ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCES

AFAAN OROMO MORPHOLOGICAL ANALYSIS: A HYBRID APPROACH

BY: KEDIR GENNA

ADDIS ABABA, ETHIOPIA

DECEMBER 2021

Addis Ababa University

አዲስ አበባ ዩኒቨርሲቲ

SEEK WISDOM, ELEVATE YOUR INTELLECT AND SERVE HUMANITY !

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCES

FOCUS LANGUAGE TECHNOLOGY

የኢንፎርሜሽን ሳይንስ ት/ቤት

School of Information Science

AFAAN OROMO MORPHOLOGICAL ANALYSIS: A HYBRID APPROACH

Kedir Genna

Advisor

Million Meshesha (PhD)

A THESIS SUBMITTED TO THE SCHOOL OF INFORMATION SCIENCES AND SYSTEMS IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS OF SCIENCE IN INFORMATION SCIENCES (LANGUAGE TECHNOLOGY)

Addis Ababa University

College of Natural Sciences

School of Information Science

Kedir Genna

Advisor: Million Meshesha (PhD)

This is to certify that the thesis prepared by Kedir Genna, titled: ***Afaan Oromo Morphological Analysis: A Hybrid Approach*** and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Information Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the examining committee:

Name	Signature	Date
Advisor: **Million Meshesha** (PhD)		
Examiner: **Michael Melese** (PhD)		
Examiner: **Solomon Teferra** (PhD)		

Abstract

This study provides relatively detailed information on developing Afaan Oromo morphological analysis system. Morphological analyzer decomposes words into its components called morphemes and annotates those morphemes with grammatical information. Although the module uses machine-learning approach on morphological analysis, it used rule-based approach to segments words into its small components, morphemes. The developed prototype focused on inflectional forms of nominals (nouns and adjectives) and verbs since the two words classes are mostly the ones that undergo inflection, they determine the inflectional characteristics of the language. The protype was developed using python programming and Hidden Markov Model (HMM). The Viterbi algorithm is used to encode the HMM model.

Then, the prototype was trained and tested using representative data. A corpus of size 4,320 nouns and 3,780 verbs are used to train the HMM model. Then the performance of the analyser was tested using 480 nouns and 420 verbs.

Generally, the accuracy of the analyzer for nouns and verbs is ***84.6 %*** and ***82.9%*** respectively. The result of the experiment was quite satisfactory, which can be improved by incorporating simple grammatical constraints and contextual information (including information encoded in tonal system) to minimize the ambiguities, words root database to reduce errors during morphemes identification and additional data to emphasis the initial probability of the model. The key limitations in this effort are limited funding opportunities, scarcity gold standard and balanced annotated data sets and inherently multiple sources of ambiguity of the language at different levels.

.

Key words: Afaan Oromo, Hidden Markov Model, Machine learning, Morphological Analysis, NLP

Dedication

I dedicate this work to my grandfather Hamda Gammada Burako.

Acknowledgment

First, I would like to thank my Almighty Allah for his countless love, care and help for me. ***Alhamdulillah***!

Then, I would like to express my deepest gratitude to my advisor Dr. Million Meshesha for his continuous follow up from the beginning to this end. His constructive comments, suggestions, guidance, and enlightening ideas were very essential. Besides this, I would like to appreciate his thoughtfulness, dedication, and willingness to pass his knowledge to his advisees. Without his sincere support, this thesis couldn't have reached this stage.

I would also like to thank Addis Ababa University PhD candidate Mr. Wakweya Olani for his linguistic expert advice on many aspects of my research.

I would like to express my gratefulness to all members of my family for their love, encouragement, as well as financial supports.

I wish to extend my acknowledgement to my friends Beshir Shaku, Desta Bati, Beyena Bedhasa, Afandi Abdi, Kemal Ebrahim, Bariso Beshir, Kedir Abdu, and those who encouraged me a lot to finish this research work.

I am also grateful to all the staff members of Addis Ababa University, especially, to staffs of school of Information Sciences for their patience and sincere services in all aspects.

Finally, I would like to thank my father Genna Hamda and my mother Dadala Turi. You both didn't get the opportunity to learn any formal education, but you gave me an opportunity. **Dad and Harmee Galatoomaa! Isin jaaladha.**

Table of Contents

List of Acronyms and Abbreviations vi
List of Tables viii
List of figures ix
CHAPTER 1 INTRODUCTION 1
1.1 Background of the study 1
1.2 Statement of the problem and Justifications 3
1.3 Research Questions 6
1.4 Objectives of the Study 7
1.5 Significance of the Study 7
1.6 Scope and Limitations of the Study 8
1.7 Research Methodology 8
1.7.1 Research Design 8
1.7.2 Research Methods 9
1.8 Organization of the Thesis 11
CHAPTER 2 LITERATURE REVIEWS AND RELATED WORKS 12
2.1 Overview 12
2.2 Basic Concepts and Terminologies in Morphology 12
2.2.1 Word 13
2.2.2 Morphemes and Allomorphs 13
2.2.3 Stems/roots (lemmas/base) 14
2.2.4 Affixes 14
2.2.5 Compounding 15
2.3 Natural Language Processing 15
2.4 Computational Morphology 17
2.5 Morphological Rules 19
2.6 Morphological Analysis 20
2.6.1 Approaches to Morphological Analysis 21
2.7 Machine Learning 22
2.7.1 Machine learning categories 23
2.8 Related Works 24
2.8.1 Language independent developments of morphological analysis 24
2.8.2 Morphological Analyzers for foreign languages 29
2.8.3 Morphological Analysis for Local Languages 37
2.9 Summary 43
CHAPTER 3 AFAAN OROMO MORPHOLOGY 45
3.1 Overview 45
3.2 Afaan Oromo Writing System 46
3.2.1 Syllables in Afaan Oromo 47
3.3 Afaan Oromo word classes (Parts of Speeches) 47
3.3.1 Open classes 49
3.3.2 Closed classes 50
3.4 Afaan Oromo Inflectional Morphology 51
3.4.1 Noun Inflections 52
3.4.2 Adjective Inflections 62
3.4.3 Verb Inflections 63

3.4.4 Adverb Inflections 69
3.5 Afaan Oromo morphotactic and Morphophonemic Properties 69
3.5.1 Assimilation 70
3.5.2 Deletion 71
3.5.3 Epenthesis 71
3.6 Summary 71
CHAPTER 4 DESIGN AND DEVELOPMENT 72
4.1 Overview 72
4.2 The Proposed Architecture of Afaan Oromo Morphological Analysis 72
4.2.1 Training Phase 73
4.2.2 Analysis Phase 75
4.3 Data Collection and preparation 76
4.3.1 Noisy Data 76
4.3.2 Stop words 76
4.3.3 Nominals and verbs selection 77
4.4 Algorithms Design for Word Segmentation: Afaan Oromo words segmenter 77
4.4.1 Rules for Afaan Oromo Segmentations 78
4.4.2 Testing Performance of the Segmenter 80
4.5 Training and Model Construction /Model Building for Morphological Analyser 81
4.5.1 Statistical Approach: HMM 81
4.5.2 Training Data 82
4.6 The Analyzer 84
4.6.1 HMM decoder: Viterbi Algorithm 84
4.7 Test Data Sets and Evaluation 84
4.8 Test results and performance of the Analyzer 85
4.9 Discussion of the Results 86
4.9.1 Sources of errors for incorrect Analysis 87
CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS 90
5.1 Conclusions 90
5.2 Recommendations and future works 92
REFERENCES 93
Appendixes 99
Appendix A: python script to remove Noisy Data from Afaan Oromo text 99
Appendix B: python Script to remove Stopwords from a text file 100
Appendix C: List of Afaan Oromo Stop Words 101
Appendix D: Sample Afaan Oromo words 104
Appendix E: Sample of Automatically annotated Afaan Oromo Words 105

List of Acronyms and Abbreviations

1PL	First Person Plural
1SG	First Person Singular
2PL	Second Person Plural
2SG	Second Person Singular
3PL	Third Person Plural
3SGF	Third Person Singular Feminine
3SGM	Third Person Singular Masculine
ABL	Ablative
ABS	Absolutive/Accusative
ACL	Association for Computational Linguistics
AI	Artificial Intelligence
BBC	British broadcasting corporations
CD-ROM	Compact disk Read-only memory
CV	Consonant-vowel (C=consonant, V= vowel)
DAFSA	Deterministic Acyclic Finite State Automata
DAT	Dative
DAWG	Directed Acyclic Word Graphs
DEF	Definitiveness
DEF/F	Definitiveness-Female
Dr.	Doctor
EMNLP	Empirical Methods in Natural Language Processing
F	Feminine (Female)
FST	Finite state transducer
GEN	Genitive
gmd	German Morphological Dictionary
HMM	Hidden Markov Models
HTB	Hindi Tree Bank
IB1	Instance Base
IGTREE	Information Gain Tree
IIIT-H	International Institute of Information Technology Hyderabad
ILP	Inductive Logic Programming
IMP	Imperative mood

IMPF	Imperfective
INSTR	Instrumental
JUSS	Jussive mood
LOC	Locative
LSTM	Long Short Term Memory
M	Masculine
MBL	Memory Based Learning
MLRS	Malta Language Resource Server
NAAL	National Association of Applied Linguistics
NERs	Named entity recognizers
NGD	Number-Gender-Definiteness
NL	Natural Language
NLP	Natural Language Processing
NN	Nominal i.e., noun& adjective
OOV	Out of Vocabulary
PBA	Paradigms Based Analyzers
PERF	Perfective
PL	Plural
POS	Part of Speech
RNN	Recurrent Neural Network
S-CRF	Semi-Markov Conditioned Random Fields
SES	Short Edit Script
SIL	Summary Institute of Linguistics
SMA	Statistical Morphological Analyzer
SUBJ	Subject form
SVM	Support Vector Machine
TAM	Tense, Aspect and Modality
TiMBL	Tilburg Memory Based Learner
URL	Uniform Resource Locator
Vs.	Versus
XML	Extensible Markup Language

List of Tables

Table 2.1: HTB Statistics

Table 3.1: Examples of nouns pluralized by oota, oolee, oolii, ilee (ota, olee, olii, ilee)

Table 3.2 Examples of nouns pluralized by -wwan, -lee

Table 3.3 Examples of nouns pluralized by -an, -een, -iin

Table 3.4 Examples of nouns pluralized by –eyyii

Table 3.5 Nominative/Subjective case makers

Table 3.6 Genitive case makers examples in Oromo nouns

Table 3.7 Examples of instrumental case makers in Afaan Oromo nouns

Table 3.8 Examples of Afaan Oromo Nouns' ablative case makers

Table 3.9 Examples of Afaan Oromo Nouns' locative case makers

Table 3.10 Regular autobenefactive verbs inflections for all persons and numbers.

Table 3.11 Active and passive verbs

Table 3.12 Verbs aspect inflection for Afaan Oromo

Table 3.13 Imperative mood for Afaan Oromo verbs

Table 3.14 Jussive mood for Afaan Oromo verbs

Table 3.15 Assimilations

Table 4.1 Statistics of Afaan Oromo collected and cleaned data

Table 4.2 Processes of nouns segmentation iteratively

Table 4.3 Lists of Afaan Oromo nouns and verbs morphemes-tag

Table 4.4 Evaluation results of Afaan Oromo Nouns and Verbs separately, on mixed and on Average

Table 4.5 Sample of incorrectly labeled because of fusional morphology

Table 4.6 Sample of incorrectly labeled because of statistical error

List of figures

Fig 4.1 General architecture of Afaan Oromo Morphological Analyzer

Fig 4.2 Transition Matrix of Afaan Oromo Nouns Morpheme Tags.

Fig 4.3 Training and Testing Environment of the Afaan Oromo Noun morphological Analyzer

CHAPTER 1 INTRODUCTION

1.1 Background of the study

Language is one of the fundamental aspects of human behavior, the chief manifestation of human intelligence and used for everyday communication by humans [1]. It is the gift that identifies human beings from the rest of life. Allen [2], stated that most of the human knowledge is recorded in linguistic form (i.e., in the form of natural language (NL) texts and utterances). Natural language serves as a means of recording information and knowledge on a long term-basis and transmitting what it records from one generation to the next generation using classical multimedia technologies. In its spoken form, it serves as a means of coordinating our day-to-day life with others through communication. Communications by human being is known as Natural language. The scientific study of languages particularly, natural languages are called Linguistics. Linguistics concerned with language and their structures. It studies the languages at different levels (i.e., phonology, morphology, syntax, and semantic level). An approach to linguistics that employs methods and techniques of computer science to manipulate Natural language is called Computational Linguistics or Natural Language Processing (NLP)[3].

NLP is a field at the intersection of computer science, artificial intelligence, and linguistics. Thus, it deals with how to program computers and use artificial intelligence approaches to process and/or understand large amounts of natural language data using linguistic knowledge. It is concerned with computational processing of natural languages in order to provide such novel products as computers that can understand everyday human speech, translate between different human languages, and otherwise interact linguistically with people in ways that suit people rather than computers[3], [4]. The goal of natural language processing is designing and building systems that will recognize, understand and generate natural languages [3]and can communicate with human using its artificial intelligence.

In every language, whether it is spoken or written, every meaningful pattern has its own structure and the elements of language should relate to each other in understandable manner. However, understanding these patterns is complicated task for machine as result of existence of relationship between large number of classes, ambiguity behavior of language at structural, semantic, and lexical levels and inflection and/or derivation of the language. These above

mentioned complications are also worse for Afaan Oromo, which is morphologically rich and resource scarce language and whose structure hasn't been studied extensively even in linguistic field of study.

Abebe [3] determined the different levels of natural language from lower level as Phonology, Morphology, Lexical, Syntactic, Semantic and Pragmatic. Among these linguistic levels, this study based on the morphological level, which is situated between phonology and syntactic [5]and study the internal structure of word [6], content of word forms and the rules for formation of grammatically right and acceptable words.

The smallest meaningful constituents which are used to from words are called morphemes. For example, if we take the word boys, we know that it is produced from {boy} and {-s}, each of which are separate morphemes and has semantic (boy) or grammatical (-s) information to add to the overall meaning of the word. However, the two separated morphemes must be combined to form the word /boys/. This process of forming words from one or more morphemes is called *word formation*. Morphemes can be either *bound morpheme* which cannot appear as word by itself or *free morpheme* which can appear as word by itself and can combine with another morpheme too. For example, in word boys {boy} is free morpheme that carry semantic content while {-s} is bound morpheme that carry grammatical information. The different variations of morphemes are called ***allomorphs***. Allomorphs have the same functions but different forms. However, like synonyms they cannot be replaced one by another. For instance, im, un, and ir in impossible, unhappy and irrational respectively indicates negative but they cannot replace each other.

It is obvious that like other systems NLP systems are developed in such a way that the output of a lower system can serve as an input to the next higher level. Thus, the output of morphological processing can be used by parser at syntax level, machine translators, spell checkers and grammar checkers.

The morphological processing can be classified into two separate tasks: morphological synthesis (generation) and morphological analysis. In Abebe [3] Morphological synthesis or morphological generation is defined as a process of returning one or more surface forms from a sequence of morpheme glosses. Morphological analysis, which is the main target of this proposal, is the process of breaking down words into its lexical components (morphemes) and tagging their grammatical features. Morphological analyzer has a vital role in NLP systems. It is used as a sub

component of NLP in applications like machine translation, dictionary (lexicon) development, and spelling and grammar checking, etc. Thus, it is the purpose of this study to explore the possibility of developing an automatic morphological analyzer useful for analyzing (parsing) Afaan Oromo words.

1.2 Statement of the problem and Justifications

Afaan Oromo is one of a major African language (Cushitic language) that is widely spoken in Ethiopia. Several varieties of Afaan Oromo are spoken in Ethiopia, Kenya, some parts of Somalia, Uganda, Tanzania and Djibouti [7], [8]. According to the 2007 census of Ethiopian population, Afaan Oromo is the mother tongue of about 33.8 percent of the country's population. Swedish encyclopedia, "The National Encyclopedia[1]" estimated number of Afaan Oromo native speakers to 24 million (0.36% of world populations). This made the language the **50**th language in the world by the number its native speakers.

Afaan Oromo is currently an official language of Oromia Regional State, medium of instruction at first and second cycles of elementary school level in Oromia regional state [9]–[11]. The language is also offered as a subject at secondary and preparatory levels and has been given as a field of study in university level [11], [12]. In addition to these, Oromo is a language of mass media and a working language in Oromia regional state and in the mass media at the federal level of the country, Ethiopia besides Amharic, which is the official federal language [8]. Language adopted Latin-based alphabet (Qubee) and use it officially for its writing system since 1991 [9], [13], [14].

Although Afaan Oromo is one of the major languages in Ethiopia, with the above described language status, no comparable research is done on NLP tasks like Morphological analysis. However, some authors conducted researches on Afaan Oromo Morphological Analysis but there were some limitations and unaddressed problems in their research.

Assefa W/Mariam [12] attempted to develop Morphological Analysis for Afaan Oromo Text. However, there some contradiction between his title, "Development of Morphological Analyzer for Afaan Oromoo Text" and the report of his work. In his report, he showed that, he followed an

[1] Retrieved from https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers on October 10,2021

unsupervised approach (the approach actually used for clustering problem), used the method was adopted from Goldsmith's [15], Linguistica[2], which was used for word segmentation in [12], [15]. This is to say that, the main tool he used, Linguistica is a tool for word segmentation (i.e., it is word decomposing tool), it cannot identify categories of word classes in the language. In unsupervised machine learning, labels of instances (i.e., morphemes in this case) are not known while in morphological analysis morpheme tags (morpheme labels) are known. This shows, this research has gap in the task it covered (i.e., absence morphological analysis) and the approach it used (unsupervised which needs huge amount of data for training to perform accurately). Therefore, our proposed work attempted to solve this problem by using hybrid approach; rule based approach to identify the boundary of morphemes (i.e., segmentations) and supervised machine learning approach to identify categories of morpheme classes (morphological analysis) in the language.

Michael Gasser [16], the author of Hornmorpho, developed morphological synthesizer and analyzer for three main Ethiopian languages; Amharic, Tigrigna and Afaan Oromo using Finite state Transducer (FST). The Afaan Oromo module of the system can handle nouns and verbs. However, the performance of Afaan Oromo module was not yet evaluated, because it is complicated by the great variation in the use of double consonants and vowels by Afaan Oromo writers, compilation of lexicon form two inconsistent dictionaries and less knowledge of Afaan Oromo morphology. Additionally, in HORNMORPHO 2.5 User's Guide [10], the author reported that the performance of the program for Oromo is inferior to that for the other languages with unknown performance in percentage. Even if it wasn't tested, it can be considered as the first attempt towards Afaan Oromo words morphological analysis in contrast to that of Assefa [12] which is limited to segmentation task. Thus, the current work proposed based on hybrid approach to develop and test Afaan Oromo nouns and verbs morphological analyzer.

Another most related and the latest effort towards Afaan Oromo morphological analysis is, the work of Moyka and Dida [17], which is proposed based on supervised machine learning approach, specifically, Memory Based Learning (MBL). He selected MBL because of its "lazy" property as lazy machine-learning methods achieved a higher accuracy than eager methods for many NLP tasks.

[2] Linguistica is a program which can be used to explore the unsupervised learning of natural language, with primarily, focus on morphology, which is to say, word-structure.

However, most of lazy learning methods including KNN which is used by Moyka and Dida [17] have the following disadvantages [18]**:**

- Lazy learners incur ***expensive computational*** costs when training data is large.
- Even if it the methods are faster in training phase, Lazy learners are ***slower*** during classification.
- Nearest neighbor classifiers assign ***equal weight*** to each attribute. This may cause confusion when there are many irrelevant attributes in the data and results into ***poor accuracy***.

  HMM: Assigns co-occurrence probability to attributes to solve this problem (see section 4.2.1 transition probabilities and emission probabilities).
- Sensitive to the local structure of the data; It doesn't not use global structure of the data.
- All lazy learning algorithms calculate the distance between instances based on all attributes; if there are many irrelevant attributes, instances that belong together may still be distant from one another. This may cause wrong classification. However eager learning methods build a model using the whole training data set and then this model is used to classify all the new query instances. The built model is completely independent of the new query instances; i.e., there no distance measuring in eager learners algorithms.
- They need ***large storage*** to save training instances since training examples are stored without modification or abstraction [19].
- They lack a principled way to choose k (numbers of nearest neighbors).

Thus, we proposed a new architecture for Afaan Oromo morphological analysis based on eager learning methods, rule-based and statistical. Rule-based approach is selected because of its accuracy and no data needed to develop rules. Statistical classifiers have exhibited high accuracy and speed, specially, when applied to large databases [18]. We used a hybrid Approach for Afaan Oromo morphological analysis. Rule-based approach was used for word segmentation while HMM was used for morph tagging/labeling.

Wakweya Olani [8] also wrote his master's thesis on linguistic aspects of Afaan Oromo Morphological Analysis (specifically, inflectional), which can be used by computational linguistic researchers to understand the morphological features of Afaan Oromo. But computational and engineering aspects of the language's morphology weren't included in his research.

There were also some articles towards some sub-stages of morphological analysis of Afaan Oromo, like stemmer by Debela and Ermias [20] and Afaan Oromo root generation system by Getachew Mamo [21]. As these works are sub phases (part of pre-processing) for morphological analysis they could contribute to our morphological analysis work to some extent. For instance, we adopted some basic concepts from these previous works to develop rules for Afaan Oromo Words segmentation.

Abebe Abeshu [22] also published on article on Afaan Oromo Automatic Morphological Synthesizer, which works in opposite direction of morphological analyzer (i.e., synthesizer generates word form from its constituents and grammatical features while analyzer breaks down word form into its morphemes and assign grammatical features).

All the above articles, related with Afaan Oromo morphological analysis, except [17], used rule-based approaches, which needs hand-crafted rules. The rules must be developed by linguistics, so it is resource consuming (i.e., human resource and time resource). The rules by themselves may contradict each other if one writes many rules.

This proposed research is concerned with Morphological Analysis using a Hybrid Approach for Afaan Oromo words specially nouns and verbs, as they are highly productive in terms of morphology. The study is proposed to tackle the challenges and limitations faced by these authors and to narrow the gaps in the development of Afaan Oromo Morphological Analysis to some extent.

1.3 Research Questions

Motivated by the above statement of problems, this study attempted to answer the following basic research questions.

- ⇒ What are the occurrence patterns (sequences or order) of morphemes in Afaan Oromo Words?
- ⇒ What are the main inflectional morphemes of Afaan Oromo words?
- ⇒ How to identify the morphemes of Afaan Oromo words for further analysis?
- ⇒ How Afaan Oromo word is analyzed morphological using machine learning algorithms?
- ⇒ What is the performance of the proposed morphological analysis prototype?

1.4 Objectives of the Study

1.4.1 General Objective

The main purpose of this study is to develop and test an automatic Morphological Analyser for Afaan Oromo words specially; Nouns, adjectives and verbs using a Hybrid Approach.

1.4.2 Specific Objectives

To achieve the above general objective and answer research questions we proposed the following sub-goals.

i. To study the morphological behavior of Afaan Oromo through literature and observation of Afaan Oromo words nature.
ii. To study the morphological analysis approaches and select the one that can fit for Afaan Oromo morphological analysis.
iii. To collect, preprocess and prepare data for training and testing models
iv. To design Afaan Oromo morphological analyzer
v. To develop Afaan Oromo morphological segmentation algorithm.
vi. To train the morphological analysis model with labeled data
vii. To evaluate the performance of the morphological segmenter and analyzer
viii. To report the findings of the study for the upcoming research area.

1.5 Significance of the Study

Morphology is the study of word formation. Word is the smallest meaning bearing unit in language. Thus, morphological processing can be recognized as the basic component of NLP. In addition, morphology takes a significant share in standardizing the grammar of a language. Generally, the results of the study are significant for linguists, teachers, and technical personnel (researchers).

For linguists, language learners, and teachers: the detailed study of the Afaan Oromo morphology would contribute to the standardization of the language. The research outcome could

be useful for preparing materials of pedagogical purpose, specifically for teaching and learning Afaan Oromo Morphology.

For NLP researchers and technical persons: Morphological analyzer have vital role in NLP systems. It is used as a sub component of NLP in applications like machine translation, dictionary (lexicon) development, and spelling and grammar checking, etc. Therefore, it simplifies development of these applications for Afaan Oromo.
It would also help those who are interested to conduct further research either on the topic or on other related issues.

1.6 Scope and Limitations of the Study

This work deals only with the syntactic aspects of Afaan Oromo morphology. It does not deal with the derivational analysis (i.e., semantic aspect) of the language except for their inflected forms. The morphological analyzer was developed for the two main categories of words in the language; nouns and verbs. The supervised machine learning approach, specifically statistical approach (HMM-hidden Markov model) is used to develop the analyser. To train and test the model, Afaan Oromo nouns and verbs are selected from raw data collected from Fana Broadcasting Corporation Afaan Oromo and BBC Afaan Oromo online archives from April 2020 to September 2020, after further data cleaning.
This research would have been exhaustive and meaningful if all word categories and aspects of morphology (including compounding, and derivational morphology) are studied. Due to time and financial constraints and absence of ready-made data for this purpose, an extensive investigation was not made.

1.7 Research Methodology

1.7.1 Research Design

In Kothari[23], research design refers to the overall strategy that we choose to integrate the different components of the study in a coherent and logical way and it is the conceptual structure within which research is conducted. In fact, it constitutes the blueprint for the collection, measurement, and analysis of data.

