ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUTE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF COMPUTER SCIENCE

PART OF SPEECH TAGGER FOR AFAAN OROMO LANGUAGE USING TRANSFORMATIONAL ERROR DRIVEN LEARNING (TEL) APPROACH

BY

MOHAMMED-HUSSEN ABUBEKER

A THESIS SUBMITED TO THE SCHOOL OF GRADUTE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCINECE IN COMPUTER SCIENCE

February, 2010

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUTE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF COMPUTER SCIENCE

PART-OF-SPEECH TAGGING FOR AFAAN OROMO LANGUAGE USING TRANSFORMATIONAL ERROR DRIVEN LEARNING (TEL) APPROACH

BY
MOHAMMED-HUSSEN ABUBEKER

Signature of the Board of Examiners for Approval

	Name	Signature
1.	Dr.Dida Midekso,Advisor	_____
2.	Sebsibe H/mariam , Co-advisor	_____
3.	_____	_____
4.	_____	_____
5.	_____	_____

Dedicated to:

1. My father Abubeker Abdellah

   (Sh/Abubeker Kanshusa)

2. My children:

   Hamad Mohammed-hussen

   Hawra’i Mohammed-hussen

Acknowledgments

First and foremost thanks to the creature of the entire universe, almighty God, ALLAH for helping me to realize this work. *Alhamdulillah!*

Second, I am heartily thankful to my advisor Dr.Dida Midekso, whose encouragement, guidance and support from the initial to the final level of this work enabled me to understand the subject. Without his encouragement and constant guidance, I could not have finished this thesis. Dr.Dida was always there to listen and to give me advice, to proofread and mark up my papers and chapters. 'Thank you Dr.'

A special thank goes to my co-advisor, Sebsibe H/mariam, who is most responsible in helping me on the technical part of my research. Sebsibe was special in that he showed me different ways to approach a research problem and the need to be persistent to accomplish any goal. Really I appreciate your friendly support Sebsibe!

I am also greatly indebted to many of my friends who directly or indirectly supported me in my thesis. Particularly many thanks go to Getachew Mamo, Solomon Asres, and Tefera Ayelew (Tafy).

It is my pleasure to express my gratitude wholeheartedly to my family. 'Thanks you' my father, Abubeker Abdellah (Sh/Abubeker Kanshusa)! Your zeal of knowledge, passion and personal integrity thought me a lot in my life. May Allah enlighten your grave! Many thank also to my mother Zukra Usman who thought me practically how to face challenges in my life. Thank you Mama, for what I know! I am extraordinarily fortunate to thank to my brothers Bayan Abubker, Sa'ud Abubeker, Abdulkadir Abubeker. Really your contribution to this works and to the whole of my life is beyond a shadow of a doubt!

Word fail me to express my appreciation to my wife Alfia Hassen whose dedication, love and persistent confidence in me, has taken the load off my shoulder. I owe her for being unselfishly let her intelligence, passions, and ambitions collide with mine.

Finally, I would like to thank everybody who was important to the successful realization of thesis, as well as expressing my apology that I could not mention personally one by one.

Table of Contents

Symbols

( ) what is inside is optional except site (source) of documents used

‘ ’ what is inside is the meaning of a word for Afaan Oromo in English

/ Separates words from their tags

Acronyms and Abbreviations

ANN	Artificial Neural Network
ATC_0	First Annotated Temporary Corpus
ATC_1	Second Annotated Temporary Corpus
ATC_2	Third Annotated Temporary Corpus
HMM	Hidden Markov Model
IR	Information retrieval
POS	Part of speech
NLP	Natural Language Processing
SR	Set of rules
TEL	Transformational Error driven Learning
WTL_i	Word list that contains Temporary corpus i. (i=1, 2, 3….)

List of Tables

List of Figures

Abstract

The purpose of this research is to develop part-of-speech tagger for Afaan Oromo using Transformational Error driven Learning (TEL) approach and compare it with other approach. Most natural language processing systems use part-of-speech (POS) tagger as a one of their component in their system. Specially, it is very significant for developing parser, machine translator, speech recognizer and search engines.

Afaan Oromo literatures on grammar and morphology are reviewed to understand nature of the language and also to identify possible tagsets. Based on this, 18 tagsets are identified and used on 223 sentences (1708 words) for the experiment.

The study customized Brill transformational error driven learning tagger for Afaan Oromo. Some template in the original Brill tagger was modified to fit Afaan Oromo morphological nature. After training data is analyzed for its appropriateness using learning curve analysis, the study used 10-fold validation method for the experiment. Moreover experiment was conducted to determine the percentage of training data for contextual and lexical rule learner. Best accuracy of the tagger was achieved when contextual rule learner training data is 35% and lexical rule learning data is 65%.This shows the morphological rule dominance over contextual rule for the language.

After modification on the templates of the Brill's tagger about 2.44% improvements over the original Brill tagger was achieved. This means 80.08% accuracy of the tagger was achieved in modifying the templates where the accuracy of the original tagger is 77.64%. Error of the modified tagger was also analyzed for further improvements using confusion matrix for the tagger.
The result obtained in both original Brill tagger and modified Brill tagger is compared with Hidden Markov Model approach (bigram and unigram approach).The comparison shows that Brill tagger is by far better than Hidden Markov Model in all the cases for Afaan Oromo i.e Hidden Markov Model accuracy for bigram approach is 70.63% and for unigram 68.08% whereas that of original Brill tagger without modification is 77.64 and 80.08% for modified Brill tagger.

Keywords: Natural Language processing, parts of speech tagging, Brill Tagger, Transformational Error driven Learning, Hidden Markov Model, Bigram, N-Gram.

CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND

Language is one of the core aspects of human behavior. It has a major role in our day to day activities. In its written form it provides a means to keep information and knowledge for long period of time and pass it to next generation. In its spoken form it can be used as a way to cooperate our day to day activities with others [4].Those languages that can be learned from environment and used for day to day communications by human beings are known as natural language [1]. Afaan Oromo, Amharic, Tigrinya, English, French and Arabic are some examples of natural language.

Language can be explored in various disciplines. Each of these disciplines can have their own set of problems and means to address them. For example, linguists are concerned with the structure of the language, how words are arranged to form sentences and why certain word arrangement will not produce correct sentence. On the other hand, computational linguists develop computational theory of the natural language and model based on concept of algorithm and data structure of computer science [8].

Computational linguistic (also known as Natural Language Processing (NLP)) is a field of study that processes natural language [3]. The goal of this field is to get computers perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech [3]. It also aims at designing and building applications that can understand, imitate language that human can use to the extent that it is possible to communicate with computer in the same way as we communicate with a person [12].

Understanding of natural language involves knowing of what concepts a word or phrase stands for and how to link those concepts together in a meaningful way. Natural language is easy for humans to learn and use, but it is the hardest for a computer system to understand. This is due to highly ambiguous nature of natural language [19]. For instance, consider English sentence “Flying plane can be dangerous”. This sentence creates difficulties to a computer software program as it lacks both knowledge of the world and experience with linguistic structures. Which interpretation is chosen? The pilot is at risk, or that the danger is to people on the ground? Should “can” be analyzed as a verb or as a noun? Which of the many possible meanings of “plane” is relevant? Depending on context, "plane" could refer to, among other things, an airplane, a geometric object, or a woodworking tool. How much and what sort of context needs to be brought to bear on these questions in order to adequately disambiguate the sentence. As we can see extensive knowledge about the outside world and the ability to relate situations is required by computer program to understand natural language. Therefore, natural language understanding and processing integrates multidisciplinary knowledge and concepts.

According to [8], natural language processing can be classified into many subfields. Some of these subfields are discussed below.

- Phonetic and Phonology: studies how words are related to the sounds. Speech synthesis, speech recognition etc are categorized under this subfield.

- Morphology: concerned with how words are constructed from more basic meaning units called morphemes. Like sentence grammar, we have also word grammar or rule that will govern how words can be formed from other words or changed by a bit modification to other words. Understanding of this word grammar is very important in many natural language processing areas. The internal structure of words is essential for information extraction and processing. Therefore, investigation and representation of word grammar is essential to facilitate searching and access of information. Morphological analyzer and synthesizer are part of this computational morphology.

- Syntactic study: deals with how words can be put together to form right sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases. Different languages may have different word categories, or they might associate different properties to the same word. Such behaviors of natural language word are studied under this subfield. Part of Speech Tagging (POS), word-category disambiguation is grouped under syntactic study of natural language.
- Semantic study: concerned with what words mean and how these meanings come together in sentences to form meaning of sentences. It is the study of context-independent meaning, the meaning a sentence has regardless of the context in which it is used. Semantic parsing of a sentence is included under this subfield.
- Discourse: studies how the instantaneously preceding sentences affect the interpretation of the next sentence. This information is especially important for interpreting pronouns and for interpreting the temporal aspects of the information conveyed.

1.2 PROBLEM STATEMENT

Computational study on local languages of Ethiopia is in its early stage even though there are some initiatives that have been undertaken. Afaan Oromo is one of the major languages of the country and it is spoken by millions of Ethiopians and people in the neighboring country [1]. Currently, the language is an official working language of the Regional Government of Oromiya and it is a medium of instructions for primary schools and training institutes in the region. In addition to these, numerous newspapers and books are published in Afaan Oromo. Moreover, several radio and television programs of the country use the language in their broad cast program. Therefore, the need of information retrieval (IR), data processing and online machine translation using Afaan Oromo is increasing from time to time.

Lack of natural language processing and tools that understand Afaan Oromo text such as part of speech tagging, morphological analyzers, and online machine translation that will translate text from any language to Afaan Oromo or vise versa is the major problem for many people that use the language as their means of information processing and retrieval.

In particular, absence of part of speech tagger for Afaan Oromo will be the main obstacle for researchers in the area of machine translation, spell checkers, dictionary compilation, and automatic sentence parsing and constructions. All of these NLP systems must use part of speech tagger as their preprocessor components for their best performance.

Research on POS for Afaan is very rare. The work in [21] which is the first attempt used one approach, Hidden Markov Model, for Afaan Oromo part of speech tagger, and evaluated its performance. Before implementing part of speech tagger for the language different approaches and algorithms have to be analyzed and evaluated for their robustness and efficiency. Hence, the aim of this thesis is to use the other approach, Transformation Error driven Learning (TEL) approach, for Afaan Oromo POS tagging and analyze its performance in comparison with some other approach.

1.3 OBJECTIVE OF THE STUDY

1.3.1 General Objective

The general objective of this study is to develop part of speech tagger for Afaan Oromo Text using TEL approach.

1.3.2 Specific Objective

The specific objectives are:

- To review literature in natural language in general and on POS of various language in particular.
- To review the nature and behaviors of Afaan Oromo letters, words and word categories and construction.
- To explore lexicon and part of speech of Afaan Oromo text.
- Adapt design for POS of Afaan Oromo text.
- Compare the result with other approaches that were used for developing POS tagger for Afaan Oromo.

- To draw conclusions based on experimental results and recommend further research area in the future.

1.4 SCOPE OF THE STUDY

The scope of the research is limited to exploring the likelihood of TEL approach to design automatic POS tagger for Afaan Oromo in comparison with Hidden Markov Model (HMM). The study also focuses on broadly categorized part-of-speech of the language for word lexical categorizations.

1.5 LIMITATION OF THE STUDY

The main limitation in conducting this study is the absence of readily available annotated corpus for NLP of the language.

1.6 RESEARCH METHODOLOGY

The following methods have been used for the accomplishment of this thesis.

1.6.1 Literature Review

Resources like books, previously conducted research papers, journals and other documents from the Internet are used for the purpose of understanding morphology of words, part of speech of language in general and Afaan Oromo in particular. Techniques and approaches appropriate for development of Afaan Oromo part of speech tagging is also reviewed. Moreover, suitable learning algorithms are explored.

1.6.2 Data Collections

It is difficult to design part of speech tagger without detail knowledge of the language. Computational knowledge is not enough by itself. Therefore, linguistics and experts in the area of Afaan Oromo have been consulted.

1.6.3 Testing and Evaluation

Due to the absence annotated corpus, small amount of training data is used in this research. A 10-fold cross validation method is used for training and testing of the system.

1.7 TOOLS AND TECHNIQUES

Various tools have been analyzed: MontyTagger, Brill Rule Based Tagger and general purpose natural language toolkit (nltk). All of these tools incorporate original Brill tagger that uses transformational rule based tagger. The Original Brill Tagger was selected in this research for testing and training. C and Perl programming language are used for coding.

1.8 APPLICATION OF RESULTS

Automatic part of speech tagging is applicable in many areas of NLP. For most of natural language processing (NLP) systems, it acts as one of the most important components. Higher forms of NLP system such as machine translations, spell checkers, speech recognitions, question and answering, semantic and syntactic parsing, automatic dictionary (lexicon) compilations, and automatic sentence constructions use part of speech tagger as their component. In addition to that of assigning word class of a given word, it can be used for statistical work such as counting the distribution of different word classes in text corpora and to provide words inflectional behaviors.

1.9 ORGANIZATION OF THE THESIS

The rest of this thesis is organized as follows. Literature review on part-of-speech, and related works in part of speech tagging of other language and Afaan Oromo language, are discussed in Chapter Two. In Chapter Three of the thesis, Afaan Oromo part of speech and the tag set of the language are discussed. Detail design process, algorithm and architecture of the Afaan Oromo part of speech tagger presented in Chapter Four. Chapter Five presents the data collection process and procedures of the experiments. Detail training, testing steps and results are also addressed in this chapter. Lastly, conclusions and recommendations of this thesis work are presented in Chapter Six.

CHAPTER TWO: LITERATURE REVIEW

2.1 INTRODUCTION

This chapter discusses survey of POS, approaches that have been used to develop POS. Researches that have been conducted on POS for languages other than Afaan Oromo and natural language processing researches conducted on Afaan Oromo are also explored.

2.2 SURVEY OF PART OF SPEECH TAGGING

Grammatical class or word group in a language to which words may be grouped based on how they are utilized in a sentence is known as part of speech [10]. Although many former scholars had their own lists of parts of speech, it was Thrax's set of eight POS which became the basis for practically all subsequent part of speech descriptions of Greek, Latin, and most European languages [10].

