ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

WORD SEQUENCE PREDICTION FOR AMHARIC

By

NUNIYAT KIFLE ABEBE

FEBRUARY 2011
ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

WORD SEQUENCE PREDICTION FOR AMHARIC

A Thesis Submitted to College of Natural and Computational Sciences of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By: NUNIYAT KIFLE ABEBE

February 2011

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

WORD SEQUENCE PREDICTION FOR AMHARIC

By: NUNIYAT KIFLE ABEBE

Name and signature of Members of the Examining Board

_____	_____	_____
Advisor	Signature	Date
_____	_____	_____
Examiner	Signature	Date
_____	_____	_____
Examiner	Signature	Date

Declaration

This thesis has not previously been accepted for any degree and is not being concurrently submitted in candidature for any degree in any university.

I declare that the thesis is a result of my own investigation, except where otherwise stated. I have undertaken the study independently with the guidance and support of my research advisor. Other sources are acknowledged by citations giving explicit references. A list of references is appended.

Signature: _____

Nuniyat Kifle

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

Advisor‟s Signature: _____

Ato Ermias Abebe

Acknowledgment

My sincere gratitude first goes to GOD for helping me in every part of my study and finish this thesis. Then I would like to thank my advisor a lot, Ato Ermiyas Abebe for his constructive comment, supervision, and patience till the completion of this study.

I would like to offer special thanks to all my family members, who encouraged, loved and supported me with all you have. I also want to say thank you very much my LT Track Teachers, classmates and friends for giving me the courage and the cheer up until the end.

Abstract

Word prediction is a popular machine learning task, which consists of predicting the next word in sequence of words. Literature shows that word sequence prediction could play a great role in real life applications including electronic based data entry. Word prediction deals with guessing what word comes after, based on some current information, and it is the main focus of this study. Even though Amharic is used by a large number of population, few works are done on the topic of word sequence prediction. Previous works on word prediction shows that statistical methods are not enough with highly inflected language and needs syntactical information.

In this study, we developed Amharic word sequence prediction following the Design science research methodology with statistical methods using Hidden Markov Model. We used around 138,000 phrases to train the model by incorporating detailed parts of speech. The experiments were done using bigram and trigram models on a window size of two, five and seven. We explained the efficacy of part of speech tag in Amharic word sequence prediction.

Evaluation was performed using developed model and keystroke savings (KSS) as a metrics. According to our experiment, prediction results using a bi-gram with detailed Parts of Speech tag model has higher KSS and performed slightly better compared to those without Parts of Speech tag. Therefore, statistical approach with detailed POS with window size of five has good potential on word sequence prediction for Amharic language.

Keywords: Word sequence prediction, Parts of Speech, n-gram

Table of Contents

List of Figures v

List of Tables vi

List of Acronyms vii

1 INTRODUCTION 1

1.1 Background to the Research 1

1.2 Statement of the Problem 3

1.3 Research Questions 4

1.4 Purpose/Objectives 4

1.5 Significance of the Study 5

1.6 Scope and Limitations of the Study 5

1.7 Methodology 5

1.8 Evaluation Mechanism 6

1.9 Organization of the Thesis 6

2 LITERATURE REVIEW 7

2.1 Introduction 7

2.1.1 What is word sequence prediction? 7

2.1.2 Word sequence prediction – Historical Background 8

2.2 Approaches to Word Sequence Prediction 12

2.2.1 Statistical Word Prediction 12

2.2.2 Knowledge Based Word Prediction 13

2.2.3 Heuristic Word Prediction 14

2.3 Language model 14

2.3.1 Statistic language modeling 15

2.3.2 Knowledge-based language modeling....17
2.3.3 Neural Language Models....17
2.4 Algorithms....18
2.4.1 Hidden Markov Model....18
2.4.2 Recurrent Neural Network....19
2.5 Related works....19
2.6 Amharic language....28
2.6.1 Amharic Grammar....28
2.6.2 Amharic Parts of Speech....28
2.7 Summary....31
3 METHODOLOGY AND ARCHITECTURE OF AMHARIC WSP....32
3.1 Methodology....32
3.2 Problem identification and motivation....32
3.3 Definition of the Objectives for a Solution....33
3.4 Design and development....33
3.5 POS Tagging....36
3.6 Demonstration....38
3.7 Evaluation....38
3.8 Communication....39
4 EXPERIMENTS & EVALUATION....40
4.1 Experiment....40
4.1.1 Experimental Setup....40
4.1.2 Experimental Procedure....40
4.2 Evaluation and Testing....42