The selection of an appropriate research design depending research problem is crucial in enabling you to arrive at valid findings, comparisons, and conclusions. Thus, when selecting a research design, it is important to ensure that it is valid, workable, and manageable.

Therefore depending on the nature of the investigation we conducted we selected Experimental research design. Experimental research (generally, Qualitative Research) are specific, well structured, have been tested for their validity and reliability, and can be explicitly defined and recognized[24].

In this study we have causes and effects and the degrees of causality is measured by co-occurrences probability.

Causes: Observations (morphemes, previous class of the predicted class)

Effects: Hidden-state (The predicted class)

1.7.2 Research Methods

Research Methods are the tools and techniques for doing research. Research methods are a range of tools that are used for different types of enquiry, just as a variety of tools are used for doing different practical jobs[25]

In this study, a number of methods (techniques) and algorithms are used for the successful completion of the study. Some of them are discussed below while other discussed in other parts thoroughly.

1.7.2.1 Literature Reviews and Discussion with Experts

Different research literatures that are considered relevant for my research from different perspectives are reviewed and adopted for my work in addition to resources like books, journal articles and other published or unpublished documents from the Internet for the purpose of understanding the behavior of human language morphology and morphology of Afaan Oromo words.

Related works in computational morphology are reviewed to identify different approaches being tested in development of morphological analysis systems; to examine and select appropriate machine learning algorithm; and to know how to develop corpus data for morphological analysis research work.

Other literatures regarding Afaan Oromo computational morphology works were reviewed to tackle the obstacles faced by those researchers and identify the problem addressed.

As we specialized in computer technology (language technology) we did not take any course related to the linguistic particularly, Afaan Oromo. However, without detail knowledge of the language, Computational knowledge is not enough to develop Afaan Oromo Morphological Analysis system. Therefore, linguists and experts in the area of Afaan Oromo were consulted.

1.7.2.2 Data Collection and Preparation

Afaan Oromo does not have publicly available annotated corpus text for the purpose of morphological analysis. The former researchers, who have done some experiments on Afaan Oromo NLP, reported that they collected Afaan Oromo text data from different documents (such as newspaper, textbooks, research papers, etc.). Machine learning approach requires large corpus data. Thus, the data was manually collected raw data from Fana Broadcasting Corporation Afaan Oromo and BBC Afaan Oromo archives from April 2020 to September 2020. Because; they are considered as addressing different issues of the community such as social, economic, technological, political, religions etc. This reduces the probability of making the corpus biased toward some specific words and domains that do not appear in everyday life[20]. After data is collected, different data preprocessing techniques (cleaning and organizing) applied on the raw data to make it suitable for a building and training Machine Learning models. In addition, it is used for model evaluation (model testing). Data preprocessing techniques of the data includes:

1. Noisy data (such as digits, special characters and punctuation) removal
2. Stopwords removals and
3. Selection of names and verbs (actually this done manually in our case to grant the selection accuracy)

Noisy data removal and stopwords removal was done automatically by using simple python scripts. Obviously, morphological analysis development needs morphologically annotated corpus (nouns and verbs). Therefore, the collected words segmented into morphemes and their grammatical features were annotated to the respective morphemes manually.

1.7.2.3 Development Tools and Techniques

To design and develop Afaan Oromo Morphological Analysis, we used python programming language. Because python is a vast language with number of modules, packages and libraries that provide multiple ways of achieving machine learning tasks. It enables us to create scalable machine learning algorithms, implement popular machine learning techniques. It is also statistically easy.

The only publically available system for Afaan Oromo Morphological Analysis is Hornmorpho [16]. However, Hornmorpho was developed using FST. For this reason, we cannot use any components of this system and another prototype system was developed to handle the morphological analysis tasks. Hybrid approach of eager learning methods was used to develop the proposed prototype system.

1.7.2.4 Evaluation

After the prototype is developed, it was tested on testing corpus using error counting technique to evaluate the performance of this analyzer. The result is represented in quantitative measure, the percentage of correctly stemmed words. The figure obtained is used to evaluate the accuracy of the analyzer.

1.8 Organization of the Thesis

The left part of this thesis is organized in five chapters. Chapter two covers literature review and related works. This chapter will try to address the basic concepts of morphology, NLP and Machine Learning along with the various approaches employed. In this chapter, we also discussed related works that are highly relevant and more related to our study as well as the state of the art of morphological analyzer and computational morphology. Linguistic structures of Afaan Oromo nominals and verbs morphology are discussed in Chapter three. Chapter four presents the general architecture of the system, along with the discussion of its components and experimentation of the system, including its results. Finally, the conclusions followed by recommendations are presented in Chapter five.

CHAPTER 2 LITERATURE REVIEWS AND RELATED WORKS

2.1 Overview

In this chapter we present review and analysis of different published works and concepts related to my main work; designing automatic Afaan Oromo text morphological analysis using supervised machine learning as well as morphological analysis development strategies and approaches. Computational and linguistic aspects of Morphological processing is also discussed to understand the techniques, methods and principles of Morphological processing of natural language (NL) especially Morphological analysis. This chapter of the paper is organized as follows.

Section 2.2 discusses the fundamental concepts in morphology such as components of morphology, terminologies related to morphology and etc. In section, 2.3 and 2.4 Natural language Processing and computational Morphology are discussed respectively. The rules that govern morphology are dealt with in the next section. Section 2.6 discusses the details of Morphological analysis, the approaches to morphological analysis those have close relevance to this study. The next section, discusses the Machine learning and its categories in detail. Section 2.8 discusses the related works that are highly relevant and more related to our study. The summary of the chapter is presented in the last section.

2.2 Basic Concepts and Terminologies in Morphology

It very important to provide almost the common functional meaning of some basic concepts and terms those frequently used in morphology to avoid ambiguity and make clarity of the technical/functional use terms. Because natural languages have their own terms or definitions to describe terminologies based on the morphological behavior of that specific language.

The term **morphology** is generally attributed to the German poet, novelist, playwright, and philosopher Johann Wolfgang von Goethe (1749–1832), who coined it early in the nineteenth century in a biological context. Its etymology is Greek "***morph***" which means "***shape, form***" [26], and morphology is the study of form or forms.

In linguistics morphology refers to the mental system involved in word formation or to the branch of linguistics that deals with **words**, their internal structure, and how they are formed from small constituents called **Morphemes** [3].

2.2.1 Word

Words are the fundamental building blocks of a language [3]. Every human language, spoken, signed, or written, is composed of words. It is a concrete word as it occurs in real speech or text. For computational purposes, a word is a string of finite characters, separated by spaces in writing for most of the languages [27].

2.2.2 Morphemes and Allomorphs

Word is produced from a limited collection of smaller units called **morphemes** [3], [4], [26], [28], [29]. Morphemes are defined as the smallest units in a language to which a meaning may be assigned [28] or, alternatively, as the minimal unit of grammatical analysis. They are the smallest meaning bearing units of words [30]. The form of a morpheme may be **free** or **bound** [19].

Free morphemes occur relatively freely (appear as word by itself) within other words or morphemes, e.g., play. We call such words *monomorphemic* because they consist of a single morph. Free morpheme can also combine with other morpheme to form words, e.g., plays. Free morphemes themselves carry lexical meaning of the word; thus sometimes-free morphemes are called **lexical morphemes**.

Bound morphemes, on the other hand, occur only in combination with other forms (cannot appear as word by itself). Their role is to modify the meaning of a lexical morpheme or specify the relationships between lexical morphemes. They are **grammatical morphemes**.

The different variation of morphemes is called ***allomorphs*** [28]. Allomorphs have the same functions but different forms. For instance, im, un, and ir in impossible, unhappy and irrational respectively indicates negative.

Grammatical/bound morphemes also may be classified as **Derivational** and **Inflectional** morphemes. Derivational morphemes change the semantic meaning of a word, they can also change the POS of the word while Inflectional morphemes change the number, gender, case etc. of nouns, adjectives etc. and the person, number, mood, tense etc. of verbs and so on.

Generally, morphemes in language are composed of **stems** and **affixes**.

2.2.3 Stems/roots (lemmas/base)

Roots carry the basic indivisible meaning of a word while a stem is the base of an inflected word. For examples in the word availabilities which can be segmented into {avail +able +ity +es}; {avail} is root because it cannot be divided further; {Available} is a stem and {Availability} is a stem of inflected word {availabilities} while {Availabilities} is an inflected word from stem {availability}

Lemma/lexeme is distinguished form from a set of morphologically related forms, chosen by convention (e.g., nominative singular for nouns, infinitive for verbs) to represent that set. Lemma can be also called the **canonical**/**base**/**dictionary**/**citation** form (word). For every form, there is a corresponding lemma. For example, in English, steal, stole, steals, stealing are forms of the same lexeme steal. The set of word-forms that belong to a single lexeme is called **Paradigm**. For instance, in the above example {steal, stole, steals, stealing} is paradigm while steal is their lexeme.

2.2.4 Affixes

Affixes are bound/grammatical morphemes which attached to base (root or stem;[3]) for the purpose of word modification and relationships specification between words. Affixes can be divided according to their attachment location:

- **Suffixes** are attached to the end of the word:
    - example: - Avail+**able**, table+**s**, go+**ing**; bold letter(s) indicates the affixes
- **Prefixes** are attached to the beginning of the word:
    - **mis**+understand **re+**employ, **en**+danger, **in**+accessible, words with are examples of prefixes.
- **Circumfixes** have two parts, one attaches in the beginning of the word and the other to the end. It is the combination of prefix and suffix that together express some feature [3].
    - for instance **un**+bliev+**able**, **em**+bold+**en**

- **Infixes** are attached in the middle of the word:
    - for instance in German: an+**zu**+fangen , ab+**ge**+fahren
- **Reduplication** is used to show an action done repeatedly and plural form adjective of in Afaan Oromo.
    - For instance, the verb cabse ‘he broke’ becomes **cac**cabse ‘He broke something into pieces’, diimaa ‘red (for masculine)’ becomes **did**diimaa ‘red ones.

As there are no prefixes, infixes, and circumfixes in Afaan Oromo, we do not deal with it any more.

2.2.5 Compounding

Compounding is using two or more words together to form a new meaning. In most of the languages, compounds words can be written together or separately. For instance: - football, chairperson, manmade.

2.3 Natural Language Processing

Language is one of the fundamental behavior of human intelligence [1]. It enables to transfer idea between people and knowledge from generation to generation orally and/or in recorded form [2]. Such a communication by human being is known as Natural Language. Natural languages scientifically studied under a field of study called linguistics. Linguistics is the science, which involves meaning of language, language context and various forms of the language. Natural languages may be studied from different perspectives such as cognitive, computational, and engineering depending on its goal. However, in this paper we are interested in computational and engineering perspectives of linguistic. These are computational linguistic and natural language processing. In many literatures, people are using computational linguistic and natural language processing interchangeably as they are synonyms. The methodologies are often related and the communities overlap. They go to the same conference, many the strongest works in both fields appear at ACL, EMNLP, NAAL, etc., and they easily discuss together about their problems and solutions. Most of the published works on this issues reported that computational linguistics and natural language processing are different in their goals and

perspectives. As defined in WIKIPEDIA, “0computational linguistic is an interdisciplinary field concerned with the ***statistical*** or ***rule-based*** modeling of natural language from a ***computational perspective***, as well as the study of appropriate ***computational approaches*** to linguistic questions.” while “Natural language processing is a subfield of computer science, information engineering and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to ***program computers*** to ***process*** and ***analyses*** large amounts of natural language data.” The association for computational linguistic also defines computational linguistic as the scientific study of language from computational perspectives. However, NLP perceive linguistic problems from engineering perspectives.

Computational linguistic interested in providing computational models of various kinds of linguistic phenomena. It focuses on issues in theoretical linguistic and cognitive science and practical outcomes of modeling human language. Its goal is to build computational models of natural language for its analysis and generation. Natural language processing (NLP) develops methods for solving practical problems involving language such as automatic speech recognition, language analysis, language understanding, language generation, machine translation, and information extraction from documents. Even if there is intersections between their components (i.e., theoretical and/or application component), they are different. However, this study is interested in both computational linguistic and natural language processing, mainly NLP.

Natural language processing (NLP) is an interdisciplinary area based on many fields of study [31] such as computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, which is used for designing and building software that can analyze, understand, and generate natural language [7]. It is a computational method that automates the translation process between computer and human languages. It enables the users to interact with computer systems as such as database management systems and expert systems using natural languages. It can also be defined as an area of active research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things [31]. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks.

In recent years, the natural language text interpretation and processing technologies have also gained an increasing level of sophistication.

NLP technologies are becoming extremely important in the creation of user-friendly decision-support systems for everyday non-expert users, particularly in the areas of knowledge acquisition, information retrieval, and language translation.

NLP is widely used in modern computer systems. It is concerned with the interactions between computers and human languages. As such, NLP is related to the area of human-computer-interaction.

There are more practical goals for NLP; many of which are related to the particular application for which it is being utilized. For example, technological motivation of building intelligent computer systems such as machine translation systems, speech understanding systems, text analysis and understanding systems, computer aided instruction systems, and a cognitive and linguistic motivation to gain a better insight into how humans communicate using natural language.

Generally, the field of NLP includes a wide variety of linguistic theories, cognitive models, and engineering approaches. Computer science is one of these fields, which provides techniques for model representation, and algorithm design and implementation.

The advances in computer science, which have made NLP possible today, are bigger and faster computers, the development of major software systems, database technology and high-level languages and operating systems.

2.4 Computational Morphology

Natural languages have intricate systems to create words and word forms from smaller units in a systematic way. The part of linguistics dealing with these phenomena is called morphology.

Morphology is the study of the internal structure of words, which are the basic building blocks of sentences[6], content of word forms and the rules for formation of grammatically right and acceptable words. It deals with the systematic correspondence between the form and meaning of words. The study of these regularities comprises the domains of inflection and derivation. Inflection concerns the expression of morphosyntactic properties, sometimes required by a

specific syntactic context. Derivation deals with the creation of new (complex) words by various morphological mechanisms such as compounding, affixation, truncation, and segmental and tonal alternations.

Morphology is also very relevant for linguistic typology (morphological classification of languages). Thus, languages may be classified according to the role and nature of their morphology [32], [33] as analytic, poly-synthetic, agglutinative language and etc.

Analytic (isolating) languages are a kind of language that does not make use of morphology. Words in analytical languages are invariable. They are composed of free morphemes and there are no morphemes to indicate grammatical information like number, gender, tense etc.

Synthetic languages are languages those has a lot of morphology. Some synthetic languages those allow the incorporation of lexical morphemes, leading to relatively complex words are called *polysynthetic* languages.

In *Fusional (Inflecting)* languages one morpheme may express more than one grammatical feature; while each bound morpheme corresponds with one grammatical feature in *Agglutinative languages*.

However, it is impossible to fit any of the languages into any of the classes, because each language is impure. That is to say, if you look hard enough, you will find inflection in mainly agglutinative languages, synthetic/inflectional in isolating languages, agglutination in inflectional languages and so on. For instance, many Indo-European languages are fusional in their inflectional system, but agglutinating in their derivational morphology [34]. Chinese, which is often considered as isolating language, also illustrates this point since it is synthetic as far as word formation (*compounding*) is concerned, but isolating as far as inflection is concerned, as it has no inflection.

Computational morphology is the study of computational analysis, synthesis, and treatment of word forms in both their graphemic (i.e., written form) and their phonemic (i.e., spoken form) for eventual use in natural language processing (NLP) applications [3]. It is intended to handle the task of morphology automatically with the use of computers and computational methods [3], [35], [36]. The major tasks of computational morphology are word surface analysis and generation (synthesis). As defined in Michael Gasser [37] Morphological analysis is the

segmentation of words into their component morphemes and the assignment of grammatical features to morphemes while morphological generation is the reverse process of morphological analysis. The function of morphological analyzer is to return all the morphemes and their grammatical categories associated with a particular word form. For a given root (stem for some languages) word and grammatical information, morphological generator will generate the particular word form of that word [38].

Computational morphological methods also give linguists the ability to create grammars and specify how word forms should be stored in lexicons. A number of systems have been developed with a wide variety of approaches to processing, for use in NLP systems including natural language generation, machine translation, information extraction and retrieval using natural language, text to speech synthesis, automatic written text recognition, grammar checking, and part-of-speech tagging.

2.5 Morphological Rules

Natural languages make use of a number of formal means for the formation of complex lexemes: compounding, affixation, reduplication, conversion, alternation, stress, and tone. The morphemes of a word cannot occur in random order. In every language, there are well-defined ways to sequence the morphemes. The morphemes can be divided into a number of classes and the morpheme sequences are normally defined in terms of the sequence of classes. These phenomena of language is governed by linguistic rules, particularly, *morphological rules*.

Morphological rules ensure the validity of the word through the morphological constraints. Based on their characteristics and applications morphological rules classified as ***Morphophonemics*** and ***Morphotactics***. Morphophonemics are the rules that apply to the phoneme substitutions. Sometimes they are called alternation rule, whereas Morphotactics are the rules that ensures to the validity of the morpheme sequences. It governs the order in which morphemes follow each other. In simple word it is the syntax of morphemes [39]. Morphotactics also serve to disambiguate the morphemes that occur in more than one class of morphemes. The analyzer uses these rules to identify the structure of words [40]

2.6 Morphological Analysis

Morphology is the study of internal structure and formation of words within a given language. It is the method of analyzing internal structure of words and how they are constructed by combining different morphemes and generating by combining different morphemes. Developing full-fledged morphological processing tools for highly synthetic languages (languages with a lot of morphology) is challenging tasks even for senior researchers [38] because of languages morphology richness and complexity.

Morphological analysis is the process of segmenting words into morphemes and analyzing the word formation. It is the process of [41]:-

⇒ Analyzing complex words, identifying their component parts (morphemes):

⇒ e.g., anti+dis+establish+ment+arian+ism

⇒ and then further Analysis of grammatical information those morphemes:

1. e.g., sings => sing [PERSON 3, NUMBER singular, TENSE present]

Shortly; the function of morphological analyzer is to return all the morphemes (i.e., word segmentation) and their grammatical categories (i.e., classification of these morphemes according to function) associated with a particular word form.

Morphological analysis is often an initial step for various types of text processing of natural languages [38], [41]. It is used as components in many NLP applications, including machine translation, spell, and grammar checker, speech recognizer, speech synthesizer, dictionary (lexicon) compilation, POS tagging, conversational systems, automatic sentence construction, information extraction, information retrieval and many others.

The Algorithm of Morphological Analysis designed for the morphological analyzing process takes input as word and returns all the possible morphological structures and definition of that word. Thus, the output is used as input for the downstream stages of NLP processing as its importance.

Morphological analysis plays a role in reduction of lexicon size, unknown word recognition using machine learned models, etc. Therefore, it avoids the storage of exhaustive lexicons and saves memory requirement. Modeling morphological analysis also helps to reduce the

vocabulary sparsity problems in the case of information retrieval and other vocabulary dependent NLP systems.

The literatures show that development of morphological analysis has been successfully done for languages like English, Chinese, Arabic and European languages using various approaches from last few years and there are a very few numbers of attempts for Ethiopian languages and still these are an ongoing process. In the next section, we present some of those approaches used in development of Morphological analyzers.

2.6.1 Approaches to Morphological Analysis

In general, there are several approaches attempted for developing morphological analyzer. A great revolution in the area of morphology started to appear in 1983 when Kimmo Koskenniemi, a Finnish computer scientist, developed a two-level morphology approach, where he tested this formalism for Finnish language [3], [42]. In this two level representation, the surface level is to describe word form as they occur in written text and the lexical level is to encode lexical units such as stem and suffixes. Even though there are many approaches for developing morphological analysis, they can be classified into two main categories based on the source of their knowledge [3], [42]: Rule-based approaches and Corpus-based approaches.

Rule based approaches: Rule-based approaches use handcrafted rules developed by incorporating sophisticated linguistic theories as its knowledge source. The rules also may be created automatically by computer programs to contain a large number of morphological, lexical and/or syntactical information [3].

Rule based approaches have their own limitations: requires linguistic experts as it use hand-written written rules; Adding a rule to the system may involve over-generation so it may harm the accuracy of the system[3]; costly and time consuming.

According to Kibur [42] and Abebe [3]morphological systems developed using rule-based approaches have the following advantages over those developed using corpus-based approaches:

i. It requires less storage (Data-compaction)

ii. High speed

iii. Better Effectiveness (better accuracy)

Corpus-based approaches: corpus-approaches use structured or unstructured data as its knowledge base, i.e., the approaches are completely based on training and testing corpora. It is called machine-learning approaches; do not strictly follow explicit theory of linguistics. Corpus-based approaches use some algorithms to learn from data (word form in our case), that is why we call it Machine learning.

Developing morphological analyzers using machine-learning approaches have the following advantages over developing using rule-based approaches:

i. **Saving human resources**: Once the system is trained, classification is done automatically with no or little human intervention

ii. **Consistent classification**: The classification is done consistently on repetitions.

iii. **Automatic rule formation**: Human resources are not needed to make rules

As we use this approach in this work, we deal with it furthermore to not loose these advantages.

2.7 Machine Learning

Machine Learning is the study of building computer systems that learn from experience. It concerned with the question of how to construct computer programs that automatically improve with experience.

" A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." [43].

Machine learning usually refers to the *changes* in systems that perform tasks associated with artificial intelligence (AI). Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc. The term change indicates the enhancement/improvement in the system because of experience.

Machine learning system may also be developed as new using some machine learning algorithms (algorithms learn from data).

2.7.1 Machine learning categories

Based on the types of training corpora they use, machine-learning approaches are divided into ***supervised*** and ***unsupervised*** main categories. In simple word the main difference between both approaches resides in the way we feed training data (examples) to our algorithm, how the algorithm uses them and the type of problems they solve.

Supervised machine learning approaches: In the case of supervised learning, the Machine Learning algorithm can be seen as a process that has to transform a particular input to a desired output. It is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. The typical use case in supervised machine learning to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.

Depending on the type of output, we have two sub-types of supervised learning: ***classification*** (When the output value belongs to a discrete and finite set) and ***regression*** (When the output value is a continuous number).

Supervised learning is the most popular category of Machine Learning algorithms. The disadvantage of using this approach is that for every training example, we have to provide the correct output, and in many cases, this is quite expensive (require a group of humans to read and tag each tweet which quite a time consuming and boring task). Collection quality tagged training data is usually a very common bottleneck for Machine Learning algorithms.

Unsupervised machine learning approaches: use heuristics or probability information generated from the training corpora to generate the machine learning model. In this approach, no sample outputs are given. Thus, the training examples only need to be the input to the algorithm, but not the desired output. The typical use case is to discover the hidden structure and relations between the training examples. A typical example of unsupervised machine learning is clustering algorithms, where we learn to find similar instances or groups of instances (clusters). According to [43] this approach reduces the cost of browsing annotated corpora.

2.8 Related Works

Computational morphology can be viewed as having three separate subtasks segmentation, clustering related words (i.e., based on stems also called Paradigm based), and labeling (tagging)[44]. Various approaches have been used for each of the tasks, ranging from rule-based [20] techniques, such as finite state transducers[45] to various unsupervised[6], [44], [46], semi-supervised [47] or supervised [41], [48]–[50] techniques which would generally deal with one or two of the subtasks. Morphological analysis, sometimes considered as labeling task, retrieves the grammatical features and properties of a morphologically inflected word[51]. For most of the techniques described, it is difficult to directly compare results due to difference in the data used and the evaluation setting itself. For instance, the results achieved by segmentation techniques are then evaluated in an information retrieval task.

This chapter will present approaches, models, tools, and techniques used to develop morphological analyser of different languages and implementations of morphological analyzers in previous works of different researchers.

After Koskenniemi [45], a Finnish computer scientist developed a general computational model for word-form recognition and generation called *Two-level morphology*; many NLP researchers and research groups have developed various language dependent and language independent models, algorithms and tools for morphological analysis using different approaches.

The following subsections will present some attempts and efforts of those researchers towards different foreign and Ethiopian local languages for nearly four decades starting from Koskenniemi [45], specially the recent ones.

2.8.1 Language independent developments of morphological analysis

In this section, we present the efforts towards language independent morphological analysis of natural languages.

2.8.1.1 Two-level Morphology

Since 1960s, linguists have used the formalism generative of phonology for inflectional morphology and morphophonology. However, this kind of formalism (formalism generative) has limited power in the aspect of descriptive. Thus, two-level model incorporates alternative formalism. Two-level morphological model was one of the major breakthroughs in the field of

morphological parsing, which is based on morphotactics and morphophonemics concepts. The advantage of two-level morphology is that the model does not depend on a rule compiler, composition or any other finite-state algorithm. The two-level morphological approach consists of two levels called *lexical* and *surface form* and a word is represented as a direct, letter-for-letter correspondence between these forms. Thus, it solves the problem of intermediate levels between the surface and lexical level. Morphological two-level model consists of two main components a Lexicon system and two-level rule. The Two- level morphology approach is based on the following three ideas:

- ⇒ Rules are symbol-to-symbol constraints that are applied in *parallel*, not sequentially like rewrite rules (rules that applied one after another in predetermined order).
- ⇒ The *constraints* can refer to the lexical context, to the surface context, or to both contexts at the same time.
- ⇒ Lexical lookup and morphological analysis are performed in tandem.

The model is developed and validated using ***Finish*** inflectional morphology.