The POS can be generally classified as: closed class types and open class types. Closed typed POS has relatively fixed membership. Some of closed classes of part of speech in English language are:

- Prepositions: on, down, up, of, over etc
- Determiners: An, a, the
- Pronouns: she, he, you, I, etc
- Conjunctions: and, but, or
- Numerals: one, two, second, third, etc.

On the other hand POS such as nouns, verbs, adjectives and adverbs are open classes because new nouns, verbs, adjectives and adverbs are frequently coined or borrowed from other languages.

Tagging is a process that accepts string of untagged word and provides appropriate tag for each of individual words in a sentence [10]. It is a method to categorize words based on their grammatical or syntactic group in a sentence or a corpus.

During tagging process, symbols (labels) are assigned to each word in the sentence that tells us the word's category in the given sentence and probably other additional information (like first person singular /plural for noun). These labels are termed as “tags”. Because tags are generally applied to punctuation, tagging requires that the punctuation marks (period, comma, etc) be separated off the words.

The input to a tagging algorithm is a string of words and a specified tag set . The output is a single best tag for each word. As an example, consider the sentence ***Falmataan kitaaba seenaa dubbise*** which means ‘Falmata read history book’. After tagging the above sentence the output may looks like: **Falmataan/NN kitaaba/NN seenaa/JJ dubbisse/VV**. The labels NN, JJ and VV are tags that imply the part of speech (NN for noun, JJ for adjective and VV for main verb) of each word.

Tagging can be done, manually or automatically. Manual tagging is done by hand and correct tag is assigned after group discussions of experts on each word’s tag. Therefore, it needs too much time for large amount of corpus. Moreover, it needs acquisition of knowledge about languages grammar and structure of sentence to be tagged. Automatically it can be done by using POS tagger software.

The significance of part of speech tagging for language processing is the large amount of information it gives about a word and its neighbors. It can be used in stemming for information retrieval (IR), since knowing a word’s part of speech can help to tell us which morphological affixes it can take. They can also improve informational retrieval applications by selecting intended words from a document. Moreover, the role of automatic assignment of part of speech in parsing, in word-sense disambiguation algorithms, speech synthesis, speech reorganization and in shallow parsing of texts to quickly find names, times, dates, or other named entities for the information extraction will not be underestimated [3].

2.3 APPROACHES TO PART OF SPEECH TAGGING

So far it has been discussed about part of speech tagging and related concepts. This section focuses on brief descriptions of different approaches that exist for part of speech tagging. Various algorithms and methods were devised to tackle part of speech tagging problem. There are several approaches that can be followed in designing part of speech tagger. Most common of these are stochastic approach, rule based approach, hybrid approach and Artificial Neural Network.

2.3.1 Stochastic Approach

Stochastic method which can also be called statistical method is the method which is based on the statistical information or on the probability of occurrence of words. Any approach that uses probability or statistic information can be grouped under this approach.

This method, in which tagging is based on frequency or probability of a given word posses a particular tag, is one of the most common approaches. Tag with higher frequency or probability is the one that will be assigned to the ambiguous word in the sentence. An individual word conditional (contextual) probability is the core idea for this approach. That is, a given word W in a sentence is assigned a tag t based on certain context (conditions), if the probability of the tag t of that given word W in that context is maximum.

The context defines the generality of the approach and as the context increases its theoretical performance increases and vise versa. However, increasing the context demands huge collection of data set which is not feasible in most cases.

The context also defines dependence of the tag of the word on its contextual words. This generally defines the N-gram model where N defines the number of words to be considered in the context (N>=0). N=0 means the tag of a word is completely independent of any of the word in context. N-gram approach calculates the frequency or probability of a given sequence of tags occurring in the sentence.

2.3.1.1 Hidden Markov Model

One of the most commonly known stochastic approaches is Hidden Markov Model approach (HMM). The basic idea used in HMM is that most likely tag of a given word in sequence of words each possessing one or more tag is chosen by calculating the probability of all possible sequence of tags and then selecting the sequence with the maximum probability [8, 10].

This approach can be used in automated tagging schema because they rely critically upon the calculation of probability on the output order or sequence. Baum-Welch Algorithm [10] which is also termed as Forward-Backward Algorithm is implemented by HMM to avoid this problem. With the help of this algorithm it can automatically learn/ train HMM.

Two assumptions are made while using HMM model.

1. Every word is not related with all the other words and their tags.
2. Every word's probability depends on the N previous tags only.

Based on the above assumption, HMM taggers select order of tag (sequence) that will maximize the formula: P(**word**|**tag**)*P (**tag** |**previous n tags**) where **word** is the word to which we are going to assign the **tag**, the probability of that tag to be for that word in the sentence. In HMM, the entire sentence tag sequence is considered rather than individual word. However, for clarity we look at a single word case example.

A bi-gram HMM tagger is the one that produces its tag result **ti** for the unknown word based on the previous tag **ti-1** given word **wi** itself [10].

Consider two sentences: **Bu'aan barumsaa mul'ataa jira**, which means 'Advantage of education is being seen' and **Ibsaan kitaaba gudda mul'ataa kenne,** which means 'Ibsaa gave a big book to mul'ataa'. HMM tagger is expected to assign the correct tag for the word **mul'ataa** with the assumption that all other surrounding words are correctly tagged as follow:

Bu'aan/NN barumsaa/NN mul'ataa /jira/VV.

Ibsaan/NN kitaaba / NN gudda /JJ mul'ataa/kenne/VV.

In the first sentence the word **mul'ataa** is used as verb and in the second it is used as a noun.

Using bi-gram approach tag, the word **mul'ataa** can be assigned by considering the neighboring words and tags. Look at the sequence in the two sentences above for words surrounding **mul'ataa**:

Barumsaa/NN mul'ataa/???

Gudda/JJ mul'ataa/???

If we are to choose between VV and NN for word the **mul'ataa** above, we are expected to select the tag that has higher probabilities:

P (VV|NN) P (mul'ataa|VV) (1)

And

P (NN|JJ) P (mul'ataa|NN) (2)

Where P (VV|JJ) and P (NN|JJ) are tag sequence P (tag| previous tag) and P (mul'ataa|VV), and P (Mul'ataa|NN) are word probabilities P (word|tag). Tag sequence P(VV|NN) for example tell us how much it is probable to get VV(verb) if the previous tag is NN(Noun) and P(NN|JJ) tell us the probability of obtaining NN(noun) when the previous tag is JJ(adjective).

Assuming that we have probabilities for the above tag sequence in our corpora as:

P (NN|JJ) =0.05

P (VV|NN) =0.25

And lexical probabilities of this word as:

P (**mul'ataa**|VV) =.003

P (**mul'ataa**|NN) =.0004

Therefore, the bi-ram approach of HMM will calculate the maximum of

P (VV|NN) P (mul'ataa|VV) = (0.25) (0.003) =.00075

P (NN|JJ) P (**mul'ataa**|NN) = (0.05) (0.0004) =.00002

Based on this calculation, the tagger assigns **mul'ataa** as VV which has maximum value.

2.3.2 Rule Based Approach

Rule based POS is a dominant approach in computational linguistics and natural language processing that uses large database that contains hand written rules necessary to remove or minimize ambiguities. These rules implement contextual information also known as context frame rule and morphological information. And thus, it provides lexical, syntactic or morphological information of ambiguous or unknown word of the language.

As discussed in [10], this algorithm uses two stages architecture. The stage that utilizes a dictionary to categorize each word in the word list as probable part of speech and the stage which uses large list of hand written disambiguation rules to assign this list to a particular part of speech. The ENGTWOL tagger [3] is such a two stage architecture tagger. In the first stage of tagging, every word passes through a two level lexicon transducer and possible values of part of speech are returned. For example, the sentence **Ibsaan barataa dha** which means

'Ibsaa is student' would return a line for every possible tag as indicated below:

Ibsaan/ NN/ SG /PROPER

Brataa /NN/SG/ COMMON

dha /VV/ AX

After all possible tags are returned a set of constraints is applied to the input sentence to discard incorrect part of speech. These constraints are mainly used to remove irrelevant tags with context rules.

Revision of tags of the neighboring or surrounding word is done using context frame rule. For instance, let **w** be a word in a lexicon and assigned more than two part of speech say **NN** (noun) and **VV** (verb).This means that the word is tagged as w/**NN/** in the first stage. The second stage of rule based tagging selects the appropriate tag contextually by looking at the surrounding. For example, if the word **w** is preceded by the word **to** (infinitive) then it will be tagged as "**to/ w/VV**" in the second stage. Such type of context frame rule will change one type of tag into another type of tag.

Another problem that arises in the first stage of tagging words based on the existing lexicon is problem of unknown words. For most languages, a given lexicon (dictionary) may not include

all words of the language under study. During tagging, a new word that doesn't belong to the lexicon may be encountered by the tagger. Contextual rules for unknown words are used to solve the problem. In English, for example, if unknown word is preceded by determiner (DT) and followed by a noun (NN), then the unknown word X will be tagged with an adjective (JJ).This can be expressed as:

DT+X+NN=X/JJ

Another rule that can be used to speculate category of words or word class is morphological information of the word itself. In English, if the word "boys" is not in the lexicon, it will be analyzed morphologically as "boy"+ "s". From this information it can be classified as plural noun and tagged as 'boys/NPL'.

In addition to context frame rule and morphological information in some systems capitalizations and punctuations can be used to get information about tag of a given word. For instance, in English and Afaan Oromo the first letter of a proper noun is always upper case the whereas there is no sense of capitalization in Amharic.

All rule based taggers have the following components that contribute for their effective tagging words.

A POS tag set: a list of possible word classes that is used during tagging process. It is the core component of the rule based tagger. All words are given unique codes based on their grammatical word categories.

Lexicon (or Dictionary): Lexicon contains words and the corresponding part of speech tags taken from the tag set. The tagger starts its processing by first looking up each word in the lexicon. For example, in English language the word 'has' is written in the lexicon as shown below with its corresponding part of speech:

Has	**V**	**MD**

Tag changing rule-rules that provide information about appropriateness of a given tag based on the context. These rules can be contextual rules, which are predefined rules based on

context, or it can be lexical rules that help the tagger to make reasonable guess. Contextual rules modify the tag of the word based on the surrounding of word whereas lexical rule uses morphological behavior the word itself as discussed above.

2.3.3 Transformational /hybrid Based Approach

This method of tagging uses both rule based and stochastic approach. As a rule based, it uses grammatical rule that determines what tag should be assigned to what word. As stochastic approach, it automatically learns and induces the rules from the given data.

Among Transformation Based tagger, Eric Brill transformational based tagger is the best known tagger [10].

Brill transformation based tagger starts by simple default heuristics and incrementally learns rules to cover special cases. This can be done by first assigning every word in the corpus its most likely tag using the initial tagger. After that lexical rules are inferred for the word if it exists. If the word is not in the lexicon a list of transformations (rules) is provided for determining the most likely tag for words not in the lexicon. Unknown words are first assumed to be nouns (proper nouns if capitalized), and then clues based upon prefixes, suffixes, infixes, and adjacent word co-occurrence are used to change the guess of most likely tag. Finally, contextual transformations are used to improve accuracy.

In this method, rules are learned by iteratively collecting errors and generating rules to correct them. General template is used to learn specific rules based on the error it encounters. Followings are some general templates:

Change tag **a** to tag **b** when:

1. the preceding/following word is tagged *z*
2. two word before/after is tagged *z*
3. one of the two preceding/following words is tagged *z*
4. one of the three preceding/following words is tagged *z*
5. the preceding word is tagged *z* and the following word is tagged *w*

6. the preceding/following word is tagged *z* and the word two before/after is tagged *w*
7. the current word is/is not capitalized
8. the previous word is/is not capitalized

For example, suppose that the initial tagger tagged 100 words in the corpus as verbs when they should have been nouns. This would produce the error triple :(Verb-Noun 100).

Suppose template number three is instantiated as the rule: *Change the tag from verb to noun if one of the two preceding words is tagged as a determiner*.

When this template is applied to the corpus, say it corrects 68 of the 100 errors. But it also creates 18 new errors by changing tags from verb to noun when they really should have been verbs. The error reduction for this template would then be 68-18=48. In this manner each candidate template is applied to the corpus and the number of corrections and errors it produced are counted. The template that yields the greatest improvement is added to the template library. Learning stops when no template can reduce the error rate by some predetermined threshold.

2.3.4 Artificial Neural Network Approach

An Artificial Neural Network (ANN) is a method of information processing paradigm that is motivated by the way biological nervous systems process information [29]. The most important feature of this paradigm is the structure of information processing systems. The system is composed of a large number of highly interconnected processing elements (neurons) working in harmony to solve specific problems. These units are highly interconnected by directed weighted links; associated with each unit as an activation value. Through this connection, this activation is propagated to other units. The interconnections of the neurons follow specific network architecture [31].

Unlike that of von Neumann machines which are based on processing-memory abstraction, Artificial Neural Networks are based on the parallel architecture of animal brains [30]. In Artificial Neural Network (ANN) learning from example is possible by configuring the system for a specific application such as pattern recognition or data classification.

This learning process can be done by adjustment for different features on the basis of the input data that is given to the network. This means the network itself is able to find properties from the input data given to it. It organizes its structure to reflect the properties of the given data. Hence, learning in this case is termed as adaptive learning. It will proceed as far as new data is available to the network. Mathematically learning process of artificial neural networks can be reformulated as function approximation tasks. This can be done using nonlinear function as approximating tools (i.e., linear combinations of nonlinear basis functions), where the parameters of the networks should be found by applying optimisation methods.

The type of learning algorithm implemented determines the way the internal structure of an ANN is changed. Based on the learning algorithm implemented and internal structure of the Network, Artificial Neural Network can have several models.

In general, three entities characterize an Artificial Neural Network: the network topology or interconnection of neurons, characteristics of individual units or artificial neurons, and the strategy for pattern learning or training.

Based on the interconnection, ANN is classified as feed-forward and feed-backward topologies.

2.3.4.1 Feed-forward Artificial Neural Network

Feed-forward Artificial Neural Network allows signals to travel forward in one direction only; from input to output. There is no feedback or backward propagation, which means the output of any inner neuron (layer), does not affect that same layer. Feed-forward Artificial Neural Network seems to be straight forward networks that associate inputs with outputs. For feed-forward networks, every neuron in a given layer receives inputs from layers below, that is, from layers nearer to the input layer, and sends output to layers above, that is, to layers nearer to the output layer. For such networks, given a set of inputs, the input vector, from the neurons in the input layer, the output vector is computed by a succession of forward passes, which compute the intermediate output vectors of each layer in turn using the previously computed signal values in the earlier layers.

