ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

Word Sequence Prediction for Afaan Oromo

Ashenafi Bekele Delbeto

A Thesis Submitted to the Department of Computer Science in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science

Addis Ababa, Ethiopia

March 2018

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

Ashenafi Bekele Delbeto

Advisor: Yaregal Assabie(PHD)

This is to certify that the thesis prepared by *Ashenafi Bekele Delbeto*, titled: *Word Sequence Prediction for Afaan Oromo* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor :	*Yaregal Assabie*(PHD)		
Examiner:	*Dida Midekso* (PHD)		
Examiner:	*Solomon Gizaw*(PHD)		

ABSTRACT

Data entry is a core aspect of human computer interaction. Text prediction is one of data entry systems to a computer and other hand held electronics device. It is a process of guessing the words which are likely to follow in a given text segment by displaying a list of the most probable words that could appear in that position. Word sequence prediction assists physically disabled individuals who have typing difficulties, speed up typing speed by decreasing keystrokes, helps in spelling and error detection and it also helps in speech recognition and hand writing recognition.

Even if Afaan Oromo is one of the major languages widely spoken and written in Ethiopia, there is no research conducted on the area of word sequence prediction. Hence, due to the absence of word sequence prediction for Afaan Oromo, people are not enjoying the core benefits of word sequence prediction. In this study, word sequence prediction model is designed and developed. We used the bi and tri-word statistics, and the bi-, and tri POS tag statistics of the language. Initially, the training corpus and user inputs are tokenized and then morphologically analyzed. Subsequently, word statistics model is built for root or stem word and POS tag statistics model is built for root or stem with tag like noun, verb, adjective, pronoun, adverb, conjunction and etc. by using training corpus. After that, the most likely probable root or stem words are suggested. Finally, lexical words are synthesized based on the proposed root or stem words.

The designed model is evaluated based on the developed prototype. Keystroke Saving (KSS) is used to evaluate systems performance. According the evaluation the primary word-based statistical system achieved 20.5% KSS, and the second system that used syntactic categories with word-statistics achieved 22.5% KSS. Therefore, statistical and linguistic rules have good potential on word sequence prediction for Afaan Oromo.

Keywords: word prediction, statistical language modeling, POS tagging, Keystroke Saving

ACKNOWLEDGEMENT

First of all, I would like to thank the Almighty God for entitling me to this opportunity.

I am deeply grateful for my advisor Dr. Yaregal Assabie for his concern, constructive comments, supervision, and encouragements on my work.

I am also grateful to Ato Dula Kefena for his support to manually tag words with their POS in the testing data and construct manual morphological analyzer and synthesizer.

Finally, I thank all my families and friends for their continuous motivation and encouragement during my stay in the university.

3TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

ACRONYM AND ABBREVIATION

AAC	Augmentative and Alternative Communication
CFG	Context Free Grammar
KSS	Keystroke Saving
KSPC	Keystrokes Per Character
KT	Total Number Of Keystrokes
KE	Effective Number of Keystroke
PDA	Personal Digital Assistant
POS	Part of Speech Tagger
PSRG	Phrase Structure Rule Grammar
Sn-gram	Syntactic n-gram

CHAPTER 1: INTRODUCTION

1.1 Background

Data entry is a core aspect of human computer interaction. Images, documents, music, and video data are entered to computers in order to get processed. There are a number of data entry techniques that include speech, chorded keyboards, handwriting recognition, various gloved techniques [1], scanner, microphone, and digital camera [2]. Keyboards and pointing devices are the most commonly used devices during human-computer interaction [3]. Because of its ease of implementation, higher speed, and less error rate, keyboard dominated text entry system [4]. However, one must master the computer keyboard in order to gain the advantage of a keyboard.

Word prediction provides better data entry performance by improving the writing mainly for people with disabilities [5, 6]. Word prediction helps disabled people for typing, speed up typing speed by decreasing keystrokes, helps in spelling and error detection and it also helps in speech recognition and hand writing recognition. Auto completion decreases misspelling of word. Word completion and word prediction also helps student to spell any word correctly and to type anything with fewer errors [7].

In general, word prediction is the process of guessing the next word in a sentence as the sentence is being entered, and updates this prediction as the word is typed [8]. Currently, word prediction implies both “word completion and word prediction” [8]. Word completion is defined as offering the user a list of words after a letter has been typed, while word prediction is defined as offering the user a list of probable words after a word has been typed or selected, based on previous words rather than on the basis of the letter. Word completion problem is easier to solve since the knowledge of some letter(s) provides the predictor a chance to eliminate many of irrelevant words [8, 9].

The task of prediction the most likely word based on properties of its surrounding context is the archetypical prediction problem in Natural Language Processing (NLP) [8]. In many NPL tasks, it is necessary to determine the most likely word, part-of-speech (POS) tag or any other token, given its history or context. Examples include part-of-speech tagging, word-sense disambiguation, speech recognition, accent restoration, context-sensitive spelling correction, and identifying discourse markers [9]. Currently, word prediction is used in many real life applications such as augmentative communication devices [10].

Afaan Oromo is one of the major languages that is widely spoken and used in Ethiopia [11]. Currently, it is an official language of Oromia regional state. It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population according to the 2008 census [11, 12]. In addition, the language is also spoken in Kenya [11]. With regard to the writing system, Qubee (Latin-based alphabet) has been adopted and it became the official script of Afaan Oromo since 1991 [11, 12, and 13]. Besides being an official working language of Oromia regional State, Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region and its administrative zones. Thus, the language has well established and standardized writing and spoken system [12, 13].

To use computers for understanding and manipulation of Afaan Oromo language, there are very few researches attempted so far. These attempts include spell checker [13], text-to-speech system [14], sentence parser [15], morphological analyzer [16], and part-of-speech tagger [17]. We understand the characteristics of the language from these researches which provide hint how to design the system.

1.2 Motivation

Afaan Oromo started to use Qube (Latin-based script) for writing system people who use Qube has difficulties typing and/or Qube use many characters compared to other languages which slow down the typing process. As to our knowledge there is no attempts research on word sequence prediction for Afaan Oromo. Hence, we are motivated to work on word sequence prediction.

1.3 Statement of the Problem

Word prediction is one of the most widely used techniques to enhance communication rate in augmentative and alternative communication [18]. A number of word prediction software packages exist for different languages to assist users on their text entry. Amharic [2, 19, 20], Swedish [21, 22], English [23], Italian [24, 25], Persian [26], Bangle [18] are some of word prediction studies conducted lately. These studies contribute in reducing the time and effort to write a text for slow typists, or people who are not able to use a conventional keyboard.

Like a number of other African and Ethiopian languages, Afaan Oromo has a very complex morphology. It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromo most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Both Afaan Oromo nouns and adjectives are highly inflected for number and gender. In contrast to the English plural marker s (-es), there are more than 12 major and very common plural markers in Afaan Oromo nouns (example: **-oota, -ooli, -wwan, -lee, -an, -een, -oo,** etc.) [27]. Afaan Oromo verbs are also highly inflected for gender, person, number and tenses. Moreover, possessions, cases and article markers are often indicated through affixes in Afaan Oromo. Since Afaan Oromo is morphologically very productive, derivations and word formations in the language involve a number of different linguistic features including affixation, reduplication and compounding [27].

Furthermore, the grammatical structure of Afaan Oromo is unique. Hence, this makes word sequence prediction unique to the language.

The purpose of this study is to design and develop word sequence prediction model for Afaan Oromo with inclusion of context information. Hence, the word predictor will propose root or stem word and morphological features internally with the aim of offering appropriate word form to the user. The developed model can be used in predictive text entry systems and writing aids.

1.4 Objective

General objectives

The general objective of this research is to design and develop word sequence prediction system for Afaan Oromo.

Specific objectives

The specific objectives of this research are to:

- Review supplementary researches conducted on Afaan Oromo and other language s with the aim to find the best approach for this study.
- Collect corpus for training and testing the model.
- Analyze the morphology of word in training corpus.
- Construct root /stem, affixes, and POS tagged corpus.
- Develop word sequence prediction algorithm for Afaan Oromo.
- Develop the prototype of the system.
- Evaluate the performance of word sequence prediction model using collected test data.

1.5 Methods

To achieve the objectives of the research we will use a number of methods.

Literature Review: A number of related works and resources will been reviewed. This consists of thesis, conference and journal articles, white papers and word prediction systems developed for other languages. The large portions of reviewed materials are conference and journal articles. The nature, background history and operational function of word prediction systems will be studied. In addition, a discussion will be made with Afaan Oromo Linguistic experts regarding the linguistic nature of the language like the grammatical structure and morphology of Afaan Oromo.

Data collection: A training corpus containing 23,400 sentences will be used to train the Afaan Oromo word sequence predictor. In addition POS tagged corpus containing 200 sentences will be used to extract representative sentences for testing by means of random sampling method. Simple random sampling method is chosen since every sentence has equal chance of being selected. These corpora are collected from newspapers (Bariisaa, Bakkalcha Oromiyaa and Oromiyaa), journals, criminal code, books, and social media like Facebook, webpages, books which are written by different authors on different issues such as politics, religion, history, fiction and love.

Prototype Development: A prototype will be developed in order to check whether our study works in accordance with the ideas and theories of word sequence prediction.

Evaluation: The prototype of developed system will be evaluated using keystroke saving.

1.6 Scope and limitation

This research will be covered with the aim to model word sequence prediction for Afaan Oromo based on statistical methods, morphological and grammatical agreement rules of the language. Statistical models of root/stem, affixes and rules of the language like part of speech will be included in this work. This study result predicts one word, it can't predict more than one word or phrase.

1.7 Application of Results

Word sequence prediction assist physically disabled individuals who have typing difficulties, speed up typing speed by decreasing keystrokes for mobile phone, computer and other hand held device users. In addition to this, the model is the applicable for other NLP applications such as spelling checkers mobile phone/PDA texting disabled users, handwriting recognition and word-sense disambiguation and furthermore, this study will be a stepping stone for further researches that can bring the aforementioned advantages for the language.

1.8 Organization of the Rest of the Thesis

The rest of this thesis is organized as follows. In Chapter 2, literature review briefly states fundamental concepts of word prediction, methods of word prediction, structure of Afaan Oromo and its grammatical rules. Chapter 3 presents researches conducted by different scholars on the topic of word sequence prediction, their approaches, and findings. In Chapter 4, architecture of the proposed word sequence prediction model, its approach, and related concepts are explained. Experiment and results are presented in Chapter 5. Finally, conclusion and future works are stated in Chapter 6.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter concentrates on major concept of word prediction and ideas associated with linguistic characteristics of Afaan Oromo. Statistical, knowledge based, and heuristics are prediction methods that are presented in order to understand basic concepts of the research area. Since the aim of this study is design and develop word sequence prediction model for Afaan Oromo, the structure of Afaan Oromo like morphological characteristics, grammatical properties, and parts of- speech of the language are discussed in respective sections of this chapter.

2.2 Structure of Afaan Oromo

2.2.1 Parts-of-Speech in Afaan Oromo

Parts-of-speech are specific classes of a word in a text or corpus. POS tagging is one of a very important task for any natural language processing activities. POS tagger is an application which assists to assign words to their suitable word class like noun, adjective, verb, etc. In many word prediction studies [17, 29] POS tagging and POS n-gram models are used to optimize word prediction task.

In Afaan Oromo, words are generally categorized in different word classes. Noun, pronoun, adjective, verb, adverb, conjunction, and preposition are the common word classes or part of speech (POS). Part-of-speech (POS) tagging is the process of marking the words in a text corresponding to a particular part of speech or lexical category, based on its definition as well as its context. That means POS tagging assigns whether a given word is used as a noun, adjective, verb, etc. POS tagging is among the most distinguished disambiguation problems. A POS tagger tries to assign the corresponding POS tag to each word in sentences, taking into account the context in which this word appears [17, and 29].

Noun

In general nouns refer to a person, an object, or abstract ideas. A noun in Afaan Oromo is treated as either male or female, though there are typically no gender markers in the words themselves. Gender can be presented through a demonstrative pronoun, a definite article, a gender-specific adjective, or the verb form [29, 30, and 31].

Example

Noun	Meaning
Barnoota	Education
Barsiisaa/ Barsiiftu	Teacher
Barataa/baratuu	Student

Pronouns

Pronouns are words or morphemes that can be used in place of nouns in a sentence. They are limited in number and they can be categorized in different sub categories like personal, possessive, interrogative, demonstrative pronouns and the like. Personal pronoun represents speaker, listener and third party in any speech. Table 2.1 shows the personal pronouns as subjects and direct objects [29, 30, and 31].

Table 2.1: Personal pronouns as subjects and direct objects

		Subject Pronouns		Direct Object Pronouns		Possessive Pronouns	
First person	singular	**ani**	I	**na**	me	**koo**	my, mine
	plural	**nuti, nu'i**	We	**nu**	us	**keenya**	our, ours
Second person	singular	**ati**	you	si	you	**kee**	your, yours
	plural	**isin**	you	**isini**	you (pl.)	**keessan(i)**	your, yours (pl.)
Third person	singular	**inni**	He,it	**isa**	him, it	**(i)saa**	his, its
	singular	**sheen**	she	**ishee**	her	**ishee**	her, hers
	plural	**isaan**	they	**isaani**	them	**(i)saani**	their, theirs

Afaan Oromo pronouns are used on various forms in order to show their role in a given sentence. English uses “he” for subjects and “him” while “he” and “him” may refer to the same person. Afaan Oromo has a number of forms for all nouns, including pronouns, though for now we will only deal with the subject (nominative) and direct object (accusative) forms[29, 30, and 31].