2.8.1.2 Morfette

Traditionally, in morphologically rich inflectional or agglutinative languages, the morphological analysis is often constrained by the use of a morphological lexicon, or a finite-state morphological analyzer. The main problem of such these approaches is inability to predict the tags of unknown words. Apart from the traditional rule-based approaches, Morfette [52] is a modular, data-driven, probabilistic system which learns to perform joint morphological tagging and lemmatization from morphologically annotated corpora using Maximum Entropy classifiers. The system is composed of ***two learning modules***, one for morphological tagging and one for lemmatization, and ***one decoding module***, which search for the best sequence of pairs of morphological tags and lemmas for an input sequence of word forms. The objective of the researchers is to use generic, minimalistic, language-independent feature sets and investigate how well such an approach generalizes to three morphologically rich languages: Romanian, Spanish, and Polish.

Maximum entropy models trained on examples to predict probability distributions over classes (i.e. morpho-tags or lemma-classes) for the current focus word form given its context as encoded in the features. For Lemmatization, they used SES (short edit script) which is the shortest

sequence of instructions (insertions or deletions). The beam search algorithm combines morpho-tag and lemma-class conditional probabilities means that the two outputs of the two learning models are integrated at decoding time and their predictions are combined into an overall scoring over morpho-tag and lemma-class pair sequences.

For evaluation, small testing sets of the above-mentioned three morphologically rich languages were selected and Error Analysis is performed manually to show how well the system performs with the minimalistic, general, and language-independent engineering efforts and identify the main source of mistakes. The model performs better in both tagging tasks for Romanian and Spanish then for Polish because of inflectionally richness and ambiguity behavior of the Slavic language.

After analysis the performance of the system (i.e., through Error Analysis), they found that some error categories are occurred in all three languages and others are either language or corpus specific. In addition, they suggest possible ways of dealing with some the common errors the system makes cross the languages. Both Morphological tagging and lemmatization errors tends to co-occur as a result of named entity, suffix ambiguity, syncretism, ambiguity function words, annotation problem and prefixal morphology.

Since the main objective of the prototype is to show how system achieves high accuracy with no language-specific feature or additional resources, for the maximum performance, it is desirable to extend it and refine with language and domain specific features.

2.8.1.3 Morphological Analysis of the Dravidian Language Family

The Dravidian family is one of the most widely spoken set of languages in the world. Over 300 million people in southern India speak it. However, [53] reported that there are very few annotated resources available to NLP researchers. Motivated by this problem they release annotated corpus for adopting statistical model (which requires annotated data) for NLP tasks particularly morphological segmentation and POS tagging. The family is morphologically rich (inflectional) and highly productive in compounding. Languages words are inflected for gender, number, person, case, Vibhakti[3], tense, aspect and modality like that of other Indian languages

[3] Vibhakti is a Sanskrit grammatical term that encompasses post-positionals and case endings for nouns, as well as inflection and auxiliaries for verbs. It is also referred as case-marker

[54]. These complexities make the morphological analysis obligatory for the Dravidian languages family. The researchers contribute in annotated corpus and evaluated models (morph segmenter and POS taggers) for the most commonly spoken Dravidian languages: Kannada, Malayalam, Tamil, and Telugu.

Primarily, their contribution is towards preparing and releasing a corrected, annotated, and open-source resource for both morphological segmentation and POS in the four languages. They called that annotated corpus *DravMorph*. They use closed-source rule-based morphological analyzers and POS taggers produced by the government of India and Indian universities and to ensure the gold standard they corrected manually the outputs from rule-based annotators. The statistics of the corpus shows the DravMorph has 4034-8600 annotated sentences, 30625-34200 tokens and 3593-4730 segmentation types per language.

Obviously, when annotated data is available, supervised approaches typically greatly outperform other approaches. Having this advantage and gold standard corpus, they apply fully supervised machine learning approach for both morphological analyser and POS tagger.

They apply Semi-Markov Conditioned Random Fields (S-CRF) to the problem since it has the ability to jointly model both segmentation and a labeling. To handle the multiple adjacent segments and labels, they allow high-order segment interaction as an extension of the standard S-CRF model.

For training, they use mixture of features standard for morphological segmentation and novel feature based linguistic properties of the Dravidian languages such as inflectional increments, orthographic features, and sandhi rules[4].

In addition to preparation of gold standard data they also standardized the POS tagging schemes across languages, using the IIIT-H[5] POS tagset (Bharati, et al, 2006), which has 23 tags.

Then they conduct their experiments on both tasks using the gold standard data of four languages under the question. They evaluated their models, compared the performance with available state-of-the-art systems, and gained encouraging results. In agglutinative languages like Dravidian family languages morphological segments marks case, gender, person, tense, aspect, and number categories indicative of the POS. For instance, tense marker only appears with verbs. These

[4] Sandhi rule is the rule describe phonological changes appear when two words meet to facilitate the pronunciation.
[5] IIIT-H stands for International Institute of Information Technology Hyderabad.

features have the potential to be more useful than the dynamics of the tagger as Dravidian word order is relatively free.

There by assuming segmentation as preprocessing for POS tagging, and using the output of the segmenter as feature significantly improves the-state-of-the-art taggers.

2.8.1.4 MorphNet

One of the recent language independent model developed to process language morphology is MorphNet [56]. MorphNet is a sequence-to-sequence recurrent neural network (RNN) model that designed to perform morphological analysis and disambiguation together.

Traditionally, analysis of morphologically complex languages has been performed separately in two stages: (i) A morphological analyzer based on finite state transducers produces all possible morphological analyses of a word; (ii) A statistical, a rule based, or neural network disambiguation model picks the correct analysis based on the context for each word. However, MorphNet eliminated the need for separate morphological analyzer and disambiguator components and provided a single, easy-to-use model.

MorphNet takes sentence as input and produces morphological analysis for a word in that sentence context-based sequence-to-sequence encoder-decoder approach. The model uses three Long Short Term Memory (LSTM) encoders, a character based encoder to obtain word embeddings; a word based bidirectional LSTM encoder to obtain context embeddings and a unidirectional LSTM encoder to obtain output embeddings of preceding word analyses. These embeddings are fed to the decoder, which is implemented as a 2-Layer LSTM and produces the correct stem and the morphological features of a target word.

The system was evaluated on datasets of seven languages (Romanian, Danish, French, Italian, Hungarian, Bulgarian and Catalan) from the Universal Dependency Treebank [57] and two available datasets (namely, TrMor2006 and TrMor2016) of Turkish language. The authors described datasets used in evaluations well and introduced their new dataset TrMor2018 for Turkish. When they were evaluating the model on available Turkish datasets, they realized that existing datasets suffer from low accuracy and small test sets, which makes model comparison difficult. To address these issues, they created a new dataset, TrMor2018 which was semi-automatically generated and manually confirmed to have above 97% accuracy.

The RNN system is trained by using back-propagation through time with stochastic gradient descent.

MorphNet was evaluated on eight different languages, and obtained state-of-the-art or comparable results for all languages except for French. They did not report the comparison result of MorphNet and previous state-of-the-art models for Catalan, Italian, and Danish because there were no reported performances results for those models.

2.8.2 Morphological Analyzers for foreign languages

In this context, foreign languages are those languages originally from other countries. They are not spoken and widely used by people of Ethiopia. All the above language independent works were efforts towards the foreign languages. In addition to the above works, it is better to analyse the language-dependent efforts since they deal with language specific features to deal with the improvement of the system.

2.8.2.1 Morphological Analyzers for Indian languages

Indian languages are lexically and grammatically similar. Lexical borrowing occurs between languages. Grammatically, there are many similarities between the languages. For instance languages have relatively free words order [53], [58] in sentence. Indian languages are morphologically synthetic; derivational and inflectional morphologies result in the formation of complex words by stringing two or more morphemes.

Fortunately, we were revising the efforts towards languages belongs to two dominant language families in Indian namely, *Indic* and *Dravidian* language families. Indic is a group of language dominantly spoken in North India. It is also dominant language family of south Asia (or the Indian subcontinent). Dravidian languages which are dominantly spoken in south Indian are also the second dominant language family in India. In this paper, language such as Tamil, Telugu, Kannada (see section 3.2.3), and Malayalam are category of Dravidian languages family while Hindi is the member of Indic language family. In the next sections (3.3.1.1, 3.3.1.2, & 3.3.1.3), we revised some of efforts toward other Indian languages including the above revised under section 3.2.3.

2.8.2.1.1 SMA: Statistical Morphological Analyzer for Hindi

Hindi language is a morphologically rich language with a relatively free word order. It has syntactic agreement of gender, number, person, and case. Because of these complexities of language features, NLP in Hindi had suffered due to the lack of coverage automatic morphological analyzers. To overcome this problem, many efforts in Hindi Morphological analysis concentrated on building rule-based systems that give all possible analysis of a given word form irrespective of its context in the sentence. Apart from these rule based approaches, Malladi & Mannem [59] developed Statistical Morphological Analyser (**SMA**) for Hindi trained on Hindi Tree Bank (HTB) and compare it with previously developed rule based systems particularly Paradigms based analyzers (PBA). In PBA, words are grouped into a set of paradigms depending on the inflections they take. Each paradigm has a set of add-delete rules to account for its inflections and words belonging to a paradigm take the same inflectional forms.
SMA predicts the lemma, gender, person and case makers, TAM (tense, aspect, and modality), and vibhakti[6] for all the words in a given sentence. Separate models were trained for each of grammatical features using data in HTB.

Data	# Sentences	# Words
Training	12,041	268,096
Development	1,233	26,416

Table 2.1 HTB Statistics

Lemma prediction was perceived from a machine translation perspective, with the characters in the input word forms treated as the source sentence and those in the lemma as the target. For gender, person, number and case prediction they built support vector machine (SVM) classification models using linear classifier based on the features set such as word, lexical category, lemma, previous word, next word, word length, last 3 or 4 characters, character n-grams of the word and etc. Other grammatical features, TAM and Vibhakti are predicted using heuristics[7] on fine grained POS tags of the input sentence. While nouns and pronouns take vibhakti, verb inflects for TAM. Both Vibhakti and TAM occurs immediately after the word in their respective word classes. Generally, the developed analyzer is robust enough to produce

[6]

[7] Heuristics is any approach to problem solving that employs a practical method that is not guaranteed to be optimal, perfect or rational but which is nevertheless sufficient for reaching an immediate, short-term goal.

analyses for OOV words and state-of-the-art statistical morphological analyzer for Hindi, which outperforms previous analyzers by a considerable margin. It achieved the accuracy of 82.03%.

2.8.2.1.2 SMA++: Statistical Morph Analyzer for Indian Languages

SMA++ [54] which is improved and extended version of SMA [58], is developed to predict the morphological attributes of Indian languages viz. Hindi, Urdu, Telugu and Tamil. As described above SMA primary focused on Hindi Language. For machine learning tasks system use some features those carry the morphological attributes of languages. Some of the feature sets are:

i. *Suffixes*: In Indian languages, inflectional morphemes carry gender, person, number, case, TAM and vibhakti of a word.
ii. *Previous morph tag and next morph tag*: to show agreement since it is an important characteristic in Indian languages. For instance, masculine noun takes masculine verb.
iii. *Part of speech (POS):* Based on the POS of the word, the set of possible inflections can be found. For instance, verbs have a set of inflections and nouns have another set.
iv. *Other features:* such as length of tokens and character type in the word form.

Those feature-sets contributed to high accuracy. The Support Vector Machine (SVM) using linear classifier was used for the ML task.

For the ML task, the class-labels for G, N, P, and C were chosen from the training data itself. For lemma, the class-labels were formed based on the edit-distance operations required to convert the given token to its lemma.

The results for each of L, G, N, P and C are shown individually, as well as in combination. For Hindi and Urdu, the Lemma+Gender+Number+Person+Case accuracy was **85.87%** and **79.16%** respectively. For Telugu, Gender+Number+Person+Case accuracy was **86.81%** and for Tamil it was **78.97%**. The researchers plan to run SMA++ for prediction of Lemma two later languages in future work. Generally, for all languages under investigation SMA++ outperformed other statistical morphological analyzers, Morfette [52] and SMA of Malladi & Mannem by after evaluated on given similar test sets.

2.8.2.1.3 Tamil language

Tamil language is a classical language, which belongs to Dravidian language family. More than 66 million people all over the world speak it. Tamil is an agglutinative, morphologically rich and complex and postpositionally inflected language; since it inflects to person, gender, and number

markings and combines with auxiliaries that indicate aspect, mood, causation, attitude etc. in verb. In Tamil, a single verb root can inflect for more than *two-thousand* word forms including auxiliaries and a single noun root can inflect for more than *five-hundred* word forms. This implies compared to verb morphological analysis noun morphological analysis is less challenging. Tamil language takes both lexical (derivational and compounding) and inflectional morphology.

Because of its morphological complexity, Tamil needs deep analysis at the word level to capture the word morphemes and its categories.

Anand Kumar et al.[60] developed morphological analyzer for Tamil language based on *sequence labeling approach*. In the proposed work, morphological analyzer problem is redefined as classification problem and solved using machine learning methodology. This is a corpus-based approach, where training and testing is performed with support vector machine algorithms. The training corpus consists of 130,000 verb words and 70,000 noun words respectively. The system is tested with 40000 verbs and 30000 nouns taken from Amrita POS Tagged corpus. The performance of the system was also compared with other systems developed using the same corpus and results showed that SVM based approach outperformed other. Methodology was implemented to all Dravidian languages.

2.8.2.2 Morphological Analyzers for European languages

This section discusses the morphological analysis of some morphological rich and more researched European languages such as German, Maltese, and English to some extent.

2.8.2.2.1 German Language Morphological Analysis: DEMorphy [61]

DEMorphy is FST-based German language morphological analyzer that takes word form as input and produces word category, gender, number, person, degree forms of adjectives, and etc. list of inflections. It is built onto large, compactified lexicons from German Morphological Dictionary (gmd) which contains lexicon words and list of their all possible analysis and is generated from Wikidumps[8] corpus by running some in-house morphological analyzer on the corpus and implemented as a Python library. gmd consists of word forms as they appear in written language

[8] Found at: https://dumps.wikimedia.org/dewiki/latest/

followed by list of its possible analysis, experimental analysis dictionary, list of all lemmas of word forms and list of all paradigms.

DEMorphy relies on a compacted form of German Morphological Dictionary which (precompiled dictionaries) will be delivered on updates by the developers.

Morphological analysis indeed is just dictionary lookup in DAFSA that is stored using DAWG (Directed Acyclic Word Graphs). DEMorphy stores word forms in DAFSA due to memory efficiency, fast lookup, fast fuzzy lookup and flexible iteration support.

DEMorphy can handle inflectional morphology, OOV using prefix and suffix analogy and other type of tokens such as e-mails, URL strings and date strings by skipping analysis and directly providing tokens type as result. Although DEMorphy is state-of-the-art morphological analyser, still there is room for further time efficiency improvements. It also better to integrate it with some other advanced features such as character level language models and named entity recognizers (NERs) to improve its performance. Compound word and Derivational analysis is also beyond the capability of the DEMorphy.

2.8.2.2.2 Morphological Analyzers for Maltese

Maltese is a Semitic language spoken in the European nation of Malta [62], the working language of the Maltese Islands and since 2004, also an official European language. It is morphologically rich and complex language with a hybrid morphological system that evolved from Arabic and Romance (Sicilian/Italian and English) [62], [63]. As a result, the language features both concatenative and nonconcatenative morphology. Concatenative morphology is characterized by ***stem-affixes*** and some orthographic alternation rules while nonconcatenative morphology is characterized by ***templatic-pattern***. A number of researchers have done countable researches at different times and using different methodologies for Maltese morphological analyzers. Among such efforts, it is worth considering the works of John [64], Ravishankar et al.[62] and Borg & Gatt[63].

John [64] developed Maltese morphological system using rule based approach as the component of a computational grammar for Maltese. The output of this analyser-generator has been

incorporated into a large online lexical resource, called *Gabra*[9]. This work was primarily focused on verb.

Ravishankar et al.[62] also developed Maltese morphological analysis as the component of rule-based machine translation system for Maltese to Arabic. An *lttoolbox*[10] paradigm system is used to represent Maltese morphology. It is toolbox for lexical processing and describes finite state transducer in XML. The paradigms are expressed as input side and output side; the transducer is made to return the lemma of the word and the corresponding tags. An lttoolbox is used as its simplicity to integrate to Apperitum[11] machine translation system and its simplicity to implement using python script. Python script is used to handle complex morphological behavior of Maltese such as alternation and templatic especially for verbs category.

Lexicon creation involves manually adding entries from a frequent list, which is generated from Maltese Wikipedia. The tagset is based on the standards of Apperitum [65] tagsets. In addition, they provided a mapping to the part-of-speech and morphological standards of the Universal Dependencies project [57].

The morphological analyser was evaluated on two corpora: the entire Maltese Wikipedia and Korpus Malti[12] [66]. As some section of Maltese Wikipedia was existed in Italian, langid.py [67] was used for preprocessing. Generally, the result of the evaluation is presented in terms of raw coverage and accuracy. The system coverage is 80% over two corpora. The accuracy; precision and recall of the system measured on a manually evaluated test set are also satisfactory, at 96.2% and 95.3% respectively.

Unlike the above researchers, Borg & Gatt [62] presented complete Maltese morphology system using unsupervised (i.e., for clustering of morphologically related words) and fully supervised (i.e., for the morphological labeling of words) method, with a particular emphasis on the problem of hybridity (mixed morphological processes existing side by side) in the morphological system.

The clustering techniques for Maltese may show different performance depending on which morphological system we are looking for: concatenative or nonconcatenative.

[9] Can be accessed from: http://mlrs.research.um.edu.mt/resources/gabra/
[10] Can be accessed from: https://github.com/apertium/lttoolbox
[11] Can be accessed from: https://www.apertium.org/index.eng.html?dir=eng-spa#translation
[12] Can be accessed from: http://mlrs.research.um.edu.mt/

In their work they analysis the clustering techniques of Borg & Gatt [46] which first identify potential affixes and then cluster words on the bases of common stem. In addition, *they improve it using the measures of orthographic and semantic similarity*. The datasets obtained from the clustering technique were split into concatenative and nonconcatenative sets.

The clustering system was evaluated in two methods: number of words removed from clusters and quality ratings of the cluster. The methods were depending on the decision of experts (native language speakers/linguistics).

Number of words removed indicates the words are not belongs to the cluster because most of the time words may be grouped together mistakenly with overlapping strings. For instance, *ittra* 'letter', *ittraduce* 'translate', and *itratat* 'treated' are clearly unrelated words but the system may incorrectly identify ***ittra*** as a potential stem in all these words. In quality rating experts are asked to rate the quality of the clusters. The correlation between both the two evaluating methods is calculated using Pearson's correlation coefficient. Perceptions of clusters quality is related to the percentage of words removed.

Generally, the clustering techniques performed best on the concatenative set than that of nonconcatenative. Because in nonconcatenative clustering there are many derived forms, which are difficult to cluster initially due to, *stem allomorphy* (root-based derivation including infixation and vowel melodies, which are unpredictable).

The authors viewed morphological labeling as a classification problem. Each morphological property is seen as a machine-learning feature, and each feature is modeled as a classifier and placed in a cascade to provide the complete label to a given word. The classification focused on the verb category. The classification also relies on basic features such as stem, prefixes, suffixes, CV pattern, and gemination.

After the classification system trained using decision trees through WEKA data mining software on over 170,000 annotated data from lexical resource, 'Gabra'; two types of evaluations were carried out to test its performance. The first was a traditional evaluation using unseen data (20,000) from 'Gabra'. The second evaluation used randomly selected words from the Maltese national corpus (MLRS—Malta Language Resource Server[13]), which were manually annotated

[13] http://mlrs.research.um.edu.mt/

with their morphological labels and has 94 verb categories (76 nonconcatenative and 18 concatenative). This was treated as a gold standard.

The evaluation was useful to determine where more representative data is needed and to assess which morphological properties were not performing adequately. Although the accuracy of the morphological classification system is exceptionally low for some of the morphological properties, the system performs well overall, and the individual classifiers can be *retrained* and *improved* as more representative data becomes available.

Generally, the demonstrations show the current approach is viable for both concatenative and nonconcatenative morphological systems and can be well suited for hybrid morphological systems.

2.8.2.2.3 English Morphological Analysis

English is morphologically near-analytic language [56] that shows a low ratio of morphemes to words and express most grammatical relations using *function words or word order* unlike that of synthetic languages (language with a high morpheme-per-word ratio). However, it is one of the well-researched languages in the area of natural language processing. Morphological systems for the language have been done as the input for higher-level package components. Here we review TANG's [68] English Morphological Analysis with Machine-learned Rules. TANG developed the language dependent algorithm that learns morphological rules (rule-building phase) and analysis (labeling phase) English word. These two components are essential to achieve accurate morphological analysis.

The author also employed machine-learning approach, which makes the use of lexical database without morphological information (unsupervised machine learning) to avoid problems such as costly human labor needed for developing of handcrafted rules, rule inconsistency and to provide additional statistical information, which can be used in morphological analysis procedure for disambiguation.

The analyzer adopted the approach proposed by Keshava & Pitler [29] in learning affix rules from wordlist and tested the approach using wordlist of different scales. To learn the affix rules, one forward lexicographic tree and one backward lexicographic tree were built using a corpus of

24,447,034 tokens. Potential affixes are recognized through a scoring procedure (i.e., probability score).

Although wide-covering and correct set of affix rules is prerequisite for accurate morphological analysis, it alone does not guarantee a successful analysis. The procedure in which the analysis is done is also crucial. In the paper, two important aspects are dealt with in terms of analysis procedure control: disambiguation and affix rule order.

Ambiguity may occur at segment level (intersectional ambiguity to decide where the morphological boundary is) or at word level (combinatory ambiguity to decide whether the word has affix or not). Authors reported experiment conducted shows that the method mentioned above improves the performance to a large extent. However, in addition to simple transitional probability; combinatory ambiguities need richer contextual information such as grammatical category (POS) and lexical meaning (context between words) for correct analysis.

In addition to affix learning rules and disambiguation many languages specific morphological features, such as morpheme application order with some exceptions are considered. English morpheme order rule can be captures by sequence of:

Lexical Morpheme ➔ Derivational Morpheme ➔ Inflectional Morpheme

The experiment shows that the analyzer has a satisfactory performance, and the result is fairly higher than many other algorithms. However, problems remain can be solved with a larger context, such as part of speech, or context between words.

2.8.3 Morphological Analysis for Local Languages

Local languages refer languages of Ethiopia include the nation's official languages, its national and regional working languages, as well as its minority. It indicates 86 individual languages indigenous to Ethiopia according to Ethnologue[14]. Although some of these languages have official status in regions or nations and are crucial for development, they remain an under-studied and under-resourced language from the NLP perspective.

In this paper, we reviewed the works on the two largest languages Afaan Oromo and Amharic, which are spoken by **33.8**% and **29.3**% of the country's population respectively according to the 2007 Ethiopian census.

[14] https://en.wikipedia.org/wiki/Ethnologue

2.8.3.1 Amharic Morphological Analysis

Amharic a Semitic language, related to Hebrew, Arabic and Syriac [41] is a working language of the Ethiopian Federal Government and some regional governments in Ethiopia. Therefore, most documents in the country (including electronic and online accessible documents) are produced in Amharic. However the language has lack of computational resources especially morphological analyser which is the essential stage for downstream NLP tasks of morphological rich and complex languages. To address lack of the language, many researchers conduct different researches at different time using different approaches ranging from FST approaches [37] to machine learning approaches [36], [41], [69].

Among these works we reviewed the work of Michael Gasser [8] which is FST-based and that of Wondwossen Mulugeta & Michael Gasser [41] which employed Machine learning approach.

Michael Gasser [37] employed FST to develop a python library called "HornMorpho", a morphological analysis and generation system for Tigrinya, Amharic, and Oromo language.

The FST that is empowered by feature structures was effectively adopted to process the unusual nonconcatenative root-and-pattern morphology. Because FST enables us to implement phonological rule (morphophonemic/alternation rule) and rule ordering (morphotactics), and a cascade of composed simpler FST, each are responsible for some aspects of morphology, can implement the two-level model with resulting single FST relating to surface and lexical level directly.

The Amharic module of HornMorpho performs the full analysis of Amharic verb and noun (nouns and adjectives) classes. For Amharic, the lexicon is compiled from the Amharic-English dictionary of Aklilu, which has 1,851 verb roots and 6,471 noun stems. The analyzer takes the word input, process using cascades of FSTs (for Romanization, gemination, alternation rules, guesser and etc.) and output a root or stem and grammatical analysis. Analyzer function can analyse a single word form or all word(s) in the file.

The system was tested on randomly selected 400 word forms (200 verbs and 200 nouns and adjectives).

System is evaluated manually by error counting approach. An output of the analyzer was considered correct only if it found all legal combinations of roots and grammatical structure for a given word form and included no incorrect roots or structures. The analysis revealed remarkably

accurate results, with 99% accuracy for verbs and 95.5% accuracy for nouns and adjectives, with a few errors due to unhandled FSTs, which can be integrated.

Wondwossen & Michael [41] employed another approach, machine learning, to learn morphology for Amharic verbs using inductive logic programming[15] (ILP). ILP was implemented in CLOG[16]. Amharic verbs have most complex and nonconcatenative morphology, also characterized by alternation rules governing the form that morphemes take in a particular environment (context). For example a single Amharic verb root can take a chain of four and five possible prefixes and suffixes respectively to show various grammatical features. The grammatical features of Amharic verbs are also shown by intercalation pattern of the consonants and the vowels (templatic-morphology).

To learn morphological rule ILP was provided manually prepared training data and background knowledge predicates that can handle stem extraction by identifying affixes, root and vowel identification and grammatical feature association with constituents of the word.