4.2.1 Keystroke Savings 42
4.2.2 Test Results 43
4.3 Discussion 47
5 CONCLUSIONS & RECOMMENDATIONS 49
5.1 Conclusions 49
5.2 Recommendations 51
6. REFERENCES 52
7. APPENDICES 58

List of Figures

Figure 2-1 key pad with multi tapping 10

Figure 2-2 keypad with QWERTY arrangement 11

Figure 4-1 Representation of bigram word sequence frequency in the corpus..... 37

Figure 4-2 Representation of bigram POS sequence frequency in the corpus 37

Figure 5-1 Demonstration of word based model 58

Figure 5-2 Demonstration of word based and POS based ranking 59

Figure 5-3 Demonstration of combined ranking 60

List of Tables

Table 2-1 summary of related works 24

Table 2-2 Amharic personal pronoun..... 27

Table 4-1 Test result from experiment 1 40

Table 4-2 Test result from experiment2..... 41

Table 4-3 Test result from experiment3..... 41

Table 4-4 Test results obtained from experiment4..... 42

Table 4-5 Test result from experiment5..... 42

List of Acronyms

AACS	Augmentative and Alternative Communication Systems
CPT	Compact Prediction Tree
DG	Distributed Generation
DSRM	Design Science Research Methodology
E-mail	Electronic Mail
HMM	Hidden Markov Model
ITU	International Telecommunication Unit
KSS	Keystroke Savings
LSTM	Long Short-Term Memory networks
MA	Morphological Analysis
NLM	Neural Language Model
NLP	Natural language processing
POS	Part of Speech
RNN	Recurrent Neural Network
SLM	Statistical Language Model
SMS	Short Messaging Service
SRILIM	SRI Language modeling toolkit
SOV	Subject Object Verb
SVO	Subject Verb Object
TDAG	Tunable Distributed Data processing model
URLs	Uniform Resource Locaters
WIC	Walta information corpus
WSP	Word Sequence Prediction

Chapter One

1 INTRODUCTION

1.1 Background to the Research

Prediction is forecasting what comes after, based on some current information [1]. Word prediction aims at easing word insertion in textual software by guessing the next word, or by giving the user a list of possible options for the next word. Word Prediction is done based on knowledge about a language and the context of the conversation. The main purpose of word prediction is to ease data entry, but it can also help dyslexic people in reducing writing errors [1], [2]. It is stated that this field of research has seen a surge of interest with the development of mobile phones and other digital devices [2].

Word prediction can be done using different techniques and approaches. Statistical information that is mainly based on word frequencies, accounting for probabilities of separated words, or by more complex language models such as Markov models [3]. Word frequencies can be acquired from the user itself, can be acquired from a corpus or through combination [1].

The other technique is knowledge based. It uses syntactic knowledge which considers part of speech tags and phrase structures, that can be statistical or can be based on hand-coded rules. Semantic knowledge can be used by assigning categories to words and finding a set of rules that constrain the possible candidates for the next word [2]. This method is not widely used in word prediction, mostly because it requires complex hand coding or may be time consuming and in efficient for real-time requirements [2].

All prediction methods require lexical data. Such data can be acquired from corpora along with word frequencies and lexical databases. A word prediction lexicon usually includes words frequencies. It may also include parts of speech and semantic data. Lexicons can be adaptable, for example updated with the user's vocabulary, and should be organized in an efficient manner [3].

Word prediction, not to be confused with word completion, is a relatively new application area. Although both of the technologies are used to facilitate typing speed, the implementations and effects

are quite different. Word completion deals with prediction of which word the user wants to type now. It starts operating as soon as the user has typed the first letter in a word; analyzing the letters [2] whereas, word prediction deals with predicting a correct word in a sentence. It saves time, keystrokes and also reduces misspelling.