Reflexive and Reciprocal Pronouns: In Afaan Oromo reflexive pronouns like myself', yourself and etc are expressed in two ways. One is to use the noun meaning "self": **of(i)** or ***if(i).*** This noun is inflected for case but, unless it is being emphasized, not for person, number, or gender***: isheen*** of ***laalti*** "she looks at herself" (base form of of), ***isheen of iifmakiinaabitte*** "she bought herself a car" (dative of of).The other possibility is to use the noun meaning 'head', ***mataa***, with possessive suffixes: ***mataakoo*** "myself", ***mataakee*** "yourself", etc [29, 30, and 31].

Demonstrative pronouns: Like English, Afaan Oromo makes a two-way distinction between proximal ("this, these") and distal ("that, those") demonstrative pronouns and adjectives. Some dialects distinguish masculine and feminine for the proximal pronouns; in the western dialects the masculine forms (beginning with k-) are used for both genders. Unlike in English, singular and plural demonstratives are not distinguished, but, as for nouns and personal pronouns in the language, case is distinguished. Only the base and nominative forms are shown in the table 2.2; the other cases are formed from the base form as for nouns, for example, ***sanatti*** "at/on/in that" (locative case).Table 2.2 shows Afaan Oromo demonstrative pronouns [29, 30, and 31].

Table 2.2: Example for Afaan Oromo demonstrative pronouns

Case		Proximal("this, These")	Distal(that, those)
Base	Masculine	**kana**	**san**
	Feminine	**tana**	
Nominative	Masculine	**kuni**	**suni**
	Feminine	**tuni**	

Interrogative pronouns: Interrogative pronouns are used when we need to ask questions about something or someone. In English, there are pronouns like who, when, what, where, whose and whom that are used to ask questions. Likewise, Table 2.3 shows Afaan Oromo interrogative pronouns [29, 30, and 31].

Table 2.3: Example of Afaan Oromo interrogative pronouns

Maal(i)	Maaliif(i)	Akkam(i)	Yoom	Eessa	Eenyu	Kan eenyu	Meeqa	Kam(i)
What	Why	How	When	Where	Who	Whose	How much/,many	Which

Verb

Verbs are doing words or action words (for instance, **fig**i "run", **rukuti** "hit", **konkolachisi**"drive", and **nyaadhu** "eat"), but some verbs show a "state of being" (for instance, **mullate**"appear" and **fedhe** "like") [29, 30, and 31].

Adjective

Adjective is a word that describes or qualifies a noun or pronoun and it appears after a word it modifies unlike to English and Amharic. It gives more information about noun or pronoun it modifies. Objects are differentiated from one another by different attributes like shape, behavior, color, etc. and this difference is described using adjective word class. Adjectives are inflected for gender, number and case in a similar technique to nouns [27]. In the following example, the underlined word is adjectives.

Example: Isheen mana <u>bareedaa</u> keessa jirti. This means she is in beauty salon. The word **bareedaa** "beauty" is the adjective.

Adverb

An adverb is a word that is used to change or qualify the meaning of an adjective, a verb, a clause, another adverb, or any other type of word or phrase with the exception of determiners and adjectives that directly modify nouns. In Afaan Oromo, adverbs come before a verb and it is used to modify verbs. Traditionally considered to be a single part of speech, adverbs perform a wide variety of functions, which makes it difficult to treat them as a single, unified category. Adverbs normally carry out these functions by answering questions such as when, how, where, in what way/how and to what extent/how much [29].

For example:**Waaqoo kaleessa dhufe**. This means Waqo came yesterday. The word kaleessa"yesterday" is the adverb.

Conjunction

Conjunction is a connecting word that is used to link words, phrases, clauses, sentences, etc. They are limited in number and can be used with verbs, nouns and adjectives [30, 31].

For example: ***fi*** /and, ***garuu, immoo***/but,***yookin***/or, etc.

Preposition

Prepositions are words that are usually used before nouns to show their relation to another part of a clause and they are limited in number [31].

For example**:** ***gara***/towards, ***haga***/until,***akka***/like etc.

2.2.2 Morphology of Afaan Oromo

Morphology is a branch of linguistic that studies and describes how words are formed in a language [12]. There are two kinds of morphology: inflectional and derivational. Inflectional morphology is concerned with the inflectional changes in words where word stems are combined with grammatical markers for things like person, gender, number, tense, case and mode. Inflectional changes do not result in changes of parts of speech. Derivational morphology deals with those changes that result in changing classes of words (changes in the part of speech). For instance, noun or an adjective may be derived from a verb.

In linguistic, the minimum unit of morphology is called morpheme. Afaan Oromo morphemes are divided into two: ***Dhamjecha walabaa***"free morphemes" and ***Dhamjecha hirkataa***"bound morphemes" [12, 33, and 34]. Free morphemes are those morphemes that can stand alone, that is, not attached to some other morpheme to give a meaning. By contrast, bound morphemes have to be attached to some other morpheme to give a meaning. For instance, ***adunyaarratti***"in the world", /***-irra-***/ and ***/-itti***/ are bound morpheme and ***adunya*** "world" is a free morpheme.

Afaan Oromo free morphemes are classified into two: ***Dhamjecha hiikaa*** "lexical morpheme" and ***Dhamjecha tajaajilaa*** "functional morpheme". Lexical morphemes are free morphemes that have their own content and stand for one meaning. For example, if we use the word ***Bishaan*** "water" as a subject or as the object of a sentence its meaning never changes. In contrast, functional morphemes generally perform some kind of grammatical role, carrying little meaning of their own. Functional morphemes specify a relationship between other morphemes. For instance, words like ***Kun****"this",****Sun*** *"that",*, and ***Kana****"this",* are all functional morphemes.

Based on their content, bound morphemes are classified into two***: hundee hirkataa***"bound root" and ***fufii***"affix" [12, 33, 34]. Bound roots are morphemes that make the most precise

and concrete contribution to the words meaning. For instance, words like ***deema, deemta, deemu, deemti, deemna***, and ***deeme***, have a bound root **deem**-"go". Most of the root words in Afaan Oromo are bound root [12, 33]. Affix is a bound morpheme that attaches to bases. The classification of Afaan Oromo affix can be done based on the position of the affix and interms of the shift of word class [12, 34].

Based on their location, Afaan Oromo affixes are classified into four***:*** ***fufiiduree***"prefix",***fufiigiddee***"infix",***fufiinaannee***"Circumfix" and ***fufiiduubee***"suffix". Again based on their grammatical functionality and the type of word class they change, affixes are classified into two: ***fufileeyaasaa***"derivational affixes" and **fufileehortee** "inflectional affixes".

Afaan Oromo verb consists of a stem and one or more suffixes. Some writers also treat the negative morpheme ***hin*** as a prefix, for example, writing ***hindubbadhu*** in place of the more usual **hindubbadhu**'I don't speak'. We'll first look at the different suffixes that are possible and then look at the stem [12, 15].

An Afaan Oromo noun consists of a stem followed by one or more suffixes. All Oromo nouns have one of the following cases: base, subject, dative, instrumental, ablative, or locative. The base form normally has no explicit suffix. The realization of the dative, instrumental, ablative, and locative cases is quite complex (in fact probably not well understood by linguists), and the program may assign too many analyses to some word forms. There are at least eight possible conjunctive suffixes and the focus suffix ***-tu***, which apparently cannot co-occur with any of the conjunctions. A noun precedes a verb with a first person singular subject; it takes the first person singular subject suffix **- *(a) n***.

Some writers treat the Oromo possessive adjectives as suffixes, though this practice seems to be less common now than it used to be: ***mannikee***; ***mannikee*** 'your house (base)' [12].

2.2.3 Morphology of Nouns

Noun inflection

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

There are singular and plural numbers in Afaan Oromo and it has different suffixes to form the plural of a noun. The use of different suffixes differs from dialects to dialects. Majority of noun plural forms are formed by using the suffix **-oota**, **-oolii**, **-ee**n, **-lee**, **-wwan**, **-yyii**, ***-eetii***, ***-ii***, and ***–oo***[27].Even though the usage of some suffixes is not common, they can be used and accepted by the speakers of the language. Linguists agree that some groups of suffixes are most preferably applied to almost all nouns, and the others are used with only some words [12, 27, and 36]. The following example shows the corresponding noun plural maker suffix.

Base form	Suffix	Inflated form	Meaning
waggaa	***-oota***	**waggoota**	Years
kitaaba	***-ota***	***kitaabota***	Books
ilma	***-an***	***ilmaan***	Sons
baatii	***-lee***	***baatiilee***	Months
aanaa	***-olee***	***aanolee***	Districts
muka	***-een***	***muk-een***	Woods
balaa	***-wwan***	***balaawwan***	Accidents

Two types of gender, that is, masculine and feminine, exist in Afaan Oromo. These are identified through gender marking suffixes, or lexically by using different words for

masculine and feminine forms.The suffix **-ttii**/ and **-tuu** are used as feminine gender marker whereas **-aa** marks masculine gender in Afaan Oromo. The distinct words for masculine and feminine like ***adaadaa*** "aunt" and ***eessuma*** "uncle" are also used in Oromo. Gender indicating words can be used for animals and they are placed immediately after or before the nouns they belong to. The most common contrastive pair of words used in this way is ***kormaa*** "male" vs. ***daltuu***"female" [36].The following example shows the corresponding noun plural maker suffix.

Stem	Suffix	Masculine	Stem	Suffix	Feminine	Meaning
barsiis	**--aa**	**barsiisaa**	**barsiis**	**--tuu**	**barsiistuu**	teacher
barat	**--aa**	**barataa**	**barat**	**--tuu**	**baratuu**	student
herreg	**--aa**	**herregaa**	**herreg**	**--tuu**	**herregduu**	accountant
barreess	**--aa**	**barreessaa**	**barreess**	**--tuu**	**barreessituu**	writer
leenjis	**--aa**	**leenjisaa**	**leenjis**	**--tuu**	**leenjistuu**	coach

Afaan Oromo has no indefinite articles (corresponding to English a, an, or some), but it indicates definiteness (English the) with suffixes on the noun: ***-icha*** for masculine nouns and **-ittii** for feminine nouns. Vowel endings of nouns are dropped before adding these suffixes: ***karaa*** "road", ***karicha*** "the road", ***nama*** "man", ***namicha*** "the man", ***haroo*** "lake", ***harittii*** "the lake". Note that for animate nouns that can take either gender, the definite suffix may indicate the intended gender: ***qaalluu*** "priest", ***qaallicha*** "the priest (masculine)", ***qallittii***"the priest (feminine)". The definite suffixes appear to be used less often than the in English, and they do not co-occur with the plural suffixes [34].

In Afaan Oromo, noun can inflate to show nominative, accusative, agentive, dative and ablative case. Nouns that are used as subject of intransitive verbs and agent of the verbs take as the inflectional morpheme for nominative case. Nominative case is marked by four different morphs of allomorphic variation occurring in complementary distribution [36].

Base form	Suffix	Inflected Forms	Meaning
Siree	**-n**	**Sireen**	Bed
Hirriba	**ni**	**Hirribni**	Sleep
Morma	**-i**	**Mormi**	Neck

Noun derivation

Derivation is a process of word formation in which one or more affixes is attached to a root word (and stem) to produce a new word known as derived word. Usually, derivation will change the part of speech of the root word to which a suffix is added. This process of word formation is very productive. Afaan Oromo is derivationally rich language [12, 13, and 34].

The process of driving a noun from the other word class is called nominalization, and the types of affixes used for this purpose is called nominalizars. In Afaan Oromo, there is a large stock of nominals derived from adjectival, verbal and nominal bases. Suffixes involved in the derivation of nouns in Afaan Oromo are classified into different groups based on the type of word class they change into nouns [12, 13, and 34].Abstract nouns are derived from other nouns by adding the suffix ***-ummaa,-eenya*** or ***-ooma*** to the noun stems. Thus, when these abstract noun formative morphemes are added to nouns, the final vowels of these words are deleted as the following set of examples illustrate.

Noun	Meaning	maker	Derived noun	Meaning
bilisa	free	**-ummaa**	b**ilisummaa**	freedom
garba	slave	**-ummaa**	**garbummaa**	slavery
fira	relative	**-ummaa/-ooma**	**firooma**	relationship

In Afaan Oromo, the nominal can be derived from the verb stem by suffixing the morphemes like **-a**,***-aa,-eenya, -tuu, -ina, -noo, -ii, -ee, -iinsa, -iisa, -umsa, -maata, -aatii***. [16] The following examples indicate the derivation of such nouns.

Verb	Meaning	Maker	Derived noun	Meaning
ibse	make it clear	**-aa**	**ibsaa**	light
lole	he fought	**-a**	**lola**	war
dhuge	he drank	**-aatii**	**dhugaatii**	drink

The suffixes –***ina*** and –***enyaa***are used to derive nouns from adjectives. The following examples indicate the derivation of nouns from adjective.

Adjective	Meaning	Maker	Derived noun	Meaning
adii	white	**eenyaa**	**addeenya**	whiteness
dhiyoo	near	**-eenya**	**dhiyeenya**	closeness
bareedaa	beautiful	**-ina**	**bareedina**	beauty

2.2.4 Morphology of Verbs

Inflection of verbs

Different inherent and agreement grammatical categories account for the inflection of verbs in Afaan Oromo. The inherent ones are aspect, mood, and voice whereas the agreement properties include person, number, gender and case. Several studies, especially the earlier ones, consider tense in the inflectional categories of verbs in Afaan Oromo. From the three major tenses present, past and future, Oromo mainly identifies between past and non-past in its morphology because the morphological markers do not distinguish each tense types. For example, present and future tenses are not distinctly marked in the morphology of the

language when tense is considered. The morphological distinctions are overtly marked on verbs point to aspect rather than tense [36].

