derived from two words: morpho - from morphology, which is the study of word formation, and *syntax* which is the study of how words are combined into larger units such as a phrase and a sentence.

> *Morphosyntactic* is a term used in linguistics to refer to grammatical categories or properties for whose definition criteria of morphology and syntax both apply, as in describing the characteristics of words. The distinctions under the heading of number in nouns, for example, constitute a morphosyntactic category: on the one hand, number contrasts affect syntax (e.g. singular subject requiring a singular verb); on the other hand, they require morphological definition (e.g. add -s for

> plural). Traditional properties such as singular, perfect, indicative, passive, accusative, third person are examples of Morphosyntax (Crystal, 2008:315).

We can make many different types of tagsets by creating abbreviated labels for different attributes in order to differentiate one from the other. Here we focus on a tagset which is morphosyntactic, i.e. the tag shows the morphological and syntactic feature of a word. When POS tags carry pieces of information by combining the morphological and syntactic information together, the tags are called ***morphosyntactic tags***. These tags are given to bound morphemes having syntactic functions.

2.2.2. Morphological process

The main focus of this thesis is concerned with words and their categorical names. Words are formed through morphological processes. Words are formed by the process of adding affixes into the stem or root of the word. This is done by inflection, derivation and compounding which we will see each of them below. Words can be processed morphologically in different ways:- through affixation, reduplication and other internal modifications (Payne, 1997:30).

> There are six basic morphological processes by which stems can be formally altered to adjust their meanings to fit their syntactic and communicational context. These six processes are (1) prefixation, (2) suffixation, (3) infixation, (4) stem modification, (5) reduplication and (6) suprafixation (also, suprasegmental modification) (Payne, 1997:29).

This affixes are added in the process of inflection, derivation and other ways of word formation. We will see each one of them in the following subsections.

2.2.2.1. Inflection

Inflectional morphology is concerned with inflectional categories that reflect grammatical processes, for example, pluralisation of nouns (Khoja, 2003:42). Inflectional affixes indicate grammatical relationships and do not change the grammatical class of the stems, the base morpheme, to which they are attached; that is, the words constitute a single pattern, (e.g. in English walk, walks, walked). Inflectional process of words creates different forms of the same lexeme (Booij, 2007:71,102). An inflectional morpheme (IM) does not change either the

grammar category or the meaning found in the word to which it applies (e.g., In English the word 'books' can be classified as: books (N) = book (N) + s (IM)). The more a language is inflectional, its POS tagset contains more number of POS tags or labels (Rama Sree et al, 2008:86). Typical inflectional operations include morphosyntactic agreement features like:

1. Person, number, gender, and case
2. Tense, aspect, mood, voice, negation, politeness

Person, number and gender are grammatical (morphosyntactic) features which need agreement of elements in a phrase or a sentence.

a. Person

According to Bussman (1996:883), ***person*** is Morphological category of the verb used to mark the singular and plural finite verb forms as, speakers referring to themselves, or to a group usually including themselves it is first person (e.g. I, we); second person (e.g. you) is when speakers typically refer to the person they are addressing, 'addressees'; and when other people, animals, things, etc. are referred to, it is third person (e.g. he, she, it, they). Person shows category used in grammatical description to indicate the number and nature of the participants in a situation. Peculiarities of person are usually marked in the verb and/or in the associated pronouns (personal pronouns). Person marking is found for the subject, and sometimes also for other dependents of the verb such as the object. Usually it is found in the above three-way distinctions (Crystal, 2008:358-359).

b. Gender

Gender is a grammatical category used for the investigation of word-classes displaying such contrasts as masculine (m, M, masc, MASC), feminine (f, F, fem, FEM) and neuter (n, neut, NEUT), etc. (Crystal, 2008:206).

c. Number

Number is a grammatical category of nouns which indicates quantity. There are also other parts of speech like adjectives, pronouns, finite verb forms which show number through agreement. The most common categories of number are singular and plural; there are also systems which

have a dual number (two) (like in Greek, Sanskrit, and Gothic) and a trialis number (three) (e.g. some South-West Pacific languages). In some languages there is a paucalis (few) for signifying a small number, as in Arabic (Bussman, 1996:819; Crystal, 2008:335).

d. Case

Bussman (1996:155-156) explains ***case*** as a grammatical category of inflected words which serves to show their syntactic function in a sentence and, depending on the function, involves government and agreement. Case systems may vary from language to language and undergo continuous change. Case identifies the syntactic relationship between words in a sentence. Cases have their values which vary from language to language. These values are referred to as morphosyntactic features. Among the different values for case for different languages can be nominative, for instance in Amharic ቤት /bet/ 'house'), accusative (ቤትን /betɨn/ 'a house'), genitive (ቤቴ /bete/ 'my house'), dative (ለቤቱ /lə-betu/ 'to the house') generally serves to indicate indirect objects, ablative - indicates various types of adverbial relations, locative - serves to identify location, instrumental - identifies the means of accomplishing the action expressed in the verb, etc.

Concerning tense, aspect, mood, etc. we are concerned with verbs. There are three important categories of natural inflection for verbs: tense, mood, and aspect. They are generally expressed in the verb by inflectional means. Payne (1997:233-234) explains tense, aspect, and mood (TAM for short) as,

> Operations that anchor or ground the information expressed in a clause according to its sequential, temporal, or epistemological orientation. Tense is associated with the sequence of events in real time, aspect with the internal temporal "structure" of a situation, while mode relates the speaker's attitude toward the situation or the speaker's commitment to the probability that the situation is true.

e. Tense

Tense is the grammatical term of the relation of the time of an event to some indication point in time, usually the moment the clause is uttered; it is the essential morphological category of the verb which expresses the temporal relation between a speech act and the event described in the utterance, i.e. which places the event spoken of in relation to the temporal perspective of the

speaker (Payne, 1997:236; Bussman, 1996:1182-1183). The ***past tense*** indicates the situation obtained before the moment of speaking, the ***present tense*** indicates the situation obtains at the moment of speaking, and the ***future tense*** shows the situation indicated is located on the time alignment after the moment of speaking.

f. Aspect

Aspect refers to the internal temporal structure of a verb or sentence meaning (Bussman, 1996:96). It is the way in which situations (states or events) are presented as to their internal temporal constituency. ***Perfective aspect*** presents a situation as completed, whereas ***imperfective*** aspect presents the situation as ongoing.

g. Mood/Mode

Payne (1997:244) describes ***mood*** as the speaker's attitude toward a situation, including the speaker's belief in its realty, or possibility. It sometimes describes the speaker's estimation of the relevance of the situation to him/herself. Mood describes the actuality of an event. The terms mode, mood, and modality are often used interchangeably. The ***indicative mood*** is typically the mood for realis (real) assertion, whereas ***subjunctive*** and ***imperative*** forms denote some sort of non-actuality.

2.2.2.2. Word formation

Language is not static; rather, it is dynamic. New entries are created and old entries are changed for many reasons through time. There are many different types of word formation processes; here we will see some of the main ones, particularly derivation and compounding.

a. Derivation

Booij (2007:71-72) describes ***derivation*** as the formation of lexemes by means of affixation, conversion, reduplication, and root-and-pattern morphology. It is an operation of word formation process by which new words/lexemes are created from a base word input. A derivational morpheme (DM) changes the category and/or the type of meaning of the form to which it applies (e.g. Modernize (V) = Modern (Adj) + ize (DM))

According to Appleyard (2006:1101) there are different forms of tense, aspect and mood. The TAM is marked by different stem shapes and person markers.

b. Compounding

Compounding is the joining of two separate words/lexemes to produce a single form, (e.g., bookcase, fingerprint, sunburn, textbook, income tax (Booij, 2007:5)). In Tigrinya there are compounds formed by combining two words through different mechanisms. (Tesfay, 2002: 65-83; Zelalem, 2009:78).

Tigrinya compounds can be right headed[5] or left headed[6] (Zelalem, 2009:78). Some tagset developers give tags for each free morpheme (lexeme) independently, considering it as a different word, but others give one tag for the whole compound, considering it as one word. For instance the word ሳልሳይ ኢድ /salsaj ʔid/ meaning 'different third part' in Tigrinya is a combound noun made of an adjective <Adj> and a noun <N>, but the question is do we categorize it, by assigning different tags, to ሳልሳይ /salsaj/ 'third' as an adjective and ኢድ /ʔid/ 'hand' as a noun or by assigning one tag to both of them as noun?

Apart from the above-mentioned morphological processes, there are also other ways of creating new words. Here we see only those which are necessary for morphosyntactic tagset development those which need to be tagged.

c. Abbreviations/short forms/acronyms

Abbreviation is the shortening form of a word or phrase to be used to represent the full form. There are several ways of making abbreviation or shortening of words, like: Initialisms or alphabetisms (e.g. TV for television), acronyms (e.g. NATO, UNICEF), clipped forms or clippings (e.g. ad from advertisement), and blends (e.g. brunch from breakfast and lunch, the Amharic word መቼት /mətʃet/ 'setting' from መቼ /mətʃe/ 'when' and የት /yət/ 'where'). Researchers on morphosyntactic tagset development should consider such short forms in texts of the language under research (Crystal, 2008:2, 27).

[5] Right-headed compound – having its rightmost element as its head (McCarthy, 2002).
[6] Left-headed compound – having its leftmost element as its head.

Acronymy and blending are uncommon means of word formation in Tigrinya. The only one "genuine" acronym is ጉዝዳ /guzda/ 'GUZDA' derived from the words ጉያ /guyya/ 'running', ዝላ /zɨlla/ 'jumping' and ዳርባ /darba/ 'throwing' (Zelalem, 2009:80)

d. Borrowing and Foreign words

Borrowing is taking over of words from one language or dialect to another language or dialect (Crystal, 2008:84). Tigrinya has borrowed many words from foreign and indigenous languages. These words can be similar in their form even meaning having some vocalic changes (Zelalem, 2009:69). E.g. The English word 'television' is borrowed directly into Ethiopian languages, for instance the Tigrinya word ቴሌቪዥን /televiʒɨn/.

Those foreign words are those words which are found written in texts other than the language's writing system, but in a foreign writing. For instance if we take the Tigrinya phrase ምሉእ ሓሳብ /mɨluʔ ħasab/ meaning sentence we can find it written as, ምሉእ ሓሳብ (sentence) in texts. Therefore the English word 'sentence' should be given the name foreign word <Fn>.

e. Backformation

Backformation is a way of word formation; word of one type (noun) is processed to form another word of a different type (verb). (Booij, 2007:40)

e.g.	English	donation (N)	donate (V)
		emotion (V)	emote(V)
	Amharic	ታይፕ /tajɨp/ (N)	ተየበ /təjjəbə/ (V)
		Type writer	he typed
	Tigrinya	በጀት /bədʒət/	በጀተ /bədʒdʒətə/
		Budget	he budgeted

2.2.3. The morphosyntactic nature of morphosyntactic tagsets

As it is defined by David Crystal (2008) in his Dictionary of Linguistics and Phonetics the morphosyntactic tagset represents syntactic features by morphological means, i.e. through the presence of bound morphemes and morphological processes. The morpheme under process goes beyond the lexeme to the phrasal or sentence structure showing syntactic features. Such morphemes may play a role both in morphology and syntax. For instance we can take nouns,

constituting a morphosyntactic category: number affects syntax (e.g. singular subject needs a singular verb); on the other hand, number requires morphological definition (e.g. add -s for plural in English).

We can see some examples from Tigrinya. In Tigrinya the subject should agree with the verb; even determiners should agree with the subject and/or object in number, gender, and person.

እቶም	ቆልዑ	ነተን	ደማሙ	ሰገሩምወን
/ʔɨt-om	k'olʕu	n-ət-ən	dəmamu	səg^{w}ig-omɨ- wwən/
The [3MPl]	children [Pl]	to the [3FPl]	cats [Pl]	chased [3MPl 3FPl]
'The children chased the cats'				

The verb /səg^{w}ig-om-wwən/ by itself means 'they chased them' indicating the subject /k'olʕu/ 'children' and the object /dəmamu/ 'cats'. That is why we need a POS tagset which is morphosyntactic for the annotation of texts.

Morphosyntactic tagsets are usually developed for the purpose of the morphosyntactic annotation of corpora. While presentations of morphosyntactic systems of various languages found in textbooks and grammars may be sufficient for many linguistic purposes, the task of assigning a morphosyntactic tag to each word in a large corpus requires a codification of such a system. The resulting tagset must exhaustively specify the range of grammatical classes (parts of speech) assumed for the language, morphosyntactic categories appropriate for particular classes, and possible values of these categories.

2.2.4. Word Classes

A ***word class*** specifies the class to which a word belongs most of the time. The assignment is made on a lexical basis without reference to a particular context. There are major word classes, and some of them have sub-classes (Jurafsky, 2008:139).

We understand that the morphosyntactic features of the language and the degree of granularity[7] of these morphosyntactic features, domain etc., decide the tags in the tag set. Before starting to develop a POS tagset for a language, a researcher should consider the following.

2.2.5. Criteria and guidelines to develop a morphosyntactic tagset

When we assess the literature concerning POS tagset design, we do not get any standard design for POS tagset development. We will see some of the characteristics, criteria, principles and guidelines stated by scholars.

2.2.5.1. Atwell's criteria for POS tagset development

There are different dimensions which should be taken into consideration when developing a POS tagset. According to Atwell (2008) these dimensions should be treated very carefully in order to get a very good POS tagset for the language under research.

> The rival tagsets display differences (and similarities) along several dimensions. These dimensions are in effect choices to be made by developers of new POS-tagsets, for English or another language; in developing a new tagset, the designer must decide how to handle each dimension. Once a researcher has decided it would be useful to add part-of-speech tags to their corpus, they must decide on the tagset: decide on the set of grammatical tags or categories, and their definitions and boundaries. (Atwell, 2008:506)

Developers of a tag-set for a corpus must also take into account a range of issues, including: mnemonic tag names; underlying linguistic theory; classification by form or function; analysis of idiosyncratic words; categorization problems; tokenisation issues: defining what counts as a word; multi-word lexical items; target user and/or application; availability and/or adaptability of tagger software; adherence to standards; variations in genre, register, or type of language; and degree of delicacy of the tag-set (Atwell, 2008:502).

[7] One of the important concerns in developing a tagset for a language is granularity - coarseness and fineness. They refer to the broad annotation and the finer annotation, respectively of any grammatical category (Sinha, 2010).

a. Mnemonic tag names

A name given to a tag should be simple, be able to be remembered easily, and shouldn't be given to other tags. A researcher engaged in POS tagset development (a linguist, computational linguist, computer scientist, information scientist, etc) has to develop the tags by using abbreviations or acronyms. For example, for using an abbreviation the designers of the BNC (British National Corpus) tagsets decided that NNS might be mistakenly interpreted as noun-singular, so instead use NN1 for singular noun and NN2 for plural noun. (Atwell, 2008:506)

b. An underlying linguistic theory

When a new tagset is developed by a linguist, they will inevitably be influenced by the linguistic theories they adopt. Here Atwell (2008:507) says, "Some corpus linguists may claim their part of speech tagsets are "theory-neutral"; but then why so many rival part of speech tagsets are abound? It is really not possible to have a theory-neutral annotation; every tagging scheme makes some theoretical assumptions."

Most developers use the traditional part of speech to develop a POS tagset for that language. But there are some others who use different linguistic theories. That is they go beyond the traditional approach to add some more features to categorize words of the language under research for the purpose they want to accomplish. Sinha (2010) also agrees with this notion.

> It is also possible that the developers are application-oriented rather than linguistic-theory oriented. For example, the Machine Learning researchers using a POS tagged corpus for their experiments are primarily concerned with Machine-Learnable tagging than with a specific linguistic theory. Therefore, such researchers will develop POS tagset accordingly. Paradoxically, this view has dominated the development of POS tagset to a large extent (Sinha, 2010: http://samarsinha.blogspot.com/2010/11/issues-in-pos-tagset design.html).

c. Definition of Parts-of-Speech, by form or function

In order to give a name to a tag, the POS definition of a word is important. Otherwise, there will be more ambiguity. Traditionally a word is defined by form, paradigmatically, or function, syntagmatically. Paradigmatically we can have the inflectional form of a word (for example: a

word is a noun if it can be inflected for number), and syntagmatically by looking its functional position in specific sentence-slots such as head of a noun phrase. (Atwell, 2008:507)

d. Handling Words with special and idiosyncratic behavior

We can find words which have special behavior to categorize them into the traditional or other linguistic theory. These words should be analyzed differently. For example, "a" is allowed a special article tag <AT> in the Brown and LOB tagsets, but is lumped in the determiner <DT> in the UPenn tagset (Atwell, 2008:507).

e. Handling categorization problems

When a linguist categorizes words in a fine-grained way, he/she should define each tag clearly without ambiguity to tag the corpus consistently. For instance in English a word can have more than one category. The word 'cut' can be a noun or verb when put in to different syntactic positions. This should be handled clearly in the naming work. This sends us to the analysis of the syntactic function of words (Atwell, 2008:508).

f. Multi-word handling

Multi-word lexical items are sometimes called idiomatic phrases, multi token word (MTW) (also commonly known as multiword expression in computational literature) (Atwell, 2008; Sinha, 2010). Such words should be treated well in order to make the corpus unambiguous and acceptable for many applications. Some schemes give tag names for each word and others give one tag name for the phrase. For example, the BNC tagset treats "for example" as a single adverb (RR21 RR22 or AV021 AV022); whereas other tagsets assume this are preposition + noun. Sometimes a token contains several POS-tags (at least at some level) and sometimes several tokens have a common POS tag (Atwell, 2008:509).

g. Target users and/or application

In POS tagset development the first purpose to consider is to satisfy the customer. With this in mind the developers should consider the application they are anticipating. Though they seem different they go hand in hand. A POS tagset could be developed for teaching and learning purpose, but this purpose goes with the customers – the learners and teachers (Atwell, 2008:509).

h. Availability and/or adaptability of tagger software

The developer of POS tagset can use a tagger in order to develop a POS tagset. If the language under research has a tagger, it is good to see the tagsets used for the application. If the language doesn't have a tagger, it is also good to see sisterly languages which have such tagsets. By using them as a spring-board, the researcher can develop POS tagset for the language under way, by extending or decreasing (decomposing) the tagset (Atwell, 2008:510).

i. Adherence to standards

For researchers who are developing a POS tagset for a language, it is good to follow some standards used for the language. The EAGLES guideline is a standard to develop POS tagsets for European languages. Linguists developing a tagset for a language without any prior POS tagset may try to conform to agreed standards of another language, better if the language is sisterly. We can say that this is the de-facto widespread adoption of an existing tagset with significant qualifications or supporters (Atwell, 2008:510).

j. Considering genre, register or type of language

A POS tagset developer can consider the type of language he/she is researching. The language can be written or spoken. They may include more informal or non-standard vocabulary and grammar. For example the LOB tagset was readily applied to the Spoken English Corpus (SEC). The International Corpus of English (ICE) and British National Corpus (BNC), corpora contain both written and spoken language, and the respective ICE and BNC tagsets cover both (Atwell, 2008:511).

k. Degree of coarseness

There are many concerns about the POS tagset development for a language. Sinha (2010) explains about the granularity of a POS tagset:

> One of the important concerns in developing a tagset for a language is granularity - coarseness and fineness. They refer to the broad annotation and the finer annotation, respectively of any grammatical category. (Sinha, 2010: http://samarsinha.blogspot.com/2010/11/issues-in-pos-tagset-design.html)

A POS tagset can be coarse, fine-grained or very fine-grained. That is, it can be developed on higher or lower level. The higher level may contain for instance, the eight parts of speech of the language, and the lower level can branch down by adding some paradigmatic and/or syntagmatic information of the language. This may result in the difference of the number of tags. For example, the main reason for the difference in number of tags, or degree of delicacy, between the LOB and UPenn tagsets was the user-group foreseen by the tagset developers (Atwell, 2008). The general corpus developers, as a principle, prefer to maximise linguistic enrichment by designing tagset in such a way that the annotation can be customised according to the needs of the application. The reason for more number of tags in a tagset is to precisely capture all linguistic criteria which describe morphosyntactic features in detail (Sinha, 2010).