2.3.4.2 Feedback Artificial Neural Network

Feedback Artificial Neural Network can have input values traveling in both (forward and backward).This can be done by introducing loops in the network [29].In the forward pass, error is calculated from outputs and used to update output weights. In backward pass, error at hidden nodes is calculated by back propagating the error at the outputs through the new weights and hidden weights are updated.

These types of topologies are very strong than feed forward in that it can distinguish data that is not linearly separable and can be more complex than feed forward topology. The Feedback networks are changing their 'state' continuously until they reach an equilibrium point; hence they are dynamic. Until the input changes the equilibrium point that was attained previously is not changed. However, in the presence of new data, the equilibrium will be changed through adaptation based on the new data presented. Fig 2.1 shows the feed-backward Artificial Neural Network. There is sense of time or memory of previous state in this network unlike that of feed forward neural network which does not have sense of time or present output is not affected with previous state.

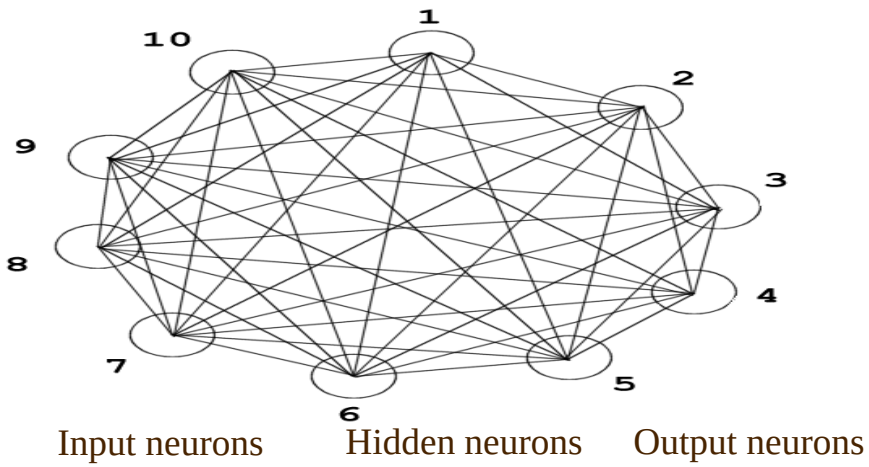

Figure 2.1 Feed backward Artificial Neural network

2.4 RELATED WORK

In this section related researches that have been conducted on natural language and particularly those works on Afaan Oromo part of speech tagger are presented.

2.4.1 Previous work on other Language

Transformation-based error driven learning, which automatically learns rules from a training corpus, is presented in the work of [23]. Most of the limitations of traditional rule based tagger were removed as it infers rules from a training corpus and it uses some corpus context. Eric Brill introduced a POS tagger in 1992 that was based on rules, where the grammar is induced directly from the training corpus without human intervention or expert knowledge [23]. Using this training corpus the system derives lexical and contextual information. This information helps the system to learn how to infer the most appropriate tag for unknown word. Based on the training corpus's tag set, the tagger can be used to annotate new, unannotated corpora. This approach needs only very small amount of linguistic knowledge built into the system [24].

Hybrid method for tagging Arabic Text [25] is another research in the area of POS tagging. It uses rule based and machine learning for Arabic language. This work attempts to improve the performance of tagging process by checking the affix patterns of word by using combination

of affix rules, the patterns of the word and a set of grammatical rules. First, rules are applied on post-position, ending of a word and patterns. After rules are applied, the errors are corrected by adopting a memory based learning method (MBL). Exceptional cases of rules and methods how to handle them is provided to the memory based learning. The memory based learning is an efficient method to integrate various sources of information and handling exceptional data in natural language processing tasks. Encouraging result was obtained in this work also.

2.4.2 Previous work on Amharic Language

In this section, previous works that have been conducted in the area of part of speech tagger for Amharic language is discussed briefly.

In automating part of speech tagging for Amharic language, the work of [11] used Hidden Markov Model approach for the language. Viterbi algorithm was used to develop the tagger. With small amount of corpus, the result obtained in the experiment shows more than 85% accuracy.

Application of multilayer preceptron neural network for tagging [12] parts of speech for Amharic language was another research work on Amharic. The researcher used three layers multi layer preceptron (MLP) network. Multi layer feed-forward network with back propagation algorithm was designed because of its capabilities of expressing non-linear decision. Neighborhood context, localized information, for tagging of words in a small tagged corpus was considered in this work. The network takes as an input set of words that fall into a window of pre-specified size centered on the target word to be tagged. Corresponding tag for the target word is an output of the system. The result of the MLP part of speech tagger was promising.

The work of [20] uses hybrid approach, both rule based and neural approach for Amharic language. Neural network output anomaly was corrected by rule based approach. Back propagation algorithm and Brill transformation based learning method are adopted for the development of Amharic tagger. Relatively, a corpus with large amount of data is used to train and test the tagger. The experimental result of this work indicates that 91% and 94% accuracy

for rule based and neural network tagger, respectively. But the result reached 98% when the experiment was conducted on the hybrid tagger.

2.4.3 Previous work on Afaan Oromo

There are a number of works that have been done in the area of natural language processing for Afaan Oromo languages, even though the actual development of working systems is yet the gap that should filled. In this section, some of natural language processing researches conducted on Afaan Oromo are discussed.

Afaan Oromo word stemmer which breaks a word in to its root and affix is one of the research areas for the language. The work in [7] discusses detail characteristics of words of Afaan Oromo language. In this work, Afaan Oromo language word stemmer algorithm was developed.

Automatic sentence parser for Afaan Oromo [2] was also another work in the area of Afaan Oromo language processing. In this work, the researcher used his own rule based tagger and morphological analyzer for the designed parser algorithm.

Another work in [8] discusses development of morphological analyzer for Afaan Oromo text. Detail nature of Afaan Oromo morphology was discussed in this work. Corpus based learning was used for algorithm to learn morphology. Evaluation of the result shows above 80 % accuracy of this work.

The work of [21] on Afaan Oromo part of tagger again is another research for in the area. Hidden Markov Model approach (statistical) was applied in this research. Part of speech tag sets of Afaan Oromo was described in detail in this work. Around 900 words were used for training purpose and around 87% accuracy was obtained.

As it is indicated in chapter one, the aim of this work is to develop part of speech tagger using transformational error driven approach and compare the result obtained by the tagger with the result that was obtained by using Hidden Markov Model.

CHAPTER THREE: WORD CLASS OF AFAAN OROMO LANGUAGE

3.1 INTRODUCTION

Qubee which is Afaan Oromo alphabet is based on the roman orthography (or Latin alphabet). Afaan Oromo phoneme range includes twenty seven consonants and ten vowels. The schematized syllable format of Afaan Oromo can be represented as: CV (V) (C), where C is a variable for 'consonant', V is a variable for 'vowel', VV represents a long vowel, and items in parentheses are optional [16]. Table 3.1 below shows Afaan Oromo consonants [1].

Table 3.1 Afaan Oromo Consonants

Labial	***Alveolar/dental***	***palatal***	***velar***	***glottal***
P	t	ch	k	‘
b	d	j	g	h
ph	x	c	q	
f	dh	sh		
v	s	ny		
m	z	y		
w	n			
	l			
	r			
	l			

Double consonant (gemination) also will cause variation in meaning of Afaan Oromo language. The word bada means 'bad' whereas **baddaa** imply 'highland'.

The language's alphabet (single letter) is constructed from one character symbol or digraph (double character symbol) like "dh", "ny", "ph", "sh". Gemination is allowed for single (one character) symbol like *ho**ff**aa* which means 'light' whereas gemination for digraph (double character) symbol is not allowed in the language. For example, ***qophii*** which means 'readiness' can not take the form qophphi**.**

The language has also five set of short vowels and five set of long vowels. Table 3.2 shows lists of these vowels. Both of these vowels can appear in initial, medial and final positions.

Table 3.2 Afaan Oromo Vowels

Short vowels	***Long vowels***
a	aa
o	oo
u	uu
e	ee
i	ii

The five long vowels can be obtained by doubling the corresponding short vowels. The difference in length of the vowel induces difference in meaning. For instance the word *'laaf**e**'* means 'to become weak' while *'laf**ee**'* is 'bone'.

Punctuation pattern in Afaan Oromo is similar to that English language and other languages that use Latin orthography [2].

3.2 LEXICAL CLASSIFICATIONS

Lexical classification is a classification based on word semantic or semantic coherence rather than synthetic meaning of the word in a sentence. In this categorization of words, positional role of the word is not considered. For example, semantic word classification or coherence for noun can be name of place, people, things or places. This means, word indicating name of place, things, people or places are categorized as noun whatever the role it plays in the sentence.

However, modern linguistic argues that word categorization should be based on the contextual position of the word it has or based on the role it plays in that sentence [21]. According to this concept, it is possible for a word to change its tag based on its context it has in the sentence. For example, if we consider the two sentences below:

1. ***Ibsaan marqa garbuu nyaate***. Which means 'Ibsaa ate barley porridge' and

2. ***Ibsaan garbuu facaasee***. This means 'Ibsa sow barley'.

In the first sentence, the word ***garbuu*** 'barley' describes the type of porridge and acts as adjective; however in the second sentence it acts as noun. Even though lexical meaning of barley is noun, based on its contextual meaning it can act as other word class (tag) in this case adjective.

In all natural languages, there is a standardized word order in a sentence. For example, in English and French language, word orders obey *Subject -Verb- Object* order. This word order changes to *Subject -Object-Verb* for Japanese and Amharic languages [21]. Afaan Oromo word order also obeys the latter order i.e. *Subject-Object-Verb*. If words of the language do not follow this standardized word order, the sentences may convey vague meaning or totally loose their meanings.

Understanding of this syntactic structure of sentence can help us to know the relationship between words which in turn leads us to categorize them correctly.

Another means that can be used for word categorization is morphological derivation of the language. Morphological derivation changes word category from one word class to another word class. The following examples illustrate the concept.

Laafaa means 'weak' -------***laafinna*** meaning 'weakness' --->adjective class was changed to noun class.

Jabaa means 'strong'------***jabeessuu*** meaning 'to make strong'---->from adjective class to verb class.

For that reason, the aim of this chapter is to identify the lexical categories of Afaan Oromo word based on their contextual position they occupy in a given sentence. This lexical classification is used to annotate the corpus that is going to be used in implementation section of this thesis.

3.3 WORD CATEGORIZATION

Words are the basic part of a given language. The arrangement of word or their combination depends on the rule or grammar of that language. The combination of these words on the bases of the language gives us sentences. The meanings of these sentences depend on each word of the sentence and the way they are arranged. However, the extent to which a given word determines the meaning of a sentence depends on the contribution of that word. All words do not 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 [12].

In order to categorize a given word into some class, there are three clues that can be used. These are: meaning of the word, the form (shape) of the word and the position or environment of the word in the sentence.

Meaning of the word criterion is used to deduce the meaning the words imply. Some of these meanings can be action meanings which are conveyed by verbs, entity words that can be implied by nouns, state words that can be expressed by adjectives or that can denote the manner in which something is done. For instance, words like **mana**, means 'house', **gurbaa** means 'boy', **laga** means 'river' can be grouped into same group or category based on the above criteria because they all refer to the name of a person, place or objects which can be generalized as entities. On the basis of this common feature (i.e. entities) the above words are grouped as noun. In a similar manner words like **figi** means 'run', **barressi** means 'write', **dubbisi** means 'read' imply common actions. Based on these common actions they can be categorized as verb.

A form or shape criterion uses morphological information of the word to be categorized. The fact that certain affixes are used with particular group of words makes this criterion possible to use. In English, abstract nouns end with bound morpheme /-ion/ like repetition, presentation, demarcation etc. In the same manner, adjectives end with /-able/ affix as readable, movable, tolerable etc. Inflection is one common form of morphological changes. This can be a change of singular noun to its plural form by adding morpheme –s at the end of the word or it can be a change of verb from its present form to its past form by adding /-ed/post fix [12].

In Afaan Oromo, nouns can be inflected by adding /-oota/, /-oolee/, /-eenn/, /eewwan/ etc by adding at the end of the noun. For example, word like **saree** which means ‘dog’ **sa’a** which means ‘cow’ form their plural as sar***oota*** ‘dogs’ and *sa**’oole*** ‘cows’ . Some verbs can add post fix /-e/to change their tense into past. For example,***qab*** which means 'to catch' can change to its past form by adding /-e / as ***qabe*** which means ‘caught’.

The other type of word categorization criterion is position of the word in the sentence. This is also known as environment of (surrounding of the word) the word in the sentence. It is the criterion that carefully looks at where the word has occurred in the sentence and word surrounding or neighboring. For example, in Afaan Oromo language, verbs come at the end of the sentence. However, adjectives come after the noun they modify. For example, **farda *diima*** meaning ‘red horse’, ***Diima*** is adjective equivalent to English red whereas **farda** is same as ‘horse’ which is noun. Therefore, if we know that ***diima*** acts as adjective in the sentence we can directly deduce that the preceding word is noun.

In general, word categories can be identified by looking at the meaning (semantic) of that word, by looking at the form (morphology) of that word or by looking at the actual position (syntactic) of that word [12].

3.4 AFAAN OROMO WORD CLASS

Recent works in [21, 1] have stated that Afaan Oromo has five word classes: noun, ad-position, adverb, conjunction and verb. Each of these classes again can be divided into other sub-classes. For instance, noun class is categorized as proper noun, common noun, pronoun,. Preposition and postpositions are sub classes of ad- positions. 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 [20].

In this work (including subclasses of the above classification) eight part of speech (i.e. noun, pronoun, adjective, preposition, adverb, conjunction, verb and introjection) and numerals are used for word categorization. Each of this word classification is discussed below in detail.

3.4.1 Noun (‘Maqaa’)

Nouns are any word that can be used to name or identify place, object or ideas. Two types of grammatical genders exist in Afaan Oromo nouns. These are masculine and feminine, and the entire nouns of the language belong to one of these gender categories [12, 14]. Similarly, there are two numbers (singular and plural) which can be identified by the morpheme it adds. Plural form of a given noun can be formed by adding suffix to the root noun. Various types of suffixes can be added to transform a singular noun to its plural form. All of these suffixes change singular noun to plural without variation in meaning. The last vowel of the singular noun is dropped before the suffix is added. The suffixes which include -oota;- [w]wan,-een, -eelee, -iin,-[a]an,-oolee,-eewwan,-iilee etc are used to form plural. Table 3.3 depicts plural information for noun.