The three main functional domains of inherent verb inflection in Afaan Oromo are aspect, mood and voice with some indications of tenses.

Aspect

Aspect is context related which morphologically distinguishes between completeness and incompleteness of an action. It is bound with situation and duration unlike tense which is just about time of an event in relation with the speech time. In Afaan Oromo, the roots or stems of verbs, usually ending in consonant, take inflectional morphemes showing distinction between perfective and imperfective aspects. The two aspects are distinguished primarily by their suffix vowel, which is ***-a*** (and its allomorphs ***-i*** and ***-u***) for the imperfect and ***-e*** (and its allomorph ***-i***) for the perfective. The continued actions are categorized as imperfective aspect whereas a short and completed action can be considered as perfective aspect [36].

The perfective verb denotes a single event that happens which is seen as a whole regardless of duration. In their two major parts showing past and non-past, Afaan Oromo verbs are marked for such distinction of aspect. Perfective aspect can, of course, be illustrated in connection with a sense of past tense. The concept of perfectness is that an action is prior to a specific moment in time whereas the imperfectness is connected with an action in process or in progress. The perfect form indicates an action is complete at a specific time in the past **iseen hojicaa tumurte "**She finished the work". Whereas imperfective aspect indicates a longer lasting action as in ***ɗufaa jira***"he is coming".

Voice

Voice is a verb form that relates action of a verb with its participants (or arguments). It tells us if the subject performs or receives the action indicated by the verb. When the subject performs the action, the voice is active whereas the form in which the subject receives the action is passive voice. Using sentence types in which the verb form is changed for the purpose of such grammatical function is inflectional. Passive formation is a syntactic process in which the subject object exchange happens so that subject in active becomes object in the passive form and vice versa. Afaan Oromo passive formation is purely morphological as it is formed by adding the morpheme **-am** on transitive verbs. Based on the lexical-functional approach, this thesis treats voice as morphological form in inflection [36].

The passive morpheme **-am**, in Afaan Oromo, is an invariable morpheme across subjects and aspects. Voice involves all valency changing verb forms including causative and middle. Here are few examples:

Voice	Root	Maker	Inflated form	Meaning
Active	***Kut-***	-	***Kute***	Cut
	Gurgu-	-	***Gurgure***	Sold
Passive	***Kut***	***-am***	***Kutame***	Was cut
		-am	***Gurgurame***	Was sold

Verb derivation

Like that of nouns, Afaan Oromo verbs are derivational. Similarly suffixes involved in the derivation of verbs (verbilizer) in Afaan Oromo are classified into different groups based on the type of word class they change into verbs [16, 34].

The suffixes **-oom-**, **-aa'-**, **-a'-** are verbalizer that used to derive verbs from nouns and adjectives. The following examples indicate the derivation of such verbs.

Noun/adjective	Meaning	Maker	*Derived verb*	Meaning
Arjaa	generous	***-oom-***	***Arjoome***	Give
gurraacha	black	***-a-***	***Gurraacha'ee***	Be black
qulqulluu	cleanness	***-aa-***	***Qulqullaa'uu***	be clean

The affixes **-at-, -am-, -sis-, -siis, -s-** are verbalizer that are used to derive verbs from adjectives and another verb. The following examples indicate the derivation of such verbs.

Verb /adjective	Meaning	Maker	Derived verb	Meaning
Mure	***Cut***	***-at-***	***Murate***	Cut for himself
diimaa	***red***	***-at-***	***diimate***	became red
hidhe	***bundle***	***-am-***	***hidhame***	to be imprisoned
dhuge	drunk	***-siis-***	***dhugsiise***	

2.2.5 Morphology of Adjectives

Adjective inflection

The inflectional categories or properties of adjectives are the same with that of nouns. Adjectives are inflected for number, gender, singulative and case like nouns; however, sometimes they are marked differently from nouns. For instance, adjectives, unlike nouns, are inflected by reduplication to mark plurality [34, 36].

When adjectives occur with nouns in sentences, number is marked on both of them. Nouns are marked for plurality, but adjectives are marked for number by reduplication of its initial

syllable (CV, CVC), or by the plural suffix-(***o)ota***. In the former way of marking plurality, the initial syllable reduplication co-occurs with the final vowel shift from **-aa** to –***oo*** when the adjectives end in long ***-aa***. The latter way of marking number in adjectives is the same with that of nouns. The suffixes***-(o)ota*** ,**-yyii** show plurality in adjectives[34,36].Here different ways of marking plural adjectives with examples:

Base form	Affix	Inflated form
dheeraa(m) **dheertu**(f) long	**dhe**	**Dhedheeraa**(m) **Dhedheertuu** (f)
Jabaa(m) **Jabduu**(f) strong	**jab**	**Jajabaa**(m) **Jajjabduu**(f)
Dureessa/rich(m)	**-yyii/-ota**	**Dureeyyi**/**duressota**
Dureettii/rich(f)	**-wwan**	**dureettiwwan**

Adjective derivation

Forming adjectives from other lexical categories is termed as adjectivization. From stative verb like /***diim-at***/ "become red" one can derive the adjective /***diim-at-aa (-tuu***)/ " reddened" . In Afaan Oromo adjectives can be formed from verbs by taking adjectivizers like /***-aa***/, /***-tuu***/, /***-eessa***/, and /***- eettii***/. The following examples indicate the derivation of such adjectives [12, 13, and 16]. Afaan Oromo adjectives are not derivatives as noun and verbs, this is because there are a few number of adjectivizers in the language [12, 13, 16].

sodaate	fearful	***-aa/-tuu***	***Sodaataa(m)/sodaattuu(f)***	coward
iyye	cry	***-essa/-etti***	***Iyyeessa(m)/iyyeettii(f)***	poor

2.2.6 Grammatical Structure

Grammar is a set of structural rules governing the composition of sentences, clauses, phrases, and words in a given natural language. These rules guide how words should put together to make sentences. Word order and morphological agreements are basic issues considered in Afaan Oromo grammar and are used as part of our word sequence prediction study. A sentence is a group of words that express a complete thought. Sentences are formed from verb phrase and noun phrase and can be classified as simple and complex sentences. A phrase is a small group of words that stands as a conceptual unit. Simple sentences are formed from one verb phrase and one noun phrase whereas a complex sentence contains one or more subordinate verbs other than the main verb, where subordinate verbs are verbs that are integrated with conjunctions. A sentence is said to be complex because it has capability to contain other sentences within it [37].Table 2.4 shows simple and complex sentences.

Table 2.4: Examples of simple and complex sentences in Afaan Oromo

Simple sentence	**Gammaachu kalessa dhufe** "Gemechu came yesterday"
Complex sentence	**Gammaachu kalessaa dhufe fi Kitaaboota isaa fudhe**" Gemechu cameyesterday and he took his books"

A subject is part of a sentence or utterance, usually noun, noun phrase, pronouns or equivalent that the rest of a sentence asserts something about and that agrees with the verb. It usually expresses an action performed by a verb. In Afaan Oromo sentence, subjects more often occur at the beginning of a sentence. The subject of a sentence should be in accordance with verb in gender, number, and person.

In the sentence, ***Roobaa intala isaa waammee*** "Roobaa called his daughter", the subject **Roobaa** "*Roobaa*" shows person, gender, number information which is third person, masculine, and singular respectively. This morphological property is reflected on the verb, ***waammee*** "called**"**.

Therefore in order to predict words in proper morphological information, morphological properties of subject of a sentence should be captured and properly used on the verb while providing word suggestions.

Object and Verb Agreement

In Afaan Oromo object of the sentence has no any grammatical relation with the subject and verb of the sentence [37, 38].

Examples:

1. *Isheen isa jaalatti*. "She likes him. **"**
2. *Inni ishee jaalata*. "He likes her. **"**
3. *Nuti isa binna*. "We buy it. **"**

Adjective and Noun Agreement

Adjectives are very important in Afaan Oromo because its structure is used in every day conversation. Oromo adjectives are words that describe or modify another person or thing in the sentence [27, 37]. Afaan Oromo adjectives should be in agreement in number and gender with the noun it modifies. Afaan Oromo adjectives may mark number (singular or plural) and gender (feminine or masculine or neutral) of a noun it qualifies and hence it should agree with number and gender of the noun [27, 36].

For example: In noun phrase ***namootabeekoo*** **"**knowledgeable men", the word ***beekoo*** is an adjective that modifies the noun ***namoota*** "men. It is marked for plural number and is reflected on the noun. It is inappropriate to write the above phrase as ***namabeekoo*** "knowledgeable man", since it shows number disagreement between the adjective and noun.

To write this incorrect grammatical format either the adjective should be marked with singular number ***namabeekaa*** knowledgeable man or the noun should be marked with plural number. Noun phrase, ***namichafurda*** "The fat man", the word ***furdaa*****"**fat**"** , is an adjective that modifies the noun **namicha** "The man**"** . It is marked with masculine gender and is in agreement with the noun. However if we take a phrase ***namichafurdoo*** **"**The fat man", the adjective is marked with feminine gender while the noun it modifies is masculine. Therefore the adjective and noun are in disagreement and to avoid this kind of inconsistency either the adjective should be marked with masculine or the noun should be marked with feminine gender. For this particular example an appropriate phrase is either ***namicha furda*** "the fat man" or ***namiti furdoo*** "The fat woman", where there is agreement in number and gender between the adjective and noun.

Adverb and Verb Agreement

Oromo adverbs are part of speech. Generally, they're words that modify any part of language other than a noun. Adverbs can modify verbs, adjectives (including numbers), clauses, sentences and other adverbs [27].

For example: In a sentence ***Guta boru dhufa*** "Guta will come tomorrow", the word **boru** "tomorrow" is an adverb that modifies the verb **dhuf**a "will come". The adverb and verb are in agreement taking imperfective tense form.

2.3 Word prediction

A number of people with physical disabilities were rocketed dramatically after world war second [39]. In order to assist them to interact with the outside world, assistant technology such as word prediction was used. Researchers dedicated to develop systems that are alternative to the users' disabilities and could augment their abilities too. Since the early 1980, the prediction systems have been in use [39].

Word prediction is about estimating what word the user is going to write for the purpose of facilitating the text production process [18, 39]. Sometimes a distinction is made between systems that require the initial letters of an upcoming word to make a prediction and systems that may predict a word regardless of whether the word has been initialized or not [39, 40]. The former systems are said to perform word completion while the latter perform proper word prediction.

Prediction refers to those systems that figure out which letters, words, or phrases are likely to follow in a given segment of a text. Such systems are very useful for user, mainly the ones with writing disabilities. The systems usually run by displaying a list of the most probable letters, words, or phrases for the current position of the sentence being typed by the user. As the user continues to enter letters of the required word, the system displays a list of the most probable words that could appear in that position. Then, the system updates the list according to the sequence of the so-far entered letters. Next, a list of the most common words or phrases that could come after the selected word would appear. The process continues until the text is completed [39].

Whenever the user types a letter or confirms a prediction, the system updates its guesses taking the extended context into account. The size and nature of the context on which the predictions are based, varies among different systems. While the simplest systems only take single word form frequencies into account, thus not at all making use of the previous context, more complex systems may consider the previous one or two word forms and/or the grammatic categories. Yet more complex systems combine these methods with other strategies such as topic guidance, recency promotion and grammatical structure.

The goal of all writing assistance systems is increasing the Key Stroke Saving (KSS) which is the percentage of keystrokes that the user saves by using the word prediction system. A higher value for KSS implies a better performance; as a result, decreasing the users effort to type a text. In other words, the amount of text to be typed needs to be as short as possible for

the user with the least effort. KSS is one of the important standard performance metrics to evaluate prediction systems [39, 41, and 42].

There are numbers of prediction systems that were developed and are developing with different methods for different languages. In these sections three major approaches are described: statistical modeling, knowledge-based modeling, and heuristic modeling (adaptive).

2.3.1 Statistical Modeling

Traditionally, predicting words has solely been based on statistical modeling of the language [43]. In statistical modeling, the choice of words is based on the probability that a string may appear in a text. Consequently, a natural language could be considered as a stochastic system. Such a modeling is also named probabilistic modeling. The task of predicting the next word can be stated as attempting to estimate the probability function P:

$$P(W_n/W_{n-1}, \dots, W_2, W_1) \quad (2)$$

A Markov model is an effective way of describing a stochastic chain of events, for example a string of words. Such a model consists of a set of states and probabilities of transitions between them. The transition probabilities represent the conditional probabilities for the next word given the previous words. For example, the probability of a transition from state AA to state AB represents the probability that B is written when the two previous words were AA.

Sequences of words extracted from the training texts are called n-grams. In this report, 1, 2, and 3-grams are named uni-, bi-, and tri-grams, respectively. The probabilities for the transitions in a second order Markov model can be estimated simply by counting the number of bi-grams and tri-grams in the training text, and by using the relative frequency as an estimate. Thus

$$P(W_n/W_{n-2}, W_{n-1}) = \frac{C(W_{n-2},W_{n-1},W_n)}{C(W_{n-2},W_{n-1})} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Where is the count of *n*-grams ,W_{n-2}, W_{n-1},and W_n are words ,$P(W_n/W_{n-2},W_{n-1})$ is probability of a word W_n given W_{n-2}, W_{n-1}previous word, $C(W_{n-2},W_{n-1},W_n)$ is frequency of word sequence W_{n-2} ,W_{n-1},W_n in a corpus, $C(W_{n-2},W_{n-1})$ is frequency of W_{n-2},W_{n-1} in a corpus[25].