2.2.5.2. Geoffrey Leech's standards for corpus annotation (Leech, 2005:29-30)

Leech (2004) asserts that the usefulness of annotated corpora depends crucially on whether the annotation has been well planned and well carried out. Therefore it is important, to see the recommended set of standards of good practice to be observed by annotators wherever possible.

a. Annotations should be separable

Annotations (labels/tags) are some added information to words or words in a text. The text and the labels should be very well seen and separable. These annotations shouldn't lose any information from the original data.

b. The annotation practices should be linguistically consensual

This issue is a debatable issue in the annotation of corpora. Leech (2005:30) expresses his uncomfortability of the issue by saying, "Even an apparently simple matter, such as defining word classes (POS), is open to considerable disagreement…. If this is reasonable, then an annotation scheme can be based on a 'consensual' set of categories on which people tend to agree." According to Leech, in order to make the annotated resource more useful and sharable, it shouldn't stick to a particular linguistic theory; this means theory neutrality is better than to stick to one theory.

c. Annotation practices should respect emergent de facto standards

In order to make the annotation more usable by the research community, it is mandatory to use a de facto standard in the annotation scheme to which researchers agree most of the time.

> By de facto standards, I mean some kind of standardisation that has already begun to take place, due to influential precedents or practical initiatives in the research community. These contrast with de iure or 'God's truth' standards, which I have just argued do not exist. 'God's truth' standards, if they existed, would be imposed from on high. De facto standards, on the other hand, emerge (often gradually) from the research community in a bottom-up manner (Leech, 2005:30).

2.2.6. Grammatical input for morphosyntactic tagset development

2.2.6.1. Morphological information

Since every POS tag developed is to be assigned for every word in the text, and in order to give them a class it is mandatory to learn about their class/POS. As it is stated in the above Leech's idea the common or consensual elements should be considered concerning the grammatical information of the language under research. Many languages commonly share the following parts of speech, as they are indicated in the EAGLES standards for morphosyntactic annotation (Leech, 1996:10) and in Jurafsky (2008:137-138). They are noun, verb, adjective, pronoun/determiner, article, adverb, adposition/preposition, conjunction, numeral, interjection, unique/unassigned, residual and punctuation. Each major category may also be sliced into well known subcategories. For example noun can be sub categorized into gender having the values masculine, feminine and neuter, and number having the values singular and plural (Leech, 1996:8; Hardie, 2003:71). This can be handled by including the morphological information of the word.

2.2.6.2. Morphosyntactic information

The morpheme under process goes beyond the lexeme to the phrasal or sentence structure showing syntactic feature. Such morphemes may play a role both in morphology and in syntax (Leech, 2008:6). We can see some examples from Tigrinya. In Tigrinya the subject should agree with the verb; even determiners should agree with the subject and/or object in number, gender, and person.

እቶም	ሰብኣይ	ነቲ	ወንበር	ሰሪሑምዎ
/ʔɨt-om	səbʔay	n-əti (nɨ-ʔɨti)	wənbər	seriħ-omɨ-wwo/
the	man	the	chair	made he it
[PnH3MSgS]	[NCMSg]	[Dt3M]	[NCSg]	[PnHM3SgO]

'The man made the chair'

Morphosyntactic tagsets are usually developed for the purpose of the morphosyntactic annotation of corpora.

While presentations of morphosyntactic systems of various languages found in textbooks and grammars may be sufficient for many linguistic purposes, the task of assigning a morphosyntactic tag to each word in a large corpus requires a codification of such a system. The resulting tagset must thoroughly specify the range of grammatical classes (parts of speech) assumed for the language under research, morphosyntactic categorical features appropriate for the particular classes, and possible values of these categories.

2.2.6.3. Information about special elements in texts

For annotation purposes everything in a text is considered as word and a label/tag should be given for each of them. These words may contain beyond the eight parts of speech to categorize the words into different classes. For instance we can have different types of multi-word expressions and numerals. All these should be addressed very well during the POS tagset development. Beyond such a scope the rest are considered as residuals.

All the above information is given as a major category, subcategory and morphological information in a single tag. The sequence of characters from left to right will represent a hierarchy of features ordered from the most general to the most specific (Hardie, 2003).

2.2.7. Types of morphosyntactic tagsets

The aim of corpus annotation is to make best use of information content so that the tagged corpus can be used for a variety of applications. But as a matter of fact, the applications are not known in advance, for this reason; the level of linguistic annotation required is also unknown. The general corpus developers, as a principle, prefer to maximize linguistic enrichment by designing tagset in such a way that the annotation can be adapted according to the needs of the

application. However, in POS tagset design, there are two schemes for granularity. The coarse-grained annotation has less number of tags than the fine-grained annotation, and assists in higher accuracy in the way of manual tagging and in efficient machine learning. Regardless of such advantages, the coarse-grained POS tagset is of less use as it does not give much appropriate information on POS. On the other hand, a fine-grained annotation give a very large number of information, but creates a difficulty for automatic tagging as it maximizes tag options for a given token leading to computational complication (Sinha, 2010: http://samarsinha.blogspot.com/2010/11/issues-in-pos-tagset-design.html).

2.2.7.1. Flat/linear POS tagset

A flat POS tagset is a non-hierarchical set of tags having few labels. Flat tagset just lists down the categories appropriate for a particular language without any condition for modularity or feature reusability. Since this type of POS tagset is not deeply hierarchical (having many tags), it is not decomposable. Hardie (2003) explains the category of flat tagset's "major categories can have subdivisions, but there are no further subcategorisations." The C5 tagset is classified as flat by Hardie as it is characterized by Cleoren (Hardie, 2003:51). They don't have more complicated paradigms. For instance Mandarin has a flat POS tagset. A flat tagset has a large number of independent categories which are not presented within the tagset as subcategories of a more general category. Hierarchical tagsets are more fine-grained, decomposable and less number of sets/lists of tags when going to a higher level (Hardie, 2003:48).

2.2.7.2. Hierarchical/fine-grained POS tagset

A tag in a POS tagset contains a string of connected elements and each tag is expressed by mnemonic letters that could be remembered easily. Each of these elements signifies a single, atomic grammatical feature. Each letter in a tag means something (e.g. <NP> means Proper Noun). With regard to the recognized structuring the tagset is a logical tagset, implying that the relations between the word categories can be represented as a hierarchical tree. Such an arrangement therefore reflects the relations between word categories. The term "hierarchical", when used of a tagset, means that the categories in that tagset are structured relative to one another. Rather than a large number of independent categories, a hierarchical tagset will contain

a small number of categories, each of which contains a number of sub-categories, each of which may contain sub-sub-categories, and so on, in a tree-like structure (Hardie 2003:48). Hardie (2003:75) explains the tree of hierarchy suggesting that of the three types of information in a tagset, major word classes which is the highest in the hierarchy, followed by subclassifications, and lastly morphological features with their values.

Leech (1997:27) suggests an identification standard for the logical tagset:

> The idea of a logical tagset is that the relations between the word categories symbolized by the tags should be representable as a hierarchical tree (not a POS tree, but a tree of features and attributes), with attributes being inherited from one level of the tree by another.

The attributes of a word category are inherited from one level of the hierarchy to the next. It is therefore useful to make sure that the tag naming consensus reflects the logical structure of the classification covered by the tagset. This is also reflected in the tagset – some tags refer to purely morphological features, whereas others are more syntactically oriented. Features are pairs of attributes and values, such as 'Number = plural'. The ordering of elements in a tag, which is decomposable, defines the position of the element in the hierarchy. Hierarchical-decomposable tagsets permit us to explore for different sections of the paradigm. When we design a hierarchical-decomposable tagset, two features are typically in proportion to where the tagset stands, which is the depth of the hierarchy and total number of tags.

> A tag is considered to be "decomposable" if the string that represents that tag contains one or shorter strings or single characters that are meaningful out of the context of the original tag and may be found elsewhere in the tagset with the same meaning. For example, any noun tag which combines an N for "noun" with other characters to indicate other features of the word is decomposable (Hardie, 2003:48).

Most of the tagsets developed for some purpose in different languages follow a hierarchical structure by developing fine-grained tags for their application and use (Sawalha, 2011:106). The advantage of fine-grained tagsets is that the words in a corpus can be well addressed and disambiguated. But the annotation process may be slow when compared with the coarse-grained tagsets (Sawalha, 2011:106).

A hierarchical tagset presents major categories at a higher level and goes down by creating other trees of lower level which inherit some of their attributes. A morphosyntactic tagset can have many layers in the hierarchy – the top level having the major lexical categories followed by types (of these categories) in the next layer and features (attributes) carrying finer details (values) placed in the lower layers. The reason for more number of tags in a tagset is to precisely capture all linguistic conditions which describe morphosyntactic features in detail (Hardie, 2003:56).

a. Categories

Categories are the primary grammatical classes to which the words belong. 'Grammatical' means grossly the parts of speech through which each individual word is recognized, (e.g., noun, verb, adjective etc.) The Category level tags are determined on the basis of the categorization features of the word. It decides on the Parts of Speech the word that it belongs to (Khoja, 2003:64; Hardie, 2003:47).

b. Types

Types are the subclasses or finer specifications of the categories, which are determined on the basis of either form or function. E.g., Common Noun, Proper Noun etc. are the subcategory of the category 'Noun'. Types are fine-grained to sub classifications of the categories. Each Type, group words which are similar in terms of their characteristics, i.e., whether they form a class on the basis of their distribution, references etc. (Hardie, 2003:56, 91).

c. Attributes

Attributes are the set of basic morphosyntactic features of a type, like, gender, number, person etc. Each Type has fixed and exhaustive set of Attributes that accumulates the possible morphosyntactic features to be attached to each word. Some attributes are mandatory and some are optional. Mandatory attributes are those which contribute to the basic meaning of the word in its grammatical specification, (e.g., Tense in Verb is mandatory (Hardie, 2003:56).

d. Values

In the hierarchy, values match up with attributes. At each level, tags are defined as morphosyntactic attribute-value pairs (e.g. Number is an attribute that can have the values Singular or Plural (Hardie, 2003:56).

According to Hardie (2003), many well-resourced languages have developed hierarchical decomposable POS tagsets. For example, Italian has a DMI tagset (Leech and Wilson 1996/1999), Urdu (Hardie 2003), Arabic (Khoja, et al., 2001), German has a STTS tagset (http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html) and C7 is an English hierarchical POS tagset which can be decomposed to higher levels.

2.3. Empirical literature

Here we present a selection of relevant literature done in other languages on the proposed domain. These languages can be put as POS tagsets developed for in a global/international context and the Ethiopian local context.

2.3.1. International/global experience on POS tagset development

Every society in the world is trying to climb to the ladder of natural language technology. This technology is developing faster and manipulating its natural language to make things easier and easier. It deals with applications like speech synthesis systems, machine translation, information retrieval, etc. In order to develop such applications every language needs a POS tagger, a parser or a tree bank, and the basis for these all is developing a POS tagset.

As we have discussed above every language, especially the well-resourced ones are developing different types of POS tagset having different sizes for diverse purposes. One of these is English, and we will see some of its known POS tagsets developed for various applications.

2.3.1.1. POS tagsets developed for English

English as it is an international lingua-franca enjoys very many POS tagsets constructed for different purposes at different times. Some of them are the following.

a. The Brown Corpus tagset

Brown University started to work on an electronic corpus in 1961, at Rhode Island. The corpus was called the Brown corpus and it containes one million words which were all publications of that year (1961). The corpus was completed in 1964 and since then has been used for hundreds of corpus work (Khoja, 2003). The Brown Corpus tagset has 87 tags. The release of the first Brown corpus represented the start of tag set design as scheme for morphosyntactic annotation of

corpora. The Brown Corpus was the first to be part of speech tagged, and it was the first major American corpus, which led to its widespread use in American computational linguistics research. It is believed that other tagsets are evolved from the Brown tagset (Atwell, 2008).

b. The Penn tree bank tagset

The Penn Treebank is a corpus consisting of over 4.5 million words of American English. The tagset contains 36 POS tags and 12 other tags for punctuation and currency symbols, all in all 48 tags. The design of their POS tagset and presenting the tagset itself, they describe a two-stage tagging process, in which the text is first assigned with the POS tags automatically and then corrected by human annotators (Marcus et al., 1993:313). Marcus et al. (1993:314) write:

> The Penn Treebank tagset is based on that of the Brown Corpus. However, the stochastic orientation of the Penn Treebank and the resulting concern with sparse data led us to modify the Brown Corpus tagset by paring it down considerably. A key strategy in reducing the tagset was to eliminate redundancy by taking into account both lexical and syntactic information. Thus, whereas many POS tags in the Brown Corpus tagset are unique to a particular lexical item, the Penn Treebank tagset strives to eliminate such instances of lexical redundancy. The Penn Treebank tagset, like many others, is based on that of the Brown Corpus.

The one thing the Penn Treebank is good at is by reducing the size of the tagset it reduces the chances of tagging inconsistencies which are found in the Brown corpus. It also allows words to be associated with more than one POS tag. In principle, annotators can tag a word with any number of tags, but in practice, numerous tags are limited to a small number of frequent two-tag combinations (Marcus et al., 1993:313).

c. The BNC (British National Corpus) tagset

BNC was developed from 1991 up to 1994. The corpus is encoded with the Standard Generalized Markup Language (SGML) to represent POS tags, and other structural properties of texts. The BNC is word class-tagged using a set of 57 tags (known as C5) which is referred as the "BNC Basic Tagset". The BNC tagset is also known as the BNC Enriched Tagset or the C7 tagset. The tagset is a larger set of grammatical word labels than the Basic Tagset which is the BNC - C5 tagset: it has 139 tags (minus punctuation tags), instead of 61. The BNC Enriched Tagset (C7) has been used for the tagging of the Core Corpus of 2 million words of spoken and

written English, yet having over 100 million words. We can take the C7 tagset as a hierarchical, logical tagset which organizes the tags by different levels from the major category level to the finest level of category (http://info.ox.ac.uk/bnc).

d. The EAGLES tagset

The Expert Advisory Group on Language Engineering Standards (EAGLES) project is one of the most known and used hierarchical POS tagsets of English. It is a program to bring the various tagsets of European languages into a common standard tagset. The Eagles projects use three levels of categories (obligatory, recommended and optional). The “obligatory" level of category is limited to parts of speech or word classes as Noun, Verb, Conjunction, etc. Here they identified 13 tags: N(Noun), V(Verb), AJ(Adjective), PD(Pronoun/Determiner), AT(Article), AV(Adverb), AP(Adposition), C(Conjunction), NU(Numeral), I(Interjection), U(Unique/unassigned), R(Residual), and PU(Punctuation).

The recommended level of category applies to well-known attributes used widely in the description of European languages: e.g.) Number, Gender, Case, Person, Finiteness, Mood, Tense, Voice, Status, Degree, Possessive, Category, and Function. These are specified below and under the part of speech headings assigned under the major category which is the obligatory level.

The "optional" level defines tags that are either syntactic, semantic, or language specific. At the optional level, the guidelines clearly have a weaker importance, and should not be regarded as compulsory in any sense, but simply as a presentation of possibilities approved by current practice. The EAGLES tagset design has a very good strength which is extensible to other languages (Hardie, 2003). The “Intermediate” option for other languages is also a strength, that any language can adapt the guide line standard to itself.

2.3.1.2. Tagsets for Arabic

Shereen Khoja (2003) has developed a POS tagset for Arabic. The tagset is very detailed and is based on the traditional Arabic categorization. It states all Arabic words which are derived from nouns, verbs or particles and all tags inherit properties from these three. It is from these three main categories that the rest of the language is derived. It contains 177 tags. All the subclasses of

these three main classes inherit properties from the parent classes. According to Khoja (2003:64):

> Arabic is very rich in categorizing words, and contains classes for almost every form of word imaginable. If all the subclasses described by Arabic grammarians were used, the size of the tagset would soon reach more than two or three hundred tags. For this reason, only the main classes and subclasses have been chosen. But because of the way all the classes inherit from others, it would be quite simple to extend this tagset to include more subclasses, or simplify it and make it smaller.