After trained the program on training set (manually prepared 216 Amharic verbs), 108 rules for affix extraction, 18 rules for root template extraction and 3 rules for internal stem alternation have been learned.

Then they combined the background predicates used for the three learning tasks and integrated all the rules learned. The program has been evaluated using a test set containing 1,784 Amharic verbs. Their system achieved the encouraging result, 86.99% accuracy to extend to more morphological attributes and word classes. However the rule coverage is bottleneck in CLOG.

2.8.3.2 Afaan Oromo Morphological Analysis

Afaan Oromo, one of a major African language (Cushitic language) that is widely spoken in Ethiopia and other African countries Kenya, Somalia, Uganda, Tanzania and Djibouti as minorities language [7], [8]. The status of Afaan Oromo was already described above in chapter 1.

Although Afaan Oromo is the largest spoken language in Ethiopia, very limited works have been done in the past in the areas of morphological analysis in relation to the language. Researchers

[15] Inductive Logic Programming (ILP) is a supervised machine-learning framework based on logic programming. In ILP a hypothesis is drawn from background knowledge and examples

[16] CLOG is a Prolog based ILP system for learning first order decision lists (rules) based on positive examples only.

conducted researches on Afaan Oromo Morphological Analysis [37] and its subtasks such as stemming [20], root generation [21] and segmenting.

Abebe Abeshu [8] also used rule based approach of paradigm based to develop Afaan Oromo Automatic Morphological Synthesizer, which works in opposite direction of morphological analyzer (i.e., Synthesizer generates word form from its constituents and grammatical features while analyzer breaks down word form into its morphemes and grammatical features).

Wakweya Olani [8], also wrote his master's thesis on linguistic theoretical aspects of Afaan Oromo Morphology, which is founded it is important to understand the morphological features of the language.

The first reported work on Afaan Oromo morphological analysis was HornMorpho [37], the morphological synthesizer and analyzer for three main Ethiopian languages; Amharic, Tigrigna and Afaan Oromo using Finite state Transducer (FST). The Afaan Oromo module of the system can handle nouns and verbs. The *lexicon* of verb and noun roots was extracted from the dictionaries of Gragg[17] (1982) and Bitima[18] (2000). In the dictionaries, there were 4,112 verb roots and 10,659 noun stems. The module also provided segmenter function handled by separate FST. Like for Amharic analyser for Afaan Oromo can handle both a single word and words in file.

The performance of Afaan Oromo module was not yet evaluated, because it is complicated by the great variation in the use of double consonants and vowels by Afaan Oromo writers, compilation of lexicon form two inconsistent dictionaries and less knowledge of Afaan Oromo morphology. However, in HORNMORPHO 2.5 User's Guide [16], the author reported that the performance of the program for Oromo is inferior to that for the other languages with unknown performance in percentage.

Another the most related and the latest effort towards Afaan Oromo morphological analysis is, the work of Moyka and Dida [17], which is proposed based on machine learning approach, specifically, Memory Based Learning (MBL) algorithm. They selected MBL because of its "lazy" property and its appropriateness for Afaan Oromo morphological analysis, as lazy machine-learning methods achieved a higher accuracy than eager methods for many NLP tasks.

[17] Gragg, G. (1982). Oromo dictionary. Michigan State University Press, East Lansing, MI, USA

[18] Bitima, T. (2000). A dictionary of Oromo technical terms. Rudiger Kopper Verlag, Koln.

Their work clearly addressed the morphological properties of Afaan Oromo word classes such as nouns, verbs and adjectives which are reviewed and investigated from most of linguistic works and computational approaches to morphological analysis.

The developed system has two main components: training and analysis phase. The training phase, which comprises necessary components that are used in the process of training the learning component of memory-based learning, contains feature extraction, memory learning and trained model sub-components. The analysis phase, the phase that maps the input into output, contains morpheme identification and morpheme extraction components.

For training and testing the system, the author developed OROLEX which is a morphological database consisting of the grammatical description of 2,270 annotated Afaan Oromo words (nouns, adjectives and verbs). From these Afaan Oromo annotated words, 17,386 instances were automatically extracted. Then the dataset partitioned into training and test dataset and the training dataset shares 90%, while test dataset shares 10% of the total dataset.

The components within training phase operate on the training dataset and produce the trained model.

Feature extraction makes/extracts instances, generates and supplies necessary information (features) from Afaan Oromo morphological database, OROLEX during the development of the model based on the concept of windowing method in a fixed length of left and right context. In this method a word is converted into fixed-sized instances of which the letter at focus position is mapped to a class denoting a morpheme boundary. To generate fixed-sized instances from OROLEX, the authors employed the following Algorithm.

Input: Annotated words

1. Define the length of window size (7-1-7).
2. Mark the middle positions of arrays as a focus letter (the focus letter represents where the first letter of a word starts at).
3. Read from the database and push one step forward each character until the right context reached(filled).
4. Put 0(zero) at the class if there are no any special character like @, &, digits and capital letters, next to the characters placed in the focus letter; if any one of those symbols exist put the value as a class (in the last index)
5. Push the previous focus letter to the left and start putting each letter (as in step 3)
6. Go until it finishes that line
7. Go to the next line and repeat the steps 3,4,5,6.

Output: - Instances

Algorithm 2.1 Instance Extraction [17]

Windowing schema converts each word form into as many instances as it has letters. Each instance focuses on one letter, and includes a fixed number of left and right neighbor letters, chosen here to be seven because the average word length in OROLEX is eight. In this way each instance spans fifteen letters, which also happens to be the longest word length in the OROLEX database.

Memory base learning is a component designed to learn an instance and then build a trained model. It adds training examples to a memory. During training, a set of examples, the training instances are presented to the memory-based learning component and those instances are added to the memory (the instance base or case base) without abstraction, selection, or restructuring.

The ***trained model***, which is the final output of the learning process, contains a set of instances called instance base or case base which is extracted from OROLEX using windowing scheme and fed to memory- based learning component. It also provides trained instances information during analysis.

In the **analysis phase**, the instances stored in trained model are used in order to map the input to the output. When new words to be classified are given to the system, the ***feature extraction*** component will deconstruct the word in a fixed-length of instance. The features are extracted as the process of feature extraction described above.

Morpheme identification is used to classify and extrapolate the class of new instances based on the output from feature extraction component.

Morpheme *extraction* is done after the appropriate morphemes are identified during morpheme identification. In morpheme extraction, reconstruction of individual instances into a meaningful (to their original word form) and insertions of identified morphemes in their segmentation point are performed.

The authors used IB1 and IGTREE memory-based learning algorithms implemented in TiMBL in order to train and test the dataset. The model was evaluated in four scenarios by default parameter settings, feature selection, parameter optimization, and interleaving feature selection and parameter optimization. They organized the presentation accuracy of each scenario into generalization accuracies. Among all the scenarios, interleaving the combination of selected features and optimal parameters of IB1 obtains the generalization accuracy of 98.86%, while IGTREE obtains the generalization accuracy of 94.14%. From this result, we conclude that the feature selection plays a vital role in getting the best accuracy and there is a trade-off between IB1 and IGTREE algorithms. IB1 usually leads to more accuracy at the cost of memory and slower computation than IGTREE.

However, most of lazy learning methods including KNN which is used by these authors have their own limitations (see section 1.2)

2.9 Summary

This chapter discussed various morphological analysis and related concepts. The discussions included in this Chapter on supervised machine learning approach with rule based approach will be applied in the experimentation chapters to develop the morphological analysis prototype and algorithms.

The chapter also discusses related works in the area of morphology that have been done for different languages. Morphological analyzers of many languages developed using different approaches such as rule-based and machine learning were discussed. All the above morphological analyzer systems developed either as the component of other system or stand-alone system.

Morphological analyzers may use generic, minimalistic, and language independent features to handle word morphology. It also uses language specific or document specific features to increase the performance. Language specific features are very important to increase the performance of language NLP specially, morphological processing because every language has unique characteristics, which it cannot share with others. Thus, the next chapter will discuss Afaan Oromo morphological behavior in detail.

The above works, related with Afaan Oromo morphological analysis used either rule-based approaches, which need handcrafted rules and compilation of lexicons; or memory based learning approaches, which is lazy learner method and need lookup database. However, eager machine learning methods used by many researchers and outperforms other approaches in the presence of an adequate corpus. Eager machine learning method also avoid problems such as costly human labor needed for developing of handcrafted rules, rule inconsistency and provide additional statistical information, which can be used in morphological analysis procedure for disambiguation.

Thus, we proposed a new architecture for Afaan Oromo morphological analysis based on eager learning methods, “rule-based and statistical”. Rule-based approach is selected because of its accuracy and no data needed to develop rules. Statistical classifiers have exhibited high accuracy and speed, specially, when applied to large databases [18]. We used a hybrid Approach for Afaan Oromo morphological analysis. Rule-based approach was used for word segmentation while HMM was used for morph tagging/labeling.

CHAPTER 3 AFAAN OROMO MORPHOLOGY

3.1 Overview

Afaan Oromo, the major languages that is widely spoken and used in Ethiopia [70], is morphologically polysynthetic[19] (agglutinative [3], [71], [72] and fusional [8]), very complex and productive like most of African and Ethiopian languages. For instance, in [22] one verb root and noun stem can take on average 230 and 25 surface forms respectively. In Afaan Oromo, all bound forms (morphemes) are affixes. In polysynthetic languages like Afaan Oromo most of the grammatical information is conveyed through affixes (specifically, suffixes) attached to the stems/roots of the target word. For instance, Afaan Oromo nominals (nouns, pronouns, and adjectives) are highly inflected for gender and number while Afaan Oromo verbs are also highly inflected for gender, person, number, and tenses (aspects). Moreover, possessions, cases, and article markers are often indicated through affixes in Afaan Oromo. As evidence, we can compare Afaan Oromo plural makers with that of English. In comparison to the English plural marker 's' ('es'), there are more than 12 major and very common plural markers in Afaan Oromo nouns (e.g. -oota, -ooli, -wwan, -lee, -an, een, -oo and so on) [70].

Morphology is the study of word formation. It is the study of the internal structure of word [6], content of word forms and the rules for formation of grammatically right and acceptable words. Word is the smallest meaning-bearing unit in language. In Afaan Oromo **inflection**, **derivation**, **reduplication,** and **compounding** are common word formation processes. Obviously, these extensive inflectional and derivational features of the language are presenting various challenges for text processing tasks in Afaan Oromo. These complications are also worse for Afaan Oromo which is resource scarce language and whose structure has not been studied extensively even in linguistic field of study.

This chapter presents the morphological behavior of Afaan Oromo, which is important in the areas of Afaan Oromo word including structure, function, nature, and categories of words. Afaan Oromo word morphology ambiguities, phonological behavior of Afaan Oromo (for alternation rule), and writing system are also analyzed and reviewed in this chapter. The chapter will help to

[19] Polysynthetic language is a language in which words tend to consist of several morphemes and allomorphy.

understand the internal structure of Afaan Oromo words and beneficial for those researchers who want to conduct further study on it. It is also beneficial for the software developers of Afaan Oromo natural language and speech processing applications.

Since our work is limited text processing, the next section revises mostly orthographic behavior of Afaan Oromo.

3.2 Afaan Oromo Writing System

The Afaan Oromo writing system is a modification to Latin writing system. The language adopted Latin-based alphabet called Qubee Afaan Oromo (shortly Qubee) and using it officially for its writing system since 1991 [9], [13], [14]. The writing system of Afaan Oromo is nearly phonetic since it is written the way it spoken, i.e. one letter corresponds to one sound. Language has also its own syllable structures and combination to form word as will be defined below in section 3.2.1. The language has its own consonants and vowels sounds. Afaan Oromo has thirty-three letters, of these, seven of them are *combined consonant letters*: CH, DH, NY, PH, SH, TS and ZH. Five of them are vowel letters: A, E, I, O and U. The combined consonant letters are known as "**Qubee dachaa**".

Afaan Oromo has five short and five long vowels, which are *sound makers* and stand by themselves. The five long vowels can be obtained by doubling the corresponding short vowels. Afaan Oromo has also *gemination* (double consonants). The difference in length of the vowel and gemination of consonants induce difference in meaning. For examples: **lafa** "earth", **laafaa** "soft", **badaa** "bad" **baddaa** **"**highland**"** and etc.

In Afaan Oromo, a single letter is constructed from one character symbol or digraph (double character symbol) like **ch**, **dh**, **ny** and etc. Gemination is allowed for single (one character) symbol like hoffaa, which means 'light/not heavy' whereas gemination for digraph (double character) symbol is not allowed in Afaan Oromo. For example, qophii which means 'readiness' cannot take the form qophphi.

Capital and small letters like English alphabet characterize the Afaan Oromo. Thus, the language writing system shares many features with English writing with some important modifications.

Afaan Oromo alphabets are:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

CH DH NY PH SH TS ZH

3.2.1 Syllables in Afaan Oromo

Syllable is a unit of organization for a sequence of speech sounds. Syllabification, the separation of a word into syllables, is language-dependent. Each language has its own structure of syllables. For example, in English more than two consonants can come consecutively in a single word as in '**scr**een'. Nevertheless, in Afaan Oromo more than two consonants cannot come together except in diagraphs (which is considered as a single letter). The schematized syllable format of Afaan Oromo can be represented in possible seven structures CV, CVV, CVC, CVVC, VC, VVC and V; where C is a variable for 'consonant', V is a variable for 'vowel' and VV represents a long vowel. A valid word can be composed from the combination of one or more type(s) of these structures.

3.3 Afaan Oromo word classes (Parts of Speeches)

Word is the smallest meaning-bearing unit in language. In SIL[20] a word is defined as a unit which is a constituent at the phrase level and above. It is sometimes placed, in a hierarchy of grammatical constituents, above the morpheme level and below the phrase level. It is a basic part of a given language.

Word class/category, sometimes called parts-of-speech, morphological class or semantic tag [73], is a category of words (lexical items or surface form) that have similar grammatical properties, play similar roles within the grammatical structure of sentences and sometimes have similar morphology in that it undergoes inflection for similar properties. For instance, verb class can take tense inflection while noun word class cannot.

The arrangement of words or their combination in sentences depends on the rule or grammar of that language. For instance, English sentences follow ***subject-verb-object*** order while those Afaan Oromo follow ***subject-object-verb*** sequence. The meanings of these sentences depend on each word of the sentence and the way they are arranged but the extent to which a given word

[20] SIL stands for summary institute of linguistics: https://glossary.sil.org/term/word

determines the meaning of a sentence depends on the contribution of that word. Not all words have equal contributions to sentence meaning. Their contribution depends on their category and their feature. Based on the category of the word; we can find out the contribution of that word. In addition to this, we can easily identify the rules to be applied to this word in case of morphological change.

Knowing the word class of a given word form also simplifies the task of syntactic or semantic parsers [74]. Word classifier can be incorporated in NLP systems such as information extraction, information retrieval, machine translation and speech processing (i.e., it can further tell us about how the word is pronounced).

In general, word categories can be identified by looking at the ***meaning*** (semantic) of that word, by looking at the ***form/shape*** (morphology) of that word or by looking at the actual ***position/environment*** (syntax) of that word in sentence. Identifying word class by looking at the meaning of Afaan Oromo word is the focus of this work as it is input for the system.

Although different schools of grammar present different classifications for the parts of speech, the grammatical categories of Afaan Oromo have undergone a series of improvement in terms of its word categories and other syntactic features. Traditionally, Afaan Oromo words are categorized into eight grammatical categories (Noun, Verb, Adjective, Adverb, Adposition, Pronoun, Conjunction and Interjection). However, recent studies categorize Afaan Oromo words in to five classes which include nouns, verb, adverb, conjunction and adposition [73]. Pronoun and adjectives are categorized under noun. Preposition and postpositions are categorized under adoptions.

From this point, one can understand that Afaan Oromo has five major grammatical categories serving as heads in phrase construction. Each of these classes again can be divided into other sub-classes. For instance, noun class is categorized as proper noun, common noun and pronoun, and Preposition and postpositions are sub classes of adpositions. The subclasses in turn can be divided into subclasses, and the subdivision process may continue iteratively depending on the level and aim of the investigation. In detail, we will see some of Afaan Oromo word categories that have contributions to our study that are actually open class word categories as justified below.

Like in most of languages, Afaan Oromo word classes can be classified as *open* and *closed* depending on their functions or lexical meaning. They are defined here in the following sections.

3.3.1 Open classes

Open word classes, sometimes called content words or lexical words, are the class of words that carry a high degree of meaning. This class is always ready to accept new word as new members of this group can easily be created and added if new things are created or new concepts developed, so that the vocabulary has to be expanded to accommodate these items. For instance, noun class is potentially infinite, since it is continually being expanded as new scientific discoveries are made, new products are developed, and new ideas are explored. For example, in the late twentieth century, developments in computer technology have given rise to many new nouns: *internet*, *website*, *URL*, *CD-ROM*, *email*, *newsgroup*, *bitmap*, *modem*, *multimedia*, etc. New verbs have also been introduced: *download*, *upload*, *reboot*, *right-click*, *double-click*, etc. On the other hand, we never invent new prepositions, determiners, or conjunctions that are closed word classes.

Words are added to open classes through such processes such as compounding, derivation, coining, and borrowing. Open word classes include:

- ✓ **Nouns:** are words or lexical items denoting any abstract (abstract noun: e.g., *home*) or concrete entity (concrete noun: e.g., *house*); a person, place, thing, idea (e.g., *happiness*), or quality. They are the most common part of speech; they are called naming words.
- ✓ **Verbs:** are words denoting an action (*walk*), occurrence (*happen*), or state of being (*be*). Without a verb, a group of words cannot be a clause or sentence.
- ✓ **Adjectives:** are modifiers of a nouns or pronouns. They make the meaning of another word (noun, pronoun, etc.) more precise.
- ✓ **Adverbs:** are the modifiers of an adjective, verb, or another adverb (e.g. *very, quite, frequently etc.*). They make language more precise.

3.3.2 Closed classes

Closed word classes, known as functional words are the class of words that have a predominantly grammatical function and relatively little meaning themselves. They act as markers or guides to the structure of a sentence. In contrast to that of open word class the closed word classes acquire new members very infrequently even if it is needed in some cases because it is very uncommon to have new function words created in the course of speech. They don't readily accept new members and very resistant to the introduction of new items. *This group of words is much fixed* and we need them so frequently to link together different parts of a sentence or to create a sense of continuity (cohesion) between different parts of a text or spoken language. So those are among the most important words to learn. These classes include:

iv. **Pronouns:** are substitutes for nouns or noun phrases (*them, he*). Pronouns make sentences shorter and clearer since they replace nouns and noun phrases.

v. **Determiners (articles, demonstratives and etc.):** are grammatical markers of definiteness (*the*) or indefiniteness (*a, an*) in some languages. The article is not always listed among the parts of speech as it can be delivered using inflections. For instance Afaan Oromo doesn't have articles.

vi. **Adposition (prepositions and postpositions):** are words that relate words to each other in a phrase or sentence and aids in syntactic context (*in, of*). They show the relationship between a noun or/and a pronoun with another word in the sentence.

vii. **Conjunctions:** are syntactic connector; links words, phrases, or clauses (e.g., *and, but, or*). They connect words or group of words.

Our work focuses on open class words (content words) rather than functional words for the following reasons:

- As the class can accept new words frequently, it can be influenced by the problems those caused by out of vocabulary (OOV) in language analysis processes.
- The frequent creation of new words in this class need to frequently update of lexicons to handle new unknown words (OOV) which is tiresome.
- The NLP tasks of closed class word elements can be simply handled using even lookup table, rule based approach, or Paradigm based approach since the elements are fixed in number.
- Some of the closed word class elements do not inflected. For instance, conjunctions in English, Amharic, and Afaan Oromo cannot take any inflection.

- Most of concepts carried by function words in Analytical languages (or nearly analytical languages like English) are delivered by inflections in inflectional languages such as Afaan Oromo.
- Some of them are not available in some languages. For example, for preposition is represent by inflection –f as in '*isaaf'* which means 'for him'
- Other closed word members are very few in numbers. For pronouns and conjunctions can be listed, sometimes they are considered as stop words in many language processing tasks.
- Content words are also huge in number compared with the functional ones.

To handle the morphological complexity Afaan Oromo content words (open class words), Machine learning approach, which learns rules from annotated huge data, is used. Thus, the next section describes the morphological behavior of content word classes: Nouns, verbs, adjectives and adverbs.

3.4 Afaan Oromo Inflectional Morphology

Inflectional morphology is concerned with the inflectional changes in words where word stems are combined with grammatical markers for person, gender, number, tense, case, and mode. Inflectional changes do not result in changes of parts of speech. Actually, Afaan Oromo has both free and bound morphemes as defined in chapter 2 of this thesis. Free morphemes in Afaan Oromo are either stems or words with zero bound morphemes. Afaan Oromo roots such as dhug- 'drink' and beek- 'know' are bound as they cannot occur on their own and are pronounceable only when other completing affixes are added to them. Roots are affixes (suffixes) takers while suffixes are grammar makers.

In Afaan Oromo, most of grammatical makers are suffixes, which are bound morphemes. Suffixes in Afaan Oromo can be categorized into three basic groups: *derivational*, *inflectional*, and *attached suffixes*.

Afaan Oromo derivational suffixes such as -achuu, -eenyaa, -ina and -ummaa are often used for formation of new words in the language following the rule of word formation such as morphophonology while inflectional suffixes comprise the most frequent and dominant suffixes such as –n, -lee, -een, -icha, -tu, -oo, -oota and –wwan. Afaan Oromo attached suffixes are *particles* or *postpositions* like -irra, -itti and –dha.

In complex word structure, certain set of suffixes conventionally come in a particular sequence before or after other suffixes. The most common order/sequence of the above major three Afaan Oromo suffixes (within a given word) is - **Root/stem➔derivational suffixes➔inflectional suffixes➔attached suffixes.**

The order is relative because one or more of these suffixes may be absent.

The focus of this work is inflectional morphology, which is characterized by inflectional suffixes. Thus, it considers inflectional suffixes that are appended to roots or stems and derived stems (i.e., derivational suffixes those change the root into a stem on which inflectional morphemes can be suffixed). The forms, functions, and patterns of occurrences of morphemes of the language described through this chapter.

3.4.1 Noun Inflections

Noun is a class of words that names and identifies things, people, animals, places, ideas (feelings), state or quality. All nouns in Afaan Oromo end in vowel with few exceptions ending in particular consonants [8]. In Afaan Oromo most of the time a sentence begins with a noun which starts with capital letter and it uses a noun as a ***subject*** followed with subject markers (-ni, or -n,). ***Direct object*** and ***indirect object*** also optionally follows the noun which is the subject of a sentence. Every word that has affixes like –***eenya***, -***ina***, -***umma*** are nouns in Afaan Oromo.

Afaan Oromo nouns are inflected to indicate different grammatical functions such as ***number***, ***gender***, ***definiteness,*** and ***case***. Inflectional suffixes are combined with stem usually resulting in a word of the same class as the original stem.

3.4.1.1 Number

In Afaan Oromo the most common number features are plural and singular. Afaan Oromo uses different suffixes to form plural of a noun. These suffixes are different from dialect to dialect[71] and they can be used alternatively by replacing one by other pluralization suffix or concatenatively one after another[3].

For instance, jabbii ‘calf’ can be inflected to Jabbilee (jabbii +-lee), jabboota (jabbii + -oota) or jabbiloota (jabbii +-lee +-oota).

However, some collective nouns exist only in plural form only like in hamaamota ‘bride servants’ and ijjoollee ‘children’ whereas some others have the same singular and plural forms

like ilkaan ‘tooth/teeth’, bishaan ‘water’, etc. Some naturally single nouns such aduu and Waaqaa are senseless if they take pluralization suffixes.

The common suffix morphemes to mark the plural form of the nouns are: **-oota –lee –wwan –yyii -een –oolii –eetii –ii –oo –iin -an.** These plural markers vary based on semantic nature of the base noun [8] and categorized based on universal usage [3]. Abebe Abeshu[3] classified the pluralization suffixes (with their variation) in four groups.

Group 1: oota, oolee, oolii, ilee (ota, olee, olii, ilee)

These suffixes delete all last vowels of citation form of nouns and then attached. The variations between long form and short form of this group suffixes are based the length penultimate syllable (vowel present in the syllable that precedes the last syllable) of a base noun. Thus, when the penultimate syllable contains short vowel, long from (oota/oolee/oolii/iilee) is suffixed but when it contains long vowel short form (ota/olee/olii/ilee) is suffixed for plurality of the nouns.

For examples:

Noun	**Plural form**	**Gloss**
Nama	Nam**oota**	Men
Gaara	Gaar**ota** gaar**olee**	Mountains
Sa’a	Sa’**oolee** sa’**oota** sa’**oolii**	Cows
Mammaaksa	Mammaaks**ota**	Proverbs
Jabbii	Jabb**iilee**	Calves
Waatii	Waat**ilee**	Calves
Ganda	Gand**oota**	Kebeles
Diina	Diin**ota**	Enemies
Gaangee	Gaang**olii** gaang**olee** gaang**otii**	Mules

Table 3.1 Examples of nouns pluralized by oota, oolee, oolii, ilee (ota, olee, olii, ilee)

Group 2: -wwan, -lee

These suffixes are used with nouns terminating in long vowels without deleting the final vowels.