Since this work mainly focuses on Amharic word prediction model that can be an input for mobile phones and other digital devices data entry, we reviewed technological background and some works done in the area.

Mobile phones were produced with 12 key pads. Since there are more characters than the number of keys, more than one character have to share one key [4]. This is why ambiguity arises when pressing a key. i.e. which of the characters mapped to the key should appear on the screen. A simple idea is to have multiple taps on the key in order to have the desired character on screen. Multi tapping was a solution but it is hectic, time consuming and the user may get tired of tapping a lot of times [4] [5]. Other than speed multi tapping has problem in moving fingers to the required key, pause time to repeat to get the same key to type another letter. Nowadays mobile phones with full key pads like a computer keyboards in QWERTY arrangement are being produced. Then by using a genetic algorithm keyboard was optimized for mobiles which puts characters with highest frequency are on starting key and less frequent on latter keys. Nevertheless, it is small in size and still didn't solve the problem in text entry [1].

Later on the dictionary based technique was introduced which uses some type of corpus to build dictionary [4]. T9 by Tegic communications and eZiText by ZiCorp are two commercial examples of such predictive text software that work by maintaining dictionary of words. T9 is used in most mobile phones for text entry. In T9, a user has to press corresponding key once for each letter with no tap and after typing. If the desired word does not appear then press a special key, usually asterisk mark(*), to switch between different alternate words correspond to the same key character sequence. This technique is letter by letter prediction [4] [5]. The eZi integrated letter by letter and word completion technique. Then text prediction methods such as statistical and knowledge-based approaches started to be investigated.

1.2 Statement of the Problem

Word prediction is a popular machine learning task, which consists of predicting the next word in sequence of words. Literature shows that word prediction could play a great role in real life applications including mobile based data entry [12].

Most technological advancements incorporate computers and hand-held devices. More interestingly these advancements and enhancements are done using one's own language. Many countries are trying to make their language in use by designing and developing prediction systems in their local language [13], [14]. This research deals with designing word sequence prediction model for Amharic language. One of the needs for Amharic word sequence prediction for mobile use and other digital devices is in order to facilitate data entry and communication in local language [15].

Word sequence prediction is a challenging task for inflected languages [13]. Inflected languages are languages that are morphologically rich and have enormous word forms. i.e. one word can have different forms. As Amharic language is highly inflected language and morphologically rich it shares this problem. This problem makes word prediction system much more difficult and results poor performance. Here the question will be what technique should be used to help the predictor to suggest the word that goes with the current conversation? So storing all forms in dictionary won't solve the problem as in English and other less inflected languages. But considering other techniques that could help the predictor to suggest the next word like a Part Of Speech based prediction should be used [16].

Very few researches have been conducted in Amharic word prediction. The first work on Amharic word prediction was "Word Prediction for Amharic Online Handwriting Recognition" by Nesredin Suleiman [17]. He used dictionary approach with no consideration of context information. The second work on Amharic word prediction was "word sequence prediction for Amharic language" by Tigist Tensou [18]. Her work considers context information and tried to add some linguistic rules. However including morphological features for Amharic affects the search space and the speed of the suggestion since these two things are the very most important parameters for a word prediction model [13], [16]. In her research she did not include these parameters and did not include user profiling or adaptation which is another important feature of prediction. Therefore, this study solves speed, search space, and user profiling or adaptation using separate lexicon considering the nature of the

Amharic Language. Finally, what flaws are identified and what possible things to come up are recommended.

1.3 Research Questions

There are two steps to perform sequence prediction model. The first is using some previously seen sequences called the training sequence and second is to use the trained sequence prediction model to perform new predictions. So, through this process the questions are

1. What language model improves Amharic word sequence prediction considering the nature of the language?
2. What technique and specific method should be applied to give efficient performance for Amharic language word sequence prediction?
3. To what extent parts of speech tagging specifically help the Amharic word sequence prediction?