The article by Alqrainy and Ayesh (2006:2787-2788) is focused on the development of tagset for automated POS tagging in Arabic. It is an extension of the "tagsets for the morphosyntactic tagging of Arabic" developed by Shereen khoja (2003). The background of the tagset and the EAGLES guidelines overview are also presented. The proposed Arabic tagset is based on the inflectional morphology system. Accordingly the POS tagset is organized in a hierarchical way.

Features play a great role in adding linguistic attributes to Arabic words which help to assign the most likely tag of the word in POS tagging system and in indicating pronunciation and grammatical function of the words. In Arabic, short vowels are not part of the Arabic alphabet. They are used in both Noun and Verb in Arabic Language. They indicate the case of the noun and the mood of the verb (Alqrainy and Ayesh, 2006).

As the researchers noted, the tagset discussed here is not being developed in accordance with the EAGLES guidelines for morphosyntactic annotation of corpora. Arabic is very different from the languages for which EAGLES was designed, and belongs to the Semitic family rather than the Indo-European. These tags contain a large amount of information and add more linguistic attributes to the word. Therefore they have developed a fine-grained tagset for Arabic (Alqrainy and Ayesh, 2006:2788-2790).

2.3.1.3. Tagsets for Hebrew

Sima'an et al. (2001) have developed POS tagset for Modern Hebrew in order to build a Treebank. They developed 31 POS tagsets and 18 syntactic tagsets for Hebrew. They have used the English tagset developed by the Penn Treebank as a model. But the tagset is extended to the agreement of the features of Hebrew. Beside the addition of features, other modifications on the

Penn tagset were motivated by occurrences or categories that are special to Hebrew or by cases where the peculiarities of the Penn tagset are insufficient for their application (Sima'an et al. 2001).

2.3.2. Local experience on POS tagset development

2.3.2.1. Amharic Tagsets

There are some POS tagsets developed for Amharic. There is also one POS tagset developed for Afaan Oromo to construct Afaan Oromo POS tagger. Most of the Amharic POS tagsets are developed for the purpose of constructing a POS tagger. Only one POS tagset is developed for the sake of POS tagset development. The next paragraphs review POS tagsets developed for Amharic.

a. Mesfin's tagset

Mesfin Getachew (2001) developed a POS tagset for Amharic in his thesis for the partial fulfillment of his Master's Degree. These POS tags are designed on the basis of the review made regarding the linguistic properties of the Amharic word classes. He developed 25 tags to be applied on the POS tagger. To some extent it is hierarchical, because it goes down from the major to the next subcategories (Mesfin, 2001:1)

b. Sisay's tagset

There is also another Amharic POS tagset developed by Sisay Fisseha Adafre (2005) containing 10 tags. This tagset is a reduced version of Mesfin's tagset. The tags are: Noun (N), Verb (V), Auxiliary verbs (AUX), Numerals (NU), Adjective (AJ), Adverb (AV), Adposition (AP), Interjection (I), Residual (R), and Punctuation (PU). He actually established this reduced version for practical reasons (Sisay, 2005). Just this shows us that it is a coarse-grained flat POS tagset which does not have any subcategory to show the morphosyntactic category of the words of Amharic. The names given for the categories are different from that of Girma and Mesfin's representation, giving abbreviation like letters.

c. Girma and Mesfin's tagset

The other POS tagset developed is that of Girma Awgichew Demeke and Mesfin Getachew (2006). This was a project done at the Ethiopian Languages Research Center (ELRC)[8], named as "The Annotation of Amharic News Documents". The annotation of 210,000 words that occur in the 1065 Amharic news documents with appropriate POS or morphosyntactic categories was done manually by the researchers at ELRC. The tagset is developed just for the annotation of these Walta Information Center (WIC) news texts.

The researchers decided the basic POS categories as nouns tagged as (N), pronouns (PRON), adjectives (ADJ), adverbs (ADV), verbs (V), prepositions (PREP), conjunctions (CONJ), numeral (NUM), interjection (INT), punctuation (PUNC) and for the unclassified (UNC). These 11 tags are the basic or obligatory ones. This tagset of 30 tags add tags representing phrases of nouns, pronouns, adjectives, verbs and numerical. These tags are represented by taking the acronyms of the categories of the words.

d. Ermias' tagset

The last one we saw is the comprehensive POS tagset developed by Ermias (forth coming). It explains what morphosyntactic tagging and annotation is and gives some notion about tagset development in general. Ermias gives a comprehensive background on Amharic, its emergence in literature, its use, its writing system, its morphology (as a morphologically rich language), and its inflectional and derivational ability. He also shows the two cultures of Amharic word classes.

Ermias set two recommendations; that is linguistic neutrality (not to be obligated to one linguistic theory, develop the tagset by starting from where there is consensus and moving to disputed areas) and adherence to defacto standards. In addition, he mentioned that learnability of the tagset by human annotator is important. He states that for the reason of linguistic related details there could be convergence across languages.

Ermias has put the use of the tags as higher level and lower level. He sets two characters long acronym, examples and Amharic equivalents for every tag. For instance a verb (VB)[9] can have the following features in its tagset.

[8] Now it is called Academy of Ethiopian Languages and Cultures (AELC).

[9] Made available by the author.

Category	Verb	
Type	Main	Auxiliary
Aspect	Perfect	Imperfect
	Imperative	
Gender	Masculine	Feminine
	Dual	
Number	Singular	Plural
	Dual	Respect
Tense	Simple past	Present Perfect
	Past Perfect	Present/Future
	Present Continuous	Past Continuous
Person	1st	2nd
	3rd	
Mood	Negative	Jussive
	Indicative	Interrogative

He analyzes the word categories as nouns, pronouns, verbs, adjectives, adverbs, adpositions (prepositions and conjunctions), Interjection (the problematic ones), punctuation (the non-phonological signs), numerals, and residuals. The tags are put by category, type, attribute and value. Then the complete list of Amharic tags is put in order with English tag names and their Amharic equivalents.

Ermias accomplishes well by analyzing morphosyntactic subcategories for words of Amharic. He attempts to represent them by giving two letters label. Still the tag is represented by capital letters. The design and development is better than the previous ones, for its detail and fineness. Anyone who wants to develop POS tagset for Ethiopian languages can follow the way Ermias did in developing Amharic POS tagset. Atwell (2008) asserts that it is good to adopt a POS tagset previously developed for that language, especially if the languages are sisterly ones.

> It may be attractive to simply adopt an existing tag set, but this still leaves the decision of which of several possible or rival tag sets to adopt, at least for English or other major European languages. If the language being studied is like a virgin, tagged for the very first time (cf. Madonna, 1984), then the researcher does not have the option to adopt an existing tag set; but they may still draw on parallels from other, more experienced language (Atwell, 2008:506).

2.4. Summary

POS tagset development is the basis for POS tagger construction. In POS tagset development, a developer has to consider the design, the characteristics, size and the purpose of the tagset being developed. The well-resourced languages developed POS tagsets for different purposes and audience. The under-resourced African, especially Ethiopian languages have developed POS tagsets in order to construct a POS tagger. Except that of Ermias' POS tagset developed for Amharic, the purpose of developing the tagsets is tagger construction.

Some of them are coarse-grained (flat/linear) tagsets, but few are especially that of Girma and Getachew's and Ermias' are designed hierarchically in a more fine-grained design. That helps to accommodate as many words as possible to assign to their word classes accurately. It is better to design the tags in a logical structure related to one another.

Since there is no POS tagset developed for its own case, most of the POS tagsets developed are prepared at a higher level. This does not help the POS tagger developer to disambiguate the inaccuracy of the labels that are used in the tagset; therefore it is basic to develop first a morphosyntactic tagset thoroughly before constructing a tagger. That is why we focus here to develop morphosyntactic tagset for Tigrinya.

Chapter three

Methodology

3.1. Introduction

The set of tags that are required in a language for inclusive morphosyntactic annotation of all word structures is the morphosyntactic tagset. Many languages in the world have developed POS tagsets for the purpose of producing a tagger, an application that annotates texts in a language.

In the earlier days and even sometimes now POS tagsets are at a higher coarse-grained level. This was just done by only taking the traditional parts of speech names to be tagged to the corpus under process. But nowadays the state of the art is developing morphosyntactic tagsets for their taggers, going beyond the morphological features to morphosyntactic ones by extending to a lower fine-grained levels.

In order to accomplish the purpose of developing a morphosyntactic tagset, one POS tagset developer should consider: the principles and criteria/guidelines for tagset development, the theoretical frame work or approach to be used when developing POS tagset, the needed grammatical input for POS tagset development and the type of the POS tagset.

In order to develop a morphosyntactic tagsets for Tigrinya, grammar books, theses and articles on Tigrinya morphology are consulted. Our works on POS tagsets will not be only the initial step towards standardized Tigrinya POS Tagging development, but also the foundation for the development of various areas of Natural Language Processing (NLP).

3.2. Theoretical framework

It is possible to develop a POS tagset following some type of linguistic theory. Atwell (2008:505) says, "Corpus linguists have tended to devise POS tagsets with very fine-grained grammatical distinctions; these POS tagsets reflect their expert interest in syntax and morphology, rather than specific predicted needs of end-users." The tagset developer should take into account that the tagset should cover aspects of the theory of language and the characteristics of that language (i.e., inflectional feature).

Beyond such a theory focused POS tagset development others develop POS tagset for some kinds of purposes/applications or target users. To do so they don't follow one linguistic theory, just they choose to remain theory neutral for the sake of their purposes.

We are developing a general purpose POS tagset for Tigrinya. Therefore we choose to be theory neutral in our POS tagset development. This helps any researcher to use this POS tagset for his/her own purpose, by decomposing or extending the existing POS tagset.

3.3. Design principles for Tigrinya tagset

In our development of POS tagset we consider two authorities who proposed criteria and guidelines for POS tagset development and corpus annotation; namely: Atwell (2008) and Leech (2005). Therefore, we take both of them and attempt to sift and select those which are very foundational to our POS tagset development. We see them here how they are applied in Tigrinya POS tagset development. In the section that follows we will outline the design features to which we adhere in devising a tagset for Tigrinya.

3.3.1. The tag name

a. Words of the same category should have the same name

Those words that have the same categorical and morphological or morphosyntactic classification should be grouped under the same name. Tag names should not vary here and there. For instance, words which are classified as noun shall have the tag name – N. This name shouldn't be given to another word class, otherwise it may create ambiguity when annotating or tagging a corpus.

b. Tag names should be remembered easily

When words are categorised into some kind of part of speech, the category name given to the word should be considered very well so that it may easily be remembered or memorized. There are many ways of labeling a tag name. Some label the first one letter or two letters or three letters as the name of the tag (the category Adjective can have the names A or Aj or JJ or Adj); others give a shorter known name and others give the first and the last consonant of the category (category Verb can be given the name V or VB or Vb); and others can give the name just

what they think is suitable (if JJ is considered suitable for the category adjective it is given that name for some reason). In our case, we consider the common shorter name of the category in Tigrinya grammar and check if that is remembered easily and holds to the most commonly known and suitable tag name. For instance the category name for verb is VB, it can be either V or VB having one letter or two, but for our own use, we will make it V, because the tag name should easily be distinguished and decomposable, and there is no problem to predict that a word tagged with V is simply a verb.

c. Every tag name should signify or indicate the place it holds

The tag should show clearly the categorical sense, so that we can predict its category/type easily. The tag names have been chosen to help linguists and NLP application developers to remember the lexical class of each word. For instance the tag name for the Tigrinya word ብዕራይ /bɨʕraj/ ‘ox’ is <NCSg> and it is read as ‘common-noun-singular’. Here, it is clearly known that where each category, type, attribute or value is starting. Therefore we will follow the same type of tag naming for Tigrinya.

3.3.2. Classification of words – by form and function

a. Definition by form

Every word in Tigrinya should have a name in order to be tagged. The main method of identifying these words is the white space in the written text. Hence, the words have different forms and they are inflectionally affected. As a result we see their inflectional forms and categorize them somewhere. The main defining approach of Tigrinya words is their form. Classifying of words by form (paradigmatically) involves morphological features of the word under process.

> The word ገዛ /gəza/ ‘house’ and ገዛታት /gəza + tat/ ‘houses’ are defined morphologically as <NCSg> - common noun-singular and <NCPl> - common noun-plural, respectively.

b. Definition by function

Classifying words by function (syntagmatically) involves morphosyntactic features of the word under process. Since Tigrinya is a morphologically rich language, the words have affixes that

show some function beyond their morphological feature, and this is defined by the function of the words morphosyntactically.

In the Tigrinya sentence መበራት ክዳና ቀዲዳ /məbrat kidana k'ədida/ 'Mebrat tore her cloth', the word ቀዲዳ /k'ədida/ 'she tore' is defined functionally, because /k'ədid + a/ tear <V> she <3FSg>, 'she tore' is defined syntactically showing that the subject 'she' and the action 'tear'.

Therefore it is imperative that we consider both form and function in our classification of words for Tigrinya as we assign their respective tag names.

3.3.3. What counts as a word?

As it is indicated in our review of literature and defined by Todd (1987:25-26) the most frequently implied meanings of 'word' are: (1) *Orthographic,* (2) *Morphological,* (3) *Lexical* and (4) *Semantic*. In our written text of Tigrinya, we consider that a word as any written token delimited by white space or punctuation. These tokens can be simplex words (words having one morpheme), complex words (words having more than one morpheme which can be decomposed into simpler morphemes), compound words, contractions, abbreviations, foreign words, numbers, dates, and symbols.

3.3.4. Multiword handling

Multiwords are compound words include idiomatic phrases and multiword expressions, found in Tigrinya texts. Some annotators give a tag name for each token of the phrase and yet others give one tag name for the compound. For instance the Tigrinya compound ፀዋር ደርሆ /s'əwar dərho/, literally means 'chicken carrier' and the idiomatic expression meaning 'briber'. The tag given to multiword in Tigrinya, if they are not concatenated by some means, the words will be taken together as one, but separately as it is shown in the Tigrinya example ፀዋር /s'əwar/ 'carrier' <Aj> and ደርሆ /dərho/ 'chicken' <N>.

3.3.5. Target users and/or application

The text to be annotated could be targeted to some type of users and/or application. This purpose determines the tag names and size of the tagset. In our morphosyntactic tagset development our target is the common user not the specialized ones and not even for one type of application, it is a

general purpose morphosyntactic tagset. We assume that it will be useful for everyone who wants to use for some purpose and application. Anyone can amend, decompose or extend the tagset for its own purpose and end.

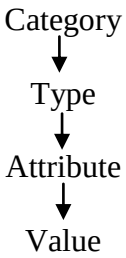
3.3.6. Adherence to standards

Since Tigrinya doesn't have any morphosyntactic tagset or tagger developed before now, there is no a standard set for the design of the morphosyntactic tagset. Therefore we shall adhere to the standards set for Amharic (because it is a sisterly language), and the EAGLES standards with some changes to comply with the Tigrinya morphosyntactic structure. In order to make the tagset as transparent and familiar as possible, the tag labels are derived from the typical morphological categories and word categories distinguished and named in the descriptive tradition of Tigrinya linguistics (Tsehaye, 1979; Amanuel, 1998; Daniel, 2000; and Tesfay, 2002).

3.3.7. Degree of coarseness

One of the fundamental issues that need to be addressed while designing a morphosyntactic tagset is its size. Generally, the assumption is – the smaller the tagset, the greater the accuracy in annotation. However, in saying so, we would not want to miss the essential categories plain in the language and at the same time also not necessarily increase the size of the tagset whenever economy can be maintained. Therefore, a middle ground has been adopted while designing a morphosyntactic tagset for Tigrinya. The coarseness of the tagset determines its size.

Since the morphosyntactic tagset being developed for Tigrinya is morphosyntactic it is by default fine-grained morphosyntactic tagset. These tags are not flat, but hierarchical, showing the category, type, attribute - value of each token feature by going down to the lower level to add some paradigmatic and/or syntagmatic information in a logical way. By doing so we maximize the linguistic enrichment describing morphosyntactic features (Hardie, 2003). The hierarchy goes down like the following.



3.4. POS names (Tags)

The names or labels given to every feature - value is given just by taking the first letter and if that is not enough we add one of the letters which better help a researcher to predict the category/attribute of the word and which does not create ambiguity with other tag names. If the tag has more than one letter, the first letter is in capital letter and the next one is in small letter. We chose such a label so that it helps any researcher locate every feature – value easily. The researcher can differentiate every capital letter as the start for the value of any category/attribute without giving any explanation.

E.g. For **N**oun - N, **A**d**j**ective - Aj, **V**erb - V, **A**d**v**erb - Av…

For **S**in**g**ular – Sg, **Pl**ural – Pl, **M**asculine – M, **F**eminine – F...

The following issues are considered when selecting the tag names. These will help us to judge whether our tags are appropriate or not.

The first issue we consider is the **conciseness** of the tag name given to every word. Everyone understands that short labels are better than long ones; and in fact the conciseness also saves space when annotating a corpus. Because of this we have tried to maintain our tags not to be more than two letters (e.g. a word categorized as noun is given the label <N> and a word categorized as preposition is given the tag <Pp>).

The second issue we ponder over is the **perspicuity** of the tag given to every word. Meaning that easily interpreted labels given to a word are better than those ambiguous ones (e.g. the tag <Pp> can be easily interpreted to mean preposition; it can't mean prepositional phrase or another thing, because everyone can understand easily the first letter in capital is the start of the name and if it has a second one as small letter that is its part and that goes with it to predict the meaning).

The last issue we are concerned with is the **analyzability** of the tags given to every word. The morphosyntactic tag given to every word should be able to be decomposed into different parts easily (e.g. the tag <VPfMSg> given to a Tigrinya verb should simply be able decomposed in to 'perfective, verb, masculine, singular').

For our purpose, we have chosen the way Khoja (2003) has put the tags for her ARBTAGS. She puts the tags as capital letter and small letter for the first and the second respectively, like Sg [singular], Pl [plural], Du [dual], etc.

3.5. What should be included in the tagset?

Hardie (2003) raises the most fundamental question about tagset design; that is, "What should the tags tell the user?" or, to put it another way, "What information should be included in a tagset?" According to Hardie (2003), "all tagsets account for major wordclass information".