Examples:

Base form	**Plural form**	**Gloss**
Gaaffii	Gaaffii**wwan**	Qestions
Murtoo	Murtoo**wwan**	Decisions
Barruu	Barruu**lee**	Written materials
Bu'aa	Bu'aa**lee**	Benefits

Table 3.2 Examples of nouns pluralized by -wwan, -lee

Group 3: -an, -een, -iin

All nouns that take this group of plural maker suffixes make plural noun by **doubling the consonant** in the last syllable. These nouns mostly end in the *consonantal phoneme* **l, m** and **r** in the case of **-an**, and **b, d, g, k, n** in the case of –**een/-iin** followed by short vowel. –**iin** and –**een** reflect only dialectal differences but their function is similar.

Note that in doubling consonants writing system matters as in the examples with *. In Afaan Oromo, more than two consonants cannot come consecutively in a single word.

Examples:

Base Forms	**Plural Forms**	**Gloss**
Eessuma	Eessumm**an**	Uncles (maternal)
Wasiila or abeeraa, adeera	Wasiill**an**/adeerr**an**	Uncles (paternal)
Beera	Beerr**an**	Old women
Gaara	Gaarr**een**/gaarr**iin**	Mountains
Muka	Mukk**een**/mukk**iin**	Trees
Mana	Mann**een**/mann**iin**	Houses
Ilma	* Ilma**an**	Sons
Farda	* Farde**en**	Horses

Table 3.3 Examples of nouns pluralized by -an, -een, -iin

Group 4: -eyyii

This plural maker suffix forms plural by dropping complex endings that form nouns like –**(e)essa**, (e)**ensa**, **(e)ettii** and attaching the plural maker –(e)**eyyii** as in:

Base forms	Plural forms	Gloss
Sooressa	soor**eyyii**	Riches
Dureettii (f.)	dure**eyyii**	Riches
Bineensa	bine**eyyii**	Wild animals
Waraabessa	waraab**eyyii**	Hyenas
Dureessa (m.)	Dureeyyii	Riches

Table 3.4 Examples of nouns pluralized by -eyyii

Although the categories of suffixes discussed above are the most common ones, it is difficult to categorize a given noun under one of the above categories as members of different categories can be alternatively attached.

In addition to the above discussed plural markers, Afaan Oromo proper noun uses associative marker fa'a, which means 'and others', (***-faa***) to identify a group referring to human [8]. For instance, Caaltuu-faa 'chaltu and others. The morpheme -faa is suffixed to someone's name in the group and suffixed on interrogative pronoun eenyu 'who' in questions. The form is eenyuun-faa 'whom and others.

3.4.1.2 Gender

There are two types of gender in Afaan Oromo: Masculine and Feminine. Limited group of nouns differed by using different suffixes for masculine and feminine form [71] because some of the nouns are distinct (distinguished semantically) as in eessuma/wasiila/abeeraa 'uncle' and adaadaa/haboo 'aunt'. Some of the nouns are epicene as in hattuu 'thief' while some of them are expressed using contrary adjectives/verbs those follows them such as kormaa 'male' and dhaltuu 'female' as in lukkuu kormaa 'rooster' and lukkuu dhaltuu 'hen' or sareen dhufe 'the dog came/masculine' and sareen dhufte 'the dog came/for feminine'. Some of language's nouns are considered as natural male or female gender. For instance, biiftuu 'sun' is considered as female while garba 'ocean' and Waaqa 'God' are considered as male gender grammatically.

Some of Afaan Oromo gender markers suffix pairs are **-ssa/-ttii, -aa/(d)tuu, icha/ittii** and etc. For examples: obboleessa 'brother'/obboleettii 'sister'.

Barataa 'student (m.)'/barattuu 'student (f.)'.

Bareedaa 'handsome (m.)'/bareedduu 'beautiful (f.)'.

Namicha ‘the man (m.)’/namittii ‘the woman (f.)’.

3.4.1.3 Definiteness

Some language use article or other determiners to express definiteness. For instance, English uses definite article ‘the’ or determiners such as ‘all’, ‘this’, ‘that’ and so on, to express definiteness. English has also indefinite article a/an.

Although Afaan Oromo has no indefinite articles (corresponding to English a/an), it indicates definiteness (English corresponding the) with suffixes on the noun: **-icha** for masculine nouns and - **ittii** for feminine nouns. Nouns inflected by these suffixes drops end vowel before adding these suffixes. For example: Nama ‘humanbeing’ becomes namicha 'the man' or namittii ‘the woman’. Afaan Oromo animate nouns that can take either gender, these definite suffixes may indicate the intended gender: qaalluu 'priest', qaallicha 'the priest (masculine)', qallittii 'the priest (feminine)'. Unlike in English, in Afaan Oromo definite suffixes do not to co-occur with the plural suffixes. Thus, definiteness suffixes of Afaan Oromo refer the **singular markers**.

Definiteness in Afaan Oromo also expressed using demonstrative pronouns like kun (this), sun (that).

3.4.1.4 Cases

Case is a grammatical category of nouns, pronouns or determiners that indicates the nature of their relationship to the verb [3] or other function-like expression in sentences. Olani[8] defined case as a grammatical relationship of nouns or pronouns to other words in a sentence. Case marking is implemented in various languages, in various ways specifically word order, inflection and adposition. For instance, English has largely lost its inflected case system except personal pronouns which is inflected for three cases (subject form, absolutive/accusative form and possessive form), and uses preposition (e.g., to school, from school, for Trump) or word order to mark cases. **The man** is here vs. I saw **the man**. Thus number of cases varies from language to language [3].

Afaan Oromo has extensive case systems, with nouns, pronouns, adjectives, and determiners all inflecting (usually by means of different suffixes) to indicate their case. Fundamentally, when we come to noun cases, Afaan Oromo case is based on changes to the noun to indicate the noun's role in the sentence. It is dominated by inflectional case system. Inflectional patterns may depend

on a variety of factors, such as gender, number, and phonological environment. In addition to, inflection case system Afaan Oromo is also characterized by *Adpositional*[21] case system. The case is understood by considering placement of the noun and the syntactic function it conveys.

Although in some literatures shows Afaan Oromo nominals (nouns, pronouns and adjectives) are inflected for four [3] to six [71] cases; they are subject to the following case forms.

1- Base form: Ø inflected: NULL inflected

Base form (noun base form) is usually given in dictionaries because it is that form of a noun, adjective or pronoun that does not have any case ending or suffix. It is used for isolated citation, a direct object, a predicate nominal and to express the different oblique cases for subject in focus.

2- Nominative case: subject form

The nominative case is used for nouns that are the subjects of clauses. It is a form of a noun, pronoun, or adjective, which is used for a subject if it is not in focus. The nominative case is marked by four different morphs of allomorphic variation occurring in complementary distribution. The allomorphs for the nominative case are -n, -ni, -i and Ø. The difference in the phonological realization of the nominative case markers arises from the phonological nature of the nouns [8].

- The marker **-n** occurs after a *terminating long vowel* of a given noun if the noun ends in a long vowel.
- Nouns *ending in short vowels* with the *preceding single consonant* drop the final vowel and add **–ni** (its variations: following certain consonants, assimilation changes either the n or that consonant) to form the nominative.
- Nouns *ending in short vowel* is *preceded by two consonants or a geminated consonant*, **-i** is suffixed.
- If the noun ends in certain consonants, the nominative is identical to the base form (**Ø** is suffixed).

[21] In **Adpositional** case systems, nouns are accompanied by words that mark case adpositions classified into preposition and postposition.

Base form	Inflected (Nominative) Forms	Meanings
Nama	Namni	Man
Mana	Manni	House
Maqaa	Maqaan	Name
Arba	Arbi	Elephant
Hirriba	Hirribni	sleep
Madda	Maddi	Source
Ilkaan	Ilkaan	Teeth/tooth
Sagal	Sagal	Nine
Gosa	*Gosti	Type/clan
Lafa	*lafti	Earth
Bifa	*bifti	Color
Bara	*barri	Year
Dhiira	*dhiirri	Male
Dhara	*dharri	False
Gaala	*Gaalli	Camel
Intala	*Intalli/intalti	Girl

Table 3.5 Nominative/Subjective case makers

*The nominative case allomorph -**ni** undergoes phonological processes and gets changed to –**ti** or *doubled last consonant* as in ***barri*** and ***gaalli***.

3- Accusative/Absolutive case: object form

Absolutive case, the unmarked noun in Oromo is an underlying/inherent noun that occurs in the object position without an inflectional suffix. Absolutive case is considered as the primitive form (base form) of nouns. Sometimes an object seems to be marked if the underlying nouns end with certain consonant. For example, Caalaan (Caalaa-nom.) Galaan**iin** (Galaan-abs) waame 'Chala called Gelan'.

4- Dative

The dative case is an *indirect* object used for nouns that represent the ***recipient*** 'to' or the ***benefactor*** 'for' of an event. It takes the position before or after the direct object with the function of telling "for whom" or "to whom" the action is done as semantic criteria [8].

The dative case can be expressed by: -

a) Lengthening of a short final vowel,
b) Lengthening of a short final vowel and adding of a suffix **-f**, **-dhaa** or **-dhaaf** (to nouns with final long vowel),
c) Adding of the suffix **-ii** to nouns terminating in consonant,
d) Adding of the suffix **-(tii)f** to a genitive construction, and
e) Adding of the suffix **–tti** (irrespective of the spelling of the noun).

There are two fundamental inflectional morphemes to mark dative case in Afaan Oromo: ***-f*** and ***-tti***. All lengthening of final vowel, *dhaa*, *dhaaf*, and *tiif* are the variations of f.

The dative form of a verb infinitive (which acts like a noun in Afan Oromo) indicates purpose.

The following Examples show each of the types.

Namich**aa** buna fidi! (Bring coffee for the man!)

Namich**aaf** buna fidi! (Bring coffee for the man!)

Saree**dhaa**(**f**) foon kenni! (Give the meat for the dog)

Nyaata muc**aaf** kenni! (Give the food for baby)

Mee loon**ii** okaa kenni! (Please give fodder for the cattle!)

Mee looniif okaa haami! (Please cut fodder for the cattle!)

Mana barumsaa**(tii)f** (for the school).

Qarshii caala**tti** kenni! (Give the money to Chala)

Uffata kee isa**tti** kenni! (Give your cloth to him)

5- Genitive

The genitive case marked on nouns (nominals, [8]) for indication of possession. Of course, genitive case is broader than possession inclusive of purpose, source, reference, etc. In Afaan Oromo, the genitive is characterized by the sequence: ***Possessed-possessor.*** It doesn't have special marker [71].

In [8], [71] genitive case is formed in two ways:-

a. by succeeding relative particle **kan/tan** and lengthening **the last vowel of possessor** nouns (or suffixing **-i** to final consonant of the possessor noun) or by leaving a final long vowel of possessor nouns unchanged and
b. By **putting side by side** the thing *possessed* and the *possessor* in that sequence and lengthening the final vowel of the possessor if it is short (or suffixing **-i** after consonant).

If the noun has more than one qualifier normally, *only the last part* of the noun phrase has the genitive marker. Vowel length is the marker of genitive case on a noun as in the table below.

Base form		Subject form		**Meaning**
Possessed	***Possessor***	***Possessed***	***Possessor***	
Mana	nam**aa**	manni	namaa	Somebody's house
Mana	namich**aa**	manni	namichaa	The house of the man
Kitaaba	barat**aa**	Kitaabni	barataa	Student's book
Mana	aanan**ii**	Manni	aananii	Milk's house
Kan	*is**aa***	*kan*	*Isaa*	*His*
Mana	nama san**aa**	Manni	Nama sanaa	The house of that man

Table 3.6 Genitive case makers examples in Oromoo nouns

6- Instrumental

The instrumental case is used for nouns that represent the instrument ('with'), the means ('by'), the agent ('by'), the reason, or the time of an event. In Afaan Oromo instrumental case is marked by **–n** and its allomorphs. The instrumental case is marked in the following ways:-

a) Lengthening of a short final vowel and adding of the suffix **–n** (or adding of the suffix **–n** to nouns with final long vowel),
b) Adding of the suffix **-dhaan** to nouns with final long vowel,
c) Adding of the suffix **-iin** to nouns terminating in consonant, and
d) Adding of the suffix **-tiin** to a genitive construction.

For instances: -

Base form	**Instrumental case**	**Glossary**
Miila	Miila**an**	By foot, with foot
Harka	Harka**an**	By hand, with hand

Base form	Instrumental case	Glossary
Ilkaan	Ilkaan**iin**	By teeth, with teeth
Halkan	Halkan**iin**	At night
Afaan Oromo	Afaan Oromo**tiin**	In Afaan Oromo
Eboo	Eboo**dhaan**/eboo**n**	By spear
Yeeroo	Yeeroo**dhaan**/yeeroo**n**	On time

Table 3.7 Examples of instrumental case makers of nouns

7- Ablative

The ablative is used to represent the ***source of an event***; it corresponds closely to English ***from***. It expresses the source, origin or from where a movement begins. It is formed in the following ways:

i. **Vowel lengthening**: when a word ends in a short vowel, this vowel is lengthened.

ii. **-dhaa**: When the word ends in a long vowel, -dhaa is added. The interesting point is that such form distinguishes ablative case from the object form of nouns in its morphology as in Adaamaa dhufe 'he came to Adama' and Adaamaadhaa dhufe 'he came from Adama'.

iii. –**ii:** When the word ends in a consonant, -ii is added (as for the genitive: tii).

iv. –**tii**: Following a noun in the genitive, -tii is added.

For instances:

Base form	Ablative case	Glossary
Keessa (inside, in)	Kess**aa**	From inside
Ala (outside, out)	Al**aa**	From outside
Jimma	Jimm**aa**	From Jimma
Gabaa	Gabaa**dhaa**	Form market
Harar	Harar**ii**	From Harar
Gadab	Gadabii	From Gadab
Mana (house) bunaa (coffe:genitive)	Mana bunaa**tii**	From café
Biyya alaa	Biyya alaa**tii**	From foreign country

Table 3.8 Examples of Afaan Oromo Nouns' ablative case makers

An alternative to the ablative is the postposition **irraa** 'from' whose initial vowel may be dropped in the process. For instance, gabaa 'market', **gabaa irraa**, **gabaarraa** 'from market'

8- Locative

The locative is used for nouns that represent general locations of events, goals or states, roughly corresponds English '***at'.*** For more specific locations, Afaan Oromo uses adpositions [3]. The locative case is marked by suffix –**tti**.

For instances,

Base form	Locative case	Glossary
Harka	Harka**tti**	In hand
Arsii	Arsii**tti**	In Arsii
Guyyaa	Guyyaa**tti**	Per day
Mana	Mana**tti**	In house
Nama	Nama**tti**	To man
Aangoo	Aangoo**tti**	By authority

Table 3.9 Examples of Afaan Oromo Nouns' locative case makers

9- Vocative

Vocative is a verbal means of calling attention or incorporating strong feelings. In Afaan Oromo, there are various ways of marking the vocative case. The common ones are using the word '**yaa'** and suffix **-na.** The suffix -na which is our focus case is appended to a noun which is two syllabic and ending in short vowel with harmonic occurrence of vowels as in nama**na** 'you guy!' and jara**na** 'you guys!'. Its full word form '**nana'** is used after nouns that end in long vowel as in gurbaa nana 'you boy!'.

3.4.2 Adjective Inflections

An adjective is a word that describes a noun, giving extra information about it. For example:

An adjective modifies a noun or a pronoun by describing, identifying, or quantifying words. In Afaan Oromo, an adjective usually follows the noun or the pronoun that it modifies.

In Afaan Oromo, the principles of nouns inflection can apply to the adjectives. Actually, word classes of nouns, pronouns and adjectives can be described under nominals [8].

3.4.3 Verb Inflections

Verb is a word that describes what a person or thing does or what happens. It is the most important part of a sentence that says something about the subject of a sentence, expresses an action (jump, stop, search), events (e.g., snow, happen), changes (e.g., involve, shrink, melt, widen) or states of being (e.g. is, be, have). Without a verb, a group of words cannot be a clause or sentence. Olani [41] classified the verbs in Oromo into three types as *action*/*stative verbs*, *auxiliary verbs* and *copula*. The action verbs can be used in different derivational forms like causative and passive constructions. Auxiliary verbs, which occur as helping verbs, can be considered as action/main verbs when they are used in the absence of another action verb in a sentence. They are functioning as helping verbs being with other action verbs in a sentence. The invariable particles functioning as copula exist in the language. They are **dha** (and its variation -**ti**) and its negative form **miti**.

Verbs undergo several *inherent* and *agreement* inflection and thus complexity of conjugational[22] occurrences is noticeable. According to many literatures [8], [20], [41], [60] in most languages including Afaan Oromo the verb shows greater morphological complexity than any other word class.

In Abeshu [3] there two criteria to identify Afaan Oromo verbs from other word classes: *syntax* and *morphology*. Syntactically Afaan Oromo verbs function as predicates in simple sentences and occur in the final positions of a sentence. In morphology case Afaan Oromo verbs shows the agreement with the number, gender and/or person of the subject and expression of the tense, aspect of the verb. However, we can also identify verbs from other word class by looking its *meaning* (*semantics*).

Generation of syntactically and semantically correct sentences requires appropriate choice among the different forms of verbs. Thus, this section mainly focuses on various inflectional forms of verbs of Afaan Oromo in terms of root or stems and their affixes.

[22] Conjugation refers to inflectional verb forms [8]

3.4.3.1 Derivations of verbs stem

Afaan Oromo verb basic stems/roots can be the basis for four derived stems [3], [8], [71]: passive, causative, autobenefactive and frequentative/intensive. These derived stems then inflected for both inherent verbs inflections and agreement inflections like that of primary verbs. Passive, causative and Autobenefactive are formed with addition of a suffix to the stem/root, yielding the derived stem that the inflectional suffixes are added to while the intensive stem is formed by reduplicating the first consonant and vowel of the first syllable. The derived stems may be formed from all verbs the meaning of which permits it. They may be also derived from nouns and adjectives. For instance: adaa**chuu** 'becoming white /autobenefactive', guraach**essuu** 'make black /causative', guraach**omuu** 'to become black'

1. **Autobenefactive**

The Afaan Oromo autobenefactive is formed by adding **–adh** to the verb root. This stem has the function to express an action done for the benefit of the agent himself.

The conjugation of a derived verb is irregular in the other persons of the present and past (-**dh** in the stem changes to **–t** for 2nd and 3rd persons and -**n** for 1st person plural) and in the singular imperative (the suffix is -**u** rather than -i). Infinitive and participles are always formed with -(a)ch, while the imperative forms have -(a)(a)**dh** instead of -(a)ch. For instances, the following Table shows the – **adh** varieties of autobenefactive verb stems derived from bit- 'buy', arg- 'find/get' and waam- 'call up on'

Person	Bitachuu (to buy)	Argachuu (to find)	Waamachuu (to call up on)
Sg. 1.p	Bit**adh**a	Arg**adh**a	Waam**adh**a
Sg. 2.p	Bit**at**ta	Arg**at**ta	Waam**at**ta
Sg. 3.p.m	Bit**at**a	Arg**at**a	Waam**at**a
Sg. 3.p.f	Bit**at**ti	Arg**at**ti	Waam**at**ti
Pl. 1.p	Bit**an**na	Arg**an**na	Waam**an**na
Pl. 2.p	Bit**at**tan	Arg**at**tan	Waam**at**tan
Pl. 3.p	Bit**at**an	Arg**at**an	Waam**at**an

Table 3.10 regular autobenefactive verbs inflections for all persons and numbers.

2. **Passive (usually called Passive voices)**

The Afaan Oromo passive corresponds closely to the English passive in function. It is formed by adding -**am** to the verb root. The resulting stem is conjugated regularly [3], [71] as in beek**am**- (be known) which is derived from beek- (know). For example:

Voice	Root	Marker	Inflected form	Glossary
Active	Kut-	-	Kute	Cut
	Bit-	-	Bite	Bought
Passive	Kut-	-am	Kutame	was cut
	Bit-	-am	Bitame	Was bought

Table 3.11 Active and passive verbs

3. **Causative**

The Afaan Oromo causative of a verb V corresponds to English expressions such as 'cause V', 'make V', 'let V'. With intransitive verbs, it has a transitivizing function. It is formed by adding -**s**, -**sis**, or -**siis** to the verb stem, except that stems ending in -l add -ch. Verbs whose stems end in **'** (apostrophe) or '**hudhaa'** drop this consonant and may lengthen the preceding vowel before adding -s. For instance, beek- 'know', beeksis- 'cause to know, inform', beeksifne 'we informed'; ka'- 'go up, get up', kaas- 'pick up', kaasis- 'make to pick up'; gal- 'enter', galch- 'put in', galchiti 'she puts in'; bar- 'learn', barsiis- 'teach', nan barsiisa 'I teach'.

4. **Intensive/frequentative**

It is formed by duplication of the initial consonant and the following vowel, geminating the consonant.

Example Waamuu (to call, invite) **waww**aamuu (to call intensively)

5. **Complex Derived Verb Stems**

The derived stems can be combined with each other in different sequences.

Example: Waamuu (to call, invite), waamamuu (to be called), waamsisuu (making to call) and etc.

3.4.3.2 Agreement properties of Verbs Inflection

Agreement properties indicate inflection of a word class for properties out of its members. These properties are dependents on persons (first, second and third), number (male and female), and gender (singular and plural) of subjects. Like in many of other languages in Afaan Oromo Genders are distinctly identified only in third person singular whereas numbers are indicated in all persons. Verbs are marked by agreement morpheme **–t-** for 2nd persons (singular and plural) and 3rd person feminine. Plural numbers of persons are marked by suffix **–n (an).**

3.4.3.3 Inherent properties of verbs Inflection

Afaan Oromo verb consists minimally of a stem, representing the lexical meaning of the verb, and a suffix, representing aspects, mood or voice [8] and subject agreement (section 3.4.3.3). Aspects, moods, and voices with some indications of tenses are inherent properties of verb[8]. These properties are the basic members of a word class triggering inflection on that word class.

1. **Aspect**

Aspectual property of Afaan Oromo verbs indicates their tenses. From the three major tenses present, past and future, Afaan Oromo mainly identifies between past and non-past tense in its morphology because the morphological markers do not distinguish each tense types (present tense and future tense).

Aspect morphologically distinguishes between completeness and incompleteness of an action. It is about the situation, time and duration of an event.

In Afaan Oromo, there are two kinds of aspects namely perfective and imperfective aspects indicated by different set of suffixes. Perfective aspect indicates the action that has been completed and imperfective aspect indicates the action that has not been completed yet but can be completed at any time. Perfectness indicates an action completed at a specific time in the past whereas imperfectness is connected with an action in process or in progress.

The perfective aspect is indicated by the suffixes **–e, -ne, -te, -tan**, and the imperfective aspect is indicated by the suffixes **–a, -na, -ta, - ti, -tu, -u**.

Aspects	Person	Root deem- 'go'	Agreement		Asp. indicator	Inflected word
			Pers.	Num.		
Perfective	1sing.	deem-	-	-	-e	deeme
	2sing.	deem-	-t	-	-e	Deemte
	3sing.m.	deem-	-	-	-e	Deeme
	3sing.f.	deem-	-t	-	-e	Deemte
	1pl.	deem-	-	-n	-e	Deemne
	2pl.	deem-	-t	-an		Deemtan
	3pl.	deem-	-	-an		Deeman
Imperfective	1sing.	deem-	-	-	-a	Deema
	2sing.	deem-	-t	-	-a	Deemta
	3sing.m.	deem-	-	-	-a	Deema
	3sing.f.	deem-	-t	-	-i	Deemti
	1pl.	deem-	-	-n	-a	Deemna
	2pl.	deem-	-t	-	-u	Deemtu
	3pl.	deem-	-	-	-u	Deemu

Table 3.12 Verbs aspect inflection for Afaan Oromo

2. Mood

Mood is the attitude of the speaker towards an utterance [8]. It is particular way of speaking. Afaan Oromo has several types of moods among those the most applicable ones are indicative, imperative and jussive. While Indicative can be perfective or imperfective in terms of aspects, imperative and jussive are imperfective.

a. Indicative mood: - the way of making statements and asking questions (yes/no question forms). In question form, the final vowel is lengthened along with intonational relevance.

Example: - Inni mana baruumsaa deeme. He went to school

Inni mana baruumsaa deemee? Did he go to school?

b. Imperative mood: is the way of giving order or command for second person singular '**ati**' and plural '**isin**'. Imperatives are marked by dependent morphemes i, u, and aa. For instances

Stem	Sing. Imperative	Pl. imperative	Glossary
Deem	Deem**i**	Deem**aa**	Go
Nyaat	Nyaadh**u**	Nyaadh**aa**	Eat
*** dhuf**	Koott**u**	Koott**aa**	Come
Bitadh	Bitadh**u**	Bitadh**aa**	Buy +autobenefactive

Table 3.13 Imperative mood for Afaan Oromo verbs

Imperative verbs are also inflected by suppletive form using a completely different word of inflection as in *dhuf- above which is not used in imperative construction but in the indicative form.

From the above table the imperative singular stems are formed by means of the suffix **–i** and **–u** (all autobenefactive verbs) while imperative plural of all stems is formed by means of **-aa**.

All negative imperatives are formed by means of suffixes **-in** for singular and **-inaa** for plural preceded by preverbal particle **hin**.

Example: - hin deemin 'don't go +singular', hin deeminaa 'don't go +plural'

c. Jussive mood: jussive mood is the way of giving permission or command for the both second persons for the action done by third persons (inni, ishiin, and isaan) or first person plural nuti. In Afaan Oromo, jussive mood shares semantic and morphological features with imperative mood. However, jussive mood is marked by the preverbal particle **haa** and the dependent morphemes **-u** or **–n (ni)** on the verb.