Table 3.3 Possible Plural Formation Affixes

Singular	***Suffix added***	***Plural***
Nama	-oota	Nam**oota**
	-een	Nam**een**
	-eelee	Nam**eelee**
	-iin	Nam**iin**
	-oolee	Nam**oolee**
	-ewwan	Nam**ewwan**

In general, more than 12 major affixes exist that can be used in plural formation for the language [15].

In Afaan Oromo language, there is no grammatical definite article which matches ‘*the*’ of English language. However, the suffix – ***(t) icha*** can be used in the same context as English language ‘*the*’ in masculine and **– *(t) ittii*** for feminine nouns. Ending vowels of nouns are dropped before these suffixes are added to the noun. For example: **mucaa** meaning ‘boy’, muc***icha*** ‘the boy’,**mana** ‘house’ ,man**icha** ‘the house’ ,durba ‘girl’ durb***itti*** ‘the girl’.

Unlike English definite article *the*, definite suffix in Afaan Oromo is used infrequently [14].

In most of the cases, Afaan Oromo nouns are found at the beginning of the sentences and they can have the role of subject or object in that sentence. For example, consider the following sentences:

Hawwiin mana barumsaa deemtee. This means ‘Hawii went to school’. **Hawii** is noun and acts as subject.

Saani marga dheede. This means ‘The cow ate grasses. Both **Sanii** and **marga** are nouns acting as subject and object respectively.

Nouns can also be derived from other word classes. Consider the following nouns which are derived from other classes [21].

Dherinna means 'height' derived from ***dheeraa*** means 'tall' which is adjective.

Qabeenyaa means ‘wealth' derived from ***Qaba*** means ‘has’which is verb.

Kufiinsa which means 'failure' derived from ***kufa*** which means 'to fail'which is verb.

3.4.2 Pronoun (‘Maqdhalaala’)

Pronouns are words that can be used in place of nouns. Similar to that of nouns, pronouns have number and gender. For example, **ishee/isii** which means ‘she’ is feminine (singular) whereas **isa** which means 'he' is masculine (singular) and **isaan** which means 'they' is plural can be masculine or feminine. Pronouns can also be categorized based on their functions and meanings in the sentence. These are personal pronoun, possessive pronoun, demonstrative pronoun, relative pronoun or reciprocal pronoun [21]. Table 3.4 illustrates personal pronouns that can be in the subject positions.

Table 3.4 Subject Position Personal Pronouns of Afaan Oromo

	1st person	***2nd person***	***3rd person***
singular	Ani(I)	Ati(you)	Isa/inni(he) Isii/ishee(she)
plural	Nuti(We)	Isin(you)	Isaan/Jarri(they)

For instance, the following sentence illustrates the above fact.

Hawwiin kitaaba dubbifte. ‘ Hawi read a book’.

Isiin kitaaba dubbifte. ‘ She read a book’.

Barssisaan baratoota qoramaatee. ‘The teacher examined the student’.

Inni isaan qoramaatee. ‘He examined them’.

Possessive pronouns are pronouns that indicate ownership of something. Table 3.4.2.2 shows personal pronouns that can be in the subject positions.

Table 3.5 Possessive Personal Pronouns of Afaan Oromo

	1st person	***2nd person***	***3rd person***
singular	Kiyya/kooti/tiyya(mine)	Kee(yours)	Kan /isa(his) Kanishee/kan isii(hers)
plural	keenya(ours)	Keessan/(yours)	Kan Isaani(theirs)

For example,

Re'een sun tiyya.-‘That goat is mine’.

Konkolaataan sun keessani –‘That car is yours’.

Hiriyyaan isii dhufee –‘Her friend came’.

Demonstrative pronouns are pronouns that are used to refer to a thing that was known previously or mentioned earlier. It can also be used to refer to the objects which are in the speaker's mind [21]. Both proximal and distal demonstrative pronouns exist in Afaan Oromo. Proximal pronouns have masculine and feminine whereas distal do not have. However, plural and singular demonstratives are not distinguished. Table 3.6 shows demonstrative pronouns in Afaan Oromo.

Table 3.6 Demonstrative Pronouns in Afaan Oromo

Proximal (this, these)	***Distal*** (that, those)
Kana/kuni(Masculine) Tana/tuni(feminine)	San /sun

In Afaan Oromo, relative pronouns are pronouns that are used to ask question for clarification [17]. Some of these are: **kam**,**eenyu**,**eenyuut** and **eenyuuf .**Consider the following sentences:

Kitaaba kam bittee? This is to mean ‘which book did you buy?’

Minjaallii kuni kan eenyuutii? Which means ‘whose table is this?’

Reflexive pronouns are pronouns that indicate the person who realizes the action of the verb is the same person who receives the action. Most of the time, reflexive pronouns are used when subject and object of the sentence is the same or when we want to emphasize the subject. Some of these pronouns in Afaan Oromo are depicted in table 3.7

Table 3.7 Afaan Oromo Reflexive Pronouns

	1st person	***2nd person***	***3rd person***
singular	Mataakiyya/ofii kiyya (myself)	Mataa keetii/ofii keetii (yourself)	Mataa isaati/ofii isaatii (himself) Mataa isiitii/ofii isheetii (herself)
plural	Mataa keenya/offi keenya(our self)	Mataa keessanii/ofii keessani(you yourself)	Mataa isaanitii/ofii isanniti(themselves)

Consider the following examples:

Ani *mataa kiyyaa* arge. This means 'I myself saw it'.

Jaallenneen ofi isiiti uffataa miicitii. This means 'Jalanne herself washed the cloth'.

Pronouns that are used to indicate that two individuals carry out some action and receive the same consequence is known as reciprocal pronoun. Afaan Oromo has '***wal***' which is used as a reciprocal pronoun. For example, in the sentence: **Isaan *wal* gargaaruu** which means 'they help each other'.

3.4.3 Verb ('Xumura')

In Afaan Oromo verbs are words that are used to indicate some action or event occurrence within time boundaries [17]. It can be transitive, intransitive, modals and auxiliary verbs. Transitive verbs are those verbs which transfer message to complement or object whereas, intransitive verbs do not transfer message to complement and hence, do not have complement or object. The following examples illustrate this fact.

Caalaan qotiyoo *bite*. This means 'chala bought an ox'. Since the action of buying was transferred to object **qotiyyo** 'ox', ***bite*** is transitive verb. However in the sentence:

Durreettiin kaleessa *deemte*. This means ‘Dureti went yesterday’, the word ***deemte*** ‘went’ doesn’t take complement or object and hence is intransitive verb.

Tenses like past, present or future also exist in Afaan Oromo verbs [18]. Some examples of Afaan Oromo verbs are: **deeme** ‘(he)went’, **barressaa jira** ‘is/am writing’,**jaalatuufi** ‘going to love’ etc.

Consider the following examples:

Ibsan mana barumsaa deeme. This means ‘Ibsa went to school’.

Tolaan xalayya barreessa jira. This is to mean ‘Tola is writing a letter’.

Hawwin boru nideemtii. This means ‘Hawi will go tomorrow’.

Auxiliary verbs are verbs that help the main verb of the sentence. Some of Afaan Oromo auxiliary verbs are:**dha,ture,jira, ta’e** and **qabda** .The following examples illustrate this.

Caaltuun barattuu *dha*.This means ‘Chaltu is a student’.

Murteysaan barsiisaa ta’e. This means ‘Murtyesa became a teacher’.

Ati barachuu barachu qabda. This means ‘you have to learn’.

In the above sentences all bolded words;dha,ta’e,qabda are auxiliary verbs.

3.4.4 Adjective (‘Ibsa Maqaa’)

Terms or words that clarify nouns are known as adjectives. Some of Afaan Oromo adjectives are shown below in table 3.8

Table 3.8 Afaan Oromo Adjectives

Descriptive	***Possessive***	***interrogative***	***Quantitative***	***Numbers and rank***
guddaa Dheeraa diimaa	Keenna Kee	Maalii Akkamii Kam	Hedduu mara	Lama Sadii Tokkofaa Afraffaa

Dhaabaan gola isaa keessa tokkoffa dha. This means ‘Dhaaba is first from his classes.’

Waaqoo Guutu nama *dheeraa* dha. This means ‘Waqo Gutu is a tall man’. Both ***tokkoffa*** and ***dheeraa*** are adjectives in the above sentences. These words describe the nouns **Dhaabaa** and **nama** respectively.

As it can be seen in the above sentences, adjectives follow the nouns they describe in Afaan Oromo .

3.4.5 Adverbs (‘Dabala/ibsa Xumura’)

Adverbs are any words that explain or modify verbs [17]. These can be adverbs of time, place, manner, frequency etc. Adverbs precede verbs they modify in Afaan Oromo. Some examples of adverbs illustrated in table 3.9

Table 3.9 Afaan Oromo Adverbs

Manner adverbs	***frequency adverbs***	***place adverbs***	***time adverbs***
dafe Malaan tasa	si’a lama Yeroo hunda	achi Gara mana	Guyyaa Har’a Boru

Consider the following examples:

Caalan dafe dhufe. This means ‘chala came quickly’. **Dafe** ‘quickly’ is an adverb.

Jaalannen boru deemti. This means ‘Jalane will go tomorrow’. **Boru** ‘tomorrow’ is an adverb**.**

Kananiisaan yeroo hunda ni mo’ata. This means ‘kananisaan wins every time. **Yeroo hunda** ‘every time’ is an adverb.

3.4.6 Preposition (‘Dur Duubee’)

Words that can have full meaning only in combination with some other words like noun, adjective or verbs are termed as preposition/postpositions. They do not take any affix and belong to closed part of speech [14]. Some prepositions in Afaan Oromo are: **akka** ‘as’,**eegasii** ‘since’,**hamma** ‘until’,**gara** ‘to’,**gadi** ‘below’,**irra** /**gubbaa** ‘on’ etc. Prepositions or postpositions may precede or follow the category to which they add syntactic meaning. Let us see the following sentences:

Murteysaan kitaaba minjaala *gubbaa* kayye. This is to mean ‘Murteysaa put the book on the table’. ‘**Gubbaa**’ in this sentence is postposition. As it can be seen it followed the category to which it adds meaning that is **minjaala** ‘table’. But in the sentence:

Murteysaan gara manaa kaate. This is to mean ‘Murteysa ran to home’. The word **gara** ‘to’ is a preposition and in this case it precedes the category to which it adds meaning.

3.4.7 Conjunction (‘Walqabsistotaa’)

A word that can be used to join or connect two phrases, clauses and sentences is known as a conjunction. Conjunctions can be divided into coordinating and subordinating conjunctions. Coordinating conjunctions are used to connect two independent clauses. Mostly these conjunctions are used when the speaker needs to lay emphasis on the two sentences equally. Some of these conjunctions in Afaan Oromo include: **garuu** ‘but’, **moo** ‘or’, **kanaafuu** ‘therefore’, **haata’u malee** ‘however/so’, **tu’ullee** ‘even though’ etc. Consider the following example:

Jaalanneen kaleessa mana barumsarra hafte *garuu* hojii mana hojjatee jirti. This means ‘Jalanne was absent yesterday from the school but she has done home work’. ‘**Garuu**’ in this sentence is coordinating conjunction. It is used to join the two sentences **Jaalanneen kaleessa mana brumsa hafte ‘**Jalanne was absent from the school yesterday**’** and **Jalannen hojii mana hojjattee ‘**Jalanne has done her home work’ which are independent.

Subordinating conjunctions are those conjunctions that are used to join main clause with subordinate clause. A subordinating conjunction is always followed by a clause [17]. Afaan

Oromo subordinating conjunctions include **yoo** ‘if’, **akka waan** ‘as if’, **wayta/yeenna** ‘when’, **hamma** ‘until’, **erga** ‘after’, **dursa** ‘before’ etc. The following example illustrates the above case:

Wayta innii dhufuu ani barressa jira. ‘Wayta’ in this sentence is used as subordinating conjunction. It joins one subordinating clause that is **inni dhufu ‘**when he was coming’ and **ani barreessa jira**, ‘I was writing’.

3.4.8 Numerals (‘Lakkofsa’)

Numerals include words that refer to number or quantity of something [21]. It can be cardinals such as **sadii** (three), **afur**(four) or it can be ordinals like **sada*ffaa*** (third) **afra*ffaa***(fourth). As discussed in[21], numerals in a sentence follow the category they describe their quantity or amount. Ordinals in Afaan Oromo are formed by adding suffix ***-ffaa*** to the cardinal numerals. Consider the following sentences:

1. **Ibsaan konkolata *lama* qaba**. This is to mean ‘Ibsa has two cars’. The word ***lama*** is cardinal numeral in the sentence. It follows the word it describes its quantity that is **konkolataa** ‘car’.
2. **Jalanneen golaa isiitii tokkoffaa baatee**. This is to mean ‘Jalane stood first from her classes’. In this case, the word **tokko*ffaa*** is ordinal which is to mean ‘first’. It is formed from cardinal **tokkoo** ‘one’ by adding affixes ***–ffaa***.

3.4.9 Introjections

These are words that are used to express special situations or events in the language. Emotions, pleasure, sorrow or suddenly happening situations have their own word expressions in many languages. These words are called introjections. Afaan Oromo’s introjections include **ishoo**(for happiness), **wayyoo**(for sadness), **ah**(for silent new events or situations happened).

3.5 AFAAN OROMO TAGSET

Detailed description of word categorization in Afaan Oromo was given in the previous section. In this section, actual tagset which will be used in this thesis is determined and discussed.

Since part of the corpus prepared for this study was adopted from [21] corpus, tagset selected for the study is also based on the work of [21]. Around 18 tagsets were identified in this work. 17 of the tag sets are adopted from [21] and one additional tag, Negative Indicator (NG) is introduced in this research. Detail description of these tagsets is given below.

3.5.1 Noun Tags

As explained in 3.3.1 Afaan Oromo nouns indicate numbers (singular and plural) and genders (masculine and feminine). Moreover, a noun can be grouped as proper or common noun. In this thesis work regardless of this detail, all nouns are assigned NN tag.