In the literature, there seem to be an established consensus on what these categories should be. Hardie (2003) lists adjective, adposition, adverb, article, conjunction, interjection, noun, numeral, pronoun/determiner, and verb, with punctuation often counted as a major class, and two classes, unique and residual. The residual describes the classes and items which do not fit elsewhere in the design of analysis. There are also distinctions which Hardie refers to as "subclassifications of the major wordclasses". These distinctions include decompositions such as nouns as common or proper, verbs as main or auxiliary, adverbs as degree or general, prepositions or postpositions, and so on. There is also morphological information, noting such features as number, person, gender and case (Hardie, 2003:27 – 28).

The categories suggested by the EAGLES guidelines as obligatory to any POS tagset are that of "major word categories," proposed at a higher level are: noun, verb, adjective, pronoun/determiner, article, adverb, adposition, conjunction, numeral, interjection, unassigned/unique, residual, and punctuation.

The recommended and optional attributes are then branched from these major word categories to a lower level hierarchically, and do not necessarily match across word classes. For example, the attribute numbered (i) (i.e. the first recommended attribute after the obligatory attribute of Major Word Category) is Type (Common/Proper) for nouns but Person (First/Second/Third) for verbs and Degree (Positive/Comparative/Superlative) for adverbs.

The recommended attributes also include number, gender, case, finiteness, tense, voice, and other miscellaneous subcategorisation features. The optional part of the recommendations

consists of similar attributes of lesser applicability, and some additional values – mainly specific to one language or a small group of languages – for the recommended attributes (Leech 1996).

The known ELRC POS tagset developed by Girma and Mesfin (2006) has 11 basic classes of words: noun, pronoun, adjective, adverb, verb, preposition, conjunction, interjection, punctuation, numeral and the unclassified which is used for words which are difficult to place in any of the classes. Different tags for phrases of Amharic are alos added, and a total of 30 POS tags have been identified.

Sisay Fisseha Adafre (2005) developed a reduced POS tagset of Mesfin's containing 10 tags. The tags are: noun, verb, auxiliary verbs, numerals, adjective, adverb, adposition, interjection, residual, and punctuation.

Ermias analyzed the word categories in a higher level as noun, verb, adjective, adverb, adposition (prepositions and conjunctions), interjection, punctuation, numeral, and residual, all of them being 9 main categories. Beyond these he goes down to the types, attributes and values of the main categories, but he does not mention the total number of tags he assigned.

When we come to the Tigrinya POS tagset we have morphological information included when labeling complex words with affixes. In addition, complex words with morphosyntactic information will also be labeled accordingly. Moreover, special elements in texts are also labeled when available.

The morphological information given at a higher level for noun <N>, verb <V>, adjective <Aj>, adverb <Av>, pronoun <Pn>, prepositions <Pp>, conjunctions <Cj>, interjection <Ij>, and determiner <Dt>. Labels will be given to special elements like punctuation <Pu>, numeral <Nu>, and we may find some words like abbreviations <Ab>, foreign words <Fn>, symbols <Sy>, etc.

Concerning the foreign words, in Tigrinya texts we find many foreign words, especially the technological and scientific ones. We will give the tag <Fn> for foreign words found in Tigrinya texts. For instance we can find a date written in Ethiopian and Gregorian calendar like ሓምለ 8/ 2005 ዓ.ም. (July 15, 2013); therefore we tag the English word 'July' as <Fn>.

Our morphosyntactic tagset will also have different types for some categories. For instance the category noun will have the noun types common and proper. And in turn the common noun will have the attributes gender and number with their values, masculine and feminine for gender, and singular and plural for number.

> For instance let's take the Tigrinya noun ዕምበባ /ʔɨmbəba/ 'flower' and its plural ዕምበባታት /ʔɨmbəba + tat/ 'flowers'. The tag given for the plural noun will be the tags given for the two morphemes, which is the tag given for ዕምበባታት /ʔɨmbəba + tat/ i.e. <NCPl> (common-noun-plural).

It is also the same for some of the categories which have morphemes added to their bases, for instance verbs and adjectives.

Concerning morphosyntactic information our morphosyntactic tagsets will accommodate by labeling the morphosyntactic features of complex words having different affixes. This will go beyond the main categories to the recommended and sometimes to the optional level, according to the EAGLES guidelines.

For instance the Tigrinya verb ሰቢሩዎ /səbiruwwo/ 'he broke it/him' will have the tag <VPf3MSg> showing that /səbir-/ is the verb <V>, /-u-/ indicating both the third person masculine singular subject and the past action shown by <VPf3MSg> and the suffix /-wwo/ indicating the third person masculine singular object 'him' shown by <3MSg>, all in all as it is indicated above it will have the tag <VPf3MSg>.

3.6. What is not included in Tigrinya POS tagset

In our Tigrinya morphosyntactic tagset (TIGTAGS), we do not include the information listed below, because our purpose is developing a morphosyntactic tagset; nothing beyond that. Therefore, no etymological information will be included in the tagset. There will not be syntactic information like syntactic roles as subject and object and the kind of complements demanded by verbs. No semantic and discourse information will be included in the tagset as well.

3.7. Size of the POS tagset

The size of the tagset can be 10 - 20 tags or could also go up to two or three thousand tags. For instance, we can see that the following range of tagsets. The Penn Treebank has 36 tags, BNC C5 has 61 tags, Brown has 77 tags, the LOB has 132 tags, London-Lund Corpus has 197 tags, the TOSCA-ICE has 270 tags, Romanian has 614 tags, and Hungarian has 2100 tags and so on. Our Tigrinya morphosyntactic tagset has the necessary tags which can accommodate most of the words of Tigrinya.

Chapter four

Tigrinya Morphosyntactic Tagset (TIGTAGS) Design

4.1. Introduction

The main question in POS tagset design is "what information should be included in the tags?" Primarily, tags should include the major word classes (categories) of the language. Since our design is fine-grained, hierarchical and decomposable, it should also show the type of the word by subclassification. Below the type when ever necessary it shows the attribute and value of the word. All these are represented with a letter or letters, showing information about the morphological and morphosyntactic feature of a word.

We have consulted the classifications of word class categories by Tsehaye Teferra (1979), Amanuel Sahle (1998), and Tesfay Tewolde (2002) to categorize and give names to words of Tigrinya. Concerning sub classifications we have also consulted Daniel Teklu's (2008) classifications. In the design process we have followed the EAGLES guidelines to some extent, and, Girma's and Ermias' way of design to a greater extent, because Tigrinya and Amharic are sisterly languages. But Tigrinya may have its own peculiarities from the above-mentioned designs and we have taken our own measure to put those into place.

4.2. Design of the Tagset

In the following sections, we will discuss the design of morphosyntactic tagset for Tigrinya. Here, Tigrinya word categories and attributes with their values are presented. The definitions for the categories are taken from Amanuel Sahle (1998), Daniel Teklu (2008) and Jurafsky (2008), unless and otherwise stated.

4.2.1. Nouns <N>

Nouns are words which signify the physical and abstract - people, place, event, substance or thing. The category noun is divided into common noun and proper noun. <**N**> is assigned to denote nouns.

While common nouns <**NC**> refer to person, place or thing, poper nouns <**NPr**> refer to a specific name of a person, place, and institution. We see below the category noun with its types and the tags given to them.

Sentence in Tigrinya:	ሽሻይ	ገዛ	ሰሪሑ
IPA transcription:	/ʃɨʃaj	gəza	səriħ-u/
Analyzed Word constituents:	Shishay	house	built-he
Assigned tag:	<NPo>	<NC>	
Gloss:	‘Shishay built a house’		

In the above example we see that ‘Shishay’ is a Proper Noun <NPo> and house is a Common Noun <NC>. The values given for every attribute is always in combination with the main category as <NC> (common noun) and <NPo> (proper noun).

Common nouns are divided in to two: countable <NCCt> and mass nouns <NCMs>. The countable and mass nouns can also be concrete <NCr> and abstract <NAs> nouns. These features are illustrated in the following.

	Countable	**Mass**	**Concrete**	**Abstract**
Tigrinya	ኦም	ፀባ	ማኪና	ፍቕሪ
IPA transcription	/ʔom/	/s’əba/	/makkina/	/fɨχ’ri/
Tag	<NCCtSg>	<NCMs>	<NCCrSg>	<NCAs>
Gloss	tree	milk	car	love

Tigrinya countable and concrete common nouns have singular <Sg> and plural <Pl> number, and mass nouns are always plural, but, abstract nouns do not have number. Nouns have gender forms with values feminine <F> and masculine <M>. The gender of most inanimate nouns is not predictable from their form. But the others which are not naturally indicated for gender should

have a determiner which shows the gender of the noun, like እቲ ኦም /ʔɨti ʔom/ ‘the <DtMSg> tree <NCCtSg>’, እታ ማኪና /ʔɨta makkina/ ‘the <DtFSg> car <NCCtSg>’.

Definiteness for nouns is indicated by determiners; otherwise we can’t tell whether a noun is definite or indefinite. Tigrinya nouns inflect for possessiveness (Di Lello, 1995: 13, 15).

The Tigrinya common noun ዘመድ /zəməd/ ‘relative’ can be inflected to be possessive by adding suffixes and its inflected form can be replaced by isolated possessive pronoun:

Tigrinya noun	**Tigrinya possessive noun**	**Gloss**	**Isolated possessive Pronoun**	**Gloss**
ዘመድ	ዘመደይ /zəməd + əy/	my relative <NPsSg>	ናተይ ዘመድ /natəy zəməd/	my relative <PnPsSg> <N>
	ዘመድና /zəməd + na/	our relative <NPsPl>	ናትና ዘመድ /natna zəməd/	our relative <PnPsPl> <N>
	ዘመድካ /zəməd + ka/	your relative <NPsMSg>	ናትካ ዘመድ /natka zəməd/	your relative <PnPsMSg> <N>
	ዘመድኪ /zəməd + ki/	your relative <NPsFSg>	ናትኪ ዘመድ /natki zəməd/	your relative <PnPsFSg> <N>
	ዘመድኩም /zəməd + kum/	your relative <NPsMPl>	ናታትኩም ዘመድ /natatkum zəməd/	your relative <PnPsMPl> <N>
		your relative <NPsMSgH[10]>	ናትኩም ዘመድ /natkum zəməd/	your relative <PnPsMSgH> <N>
	ዘመድክን /zəməd + kɨn/	your relative <NPsFPl>	ናታትክን ዘመድ /natatkɨn zəməd/	your relative <PnPsFPl> <N>
		your relative (honourific) <NPsFSgH>	ናትክን ዘመድ /natkɨn zəməd/	your relative (honourific) <PnPsFSgH> <N>
	ዘመዱ /zəməd + u/	his relative <NPsMSg>	ናቱ ዘመድ /natu zəməd/	his relative <PnPsMSg> <N>
	ዘመዳ /zəməd + a/	her relative <NPsFSg>	ናታ ዘመድ /nata zəməd/	her relative <PnPsFSg> <N>
	ዘመዶም /zəməd + om/	their relative <NPsMPl>	ናታቶም ዘመድ /natatom zəməd/	their relative <PnPsMPl> <N>
		his relative (honourific) <NPsMSgH>	ናቶም ዘመድ /natom zəməd/	his relative (honourific) <PnPsMSgH> <N>
	ዘመደን /zəməd + ən/	their relative <NPsFPl>	ናታተን ዘመድ /natatən zəməd/	their relative <PnPsFPl>
		her relative (honourific) <NPsFSgH>	ናተን ዘመድ /natən zəməd/	her relative (honourific) <PnPsFSgH> <N>

Table 3: Tigrinya noun inflected for possession and possessive pronoun (Di Lello 1995:15)

[10] <H> indicates respect and honor.

According to Di Lello (1995), the suffix added to the noun, which shows possession (e.g., /-əy/ 'my') is the same as the isolated possessive pronoun (e.g., /natəy/ 'mine'); and it is the same for the other possessive noun suffixes. Tesfay (2002:121) calls these suffixes as possessive endings which indicate pronouns.

Verbal nouns, which are infinitives or gerundives, are words which look like verbs, but they are not verbs because they function as a noun. In Tigrinya verbal nouns are included here as derived nouns from verbs. All Tigrinya verbal nouns start with "ምᎅ-" /mɨ-/ meaning "to-" or "-ing" Tsehaye, 1979:173). For instance the Tigrinya word

Tigrinya	**ምᎅᎁᎋᎀ**
Transcription	**/mɨ-blaʕ/**
Assigned tag	**<Nv> (verbal noun)**
Gloss	**'eating' is derived from the verb ᎁᎇᎀ /bəlʔə/ 'he ate'.**

If we want to make the noun definite, we have to add a determiner before it. For instance 'the house' is translated in to Tigrinya as ᎎᎢ ᎃᎋ /ʔɨt-i gəza/. Here ᎎᎢ /ʔɨt-i/ 'the' [Dt3MSg] is the definite marker pointing to the house. We will see more about determiners later.

Category	**Noun <N>**	
Type	Common <NC>	Countable <NCCt>
		Mass <NCMs>
		Concrete <NCCr>
		Abstract <NCAs>
	Proper <NPo>	
	Verbal noun <Nv>	
	Possessive <NPs>	
Gender	Masculine <NM>	
	Feminine <NF>	
Number	Singular <NSg>	
	Plural <NPl>	
	Honorific/respect <NH>	

Table 4: Nominal Features of Tigrinya

4.2.2. Verbs <V>

Verbs signify actions or processes of events (Daniel 2008:146; Jurafsky 2008:4). They contain inflections for aspect, mood, and tense. Tigrinya verbs can be inflected to be perfective, imperfective, gerundive, imperative and jussive. They also have additional affixes that show person, number, and gender for agreement with the subject and/or the object (Tesfay 2002:110). The agreement on the sentence depends not on the verb itself, but on some other associated component - on the characteristics of the subject.

Tigrinya	ፀጋይ	ንኣብርሀት	ፀዊዑዋ		
Transcription	/s'əggaj	nɨ + ʔabrɨhət	s'əwwiʃ + u	+	wwa/
Constituents	Tsegay	to Hagos	called	he	her
Assigned tag	<NPoS>	<NPoFO>	<VPt>	<3MSgS[11]>	<3FSgO[12]>
Gloss	'Tsegay called Abrehet'				

In the above example we see that the verb ፀዊዑዋ /s'əwwiʃuwwa/ 'called' agrees with the subject ፀጋይ /s'əggay/ 'Tsegay' which is <NPoS> and with the object ኣብርሀት/ʔabrɨhət/ Abrehet which is <NPoO>. Regarding the preposition attached to the object we will see it under the Phrases section.

Verbs in Tigrinya can be **transitive** and **intransitive**. Tigrinya **transitive** verbs are words expressing an action carried from the subject to the object and require direct object to complete meaning.

Tigrinya	ሽሻይ	ንሰሰን	ወቒዑዋ
Transcription	/ʃɨʃay	nɨ + səsən	wəχ'iʕ + u + wwa/
Word constituents	Shishay	to Sesen	hit
Assigned tag	<NPo3MSg>	<NPo3Sg>	<VPtTr 3MSgS 3FSgO>
Gloss:	'Shishay hit Sesen'		

[11] The tag S shows the subject agreement.
[12] The tag O shows the object agreement.

Intransitive verbs on the other hand do not require a direct object.

Tigrinya	ሽሻይ	ደቂሱ
Transcription	/ʃɨʃaj	dək'k'is + u/
Constituents	Shishay	slept
Assigned tag	<NPo3Sg>	<VPfIt 3MSgS>
Gloss	'Shishay slept'	

Since the action didn't pass to another object, the verb does not need any object. Therefore the verb ደቂሱ 'slept' is intransitive.

Tense in Tigrinya is past, present and future locating a particular situation in a particular time. Past and future tenses situate an event in point of time before and after time of utterance respectively. Present tense is an event at time of utterance.

	Past	Present	Future
Tigrinya	በልዐ	ይበልዕ (ኣሎ)	ክበልዕ (እዩ)
Transcription	/bəlʕə/	/jɨ-bəllɨʕ/ (/ʔallo/)	/kɨ-bəllɨʕ/ (/ʔɨjju/)
Assigned tag	<VPt3MSg>	<VPr3MSg>	<VFu3MSg>
Gloss	'He ate'	'he eats'	'he will eat'

Tigrinya verbs also have **active** and **passive voice**. Those perfective and gerund verbs are active and they can be changed into passive by adding the prefix ተ- /tə-/. The other types of verbs are free of such inflections (Girmay, 1991:39).

Active verb Tigrinya	Gloss	Passive verb Tigrinya	Gloss
በልዐ /bəlʕə/	**'he ate'**	**ተበልዐ /təbəlʕə/**	**'it is eaten'**
ወሰደ /wəsədə/	**'he took'**	**ተወሰደ /təwəsədə/**	**'it is taken'**
<VPfAt3MSg>		<VPfPv3MSg>	

To express **negation** in Tigrinya most of the time we use the affix ኣይ....ን /ʔay---n/. It is used to negate verbal and nominal elements. The perfect and imperfect forms of the verb are negated by the affix ኣይ....ን /ʔaj---n/ (Tsehaye 1979). This is illustrated in the following examples:

Positive verb Tigrinya	Gloss	Negative verb Tigrinya	Gloss
ቆረፀ /kʾorəsʾə/	ʿhe cutʾ	**ኣይቆረፀን /**ʔaj-kʾorəsʾə-n/	ʿhe didnʾt cutʾ
<VPf3MSg>		<VPf3MSgNg>	
ተሰበረ /təsəbərə/	ʿit is brokenʾ	**ኣይተሰበረን /**ʔaj-təsəbərə-n/	ʿit is not brokenʾ
<VPv3MSg>		<VPv3MSgNg>	
ኣብልዐ /ʔablɨʕə/	ʿhe fedʾ	**ኣይኣብልዐን /**ʔaj-ʔabəlʕə-n/	ʿhe didnʾt feedʾ
<VCa3MSg>		<VCa3MSgNg>	

Tigrinya verbs in their mood can be conjugated for imperative/jussive, interrogative, negative, indicative, and subjunctive. These mood forms can also be in their present or past tenses. For instance we will see the following.