Examples

Persons	Stem	Jussive with preverbal particle 'haa'	Gloss
3sing. M	Deem	Haa deem**u**	Let him go
3sing. F	Deem	Haa deem**tu**	Let her go
3pl.	Deem	Haa deem**ani**	Let them go
2pl.	Deem	Haa deem**nu**	Let us go

Table 3.14 Jussive mood for Afaan Oromo verbs

The negation of jussive form is formed by preverbal particle with connection of suffix **-in** for all third persons as in ***hin*** deem***in***. There is no negation jussive form for first person plural.

3. **Voice**

Voice refers if the subject performs or receives the action indicated by the verb. When the subject performs the action, the voice is **active** whereas the form in which the subject receives the action is **passive** voice. Subject in active becomes object in the passive form and vice versa.

Using sentence types in which the verb form is changed for the purpose of such grammatical function is inflectional. Passive formation in Afaan Oromo is purely morphological as it is formed by adding the morpheme **-am-** on transitive verbs. The Afaan Oromo passive corresponds closely to the English passive in function. In both perfective and imperfective aspects, the morpheme **-am**- invariably marks the passive voice in contrast with the unmarked active form [8].

Examples:- beekte 'you/she knew' beek**am**te 'you were known or she was known'

3.4.4 Adverb Inflections

An adverb is a word that is used to give information about a verb, adjective, or other adverb. They can make the meaning of a verb, adjective, or other adverb, and usually precede the verbs they modify or describe. An adverb indicates time, manner, place, cause, or degree and answers questions such as how, when?, Where? and how much?.

3.5 Afaan Oromo morphotactic and Morphophonemic Properties

In Afaan Oromo, affixes are the most grammar makers. Language's morphology is characterized by patterns of occurrences of morphemes (stem-affixes) and some orthographic alternation rules governing the form that morphemes take in particular environments. The alternation can happen either at the stem-affix intersection points or within the stem itself. The constraints imposed on the order in which morphemes are combined are governed by **morphotactics**. Morphotactics is responsible for governing the rules for the combination of morphemes into larger entities [3].

Orthographic alternation rule is governed by **morphophonemic**. Morphophonemic is influences of phonology to change shape of morphemes at boundary when concatenated to form words. It indicates the change that takes place between the boundary of stems and inflectional or derivational suffixes. The rules of morphophonemics in Afaan Oromo operate on consonant-consonant sequences, consonant-vowel or vowel-vowel sequences across morpheme boundaries.

Morphophonemic processes such as assimilation, deletion, epenthesis, and reduplication are prevalent in Afaan Oromo.

3.5.1 Assimilation

The phonemes that come next to each other at morpheme boundary may take the form of the previous or next letter of morpheme because of their similarity to some extent. The following table summarizes these changes with examples.

Combinations Of Phonemes	Results	Examples
d+s	ch	duud +sa =duucha
dh+s	ch	nyaadh+sisa= nyaachisa
dh+n	n	fuudh+na = fuuna
d+n	nn	did+na = dinna
t+n	nn	dhaloot+ni = dhaloonni
t+ch	ch	hojjet+siisa = hojjechiisa
x+s	cc	fix+siise = ficcisiise
d+s	ch	fid+siise = fichisiise
t+dh	dh	barat+dhu = baradhu
dh+t	t	fuudh+tan = fuutan
l+s	lch	awwaal+sise = awwaalchise
b+t	bd	qab+da= qabda
s+t	ft	baas+te = baste
d+t	dd	did+te = didde
l+n	ll	gal+ne = galle
g+t	gd	fig+te = fiigde
x+t	xx	fix+te = fixxe
c+t	cc	boc+te = bocce
j+t	jj	ifaaj+te = ifaajje
r+n	rr	abaar+ne = abaarre
s+n	fn	baas+ne = baafne

Table 3.15 Assimilations (taken from[3])

3.5.2 Deletion

To avoid the inconvenience of speaking, phonemes at morphemes boundaries are deleted.

Example: mana+oota = manoota ‘houses’

In verbs, deletion usually takes place in stems ending with ‘h, dh, hudhaa(‘) ’ as in hodh+te=hoote and bah+te=baate

3.5.3 Epenthesis

Epenthesis is insertion of a vowel to avoid impermissible occurrence of consonant based on phonotactic constraints of the language.

Examples: mars+ne = marsine ‘we revolved’ sirb+te = sirbite ‘you-$_{sing}$/she sung’

3.6 Summary

This chapter has discussed the morphological features of Afaan Oromo. Morphological behavior of Afaan Oromo yields significant challenges in NLP since it is morphologically derivational, inflectional and fusional [8]. Fusional morphology (also called inflectional morphology) is a term, which is used for a morphological system in which one morpheme, usually an inflectional affix, expresses several different meanings or grammatical functions. Afaan Oromo is suffixal language, which is characterized by concatenative morphology. It also used some preverbal particles[8] such as **hin** (negation proclitic) and **haa** (for jussive mood) in connection with some dependent suffixes. In inflectional morphology, inflectional suffixes are combined with stem usually resulting in a word of the same class as the original stem. The combination orders and boundary changes of morphemes are governed by morphotactics and morphophonemics rules respectively. While Afaan Oromo nouns are inflected to indicate *number*, *gender*, *definiteness,* and *case,* verbs are inflected to indicate both *inherent* (aspects, mood, and polarity (affirmative and negative)) and *agreement* (person, number and gender) properties. Verbs also inflected for various derivational categories such as voice, causative, autobenefactive and intensive/frequentative. In Afaan Oromo, the verb shows greater morphological complexity than any other word class. The next chapter will discuss the design and development of Morphological analysis prototype, which is the principal part of this study.

CHAPTER 4 DESIGN AND DEVELOPMENT

4.1 Overview

The previous two chapters discussed the approaches to develop morphological analyzers and Afaan Oromo morphological features in detail. This chapter discusses the design and implementation of Afaan Oromo Morphological Analyzer and its subcomponents starting from data preparation to evaluation of the system. In the following sections, the details about the techniques and the models developed for the proposed solution are described. In section 5.2 The Proposed Architecture of Afaan Oromo Morphological Analysis is presented. Section 5.3 presents data collection and preparation process including noise entity removal. Section 5.4 shows the algorithms designed for words segmentation (i.e., considered as word normalization and standardization processes for the main module, Morphological Analysis) and the last section demonstrates the training and model building for the language morphological analysis.

4.2 The Proposed Architecture of Afaan Oromo Morphological Analysis

The morphological analysis of Afaan Oromo words involves many processing steps. These steps, described below, are executed sequentially where each step depends on the results of previous step. For each Afaan Oromo word, a special module divides the word into two parts: a valid stem/root and suffixes. The full analysis means all possible analyses of the word such as, all possible root/stems, affixes, the morphological features of each morphemes represented by a morphological tag.

The proposed morphological analyzer has two main phases: Training Phase and Analysis Phase.

The following figure (fig. 4.1) depicts the new proposed Afaan Oromo Morphological analyser architecture. Training (left side) and testing (write-side) phase separated by dashed-line. The components of both phases were discussed thoroughly in following sections.

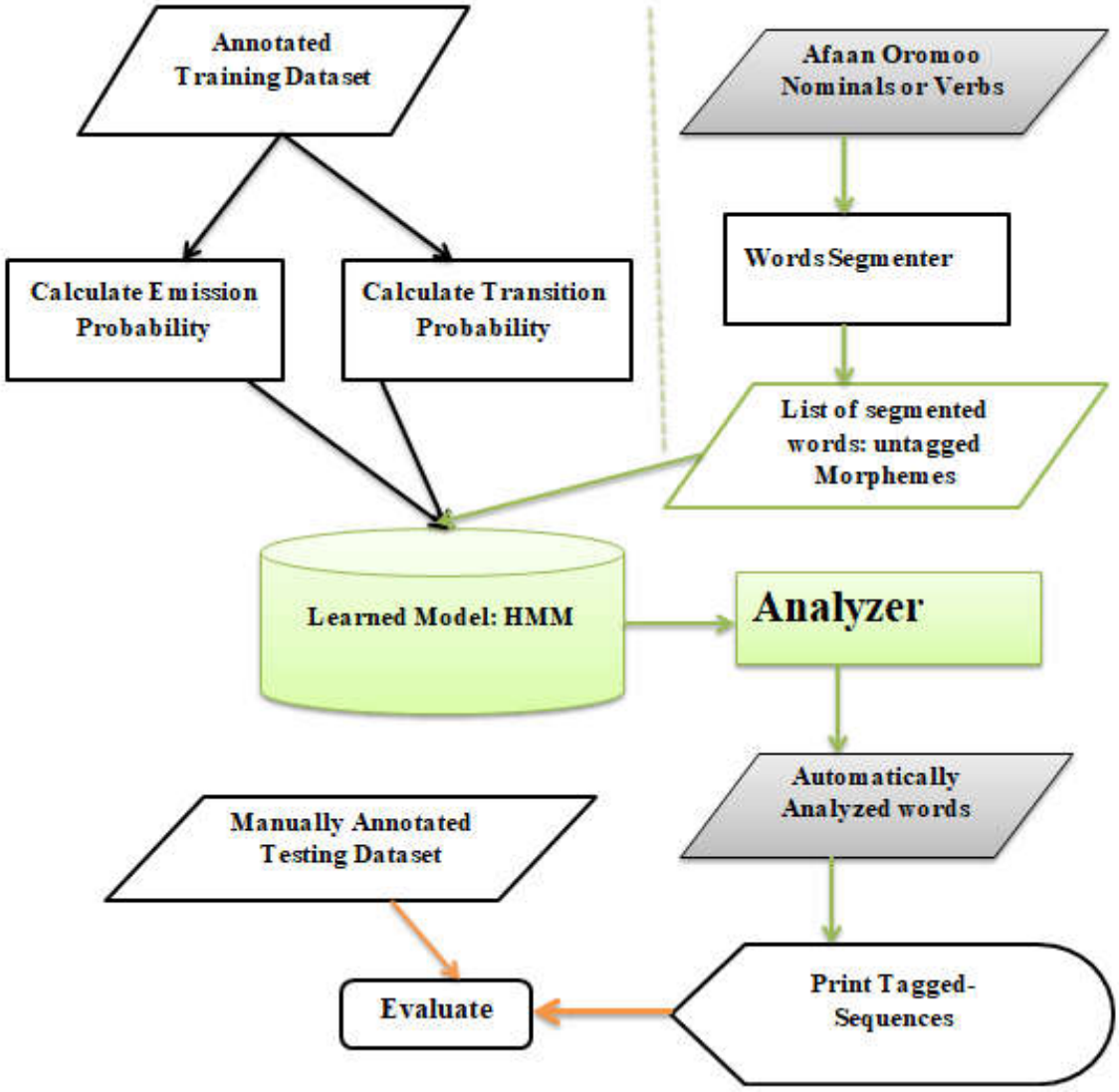


Fig 4.1 General architecture of Afaan Oromo Morphological Analyser System

4.2.1 Training Phase

The Analyzer trained to learn based on the training data set using stochastic concepts to produce trained model. The Training phase contains input data, calculated probabilities (Emission probability and Transition probability), and Trained Model (HMM).

Input Data is 90% of manually labeled corpus (4,800 nouns and 4,200 verbs) which is used for training the model. It is called training data.

Calculated Probabilities our model uses two major probabilities as source of knowledge:

Emission probabilities and Tag transition probabilities

1. *Emission Probability* is the probability of a word appearing depends only on its *own tag* and is independent of neighboring words and tags. It is morpheme likelihood probability.

2. *Transition Probability* is the probability of a tag depends only on the *previous tag (bigram HMM)* occurred rather than the entire previous tag sequence i.e., shows Markov Property.

   The Train Model (HMM) is the encoded Transition probabilities Matrix and Emission probabilities Matrix. For instance, the following figure shows the Transition Probability Matrix Afaan Oromo Nouns Morpheme Tags.

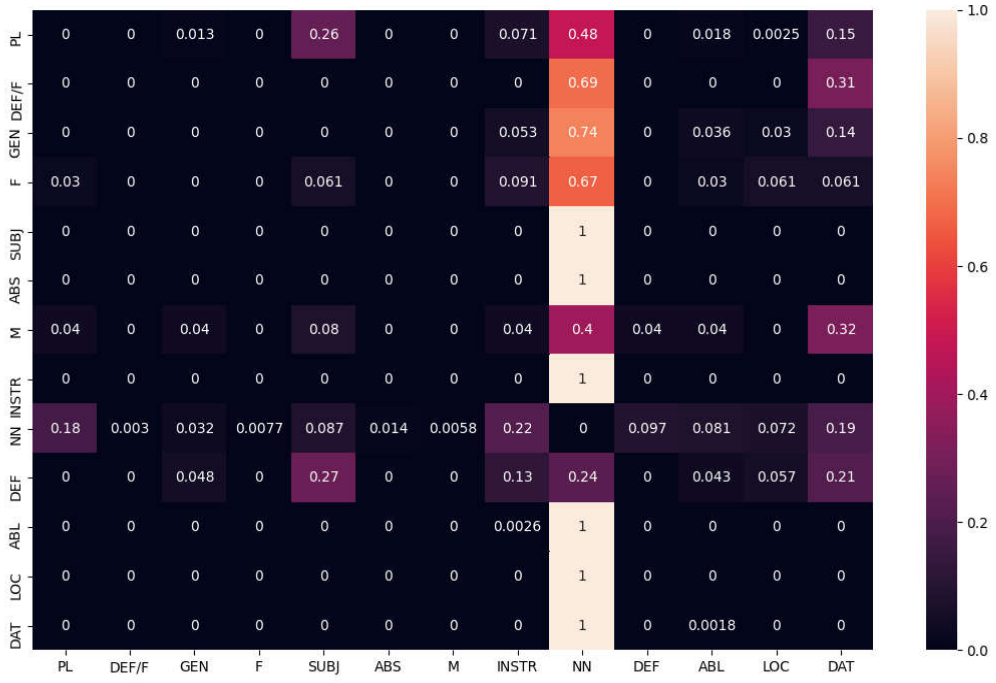


Fig 4.2 Transition Matrix of Afaan Oromo Nouns Morpheme Tags.

The above transition p(ti|ti-1) matrix indicates the probability of all possible sequences of tags. In the figure each row represents the previous tag, t_{i-1}, while each column represents the current tag t_i. In this matrix table count starts from 0, rows counted from up to bottom and column counted from left to right. For instance probability at (row 0, column 0) which is p(PL,PL) and indicates the first upper cell at the left side of the table is 0. This means there no probability of getting one plural indicator after another plural indicator. Accordingly probability of getting instrumental after nominal root p(INSTR,NN) which is at row 8 (9^{th} row or NN) and column 7 (8^{th} column or INSTR), 0.22. The 9^{th} row (row of NN is a root of the word in our case) shows each morpheme

tag may occur after the Nominal and the 9th column shows every morpheme may occur at the end of every word and immediately before the root of the next word root in a given sequence. The probability of occurrence is presented in number. Full probability (probability 1) represents no morpheme occur after a morpheme tag that means that tag is always occurred at the end of a given word. Zero probability represents we cannot get t_1 after t_{1-1}. The probability between 0 and 1 ($0<P<1$) show that, we may get another morpheme tag and root of the next word after a given tag.

4.2.2 Analysis Phase

In Analysis phase unlabeled corpus of Afaan Oromo nouns or verbs are taken as input; then for each word, a segmenter divides the word into two parts: a valid root and suffixes. After that, the analyzer takes segmented word and labels it with a morphological tag depending on the knowledge of the model. The Analysis Phase contains Input data, a segmenter, an analyzer, and output data.

Input Data is 10% of unlabeled corpus (4,800 nouns and 4,200 verbs) which is used for testing the analyzer. It is the complement training data, also called testing data.

Segmenter is one of the most important module in the morphological analysis; divides word in input data and produce sequence morphemes to feed the analyzer for further process (to analysis). Our morphological segmenter is developed using rule-based approach (details discussed in section 4.4).

Analyzer uses trained Model HMM (Transition Matrix, Emission Matrix) and take a sequence of morphemes in w*ords* as input, to find the most probable sequence of corresponding *morphemes' tags.* It uses the trained model as a basis for mapping input to output.

Output Data is the annotated version input data that is automatically labeled by the analyzer.

Evaluator is a simple module used to test the performance of the developed model by comparing manually annotated test data with automatically analyzed data by the developed analyzer and it presents the result in accuracy.

4.3 Data Collection and preparation

Every NLP, especially morphological analyser needs large corpus if the development follows machine-learning approach. However, such a large Afaan Oromo corpus is not readily available. Thus, we collected raw data from Fana Broadcasting Corporation Afaan Oromo and BBC Afaan Oromo publically online archives from April 2020 to September 2020. Because; they are considered as addressing different issues of the community such as social, economic, technological, political, religions etc. This will reduce the probability of making the corpus biased toward some specific words and domains that do not appear in everyday life[20]. After data is collected, different data preprocessing techniques (cleaning and organizing) applied on the raw data to make it suitable for a building and training Machine Learning models. In addition, it is used for model evaluation (model testing). Data preprocessing techniques of the data includes:

1. Noisy data (such as digits, special characters and punctuation) removal
2. Stopwords removals and
3. Selection of names and verbs (actually this done manually in our case to grant the selection accuracy)

4.3.1 Noisy Data

Noisy data such as digits, special characters, and punctuation marks used in both Afaan Oromo and English languages are the same and are used for the same purpose with the exception of apostrophe. Apostrophe mark (‘) in English shows possession but in Afaan Oromo it is used in writing to represent a glitch (called hudhaa) sound. It plays an important role in the Afaan Oromo reading and writing system. For example, it is used to write the word in which most of the time two vowels are appeared together like “du’a” to mean, (“die”) with the exception of some words like “har’a” to mean “today” and “mul’ataa” to mean “the visible one” which is identified from the sound created.

Appendix A shows python script we used for the removal of noisy data (punctuation and digits).

4.3.2 Stop words

Stop words are common words that are present in text but contribute less in the meaning of a sentence. Such words are not at all important or do not contain important significance for the purpose of information retrieval or natural language processing. Usually, these words are filtered

out from search queries because they return a vast amount of unnecessary information. They are functional words that do not contribute that much for the meaning of the sentence.

Actually, stop words we used were compiled from two sources (see Appendix C):

1. 78 adopted from Debela Tesfaye [71] master's thesis.
2. 264 compiled manually from our corpus (noisy data free corpus) and approved by linguist.

4.3.3 Nominals and verbs selection

After data collected from the online archives that raise these wide ranges of issues are cleaned, the balanced corpus of Afaan Oromo nouns and verbs is prepared manually. Balanced corpus is a corpus that represents the words that are used in a language and needed to process natural language processing tasks[71], [74] like morphological analysis. Then both corpuses (nouns and verbs) are segmented and divided into two data sets for analysis. Namely, training data and testing data.

The statistics of the raw collected data and selected nominals and verbs are presented in the following table.

Nº	Data source	Nº of Tokens (items) in raw collected data	Nº of Tokens after noises and stop words removal	Nº of non-redundant words in cleaned corpus	Nº of unique in merged file	Nº of nouns and adjectives in %	Nº of verbs in %
1	FBC Afaan Oromo	167,916	135,772	22,674	48,604	48.9%	44.2%
2	BBC Afaan Oromo	288,992	226,950	34,675			
3	**Total**	**456,908**	**362,722**	**57,349**	**48,604**		

Table 4.1 statistics of Afaan Oromo collected and cleaned data

4.4 Algorithms Design for Word Segmentation: Afaan Oromo words segmenter

In Afaan Oromo, segmentation of words into a stem (a root of inflected word) and zero or more suffixes is necessary to achieve good performance when building a morphological analysis system. Those segments are the result of running the segmentation models on the input raw

Afaan Oromo text words. The segments become the unit of analysis when doing classification where it can be a morph segment (a stem or a suffix) or a word when no segmentation is conducted.

Stem types of Afaan Oromo verbs can be causative, passive, intensive and autobenefactive.

Afaan Oromo nouns suffixes indicate numbers, definiteness, genders and cases (nominative, accusative/absolutive, dative, genitive, instrumental, ablative, locative or locative) while verbs suffixes indicate persons, aspects (perfective or imperfective) and moods (indicative, imperative or jussive).

Thus, to make our prototype completely automatic, the segmentation module is developed using rule-based approach for its accuracy. Rule based approach is also simple if the developer has enough knowledge about the language as it takes the properties of the language into account.

Our module takes a word(s) and list of possible morphemes as input and segments into its root/stem and affixes according to a rule-based algorithm.

4.4.1 Rules for Afaan Oromo Segmentations

Rules for Afaan Oromo words segmenter developed based on the following concepts: -

A. Concepts taken from Debela stemmer [71] which basically adopted some basic concepts from porter stemmer[75]. However, it is not possible to apply rules of stemmer to word segmenter, as their function is different from each other. Stemmers strip all affixes from a given word and produce words stems (root of inflected word) while segmenters separate each morpheme (a stem and affixes). Thus, our module keeps all the valid morphemes of a given word for further processing rather than stripping affixes.
B. Concepts learned while studying the morphological features of Afaan Oromo words
C. Concepts observed from sample words.

Considering the above concepts, we developed language dependent rules to segment words starting from right-end by separating the possible morphemes systematically one after another until a valid stem is reached because suffixes order and co-occurrences rules are applicable in Afaan Oromo inflectional morphology. For instance, the most common inflectional morphemes application order/sequence of Afaan Oromo noun is *Root/Stem* ➔*Definiteness/Number/Gender* ➔ *Case*. The order is relative because one or more of these *suffixes* may be absent.

In addition to patterns of occurrences of morphemes, we tried to handle some orthographic alternation rules governing the form that morphemes take in particular environments which can happen either at the stem-affix intersection points or within the stem itself (section 4.5).

4.4.1.1 Algorithm: Afaan Oromo nouns and verbs segmentation

Based on the above-mentioned general morphological properties of the language, Afaan Oromo word segmentation algorithm is developed. For instance, the following is algorithm that separate morphemes from Afaan Oromo words.

Name: segmenter Algorithm

Input: Nominals or Verbs, List of Morphemes

Output: segmented words

1. Read from nominals file
2. apply morphophonemic (orthographic alternation) or Assimilation rules to the words (if any)
3. Read list of morphemes *#sorted by length as the longest suffix removed first*
4. If last part of a word match with one of morphemes then stop comparison and divide word into stem and morpheme
5. Apply Epenthesis rules to the stem (if any)
6. If length of stem is greater than 3 then repeat the step 4
7. If there no match, go to the next line and repeat the steps 2,3,4,5,6
8. Put morphemes in proper order next to root of the word (last removed first written, LIFO)

Algorithm 4.1 Afaan Oromo word segmenter algorithm

The above algorithm removes ***case*** or ***Aspect*** first since it attached at the end position then either of plural, definiteness, gender or person indicators as they occurred alternatively after zero or more case morphemes removed. For instance the following table summarizes the process of nouns segmentation.

№	Word	Case-segmented	Caseless stem	NGD-segmented	Segmented word
1	Baratoonni	Baratoot ni	baratoot	Barat oota	Barat oota ni
2	Haramaayatti	haramaaya tti	haramaaya	haramaaya	haramaaya tti
3	Filannichaaf	filannich aaf	filannich	Filan icha	Filan icha aaf
4	Biyyoolessaa	biyyooless aa	biyyooless	Biyyool essa	Biyyool essa aa
5	Gidduugalichi	gidduugalich i	gidduugalich	Gidduugal icha	Gidduugal icha i
6	keellaawwanirratti	keellaawwan rratti	keellaawwan	Keellaa wwan	Keellaa wwan rratti

Table 4.2 processes of nouns segmentation iteratively

4.4.2 Testing Performance of the Segmenter

5.4.2.1 Test Data and evaluation technique

As wide-covering and correct set of affix rules is prerequisite for accurate morphological analysis and guarantees a successful analysis, we tested our algorithm using wordlist that contains 2300 (1300 nouns and 1000 verbs) unique segmented words taken from Afaan Oromo nouns and verbs corpus. Error counting technique was employed to evaluate the performance of the segmenter. The errors were analyzed and classified into two different categories: under segmenting and over segmenting errors. Then the performance of the segmenter was presented by accuracy that means the number of correctly segmented words divided by the total number of words.

$$\textbf{Accuracy} = \frac{\text{number of correctly segmented words}}{\text{total number of tested words}}$$

Accordingly, the overall performance of the segmenter is 95.7% and 93.2% for nouns and verbs respectively.

5.4.2.1 Sources of errors for the segmenter

Although the experiment shows that, the segmenter has a satisfactory performance; it is not perfect because of the errors caused by the following reasons:

a. **Ambiguity: -** ambiguities may be occur at segment level or at word level (ambiguity to decide whether the word has affix or not). For instance, bolded letters 'aa' in caal**aa** may implies (1) chala, or (2) for chala/chala's.

b. **Rule's limitation**: although rule-based approach was selected for its accuracy; developing **(**handling**)** all specific, exceptional and detail rules is difficult as it need details and all exceptional properties of Afaan Oromo nouns and verbs.

4.5 Training and Model Construction /Model Building for Morphological Analyser

Word analysis is all about breaking up words into its components (morphemes), list all the features for each component, and look for unusual combinations. All most all of Afaan Oromo bound morphemes, are suffixes that are divided into derivational and inflectional morphemes. The derivational morphemes are bound morphemes that create new words and change the lexical category of a word while inflectional morphemes serve a grammatical role in the language. The latter, which is our focus, cannot create new words in the language and change the lexical category in the language.