1.4 Purpose/Objectives

The general objective of this research is to develop Amharic word sequence prediction model using Part of Speech Tag for mobile use and portable digital devices.

The specific objectives of the research are:

1. To explore previously proposed related works in order to grasp basic knowledge about word sequence prediction
2. To Study the nature of Amharic language for the purpose of sequence prediction.
3. To explore the ways to improve Amharic word sequence prediction efficiency by adding features like Part of Speech Tag.
4. To design Amharic word prediction model
5. To develop a prototype to demonstrate the effectiveness of the designed model.
6. To evaluate the performance of the proposed model.
7. To communicate the result.

1.5 Significance of the Study

Designing Word sequence prediction for Amharic language is challenging to develop as it is necessary to study the neighboring characters, the context or meaning of the word, the structures, occurrences of words and the language behavior itself. However, its usage can improve writing and data entry. Also, this research output has a great importance for other researches done for Amharic language in the area of Amharic grammar checker, Amharic spell checker and other Amharic word prediction related researches. It points and gives direction on how statistical approach with Part of Speech Tag contributes for Amharic Word sequence prediction. In addition to that, it shows in what way to use the model for Amharic word sequence prediction. The study can be a baseline for other related studies. Generally, it addresses the issue to explore the means for effective Amharic text writing, reducing the input burden of the system, helping people with language impairment and facilitating data entry for Amharic language in mobile and portable digital devices.

1.6 Scope and Limitations of the Study

The scope of this study is restricted to the analysis and development of appropriate model for word sequence prediction for Amharic language. We developed Amharic word sequence prediction model using statistical approach incorporating syntactic information which in our case is part of speech tag. We tried to show the importance of part of speech tag in order to have better prediction. The corpus used had an impact on our keystroke savings since statistical approach needs huge data for the training. The focus in this study is only on achieving efficiency of the word sequence prediction model for Amharic language. Prediction way and the user interface issue are out of scope. Due to time limitation we couldn't include agreements between part of speech tags.

1.7 Methodology

The study follows Design Science Research Methodology [19]. In design science a problem is assessed and based on that problem artifact is proposed. Then the research design will be validated in order to understand more and to improve the design. These steps help better to design word sequence prediction model for Amharic language. Since DSRM includes approaches, techniques, tools, algorithms and evaluation mechanisms in the process [20], [21], we followed statistical approach with statistical language modeling and built Amharic prediction model based on information from Part of

Speech tagger. The statistics included in the systems varies from single word frequencies to part-of-speech tag n-grams. That means it included the statistics of Word frequencies, Word sequence frequencies, Part-of-speech sequence frequencies and other important information. We used python programming language and its libraries for both the part of speech tagger that was developed and the Amharic word sequence predictor module. Different literatures and previous works done were reviewed in order to have a good background about the topic, improve and introduce new concepts and theories. The methodology is explained in detail and discussed in chapter three.

1.8 Evaluation Mechanism

The system is evaluated using Keystroke Savings. Since the major goal of word prediction system is to reduce the number of keystrokes, the primary and well-known evaluation metrics for word prediction is keystroke savings which estimates the saved effort percentage [22]. By comparing the obtained KSS for tri-gram and bi-gram with different POS parameters, the model that shows maximum keystroke saving is considered as better model.

1.9 Organization of the Thesis

This thesis is organized in 5 chapters. Chapter 2 deals with literature review and related works. In Chapter 3, the design and methodology for Amharic word prediction model is discussed briefly with its Architectures. Chapter 4 deals with Experiments and Implementation in order to show the result of the developed model. Finally, in Chapter 5 conclusion and recommendations for future work are presented.

Chapter Two

2 LITERATURE REVIEW

2.1 Introduction

Many theories have been proposed and researches have been conducted to explain what sequence prediction is about and what word sequence prediction specifically covers. The literature covers a wide variety of such theories and concepts. This review however focuses on major concepts which emerge repeatedly throughout the literature reviewed. These concepts and subtopics are overview of word sequence prediction, its Historical Background, the different approaches used in Word sequence prediction, fundamental concepts of word sequence prediction and grammatical properties of Amharic language with parts of speech tag. Finally, the works related to word sequence prediction both in Amharic and other languages that are related to Amharic in their language properties are explored.