	Imperative	**Interrogative**	**Indicative**	**Subjunctive**
Tigrinya	ብላዕ	በሊዖዶ	ይበልዕ ኣሎ	እንተዝበልዕ
Transcription	/bɨlaʕ/	/bəliʕu-do/	/yɨbəllɨʕ ʔallo/	/ʔɨntəzzɨ-bəllɨʕ/
Tag	<VIm2MSg>	<VIg3MSg>	<VIc3MSg>	<VSb3MSg>
Gloss	‘you eat’	‘did he eat?’	‘he is eating’	‘if he eats’

Nominals can also have negative features.

Tigrinya	ክፍቲ	ኣይ-ክፍት-ን
Transcription	/kəfti/	/ʔaj-kəfti-n/
Tag	<NC>	<NCNg>
Gloss	‘cattle’	‘not cattle’

The following summarizes what we have said up to now about verbal features of Tigrinya.

Category	Verb <V>
Type	Main <VMn>
	Auxiliary <VAx>
Aspect	Perfect <VPf>
	Imperfect <VIp>
	Gerundive <VGe>
	Infinitive <VIf>
Gender	Masculine <M>
	Feminine <F>
Number	Singular <Sg>
	Plural <Pl>
	Honorific/respect <H>
Transitiveness	Transitive <Tr>
	Intransitive <It>
Tense	Past <VPt>
	Present <VPr>
	Future <VFu>
Person	First <1>
	Second <2>
	Third <3>
Mood	Imperative/Jussive <Im>
	Interrogative <Ig>
	Negative <Ng>
	Indicative <Ic>
	Subjunctive <Sb>
Voice	Active <At>
	Passive <Pv>

Table 5: **Verbal features of Tigrinya**

4.2.3. Pronouns <Pn>

Pronouns are words that function as nouns. Some scholars put them under noun, because they function like a noun. There is no consensual agreement concerning where to put pronouns. Since

they are closed and the traditional grammars of Tigrinya put them independently; accordingly, we have chosen to treat them independently. Tigrinya pronouns are few in number and could be isolated or suffixed to a noun or a verb. When they are suffixed to a noun, it becomes possessive noun. They can be distinguished as personal, reflexive, reciprocal, demonstrative/definite, indefinite and interrogative (Tesfay 2002).

Pronoun	**Tigrinya**	**Transcription**	**Tag**	**Gloss**
Personal	ንሱ	/nɨssu/	<PnPe3MSg>	‘he’
Reflexive	ባዕሉ	/baʕlu/	<PnRf3MSg>	‘himself’
Reciprocal	ንሕድሕዱ	/nɨħɨdħɨdu/	<PnRc3MPl>	‘each other’
Demonstrative	እቲ	/ʔɨti/	<PnDm3MSg>	‘that’
Indefinite	ገለ	/gələ/	<PnId>	‘something’
Interrogative	መን	/mən/	<PnIg>	‘who’
Possessive	ናቱ	/natu/	<PnPs3MSg>	‘his’

We have polite/honorific forms in the pronominal system of Tigrinya; these are, **ንስኹም** /nɨssɨχum/ ‘you’ or **ንሶም** /nɨssom/ ‘you’ for 2MSgH, **ንስኽን** /nɨssɨχɨn/ ‘you’ or **ንሰን** /nɨssən/ ‘you’ for 2FSgH, **ንሶም** /nɨssom/ ‘he’ for 3MSgH, and **ንሰን** /nɨssən/ ‘she’ for 3FSgH [13]. The polite/honorific forms also show plurality in some contexts. These different types of pronouns are tagged as follows:

Tigrinya		Tag	Gloss
ንስኹም	/nɨssɨχum/	<PnPeH2MSg> or <PnPe2MPl>	‘you’ or ‘you’ plural
ንሶም	/nɨssom/	<PnPeH2MSg> or <PnPe2MPl>	‘you’ or ‘you’ plural
ንስኽን	/nɨssɨχɨn/	<PnPeH2FSg> or <Pne2FPl>	‘you’ or ‘you’ plural
ንሰን	/nɨssən/	<PnPeH2FSg> or <PnPe2FPl>	‘you’ or ‘you’ plural
ንሶም	/nɨssom/	<PnPeH3MSg> or <PnPe3MPl>	‘he’ or ‘they’ plural
ንሰን	/nɨssən/	<PnPeH3FSg> or <PnPe2FPl>	‘she’ or ‘they’ plural

Tigrinya pronouns are also inflected for nominative, accusative and genitive case. Some scholars may consider this possessiveness with the possessive marker noun /-ə**y**/ (Tesfay 2002). The accusative form of pronoun indicates object and the nominative to a subject in a sentence.

[13] These are almost the same with the Amharic **እርስዎ** /ʔɨrsɨwo/ and **እርሳቸው** /ʔɨrsačč**ə**w/.

Pronoun	Tigrinya		Tag	Gloss
Nominative	ኣነ	/ʔanə/	<PnNm1NtSg>	‘I’
Accusative	ንዓይ	/nɨʕay/	<PnAc1NtSg>	‘to me’
Genitive	ናተይ	/natəy/	<PnGn1NtSg>	‘mine’

Subject and/or object pronouns are also indicated in the verb with out being overt or being droped from a sentence. Let us see the following sentence.

Tigrinya	ንመድህን	ደብዳበ	ፅሒፈላ			
	/nɨ + mədhɨn	dəbdabbə	s’ɨħif +	ə +	ll +	a/
	To Medhin	(a) letter	wrote	I		
Tag	<Pp> <NPoSg>	<NCSg>	<VPf>	<Pn1SgS>	<Pp	3FSgO>
Gloss	‘I wrote a letter to Medhin’					

In the above sentence we do not see the subject, but it is indicated in the verb for agreement with it. We indicate this agreement by putting the subject tag <S> with the verb. We can also sentences which do not show the subject and the object overtly.

Tigrinya	ሎሚ	ደብዳበ	ፅሒፉሎም			
	/lomi	dəbdabbə	s’ɨħif +	u +	ll +	om/
	today	(a) letter	wrote	he		them
Tag	<AvTm>	<NCSg>	<VPf>	<Pn3SgS>	<Pn3MPlO>	
Gloss	‘Today he wrote a letter to them’					

Here again we see that the subject and the object are dropped from the sentence, but indicated in the verb for their agreement. Therefore we show the subject and the object by putting the tags <S> and <O> into the verb tag respectively.

Category	Pronoun <Pn>
Type	Personal <PnPe>
	Reflexive <PnRf>
	Reciprocal <PnRc>
	Possessive <PnPs>
	Demonstrative <PnDm>
	Indefinite <PnId>
	Interrogative <PnIg>
Case	Nominative <PnNm>
	Accusative <PnAc>
	Genitive <PnGn>
Gender	Masculine <M>
	Feminine <F>
Number	Singular <Sg>
	Plural <Pl>
	Honorific <H>
Person	First <1>
	Second <2>
	Third <3>

Table 6: **Pronominal features of Tigrinya**

4.2.4. Adjectives <Aj>

Adjectives modify nouns, specifying attributes and qualities of things or people. They can be identified by their function and distribution in a sentence. They can function as modifiers of nouns by following them. They can describe the dimensions, color, age, value, and position, the physical and human properties of nouns they follow.

Adjectives may attach plural forms or they may add affixes to show the agreement of the gender and number of the nouns they modify. The Tigrinya word ቀይሕ /kʼəyyɨħ/ and ቀያሕ /kʼəyyaħ/ both meaning ‘red’, but different for masculine and feminine. They modify the nouns which follow them accordingly. We can say ቀይሕ ወዲ /kʼɐyyɨħ wəddi/ ‘red boy’ ቀያሕ ጓል /kʼəyyaħ gʷal/ ‘red girl’.

And they can also be pluralized, ሓፀርቲ ኣንስቲ /ħasʾərti ʔanɨsti/ ‘short women’ ቆፀልቲ መፃሕፍቲ /kʾo sʾsʾəlti məsʾaħɨfti/ ‘green books’.

	Tigrinya		Tag	Gloss
Singular	ቀይሕ ወዲ	/kʾeyyɨħ wəddi/	<AjMSg NCMSg>	red boy
Plural	ሓፀርቲ ኣንስቲ	/ħasʾərti ʔanɨsti/	<AjPl NCFPl>	short women
Feminine	ፅብቕቲ ጓል	/sʾɨbbɨχʾti g^{w}al/	<AjFSg NCFSg>	beautiful girl
Masculine	ገፊሕ ክፍሊ	/gəffiħ kɨfli/	<AjMSg NCSg>	wide room

There are also adjectives which do not show gender dinstiction, for instance, ሰነፍ /sənəf/ ‘lazy’. The gender of such adjectives is identified by the preceding determiner or the following verb (እቲ ሰነፍ ቆልዓ /ʔɨti sənəf k'olʕa/ ‘the lazy child’).

The following table shows the adjective features of Tigrinya.

Category	**Adjective <Aj>**
Gender	Masculine <M>
	Feminine <F>
Number	Singular <Sg>
	Plural <Pl>
Person	First <1>
	Second <2>
	Third <3>

Table 7: **Adjective features of Tigrinya**

The definiteness of adjectives is not shown by affixes attached to them, but by adding independent conformable determiners before them.

4.2.5. Adverbs <Av>

Adverbs function to specify the mode of action of the verb. They refer to the duration, place, direction, or situation of actions which have been completed, and which are in progress, or which will occur in the future. They are modifiers of verbs, adjectives, sentences or clauses, and other adverbs. There are different types of adverbs, like:

Degree adverb (modifies adjective)	ኣዝዩ ነዊሕ	/ʔazzɨyu nəwwiħ/	‘very long’
Temporal adverb (signifies time)	ትማሊ	/tɨmali/	‘yesterday’
Manner adverb (modifies verb)	ብቕልጡፍ	/bɨχ’ɨlt’uf/	‘quickly’
Directional adverb (indicating location/direction)	ንየማን	/nɨjəman/	‘to the right’

Category	**Adverb <Av>**	
	Degree <AvDg>	
	Temporal <AvTm>	
	Manner <AvMr>	
	Directional <AvDr>	
	Locative <AvLc>	

Table 8: **Adverb features of Tigrinya**

4.2.6. Prepositions <Pp>

Traditional grammarians classify prepositions and conjunctions separately. But some modern Tigrinya grammar books classify them as adpositions. For our own convenience and since there is no consensus, we chose to follow the traditional classification and put them separately. Prepositions express the relationship between a person, thing, or event. They come before nouns or adjectives as separate word or concatenated to the following morpheme. We treat their tags according to their nature; if they come differently before the noun we give them a tag for themselves alone as preposition <Pp>. Otherwise, they will be taken as prepositional phrase.

Tigrinya		Tag	Gloss
ትሕቲ ዓራት	/tɨħti ʕarat/	<Pp> <NCSg>	‘below bed’
ካብ ዕዳጋ	/kab ʕɨdaga/	<Pp> <NCSg>	‘from market’
ብበትሪ	/bɨbətri/	<PNPp>	‘with a stick’
ናብ ቤት ትምህርቲ	/nab bet tɨmhɨrti/	<Pp> <NCSg> <NCAs>	‘to school’
ክሳዕ ሎሚ	/kɨsaʕ lomi/	<Pp> <AvTm>	‘until today’
ናይ ራሄል	/nay rahel/	<Pp> <NPo3Sg>	‘Rahel’s’

Category	**Preposition <Pp>**

Table 9: Preposition feature of Tigrinya

4.2.7. Conjunctions <Cj>

Conjunctions in Tigrinya connect words, phrases or sentences. They can be classified into coordinating and subordinating conjunctions. Coordinating conjunctions adjoin elements of equal rank.

Tigrinya	ሰብኣይን ሰበይትን			
	/səbʔay +	ɨn	səbəyt +	ɨn/
Tag	<NCSg>		<NCSg>	
Gloss	‘a man and a woman’			

The -ን /-ɨn/ morpheme that attaches to each conjunction connect the two words ‘man’ and ‘woman.’ When giving a tag for coordinating pronouns we give them a general tag for the pronoun, because we do not detach the coordinating pronoun bound morpheme from the free morpheme. And that is not our focus for this research.

Subordinating conjunctions, on the other hand, conjoin elements by making one of them dependent to the other sentences.

ፀጋይ ምስመፀ በርሀ ከይዱ

/s’əggaj mɨsməs’ə bərhə kəjdu/

‘Berhe went when Tsegay came’

Here we see that there are two sentences ፀጋይ መፂኡ /s’əggay məs’iʔu/ ‘Tsegay came’ and በርሀ ከይዱ /bərhə kəjdu/ ‘Berhe went’. These two sentences are joined by adding /-mɨs/ inserted before the first sentence. Here the morpheme ምስ /mɨs-/ ‘when’, the subordinate conjunction, is conjoining the two sentences. Let’s see another example; in the Tigrinya sentence, ናብ እትኾዱ ክኾይድ እየ /nab ʔɨttɨχədo kɨχəjjɨd ʔɨjjə/ ‘I will go wherever you go’, we see the prepositional morpheme ናብ /nab/ ‘wherever’ working as a subordinate conjunction, because it joins two dependent sentences.

Tigrinya	ናብ	እትኸዶ	and	ክኸይድ እየ
Transcription	/nab	ʔɨttɨχədo/		/kɨχəjjɨd ʔɨjjə/
Tag	<CjSu>	<V2MSg>		
Gloss	‘wherever	you go‘		I will go

Category	**Conjunction <Cj>**
Type	Coordinating <CjCo>
	Subordinating <CjSu>

Table 10: Conjunction features of Tigrinya

4.2.8. Determiners <Dt>

Determiners indicate some information about the definiteness of a thing or a person. Whether the thing or person referred to is familiar to the hearer and the speaker is not familiar, it is indicated by determiners. We can see this from the following instances:

Tigrinya Definite		Tag	Gloss
እቲ	/ʔɨt-i/	<DtDf3MSg>	‘that’
እታ	/ʔɨt-a/	<DtDf3FSg>	‘that’
እዚኦቶም	/ʔɨziʔat-om/	<DtDf3MPl>	‘these’
እዚኣተን	/ʔɨziʔat-ən/	<DtDf3FPl>	‘these’

Tigrinya Indefinite		Tag	Gloss
ሓደ	/ħad-ə/	<DtIf3MSg>	‘a’
ሓንቲ	/ħan-ti/	<DtIf3FSg>	‘a’
ዝኾነ	/zɨχon-ə/	<DtIf3MSg>	‘somebody’
ዝኾና	/zɨχon-a/	<DtIf3FPl>	‘somebody’

In Tigrinya, /ʔɨt-/ is the stem common to both the nominative and the accusative forms of the demonstrative. The subject is indicated by the nominative determiner. For instance, in the sentence እቲ መምህር ነቲ ተምሃራይ ሓጊዝዎ /ʔɨti məmhɨr nəti təmharay ħaggizɨwwo/ ‘the teacher helped the student’, the subject መምህር /məmhɨr/ ‘teacher’ is indicated as definite by the

preceding determiner እቲ /ʔɨti/ ‘the’. The object ተምሃራይ /təmharay/ ‘student’ is also indicated by the preceding determiner ነቲ /nəti/ ‘the’. Determiners agree with their head in person, gender and number.

Tigrinya Nominative		Tag	Gloss
እቲ	/ʔɨt-i/	<DtDfNm3MSgS>	‘that’
እቶም	/ʔɨt-om/	<DtDfNm3MPlS>	‘those’ (masculine)
እተን	/ʔɨt-ən/	<DtDfNm3FPlS>	‘those’ (feminine)
Tigrinya Accusative		**Tag**	**Gloss**
ነቲ	/nəti/ (/nɨ-ʔɨt-i/)	<DtDfAc3MSgO>	‘to that’
ነቶም	/nətom/ (/nɨ-ʔɨt-om/)	<DtDfAc3MPlO>	‘to those’
ነተን	/nətən/ (/nɨ-ʔɨt-ən/)	<DtDfAc3FPlO>	‘to those’

Determiners which are conjugated for genitive case, usually indicating possessiveness, are the concatenation of the preposition ናይ /nay-/ ‘of’ and the nominative form of the determiner.

Tigrinya Genitive		Tag	Gloss
ናይቲ	/nayti/ (/nay-ʔɨt-i/)	<DtDfGn3MSg>	‘of that’
ናይቶም	/naytom/ (/nay-ʔɨt-om/)	<DtDfGn3MPl>	‘of those’
ናይተን	/naytən/ (/nay-ʔɨt-ən/)	<DtDfGn3FPl>	‘of those’

Category	**Determiner <Dt>**
Gender	Masculine <M>
	Feminine <F>
Number	Singular <Sg>
	Plural <Pl>
Case	Nominative <Nm>
	Accusative <Ac>
	Genitive <Gn>

Table 11: Determiner features of Tigrinya

4.2.9. Interjections <Ij>

David Crystall (2008: 249) explains the category interjection as;

> A term used in the traditional classification of parts of speech, referring to a class of words which are unproductive, do not enter into syntactic relationships with other classes, and whose function is purely emotive, e.g. Yuk!, Strewth!, Blast!, Tut tut! There is an unclear boundary between these items and other types of exclamation, where some referential meaning may be involved, and where there may be more than one word, e.g. Excellent!, Lucky devil!, Cheers!, Well well!

Since we have seen the problem encountered in Amharic is the same with interjections in Tigrinya, we have quoted the idea of Ermias as it is.

> Interjections are one of the most problematic areas of tagging. To start with, their definition is unclear. They are generally described as those words which are not morphologically productive. But this does not seem to work with some emotives like "hm" in English (which may duplicate the last sound like "hmm", "hmmmm", etc) or "ill" (for ululation) in Amharic (which may duplicate the last sound like "illll", "illllllll", etc).
> Differentiating between "interjections" and "exclamations" is also a problem. But most importantly they don't enter into "syntactic relation with other words" surrounding them. Baye argues that interjections are not word classes; they are rather parts of discourse. This won't however help us in POS tagging and we have to look for a mechanism by which we can easily tag such words.
> Kawata suggests a way out by saying that
> ... if it can be described as other than an interjection, describe it so. In the absence of an adequate tag, such a fragment would have to be put under residual for the future development, or be forced to be classified as an interjection.
> Taking his example, "Lucky devil!" may be tagged as an adjective followed by a noun. "Look!" may also be tagged as a verb.
> Thus, if we can tag a word appearing in an emotive context using any of the major tags, it is preferable to do so. If and only if this fails, we resort to tagging it as an interjection. If in doubt, always assign a residual tag (Ermias, forthcoming).