Our system was developed using two mixed approaches namely: -

1. ***Rule based approach*** for breaking up words into morphemes (already discussed in section 4.4)
2. ***Statistical approach*** for listing all the corresponding features of the morphemes.

4.5.1 Statistical Approach: HMM

The central idea of statistical machine learning of natural language is that, given a corpus as an input to the system, it analyzes each word in the text and generates a report. All stochastic taggers simply 'pick the most- likely tag for the item' [73] based on the Bayesian framework. The system uses neither dictionary nor morphological rules particular to the language. Nowadays, stochastic models are frequently used in NLP as almost any speech and language-processing problem can be recast as:

"Given N choices for some ambiguous input, choose the most probable one".

Thus, among many statistical models we selected Hidden Markov Model (HMM) for its ability to solve many ambiguity problems[4].

HMM is the statistical model that has been mostly used in tagging process. It allows us to consider about both observed events (like words, morphemes that we see in the input) and hidden events (like part-of-speech tags or morpheme tags) that we think of as causal factors in our probabilistic model.

The general idea behind HMM is that: -

If we have a sequence of items which we can directly observe, each with one or more potential tags, then we can choose the most likely sequence of tags which is 'hidden' from the observer of the text by calculating the probability of all possible sequences of tags, and then choosing the sequence with the highest probability.

The probability of a tag sequence (***hidden states***) is generally a function of the probability of a morpheme being assigned a particular tag from the list of all possible tags (most frequent tag) and the probability that one tag follows another (n-gram).

hidden-tag=P(morpheme/tag) * P(tag/previous-tag)

From the above formula there are two kinds of probabilities in HMM emission probabilities (morpheme likelihood probabilities) p(mi|ti) and tag transition probabilities p(ti|ti-1).

4.5.2 Training Data

4.5.2.1 Manual Preparation of Gold-standard Training data

After we collected raw data from Fana Broadcasting Corporation Afaan Oromo and BBC Afaan Oromo archives from April 2020 to September 2020 and selected noun and verbs, we manually tagged by considering context of morphemes in the words. This tagged corpus is used for training the analyzer and evaluating its performance. The total tagged corpus consists of 4,800 and 4,200 nouns and verbs respectively.

Among these data, 90**%** of them were selected from the top part of data and used for ***training*** HMM for both nouns and verbs. Even if the performance of machine learning model highly depends on data size, when small data was used in machine learning, data splitting algorithms

can improve generalization performance [76]. However, there is no optimal suggested split percentage of data into training and testing dataset, therefore one can choose a split percentage that meets his/her project's objectives. Based on this, we consider our ***training set representative,*** we split our data into "Train: 90%, Test: 10%" in its normal order (not randomly), that means in machine learning, models' accuracy are increased as the amount of training data increase [77]. That is why we used 90% our manually annotated data for training. The same ration/share of training and test dataset is also used in Moyka and Dida [17] and Yitayal [19].

4.5.2.2 Afaan Oromo morphemes Tag sets

Since there is no tag set prepared for the purpose of Afaan Oromo morphological analysis, thirteen noun tags and twelve verb tags have been identified for the study. Thus, the words selected for training and evaluation of the system are segmented and tagged base on these morphemes tags.

No	Nouns		verbs	
	Tag	Description	Tag	Description
1	'NN'	Nominals = noun || adjective	'VB'	Verb
2	'PL'	Plural	'1SG'	1st person singular
3	'M'	Male	'1PL'	1st person plural
4	'F'	Female	'2SG'	2nd person singular
5	'DEF'	Definitiveness (default Male)	'2PL'	2nd person plural
6	'DEF/F'	Definitiveness-Female	'3SGF'	3rd person singular feminine
7	'ABS'	Absolutive/Accusative	'3SGM'	3rd person singular masculine
8	'SUBJ'	Subjective/Nominative	'3PL'	3rd person plural
9	'LOC'	Locative	'PERF'	Perfective
10	'DAT'	Dative	'IMPF'	Imperfective

11	'INSTR'	Instrumental	‘JUSS’	Jussive mood
12	'ABL'	Ablative	‘IMP’	Imperative mood
13	'GEN'	Genitive		

Table 4.3 lists of Afaan Oromo nouns and verbs morphemes-tag

4.6 The Analyzer

4.6.1 HMM decoder: Viterbi Algorithm

In training phase, machine will learn and develop model based on the training data and employed algorithm for the purpose. Thus, we selected HMM among other stochastic models.

Given an input as HMM (Transition Matrix, Emission Matrix) and a sequence of morphemes in w*ords*, find the most probable sequence of states *(morphemes tags in our case)*.

The two major assumptions followed while decoding tag sequence using HMMs:

3. The probability of a word appearing depends only on its *own tag* and is independent of neighboring words and tags.

4. The probability of a tag depends only on the *previous tag (bigram HMM)* occurred rather than the entire previous tag sequence i.e., shows Markov Property.

Viterbi Algorithm is the decoding algorithm used for HMMs. It is used to decode the hidden sequence of tags for a given sequence of observation of morphemes in a word.

4.7 Test Data Sets and Evaluation

In testing phase, the model will take word input and analysis it according to learning from training data. Accordingly, the test set for this model was the list of valid Afaan Oromo nouns and verbs, which was selected purposely from high-inflected (longer words), medium inflected, and less inflected (shorter words) words. We selected 480 nouns and 420 verbs of Afaan Oromo. Then the performance of the analyzer was evaluated by comparing the result of the analyser with that of manually tagged (gold standard) data. The evaluation result of the model performance is presented in accuracy.

4.8 Test results and performance of the Analyzer

After train the model, the evaluation of its performance is core step. The algorithm of the analyser was tested with 10% of manually tagged corpus for both nouns and verbs. Therefore, performance of the system measured using accuracy which was calculated by comparing the result of the analyser with that of manually tagged. If the result of the algorithm does not much the manual one, the result is counted as incorrect, otherwise it is counted as correctly analyzed.

Accuracy is dividing the correctly analyzed words by total test words in test data.

$$\textbf{Accuracy} = \frac{\text{number of correctly analyze words}}{\text{total number of tested words}}$$

The model is tested and show an accuracy of **84.6 %** and **82.9%** correctly analyzed nouns and verbs respectively. All information about the evaluation result presented in the following table.

Type of Accuracy Measured	Nº-of tested data	Nº Correctly analyzed words	Accuracy
Nouns	480	406	84.6%
Verbs	420	348	82.9%
Mixed	900	754	83.8%
Average			**83.7%**

Table 4.4 Evaluation results of Afaan Oromo Nouns and Verbs separately, on mixed and on Average

Sample part of the training and testing environment of the Noun morphological tagger was illustrated below.

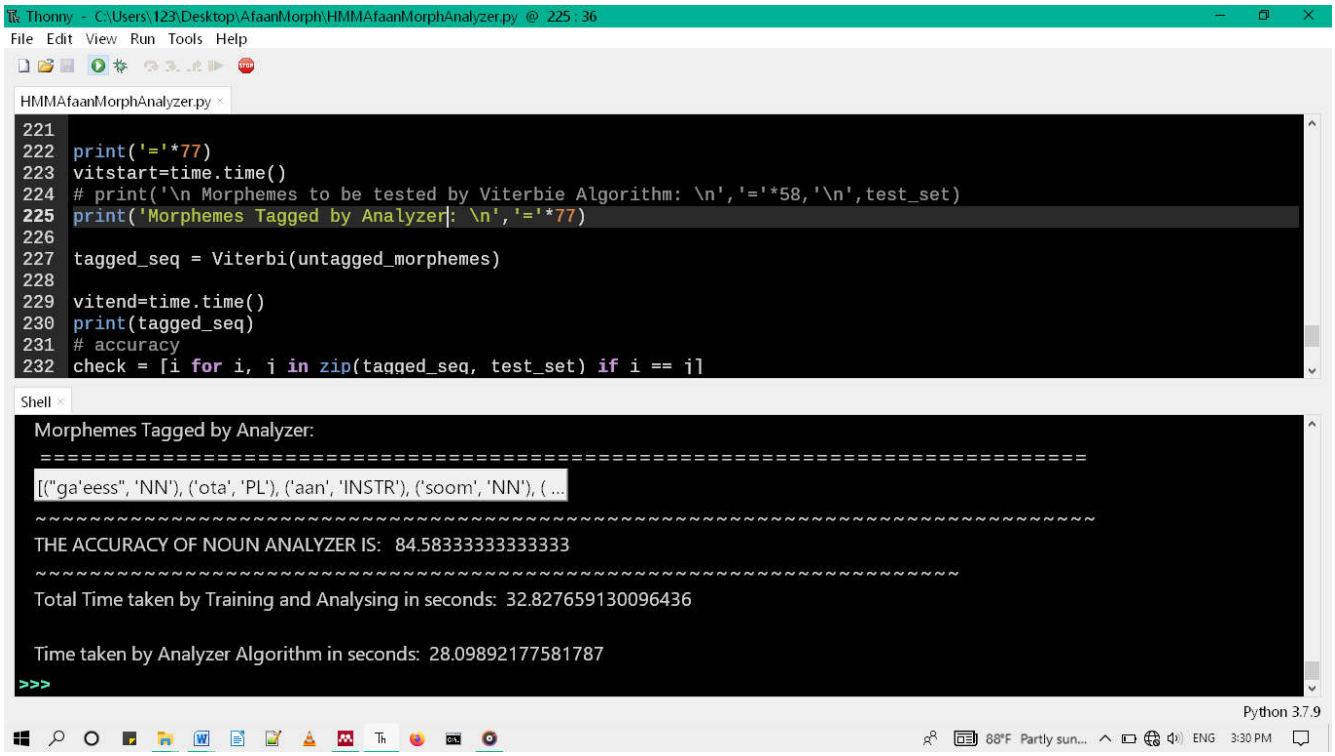


Fig 4.3 Training and Testing Environment of the Afaan Oromo Noun morphological Analyzer

4.9 Discussion of the Results

The proposed Afaan Oromo morphological analyzer model is evaluated to analysis the new words that were not found in the training dataset based on the knowledge of the trained model. The experiments performed separately to label both nouns and verbs morphemes; the accuracy was used as evaluation method. The accuracy of the model is evaluated for nouns and verbs separately, for both on mixed and on average, the result presented in Table 4.4.

Accuracy nouns/verbs analyzer indicates the percentage of nouns/verbs analyzed correctly by the model. The mixed accuracy indicates the percentage of both verbs and nouns analyzed correctly. Average accuracy calculated by adding nouns and verbs accuracy the divided by 2.

Although we get satisfactory result, the model must perform than what it did. Thus, to learn the reasons for these accuracies we do some assessments on our output data (Automatically analyzed data) and observe the following source of errors.

4.9.1 Sources of errors for incorrect Analysis

1. **Words Ambiguity**

A. **Morpheme level ambiguity:** as Afaan Oromo is morphologically fusional language[8] (also called inflectional), unlike in agglutinative language one morpheme, usually a single inflectional suffix, expresses several different meanings or grammatical functions. For instance case indicator suffix "**aa**" in jiraataa, Shirkaa, and Tolaa, indicates Genetive(GEN), Ablative(ABL), and Dative(DAT) case respectively. Afaan Oromo verbs are marked by agreement morpheme **–t-** for 2nd persons (singular and plural) and 3rd person feminine. Plural numbers of persons are marked by suffix **–n (an).**

In addition to this, its high number of allomorphs characterizes Afaan Oromo.

Correctness	Manually Analyzed	Automatically labeled
(a) Correctly annotated words	('caams', 'NN'), ('aa', 'ABL'),	('caams', 'NN'), ('aa', 'DAT'),
	('barnoot', 'NN'), ('icha', 'DEF'), ('aa', 'GEN'),	('barnoot', 'NN'), ('icha', 'DEF'), ('aa', 'DAT'),
	('baanki', 'NN'), ('tti', 'LOC'),	('baanki', 'NN'), ('tti', 'DAT'),
	('jiraat', 'NN'), ('aa', 'GEN'),	('jiraat', 'NN'), ('aa', 'DAT'),
	('shirk', 'NN'), ('aa', 'ABL'),	('shirk', 'NN'), ('aa', 'DAT'),
	('xaaliyaan', 'NN'), ('ii', 'ABL')	('xaaliyaan', 'NN'), ('ii', 'GEN'),
	('walakkaa', 'NN'), ('tti', 'LOC'),	('walakkaa', 'NN'), ('tti', 'DAT'),
	('lixaa', 'NN'), ('tti', 'LOC'),	('lixaa', 'NN'), ('tti', 'DAT'),
(b) Incorrectly annotated words	('boongaa', 'NN'), ('tti', 'LOC'),	('boongaa', 'NN'), ('tti', 'DAT'),
	('hidhann', 'NN'), ('oo', 'GEN'),	('hidhann', 'NN'), ('oo', 'ABL'),
	('kibbaa', 'NN'), ('tti', 'LOC'),	('kibbaa', 'NN'), ('tti', 'DAT'),

Table 4.5 Sample of incorrectly labeled because of fusional morphology

(a): Automatically labeled morphemes are correct linguistically but different from the manual one and considered as incorrect by model ***(b)*** Automatically labeled morphemes are incorrect and also different from the manual one

These problems can be solved by integrating simple NLP application such as NER and accessing complete contextual information.

B. **Word Level ambiguity: Homograph:** When the same form representing multiple word or sense, the phenomenon is called homographs. For instance

   In Afaan Oromo verb "**baate"** have two interpretations based on the context: (1) he carried and (2) she came out

2. **Statistical Error: Model over fitting**

   For example, the following wrongly labeled are because of statistical error

Manually annotated sequences	**Automatically annotated sequences**
('taph', 'NN'), ('icha', 'DEF'), ('qix', 'SUBJ'), ('aan', 'NN'), ('biiznas', 'ABS'), ('ii', 'NN'), ('dhimm', 'ABS'), ('icha', 'NN'), ('aaf', 'DAT'), ('gadaa', 'NN'), ('tiin', 'INSTR'),	('taph', 'NN'), ('icha', 'DEF'), ('qix', 'NN'), ('aan', 'INSTR'), ('biiznas', 'NN'), ('ii', 'GEN'), ('dhimm', 'NN'), ('icha', 'DEF'), ('aaf', 'DAT'), ('gadaa', 'NN'), ('tiin', 'INSTR'),

Table 4.6 Sample of incorrectly labeled because of statistical error

Finally, we conclude that the statistical model can handle every unseen data depending on the statistics sequences. So, it can handle out of vocabulary problem (OOV) problem if it is provided well with enough training dataset. However, it also needs integration of rule based or lexeme-based approach to handle irregularities and ambiguities behavior of language and contextual (document) particularity.

In Moyka and Dida [17], the only evaluated morphological analyser for Afaan Oromo, the generalized accuracy of 98.86% and 94.14% was achieved by interleaving the combination of selected features and optimal parameters of IB1 and IGTREE algorithm respectively from 2,270 annotated words. Among these annotated words, 10% of dataset was used as testing data. Even if this work seems outperform our work, it is difficult to directly compare the two works because of

difference in data and features of data used. In order to compare and select the best performing morphological analyzer for Afaan Oromo, developing morphological analyzer using all available approach and testing them using the same (identical) dataset is important. However, we can compare the results obtained and parameters used in both works. A comparison of Afaan Oromo morphological analyzer with memory based learning approach and hybrid approach leaded by statistical approaches in terms of accuracy is shown below in table 4.7.

Approaches	**Algorithms Or Models**	**Size of dataset**	**Training (Testing) Instances**	**Accuracy**
Memory based Learning [17]	IB1	2,270 Nouns, adjectives and verbs	15647 Training 1739 Testing	98.86%
	IGTREE			94.14%
Hybrid Approach (RB leaded by Statistical Approach)	HMM	4800 nouns and adjectives	9394 training, 1040 testing	84.6%
		4200 verbs	8768 training, 983 testing	82.9%

Table 4.7 a comparison: the summary of results of MBL approach and Our Work.

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This thesis presented original work on the analysis of Afaan Oromo words employing a combination of supervised machine learning approach and rule-based approach, for tagging and segmentation respectively. In this thesis, we also tried to analysis all morphological features of Afaan Oromo and different approaches employed for the development of morphological analyzer of different languages to select the one, which is suitable for Afaan Oromo.

Through this thesis, we answered the questions which were raised in this thesis and achieved our goal which were proposed early. Accordingly in chapter three of this paper we tried to answer the first question from the literatures. Almost all of Afaan Oromo morphology related literatures states that, "in complex word structure, certain set of suffixes conventionally come in a particular sequence before or after other suffixes". The most common order/sequence of the major Afaan Oromo suffixes/morphemes within a given word is:-

Root/stem➔derivational suffixes➔inflectional suffixes.

The order is relative because one or more of these suffixes may be absent except the first free morpheme. As the main focus of this work, is inflectional morphology, which is characterized by inflectional suffixes/morphemes. Thus, we referred and analyzed many related literatures with Afaan Oromo inflectional morphology to answer the second question. The forms, functions, morphological behaviors and patterns of occurrences of inflectional morphemes of Afaan Oromo nouns, adjectives and verbs are described chapter three. Depending on the information in chapter three, we collected and prepared data (see section 4.3). The main Afaan Oromo inflectional morphemes with its probability of co-occurrences are also systematically analyzed from our annotated training data and illustrated in chapter four (see fig 4.2). In addition to affix learning rules and disambiguation, many languages specific morphological features such as morpheme application order with some exceptions are considered.

The last three questions were practically answered and reported in chapter four. In this chapter the proposed architecture of Afaan Oromo morphological analyser is well represented. The system has two phases: Training phase and Analyzer phase. In the training phase the model trained to learn based on the training data set using stochastic concepts to produce trained model.

The analysis phase of proposed system will take segmented words that are produced at segmentation module and labels the morphemes based on lexicon probability (emission probability) and transitional probability of the morphemes learned from training. Respectively, a corpus of size 4,320 and 3,780 nouns and verbs are used to train the HMM model. Then the performance of the analyser was tested using 480 nouns and 420 verbs of Afaan Oromo. The accuracy of the analyzer for both nouns and verbs is 84.6 % and 82.9% respectively, which is quite satisfactory result. However, problems remains can be solved with adding detail information such as NER and taking account of the regular phonological or orthographical alternations due to morphological, and morphophonological processes involved. The inflectional category of words is often taken from a morpheme that serves as a "head of a word's morphemes" as in gadaa**tiin** which has two case morphemes (GEN-INSTR). However, our system considers only the head morphemes (tiin-INST) rather than GEN-INSTR.

The key limitations in this effort are limited funding opportunities, scarcity of gold standard and balanced annotated data sets, language's morphological complexity (i.e., Afaan Oromo is morphologically, polysynthetic and fusional) and inherently multiple sources of ambiguity of the language at morpheme level and word level.

Depending on the limitations and weaknesses of the model and challenges we faced during the research work, ultimately, we put our suggestions in following section.

5.2 Recommendations and future works

As a number of additional tasks need to be done to develop fully-fledged morphological analysis system for Afaan Oromo; possible directions for further research emanating from the work presented in this thesis, include the following:

1. The inclusion of other model of Viterbi algorithm such as trigram model and other approach includes genetic algorithms and neural networks to improve the performance of the Afaan Oromo word analysis system and to compare the results of approaches.
2. As machine-learning approach requires standardized, large corpus size, more corpus size in the gold standard is needed for developing an improved morphological analyzer.
3. In addition to simple emission and transitional probability, combinatory ambiguities need richer contextual information such as grammatical category (POS) and lexical meaning (context between words) for correct analysis.
4. For better segmentation, which causes better analysis, the system also needs detail information (sub-categorization of a word) such as types of nouns in noun category. For instances, person nouns do not inflect for pluralization, case suffix ***–tti*** indicates dative for noun and locative for place names.
5. It also possible to improve the performance of segmenter by using Word Net to identify weather certain end of word is a part of that word or bounded (inflectional) morpheme.
6. This study includes only most common inflectional affixes used in the language. However, one can conduct more research on other types of affixes such as derivational and compound words.
7. Because of the time and resource limitations, we cannot consider other categories of word classes than nouns and verbs in the language. Therefore, more research has to be done to include all categories of word classes in the system.
8. The procedure followed in this study to develop Afaan Oromo morphological analyser can be adopted in developing morphological analyzer for other local languages especially for those have the same morphological behaviors.
9. The system being developed in this study is just a prototype. Further research should be undertaken to develop a full-fledged morphological analyzer that can be easily integrated into different NLP applications.

REFERENCES

[1] S. Bird, E. Klein, and E. Loper, "Natural Language processing in Python," p. 479, 2008, doi: 10.1007/s11051-014-2693-7.

[2] J. Allen, *Natural Language Understanding*, 2nd ed. California: Benjamin-Cummings Publishing Company Inc., 1996.

[3] A. Abeshu, "Automatic Morphological Synthesizer for Afaan Oromoo," *Master's thesis, AAU*, 2010.

[4] D. Jurafsky and J. H. Martin., "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition," *Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. speech recognition.*, 2006.

[5] R. Wiese, "A Two-Level Approach to Morphological Structure," *J. Ger. Linguist.*, vol. 20, no. 3, pp. 243–274, 2008, doi: 10.1017/s147054270800010x.

[6] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Assoc. Comput. Linguist.*, vol. 27, no. 2, pp. 153–198, 2001, doi: 10.1162/089120101750300490.

[7] A. Sani, "Afaan Oromo Named Entity Recognition Using Hybrid Approach," *Master's thesis*, 2015.

[8] W. Olani, "Inflectional Morphology in Oromo," *Master's thesis, AAU*, 2014, [Online]. Available: http://www.sfu.ca/~mtaboada/docs/Julian_Brooke_MA_.

[9] W. Tegegne, "The Development of Written Afan Oromo and the Appropriateness of Qubee , Latin Script , for Afan Oromo Writing," *Issn 2224-3178*, vol. 28, no. 1976, pp. 8–14, 2016, [Online]. Available: https://www.iiste.org.

[10] H. Alemayehu, "The Structures of Nominal Clauses in Afan Oromo A Minimalist Approach," *Master's Thesis*, 2007.

[11] K. A. Ouba, "The Analysis of Morphological Properties of Arsi-Bale Word Classes with Special Reference to Arsi-Bale Noun Morphology," *JETIR*, vol. 6, no. 1, pp. 649–674, 2019.

[12] A. Woldemariam, "Development of Morphological Analyzer for Afaan Oromoo Text," 2005.

[13] A. Tuni, S. Tumsa, D. P. Sharma, B. Singh, and M. A. Bochicchio, “Afaan Oromo Hybrid Modelling: A Case based Optimized Intelligence in Information Retrieval System’s Localization,” *Psychol. Educ.*, vol. 58, no. 2, pp. 9273–9283, 2021.

[14] W. Tegegne, “The Politics of Qubee in Ethiopia: Rationales and Criticisms to the Adaption of Qubee as an Official Orthography of Afaan Oromoo,” *Gadaa Journal/Barruulee Gadaa*, vol. 2, no. 1, pp. 18–38, 2019.

[15] J. Goldsmith, “An algorithm for the unsupervised learning of morphology,” *Nat. Lang. Eng.*, vol. 1, no. 1, pp. 353–371, 2001, doi: 10.1017/S1351324905004055.

[16] M. Gasser, “HORNMORPHO 2.5 User ’s Guide,” *Hum. Lang. Technol. Democr. Inf.*, pp. 1–55, 2012.

[17] M. Degefa and D. Midekso, “Morphological Analyzer for Afaan Oromoo Using Machine Learning,” Addis Ababa University, 2020.

[18] H. Bhavsar and A. Ganatra, “A Comparative Study of Training Algorithms for Supervised Machine Learning,” *Int. J. Soft Comput. Eng.*, vol. 2, no. 4, pp. 74. – 81, 2012.

[19] Y. Abate, “Morphological Analysis of Ge’ez Verbs Using Memory Based Learning,” Addis Ababa University, 2014.

[20] D. Tesfaye and E. Abebe, “Designing a Rule Based Stemmer for Afaan Oromo Text,” *Int. J. Comput. Linguist.*, vol. 1, no. 2, pp. 1–11, 2010, [Online]. Available: http://www.cscjournals.org/csc/manuscriptinfo.php?ManuscriptCode=69.70.63.72.41.50.102.

[21] G. Mamo, “OroRoots: Rule-Based Root Generation System for Afaan Oromo,” *Int. J. Sci. Eng. Res.*, vol. 8, no. 3, pp. 1212–1214, 2017.

[22] A. Abeshu, “Analysis of Rule Based Approach for Afan Oromo Automatic Morphological Synthesizer,” *Sci. Technol. Arts Res. J.*, vol. 2, no. 4, pp. 94–97, 2013, doi: http://dx.doi.org/10.4314/star.v2i4.16.

[23] C. R. Kothari, *Research Methodology: Methods and Techniques*, Second Rev. New Delhi: New Age International (P) Ltd., Publishers, 1990.

[24] R. Kumar, *RESEARCH METHODOLOGY: a step-by-step guide for beginners*, 3rd Editio. SAGE Publications Ltd, 2011.

[25] N. Walliman, *Research Methods: The Basics*. London: Routledge, 2011.

[26] A. Carstairs-Mccarthy, *An Introduction to English Morphology: Words and Their*

Structure. Edinburgh University Press, 2002.

[27] R. Herbrich and T. Graepel, *Handbook of Natural lannguage Processing*, 2nd ed. Taylor and Francis Group, LLC, 2010.

[28] M. Haspelmath and A. D. Sims, *Understanding Morphology*, 2nd ed. London: Hodder Education, an Hachette UK Company, 338 Euston Road, London NW1 3BH, 2010.