2.1.1 What is word sequence prediction?

We can easily define the term word sequence prediction once we capture the essence of its essential components which are "sequence" and "prediction". A sequence is a finite or infinite list of terms (or numbers or things) arranged in a definite order, that is, there is a rule by which each term after the first may be found [23]. Prediction is concerned with guessing the short-term evolution of certain phenomena [1]. Forecasting tomorrow's temperature at a given location or guessing which asset will achieve the best performance over the next month could be examples of prediction problems. One must predict the next element of an unknown sequence given some knowledge about the past elements and possibly other available information. The entities involved in forecasting task are the elements forming the sequence, the criterion used to measure the quality of a forecast, the protocol specifying how the predictor receives feedback about the sequence, and any possible side information provided to the predictor [24]. Therefore, word sequence prediction is forecasting or guessing the next word the user intends to write or to insert based on some previous information [24].

The primary users of word prediction systems have traditionally been physically disabled persons. For people having motor impairments making it difficult to type, a word prediction system would optimally reduce both the time and effort needed for producing a text, as the number of keystroke decreases.

Lately word prediction techniques have entered new domains. In particular, it has come to be an integrated part of many cellular phones as an assisting aid in the popular text message service. Since cellular phones are smaller in size [4], the text input method was by sharing the standard keys for dialing numbers. That way, a key along with a number, represents three or more characters. But this has changed after word prediction technique introduced.

2.1.2 Word sequence prediction – Historical Background

There is no specific date or time that shows the introduction of word sequence prediction [25]. However literatures shows that after the value of text based information continued to increase so did the need for word sequence prediction [1], [27]. Therefore, word sequence prediction becomes one of popular text or data entry techniques. Sequence prediction first started at a character level in which predicting the sequence is based on the character of the previous one [4]. T9 by Tegic Communications (www.tegic.com) and eZiText by Zi Corp (www.zicorp.com) were the first commercial products that started word sequence prediction at a character level [8]. Then eventually grew to word level in which predicting the sequence is based on the previous word as natural language processing becomes very popular and widely in use [4]. Various sequence prediction models have started to be investigated in the last two decades. There are actually numerous models that have been proposed by researchers such as all-k-order Markov, LSTM, TDAG and CPT. these models utilize various approaches. Some of them use for example neural networks, pattern mining and probabilistic approach [3], [26].

One of the most defining qualities of the modern man compared to other creatures on the earth is our ability to use highly complex language. Written language eventually allowed us to do all these things and more, especially for speech and hearing-impaired individuals; the access to a writing system actually gave them a chance to participate in the general social forum. Text prediction provides better data entry performance by improving the writing mainly for people with disabilities [3].

Using computers and hand-held devices gives us the ability to put previously theoretical knowledge such as mathematics, linguistics and physics into practical use with their immense computing power. Especially linguistics has seen a real jump in application areas in the last few years in such systems as augmentative and alternative communication systems, word completion, expert systems, information retrieval systems and many more [27]. AACS are the broad term used to describe the many communication methods which support or replace speech. Augmentative and Alternative Communication can help people express their needs, hopes and ideas, to connect with their family and friends, to communicate in the workplace, access education, understand more about the world around them [28]. This thesis deals with a certain type of AACS system called word sequence prediction, which can be of great use to functionally impaired users.

One of the word sequence prediction application is in SMS which is a text messaging service component of mobile phone, web, or mobile communication systems that allows exchange of up to 160character messages, including broadcast messages [29]. Most of the SMS messages are mobile to mobile. According to the latest United Nations telecom agency (International Telecommunication Union-ITU) report, there are over 6 billion mobile phone Subscribers in the world [www.bbc.co.uk].

When using this service there are different techniques to insert the text/data we intend to write. We will try to discuss some, such as writing using 12-key keypad phones, using multi-taping in reduced keys, using full-key keypad and predictive text entry methods.