Initials in names like **Obbo** 'Mr.', **Adde** 'Mrs.', **Durbee** 'Miss.'etc and directions like **baha** 'east', **dhiha** 'west', **kabaa** 'north', **kibba** 'south' are also assigned NN tag.

Nouns joined with post positions and conjunctions are given another sub tag. Nouns that are joined to their postpositions are given NP tag whereas nouns that are joined with their conjunctions are assigned NC tag. Consider the following example to illustrate these assignments:

Hawwiin fardaan**/NP gaba deemte**.Which means 'Hawwi went to the market by horse'. **Farda** is a noun joined with postposition **'–an'** and hence given tag NP.

Jaalanneefii/NC ibsaan kalessa dhufan. Which means 'Jalanne and Ibasa came yesterday'. **Jaalannee** is noun joined with conjunction '**-fi**' and it is assigned tag NC.

Sorrettinis/NC dhufte. This means 'Sorretti also came'. In this sentence **Sorretti** is noun joined with conjunction '**-nis**'. Therefore, it is assigned tag NC.

3.5.2 Pronoun Tags

Regardless of gender and number, all types of pronouns: personal, possessive, demonstrative, reflexive and reciprocal pronouns are assigned tag PP.

Pronouns that are joined to their post positions are assigned tag PS. Those pronouns joined with conjunctions are assigned tag of PC. The following example illustrates this fact:

Isaafii/PC isiin waliin deeman. Which mean ‘He and she went together’. In this sentence, pronoun **Isaa** ‘he’ is joined with conjunction ‘-fi’ and assigned tag PC.

Fedhii isiitiin/PS hafte. Which means ‘she was absent by her interest. **Isii** is pronoun joined with postpositons **‘-tiin’** hence assigned tag PS.

3.5.3 Verb Tags

Both transitive and intransitive verbs are assigned a tag VV. Auxiliary verbs of the language are given tag AX. The following sentences illustrate this.

Hundeen doctora ta’e/AX.Which means ‘Hundee became a doctor. **Ta’e** in this sentence has equivalent meaning of ‘became’ which is auxiliary and assigned tag of AX.

Hawiin xalyaa barresite/VV. This means ‘Hawi wrote a letter’. **Barreesite** has equivalent meaning of ‘(she) wrote’ and hence assigned tag VV.

3.5.4 Adjective Tags

All groups of adjectives regardless of their specific features are assigned tag JJ.

Adjectives joined with their conjunction are assigned tag JC.

Jaalannee qal’oofii/JC dheertuudha. This means ‘Jalanne is thin and tall’. Adjective **qal’oo** ‘thin’ is joined with conjunction ‘-fi’ hence assigned tag JC.

3.5.5 Adverb Tags

All categories of adverbs in Afaan Oromo are assigned tag AD. Days of the week that are used as adverb in a sentence are also assigned tag AD. For example:

Boontuun gaafa Dilbataa/AD dhufti. This means ‘Boontuu will come on Sunday’. The word **Dilbataa** ‘on Sunday’in this sentence acts as adverb and therefore assigned tag f AD.

Jimaata/AD Jimaata/AD ani sagada deema. This means ‘I go worshipping every Friday.’ **Jimaata jimaata** equivalent to ‘every Friday’ is used as adverb and assigned tag AD.

3.5.6 Preposition Tags

All prepositions that are not joined with other words are assigned tag PR. Consider the following example:

Hurrisaan gara/PR mana barumsaa deeme. This means ‘Hurisa went to school’. The word **gara** in this sentence has equivalent meaning of ‘to’ and acts as a preposition. Hence it is assigned tag PR.

Qoteebulaan wa’ee/PR misoomaa irratti hirmaachu qaba. This means ‘Farmers should participate in development.’ **Waa’e** ‘in’ is used as a preposition in this sentence and assigned tag PR.

3.5.7 Numerals Tags

Ordinal numerals in Afaan Oromo are given tag ON. Since most of the time they are used as adjectives, cardinal numerals are treated as adjective tag sets and given tag JN.

3.5.8 Conjunctions Tags

Conjunctions that are not joined with other words are assigned tag CC. Those that are joined with other words are treated as described in the previous sections. The following examples illustrate tag of unjoined conjunctions:

Jalanneen yookin/CC Ibsaan ni dhufa. This means ‘Jalanne or Ibsa will come’. **Yookin** is equivalent of ‘or’ is a conjunction.

Sorrettin nirafti Caaltuun ammo/CC ni qu’attii. This means ‘Sorretti sleeps whereas Chaltu studies.The word **ammo** which is equivalent of ‘whereas’ is a conjunction.

Akkuma/CC ani dhufeen ini deeme. This means ‘As soon as I came he went out.’ The word **akkuma** ‘as soon as’ is also a conjunction.

3.5.9 Introjections Tags

Afaan Oromo introjections are assigned tag II.

Ishoo/II! Baga dhuftee.This means 'You are welcome!' The word **ishoo** is used to express pleasure of something and is used as introjection word. **Wayyoo/ II! nan irranfadhe**. This means 'Sorry I forgot!' **Wayoo** is introjection and assigned tag of II.

3.5.10 Punctuations Tags

All punctuation words and other unique symbols are assigned tag PN in this thesis work. Consider the following sentences:

Mul'ataan dhufe./PN. This means 'Mul'ata came.' The punctuation word (.) is assigned tag PN.

Ishoo! /PN baga dhufete. The punctuation symbol (!) is also assigned tag PN.

3.5.11 Negation tags

The word **hin** in Afaan Oromo is used as negation of something and assigned tag NG. For example:

Dhufe meaning '(he) came.' Whereas **hindhufne** meaning '(he) didn't come.'

Nyaate meaning '(he) ate' whereas **hinnyaane** is to mean 'he didn't eat'.

NG tag is assigned for the word hin in this thesis work. Consider the following sentence:

Tolasaan mana barumsaa hin/NG deemne. This means 'Tolasa did not go to school'. The word **hin** indicates negation and is assigned tag NG. Summary of all the above tag is shown in table 3.10 below.

Table 3.5.1 Afaan Oromo Tags set Summary

No	***Tag***	***Assigned to***

1	**NN**	Nouns not joined with other.
2	**NP**	Nouns joined to post-positions.
3	**NC**	Nouns joined to conjunctions.
4	**PP**	Pronouns not joined with other.
5	**PS**	Pronouns that coexist with postpositions.
6	**PC**	Pronouns that exist with conjunctions.
7	**VV**	Main verbs.
8	**Ax**	Auxiliary verbs.
9	**JJ**	Adjectives not joined with other.
10	**JC**	Adjective joined with other.
11	**JN**	All numerals
12	**AD**	All adverbs
13	**PR**	Prepositions have.
14	**ON**	Ordinal numeral.
15	**CC**	Conjunction not joined with other.
16	**II**	All introjections.
17	**PN**	Punctuation marks.
18	**NG**	Negative indicator (hin).

CHAPTER FOUR: DESIGN OF AFAAN OROMO POS TAGGER

4.1 INTRODUCTION

Afaan Oromo POS tagger is a program that assigns part of speech to Afaan Oromo words according to the context of that word in a sentence. As it is illustrated in chapter 2, Brill transformational rule based POS tagger which is used in this work, uses lexical and contextual rule to assign part of speech to a given word. This POS tagger first uses statistical techniques to extract information from the training corpus and then uses a program to automatically learn rules which reduce the faults that would be introduced by statistical mistakes. Hence, Brill rule based tagger can be considered as a hybrid approach [23].

Once the tagger is trained, untagged corpus can be tagged by first assigning the tag based on the lexical rule that was learned during training and then the incorrect outputs which are obtained through lexical rule application is corrected by applying the contextual rule that was also learned during the training phase.

In section 4.2, general framework of Brill transformation error driven learning (TEL) and how each components work is illustrated. Next, the generalized learning algorithm for Afaan Oromo POS tagger is explored. In the last section of the Chapter the architecture of the tagger is presented.

4.2 TRANSFORMATION-BASED ERROR-DRIVEN LEARNING

The general framework of Brill's rule based learning is known as transformation based error driven learning (TEL). It is to imply that the tagger is based on transformations or rules and learns by detecting errors occurred in the previous steps. In this section overall description of how TEL works is presented. Different components of TEL are also briefly discussed.

Transformational based error driven learning takes unannotated corpus as input which goes through the initial state tagger. This initial state tagger assigns a tag that is most likely. This gives a temporary corpus as an output.

The temporary output corpus produced by the initial state tagger is then compared with the target corpus which was manually tagged and assumed to be correct. The corpus passes through the learner iteratively to derive rule for transformations. Each of these derived rules is examined by applying it to the temporary corpus and comparing the result with the target corpus. Based on this comparison a set of transformation rules (which are thought to be best) is produced as ordered list of rules. The process continues until performance is improved.

This transformation error driven learning (TEL) is used twice in Brill's tagger. First for lexical rules derivation which are later used to tag unknown words and second for derivations of contextual rules that are used to improve the accuracy of the tagger by considering the environment of the word. Each of these rules (lexical and contextual) use two corpora, the target corpus, and a temporary corpus whose tags are changed step by step to make it similar with the target corpus as much as possible. Figure 4.1 shows the frame work for TEL adopted from the original Brill rule based tagger.

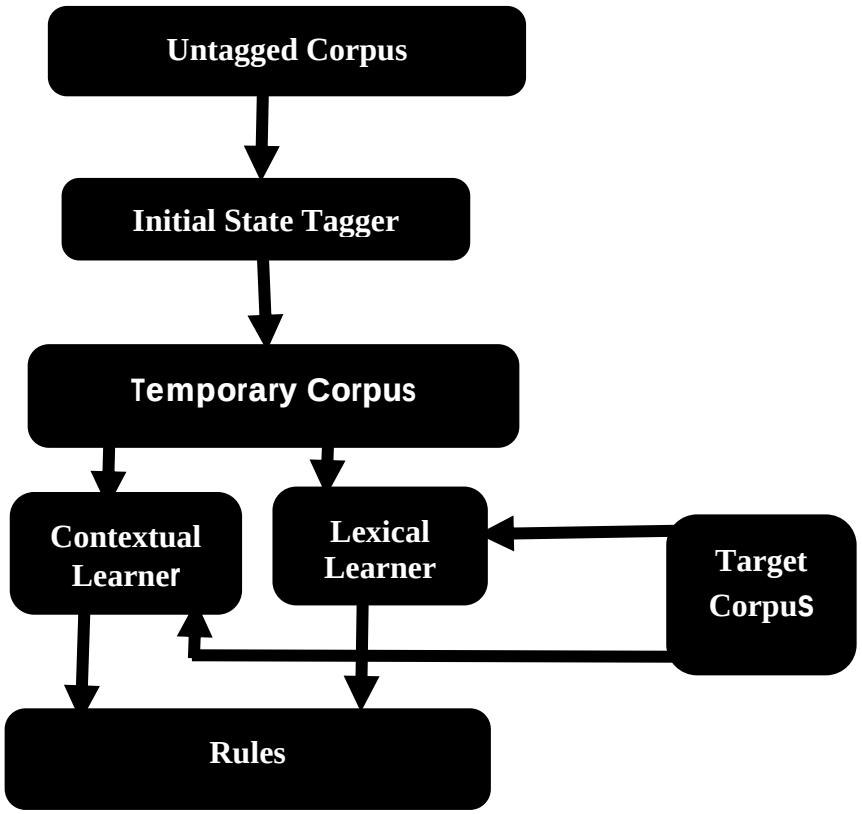

Figure 4.1 Transformation Based Error-driven Learning of Brill's Tagger [23].

As it can be seen from the figure, there are two learner modules, lexical learner and contextual learner. The lexical learner module uses a list of words containing information about the frequencies of tags, which is derived from the target corpus whereas the temporary corpus for the lexical learner is a list of words which is similar to the target corpus but tagged in some way [initially all are tagged as noun]. On the other hand, the target corpus for the contextual learner is the manually tagged corpus which is already running and the temporary corpus is same corpus which was tagged in different way than the target corpus.

The rule part consists of a condition (the trigger and possibly a current tag) and a resulting tag. These rules are initialized from already defined transformation templates. Originally, these templates are defined as uninitialized variables which can later be instantiated to some rule during the training. Below is the form of these template rules:

1. If Trigger, then change the tag X to the tag Y
2. If Trigger, then change the tag to the tag Y where X and Y are variables.

In template (1) the interpretation is change the current tag X of a word to the resulting tag Y if the rule triggers the word whereas in template (2) it is to mean that change the word tag regardless of its current tag to the resulting tag Y when the rule triggers the word. Based on this, a set of all rules (SR) is generated from the existing predefined templates.

4.2.1 Learning Process

As it was stated above, during the training of Brill's tagger, both lexical and contextual rules are generated as output of the training process. First, annotated temporary corpus (ATC_0) was given as input to the learner. The learner module finds a set of rules (SR) that can be applied to ATC_0 to get better temporary corpus ATC_1 (temporary corpus is better if it is closer to the target corpus than the previous temporary corpus). Now if we assume the above rule to be R1, then by applying R1 to ATC_0 it gives us ATC_1 (annotated temporary corpus).This is repeated again with ATC_1 and it extracts rule R2 that gives better annotated temporary corpus ATC_2 when applied to ATC_1. The successive improvement of temporary corpus to make it closer to the target corpus continues iteratively to get R3, R4, R5 …etc and the corresponding annotated

temporary corpus ATC_3, ATC_4, ATC_5…are generated until some threshold value for the number of iterations is reached (the threshold value for iteration can be varied). Finally, all best rules R1, R2, R3…, obtained are given to the tagger as ordered list of output.

This learning component has two sub-components: lexical rule learner and contextual rule learner. Each of these sub-components uses TEL for learning the rule. In the following consecutive sub-sections these two sub components of the learner are discussed in detail.

4.2.1.1 Lexical Rule Learner

The main aim of lexical rule learner is to get a rule that would assign most likely tag to the word in a language. Ideally, this requires consideration of all the texts in that language. This is difficult in real implementation as it is to mean determining most likely tags for unknown words of the language provided that tags for relatively small set of words are known .

To solve this difficulty, the lexical rule learner of Brill tagger uses a statistical method. Half of the total training corpus (with some additional annotated if possible) is used by this sub component of the learner. Besides this, the learner uses three different lexicons (smallwordtaglist, bigbigramlist and bigwordlist) which are constructed from already annotated corpus. Fig 4.2 shows inputs and output of the lexical learner components.