[29] S. Keshava and E. Pitler, "A Simpler , Intuitive Approach to Morpheme Induction," in *Proceedings of 2nd Pascal Challenges Workshop*, 2005, pp. 31–35.

[30] B. Can and S. Manandhar, "Probabilistic hierarchical clustering of morphological paradigms," *EACL 2012 - 13th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc.*, pp. 654–663, 2012.

[31] G. G. Chowdhury, "Natural Language Processing," *Lect. note,Department Comput. Inf. Sci. Univ. Strat.*, vol. 39, no. 1, pp. 60–62, 1996, doi: 10.1145/234173.234180.

[32] M. Haspelmath, "Universals of Language Today," *Stud. Nat. Lang. Linguist. Theory*, no. September 1998, p. 301, 2009, doi: 10.1007/978-1-4020-8825-4.

[33] B. Comrie, *Language Universals and Linguistic Typology: SYNTAX AND MORPHOLOGY*, Second. The University of Chicago Press, 1989.

[34] G. Booij, *The Grammar of Words: An Introduction to Linguistic Morphology*. Oxford University Press., 2007.

[35] B. Desta, "Design and Implementation of Automatic Morphological Analyzer for Ge ' ez Verbs," *Master's Thesis*, pp. 1–144, 2010, [Online]. Available: aau.etd.edu.et.

[36] T. B. Bati, "Automatic Morphological Analyzer: An Experiment Using Unsupervised and Autosegmental Approach," 2002.

[37] M. Gasser, "HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya," *Conf. Hum. Lang. Technol. Dev.*, pp. 94–99, 2011.

[38] P. J. Antony and K. P. Soman, "Computational Morphology and Natural Language Parsing for Indian Languages : A Literature Survey," *Ijcset*, vol. 3, no. 4, pp. 136–146, 2012, [Online]. Available: http://www.ijser.org.

[39] I. I. Ayogu, A. O. Adetunmbi, and N. C. Kammelu, "Finite State Concatenative Morphotactics : The Treatment of Igbo Verbs," *Int. J. Comput. ICT Res.*, vol. 7, no. 1, pp. 70–80, 2013.

[40] A. Kumar, "Morphology based Prototype Statistical Machine Translation for English to

Tamil Language," *PhD Diss.*, 2013, [Online]. Available: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf%5Cnhttp://nlp.amrita.edu:8080/project/mhrd/ms/Tamil/AnandkumarPhdSlides.pdf.

[41] Mulugeta W. & Michael G., "Learning Morphological Rules for Amharic Verbs Using Inductive Logic Programming," *Work. Lang. Technol. Norm. Less-Resourced Lang.*, vol. 8, pp. 7–12, 2012.

[42] K. Lisanu, "Design and Development of Automatic Morphological Synthesizer for Amharic Perfective Verb Forms," *Master's Thesis*, 2002.

[43] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill Science/Engineering/Math, 1996.

[44] H. Hammarström and L. Borin, "Unsupervised learning of morphology," *Assoc. Comput. Linguist.*, vol. 37, no. 2, pp. 309–350, 2011, doi: 10.1162/COLI_a_00050.

[45] K. Koskenniemi, "Two-level Morphology A general computational model for word-form recognition and production," 1983.

[46] C. Borg and A. Gatt, "Crowd-sourcing evaluation of automatically acquired, morphologically related word groupings," *Proc. 9th Int. Conf. Lang. Resour. Eval. Lr. 2014*, pp. 3325–3332, 2014.

[47] O. Kohonen, S. Virpioja, and K. Lagus, "Semi-supervised learning of concatenative morphology," no. July, pp. 78–86, 2010.

[48] A. Clark, "Supervised and Unsupervised Learningof Arabic Morphology," *A. Soudi, A. van den Bosch G. Neumann (eds.), Arab. Comput. Morphol.*, pp. 181–200, 2007, doi: 10.1007/978-1-4020-6046-5.

[49] G. Durrett and J. DeNero, "Supervised learning of complete morphological paradigms," *NAACL HLT 2013 - 2013 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Main Conf.*, pp. 1185–1195, 2013, [Online]. Available: https://www.aclweb.org/anthology/N13-1138.

[50] T. Ruokolainen, O. Kohonen, S. Virpioja, and M. Kurimo, "Supervised morphological segmentation in a low-resource learning setting using conditional random fields," *CoNLL 2013 - 17th Conf. Comput. Nat. Lang. Learn. Proc.*, pp. 29–37, 2013.

[51] V. P. Abeera *et al.*, "Morphological analyzer for Malayalam using machine learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6411 LNCS, no. November

2017, pp. 252–254, doi: 10.1007/978-3-642-27872-3_38.

[52] G. Chrupała, G. Dinu, and J. Van Genabith, “Learning morphology with Morfette,” *Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008*, pp. 2362–2367, 2008.

[53] A. Kumar, R. Cotterell, L. Padro, and A. Oliver, “Morphological Analysis of the Dravidian Language Family,” in *Proceedings ofthe 15th Conference ofthe European Chapter ofthe Association for Computational Linguistics*, 2017, vol. 2, pp. 217–222.

[54] S. Srirampur, R. Chandibhamar, and R. Mamidi, “Statistical Morph Analyzer (SMA++) for Indian Languages,” in *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 2014, pp. 103–109, doi: 10.3115/v1/w14-5312.

[55] A. Bharati, R. Sangal, D. M. Sharma, and L. Bai, “Anncorra: Annotating corpora guidelines for POS and chunk annotation for Indian languages,” *LTRC-TR31*, 2006.

[56] E. Akyürek, E. Dayanık, and D. Yuret, “MorphNet: A sequence-to-sequence model that combines morphological analysis and disambiguation,” *Trans. Assoc. Comput. Linguist.*, vol. 7, 2018, doi: 10.1162/tacl_a_00286.

[57] J. Nivre *et al.*, “Universal dependencies v1: A multilingual treebank collection,” *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, pp. 1659–1666, 2016.

[58] D. K. Malladi and P. Mannem, “Statistical Morphological Analyzer for Hindi,” in *International Joint Conference on Natural Language Processing*, 2013, pp. 1007–1011.

[59] D. K. Malladi and P. Mannem, “Context based statistical morphological analyzer and its effect on Hindi dependency parsing,” in *Proceedings ofthe Fourth Workshop on Statistical Parsing ofMorphologically Rich Languages*, 2013, pp. 119–128.

[60] M. Anand Kumar, V. Dhanalakshmi, K. . Soman, and S. Rajendran, “A Sequence Labeling Approach to Morphological Analyzer for Tamil Language,” *Int. J. Comput. Sci. Eng.*, vol. 02, no. 06, pp. 1944–1951, 2010.

[61] D. Altinok, “DEMorphy, German Language Morphological Analyzer,” *arXiv:1803.00902v1 [cs.CL]*, 2018, [Online]. Available: https://arxiv.org/abs/1803.00902.

[62] V. Ravishankar, F. M. Tyers, and A. Gatt, “A morphological analyser for Maltese,” *Procedia Comput. Sci.*, vol. 117, pp. 175–182, 2017, doi: 10.1016/j.procs.2017.10.107.

[63] C. Borg and A. Gatt, “Morphological Analysis for the Maltese Language : The challenges of a hybrid system,” *Proc. ofThe Third Arab. Nat. Lang. Process. Work.*, pp. 25–34, 2017.

[64] C. John J, “A Computational Grammar and Lexicon for Maltese,” *Master’s Thesis,*

Chalmers Univ. Technol. Gothenburg, Sweden, 2013.

[65] F. M. Tyers, F. Sánchez-martínez, S. Ortiz-rojas, and M. L. Forcada, "Free / Open-Source Resources in the Apertium Platform for Machine Translation Research and Development," *Prague Bull. Math. Linguist.*, no. 93, pp. 67–76, 2010, doi: 10.2478/v10108-010-0015-5.Bereitgestellt.

[66] A. Gatt and S. Ceplo, "Digital corpora and other electronic resources for Maltese," *Corpus Linguist. 2013*, pp. 96–97, 2013.

[67] M. Lui and T. Baldwin, "langid . py : An Off-the-shelf Language Identification Tool," no. July, pp. 25–30, 2012.

[68] X. TANG, "English morphological analysis with Machine-learned rules," *PACLIC 20 - Proc. 20th Pacific Asia Conf. Lang. Inf. Comput.*, pp. 35–41, 2006.

[69] M. Abate and Y. Assabie, "Development of Amharic Morphological Analyzer Using Memory-Based Learning," *PolTAL 2014, LNAI 8686*, pp. 1–13, 2014.

[70] K. Kula, V. Varma, and P. Pingali, "Evaluation of Oromo-English Cross-Language Information Retrieval," *Int. Jt. Conf. Artif. Intell. (IJCAI)-2007*, 2008.

[71] D. Tesfaye, "Designing a Stemmer for Afaan Oromo Text : A Hybrid Approach," *Master's Thesis, AAU*, 2010.

[72] S. O. Gonfa, "Word Sense Disambiguation for Afaan Oromo: Using Knowledge Base," *Master's Thesis, St. Mary's Univ.*, 2018.

[73] G. Mamo, "Part-of-Speech Tagging for Afaan Oromo Language," *Master's Thesis*, 2009, doi: 10.18860/ling.v5i1.609.

[74] G. Mamo and M. Meshesha, "Parts of Speech Tagging for Afaan Oromo," *Int. J. Adv. Comput. Sci. Appl.*, vol. 1, no. 3, 2011, doi: 10.14569/specialissue.2011.010301.

[75] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 40, no. 3, pp. 211–218, 2006, doi: 10.1108/00330330610681286.

[76] D. E. Birba, "Study of Data Splitting Algorithms for Machine Learning," STOCKHOLM, SWEDEN 2020, 2020.

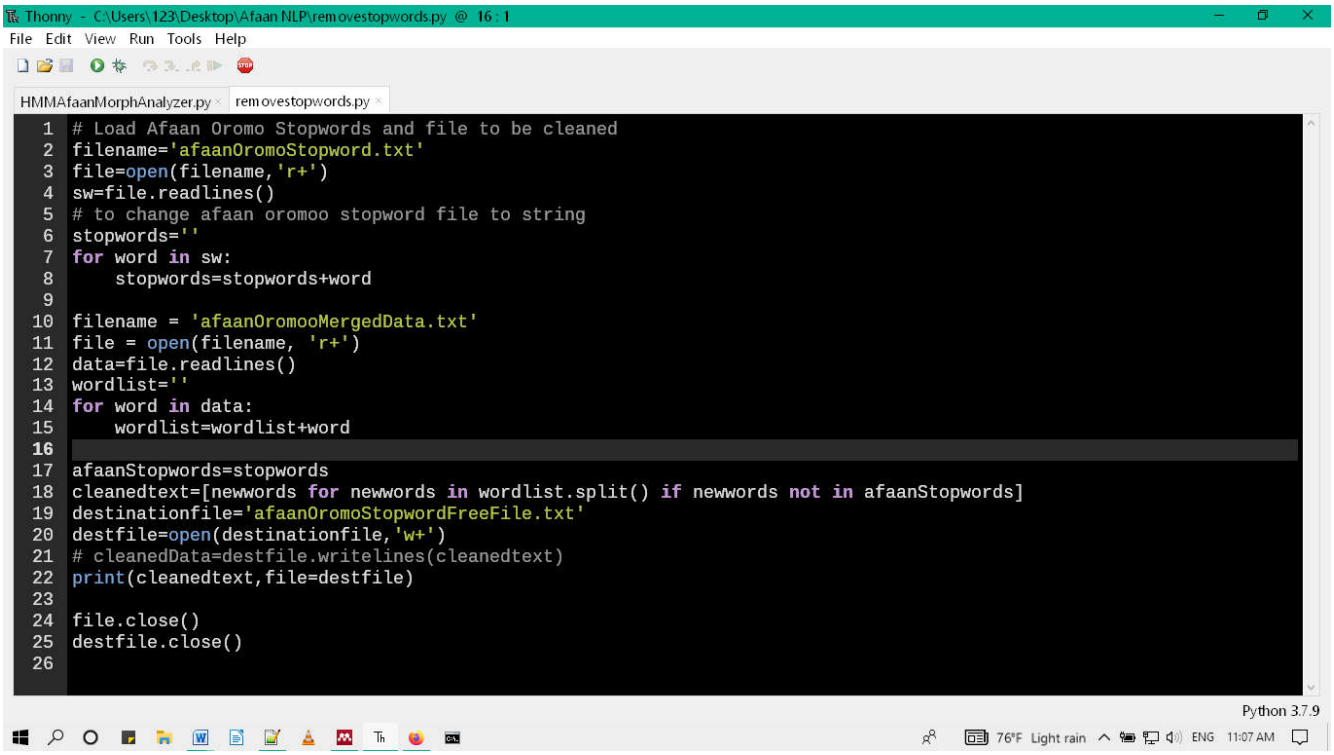
[77] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, no. 4, pp. 1–16, 2021, doi: 10.3390/molecules26041111.

Appendixes

Appendix A: python script to remove Noisy Data from Afaan Oromo text

```
# import very neccessary packages and modules
import re
from string import digits

# load text
filename = file to be processed
file = open(filename, 'r+')
text = file.readlines()
filename2="NoisyFreeWords.txt"
file2=open(filename2,'w+')

for txt in text:

    totalcount=totalcount+len(txt.split()) #Total number of words, punctuations and digits
    punctions='\!"#$%&()*+,./:;<=>?@[\\]^_-`{|}~' # list of all punctuations apostrophe(') is excluded
    pd=punctions+digits
    remove_punctuations_and_digits = str.maketrans('', '', pd)
    pdfreefile = txt.translate(remove_punctuations_and_digits) #Noisy free data

    file2.writelines(pdfreefile)     # to write the file content of file in  file2
    words=pdfreefile.split()
    wordscount=wordscount+len(words)      #number of  net words
file.close()
file2.close()
print("\nTotally, there are ".upper(),totalcount," items in this FILE".upper())
print("\nnetly, There are ".upper(),wordscount," words in this file".upper())

```

Appendix B: python Script to remove Stopwords from a text file



Appendix C: List of Afaan Oromo Stop Words

A) List Afaan Oromoo Stop words adopted from Debela Tesfaye's[71] master's thesis (78 stopwords):

silaa	keessattuu	osoo	hoggaa	ennaa	tanaaf
hoo	yommuu	saniif	otuu	yoo	simmoo
yammuu	malee	moo	yookiin	hogguu	yookinimoo
oggaa	otuullee	fi	ani	itumallee	hanga
kanaafuu	eegana	eegasii	yommii	ituu	akkam
akka	jechaan	akkuma	odoo	otumallee	waan
innaa	yoom	kanaaf	immoo	yookaan	jechuun
ituullee	kan	isheen	tahullee	inni	isaan
kun	yeroo	jechuu	tawullee	eega	isaa
utuu	sun	akkasumas	isiin	tanaafuu	booda
akkum	yemmuu	booddee	yoommuu	ammo	ofii
koo	dura	tanaafi	kanaafi	ta`ullee	illee
henna	waggaa	erga	otoo	garuu	

B) List Afaan Oromo we compiled manually from our corpus (264 stopwords):

ittiin	osoo	mala	achi	biraa	ofisaa
simmoo	hamma	kanneen	kuni	addaan	inumaa
maal	jidduutti	kan	inni	tana	akkas
akkasuma	irra	silaa	akkasumsa	ofii	ta'us
kamuu	ammoo	adda	hammam	jala	natti
innis	kamittuu	haaji	kanuma	jechaan	walii
gad	booda	ta'ellee	tahullee	sana	ishiin
abboo	laata	dura	ta'ullee	ittin	ammumaa
anis	ta'us	kamiituu	ta'es	saniif	hammas
haata'u	ani	takkaa	mitii	uf	sababa

boodas	keenya	koo	akkasi	oli	jechuu
eegasii	kinooti	akkanaa	eega	ituu	kamiinis
kunniin	sunisasi	kun	nan	akkanaatti	erga
nutti	bira	akkam	hangam	sirritti	nuu
boru	sunis	isin	kee	obbo	siif
koon	sanaa	malees	ammas	akka	isaanii
kam	moo	hoggaa	amma	si'a	ishiis
akkaata	yoota’u	eeyyee	ofiin	isheen	ta`ullee
isa	utuu	ergaa	waliin	taatus	hoo
innaa	gara	itumallee	as	hanga	si
waggaa	hogguu	henna	hundaafuu	jechuun	isiin
immoo	isiniin	asi	tun	nuun	obboo
yookan	yookiin	maraa	ana	teenya	ammaan
isaati	oggaa	kaan	gama	kannen	akanaa
dha	ammammoo	eenyumaa	isaatti	akkasii	yoon
walbira	ta’ullee	naa	waan	gadi	ammaa
otoo	achuma	kara	ennaa	akkan	tanaaf
waanin	yommii	yookin	tanaafi		ti kanaafuu
yommuu	nin	keessattuu	homaa	kana	akkaataa
maaliif	waloo’	akkum	odoo	akkasumaas	yoommuu
kamiin	fa’a	otuullee	mii	yookinimoo	kanaaf
addaa	tanaafuu	sanaan	kum	yeroo	tan
akkuma	walqixxee	waa’ee	maaloo	isaaf	keessan
hedduu	hanaga	malee	ittii	kiyya	anaaf
sanii	biroo	yeeruma	ammatti	isaan	sun
ittuu	kooti	ittti	tasa	kunis	isaa
qofa	yommu	akki	kanafiis	illee	amman
kanaa	isarra	tawullee	garuu	haa	baga
waa	kunneen	hammuma	kanaafi	kanattii	kanaafan
addee	ishii	na	yookaa	isaanin	kunoo
yoo	anaa	yookaan	eenyuyyuu	ammallee	otumallee

nuti	eda	maalif	irraa	tahus	naaf
mitis	yoomii	ati	otuu	ishee	kanas
ofirra	booddee	yoom	kami	yoonaa	nu
eegaa	wal	itti	hunduma	yammuu	nuuf
taanan	eegana	akkaatuma	akkuman	aadde	haga
kanaan	al	of	akkasumas	qofaaf	ofiis
ammo	isanii	ol	oltuumma	yaa	kaanaf
kanatu	yemmuu	ituullee	akkaa	hagas	fi
miti	yeruma	hamman	hin	ittis	ni
akkamiin	akkanaatin				

Appendix D: Sample Afaan Oromo words

noottiif	nagaadhaan	riizoortiitiin	reestooraantotaa
riifaraalatti	waancadhaaf	heektaraa	shaneen
daawwachiisummaadhaan	keessaatiif	xalayichaan	hanqinootni
loowwaniif	ijoorraa	galgalaan	miidiyaa
daandiileen	peeruutti	afgaaffiin	afaarirraa
humnoota	yakkicha	yeeroon	ka'uumsi
gaariidhaan	ilbiisota	waggoonni	finfinneetti
qobbootti	dhiigaaf	dhaadannoo	jabilootaaf
diinagdeen	shororkaadhaan	itopphiyaan	jabbiilee
maanguddootarratti	bishaaniirratti	iyyannoodhaan	muumm yaa'onni
magaalicharratti	furaadhaaf	mari'ataan	funyaaniin
atleetota	gaafatamaan	eeruuwwan	yakkaan
paartiin	qotiyyoodhaaf	maanuufaakchariingii	saayintistootaan
biyyaaleessaatiif	dirqamaatti	intalijansii	ichaa ummanni
qilleensaarraa	cimdii	waraanichi	muummeen
ilaalchaan	simbirrootaan	namootarraa	eebbisaaf
kilaasteeraan	waltajjiif	tikaaf	eenyuuf
biyyaaleessaatiin	kireessaan	imimaanesituun	seerri
istiyaadiyeemii	wal'aansaaf	onkoloolessi	herreegni
godinaa	imaammanni	koroonaavaayrasiif	eeyyamni
firiijiin	galaanaa	jabinaan	kennicha
madaalawaan	hordoftootaan	buttaajiraa	piroojeektoota
hawwataa	guutuu	jeequmsatti	haalota
amabaasaaddarri	tawaahdootiif	yaadessaa	galaanirratti
meeshaaleerraa	tokkotti	biqiltoota	

Appendix E: Sample of Automatically annotated Afaan Oromo Words

[("ga'eess", 'NN'), ('ota', 'PL'), ('aan', 'INSTR'), ('soom', 'NN'), ('aaf', 'DAT'), ('tokkoo', 'NN'), ('tiif', 'DAT'), ('olaan', 'NN'), ('aan', 'INSTR'), ('gazexessit', 'NN'), ('oota', 'PL'), ('ni', 'SUBJ'), ('taph', 'NN'), ('icha', 'DEF'), ('qix', 'SUBJ'), ('aan', 'NN'), ('biiznas', 'F'), ('ii', 'ABL'), ('dhimm', 'NN'), ('icha', 'DEF'), ('aaf', 'DAT'), ('gadaa', 'NN'), ('tiin', 'INSTR'), ('dhaabbat', 'NN'), ('aa', 'DAT'), ("bal'aa", 'NN'), ('dhaan', 'INSTR'), ('maaykiroosooftii', 'NN'), ('dhaaf', 'DAT'), ('meeq', 'NN'), ('aan', 'INSTR'), ('olol', 'NN'), ('aaf', 'DAT'), ('amanamummaa', 'NN'), ('dhaan', 'INSTR'), ('gambeelaa', 'NN'), ('tiin', 'INSTR'), ('ollaa', 'NN'), ('rraa', 'ABL'), ('hooggans', 'NN'), ('aan', 'INSTR'), ('kaansar', 'NN'), ('iin', 'INSTR'), ('kominikeeshin', 'NN'), ('iin', 'INSTR'), ('tuurizim', 'NN'), ('iif', 'DAT') , ('demookraasii', 'NN'), ('tiin', 'INSTR'), ('demookraasii', 'NN'), ('dhaan', 'INSTR'), ('vaayires', 'NN'), ('iin', 'INSTR'), ('keessummeess', 'NN'), ('aan', 'INSTR'), ('hanqin', 'NN'), ('icha', 'DEF'), ('aa', 'DAT'), ('haadh', 'NN'), ('olee', 'PL'), ('een', 'INSTR'), ('ameerikaa', 'NN'), ('dhaan', 'INSTR'), ('gammach', 'NN'), ('uuf', 'DAT'), ('gammachuu', 'NN'), ('dhaaf', 'DAT'), ('gammachuu', 'NN'), ('dhaan', 'INSTR'), ('kantiibaa', 'NN'), ('tti', 'DAT'), ('kotab', 'NN'), ('eef', 'DAT'), ('callis', 'NN'), ('aan', 'INSTR'), ('laaffis', 'NN'), ('aan', 'INSTR'), ('ilmasaa', 'NN'), ('tiif', 'DAT'), ('ummam', 'NN'), ('aan', 'INSTR'), ('falmii', 'NN'), ('dhaan', 'INSTR'), ('cir', 'NN'), ('oo', 'ABL'), ('quuqamt', 'NN'), ('oota', 'PL'), ('ni', 'SUBJ'), ('beellam', 'NN'), ('icha', 'DEF'), ('tti', 'DAT'), ('lafee', 'NN'), ('tti', 'DAT'), ('dinqisiifatt', 'NN'), ('oota', 'PL'), ('ni', 'SUBJ'), ("baay'in", 'NN'), ('ni', 'SUBJ'), ('boongaa', 'NN'), ('tti', 'DAT'), ('abdachiisaa', 'NN'), ('dhaaf', 'DAT'), ('abdachiisaa', 'NN'), ('dhaan', 'INSTR'), ('wagg', 'NN'), ('icha', 'DEF'), ('aa', 'DAT'), ('dogoggor', 'NN'), ('aan', 'INSTR'), ('xiinsamm', 'NN'), ('uun', 'INSTR'), ('qorannoo', 'NN'), ('tiif', 'DAT'), ('saamichaan', 'NN'), ('icha', 'DEF'), ('aan', 'INSTR'), ('haaroms', 'NN'), ('icha', 'DEF'), ('wallaans', 'SUBJ'), ('aaf', 'NN'), ('dant', 'F'), ('aaf', 'INSTR'), ('ogg', 'NN'), ('eessa', 'M'), ('i', 'SUBJ'), ('ogg', 'NN'), ('eessa', 'M'), ('icha', 'DEF'), ('i', 'SUBJ'), ('ogg', 'NN'), ('eessa', 'M'), ('icha', 'DEF'), ('ogg', 'SUBJ'), ('eessa', 'NN'), ('icha', 'DEF'), ('aan', 'INSTR'), ('ogg', 'NN'), ('eessa', 'M'), ('icha', 'DEF'), ('aaf', 'DAT'), ('ogg', 'NN'), ('eessa', 'M'), ('aaf', 'DAT'), ('barnoot', 'NN'), ('icha', 'DEF'), ('aa', 'DAT'), ('xaaliyaan', 'NN'), ('ii', 'GEN'), ('lukk', 'NN'), ('uun', 'INSTR'), ('lub', 'NN'), ('ni', 'SUBJ'), ('saamunaa', 'NN'), ('dhaaf', 'DAT'), ('dhibbeent', 'NN'), ('aan', 'INSTR'), ('dhibbeent', 'NN'), ('aaf', 'DAT'), ('dhibbeentaa', 'NN'), ('dhaan', 'INSTR'), ('fooram', 'NN'), ('ii', 'GEN'), ('garaa', 'NN'), ('rraa', 'ABL'), ('leencaa', 'NN'), ('tti', 'DAT'), ('biyyaa', 'NN'), ('tiif', 'DAT')]