The probabilities for the transitions in a first order Markov model can be estimated simply by counting the number of uni-grams and bi-grams in the training text, and by using the relative frequency as an estimate. Bi-gram probability is computed using (Eq.4)[25].

$$P(W_n/W_{n-1}) = \frac{C(W_{n-1}, W_n)}{C(W_{n-1})} \qquad (4)$$

where ,C is the count of n-grams,W_{n-1}, W_n are words , $P(W_n|W_{n-1})$ is probability of a word W_n given W_{n-1}, $C(W_{n-1},W_n)$ is frequency of word sequence $W_{n-1}W_n$ in a corpus, $C(W_{n-1})$ is frequency of W_{n-1} in a corpus.

The m^{th} order Markov model requires m+1-grams to be extracted from the training texts in order to calculate.

In such a stochastic problem, we use the previous word(s), the history, to predict the next word. To give reasonable prediction to the words which appear together, we try to use the Markov assumption that only the last few words affect the next word [25]. So if we construct a model where all histories restrict the word that would appear in the next position, we will then have an $(n-1)^{th}$ order Markov model or an n-gram word model [25].

The statistical information and its distribution could be used for predicting letters, words, phrases, and sentences [39].

2.3.1.1 Letter Prediction

Prediction approaches have applied into current electronics devices such as computer, cell-phones and PDAs. Letter prediction could be used as an aiding tool to enter a text on Short Message Service (SMS), to chat on Instant Message, and to write an email. Because of being portable, such systems could not have a single button for a letter. So, a text should be entered with a limited number of keys (e.g., in cell-phones, a text is written with only 9 buttons on the phone) [44]. This means that a key should carry three or four letters. The reduced keyboard makes it hard for the user to enter a text; so, the letter prediction method would be an efficient way. The reason to have such a system is that the user will need to press only one button for each character on the mobile phone [39, 44]. But how the letters would be disambiguated in a single button? Three methods are presented:

The Lexical-based Predictive Text Entry Method: One of the data entry methods is to press one key for one character. This needs a system that matches the key arrangement to the corresponding words in a lexical dictionary. In this method, the most frequent words that match with the key sequence will be displayed. In fact, the method prefers simply targets at disambiguating the sequence of letters than predicting them. The user can look through the resulting word list and choose the proposed word if the key sequence corresponds to two or more words [45, 46, and 47].The keystrokes per character (KSPC) for such systems are greater than one.

Word Wise: was introduced and developed by Eatoni Ergonomics [48]. Auxiliary key is used. In this approach, a key is selected as an auxiliary key that is pressed simultaneously with the corresponding key to the character. Both the auxiliary key and the letters that need to be pressed simultaneously by the key are red. For example, Key 1 is the auxiliary key, and letters c, e, h, l, n, s, t, y are red[40].

Letter Wise: Letter Wise is also another method that developed by Eatoni Ergonomics. The letter diagram probabilities are considered. The most likely letters are selected by the system. The letters disambiguating are based on the previously entered characters rather than on the lexical dictionary in itself, due to this fact the amount of memory required for the methods is small. Entering new words is to the system so easy [44].

2.3.1.2 Word Prediction System

Word prediction systems estimate the words that the user is going to use. These words are proposed in a list to the user that might be used in that position. The method attempts to ease writing a text to the user. Statistical language modeling is broadly used in these systems. Markov assumption is used as a base line for statistical word prediction which only the last n-1 words of the history have effect on the next word [47].Thus, the model could be named n-gram Markov model. Word frequency and word sequence frequency, usually, are the methods that are frequently used in word prediction systems, particularly for the ones that are established commercially.

Word sequence Frequency: Constructing a dictionary that contains words and their corresponding relative frequency of occurrence is the most common and simplest word prediction methods. It comes up with the most frequent words begin by this string in similar order they are stored in the system. This method may require some improvement by a user in order to amend its concordance when used to inflected words since context information are not considered. In addition, this method uses unigram model with fixed lexicon and it provided with the same proposal for similar sequences of letters. To increase word prediction accuracy result, sign about recency of use of each word may be involved in the lexicon. In this way, the prediction system is capable to provide most recently used words among most likely words. This method provides access to adaptation of each word to a user‘s vocabulary by updating frequency and recency of each word used [6, 39].

Most likely words that start with the similar characters are offered when a user has typed the beginning of a word. If the required word is not accessible among options presented by the system, a user may continue typing , else the required word is used from the given list and it may automatically adapt to user‘s lexicon by means of simply updating frequencies of words used and assigning an initial frequency for new words added to the system. In order to improve the result of this approach, recency value is stored in a dictionary each word with their corresponding frequency information. The outputs found with recency and frequency based methods are better than the ones based on frequency alone. However, this method requires storage of more information and increases computational complexity [6, 49].

Word Frequency: A unigram word model is used for early systems. It was cleared that some of the suggestions are not suitable in that position of a sentence. If a context is taken into account, suggestions will be better. As a result, researchers enforced to develop systems that used history as a hint for presence of the next words. A model that considers only the previous word in order to predict the next word in current position of the sentence being typed was named bigram word model or first order Markov model. If the last two words were considered in order to predict the next word, then it was called trigram word model or second order Markov model [39].

2.3.2 Knowledge Based Modeling

The systems that only dedicated to use statistical modeling for prediction often present words that are syntactically, semantically, or pragmatically inappropriate. Then they enforce a heavy cognition load on the user to choose the proposed word and decrease the writing speed as a result [50, 51]. If the system minimizes improper words from the prediction list, it will provide more comfort and confidence to the user. The linguistic knowledge that could be used in prediction systems is syntactic, semantic, and pragmatic.

2.3.2.1 Syntactic Prediction

Syntactic prediction is a method that tries to present words that are appropriate syntactically in a particular position within the sentence[25]. This means that knowledge from the syntactic structure of the language is used.A method that tries to present words appropriate statically that position of the sentence is called syntactic prediction. In syntactic prediction, part-of-speech (POS) tags of all words are identified in a corpus and the system has to incorporate the syntactic knowledge for prediction. Statistical syntax and rule based grammar are two general syntactic prediction methods that will be presented in more detail [39].This method includes various types of probabilistic and parsing methods such as Markov model and artificial neural network**.**

Statistical Syntax: The sequence of syntactic categories and POS tags are used for predictions in this approach. In this method, the appearance of a word is based upon the correct usage of syntactic categories. It means the Markov assumption about n-gram word tags is used. In the simplest method, the POS tags are adequate for prediction. Therefore, a probability would be allocated to each candidate word by guessing the probability of having this word with its tag in the current position and about the most probable tags for the previous word(s) [39, 42].

Statistical Syntax uses the sequence of syntactic categories and POS tags for predictions. The appearance of a word in this method is based upon the correct usage of syntactic categories. In other words, the Markov assumption about n-gram word tags is used. The most frequent tag for a particular word is used when producing surface words. Bi-gram and tri-gram probabilities for the next tags are computed using (Eq.5) and (Eq.6) respectively [25].

$$P(t_n/t_{n-1}) = \frac{P(t_{n-1}, t_n)}{P(t_{n-1})} \qquad (5)$$

Where t_{n-1}, t_n are tag a given corpus, P (t_{n-1},t_n) is a probability of tag t_{i-1} and t_i sequence in a given corpus, P (t_{n-1}) is probability of tag t_n, P (t_n/t_{n-1}) is probability of tag t_n after tag t_{n-1}.

$$P(t_n/t_{n-2}, t_{n-1}) = \frac{P(t_{n-2}, t_{n-1}, t_n)}{P(t_{n-2}, t_{n-1})} \qquad (6)$$

Where t_{n-2}, t_{n-1},and t_n are tags in a given corpus, P (t_{n-2},t_{n-1},t_n) is a probability of tag t_{n-2},$t_{n-1}$and t_n sequence in a given corpus, P (t_{n-2},t_{n-1}) is a probability of tag $t_{n-1}$and t_n sequence, P (t_n/t_{n-2},t_{n-1}) is probability of tag t_n after tag t_{n-2} and tag t_{n-1}.

In another approach, the system attempts to estimate the probability of each candidate word according to the previous word and its POS tag, and the POS tag of its preceding word(s).In addition, the system uses word bigram and POS trigram model [39].

A linear combination model of POS tags tries to estimate the probability of POS tag for the current position according to the two previous POS tags. Then, it attempts to find words that have the highest probability of being in the current position according to the predicted POS tag. Then, it combines this probability with the probability of the word given the previous word. So, there are two predictors in which one predicts the current tag according to the two POS tags and the one that uses bigram probability to find the most likely word [39, 43]

Rule-based Grammar: Rule based grammar is another approach in which syntactic word prediction would be made by using the grammatical rules of the language. The current sentence will be parsed by using the grammar of the language to reach to its categories. The parsing method can be either top-down or bottom-up [51]. Context Free Grammar (CFG), Phrase Structure Rule Grammar (PSRG), and Head-driven Phrase Structure Grammar (HPSG) are the methods that could be used in prediction systems based on grammatical rules.

2.3.2.2 Semantic Prediction

Some of the predicted items in the prediction list, there is a possibility to be wrong semantically even though they are syntactically right. If suggesting the words are syntactically and semantically correct, they will increase the accuracy of the predictions [42, 51]. To achieve the goal, a great semantic knowledge is tagged to the words and phrases in a corpus. Frequently in semantic prediction presence of specific word with exceptional content is a clue to increase the probability of appearing other words that have semantic relationships to that word [39].

Semantic prediction used two methods. One of these methods is lexical source like WordNet in English which measures the semantic probability of words to get assured that the predicted words are semantically related in that context. The other method is lexical chain that assigns the highest priority to the words which are related semantically in that context; the unrelated words to that context would be removed from the prediction list.

2.3.2.3 Pragmatics Prediction

The predictions can be affected by pragmatics. Adding the method to the prediction procedure attempts to filter the words that are possibly correct syntactically and semantically, but wrong according to discourse. The pragmatic knowledge is also tagged to the words in a corpus. Suggesting the words that are correct pragmatically would increase the accuracy of predictions as well [42, 53].

2.3.3 Heuristic Modeling

Predictions become more appropriate for specific users when the adaptation methods are used. In this approach, the system adapts every individual user [41, 42 and 43]. Short term learning and long-term learning are the two general methods that make the system adapted to the users. They will be presented in this section.

2.3.3.1 Short-term Learning

In this approach, the system adapts to the user on a current text that is going to be typed by an individual user. Recency promotion, topic guidance, trigger and target, and n-gram cache are the methods that a system could use to adapt itself to a user in a single text. The methods are commonly used in prediction systems [54].

Recency Promotion: The term "recency" has come from cognitive psychology. This concept means a word that has already occurred in a text will be given a higher probability of use; thus, more likely to be used in that text again. Such a method usually assigns dynamically higher probabilities to the words that recently are used in the text; so, it does not only take into account what words have been typed; but further, how recent they have been used [54].

Topic Guidance: This approach is a way of adapting the predictor to the overall subject of the current text. To do so, the general lexicon is complemented with a domain specific lexicon that contains words which are frequently occurring within certain domains, though not very common in general [39, 43].

Trigger and Target: In this method, the appearance of a word is highly correlated with other word sequences. It means when the word A, the trigger, occurs in the text, it triggers

the word B, the target. Then, it causes the target word's probability estimation to change [39, 43].

N-gram Cache: It is assumed that if a word is used once, it is more likely to be used again. In other words, the previous use of a word in a context increases the probability of that word to be used again. Using n-gram cache is a way to capture the most common words and sequences that are frequently used. These words would be put in the cache to get an increased probability [39, 50].

2.3.3.2 Long-term Learning

In this method, the system gets adapted to the user by considering not only the current text, but previous texts that are produced by the user. As a result, gradually by using the system more, it adapts to the user heuristically [43]. Some of the methods for heuristics adaptations that are language specific are adding new words, automatic capitalization, providing inflected form of words, and compounding.

Adding New Words: Heuristic adaptation may involve adding new words to the lexicon of the system whenever the user types unknown words to the system. The added new words could be called in the prediction list for future use [54].

Automatic Capitalization: Depending on the language that the system is running for, some letters should be capitalized. For example, the first letter of a word at the beginning of a sentence and also proper words must be capitalized. Automatic capitalization allows the user to save more keystrokes [54].

Providing Inflected Form of Words: For some languages which are very inflected such as German, French etc, the prediction system would be more efficient to the user if the system takes the inflected forms of the words into account. The result is having higher percentage of KSS [55].

Compounds: Compounding is a method to make new words from other words. Compound words are written as a single unit. Compounds are numerous in languages like German, French etc. Adding such a method to the prediction systems allows the user to write compound words more easily with higher KSS [55].

2.4 Evaluation Techniques for Word Prediction

There are four standard performance metrics to evaluate word prediction system. Those are keystrokes saving (KSS), Hit rate (HR), Keystrokes until completion and Accuracy (Acc).