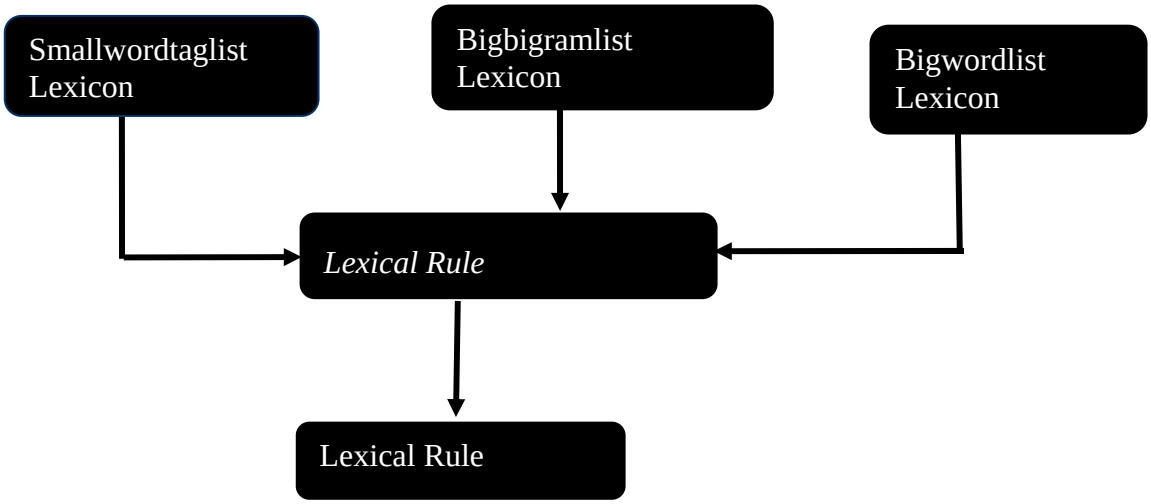

Figure 4.2 Lexical Rule Learner Components

The smallwordtaglist lexicon contains list of word and tags frequency. It shows the number of occurrence (frequency) of a word and its tag in the tagged corpus used for training.

Based on the above concept, Freq (W|T) which refers to frequency of a word with specific tag T and Freq (W) which refers to the frequency of the word W in that corpus can be used to estimate the probability that the word W is tagged with tag T as:

P (T|W) = Freq (W|T) / Freq (W) where P (T|W) is the estimated probability.

On the other hand, Freq (W|T) can be used to directly compute the most likely tag for the word W. This can be done by taking the tag with the highest frequency.

The bigwordlist lexicon contains list of all words found in the unannotated corpus arranged by decreasing frequency of word tags.

The bigbigramlist is a list consisting of all word pairs occurring in the unannotated corpus. This list of bigbigramlist doesn't contain frequency of the words rather it contains information that tells us if a given word pair occurs in the unannotated corpus or not. Both the bigwordlist and bigbigramlist are used to determine the trigger condition or to make it sure.

After these lexicons are provided, the learner creates a word list from the smallwordtaglist which contains list of words that their tagging information has been removed. The initial state tagger allocates every word the most likely tag which is NN in the case of Afaan Oromo tagger. At this stage, wordlist that contains temporary corpus WTL_0 is created. The generation of WTL_i is similar to $ATC_{i\ ,i=1,2,3...}$.Then all permissible rules SR are created from all instantiations of the predefined lexical templates. Moreover, the score for every R in SR is computed as follows:

1. For the rule template: if Trigger then change tag X to tag Y

For W a word in WTL_i with current tag X satisfying the trigger condition the score of R is given as:

Score(R) =P (Y|W) - P (X|W) where P (Y|W) is the probability of word W to have new tag Y and P (X|W) is the probability of word W to have old tag X. The total score of rule R can be obtained by computing all the scores for word W through the WTL_i that satisfies the trigger condition as:

$$\text{TotalScore}(R) = \sum_{w} P(Y|W) - P(X|W)$$

2. For the rule template: if Trigger then changes current tag to tag Y

For W a word in WTLi satisfying the trigger condition the score of R is given by:

Score(R) = P (Y|W) - P (Currenttag of W |W) and the total score for R is then calculated by adding all the word scores. i.e

$\text{TotalScore}(R) = \sum_{w} P(Y|W) - P(\text{Current tag of } W\ |W)$,where the sum runs over all W in WTL_i that satisfy the trigger condition.

In general, the score for rule R can be computed by subtracting probability of new tag of word W from the probability of old tag of word W as:

P (Newtag|W)-P (oldtag|W)

If the result of this calculation is positive the new tag is the most likely tag than the old tag, otherwise the old tag is the most likely tag. Note that the trigger condition is tested using bigwordlist and bigbigramlist, and the estimated probabilities are computed from the frequencies in smallwordtaglist.

Some of the templates used in lexical rule learner are given below.

1. Change the most likely tag to Y if the current word has suffix/prefix x.
2. Change the most likely tag to Y if ***deleting /adding*** the suffix x, |x|=< 4 results in word.
3. Change the most likely tag from X to Y if ***deleting/adding*** the prefix x, |x|=<4 results in word,|x| is length of x.
4. Change the most likely tag from X to Y if word W ever appears immediately to the left/right of the word.

For the original Brill tagger, template number 2 and 3 adding /deleting of suffix/prefix considered only up to 4 characters. However considering the nature of Afaan Oromo, the templates were modified to extend deletion/addition of the suffix/prefix up to six characters as:

- Change the most likely tag to Y if ***deleting /adding*** the suffix x, |x|=<6 results in word.

- Change the most likely tag from X to Y if ***deleting/adding*** the prefix x, |x|=<6 results in a word,|x| is length of x. This is because some words in Afaan Oromo can have more than 4 suffixes. Consider the following example:
    - **bektan** which means ‘you knew it’ whereas bektan***iirtu*** which means ‘you have known it’ has five suffix.
    - **nama** which means ‘man’ whereas **nam*eewwan*** has six suffix***(eewwan).***

After the lexical rules are learned, the training corpus for the contextual training is initially tagged using these rules. In the next section contextual rule learner component is discussed in detail.

4.2.1.2 Contextual Rule Learner

The contextual rule learner component is used to tag unknown word in the corpus based on the context (environment). After lexical rule learner component has learned method for predicting the tag of unknown word based on some rules the most likely tags for words in the annotated corpus the contextual rule is used for further disambiguation and better accuracy.

The contextual rule learner module takes initially tagged text. The initial state tagger takes untagged text (second half of the corpus where tag information was removed) and lexicons termed as training lexicon, bigbigramlist (same as used in the lexical learning module above) and the lexical rules obtained during lexical rule training process. Diagrammatically these inputs and outputs of the contextual rule learner are shown in figure 4.3 below.

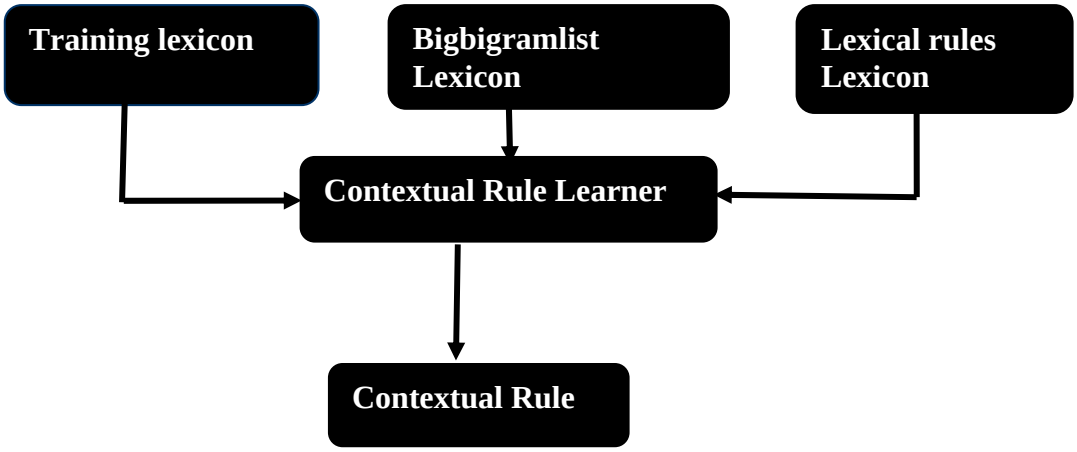

Figure 4.3 Contextual Rule Learner Components

Training lexicon consists of list of words with their possible tags attached to each word. With the help of these lexicons, the initial state tagger assigns most likely tags to the word in the untagged text. Known words (words found in the training lexicon) are assigned with the most frequent tag.

Unknown word tag is computed by using the lexical rule. First, all words are assigned most likely tag, i.e. NN and then lexical rules are applied in the order they appear in the lexicon. Some of the triggers of the lexical rule learner are listed in the bigbigramlist lexicon. Therefore, bigbigramlist is also referred this time to find the triggers and apply the rule based on that trigger.

The output (temporarily tagged corpus say CT_0) of initial state tagger which is a tagged text is given as input to the contextual rule learner. As the target corpus, the second half which is manually tagged corpus is given to the learner. In a similar way to that of lexical rule leaner the contextual rule learner also generates a set of all permissible SR rules. These sets of rules in the contextual rule learner are totally different from the lexical rule learner for it uses different transformation templates. The trigger in this case depends on the context (environment) of that word but not on the morphology of the word. Some of these templates are listed below:

1. The preceding/following word is tagged with Z.

2. One of the two preceding/following words is tagged with Z.
3. One of the three preceding/following words is tagged with Z.
4. The preceding word is tagged with Z and the following word is tagged with V.
5. The preceding/following two words are tagged with Z and V.
6. The word that is two words before/after is tagged with Z.
7. The current word is tagged with Z

Scores on the temporary corpus CT_0 are computed for all rules in PR if the trigger condition is satisfied for the rules. Score of rule R can be computed simply by comparing the word W tag in CT_1 when rule R_1 is applied with the correct tag of same word in the target corpus. If the applied rule R1 corrects the tag of the word, the score of R1 is +1. However, if the applied rule R1 introduces error, the score for R1 is -1. In all other cases, the score for the rule is 0. Hence, the score for rule R is generalized as:

Score(R) = number of errors corrected - number of errors introduced

The rule with the highest score is selected and put on output list. After this, the learner applies R1 to CT_0 and produces CRT_1, on which the learning continues. The process repeatedly continues (putting one rule on the output list in each step based on their score) until the applied rule shows no improvement on the state of the corpus or some threshold value of error is achieved. The generalized learning algorithm is given in the next sub-section.

4.2.2 The Learning Algorithm

Transformational rule based learning algorithm for both of the contextual and lexical rule can be generalized as follow:

1. Initialization: This step initializes all the words of the corpus to some tag.
2. Learning

a) Calculate repeatedly the error score of each candidate rule (difference between the number of errors before and after applying the rule).

b) Pick the rule with higher score.

c) Add it the rule set SR

d) Use these rules to tag the initialized text.

e) Repeat step d above until no rule has a score above a given threshold, or until applying new rules make no change on the state of the tag.

f) End with lexical and contextual rule output.

In the next section the general architecture of Brill tagger, which is trained and ready to tag, is presented.

4.3 BRILL TAGGER ARCHITECTURE

Trained Brill tagger takes untagged text and rules as inputs. These rules which were already learned during the training are applied on raw text that is to be tagged. Tagged texts based on the learned rules are produced as outputs. The adapted architecture from the original Brill tagger with a bit modification for Afaan Oromo is shown in figure 4.4.

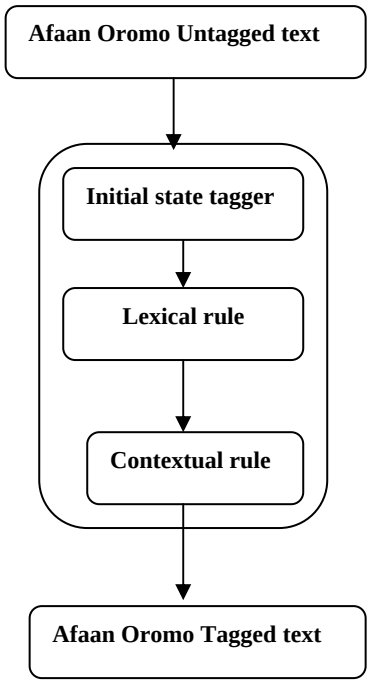

Figure 4.4 Adapted Brill Tagger for Afaan Oromo

Afaan Oromo untagged corpus is passed through initial state tagger. The initial state tagger tags all the words of the corpus with their most likely tag which is noun in this case. The lexical rules are applied on the initially annotated corpus to tag the words correctly. Then the contextual rules are applied to improve the correctness of the tag more. Finally, tagged corpus is produced as out put of the tagger.

CHAPTER FIVE: IMPLEMENTATION AND PERFORMANCE ANALYSIS

5.1 INTRODUCTION

In the previous Chapter, the design of Afaan Oromo part of speech tagging and algorithms were discussed. Data preparation and the results of the experiment conducted are discussed in this Chapter.

5.2 CORPUS PREPARATION

Corpus is a Latin word for 'body'. It is to mean any body of text. However, the term corpus has a particular meaning or connotation when used in perspective of modern linguistics. According to [13] corpus is expected to have the following features:

- Sampling and representativeness - many natural languages have large number of words and it is difficult to prepare the corpus that constitutes all the words in the language. Sample words are taken and used which can be representative of the other words. The sample has to represent variety of the words and their morphological and structural variation.
- Machine readable - nowadays many corpora are also expected to be machine readable even though it is not always true.

Corpus may exist in two different forms: unannotated and annotated corpus. Annotated corpus is a collection of texts that contains grammatical or linguistic information. Whereas unannotated corpus is a collection of text without linguistic information. These are texts that exist as normal, untagged words.

Annotated corpus can be used for various purposes. In linguistics, properly annotated (tagged) corpus can be used to study linguistic features such as morphology and phonology of a language. It can be used as part of speech taggers and parsers training in computational linguistics. Thus, annotated corpus plays a great role in automatic part of speech tagging and parsing. The corpus will be provided to the system as training data so that the system can learn

/adapt some pattern from the pattern in the corpus for each word or sentence. The size of the corpus affects the learning tendency of the system. Larger size of corpus provides greater learning tendency for the system. As a result, accuracy of the system will be better to automatically assign part of tagging.