According to Baye (2008: 94-96) Interjections express fright, exclamation, anger, etc. When someone expresses his/her emotions, he/she uses emotive words. Therefore Baye put these words as emotives (emotive words), expressing emotions of a person.

Interjections (emotives) do not have grammatical function; they do not have syntactic relation with other words, but their function is to emote or belittle an idea. Since the interjections we find in Tigrinya texts express different types of emotions are emotives, we may tag them as <Ij>.

Type	Tigrinya	IPA	Gloss
Brag	ወዲኣይተ!	/wədiʔajtə/	Bragging
Pain	እህ!	/ʔɨh/	expressing pain
Amazement	ዋእ!	/waʔ/	expressing astonishment
Fright	ኡይ!	/ʔuj/	a shout when frightened or in need of help

Category	Interjection <Ij>

Table 12: Interjection features of Tigrinya

4.2.10. Punctuation <Pu>

Punctuations are non-phonological values in texts. They can be placed at the beginning or at the middle or at the end of a word or a sentence. As it is indicated in our review of literature Tigrinya has punctuations which are used from the early stages of the literature. Some of the punctuation marks have changed into modern ones like the question mark is '፧' changed to '?'.

E.g. we can find the following punctuation marks in Tigrinya texts.

- ፡ Is used to write the amount of money to separate coins, abbreviations
- ። is used for space among words [not in modern use], it is also used to write time in numbers, separating hour from minute, and minute from second
- ። is used as a full stop,
- ፤ is used as a semi colon,
- ፥/፣ is used as a comma,
- ፦ is used as a preface colon,
- ፧ is used as a question mark [not in modern use],
- ? question mark [modern use],

፨ is used as a paragraph separator [not in modern use], and

, is used to separate numbers every three digit.

In addition to these there are also some modern punctuation marks used in Tigrinya texts like the following.

! Exclamation mark suggests excitement or emphasis in a sentence.

- We use a hyphen when adding a prefix to some words.

-- or —used when making a brief interruption within a statement and an additional comment

" encloses a direct quotation

' combines two words to make a contraction, and indicates a quotation within a quotation

( ) to clarify, to place an afterthought, or to add a personal comment

[ ] to signify an editor's note in a regular piece of writing

{ } are most widely used in denoting a numeric set in mathematics

/ to separate ‘and’ and ‘or’

… to indicate an idea continues.

Though there are many ways of giving tags to punctuation marks, some give by the place they hold (at the beginning or in the middle or at the end), others by their individual name, and yet others give the comprehensive name generally as punctuation. Since the punctuation is too much to give such detail labels for punctuation marks, we have chosen to give the punctuation marks’ comprehensive name at the upper level of the category, punctuation <Pu>.

Category	**Punctuation <Pu>**

Table 13: punctuation feature of Tigrinya

4.2.11. Numerals <Nu>

Tigrinya numerals can be cardinal or ordinal. They may function as nouns, pronouns, determiners or adjectives. Numerals can be conjugated for person, number and gender. They do not indicate definite quantifiers. They show the definiteness of a thing or person only when a determiner is added before them.

For instance,

Tigrinya	እተን	ክልተ	ኣንስቲ
	/ʔɨt-ən	kɨltə	ʔanɨsti/
Tag	<PnDm3FPl>	<NuCa>	<NC3FPl>
Gloss	‘the	two	women’

In the above example we see that the numeral ክልተ /kɨltə/ ‘two’ becomes definite by the demonstrative እተን /ʔɨtən/ ‘the’ added before it, which indicates the plural noun ኣንስቲ /ʔanɨsti/ ‘women’.

The numeral for “one”, agrees for number and gender, but the others do not. The feminine form is used only with feminine nominals. The masculine form is used with masculine nominals and independently in counting. The numeral hundred, thousand, million and billions has a plural form (Tsehaye, 1979:212,214).

Tigrinya		**Tag**	**Gloss**
ሓደ	/ħadə/	<NuCaMSg>	‘one’
ሓንቲ (ሓንቲት)	/ħanti/ (/ħantit/)	<NuCaFSg>	‘one’
ሚእቲ	/miʔti/	<NuCaSg>	‘hundred’
ኣማኢት	/ʔamaʔit/	<NuCaPl>	‘hundreds’
ሽሕ	/ʃɨħ/	<NuCaSg>	‘thousand’
ኣሽሓት	/ʔaʃħat/	<NuCaPl>	‘thousands’

Ordinal numerals are derived from their cardinal numerals. The ordinal numerals from first to tenth have singular and plural forms, the others don’t.

Tigrinya		**Tag**	**Gloss**
ቀዳማይ	**/k’ədda m-ay/**	**<NuOrMSg>**	**‘first’**
ቀዳመይቲ	**/k’əddam-əyti/**	**<NuOrFSg>**	**‘first’**
ቀዳሞት	**/k’əddam-ot/**	**<NuOrPl>**	**‘first’**
ሓምሻይ	**/ħamʃ-ay/**	**<NuOrMSg>**	**‘fifth’**

ሓምሻይቲ	**/ħamʃ-əyti/**	**<NuOrFSg>**	**‘fifth’**
ዓስራይ	**/ʕasr-ay/**	**<NuOrMSg>**	**‘tenth’**
ዓስረይቲ	**/ʕasr-əyti/**	**<NuOrFSg>**	**‘tenth’**
ዓስረዎት	**/ʕasr-əwot/**	**<NuOrPl>**	**‘tenth**

Tigrinya has distinctive forms for the fractions one-half to one-tenth. Other fractions like ‘two third’ and ‘three fourth’ can be formed by placing the cardinal numeral before the fraction.

Tigrinya		**Tag**	**Gloss**
ሲሶ	/siso/	<NuN>	‘one third’
ሕምሺት	/ħɨmmɨʃit/	<NuN>	‘one fifth’
ክልተ ሲሶ	/kɨltə siso/	<NuCaN>	‘two third’
ሰለስተ ርብዒት	/sələstə rɨbʕit/	<NuCaN>	‘three fourth’
ሳልሳይ ኣፍ (ሳልሳይ ኢድ)	/salsay ʔaf/ (/salsay ʔid/)	<NuCaN>	‘one third’

When we sum up the Numeral category:

Category	**Numeral <Nu>**
Type	Cardinal <NuCa>
	Ordinal <NuOr>
Function	Noun <N>
	Pronoun <Pn>
	Determiner <Dt>
Gender	Masculine <M>
	Feminine <F>
Number	Singular <Sg>
	Plural <Pl>
Definiteness	Definite <Df>
	Indefinite <Id>

Table 14: Numeral features of Tigrinya

4.2.12. Phrases <P>

A phrase is a group of words, without both a subject and predicate, conveying an idea which is not complete. Phrases combine words into a larger unit that can function as a sentence element.

According to crystal (2008:367) a phrase refers to a single grammatical element of structure typically containing more than one word, and lacking the subject–predicate structure typical of clauses. In syntax, phrases consist of minimally a head; and it is the syntactic category of the head that determines the category of the phrase. There are different types of phrases. A phrase with an adjectival head is an adjective phrase <AjP>, a phrase with a noun as head is a noun phrase <NP>, etc. For instance, ***very good*** is an <AjP> with the adjective ***good*** as its head, and ***hard work*** is a <NP> with the noun ***work*** as its head.

Girma and Mesfin (2006:7-9) discovered different types of phrases, after they tagged the WIC (Walta Information Center) news texts of Amharic. They discovered words or morphemes attached with nouns, pronouns, adjectives, verbs and numerical phrases and they have assigned tags for each of them.

Tigrinya phrases can be found attached, as proclitic[14] or enclitic[15], or separated. If we take some prepositions and conjunctions as an instance we can find them attached to another word or separated.

Tigrinya	ናይ ፀጋይ	ንፀጋይ
IPA	/naj sʼəggaj/	/nɨ-sʼggaj/
Tag	<Pp> <NPo>	<NP>
Gloss	ʻof Tsegayʼ (Tsegayʼs)	ʻto Tsegayʼ

As the above example shows, if the phrase is separated the tag will be assigned as separated like <Pp> <NPo>; if the phrase is found attached, so the tag will be assigned as one <NP>. Since the main category label is put at first, for our phrasal tags we put <P> before the other attributes and values. Next we will see different phrases which are found in Tigrinya texts.

[14] Proclitics are words or morphemes which depend upon a following word (Crystal, 2008:80).
[15] Enclitics are words or morphemes whichdepend upon a preceding word (Crystal, 2008:80).

4.2.12.1. Phrasal nouns

Phrasal nouns are words or morphemes attached with nouns.

Noun and preposition

Tigrinya	**ን-ዕዳጋ**
IPA	**/**nɨ-ʔɨdaga**/**
Tag	<PNPp>
Gloss	**‘**to market**’**

Noun and conjunction

Tigrinya	**ግደይ-ን**
IPA	**/**gɨdəj-ɨn**/**
Tag	<PNCj>
Gloss	**‘**Gidey and**’**

Noun, preposition and conjunction

Tigrinya	**ን-ስላስ-ን**
IPA	**/**nɨ-sɨllas-ɨn**/**
Tag	<PNPpCj>
Gloss	**‘**to Sillas and**’**

4.2.12.2. Phrasal pronouns

Phrasal pronouns are words or morphemes attached with pronouns.

Pronoun and preposition

Tigrinya	**ን-ርእሱ**
IPA	**/**nɨ-rɨʔsu**/**
Tag	<PPnPp>
Gloss	**‘**to himself**’**

Pronoun and conjunction

Tigrinya	**እዚ-ን**
IPA	**/ʔɨzi-n/**
Tag	<PPnCj>
Gloss	**‘**this and**’**

Pronoun, preposition and conjunction

Tigrinya	**በዚ-ን (ብ-እዚ-ን)**
IPA	**/**bəzin**/ (/**bɨ-ʔɨzi-n**/**
Tag	<PPnPpCj>
Gloss	**‘**with this and**’**

4.2.12.3. Phrasal verbs

Phrasal verbs are words or morphemes attached with verbs.

Verb relative

Tigrinya	**ዘለዎ (ዝ-ኣለ-ዎ)**
IPA	**/**zəlləwwo**/ (/**zɨ-ʔallə-wwo**/)**
Tag	<PVRl>
Gloss	**‘**the one who has**’**

Verb and preposition

Tigrinya	**ከም-ዝበሎ**
IPA	**/**kəm-zɨbəlo**/**
Tag	<PVPp>
Gloss	**‘**as he said**’**

Verb and conjunction

Tigrinya	**መጺኡ-ን**
IPA	**/**məsʼiʔu-n**/**
Tag	<PVCj>
Gloss	**‘**he came and**’**

Verb, preposition and conjunction

Tigrinya	**ከ-ገልግሉ-ን (ክ-ኣገልግሉ-ን)**
IPA	**/**kə-gəlgɨlu-n**/ (/**kɨ-ʔagəlgɨlu-n**/**
Tag	<PVPpCj>
Gloss	**‘**to serve and**’**

4.2.12.4. Phrasal adjectives

Phrasal adjectives are words or morphemes attached with adjectives.

Adjective and preposition

Tigrinya	**ብ-ሓጺር**
IPA	**/**bɨ-ħasʼsʼir**/**
Tag	<PAjPp>
Gloss	**‘**with a short**’**

Adjective and conjunction

Tigrinya	**ኢትዮጵያዊ-ን**
IPA	**/**ʔitjopʼɨjawi-n**/**
Tag	<PAjCj>
Gloss	**‘**an Ethiopian and**’**

Adjective preposition and conjunction

Tigrinya	**ብ-መንፈሳዊ-ን**
IPA	**/**bɨ-mənfəsawi-n**/**
Tag	<PAjPpCj>
Gloss	**‘**with spiritual and**’**

4.2.12.5. Phrasal numerals

Phrasal numerals are words or morphemes attached with numerals.

Numeral and preposition

Tigrinya	**ብ-ሰለስተ**
IPA	**/**bɨ-sələstə**/**
Tag	<PNuPp>
Gloss	**‘**with three**’**

Numerical and conjunction

Tigrinya	**ሓደ-ን**
IPA	**/**ħadə-n**/**
Tag	<PNuCj>
Gloss	**‘**one and**’**

Numerical, preposition and conjunction

Tigrinya	**ን-ሓምሻይ-ን**
IPA	**/**nɨ-ħamʃaj-ɨn**/**
Tag	<PNuPpCj>
Gloss	**‘**to fifth and**’**

When we summarize the phrasals it looks as follows:

Category	Phrase <P>	Definition of the tag
Type	Phrasal noun <PN>	Noun and preposition <PNPp>
		Noun and conjunction <PNCj>
		Noun, preposition and conjunction <PNPpCj>
	Phrasal pronoun <PPn>	Pronoun and preposition <PPnPp>
		Pronoun and conjunction <PPnCj>
		Pronoun, preposition and conjunction <PPnPpCj>
	Phrasal verb <PV>	Verb and preposition <PVPp>
		Verb and conjunction <PVCj>
		Verb, preposition and conjunction <PVPpCj>
	Phrasal Adjective <PAj>	Adjective and preposition <PAjPp>
		Adjective and conjunction <PAjCj>
		Adjective, preposition and conjunction <PAjPpCj>
	Phrasal numeral <PNu>	Numeral and preposition <PNuPp>
		Numeral and conjunction <PNuCj>
		Numeral, preposition and conjunction <PNuPpCj>

Table 15: phrasal features of Tigrinya

4.2.13. Compound words <Cp>

Compound words are very well discussed in page 19 under 2.2.2.1.2 b. When muti-words are found attached with '-', or attached with eachother, we will take them as one word and assign one tag; but when they are found separated with a white space, they will be treated as different words and different tags will be assigned accordingly.

When compound words are attached with '-':

Tigrinya	ፀዋር-ደርሆ
IPA	/sˈəwar-dərho/
Tag	<CpN>
Gloss	'briber'

When compound words are attached with eachother:

Tigrinya	ልቢወለድ
IPA	/lɨbbiwəlləd/
Tag	<CpN>
Gloss	‘novel’

When compound words are separated with a white space:

Tigrinya	ልቢ ወለድ
IPA	/lɨbbi wəlləd/
Tag	<NC> <NC>
Gloss	‘novel’

Category	Compound <Cp>

Table 16: compound feature of Tigrinya

4.2.14. Contractions <Cn>

Contraction refers to the process or result of phonologically reducing a linguistic form so that it comes to be attached to an adjacent linguistic form, or fusing a sequence of forms so that they appear as a single form. (Crystal, 2008:111) In Tigrinya texts it may be found contracted with the punctuation apostrophe (’).

Tigrinya	ኣነ’ውን (ኣነ እውን)	
IPA	/ʔanə + ’ + wwɨn/	(/ʔanə ʔɨwwɨn/)
Tag	<Pn> <Cn Cj>	<Pn> <Cj>
Gloss	‘me too/and me’	

Category	Contraction <Cn>

Table 17: Contraction feature of Tigrinya

4.2.15. Symbols <Sy>

Symbols are characters which are not used as punctuation marks and neither are alphabets of the language. They are like the characters: $, @, &, +, % etc. Symbols can be mathematical, currency or other symbols found in the texts of Tigrinya.

Tigrinya	1. 10 + %	2. $ + 100 + . + 00
Tag	<NuCa> <Sy>	<Sy> <NuCa Pu NuCa>
Gloss	‘ten percent’,	‘hundred dollars’

Category	Symbol <Sy>

Table 18: Symbol feature of Tigrinya

4.2.16. Abbreviations <Ab>

Abbreviations (short forms/acronyms) are well discussed under 2.2.2.2.c. These abbreviations can be found in Tigrinya texts formed with ‘.’, ‘/’ or by taking the initials only.

Tigrinya	1. ማልት (ማሕበር ልምዓት ትግራይ)	2. ወ/ሮ (ወይዘሮ)
IPA	/malɨt/ (/maħbər lɨmʃat tɨgraj/)	/wə/ro/ /wəjzəro/
Tag	<Ab> <NC> <NC> <NPo>	<Ab> <Adj>
Gloss	‘TDA (Tigray Development Association)’,	Mrs. (Mistress)

Category	Abbreviation <Ab>

Table 19: Abbreviation feature of Tigrinya

4.2.17. Foreign words <Fn>

These words are those types of words inserted into the texts of Tigrinya from foreign languages for some reason. These words are labeled as <Fn>. For instance if we take the Tigrinya phrase ትምህርታዊ ምብፃሕ /tɨmhɨrtawi mɨbs’aħ/ ‘academic visit’, we can find it written as ትምህርታዊ ምብፃሕ (academic visit) in texts. Therefore the English words ‘academic’ and ‘visit’ should be given the tag name foreign word <Fn>.

Category	Foreign word <Fn>

Table 20: Foreign word feature of Tigrinya

4.2.18. Residuals <Rs>

These residuals are those which are found in Tigrinya texts and those which are not classified into one of the above categories, left without tag as left over.

Category	Residual <Rs>

Table 21: Features for the residuals of Tigrinya words

4.3. Summary of the morphosyntactic tagset

We have developed 18 tags for Tigrinya at the higher major category level. Each level has its features under it. When we see the whole tagset, it looks like the following.