The KSS is referred to the percentage of keystrokes that the user saves by using the word prediction system and is calculated by comparing two kinds of measures: the total number of key strokes(KT) needed to type the text without the help of the word prediction and the effective number of keystrokes (KE) saved using word prediction. Hence,

$$KSS = \frac{KT-KE}{KT} * 100 \text{----------(1) [26]}$$

Therefore, the number of keystrokes to type texts taken from the test data with and without sequence word prediction program will be counted to calculate keystroke savings accordingly. The obtained KSS will be compared for word based and POS based models. A higher value for keystrokes implies better performance [26].

Hit rate (HR): The percentage of times that the intended word appears in the suggestion list. A higher hit rate implies a better performance [62].

Keystrokes until completion (KuC): The average number of keystrokes that the user enters for each word, before it appears in the suggestion list. The lower the value of this measure the better the algorithm [62].

Accuracy (Acc): The percentage of words that have been successfully completed by the program before the user reached the end of the word. A good completion program is one that successfully completes words in the early stages of typing [62].

2.5 Summary

In this chapter, we have reviewed linguistic characteristics of Afaan Oromo like part -of -speech, morphology and grammar. We understand that Afaan Oromo nouns are inflated for number, gender and case, verbs are inflated for number, gender, tense voice and aspect and adjectives are inflated for number and gender. In addition to this, we have discussed different word prediction approaches like statistical, knowledge based, and heuristics with strength and weakness of one over another. The hybrid of statistical and knowledge based modeling are adopted.

CHAPTER 3: RELATED WORK

3.1 Introduction

This chapter is dedicated to present related work on word or text prediction. The approaches used and the result obtained are included in it. Word prediction for Amharic, Russian, English, Persian, and Hebrew language are some of research conducted in the area that we exhaustively reviewed for this work in order to understand and identify appropriate approaches for Afaan Oromo.

3.2 Word prediction for Amharic language

Alemebante Mulu and Goyal [19] performed a research on Amharic Text Prediction System for Mobile Phone. In this work, they have designed text prediction model for Amharic language: a corpus of 1,193,719 Amharic words, 242,383 Amharic lexicons and a list of names of persons and places with a total size of 20,170 has been used. To show the validity of the word prediction model and the algorithm designed, a prototype is developed. Amharic text prediction system describes the data entry techniques that are used to enter data into mobile devices, such as a smartphone. Data entry could be either predictive or non-predictive in which the first two characters is written and listed down all predicted word, based on the frequency of the word as well as going the alphabetical order if the frequency is the same. The experiment is tested by a database or lexicon of Alembante Mulu also conducted to measure the accuracy of the Amharic text prediction engine and finally the prediction accuracy achieved 91.79%.

Tigist Tensou [20] performed a research on word sequence prediction for Amharic. In this work, Amharic word sequence prediction model is developed using statistical methods and linguistic rules. Statistical models are constructed for root/stem, and morphological

properties of words like aspect, voice, tense, and affixes are modeled using the training corpus. Consequently, morphological features like gender, number, and person are captured from a user‘s input to ensure grammatical agreements among words. Initially, root or stem words are suggested using root or stem statistical models. Then, morphological features for the suggested root/ stem words are predicted using voice, tense, aspect, affixes statistical information and grammatical agreement rules of the language. Predicting morphological features is essential in Amharic because of its high morphological complexity, and this approach is not required in less inflected languages since there is a possibility of storing all word forms in a dictionary. Finally, surface words are generated based on the proposed root or stem words and morphological features. Word sequence prediction using a hybrid of bi-gram and tri-gram model offers better keystroke savings in all scenarios for their experiment. For instance, when using test data disjoint from the training corpus, 20.5%, 17.4% and 13.1% keystroke savings are obtained in hybrid, tri-gram and bi-gram models respectively. Evaluation of the model is performed using developed prototype and keystroke savings (KSS) as a metrics. According to their experiment, prediction result using a hybrid of bi-gram and tri-gram model has higher KSS and it is better compared to bi-gram and tri-gram models. Therefore, statistical and linguistic rules have quite good potential on word sequence prediction for Amharic language.

Nesredin Suleiman [2] performed a research on word prediction model for Amharic online hand writing recognition. In this work, he designs the model using: a corpus of 131,399 Amharic words is prepared to extract statistical information that is used to determine the value of N for the N-gram model, where the value two (2) is considered as a result of the analyses made a combination of an Amharic dictionary (lexicon) and a list of names of persons and places with a total size of 17,137 has been used. To show the validity of the word prediction model and the algorithm designed, a prototype is developed. Experiment is also conducted to measure the accuracy of the word prediction engine and a prediction accuracy of 81.39% is achieved. Analyses have been done on the corpus prepared. These

analyses are used to get information like the average word-length of Amharic language; the most frequently used Amharic word-length and the like. These information have been used to decide the core element of word prediction engine which is N for N-gram model, where N is the number of characters after which the prediction process starts. Based on the analyses done, the value of N has been decided to be two (N=2).

3.3 Word Prediction for English

Antal van den Bosch [58] proposed classification-based word prediction model based on IGTREE. A decision-tree induction algorithm has been favorable scaling abilities. Token prediction accuracy, token prediction speed, number of nodes and discrete perplexity are evaluation metrics used for this work. Through a first series of experiments they demonstrate that the system exhibits log-linear increases in prediction accuracy and decreases in discrete perplexity, a new evaluation metric, with increasing numbers of training examples. The induced trees grow linearly with the amount of training examples. Trained on 30 million words of newswire text, prediction accuracies reach 42.2% on the same type of text. In a second series of experiments we show that this generic approach to word prediction can be specialized to confusible prediction, yielding high accuracies on nine example confusible sets in all genres of text. The confusible-specific approach outperforms the generic word-prediction approach, but with more data the difference decreases.

Agarwal and Arora [59] proposed a Context Based Word Prediction system for SMS messaging in which context is used to predict the most appropriate word for a given code. The growth of wireless technology has provided alternative ways of communication such as Short Message service (SMS) and with tremendous increase in mobile Text Messaging, there is a need for an efficient text input system. With limited keys on the mobile phone, multiple letters are mapped to same number (8 keys, 2 to 9, for 26 alphabets). The many to one mapping of alphabets to numbers gives us same numeric code for multiple words.T-9

system predicts the correct word for a given numeric code based on frequency. This may not give us the correct result most of the time. For example, for code ‘63’,two possible words are ‘me’ and ‘of’. Based on a frequency list where ‘of’ is more likely than ‘me’, T-9 system will always predict ‘of’ for code ‘63’. So, for a sentence like ‘Give me a box of chocolate’, the prediction would be ‘Give of a box of chocolate’. The sentence itself indeed gives us information about what should be the correct word for a given code. Consider the above sentence with blanks, “Give _ a box _ chocolate”. The current systems for word prediction in Text Messaging predict the word for a code based on its frequency obtained from a huge corpus. However, the word at a particular position in a sentence depends on its context and this intuition motivated them to use Machine Learning algorithms to predict a word, based on its context. The system also takes into consideration the proper English words for the codes corresponding to the words in informal language. The proposed method uses machine learning algorithms to predict the current word given its code and previous word’s part of speech (POS). The training was done on about 19,000 emails and the testing was done on about 1,900 emails, with each email consisting of 300 words on average. The results show 31 % good improvement over the traditional frequency based word estimation.

Trnka[60] conducted a research on topic Adaptive Language Modeling for Word. AAC devices are highly specialized keyboards with speech synthesis, typically providing single-button input for common words or phrases, but requiring a user to type letter-by-letter for other words, called fringe vocabulary. Word prediction helps speed AAC communication rate. The previous research conducted by different scholars using ngram models. At best, modern devices utilize a trigram model and very basic recency promotion. However, one of the lamented weaknesses of ngram models is their sensitivity to the training data. The objective of this work is to develop and integrate style adaptations from the experience of topic models to dynamically adapt to both topically and stylistically. They address the problem of balancing training size and similarity by dynamically adapting the language model to the most topically relevant portions of then training data. They present the results

of experimenting with different topic segmentations and relevance scores in order to tune existing methods to topic modeling. The inclusion of all the training data as well as the usage of frequencies addresses the problem of sparse data in an adaptive model. They have demonstrated that topic modeling can significantly increase keystroke savings for traditional testing as well as testing on text from other domains. They have also addressed the problem of annotated topics through fine-grained modeling and found that it is also a significant improvement over a baseline ngram model.

3.4 Word Prediction for Persian Language

Masood Ghayoomi and Seyyed Mostafa Assi[61]studied word prediction for Persian language. In this study, they designed and developed based a system on a Statistical Language Modeling. The corpus contained approximately 8 million tokens. The corpus is divided in to three sections: one was the training corpus that contained 6,258,000 tokens, and 72,494 tokens; the other section was used as the developing corpus which contained 872,450 tokens, and the last section was used as the test corpus which contained 11,960 tokens. The user enters each letters of the required word; the system displays a list of the most probable words that could appear in that position. Three standard performance metrics were used to evaluate the system including keystroke saving, the most important one. The system achieved 57.57% saving in keystrokes. Using such a system saved a great number of keystrokes; and it led to reduction of user's effort.

Ghayoomi and Daroodi [26] studied word prediction for Persian language in three approaches. Persian is a member of the Indo-European language family and has many features in common with them in terms of morphology, syntax, phonology, and lexicon. This work is based on bi-gram, tri-gram, 4-gram models and it utilized around 10 million tokens in the collected corpus. Using Keystroke Saving (KSS) as the most important metrics to evaluate systems' performance, the primary word-based statistical system achieved 37% KSS, and the second system that used only the main syntactic categories with word-statistics

achieved 38.95% KSS. Their last system which used all of the available information to the words get the best result by 42.45% KSS.

3.5 Word Prediction for Russian Language

Hunnicutt et al. [56] performed a research on Russian word prediction with morphological support as a co-operative project between two research groups in Tbilisi and Stockholm. This work is an extension of a word predictor developed by Swedish partner for other languages in order to make it suitable for Russian language. Inclusion of morphological component is found necessary since Russian language is much richer in morphological forms. In order to develop Russian language database, an extensive text corpora containing 2.3 million tokens is collected. It provides inflectional categories and resulting inflections for verbs, nouns and adjectives. With this, the correct word forms can be presented in a consistent manner, which allows a user to easily choose the desired word form. The researchers introduced special operations for constructing word forms from a word‘s morphological components. Verbs are the most complex word class and algorithm for expanding root form of verbs to which their inflectional form is done. This system suggests successful completion of verbs with the remaining inflect able words.

3.6 Word Prediction for Hebrew Language

Netzer et al. [57] are probably the first to present results of experiments in word prediction for Hebrew. They developed a NLP-based system for Augmentative and Alternative Communication (AAC) in Hebrew. They used three general kinds of methods: (1) Statistical methods: based on word frequencies and repetition of previous words in the text. These methods can be implemented by using language models (LMs) such as the Markov model, and unigram/bigram/trigram prediction, (2) Syntactic knowledge: part of speech tags (e.g. nouns, adjectives, verbs, and adverbs) and phrase structures. Syntactic knowledge can be

statistical-based or can be based on hand-coded rules and (3) Semantic knowledge: assigning categories to words and finding a set of rules that constrain the possible candidates for the next word. They used 3 corpuses of varying length (1M words, 10M words, 27M words) to train their system. The best results have been achieved while training a language model (a hidden Markov model) on the 27M corpus. They applied their model on various genres including personal writing in blogs and in open forums in the Internet. Contrary to what they expected, the use of morpho-syntactic information such as part of speech tags didn't improve the results. Furthermore, it decreases the prediction results. The best results were obtained using statistical data on the Hebrew language with rich morphology. They report on keystroke saving up to 29% with nine word proposals and 34% for seven proposals, 54% for a single proposal.

3.7 Summary

In this section, we have discussed works related to word sequence prediction for different languages. We understand that languages have their own linguistic characteristics requiring specific approaches to word prediction. Hence, the research conducted on one language cannot be directly applied to other languages. The hybrid of statistical and knowledge based modeling are adopted to the language. Therefore, the aim of this study to design and develop word sequence prediction model for Afaan Oromo by taking the unique features of the language into consideration.

CHAPTER 4: AFAANOROMO WORD SEQUENCE PREDICTION SYSTEM

4.1 Introduction

This Chapter presents details of the Afaan Oromo word sequence prediction system Architecture of the proposed word sequence prediction system. The system architecture and algorithms design consider characteristics of the language: especially morphology of nouns, verbs and adjectives, grammatical structures like subject verb agreement, noun adjective agreement, object verb agreement.

4.2 System Architecture

The architecture has two major parts these are language modeling and generation of predicted words. The system tokenizes and morphologically analyzes training corpus and user input. After that, from morphologically analyzed corpus language model is constructed like root/stem sequence and root/stem with tag. Finally, word sequence prediction component list the most likely probable root/ stem. The bi and tri- gram word statistics and bi and tri gram POs tag statistics are used to get the most expected root or stem, and inflated words and then morphological generator component synthesize the surface word form from root /stem word. Figure 4.1 shows the architecture of the system.