There is no such large size corpus which is already prepared for Afaan Oromo language. Preparation of this large size corpus is expensive and time consuming task. As a result of this, a corpus of 233 sentences (1708 words) is prepared for this experiment. Half of the corpus was taken from the work of [21].

The corpus was collected from different magazines, bulletins and Oromo news papers. As it is discussed in [21] bulletins, magazines and newspapers contain many social, economical, technological and political affairs of a certain society. Hence, they are good source for collected representative corpus for natural language processing.

The collected corpus was manually tagged with the help of experts in the field. During tagging, the number of tag sets discussed in Chapter Three and the contextual position of the word was considered. This manually tagged corpus was used to train the tagger and evaluate its performance.

After tagging it manually, the entire corpora were divided into two sub corpus as 10% for testing and 90% for training. An experiment was conducted to determine the best percentage of training data for lexical and contextual rule learner which gives the tagger good performance. The result of the experiment shows that the best accuracy is achieved when the percentage of lexical rule learner data is 65% and contextual rule learner is 35% of the entire training data. This means that the above 90% training corpus is divided into two as 65% for lexical rule learner and 35% for contextual learner training. Table 5.1 shows the result of the experiment.

Table 5.1 Percentage of Corpus for Lexical Rule Learner and Accuracy of Tagger

Corpus percentage for lexical rule learner (%)	***Corpus percentage for contextual rule learner (%)***	***Accuracy***
55	45	64.7
60	40	65.6
65	35	67.4
70	30	63.8
75	25	66.5
80	20	66.1

5.3. LEXICON PREPARATION

Several lexicons were prepared from the manually tagged sample corpus. One of these lexicons is a lexicon that contains list of word and tag frequency. Such a lexicon was prepared as a text file and is termed as smallwordlist. Table 5.2 is taken from the smallwordlist that was prepared for the training of Afaan Oromo tagger.

Table 5.2 Smallwordlist Lexicon of Afaan Oromo Tagger

Word	***Tag***	***Frequency***
.	PN	93
hin	NG	11
waan	PR	9
wal	PP	7
akka	PR	6
isaa	PP	5
keessatti	PR	4

Another lexicon that was prepared for this study is a lexicon that contains list of all words found in the unannotated corpus, arranged by decreasing frequency of word tag known as bigwordlist. Table 5.3 shows sample bigwordlist lexicon prepared.

Table 5.3 Bigwordlist Lexicon of Afaan Oromo Tagger

Order	***Word list***
1	.
2	hin
3	kan
4	waan
5	akka
6	jira
7	qaba
8	yeroo
9	wal
10	isaa

Bigbigramlist lexicon is also prepared from the sample corpus. This is a lexicon that consists of information about word pair that occurs in the untagged corpus. Table 5.4 shows sample Bigbigramlist lexicon taken from the prepared corpus.

Table 5.4 Bigbigramlist lexicon of Afaan Oromo tagger

Order	***Pair of words in the training corpus***
1	mataan isaatuu
2	demokiraasii faayidaa
3	fakkaatu qaba
4	waraanuun safuu
5	Qaroominni kunuunsi
6	darbuu kanaan
7	hin adeemu
8	jiruu ilaala
9	abba fayyisaa
10	Akka itti

A lexicon called training lexicon that consists of list of word with their possible tags attached in their order was also prepared. This lexicon is used during the training phase by the contextual learner component. Table 5.5 shows sample training lexicon taken from the prepared corpus.

Table 5.5 Training Lexicon of Afaan Oromo Tagger

No	***Word***	***Possible tags in order***
1	qorannoo	NN JJ
2	Abbaan	JJ NN
3	jireenyaafi	NC JC
4	yaadati	VV JJ

5	jira	AX VV
6	guyyaa	NN AD JJ
7	isa	PP
8	jiru	AX
9	qaroominni	NN
10	kabaju	VV

5.4 EXPERIMENTS AND RESULTS

5.4.1 Learning curve and its analysis

In order to see the convergence of the training process, we have conducted learning curve analysis.

The entire training data set was divide into ten equal sizes (each size is 10% of the total training set).The accuracy of the tagger was tested starting by the first 20% of the data and repeating the process by adding 10% to the previous data until the entire training corpus(100%)is used. For every 10 % data added the accuracy variation is recorded. Table 5.6 shows the performance of the system for a given percentage of training data and Figure 5.1 shows the learning curve analysis.

Table 5.6 Accuracy variation with Training data percentage

Training data percentage	*Accuracy of tagger*
20	45
30	47.8
40	50.5
50	54.3
60	55
70	60.6
80	64.5
90	68.9
100	70.7

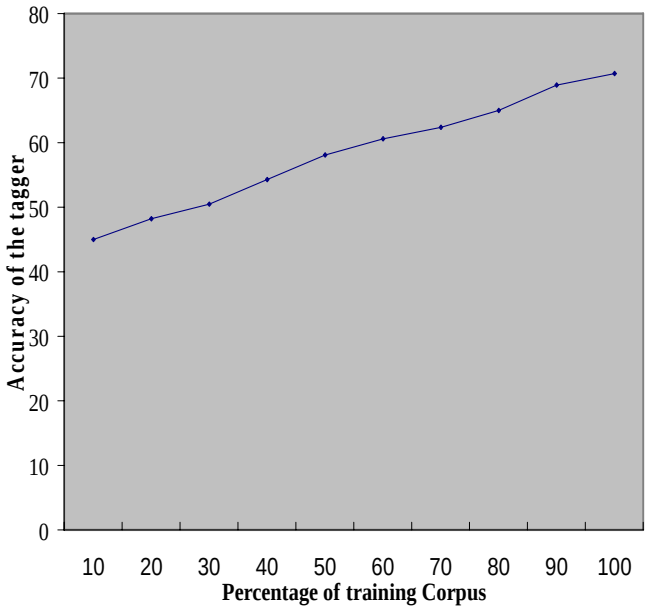

Figure 5.1 Learning Curve of the Tagger

As it can be seen, the learning curve shows that the training data is not sufficiently large enough. As a result we decided to utilize appropriate evaluation technique in our experimentation which is 10-fold validation. This is done by dividing randomly the entire testing corpus into ten test sets. The nine fold is used for training whereas the remaining tenth fold was used for testing the tagger that was trained on the previous nine folds. The process was repeated ten times by taking other nine as training and tenth one as testing corpus. As it is shown in Table 5.7, after each training phase, the tagger was tested average on 169 Afaan Oromo words. Each of the corresponding training set contains an average of 1708 words. Finally, the result obtained on test data set was compared with the corresponding test set which was manually tagged. Table 5.7 summarizes the results of this evaluation for the original Brill tagger as well as modified tagger for Afaan Oromo texts.

Table 5.7 Summary of the Result Obtained by the Original Brill Tagger

Test set	***No of words***	***Original Brill tagger Accuracy***	***Modified Brill tagger Accuracy***
1	144	87	93
2	198	66.4	71
3	165	76	81
4	177	73	79
5	185	73.4	77
6	151	86.5	91
7	114	89.4	93.2
8	178	74.1	77
9	226	76.4	80
10	191	74.2	78
Average Accuracy		**77.64**	**80.08**

As it can be seen in Table 5.7 the modified tagger improved the accuracy of the original Brill tagger by 2.44%.

5.5. PERFORMANCE ANALYSIS

In order to analyze the performance of the transformational error driven learning for Afaan Oromo POS, we have considered the case where we obtained the least performance in the 10-fold validation for the modified tagger and a confusion matrix is developed. As we have 18 tags and some of them are rarely occurring in the script, we have clustered the rarely occurring tags into one as others as shown in table 5.8.The tag for punctuation (PN) is not considered in this analysis for it has few lists which belong to this class and simply found in the lexicon. In other words, the tagger can simply refer the lexicon for PN category for there are only few symbols which belong to this class. Therefore, miss classifications of PN tag will not almost occur.

Table 5.8 Lists of Tags and Their Occurrence in the Training and Test Data

NO	***Tag***	***New Tag***	***F1***	***F2***	***F3***
1.	NN	NN	406	356	50 (12.32%)
2.	VV	VV	404	350	54 (13.37%)
3.	JJ	JJ	231	212	19 (8.23%)
4.	PR	PR	167	146	21 (12.57%)
5.	AX	AX	136	124	12 (8.82%)
6.	PP	PP	124	109	15 (12.10%)
7.	AD	AD	92	78	14 (15.22%)
8.	NG	Others	35	32	3 (8.33%)
9.	NP		28	25	3 (10.34%)
10.	CC		26	24	2 (7.69%)
11.	NC		24	22	2 (8.33%)
12.	JN		12	11	1 (8.33%)
13.	JC		8	7	1 (12.5%)
14.	PS		5	5	0 (0%)
15.	ON		4	4	0 (0%)
16.	II		4	4	0 (0%)
17.	PC		2	1	1 (50%)
Total			**1708**	**1510**	**198**

		Predicted Tag

Where F1, F2 and F3 are frequencies of tags in the entire corpus, training data set and test data set respectively.

Seven tags are found to be the most frequent tags of all 17 tags available. These seven tags are: **NN**(noun),**VV**(verb),**JJ**(adjective),**PR**(prepositions),**AX**(auxiliary verbs),**PP**(un joined pronouns),and **AD**(adverbs).The remaining tags are grouped together as others. The percentage of all the tags available in the test data is also computed to see the representative ness of the test data for the taglists. Except for few of the tags the frequency of the tags in the test data is good (percentage of entire test data is found to be **11.59).**

Table 5.5.2 shows the confusion matrix that shows the number of words tagged as X where it is actually needs to be tagged Y for any possible tags in the language.

Table 5.9 Confusion Matrix of the Tagger

Desired tags		NN	VV	JJ	PR	AX	PP	AD	Other	Total	Errors
	NN	**40**	7	3	0	0	0	0	0	50	10 (17%)
	VV	8	**45**	1	0	0	0	0	0	54	9 (15%)
	JJ	10	3	**6**	0	0	0	0	0	19	13 (22%)
	PR	5	1	0	**13**	0	1	1	0	21	8 (13%)
	AX	2	1	0	0	**9**	0	0	0	12	3 (5%)
	PP	0	1	0	1	0	**13**	0	0	15	2 (13%)
	AD	4	2	2	0	0	0	**6**	0	14	8 (13%)
	Other	0	5	0	0	0	0	0	**8**	13	5 (8%)
	Total	69	65	12	14	9	14	7	8	**198**	
	Error	**29**	**20**	**6**	**1**	**0**	**1**	**1**	**0**	**58**	
		50%	**34%**	**10%**	**2%**	**0%**	**2%**	**2%**	**0%**	**100%**	

As it can be seen in the above confusion matrix table, of the whole 198 tags about 140 tags are correctly tagged and 58 tags are tagged incorrectly. Most of the words (29 words which account about 58% of the total wrongly tagged words) are incorrectly assigned to NN tag. This error can be due to the fact that the initial tagger assigns all words NN tag and only if the matching rules are available in the rule list (lexical or contextual rule) that transformation to the correct tag may occur. In other words, if rules are not found in lexicon the word is tagged as NN by the initial state tagger. Therefore, those words that did not get transformation rule will remain as NN without change. The other reason for the NN tag to be the most wrongly assigned tag is the nature of those nouns that can be used as adjectives. As it is illustrated in the confusion matrix, 52% of JJ (adjective) also wrongly tagged to NN. This confusion is observed in nouns that are used as adjectives. The analysis of some of these words is indicated as below.

Word	*Desired (correct) tag in the sentence*
Bulchinsa	**Bulchiinsa/JJ**mootummaa/JJnaannoo/JJOromiyaatti/NN raawwiin/JJpaakeejii/JJ qulqullina/JJ barnootaa/NN haala/AD gaari/JJ irratti/PR akka/PR argamu/VV Biiroon/JJ Barnoota/JJ Oromiyaa/NN ibse/VV. /PN.

Output by the system in the sentence

Bulchiinsa/NN Mootummaa/NN Naannoo/NN Oromiyaatti/NN raawwiin/NN paakeeji/NN qulqullina/NN barnootaa/NN haala/JJ gaari/NN irratti/PR akka/PR argamu/VV Biiroon/NN Barnoota/NN Oromiyaa/NN ibse/VV. /PN

The word ***bulchinsa*** was assigned NN tag wrongly by the tagger where it should have been JJ.

Close observation of the above sentence shows that the tagger failed to transform NN to JJ for the word **Bulchinsa** which is used as adjectives (Even though it is noun originally). For such cases especially where a number of nouns used as adjectives appear in sequence the tagger has no transformations rule from NN(noun) to the JJ(adjectives).The followings are possibly correct transformation rules that are found in both lexical and contextual rule files of the tagger to change a word tag from NN to JJ .

1. baa hassuf 3 JJ (change the tag of the word to JJ if it has 3 character suffix ,ba).
2. la hassuf 2 JJ (change the tag of the word to JJ if it has 2 character suffix la).
3. daa hassuf 3 JJ (change the tag of the word to JJ if it has 3 character suffix ,daa)
4. maan hassuf 4 JJ(change the tag of the word to JJ if it has 4 character suffix ,maan).

5. JJ NN NEXTTAG JJ (change the NN tag to JJ of the word if it has next word tag JJ).

The above 1, 2, 3, 4 rules are lexical rules that are only based the suffix of a word. If we consider the word ***bulchinsa*** it has no suffix which corresponds to either of the above cases (baa,la ,daa or maan).Rule 5 is contextual rule and it applies to the word if and only if the word next **bulchinsa** (that is **Motummaa**) is tagged as JJ. However, the word **Motummaa** was initially tagged as NN by initial state tagger and there is no other rule that transforms the word **Motummaa** to JJ. Therefore rule 5 would not be applicable. As a result, there is no transformation rule that changes the NN tag of word **bulchinsa** to JJ.

Similarly, words (**Mootummaa/NN, Naannoo/NN, Biiroon/NN, Barnoota/NN, raawwiin/NN, paakeeji/NN and qulqullina/NN**) in the same sentence above are wrongly tagged for the same reason, lack of transformation rule to JJ.

About 7 NN tags are wrongly tagged to VV (verb). Analysis for some of these words is also given below.