Tigrinya morphosyntactic tagset (TIGTAGS)

Category 1	**Noun <N>**	
Type	Common <NC>	Countable <NCCt>
		Mass <NCMs>
		Concrete <NCCr>
		Abstract <NCAs>
	Proper <NPo>	
	Verbal noun <Nv>	
	Possessive <NPs>	
Gender	Masculine <NM>	
	Feminine <NF>	
Number	Singular <NSg>	
	Plural <NPl>	
	Honorific/respect <NH>	
Category 2	**Verb <V>**	
Type	Main <VMn>	

	Auxiliary <VAx>	
Aspect	Perfect <VPf>	
	Imperfect <VIp>	
	Gerundive <VGe>	
	Infinitive <VIf>	
Gender	Masculine <M>	
	Feminine <F>	
Number	Singular <Sg>	
	Plural <Pl>	
	Honorific/respect <H>	
Transitiveness	Transitive <Tr>	
	Intransitive <It>	
Tense	Past <VPt>	
	Present <VPr>	
	Future <VFu>	
Person	First <1>	
	Second <2>	
	Third <3>	
Mood	Imperative/Jussive <Im>	
	Interrogative <Ig>	
	Negative <Ng>	
	Indicative <Ic>	
	Subjunctive <Sb>	
Voice	Active <At>	
	Passive <Pv>	
Category 3	**Pronoun <Pn>**	
Type	Personal <PnPe>	
	Reflexive <PnRf>	
	Reciprocal <PnRc>	

	Possessive <PnPs>	
	Demonstrative <PnDm>	
	Indefinite <PnId>	
	Interrogative <PnIg>	
Case	Nominative <PnNm>	
	Accusative <PnAc>	
	Genitive <PnGn>	
Gender	Masculine <M>	
	Feminine <F>	
Number	Singular <Sg>	
	Plural <Pl>	
	Honorific <H>	
Person	First <1>	
	Second <2>	
	Third <3>	
Category 4	**Adjective <Aj>**	
Gender	Masculine <M>	
	Feminine <F>	
Number	Singular <Sg>	
	Plural <Pl>	
Person	First <1>	
	Second <2>	
	Third <3>	
Category 5	**Adverb <Av>**	
	Degree <AvDg>	
	Temporal <AvTm>	
	Manner <AvMr>	
	Directional <AvDr>	
	Locative <AvLc>	

Category 6	**Preposition <Pp>**	
Category 7	**Conjunction <Cj>**	
Type	Coordinating <CjCo>	
	Subordinating <CjSu>	
Category 8	**Determiner <Dt>**	
Gender	Masculine <M>	
	Feminine <F>	
Number	Singular <Sg>	
	Plural <Pl>	
Case	Nominative <Nm>	
	Accusative <Ac>	
	Genitive <Gn>	
Category 9	**Interjection <Ij>**	
Category 10	**Punctuation <Pu>**	
Category 11	**Numeral <Nu>**	
Type	Cardinal <NuCa>	
	Ordinal <NuOr>	
Function	Noun <N>	
	Pronoun <Pn>	
	Determiner <Dt>	
Gender	Masculine <M>	
	Feminine <F>	
Number	Singular <Sg>	
	Plural <Pl>	
Definiteness	Definite <Df>	
	Indefinite <Id>	
Category 12	**Phrase <P>**	
Type	Phrasal noun <PN>	Noun and preposition <PNPp>
		Noun and conjunction <PNCj>

		Noun, preposition and conjunction <PNPpCj>
	Phrasal pronoun <PPn>	Pronoun and preposition <PPnPp>
		Pronoun and conjunction <PPnCj>
		Pronoun, preposition and conjunction <PPnPpCj>
	Phrasal verb <PV>	Verb and preposition <PVPp>
		Verb and conjunction <PVCj>
		Verb, preposition and conjunction <PVPpCj>
	Phrasal Adjective <PAj>	Adjective and preposition <PAjPp>
		Adjective and conjunction <PAjCj>
		Adjective, preposition and conjunction <PAjPpCj>
	Phrasal numeral <PNu>	Numeral and preposition <PNuPp>
		Numeral and conjunction <PNuCj>
		Numeral, preposition and conjunction <PNuPpCj>
Category 13	**Compound <Cp>**	
Category 14	**Contraction <Cn>**	
Category 15	**Symbol <Sy>**	
Category 16	**Abbreviation <Ab>**	
Category 17	**Foreign word <Fn>**	
Category 18	**Residual <Rs>**	

Table 22: Tigrinya morphosyntactic tagset

When we count all the tags which are assigned for every major category, they all become **139 tags**. These can also be reduced to the upper level of the category and become **105 tags**. And again these tags can also be reduced to the higher main category level and become **18 tags.** Any person can reduce the tagset by decreasing some features which he/she thinks are not relevant for his/her purpose, and can also extend to a more finer level and increase the number of tags to his/her purpose.

Chapter Five

Conclusion and Recommendations

5.1. Conclusion

Early in our research we realized that no attempt has been made to develop a morphosyntactic tagset for Tigrinya. The basis for many NLP applications, the morphosyntactic tagset development, was being constructed up to now for Tigrinya. Since there is almost no NLP resource, we think that Tigrinya is lucky to start with a comprehensive morphosyntactic tagset development. We have analyzed the morphosyntactic tagset by referring different grammar books of Tigrinya. We have attempted to see the features, which are important for the main and subcategories of the words in a written text. Since the tagset is morphosyntactic, it is also essential to analyze the agreement of morphemes in words which are related with subject and object of the context. Eventually names/tags are given for every word in the text, and if a word is not given a POS name, then it is labeled as Residual <Rs>.

Now, Tigrinya has 18 POS tags at a higher coarse-grained level and 105 morphosyntactic tags to the next lower level. Starting from this we can extend to the level we wish adding the features more and more. When we come to next lower level we get 139 tags.

We consider that this Tigrinya morphosyntactic tagset is comprehensive in that they it considers the morphosyntactic features of Tigrinya and any researcher who wishes to construct NLP application can use it very well. And we believe that the tag names are distinct, clear and analyzable.

In the process of Tigrinya morphosyntactic tagset development, we couldn't find a comprehensive literature concerning morphosyntactic tagset development, except those of Atwell (2008) and Ermias (forthcoming). These materials helped us a lot to develop the morphosyntactic tagset for Tigrinya. The other challenge in the process was that the grammar books found written in the language are not to the extent as comprehensive as we wish them to be for our purpose. For this reason we have tried to accommodate different ideas from different Tigrinya grammar books to construct a tag for the feature.

We have tried to follow some type of standard from EAGLES and Ermias to develop morphosyntactic tagset for Tigrinya, and it was worth following. We have found that guidelines and criteria set by Atwell and Leech were also worth to be assumed as our guidelines and criteria.

5.2. Recommendations

We have tried to accomplish our objective of analyzing and developing morphosyntactic tagset for Tigrinya. Though we have said that our morphosyntactic tagset is comprehensive, we have accommodated multi-word expressions as separate words and are labeled accordingly. We did this for the reason that there is no a thorough research done on multi-word (compound words). The tagset is tested by manual annotation. Therefore we recommend that:

- Since this morphosyntactic tagset is the basis for POS tagger and other NLP applications construction, we have paved the way to construct a POS tagger and other NLP applications for Tigrinya. Therefore we recommend that any future researcher on Tigrinya NLP resources to construct a POS tagger, even a tree bank.
- There are different types of formula, especially in mathematics and the sciences; therefore they need a thorough study and assign labels to them.
- Those words which start with ተ- /tə-/ need more study, and accordingly a researcher may reach to a detailed list of features for each group of words. As a result one could be able to label them easily.
- Dates create ambiguity when they are given a tag name. Therefore they need to be studied and analyzed very well to deal with them and give them a clear and simple tag.
- It is our anticipation that researchers will do a comprehensive and well designed grammar in Tigrinya. This helps Tigrinya Computational Linguistics researchers who would like to devise some type of NLP resources for the language, because the main focus of Computational Linguistics or NLP is processing natural language in the computer**.**
- This Tigrinya morphosyntactic tagset (TIGTAGS) can also be a model for sisterly languages and other indigenous languages to develop a morphosyntactic tagset.

References

Agić, Željko, Marko Tadić and Zdravko Dovedan 2009. *Tagset Reductions in Morphosyntactic Tagging of Croatian Texts.* In Future 2009: "Digital Resources and Knowledge Sharing", University of Zagreb, Zagreb, Croatia.

Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (2001). *An Introduction to Language and Communication.* (5th edition). London, England: The MIT Press.

Allwood, J. L., & Hendrikse, A. P. *Developing a tag set and tagger for the African languages of South Africa with special reference to Xhosa.* Sweden and RSA: University of Gothenburg and University of South Africa.

Alqrainy, S. (2008). *A Morphological - Syntactical Analysis Approach for Arabic Textual Tagging.* Doctor of Philosophy thesis. De Montfort University.

Alqrainy, S., & Ayesh, A. (2006). *Developing a Tagset for Automated PoS Tagging of Arabic. WSEAS TRANSACTIONS on COMPUTERS , 5* (11), 2787 – 2792.

Amanuel Sahle (1998). *ሰዋስው ትግርኛ ብሰፊሑ። 'Sewasiw Tigrinya Bisefihu'* (*A Comprehensive Grammar of Tigrinya*). Lawrenceville, New Jersey and Asmara: The Red Sea Press.

Appleyard, D. (2006). *"Tigrinya". In Brown, Keith and Sarah Ogilvie (editors) (2009). Concise encyclopedia of languages of the world.* Oxford, UK: Elsevier Ltd.

Aronoff, M., & Rees-Miller, J. (Eds.). (2003). *The Handbook of Linguistics.* UK: Blackwell Publishers Ltd.

Atwell, E. (2008). *Development of tag sets for part of speech tagging. In: Corpus Liuistics: An International Handbook.* (A. Ludeling, & M. Kyto, Eds.) 1, 501–526.

Bar-Haim, R., Simaan, K., & Winter, Y. (2008). *Part-of-speech tagging of Modern Hebrew text. Natural Language Engineering* , 14, 223–251.

Bassano, F. d. (1918). *Vocabulario tigray-italiano e repertorio italiano-tigray.* Roma, Italy: Luigi.

Baye, Y. (2008). *የአማርኛ ሰዋስው። Ye'amarinya sewasiw. 'Amharic Grammar'. (2nd edition). , .* Addis Ababa: Eleni publishing private limited company.

Bender, M. L. (1976). *"Two Ethio-Semitic Languages." In Language in Ethiopia.* London: Oxford University Press.

Booij, G. (2007). *The Grammar of Words: An Introduction to Linguistic Morphology. (2nd edition).* UK: Oxford University Press.

Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics.* (G. Trauth, & K. Kazzazi, eds.) London, UK: Routledge.

Caressa, F. (1935). *Vocaboli delle lingue amarica, galla, tigrina. Manuale linguistica per l'Africa Orientale Coi principali.* Torino, Italy.

Chefena Hailemariam, Kroon, S., & Walters, J. (1999). "Multilingualism and Nation Building: Language and Ecucation in Eritrea.*"* In *Journal of Multilingual and Multicultural Development.* 20 (6), (479).

CIA. (2012). *The World Factbook: Ethiopia.* Washington, DC: The Central Intelligence Agency. *https://www.cia.gov/library/publications/the-world-factbook/fields/2075.html#er*.

CIA. (2012). *The World Factbook: Ethiopia.* Washington, DC: The Central Intelligence Agency. *https://www.cia.gov/library/publications/the-world-factbook/geos/et.html*.

Cimino, A. (1904). *Vocabulario italiano-tigrai e tigrai-italiano.* Asmara.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics, (6th edition).* Oxford, UK: Blackwell Publishing.

CSA. (2010). *Population and Housing Census of 2007.* Addis Ababa, Ethiopia: Ethiopia Central Statistical Agency. *http://www.csa.gov.et/index.php?Itemid=590*.

Daniel, M. (1998). *"The Origin and Development of Tigrinya Language". In Proceedings of the Tigrinya Language Symposium*. Mekele.

Di Lello, R. (1995). *A Peanian Analysis of Tigrinya. Metalogicon (1995) , VIII* (1).

Ermias Abebe Kassa (forth coming). *Tagsets for the Morphosyntactic Annotation of Amharic Text.* Addis Ababa University, School of Information Science, Addis Ababa, Ethiopia.

Esayas Tajebe (2003). *Nominal Functional Categories in Tigrinya.* Master's Thesis, Addis Ababa University, School of Graduate Studies. Addis Ababa, Ethiopia.

Gallina, F. (1894). *Indovinelli tigrai*. (Vol. I). Rome.

Gasser, M. (2009). *Semitic Morphological Analysis and Generation Using Finite State Transducers with Feature Structures*. Conference of the European Chapter of the Association for Computational Linguistics, 12., (pp. 309-317).

Gasser, M. (2010). *Expanding the Lexicon for a Resource-Poor Language Using a Morphological Analyzer and a Web Crawler*. Conference on Language Resources and Evaluation, 7.

Getachew Mamo (2009). *Part-of-Speech Tagging for Afaan Oromo Language.* Master's thesis, Addis Ababa University. Addis Ababa, Ethiopia.

Girma Awgichew Demeke and Mesfin Getachew (2006). *Manual annotation of Amharic news items with part-of-speech tags and its challenges. ELRC Working Papers.* Ethiopian Languages Research Center, Addis Ababa University. II (1).

Girma Awgichew Demeke (2006). *Understanding Word Classes.* In *Issues on Lexicography.* Thomas Belay, Davi, A. U., & Girma Awgichew Demeke (Eds.). Workshop organized by the Ethiopian Languages Research Center, Addis Ababa University. Addis Ababa, Ethipia: Addis Ababa University Printing Press.

Girmay Berhane (1991). *Issues in the Phonology and Morphology of Tigrinya.* PhD Thesis presented in partial fulfillment for the degree of Doctor of Philosophy. Montreal, Canada: University of Quebec a Montreal.

Hardie, A. (2003). *Developing a tagset for automated part-of-speech tagging in Urdu. Corpus Linguistics.*

Hardie, A. (2003). *The computational analysis of morphosyntactic categories in Urdu.* Thesis submitted for the degree of PhD Department of Linguistics and Modern English Language, Lancaster University. Lancaster, UK.

Indukhya, N., & Damerau, F. J. (2010). *Handbook of Natural Language Processing* (2nd edition). USA: Taylor and Francis Group.

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing: An Introduction toNatural Language Processing, Speech Recognition, and Computational Linguistics* (2nd edition). Upper Saddle River, NJ: Prentice-Hall.

Kane, T. L. (2000). *Tigrinya – English Dictionary.* Springfield, U.S.A.: Dunwoody Press.

Keffyalew, G. (2012). *The Alienable-Inalienable Asymmetry: Evidence from Tigrinya. Selected Proceedings of the 42nd Annual Conference on African Linguistics.* In R. M. Michael (Ed.). (pp. 161-182). Somerville, MA: Cascadilla Proceedings Project.

Khoja, S. (2003). *APT: Arabic Part-of-speech Tagger.* A PhD Dissertation. Computing Department, Lancaster University. Lancaster, UK.

Khoja, S., Garside, R., & Knowles, G. (2001). *An Arabic Tagset for the Morphosyntactic Tagging of Arabic. Corpus Linguistics 2001*. Lancaster University. Lancaster, UK.

Leech, G. (1997). *A brief users' guide to the grammatical tagging of the British National Corpus.* UCREL, Lancaster University. Lancaster. http://www.natcorp.ox.ac.uk/docs/gramtag.html.

Leech, G. (2005). *Developing linguistic corpora: a guide to good practice.* Oxford, UK: Oxbow Books.

Leech, G., & A., W. (1996). *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora.* Department of Linguistics and Modern English Language, Lancaster University. Lancaster, United Kingdom: Lancaster University. http://www.ilc.pi.cnr.it/EAGLES96/annotate/.

Leonessa, M. d. (1928). *Grammatica analitica della lingua tigray*. Rome.

Leslau, W. (1941). *Documents Tigrinya (ethiopien septentrional): grammaire et texts.* Paris: Librairie C. Klincksieck.

Lüdeling, A., & Kytö, M. (2008). *Corpus linguistics: an international handbook.* (Vol. 1). Berlin, Germany: Walter de Gruyter GmbH & Co. KG.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). *Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics , 19* (2), 313-330.

McCarthy, A. C. (2002). *An Introduction to English Morphology: Words and Their Structure.* Edinburgh, Scotland: Edinburgh University Press.

Mesfin Getachew (2001). *Automatic part of speech tagging for Amharic language: an experiment using stochastic hmm.* Master's thesis, Addis Ababa University. Addis Ababa, Ethiopia.

Microsoft Corporation (1993-2008). Mictosoft Encarta 2009.

Mueller, M. (2009). *A part of speech tag set for written English from Chaucer to the present.* NUPOS.

Payne, T. E. (1997). *Describing Morphosyntax: A Guide for field linguists.* UK: Cambridge University Press.

Rama Sree, R., Rao G., U. M., & K.V., D. M. (2008). *Assessment and Development of POS Tag set for Telugu.* Proceedings of W5: The 6th workshop on Asian Language Resources (ALR)., (pp. 85-88). Hyderabad.

Bar-Haim, R., Simaan, K., and Winter, Y. (2008). *Part-of-speech tagging of Modern Hebrew text.* Natural Language Engineering, (14), 223–251.

Santorini, B. (1991). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.*

Sawalha, M. S. (2011). *Open-source Resources and Standards for Arabic Word Structure Analysis: Fine-grained Morphological Analysis of Arabic Text Corpora.* Presented for the degree of Doctor of Philosophy. University of Leeds, School of Computing. Leeds.

Selamawi, T. (2011). *Phonological Description of Eastern Tigrinya.* Master's degree thesis, Addis Ababa University. Addis Ababa, Ethiopia.

Sima'an, K. A., Winter, Y., Altman, A., & Nativ, N. (2001). *Building a Tree-Bank of Modern Hebrew Text. In: Traitment Automatique des Langues (2001).* In *Traitment Automatique des Langues (2001)* (pp. 347 - 380).

Singha, K. R., Purkayastha, B. S., Singha, K. D., & Roy, A. (2011). *Developing a Tagset for Manipuri Part of Speech Tagging. Journal of Computer Science and Engineering , 5* (1).

Sinha, S. (2010). "Issues in POS Tagset Design". In: Indian Languages and Part-of-Speech Annotation. *Linguistic Data Consortium for Indian Language*, (pp. 7-13). Mysore. http://samarsinha.blogspot.com/2010/11/issues-in-pos-tagset-design.html

Sisay Fissaha Adafre (2005). *Part of speech tagging for Amharic using conditional random fields. In Workshop on Computational Approaches to Semitic Languages.* ACL (2005). *Workshop on Computational Approaches to Semitic Languages*, (pp. 47–54).