Mobile phone comes with a 12-key keypad which has 26 English characters mapped on. There are more characters than the number of keys so more than one character has to share one key. This is why ambiguity arises when pressing a key. The 12-key keypad is still in use and comes with many popular brand mobile phones, including Samsung, Nokia, Motorola and Sony Eriksson. A simple idea to decrease the ambiguity was to have multiple taps on the key within specified time in order to have the desired character on screen. Hence, Multi tapping was provided as a solution to reduce the vagueness. But still it was chaotic, time-consuming and the user may get tired of tapping a lot of times. In regular multi-tapping technique, what we do is first find the key on which the required character is located. Then press the key and tap it one less times the index of character on that key, with the first character having index 0 [4]. For example, in order to type "v" one needs to press key number 8 once and tap the same key twice. Similarly, in order to type "m", we simply

need to press key number 6 with no tapping. The ambiguities in reduced keypads are first multiple characters share same key.

Secondly, the ambiguity arises when typing two consecutive characters of a word that appears on the same key. To Disambiguate these two using multi-tapping we must tap twice to select required character on the key and then pause to get another character from the same key respectively [1] To resolve the multi tapping issue present in the 12-key keypad mobile phones researchers introduced mobile phones with full key keypads that look like keyboards. Mostly, full-key keypad has QWERTY arrangement of characters [1]. To have all the characters on the keypads in device like mobile is unlikely since they are small in size. Yet the use of mobile phones with both types of keypads is equally popular.

In Amharic language there are 33 characters. So using multi tapping is even more complex since the characters are more in Amharic language. It is shown in the figure 2.2 that the Amharic characters "ሀ" "ለ" and "መ" shares the key pad number 2. The character "ሀ" has seven forms which are "ሀ", "ሁ", "ሂ", "ሃ", "ሄ", "ህ" and "ሆ". This means to write a word that contains "ሁ" the user must first select "ሀ" from the keypads. The same is true for all the 33 characters. Therefore, for writing one word a user has to tap many times. Following the QWERTY keypad arrangement just like in English language some keypads were developed in Amharic language as it is shown in the figure 2.3. But the challenge is still there.

Figure 2-1 key pad with multi tapping

Figure 2-2 keypad with QWERTY arrangement

As a result, predictive text entry method is introduced. This is the method that solves all the ambiguity issues arose. This method can be used in both key-based and pen-based paradigms. Mostly it is dictionary-based method. Dictionary based techniques use some type of corpus to build dictionary. T9 was one example that used letter-by-letter prediction which is used in most mobile phones. In T9, a user has to press corresponding key once for each letter with no other tap. After typing, if the desired word does not appear then press a special key, usually *, to switch between different alternate words that correspond to the same key character sequence [1], [4]. This is letter-by-letter word prediction technique.

The eZiText is a combination of T9 and word completion techniques. Word Completion feature suggests the user most probable postfix character sequence automatically after each typed character sequence while typing. [4]

In recent years, dictionary-based disambiguation mechanisms have appeared in various forms. These include N-gram frequencies, syntactical information, or other statistical information of letter and word frequencies. Based on previously written character(s) N-gram character prediction could be used to predict the next character (N^{th}) which is going to be written at the writing space.[3] For example in English language, due to the high frequency of words such as “the, then, them, then, these, etc.” the character “e” has a high probability to appear with the context “th” than any other character whereas character “z” has zero probability to appear with the context “th” [30].

Word prediction system is more effective than character prediction system because word prediction system relies upon several words of past context whereas character prediction relies only upon characters in the current word [23]. Having some previous context will enable the system to predict the next character based on the probability associated with it. Currently there are a number of predictive systems developed by different companies for mobile.

However, dictionary-based techniques are not such a good text entry method especially for morphologically rich languages such as Amharic languages. Because morphologically rich languages have many word forms. One word can have countless inflections and derivations [3], [24]. Researches have been conducted to complete a word a user is currently typing using dictionary of words with their frequency. One of the drawbacks in the existing approach is it is impractical to capture all word forms due to the language's rich morphology. Moreover, it doesn't consider context information. This results in syntactically wrong word proposal causing extra cognitive load to adjust suggested words to appropriate form as well as causing reduction in speed of text entry.