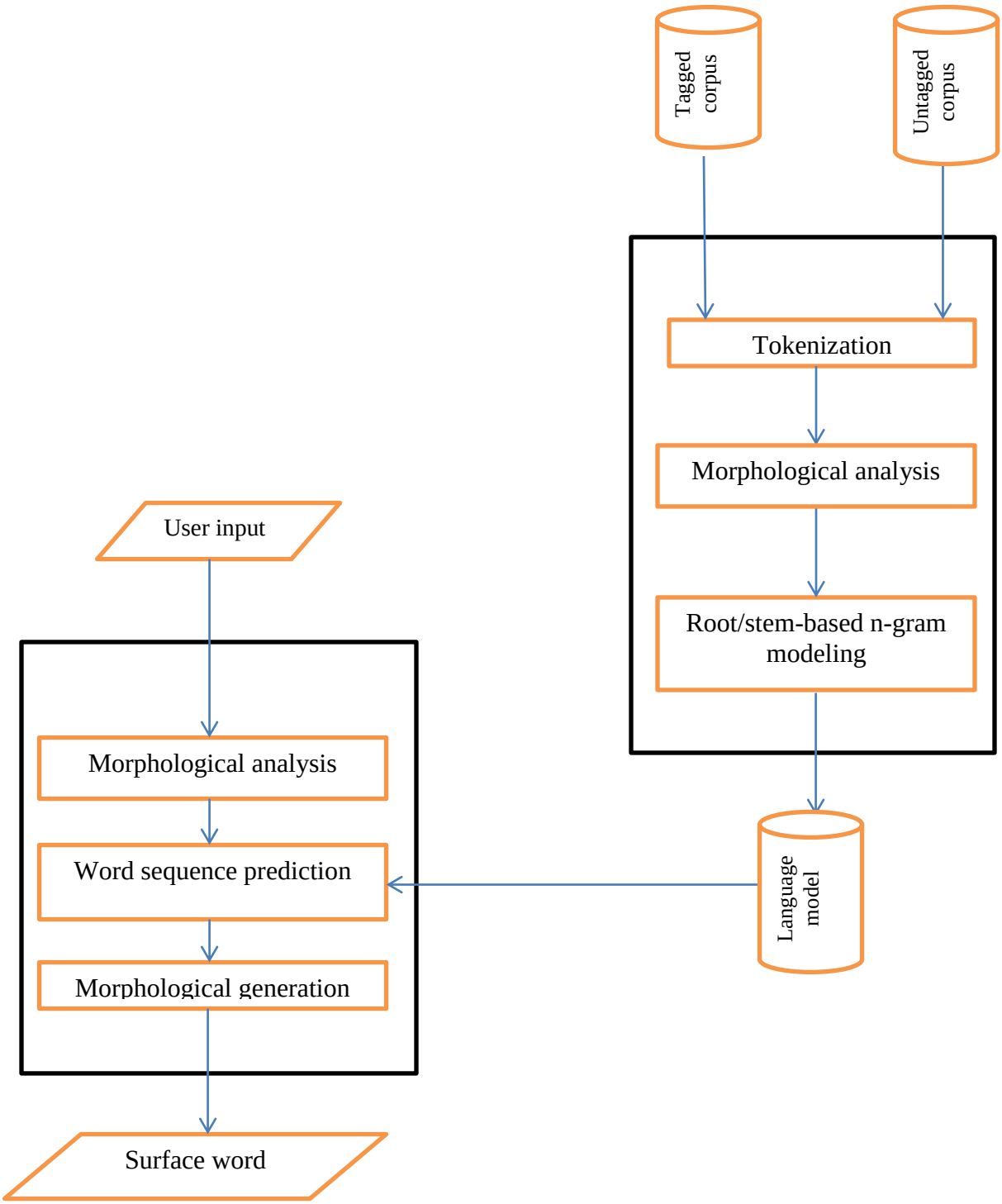

Figure 4.1: The architecture of Afaan Oromo word sequence prediction system

4.3 Language Modeling

Language model is a storage consisting of statistical information which serves as a knowledge base when predicting suitable words.

4.3.1 Tokenization

This component accept block of text and then split it into single words, numbers, symbols and punctuation marks. In Afaan Oromo, like any other language that used Latin script, the blank space, parenthesis, brackets, quotes, etc are used to show the end of one word. Besides, sentence boundaries and use of punctuations marks are almost similar to English language (i.e. a sentence may end with a period (.), line break, a question mark (?), or an exclamation point). Thus, space and punctuation marks are used as the explicit delimiters or token separator. When a space is come across, the word after the space becomes a token. The output of this component is used as input for morphological analysis.

4.3.2 Morphological analysis

Morphological analysis is the process of assigning each word found in a corpus to their morphemes which can be affix, root, stem, etc. It is useful to annotate words to their root form and other required morphological information.

The task of this module is to accept a list of words and then decompose each word into root/stem and affixes. The extracted root/stem words are used to build language model like root/stem word sequence and root/stem word with tag. As we have seen in previous chapters, Afaan Oromo has a very complex morphology. It is ideal to store all forms of words in probabilistic models.

Morphological analyzer is a program used to analyze a separate word or words in a file to their component forms. For this study morphological analyzer is prepared manually because preparing automates morphological analyzer beyond the scope of this paper. Sample of morphological analysis and s data used for training is presented in appendix 1.

4.3.3 Root/stem based n-gram modeling

The language model of word sequence prediction is based on two Markov models; one for root/stem form of words is suggested and the other for proposed root or stem tags. The two models interact, but they are separate.

The idea of this scheme is first obtain probability estimation for the tag of the next word, using the tag Markov model, and then use the word Markov model to get probability estimation for the next word. In the second step, the tag probability estimation is taken into account in order to promote words with a likely tag according to the tag Markov model.

Based on related works experience and characteristics of n-gram models, we have chosen to employ the first and second order Markov model for both tags and words, simply because they require a reasonable amount of implementation work and storage space.

Words Sequence modeling

In this component, n-length root/stem sequence is extracted from training corpus. Bi-gram, tri-gram and hybrid of bi-gram and tri-gram statistical models are constructed for root/ stem sequence. Each n-gram model is separately kept in its own repository and they hold root or stem word sequences for each value of n with their probability of occurrence in the corpus. Probabilities of all unique root or stem sequences with this respective value of n is calculated by counting occurrence of n word sequences and n-1 word sequences in the

corpus gram models, and then calculating their ratio tri-gram probabilities are computed using (Eq.4).

For example: Probability of a word given previous two words, where n=3(tri-gram) is calculated as shown below:

$$P(Kuma/mallaqa\ qarshii) = \frac{C(mallaqa\ qarshii\ kuma)}{C(mallaqa\ qarshii)}$$

$$C(mallaqa\ qarshii\ kuma) = 127$$

$$C(mallaqa\ qarshii) = 161$$

$$P(Kuma/mallaqa\ qarshii) = \frac{C(mallaqa\ qarshii\ kuma)}{C(mallaqa\ qarshii)} = \frac{127}{161} = .7789.$$

Algorithm 4.1 shows the algorithm to construct n-gram root or stem words probabilistic model.

```
BEGIN
INPUT   List of root or stem word  text file  from
        morphological analyzer and a length-n//length-n number of
        words in sequence
OUTPUT  root or stem  word sequence n-gram  probability
        model
        // word sequence means root or stem word sequence
        FOR EACH sentence in text file
                  EXTRACT  length-n  word sequence and STORE
        in buffer
        INITIALIZE a dictionary  to have zero count
        FOR EACH word sequence :
              INCREMENT  the  sequence's count in dictionary
        CALCULATE the probability  length-n word sequence
        and store in buffer with their corresponding word
        sequence
END
```

Algorithm 4.1: Root / stem sequence

Root/stem tagging

The task of this is component accept root/stem from morphological analyzer and count a unique occurrence of root/stem with their corresponding POS on training corpus. It stores the frequency of each root/stem with their respective POS. The most frequent POS for a particular root or stem is used when suggesting most probable features. Algorithm 4.2 show root/stem tagging.

```
BEGIN
INPUT   Tagged training Corpus
OUTPUT  Root/stem with tag n-gram probability model
        FOR EACH word in Tagged training Corpus
                    EXTRACT root/stem with tag and Store in
        buffer
        INITIALIZE  a dictionary to have zero count
        FOR EACH  root/stem with tag
                    INCREMENT  the root/stem  and tag
                     Sequence's count dictionary
        CALCULATE   the   probability   root/stem   with   tag
        sequence and store in buffer corresponding root/stem
        with tag sequence
END
```

Algorithm 4.2: Root/stem tagging

4.4 Generation of Predicted Words

4.4.1 Morphological Analysis

This component analyzes Afaan Oromo -n previous words accepted text from a user and extracts required morphological features like root/stem and affix. Context information like gender, number, person definiteness for nouns in addition to that tense for verbs is captured from a user‘s input to in order to predict appropriate morphological features for the coming root /stem word. When a user enters a text, the system identifies the last –n words and morphologically analyzes each word found in it. As we stated in previous section, we used manual morphological analysis.

4.4.2 Word Sequence Prediction

Word Sequence Prediction can be based on either text statistics or linguistic rules. Its component tries to suggest list of probable words to the user. Two Markov models can be included: one for word classes (POS tag unigrams, bigrams and trigrams) and one for words (word unigrams and bigrams). Bi-gram model predicts root or stem word based on previous single word from current position, whereas tri-gram model predicts root or stem word based on preceding two words. Hybrid model is a linear combination of those models. In other word, it predict based on words class and words in combine.

Suppose the user is typing a sentence and the following sequence has been entered so far from left to right according to the Afaan Oromo writing system:

W_{i-2} $W_{i-1}W_i$

Where W_{i-2}, $W_{i-1}$and W_i are the most lately completed words, and Wi is the current word that is going to be predicted. Let W be the set of all words in the lexicon that would likely

appear in that position. Our statistical prediction algorithm first attempts to estimate the probability of each candidate word's POS, (t W_i), according to the previous tags (t W_{i-2}) and (t W_{i-1}). Then, it tries to estimate the probability of the candidate root or stem word in the current position, (W_i), according to the previous root or stem words (W_{i-2}) and (W_{i-1}); i.e., P(w_i, tW_i | W_{i-2}, tW_{i-2}, W_{i-1}, tW_{i-1}).Then the algorithm selects the 15 most appropriate words from W that are likely to be the user's intended root or stem words. The general approach is to estimate the probability of each candidate word, Wi W, being the user's required word in that context based on the POS tags of the preceding words. The algorism 4.3 shows word sequence prediction.

```
BEGIN
INPUT   Root/stem word file and user input
OUTPUT  Proposed root/stem list
        Extract the last length –n root/stem sequence from
        user input and store in buffer
        FOR EACH sentence in root/stem file
           IF the last length –n root/stem sequence
           from user in sentence and the next word exist
            EXTRACT length –n+1 root/stem sequence
            from sentence and STORE in buffer
        READ root/stem with tag probability model
        READ root/stem sequence probability model
        For EACH root/stem sequence:
             EXTRACT last word
             IF the number of proposed root/stem
             list<15
               ADD the last n=word to the proposed root/stem
                     list
END
```

Algorithm 4.3: Word sequence prediction

4.4.3 Morphological Generation

The morphological generation mainly deals with the concatenation of corresponding suffixes with the root word to form a word of specific grammatical category. The input of the morphological generator would be the root word which then inflects this word to the morphology of the respective. In other word, it is to produce the inflected form of a word according to the features and values. From practical point of view, morphological generation is the inverse process of analysis, namely the process of converting the internal representation of a word to its surface form. The same rule definitions can be used to generate the desired word form as used for analysis. The only difference will be the direction of execution order of the elements in the rule definition.

This component accept root or stem from Word Sequence Prediction component and affix and then list of proposed surface word. Surface word is a morphologically suitable word that the user intends to type.

For this study, we used manual morphological generator because automatic morphology is beyond the scope.

CHAPTER 5: EXPERIMENTATION

5.1 Introduction

Developing a prototype is one of the objectives of this work. It is used to demonstrate the validity and usability of the proposed word sequence prediction system. For bi-gram, tri-gram, hybrid of bi-gram and tri-gram models word and bi-gram, tri-gram, hybrid of bi-gram and tri-gram models POS, prototypes are designed and developed in order to demonstrate as well as evaluate the developed Afaan Oromo Word Sequence Prediction Model.

This Chapter dedicated to present about the tools and development environments used to implement the prediction algorithm, corpus collection, implementation test result and discussion.

5.2 Corpus Collection

To accomplish a task of Afaan Oromo word sequence prediction, it requires statistical information such as the frequency of occurrence of root or stem and corresponding **POS** tag. This can be achieved by using a corpus. Since the Afaan Oromo word corpora are not available easily, we prepared the corpora from various sources that include newspapers(Bariisaa, Bakkalcha Oromiyaa and Oromiyaa), journals, criminal code ,books, social media like Facebook, webpages, books which are written by different authors on different issues such as politics, religion, history, fiction and love. Around 50 different files, collected from various sources mentioned above, are provided to the tool and corpus consists of 23,400 sentences and a total of 312,208 words are generated filter 49,143 unique words. All the 50 files are converted to txt format in order to be used by the visual studio C# tool. All the collected words which we refer as: “Ashenafi bekele Word Sequence Prediction Corpus” that is used in this work is, thus, composed of 49,143 unique words. The size of the corpus affects the response time. The testing is done using Afaan Oromo BBC news text

having a total of 200 sentences. In order to implement the proposed model, we need Afaan Oromo POS tagged corpus for testing but there is no any available POS tagger. Due to the fact that POS tag was prepared manually with the help of [16] and linguistic expert. Totally, 200 tagged sentences were used for testing in this study. However, we believe the sentences used are representative. Sample of training data used for training is presented in appendix 2.

5.3 Implementation

We used different tools and developing environments in order to implement the algorithms and to do necessary experiment on the system. The prototype of Afaan Oromo word sequence prediction is developed using Microsoft visual 2015 C#. The main aim of the prototype of the system is to demonstrate and test the developed word sequence prediction model.