Word	***Desired (correct) tag in the sentence***
1. **Deskii**	Tuulaa/NN waraqaatu/NP **deeskii/NN** koo/PP irra/PR jira/AX ./PN
2. **Siree**	Raadiyoonii/NN **siree/NN** koo/PP bukkeen/AD bana/VV ./PN

Output by the system in the sentence

1. Tuulaa/NN waraqaatu/VV **deeskii/VV** koo/PP irra/PR jira/AX ./PN
2. Raadiyoonii/NN **siree/VV** koo/PP bukkeen/VV bana/NN ./PN

Possible list of transformation rules that transforms a word from NN to VV found in rule list files of the tagger are the following:

1. NN ii fhassuf 2 VV(change from NN to VV if word has 2 suffix as ii).
2. NN e fhassuf 1 VV (change from NN to VV if word has 1 suffix as e).

3. e char VV (change from NN to VV if character a appears anywhere).
4. u char VV (change from NN to VV if character u appears anywhere).
5. NN hin fgoodright VV (change the tag of the word if the word ever appears to the right of the word 'hin').
6. NN VV NEXTTAG AX (change the NN tag to VV of the word if it has next word tag AX).
7. NN VV PREVTAG AD(change the NN tag to VV of the word if it has next word tag AD)
8. NN VV PREVTAG hin (change the NN tag to VV of the word if it has previous word is 'hin').

Rules 1, 2 transform the above word **deskii** and **siree** from NN to VV respectively. In similar way there are also other words that transform their tags to VV due to the above rules. Therefore, wrong transformation from NN to VV can be due to some rules that force some words wrongly to be tagged as VV.

The wrong tag of VV to NN for some words is due to lack of the transformation rules that can transform initial tag of a word NN to VV. Consider the word **bana** in the above sentence 2, out put by the system, that was wrongly tagged as NN :

- Raadiyoonii/NN **siree/VV** koo/PP bukkeen/VV bana/NN. /PN where it should have been tagged as VV.

From the above 8 transformations no rule is applicable to the word to transform it to VV tag. Therefore, it remains unchanged as NN.
Generally many errors of the tagger is found to be due to the lack of transformation rules or due to errors on the rule themselves to correctly transform some words.

5.6 COMPARISON WITH HIDDEN MARKOV MODEL

Comparison with the other tagger developed using hidden Markov model is also done. The comparison of the two approaches was made exactly on the same data corpus with similar cross validation method (ten-fold validation). The result obtained using hidden Markov model

is 70.63% accuracy in bigram model and 68.08 in unigram model, whereas that of modified Brill tagger accuracy is 80.08%. This means that, the result obtained by Brill tagger is much better than HMM for Afaan Oromo language. Summary of the result is depicted in Table 5.10 below.

Table 5.10 Summary of the result of Brill and HMM Comparison

NO	***No. of words***	***Original Brill Tagger Accuracy***	***Modified Brill Tagger Accuracy***	***Accuracy of HMM***	
				Bigram	***Unigram***
1	144	87	93	75.5	74
2	198	66.4	71	67	61.3
3	165	76	81	70	67.7
4	177	73	79	67.5	64.4
5	185	73.4	77	64	63.9
6	151	86.5	91	73	69.9
7	114	89.4	93.2	83	83.3
8	178	74.1	77	68.9	65
9	226	76.4	80	73.1	70.3
10	191	74.2	78	64.3	61
Average		**77.64**	**80.08**	**70.63**	**68.08**

5.7 DISCUSSION

The result obtained in the experiment of transformational error driven rule based (TEL) tagger is a promising approach for Afaan Oromo POS tagger. As this work is groundwork for the language, a small size corpus was used. This limited the number of rules that can be generated from the corpus. As a result of this small set of rules that was generated by the training, the tagger accuracy decreases as the number of input words for the tagger increases. Even though the result obtained by this experiment is interesting which is 80.08% on average, a much better accuracy could also be achieved if the tagger is trained on a large size corpus.

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATIONS

6.1 CONCLUSIONS

Natural Language Processing (NLP) such as POS tagging which annotates natural language sentences or texts with their categories is one of the research areas. In this thesis work transformational error driven learning approach was used in designing Afaan Oromo POS tagger.

In this work gender, number, tense, definiteness of a given word were not considered hence the tagger does not identify this features of the word.

The tagger used 18 tag sets and trained on small corpus approximately 1708 words. The result found in the experiment is relatively good, 80.08% of the total word are correctly tagged.

Even though it has limitations on generating transformation rules for some words in the language, Brill tagger is found to be better for Afaan Oromo than Hidden Markov Model. This can be due to the rule nature of Afaan Oromo.

It was also shown in this research that the Brill tagger accuracy varies with the percentage of the corpus data used for lexical and contextual rule learning training. Best accuracy can be obtained by taking more data to the lexical rule learner (all above 60% cases show good performance). From this it can be generalized that lexical (morphological) rules are more dominant in the language than contextual rules.

6.2 RECOMMENDATIONS

As it is discussed above, development of POS tagger for Afaan Oromo is in its infant stage. Hence, there are several areas of research for Ethiopian language and Afaan Oromo in particular that should be recommended for future researchers in the area of natural language processing. These recommendations are listed below.

- Standardized and readily available corpus is very important for natural language processing. Clearly stating, it affects the accuracy of the works in this area. Therefore, preparation of standardized corpus is recommended for further researches in the area of natural language processing for Afaan Oromo.
- In this work, a small size of corpus was used for training the tagger. Therefore, another research that uses large size corpus is recommended to improve the accuracy of the tagger.
- Another approach for POS tagging of Afaan Oromo can be used such as neural network, hybrid approach that uses Hidden Markov Model and rule based approach or transformational error driven learning and neural network approach.
- Extending this work so that the tagger identifies word features like genders, numbers, persons, tenses etc can also be future research area.

REFERENCES

[1]. Assefa W/mariam "Developing Morphological Analysis for Afaan Oromo Text", Thesis, School of Graduate studies, Addis Ababa University, 2005.

[2]. Dirriba Megersa "Automatic Sentence Parser for Oromo Language", Thesis, School of Graduate studies, Addis Ababa University, 2001.

[3]. Daniel Jurafsky & James H. Martin "An introduction to speech recognition, computational linguistics, and natural language processing", Prentice Hall, 2002

[4]. Kibur Lisanu "Design and Development of Automatic Morphological Synthesis for Amharic Perfective verb", Thesis, School of Graduate studies, Addis Ababa University, 2002.

[5]. Morka Mekonnen "Text to speech synthesis for Afaan Oromo", Thesis, School of Graduate studies, Addis Ababa University, 2001.

[6]. http://www.aaai.org/AITopics/pmwiki/pmwiki.php (last visited on sept 11 2008)

[7]. Wakshum Mekonnen "Development of stemming Algorithm for Afaan Oromo", Thesis, School of Graduate studies, Addis Ababa University, 2000.

[8] Allen J. "Natural Language Understanding", Benjamin Cummings, 1995

[9] Guy De Pauw1, Gilles-Maurice de Schryver et al "Data-Driven Part-of-Speech Tagging of Kiswahili", School of Computing and Informatics, University of Nairobi, Kenya, 2003.

[10] Daniel Jurafsky & James H. Martin "Speech and Language Processing", Prentice Hall, 2000.

[11] Mesfin Getachew, "*Automatic Part Of Speech Tagging for Amharic Language*: An Experiment Using HMM Approach", Thesis, School of Graduate studies, Addis Ababa University, 2001.

[12]Yenewodim Biadgie “Application of Multilayer Neural Network for Tagging Parts of Speech for Amharic Language”, Thesis, School of Graduate studies, Addis Ababa University, 2006.

[13]Tony McEnery&Andrew Wilson “Corpus Linguistic”, Edinburgh University, 2001.

[14]http://www.absoluteastronomy.com/topics/Oromo_language last visited on Monday, February 02, 2009.

[15]Bj¨orn Gamb¨ackGunnar Eriksson “Natural Language Processing” ,School of Information Studies for Africa.

[16]http://www.lmp.ucla.edu/Profile.aspx last visted on Monday, February 02, 2009.

[17]Irresoo Nagi “Caassefama Afan oromo”, Kuraz International, Addis Ababa, 2006.

[18] R *Dale et al* “Handbook of Natural Language Processing”, *2000.*

[19] http://research.microsoft.com/en-us/groups/nlp/ last visted on Monday, February 02, 2009.

[20] Solomon Asres “Amharic Part-of-Speech Tagging Using Hybrid Approach (Neural Network and Rule-Based)”, Thesis, School of Graduate studies, Addis Ababa University,2008.

[21]Getachew Mamo “Automatic Part Of Speech Tagging for Afaan Oromo Language”, Thesis, School of Graduate studies, Addis Ababa University, 2009.

[22]B. Green and G. Rubin, “Automated Grammatical Tagging of English”, Department of Linguistics Brown University, 1971.

[23]Brill Eric, “A Simple Rule-Based Part of Speech Tagger”, 3rd Conference of Applied Computational Language (ACL), Trento, Italy, 1992.

[24] Brill E., “Transformation-Based Error-Driven Learning and Natural Language Processing: a Case Study in Part-of-Speech Tagging, Computational Linguistics”, Vol.21, No. 4, pp. 543-565, 1994.

[25] Yamina Tlili-Guiassa, “Hybrid Method for Tagging Arabic Text”, Laboratoire de Recherche en Informatique LRI, University Badji Mokhtar Annaba Sidi Ammar BP 12 Annaba Algeria, Algeria journal of Computer Science 2 (3): 245-248, 2006 ISSN 1549-3636 © 2006 Science Publications

[26] Daniel Tianhang Hu, “Development of Part of Speech Tagging and Syntactic Analysis Software for Chinese Text”, Massachusetts Institute of Technology, 2001

[27]Baye Yimam, “the phrase structure of Ethiopian Oromo”, University of London, School of Oriental Studeis, Ph.d Dissertation,London,1986.

[28]Mengesha Rikitu. “How to read Oromiffa and use its garammer”,Magic Press, London, 1993.

[29]http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/repor.html visited on Aug 3 2009.

[30] http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html visited on Aug 3 2009.

[31] Haykin S “*Neural Networks*”, Macmillan College Publishing Company, Inc., 1994.

[32] http://www.dtreg.com/mlfn.htm.

[33] M. Marchesi, N. Benvenuto, G. Orlandi, F. Piazza, A. Uncini, .*Design of Multi-Layer Neural Networks with Power-of-Two Weights*. Proceedings of ISCAS-90, IEEE.

APPENDIX A: SAMPLE LEXICAL RULE

NN fi fhassuf 2 NC

NP a fchar NN

daa hassuf 3 JJ

maan hassuf 4 JJ

VV aa fhassuf 2 NN

NN k fhaspref 1 PP

NN sa fhaspref 2 JJ

maan hassuf 4 JJ

PP r fchar PR

baa hassuf 3 JJ

la hassuf 2 JJ

anii hassuf 4 VV

NNP o fchar NN

lee hassuf 3 NN

f hassuf 1 VV

hin goodright VV

NN a fhaspref 1 JJ

ab haspref 2 NN

e haspref 1 PR

wal haspref 3 PP

VV q fhaspref 1 AX

am haspref 2 AD

2 char JN

u char VV

NN ii fhassuf 2 VV

NN e fhassuf 1 VV

e char VV

u char VV

. goodleft VV

APPENDIX B: SAMPLE CONTEXTUAL RULE

NN VV PREVTAG PN

VV NN NEXTTAG JJ

NN VV PREVTAG hin

PP NN NEXTTAG JJ

NN VV NEXTTAG AX

NN VV PREVTAG AD

NN VV PREVTAG PN

VV NN NEXTTAG JJ

NN VV PREVTAG hin

PP NN NEXTTAG JJ

NN VV NEXTTAG AX

NN VV PREVTAG AD

PN VV PREVTAG AD

PP VV PREVTAG AD

VV NN NEXTTAG JN

PP NN NEXTTAG JN

PN NN NEXTTAG JN

PN VV PREVTAG AD

PP VV PREVTAG AD

VV NN NEXTTAG JN

PP NN NEXTTAG JN

APPENDIX C: SAMPLE CORPUS

Gahee/JJ dubartooti/NN baadiyyaa/JJ wabii/JJ soorataa/NN mirkaneesuuf/VV qabataafi/JC murteessaa/JJ ta'e/AX cimsuudhaaf/VV qaamoleen/NN dhimmi/JJ ilaalatu/VV xiyeeffatanii/AD hojjechuu/VV akka/PR qaban/AX ibsame/VV ./PN

Kunuunsi/JJ qabeenya/NN uumamaafi/JC eegumsi/JJ naannawaa/NN wabii/JJ midhaan/NN nyaataa/JJ Tuulaa/NN waraqaatu/VV deeskii/VV koo/PP irra/PR jira/AX ./PN

mirkaneessuuf/VV shoora/NN olaanaa/JJ akka/PR gumaachu/VV ittigaafatamaan/JJ abbaa/NN Taayitaa/NN eegumsa/JJ Naannawaa/NN ibsame/VV ./PN

Kun/PP kakuu/VV Oromoon/NN qabudha/AX ./PN

Guyyaan/NN kun/PP sadarkaa/JJ adduyaattis/NC ta'ee/AX sadarkaa/JJ biyyaa/NN keenyaatti/JJ yeroo/AD jalqabaatiif/AD kabajameera/VV ./PN

Guyyichi/NP guyyaa/NN muddamsaa/NN ta'uuf/VV ./PN

Michuun/NP koo/PP kompyuutara/NN isaa/PP irra/PR jira/VV ./PN

Galmootan/NP wallitti/PP fuunaanee/VV kaa'a/VV ./PN

Galmeen/NP barbaachisaan/NP badanii/VV jiru/AX ./PN

Buna/NN kan/PP addeessutu/VV galmechaa/NN ira/PR jira/AX ./PN

Galama/NN gara/PR biraan/JJ kiisii/NN xarapheezzaa/NN keessaa/AD arge/VV ./PN

Waraana/NN geedaramuuf/VV guddaa/JJ kennuufi/VV fudhatu/VV keessa/PR hojjate/VV ./PN

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all sources of Materials to the thesis have been duly acknowledged.

Mohammed-hussen Abubeker

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

DIDA MIDEKSO (PhD)

SEBSIBE H/MARIAM

Addis Ababa, Ethiopia
February, 2010