Tekeste, T., Daniel, M., Tsehaynesh, G., Tsegay, W., Tadese, T., & Tesfay, T. (1989 E.C.). *መዝገበ ቃላት ትግርኛ ብትግርኛ. 'Tigrinya Dictionary'.* Addis Ababa: ንግድ ማተሚያ ድርጅት::

Tesfay Tewelde Yohannes (2002). *A Modern grammar of Tigrinya.* Roma, Italy: Via G. Savonarola.

Todd, L. (1987). *Introduction to Linguistics. Longman.* UK: York Press.

Trauth, G., & Kazzazi, K. (1996). *Routledge Dictionary of Language and Linguistics.* London, UK: Routledge 11 New Fetter Lane.

Trauth, Gregory and Kerstin Kazzazi editors 1996. *Routledge Dictionary of Language and Linguistics.* Routledge 11 New Fetter Lane, London, UK.

Tsegay Woldemariam (1974 E.C.). *የትግርኛ ስሞች አመሰራረት። Yetigrinya Simoch Ameseraret.* 'Nominalization in Tigrinya'. B.A. Thesis. Addis Ababa University: Addis Ababa, Ethiopia.

Tsehaye Teferra (1979). *Reference grammar of Tigrinya.* Georgetown University, Washington, D.C., USA.

Weldu M. Weldeyesus (2004). *Case Marking Systems in Two Ethiopian Semitic Languages*. *Colorado Research in Linguistics*. 17, (1). Boulder: University of Colorado.

Zelalem Leyew (2009). *Lexical Development in Tigrinya*. In, *Journal of Education for Development.* III (II). Addis Ababa University.

Websites referred

http://www.ling.upenn.edu/courses/ling202/WritingSystem.html

www.omniglot.com/writing/syllabic.htm

http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

http://info.ox.ac.uk/bnc

https://www.cia.gov/library/publications/the-world-factbook/geos/et.html

http://www.csa.gov.et/index.php?Itemid=590

http://www.ilc.pi.cnr.it/EAGLES96/annotate/

http://www.natcorp.ox.ac.uk/docs/gramtag.html

http://samarsinha.blogspot.com/2010/11/issues-in-pos-tagset-design.html

Appendices

Appendix I

Tigrigna consonants (alphabet) with their phonetic description.

IPA	Description	ə	u	i	a	e	0/ɨ	o
p	voiceless bilabial stop	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
b	voiced bilabial stop	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
p'	glottalized bilabial stop	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
m	bilabial nasal	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
f	voiceless labio-dental fricative	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
w	voiced labiovelar approximant	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
t	voiceless alveolar stop	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
d	voiced alveolar stop	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
t'	glottalized alveolar stop	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
s'	glottalized alveolar affricate	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
n	alveolar nasal	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
s	voiceless alveolar fricative	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
z	voiced alveolar fricative	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
r	alveolar trill	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
l	lateral alveolar approximant	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ʧ	voiceless palatal affricate	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
ʤ	voiced palatal affricate	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
ʧ'	glottalized palatal affricate	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
ɲ	palatal nasal	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
ʃ	voiceless palatal fricative	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
ʒ	voiced palatal fricative	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
j	voiced palatal approximant	የ	ዩ	ዪ	ያ	ዮ	ይ	ዮ
k	voiceless velar stop	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
k^{w}	voiceless labialized velar stop	ኰ		ኵ	ኳ	ኴ	ኵ	
g	voiced velar stop	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
g^{w}	voiced labialized velar stop	ጐ		ጕ	ጓ	ጔ	ጕ	
k'	glottalized velar stop	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
k'^{w}	glottalized labialized velar stop	ቈ		ቊ	ቋ	ቌ	ቍ	

χ	voiceless velar fricative	ⵠ	ⵠⵓ	ⵠⵉ	ⵠⴰ	ⵠⴻ	ⵠⵃ	ⵠⵄ
χ^{w}	voiceless labialized velar fricative	ⵠⵎ		ⵠⵓ	ⵠⵔ	ⵠⵖ	ⵠⵓ	
χ'	voiced glottalized velar fricative	ⵖ	ⵖⵓ	ⵖⵉ	ⵖⴰ	ⵖⴻ	ⵖⵃ	ⵖⵄ
χ'^{w}	voiced labialized velar fricative	ⵗⵓ		ⵗⵓ	ⵗⴰ	ⵗⴻ	ⵗⵓ	
ħ	voiceless pharyngeal fricative	ⵀ	ⵀⵓ	ⵀⵉ	ⵀⴰ	ⵀⴻ	ⵀⵃ	ⵀⵄ
ʕ	voiced pharyngeal fricative	ⵄ	ⵄⵓ	ⵄⵉ	ⵄⴰ	ⵄⴻ	ⵄⵃ	ⵄⵄ
ʔ	glottal stop	ⴰ	ⴰⵓ	ⴰⵉ	ⴰⴰ	ⴰⴻ	ⴰⵃ	ⴰⵄ
h	voiceless glottal fricative	ⵀ	ⵀⵓ	ⵀⵉ	ⵀⴰ	ⵀⴻ	ⵀⵃ	ⵀⵄ

Appendix II

Tigrinya sample text tagged with the developed morphosyntactic tagset

This text is taken from www.tigraionline.com, and to see how our tagset is working we have tagged manually the first paragraph from this text.

ድሕሪ<Aj>	ሃንደበታዊ<Aj3MSg>	መስዋእቲ<NCAsSg>	ክቡር<Aj3MSg>	መራሒና<NPs3MPl>
dɨ̈ħri	*handəbətawi*	*məswaʔti*	*kɨ̈bur*	*məraħina*
After	(a) suddenly	sacrifice (of)	honourable	our leader

ቀዳማይ<NuOrAjMSg>	ሚኒስተር<NCSg>	መለስ<NPo>	ዜናዊ<NPo>	ህዝቢ<NCMsSg>
k'əddamay	*ministər*	*məlləs*	*zenawi*	*hɨ̈zbi*
prime/first	minister	Meles	Zenawi	people (of)

ትግራይ<NPo>	መሪር<Aj3MSg>	ሓዘኑ<NPsAsMSg>	ብምግላፅ<PNvPp>	ምስ<Pp>	ካልኦት<Aj3Pl>
tɨ̈gray	*mərrir*	*ħazənu*	*bɨ̈mɨ̈glas'*	*mɨ̈s*	*kalʔot*
Tigray	deep	his condolence	by expressing	with	others

ኣሓት<NCCtFPl>	ህዝብታት<NCMsPl>	ኢትዮጵያ<NPo>	ኮይኑ<VAxMSgS>	ዕላማታት<NCPl>
ʔaħat	*hɨ̈zbɨ̈tat*	*ʔityop'ɨ̈ya*	*koynu*	*ʔɨ̈lamatat*
sisters	peoples (of)	Ethiopia	being he/it	goals (of)

መለስን<PNPoCj>	ባዕሉ<PnPeGe>	ዝጀመሮም<VMnIp3MSgIt>	ናይ<Pp>	ነዊሕ<Aj3MSg>
məlləsɨ̈n	*baʕlu*	*zɨ̈dʒəmmərom*	*nay*	*nəwwiħ*
Meles and	himself	he started	of	long

እዋናት<NCPl>	ዓበይቲ<AjPl>	ፕሮጀክታት<NCPl>	ንምዝዛምን<PNvPpCj>	ኣብ<Pp>	ሽቶኦም<NC3MPl>
ʔɨ̈wanat	*ʕabbəyti*	*pɨ̈rodʒəktɨ̈tat*	*nɨ̈mɨ̈zɨ̈zamɨ̈n*	*ʔab*	*ʃɨ̈toʔom*
times	bigs	projects	to finish and	to	their goals

ንምብፃሕን<PNvPpCj>	ካብኡ<Av3MSg>	ዘድሊ<VMnIp3MSg>	ዘበለ<VMnIp3MSg>	ኹሉ <Aj3MSg>
nɨ̈mɨ̈bs'aħɨ̈n	*kabʔu*	*zədɨ̈li*	*zəbbələ*	*χullu*
and to reach	from there	necessary	thing	every

ከም<Pp>	ዓቕሙን<PNC3MSgCj>	ክእለቱን <PNC3MSgCj>	እጃሙ<NC3MSg>	እናወፈየን <PVMnPf3MSgPpCj>
kəkkəm	*ʕaχ'mun*	*kɨ̈ʔɨ̈lətun*	*ʔɨ̈dʒdʒamu*	*ʔɨ̈nnawəffəyən*
as of	his potential	his ability	his part	by rendering and

ንቐፃሊ'ውን<PAj3MSgPpCj>	ከምዘወፍን<PVMnIp3MSgIcPpCj>	ናይ<Pp>	መሪሕነት<NCAs>	ተርኡ<NCPs3M>
nɨ̈χ'əs's'aliwwɨ̈n	*kəmzəwəffɨ̈n*	*nay*	*mərɨ̈ħnnət*	*tərɨ̈ʔu*
and for the future	that he will render	of	leadership	his role

ከምዝፃወትን<PVMnIp3MSgItPpCj>	ብህልኸን<PNCAsPpCj>	ሕራነን<PNCAbCj>	እናገለፀን<PVMnPf3MSgItPpCj>
kəmzɨ̈s's'awətɨ̈n	*bɨ̈hɨ̈llɨ̈χɨ̈n*	*ħɨ̈rranən*	*ʔɨ̈nnagələs'ən*
that he will play	with persistence	and determination and	by expressing and

ቃል<NCSg>	እናኣተወን<PVMnPf3MSgItPpCj>	ይርከብ<VMnIp3MSg>።<Pu>
k'al	*ʔɨnnaʔatəwən*	*yɨrɨkkəb*
pledge	by making	is found .

ኣብ'ዚ ሕዚ እዋን ብተመሳሳሊ ህልኽን ሕራነን ኣብ ወፃኢ ዝነብሩን ዝመሃሩን ምሁራት ተጋሩን ፈተውቲ ትግራይን ንተማሃሮ ትግራይ ብትምህርቲ፣ ብሞያ ሳይንስን ቴክኖሎጅን፣ ስግግር ቴክኖሎጅን ካልኦትን ተማሃሮ ትግራይ ዓለምለኻ ተወዳደርቲ ክኾኑን ዓለምና በፂሓቶ ኣብ ዘላ ብርኪ ዕብየትን ምዕባለን ተዋሳእቲ ኮይኖም ናይ ባዕሎም ተራ ዝፃወቱን ዓለምለኻ ፍልጠት ዝዓጠቑን ዝወነኑን ብቑዓት ሳይንቲስታት ዝፈርዩሉ መንገዲ ብምንዳይ TSAN ዝብል ውዳበ ኣጣይሾም ይርከቡ።

ቅድሚ ሓደ ዓመት "ክረምታዊ ወፍሪ ተጋሩ" ብዝብል ስያመ ናብ ትግራይ መፂአም ናፃ ናይ ትምህርቲ፣ ስልጠናን ምምኻርን ግልጋሎት ዝሃቡ እንትኾኑ ሕዚ ድማ ዓሳኣቴ TSAN (Tigreans Scientific and Academic Network) ዓለም ለኻ ሳይንሳውን ኣካዳምያውን ኔትዎርክ ተጋሩ ብዝብል ዓለምለኻዊ ኣካዳምያዊ ኔት ዎርክ ኣብ ወፃኢ ዝርከቡ ምሁራት ተጋሩ፣ ተማሃሮ ዩኒቨርሲቲን መምህራንን ዝሓቖፈ ውዳበ ብምጥያሽ ዓብይ ምንቕስቓስ እናገበሩ ይርከቡ። ቀንዲ ዕላማ ናይቲ ውዳበ ድማ ንሃገርናን ክልልናን ማእኸል ምርምርን መፅናዕትን ከምኡ'ውን ማእኸል ምይይጥ መድረኽ ንምግባር ዝዓለመ ምኻኑ ኣብቲ ዝወፀ መግለፂ ምርዳእ ይከኣል።

ብተወሳኺ'ውን ዝለዓለ ሳይንስን ቴክኖሎጅን ዝወነነ ሕብረተሰብ ክህሉ፣ ኣብ ዓለምለኻ ዕዳጋ ተወዳዳራይ ዝኸውን ዝተማህረ ሓይሊ ሰብ ንምፍራይ ከምዝሰርሕ ተገሊፁ ኣሎ። ልዕሊ 300 ኣባላት (ቀዳማይ ዲግሪን ልዕሊኡን) ዘለዎም ዝርከብዎ እዚ ውዳበ'ዚ ኣብ ሃገርና ኢሉ'ውን ኣብ ክልልና ቅልጡፍ፣ ቀፃልን ኩለመዳያዊ ምዕባለ ንክመፅእ ብሳይንስን ቴክኖሎጂን ዝተደገፈን ኣብ ፅኑዕ መሰረት (ባይታ) ዝተመስረተን ንክኸውን ዘለዉና ዕቑራት ፀጋታትን ዝሰልጠነ ሓይሊ ሰብን ብምውዳብን ምትሕብባርን ኣብ ዝሓፀረ ግዘ ዝለዓለ ዕብየትን ዝላን ዘመዝግብ ዘሎ ኔትዎርክ ምኻኑ ምርዳእ ይከኣል።

እዚ ፕሮግራም'ዚ መቐፀልታ ናይቲ ብ2010 ኣ.ፈ. ዝተገበረ ወፍሪ ተጋሩ ኣብ ወፃኢ ዝነብሩ ተመሃሮ ዩኒቨርሲትን መምህራንን ናብ ትግራይ መፂአም ግልጋሎት ክረምታዊ ማእቶት ማለት ኣብ ምምሃር ምስትምሃር፣ ኣብ ስግግር ቴክኖሎጂ፣ ቱቶሪያል፣ ግልጋሎት ምኽሪ ምሃብን ብምስታፍን ዕዉት ወፍሪ ምክያዶም ካብቲ ዝተዋህበ መግለፂ ምርዳእ ይከኣል። ኣብ ዝቐፅል ክረምቲ'ውን ተመሳሳሊ ወፍሪ ናፃ ግልጋሎት ትምህርቲ፣ ትምህርታዊ ምብፃሕ (academic visit)፣ ፓናል ምይይጥ፣ ኮንፈረንስ፣ ምርኢት ሳይንስ፣ ክረምታዊ ትምህርቲ፣ ቱቶሪያልን ስራሕቲ ምምኻርን ስልጠናን ከምዘካይድ ካብቲ መግለፂ ምርኣይ ይከኣል።

ብተወሳኺ ድማ ናፃ ናይ ትምህርቲ ዕድል (SCHOLARSHIP) ከመይ ከም ዝርከብ፣ ኣበይ ከም ዝርከብ፣ መኣዝ ከምዝርከብን ዝተፈላለዩ ብሮሸራት ከምዘዳለዉ ሓቢሮም'ዮም። "ተፅግበካያ ቕጫስ ኣብ መቐሎኣ እንተላ ትፍለጥ" ከም ዝበሃል ነዚ ሰናይ ተበግሶ'ዚ ኩሉ ምሁር ትግራዋይን ፈታዊ ትግራይን ብዕቱብ ክንቀሳቐስሉ ዝግባእን ክነኣድ ዝግበኦ ቅዱስ ሓሳብን ወፍርን ምኽኑ እንዳሓበርና ኣብ ሙሉእ ዓለም ዘለኹም ተጋሩ፣ ምሁራት ተጋሩን ፈተውቲ ህዝቢ ትግራይን ውዳበታትኩም ብምጥንኻር ንዓወት እዚ ወፍሪ እዙይን "ወፍሪ መለስ ንዕብየትን ዝላን" ካባኹም ዝድለ ዘበለ ኹሉ

ከተበርከቱን ኣባል እዚ ሳይንሳውን ኣካዳምያውን ኔትዎርከ ክትኮኑ እንዳተላቦና ብሸም ህዝቢ ትግራይን ብሄራዊ ክልላዊ መንግስቲ ትግራይን መፀዋዕታና እንዳቐረብና ኣብ ኩሎም ወፍርታትኩምን ምንቅስቓስኩምን ኣብ ጎንኹም ከምዝተሰለፍና ከነረጋግፀልኩም ንፈቱ።

“ዓወት ንወፍሪ መለስ ንዕብየትን ዝላን”

ቢሮ ርክብ ህዝብን መንግስትን ክልል ትግራይ

መቐለ

ድሕሪ <Aj> ሃንደበታዊ <Aj3MSg> መስዋእቲ <NCAsSg> ክቡር <Aj3MSg> መራሒና <NPs3MPl> ቀዳማይ <NuOrAjMSg> ሚኒስተር <NCSg> መለስ <NPo> ዜናዊ <NPo> ህዝቢ <NCMsSg> ትግራይ <NPo> መሪር <Aj3MSg> ሓዘኑ <NPsAsMSg> ብምግላፅ <PNvPp> ምስ <Pp> ካልኦት <Aj3Pl> ኣሓት <NCCtFPl> ህዝብታት <NCMsPl> ኢትዮጵያ <NPo> ኮይኑ <VAxMSgS> ዕላማታት <NCPl> መለስን <PNPoCj> ባዕሉ <PnPeGe> ዝጀመሮም <VMnIp3MSgIt> ናይ <Pp> ነዊሕ <Aj3MSg> እዋናት <NCPl> ዓበይቲ <AjPl> ፕሮጀክታት <NCPl> ንምዝዛምን <PNvPpCj> ኣብ <Pp> ሽዮኣም <NC3MPl> ንምብፃሕን <PNvPpCj> ካብኡ <Av3MSg> ዘድሊ <VMnIp3MSg> ዘበለ <VMnIp3MSg> ኹሉ <Aj3MSg> ከከም <Pp> ዓቕሙን <PNC3MSgCj> ክእለቱን <PNC3MSgCj> እጃሙ <NC3MSg> እናወፈየን <PVMnPf3MSgPpCj> ንቐፃሊ'ውን <PAj3MSgPpCj> ከምዘወፍን <PVMnIp3MSgIcPpCj> ናይ <Pp> መሪሕነት <NCAs>ተርኡ<NCPs3M>ከምዝፃወትን<PVMnIp3MSgItPpCj> ብህልኽን <PNCAsPpCj> ሕራነን <PNCAbCj> እናገለፀን <PVMnPf3MSgItPpCj> ቃል <NCSg> እናኣተወን <PVMnPf3MSgItPpCj> ይርከብ <VMnIp3MSg> ። <Pu>