Our system generates word sequence predictions in a manner that closely matches Afaan Oromo typing systems. User type text in text box (input) and when space bar or one of delimiters is pressed, the system predicts 15 likely single next words and shows them in equal-sized suggestion list box, with the most likely suggestion in the top of the list. Next, a user clicks his or her preferred word from a given list of word options instead of typing each character and then click add button. However, if the required word is not listed in a given option, then a user continues typing in usual writing method. Figure 5.1 shows User Interface of Word Sequence Prediction.

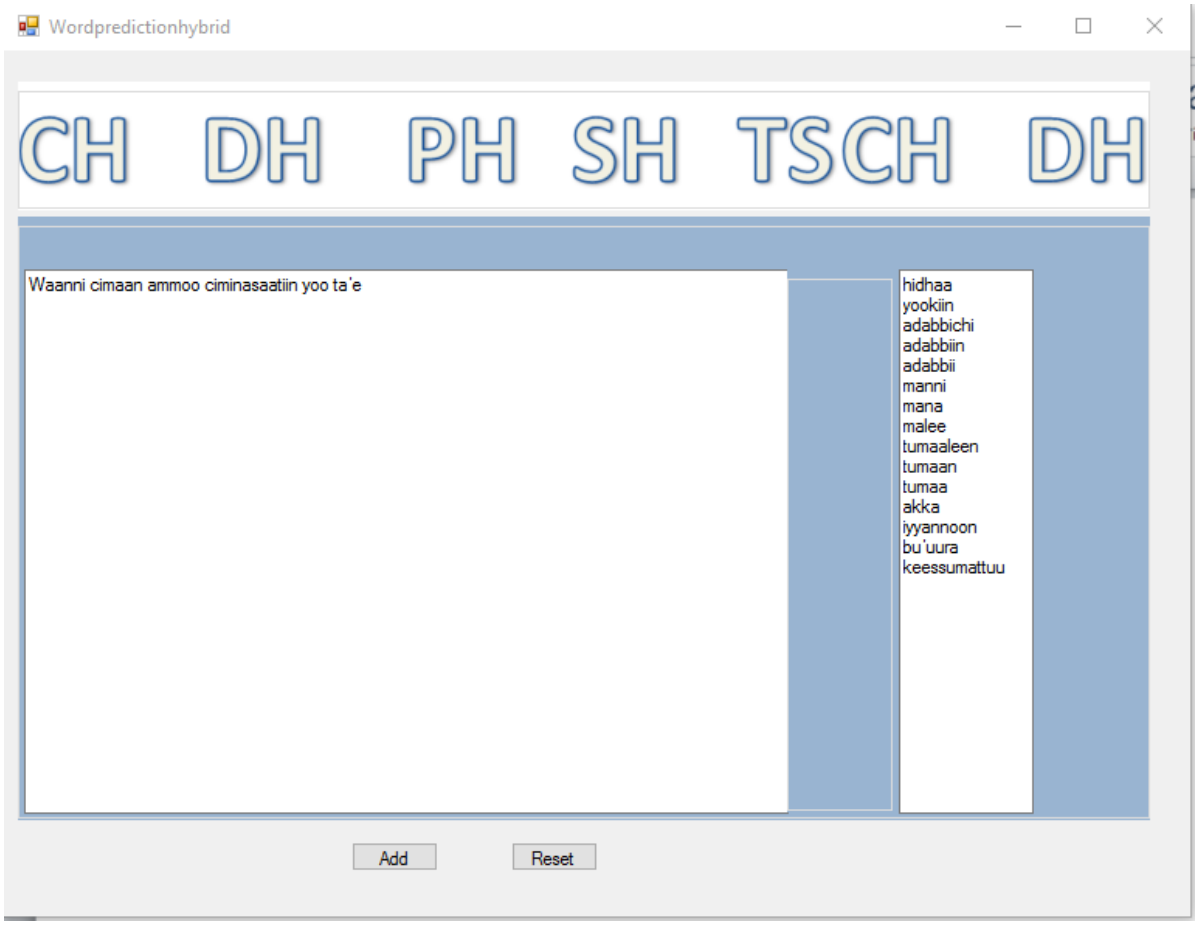

Figure 5-1: User Interface of Word Sequence Prediction

5.4 Test Results

After the experiment of the system, the result of the experiment is mentioned in this sub section, and it is explained how the result of the experiment was interpreted.

We developed and tested our system in two different scenarios. The first test used only bi-, tri- and hybrid of bi-and tri word models; we called it case one. The second test was tested using the described n-gram word models along with the words' POS bi-, tri- and hybrid of bi- and tri of the main syntactic categories; we called it case two.

Keystroke saving (KSS) is one of standard performance metrics to evaluate performance of word prediction. As described in Chapter 2, performance evaluation metrics assess the accuracy of the system prediction based on the number of keystrokes saved. The experiment conducted in this research exhibits obtained results based on keystroke savings (KSS) and using bi-gram, tri-gram, and hybrid models. Keystroke Saving (KSS) is referred to the percentage of keystrokes that the user saves by using the word prediction system. Which means estimates saved effort percentage which is calculated based on (Eq.1) by comparing total number of keystrokes needed to type a text (KT) and effective number of keystrokes using word prediction (KE). A higher value for keystroke saving implies a better performance. Table 5.1 shows summary of the test results

Table 5.1**:** Summary of the test results

Testing data	Model	**KT**	**KE**	**KSS**
case one	Bi-gram	1700	1504	13%
	tri-gram	1700	1446	17.6%
	Hybrid model	1700	1360	20.5%
case two	Bi-gram	1700	1504	13.1%
	tri-gram	1700	1422	19.5%
	Hybrid model	1700	1388	22.4%

5.5 Discussion

The result of the experiment is illustrated on table 5.1. When Testing is done using case one, 20.5%, 17.4% and 13.1% keystroke savings are obtained in hybrid, tri-gram and bi-gram models respectively. Hybrid of bi-gram and tri-gram is the highest in case one. In case two, 22.4%, 19.4% and 13.1% keystroke savings are obtained in hybrid, tri-gram and bi-gram models respectively. Word sequence prediction using a hybrid of bi-gram and tri-gram case one provides higher than using a hybrid of bi-gram and tri-gram System one keystroke savings. Hybrid of bi-gram and tri-gram in case two is the highest.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis work discusses the design and implementation of a sequence word prediction for Afaan Oromo using the bi and tri-word statistics, and the bi-, and tri- POS tag statistics of the language. The work also compares a system that solely uses word statistics with the designed systems that use word statistics as well as POS tags information. Word sequence prediction is one of text entry systems ,it help user to write on text input area especially for a person who has writing disability .As we have stated in Chapter 3 a number of researches have been conducted on various languages. Even if there are different researches in Afaan Oromo, there is no work on the topic of word sequence prediction that considers both syntax and word information.

In this study, the system allows user to write a sentence according to Afaan Oromo writing system and the system lists the most likely probable next words to be typed by a user, based on previous word history. It filters words in predictions list that is syntactically inappropriate in a particular position within the sentence. Thus, it would increase the user's confidence to enable him or her to select words from the prediction list that can result in better written sentences, along with imposing a lower cognition load on him or her. This is done using n-gram statistical language based on two Markov language models one for tag or part of speech of words, the other for words which are developed using manually tagged corpus.

The designed model is evaluated based on developed prototype. Keystroke Saving (KSS) is used to evaluate systems performance. According to the evaluation, the primary word-based statistical system achieved 20.5% KSS, and the second system that used syntactic categories with word-statistics achieved 22.5% KSS. Therefore, statistical and linguistic rules have good potential on word sequence prediction for Afaan Oromo.

We hope the application of this study assists physically disabled individuals who have typing difficulties, speed up typing speed by decreasing keystrokes for mobile phone, computer and other hand held device user. This work plays an important role in other NLP applications like spelling checkers, handwriting recognition and word-sense disambiguation.

6.2 Contribution of the Thesis

The contributions of this thesis work are summarized as follows:

- We proposed architecture for Afaan Oromo word sequence prediction system
- We identified hybrid word prediction approach that is suitable for Afaan Oromo word sequence prediction.
- We developed and adopted algorithms for Afaan Oromo word sequence prediction.

6.3 Future work

There are a number of holes for improvement and modification for Word sequence prediction of Afaan Oromo. Below are some of the recommendations we propose for future work.

1. Afaan Oromo is morphologically complex language as we have discussed in Chapter 2. Moreover, there is no morphological analyzer and synthesizer tool that for Afaan Oromo. In this work, we used manual morphological analyzer and synthesizers. Hence, we recommend for development of morphological analyzer and synthesizer tool for the language.

.

2. There is no POS tagger or adequate POS tagged Afaan Oromo corpus makes this works hard to keep morph-syntactic agreement complete .But POS tagged test data is required and used in this study to evaluate the proposed model. However, Afaan Oromo word sequence prediction can be optimized if good Afaan Oromo POS tagger is incorporated and if the model is enriched with POS.

3. Like any other NLP application word sequence prediction demand adequate in quality and quantity of training data. A model which is tested by corpus with low quality or too small in size, it provides wrong word sequence prediction output.
   We have used more than 23 thousand sentence, 60% of the data is collected by from hard copy and the others collected 40% from web site. During scanning or converting hard copy to soft copy some scripts are changed to meaning less the symbols and others changed to other scripts for examples script **m** to **nn** or vice versa. The documents which are collected from web or social media they contain misspelled words and grammatically incorrect sentences. Correcting errors manually is tedious and time consuming preprocessing task. Hence, we recommend that spelling checker and grammar included.

4. In this work, the system is developed based on word frequency and syntax statistics. But it is not necessarily list correct proposal. Therefore, we recommend considering recency of words other syntax and semantic methods along with highest frequency to make more precise feature prediction in future studies of this topic.

5. We used Keystroke saving to evaluate the developed world sequence prediction system in this work. Single evaluation metric may not be sufficient to evaluate a model. Due to this fact, other evaluation metrics can also be used and we suggest considering other evaluation metrics in future studies.

REFERENCES

[1] Barry McCaul and Alistair Sutherland, “Predictive Text Entry in Immersive Environments”, Proceedings of the IEEE Virtual Reality 2004 (VR'04), P: 241, 2004

[2] Nesredin Suleiman, “Word Prediction for Amharic Online Handwriting Recognition”, Unpublished MSc Thesis, Addis Ababa University, 2008.

[3] “19Text-Input”,Icie.cs.byu.edu/UIBook/19-TextInput.pdf,Last Referenced Date: June 4, 2017

[4] Kumiko Tanaka-ishii, “Word-Based Predictive Text Entry Using Adaptive Language Models”, Natural Language Engineering 13 (1): 51–74. 2006 Cambridge University Press, 15 February 2006

[5] Nicola Carmignani, “Predicting words and sentences using statistical models,” Language and Intelligence Reading Group, date: July5, 2006

[6] Garay-Vitoria Nestor and Julio Abascal, “Text Prediction Systems: A Survey”, Universal Access in the Information Society, 4(3): 188-203,2006

[7] Auto Complete [Online Document], Available at: http:// en.wikipedia.org/ wiki/Auto complete Referenced Date: September 15, 2016

[8] Lesher G, Moulton B., and Higginbotham D.,“Effects of N-gram Order and Training Text Size on Word Prediction,” in Proceedings of (RESNA’99) Annual Conference, Arlington, VA, pp. 52-54, 1999.

[9] Even-Zohar Y. and Roth D., “A Classification Approach to Word Prediction”, in Proceedings of The 1st North American Conference on Computational Linguistics (NAACL' 2000), pp. 124-131, 2000.

[10] Koester H. and Levine S., “Modeling the Speed of Text Entry With a Word Prediction Interface ”,IEEE Trans. on Rehabilitation Engineering, vol.2, no. 3, pp. 177-187, September 1994.

[11] Tesfaye Guta Debela,"Afaan Oromo Search Engine", Unpublished MSc Thesis, Addis Ababa University, 2010.

[12] Debela Tesfaye", Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach", unpublished MSc Thesis, Addis Ababa University ,2010

[13] Gaddisa Olani Ganfure and Dida Midekso, "Design And Implementation Of Morphology Based Spell Checker", International Journal Of Scientific & Technology Research, December 2014 pp118-125

[14] Morka Mekonnen,"Text to speech system for Afaan Oromo" ,Unpublished MSc Thesis, Addis Ababa University, 2001.

[15] Diriba Magarsa,"An automatic sentence parser for Oromo language", Unpublished MSc Thesis, Addis Ababa University, 2000.

[16] Assefa W/Mariam, "Developing morphological analysis for Afaan Oromo text", Unpublished MSc Thesis, Addis Ababa University, 2000.

[17] Abraham Tesso Nedjo and Degen Huang, "Automatic Part-of-speech Tagging for Oromo Language Using Maximum Entropy Markov Model (MEMM)", Journal of Information & Computational Science pp. 3319–3334, July 1, 2014

[18] Md. Masudul Haque and Md. Tarek Habib ,"Automated Word Prediction in Bangla Language Using Stochastic Language Models", International Journal in Foundations of Computer Science & Technology (IJFCST) Vol.5, No.6,pp 67-75, November 2015

[19] Alemebante Mulu and Vishal Goyal, "Amharic Text Predict System for Mobile Phone", International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 4, Jul-Aug 2015.

[20] Tigist Tensou, "Word Sequence Prediction for Amharic Language", Unpublished MSc Thesis, Addis Ababa University, 2014.

[21] Johannes Matiasek, Marco Baroni, and HaraldTrost, "FASTYA multi-lingual approach to text prediction", In Computers Helping People with Special Needs, pp. 243-250. Springer Berlin Heidelberg, 2002.

[22] Alice Carlberger, Sheri Hunnicutt, John Carlberger, Gunnar Stromstedt, and Henrik Wachtmeister, “Constructing a database for a new Word Prediction System,” TMH-QPSR 37(2): 101-104, 1996.

[23] Sachin Agarwal and Shilpa Arora, “Context based word prediction for texting language”, In Large Scale Semantic Access to Content (Text, Image, Video, and Sound),pp. 360-368, 2007.

[24] Carlo Aliprandi, Nicola Carmignani, NedjmaDeha, Paolo Mancarella, and Michele Rubino, “Advances in NLP applied to Word Prediction”, University of Pisa, Italy February, 2008.

[25] Aliprandi Carlo, Nicola Carmignani, and Paolo Mancarella, “An Inflected-Sensitive Letter and Word Prediction System”, International Journal of Computing and Information Sciences, 5(2): 79-852007

[26] Masood Ghayoomi and Ehsan Daroodi, “A POS-based word prediction system for the Persian language”, In Advances in Natural Language Processing, pp. 138-147, Springer Berlin Heidelberg, 2008.

[27] G. Q. A. Oromo, “Caasluga Afaan Oromo Jildi I”, Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia, pp. 105-220 (1995).

[28] Keith Trnka, John McCaw, Debra Yarrington and Kathleen F. McCoy, “Word Prediction and Communication Rate In AAC”, IASTED international conference Tele health and assistive technology ,April 16-18 2007 ,Maryland USA

[29] Getachew Mamo Wegari and Million Meshesha, “Parts of Speech Tagging for Afaan Oromo”, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence.

[30] Getachew Emiru, "Development of Part Of Speech Tagger Using Hybrid Approach" Unpublished MSc Thesis, Addis Ababa University ,2016

[31] Mohammed Hussen Abubeker, "Part-Of-Speech Tagging For Afaan Oromo Language Using Transformational Error Driven Learning (Tel) Approach",

Unpublished MSc Thesis, Addis Ababa University, 2010.

[32] Aberra Nefa, “Oromo verb inflection”, Unpublished MA Thesis, Addis Ababa University,2000.

[33] Baye Yimam, “The Phrase Structure of Ethiopian Oromo”, Unpublished Doctoral Thesis, University of London, 1986.

[34] Addunyaa Barkeessaa, “Sanyii Jechaa fi caasaa Isaa (Word and Its structure)”, Alem,2004.

[35] Michael Gasser, Hornmorph User's Guide, 2012.

[36] Wakweya Olani, “Inflectional Morphology in Oromo,”

[37] Debela Tesfaye,”A rule-based Afan Oromo Grammar Checker”, IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 8, 2011.

[38] C. G. Mewis.” A Grammatical sketch of Written Oromo”, Germany: Koln,pp. 25-99 (2001).

[39] Masood Ghayoomi and Saeedeh Momtazi,“An overview on the existing language models for prediction systems as writing assistant tools”, In Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, pp. 5083 5087,IEEE, 2009

[40] Klund, J. and Novak, M. (2001). If word prediction can help, which program doyouchoose?Availableat:http://trace.wisc.edu/docs/wordprediction2001/index.htm?

[41] M. E. J. Woods, “Syntactic Pre-Processing in Single-Word Prediction for Disabled People”, Unpublished Doctoral Thesis. dissertation, University of Bristol, Bristol, 1996

[42] A. Fazly, “The Use of Syntax in Word Completion Utilities,” Unpublished MSc, University of Toronto, Canada, 2002

[43] E. Gustavii and E. Pettersson, “A Swedish Grammar for Word Prediction ”, Unpublished MSc ,Uppsala University, Stockholm, 2003

[44] J. Hasselgren, E. Montnemery, P. Nugues, and M. Svensson, “HSM: Apredictive text entry method using bigrams”, 10th Conference of EACL, In Proceedings of the Workshop on Language Modeling for Text Entry Methods, Budapest, Hungary, pp. 59- 99, 2003

[45] C. L. James, and K. M. Reischel, “Text input for mobile devices: Comparing model prediction to actual performance”, In Proceedings of CHI-2001, ACM, New York, pp. 365-371, 2001

[46] Zi Corporation, eZiText. Technical report, 2002.http://www.zicorp.com

[47] LexicusDivision,iTap.Technicalreport,Motorolla,2002.http://www.motorola.com/lexicu

[48] http://www.eatoni.com

[49] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to

[50] Garay-Vitoria Nestor, and Julio Abascal, “Word prediction for inflected languages”, Application to Basque language, 1997.

[51] R. Rosenfeld, “Adaptive Statistical Language Modeling: A Maximum Entropy Approach”, Unpublished PhD dissertation, Canegie Mellon University, Pittsburgh,1994

[52] S. Hunnicutt and J. Carlberger, “Improving word prediction using markov models and heuristic methods”, Augmentative and Alternative Communication, vol. 17, pp. 255-264, 2001

[53] FerranPla and Antonio Molina, “Natural Language Engineering: Improving part of speech tagging using lexicalized HMMs_ 2004”, Cambridge University Press, United Kingdom, 2000

[54] http://www.gusinc.com/wordprediction.html

[55] S. Hunnicutt and J. Carlberger, “Improving word prediction using markov models and heuristic methods”, Augmentative and Alternative Communication, vol. 17, pp.

255-264, 2001.

[56] Sheri Hunnicutt, Lela Nozadze, and George Chikoidze, “Russian word prediction with morphological support”, In 5th International symposium on language, logic and computation, Tbilisi, Georgia, 2003.

[57] Yael Netzer, Meni Adler, and Micheal Elhadad, “Word Prediction in Hebrew: Preliminary and Surprising Results”, ISAAC, 2008.

[58] Antal van den Bosch, “Scalable classification-based word prediction and confusible correction”, TAL. Volume 46 – n° 2/2005.

[59] Sachin Agarwal and ShilpaArora, “Context based word prediction for texting language,” Conference RIAO, 2007.

[60] Keith Trnka, “Adaptive Language Modeling for Word Prediction,” Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume), pages 61–66, Columbus, June 2008.

[61] Masood Ghayoomi and Seyyed Mostafa ,”Word prediction in Running Text: A Statistical Language Modeling For the Persian Language ”,Proceeding of the Australasian Language Technology Workshop 2005,pages 57-63 Sydney ,Australia December 2005.

[62] Afsaneh Fazly and Graeme Hirst, “Testing the Efficacy of Part-of-Speech Information in Word Completion”, Proceedings of EACL 2003 Workshop on Language Modeling for Text Entry Method.

[63] Keith Trnka and Kathleen F. McCoy,” Evaluating Word Prediction: Framing Keystroke Savings ”,Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 261–264,Columbus, Ohio, USA, June 2008. c 2008 .

APPENDICES

Appendix 1. Sample of Morphological analysis

Form	Root/Stem	Affix
aanaani	aannan	_i
aanaanii	aannan	_ii
aanaanis	aannan	_in
aanaanq	aannan	_q
beeka	beek	_a
beekaa	beek	_aa
beekoo	beek	_oo
beekaadhaaf	beek	_aa_dhaaf
beekaadhaan	beek	_aa_dhaan
beekaafi	beek	_aa_fi
beekaan	beek	_aan
beekaat	beek	_aat
beekaatii	beek	_-aatii
beekama	beek	_ama
beekamaa	beek	_amaa
beekamaadhaatii	beek	
beekamaaf	beek	_amaa-f
beekamaan	beek	_amaa_n
beekamaatii	beek	_amaa_tii
beekaman	beek	_aman
beekamaniif	beek	_aman_fii
beekame	beek	_ame
beekamee	beek	_amee
beekameedha	beek	_amee-dha
beekameera	beek	_amee_ra
beeke	beek	_e
beekee	beek	_ee
beekeera	beek	_eera
beekellee	beek	_ellee

beeketti	beek	_etti
beeki	beek	_i
beekiiniif	beek	_ii_niif
beekin	beek	_in
beekini	beek	_inni
deema	deem	_a
deemta	deem	_ta
deemti	deem	_ti
deemtu	deem	_tu
deemu	deem	_u
deemna	deem	_na
Deemuu	deem	_uu
arguu	arg	_uu
argita	arg	_ta
arga	arg	__a
argiti	arg	_iti
argitu	arg	_itu
argu	arg	_u
argina	arg	_ina
barsiisaa	barsiis	_aa
barsiistuu	barsiis	_tuu
barsiisu	barsiis	_u
nama	nam	_a
namni	nam	_ni
namoota	nam	_oota
namaaf	nam	_aaf
namicha	nam	_icha
namichaa	nam	_ichaa
namichaaf	nam	_ichaaf
namichaan	nam	_ichaan
namichaanis	nam	_ichaanis
namichaas	nam	_ichaas
namichaatiin	nam	_ichaa-tiin
namichas	nam	_icha
namichi	nam	_ichi
namichii	nam	_ichii
namichis	nam	_ichis

namiiti	nam	_iiti
namiitu	nam	_iitu
namini	nam	_ini
namitti	nam	_itti
namittii	nam	_ittii

Appendix 2. Sample of Training Data

Bakka/PR buutuun/JJ Biirichaa/NN Aadde/NN Faantuu/NN kaleessa/AD meeshaa/NN qorannoo/ JJ dhiigaa/NN gargaarsa/VV waldaa/JJ misiyoonota/NN addunyaan/JJ hojjechaa/VV jira/ AX ./PN 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/VVxiyeeffatanii/AD hojjechuu/VV akka/PR qaban/AX ibsame/VV ./PN

Amajjii/NN ooluuf/VV hanga/PR ammaatti/AD ajjeechaan/VV raa'wate/VV hedduu/JJ hordofee/ VV dhugaan/NN jala/PR ka'udhaan/VV argama/AX ./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
Gumaa/NN daa'imman/NN adnyaa/NN dammaqinaan/AD to'annaa/VV poolisootiNN godhanVV qofaadha/VV ./PN mirkaneessuuf/VV shoora/NN olaanaa/JJ akka/PR gumaachu/VV ittigaafatamaan/JJ abbaa/NN Taayitaa/NN eegumsa/JJ Naannawaa/NN ibsame/VV ./PN Dhalattoota/NN uumamni/NN hirmaataa/VV biyyasaanii/JJ loogiin/VV raaw'atuVV hawaasichi/ NN barumsa/NN barbaachisaVV/./PN Kun/PP kakuu/VV Oromoon/NN qabudha/AX ./PN

Biyyeefi/NC bishaanii/NN isheetiin/PP taasifamne/VV kanuma/PP dhuumamuu/VV seenaniiVV yaadatama/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

Sareen/NN yeroo/AD dheeraaf/JP halkan/NN dutti/VV walirraa/PP baasuun/VV xiqqaanAD isaatti/PP kanneen/PR ayyaanichaafi/NC badhaadhina/NN jiru/AX /.PN
Guyyichi/NP guyyaa/NN muddamsaa/NN ta'uuf/VV ./PN Billachi/NN isaanitti/PP namite/vV maraguuf/VV dhandhamatti/VV fixuun/VV ./PN Michuun/NP koo/PP kompyuutara/NN isaa/PP irra/PR jira/VV ./PN

Midhaan/NN fagaachuun/VV oomishamu/VV qopheesadhu/VV ./PN Yoo/CC wal/PP hin/AX lolan/VV waraana/NN of/PP harkaa/NN qabu/AX taanaan/AX ofirraa/ PP garagalchanii/VV wal/PP rukutu/VV malee/CC ittiin/PS wal/PP waraanuun/VV safuu/AD dhorkaa/VV tureedha/AX ./PN Galmootan/NP wallitti/PP fuunaanee/VV kaa'a/VV ./PN

Bifti/JJ qonnaa/NN kanaa/PP boodaa/PR akka/PR geedaramuuf/VV jiru/AX bu'awwan/JJ qorannoo/ NN saayinsii/JJ addeessu/VV ./PN Galmeen/NP barbaachisaan/NP badanii/VV jiru/AX ./PN Waraana/NN geedaramuuf/VV guddaa/JJ kennuufi/VV fudhatu/VV keessa/PR hojjate/ Waajirichi/NN dhaabbilee/NN 23/JN keessattiPR jijjiirama/VV hojii/NN qoratee/VV hojjeessuuf/ VV sochiirra/VV akka/PR jiru/AX beeksisaniiru/VV ./PR Bara/AD 90/JN eegalee/PR lalisaa/VV jira/AX ./PN Bishaan/NN jireenyaafi/JC waan/PR barbaachisaa/AX guddaa/JJ akka/PR ta'e/AX beekamaadha/ AX ./PN Waggaa/NN 12/JJ booda/PR ,/PN dubartii/NN jalqabaa/JJ sanyii/JJ gurraachota/NN kessa/PR pirezidaantii/NN yuunvarstii/JJ taatee/AX hojjate/VV ./PN kan/PR duulee/JJ hin/AX beekne/AX hidhataa/VV bula/AX ./PN Waanni/PP cimaan/JJ ammoo/CC ciminasaatiin/JJ yoo/CC itti/PP fufe/VV kan/PR maqaan/JJ isaa/PP tolee/JJ mul'atu 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

/AX ./PN

Submitted by:

ASHENAFI BEKELE _____ _____

Student Signature Date

Approved by:

1. DR.YAREGAL ASSABIE_____ _____

Advisor Signature Date

2. _____ _____ _____

Examiner, Signature Date

3. _____ _____ _____

Chairman, Signature Date

Department's Graduate Committee