2.2 Approaches to Word Sequence Prediction

Word prediction methods can be classified as statistical, knowledge based and heuristic (Adaptive) modeling. Most of existing methods employ statistical language models using word n grams and POS tags.

2.2.1 Statistical Word Prediction

In statistical modeling, the choice of words is based on probability that a string may appear in a text. [31] The statistical information and its distribution could be used for predicting letters, words, and phrases. Statistical word prediction is made based on Markov assumption in which only the last n-1 words of the history affects succeeding word and it is named n-gram Markov model. It is based on learning parameters from large corpora. However, one of the challenges in this method is when a language that is being written with the help of word prediction system is of a different style than the training data [32], [33]. Word frequency and word sequence frequency are commonly used methods in word prediction. The early predictive systems use frequency of each word independently to complete a word in the current position of a sentence being typed without considering context information. In other words, the system uses unigram word model with a fixed

lexicon and same suggestion is offered for a particular sequence of letters. However, prediction is better if context is taken into account [1].

2.2.2 Knowledge Based Word Prediction

Word prediction systems that merely use statistical modeling for prediction often present words that are syntactically, semantically, or pragmatically inappropriate and impose a heavy cognition load on users to choose the intended word in addition to performance decrease in writing speed. To solve the problem syntactic, semantic and pragmatic linguistic knowledge can be used in prediction systems. [14]

2.2.2.1 Syntactic Knowledge for Word Prediction

In this approach, Parts-of-Speech (POS) tags of all words are identified in a corpus and the system uses this knowledge for prediction. This approach requires a set of linguistic tools such as POS taggers and Lemmatizes. However, these are not available in all languages. Statistical syntax and rule-based grammar are two general syntactic prediction methods, where statistical syntax uses the sequence of syntactic categories and POS tags for prediction. Therefore, a probability would be assigned to each candidate word by estimating the probability of having this word with its tag in the current position and using most probable tags for previous one or more words. In rule-based grammar, syntactic prediction is made using grammatical rules of the language. A parser will parse current sentence according to grammar of the language to reach its categories [14].

Syntactic prediction using probability table takes syntactic information inherent to natural languages into account. This approach makes use of probability of appearance of each word and relative probability of appearance of every syntactic category after each syntactic category. These systems offer words with most probable syntactic categories at the current position of a sentence and results are usually better than the ones obtained using purely frequency-based word prediction methods.

2.2.2.2 Semantic Knowledge for Word Prediction

Semantic prediction works by semantically analyzing sentences as they are being composed, where each word has an associated semantic category or a set of semantic categories. The working method, complexity, dictionary structure, adaptations, etc. are very similar to syntactic approach using grammars. It provides comparable result to syntactic approaches though it has much higher

complexity, and due to this these methods are not commonly used [34]. In semantic word prediction, Lexical source and Lexical chain are two methods that are used. The first method is lexical source, like Word Net in English, measures probability of words to get certain that predicted words is related in that context. The second method is lexical chain that assigns highest priority to words which are related semantically in that context with removal of unrelated words to that context from the list of predictions.

2.2.2.3 Pragmatics Knowledge for Word Prediction

Predictions can be correct syntactically or semantically but wrong according to discourse. Pragmatics affects capability of the predictor and taking this knowledge while training the system enhances accuracy of predictions [35].

2.2.3 Heuristic Word Prediction

Heuristic (adaptation) method is used to make more appropriate predictions for a specific user and it is based on short term and long-term learning. In short term learning, the system adapts to a user on 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 to a user in a single text. However, in long-term prediction the previous texts that are produced by a user are considered [36].

2.3 Language model

For a word prediction system to make predictions it needs access to some kind of language model [5]. A language model is an approximate description that captures patterns and regularities present in natural language and is used for making assumptions on previously unseen language fragments [25], [31]. There are different language modeling techniques; neural language modeling, statistical language modeling and knowledge-based language modeling. A statistic language model is based on data automatically extracted from large corpora. A knowledge-based language model, on the other hand, relies on manually defined linguistic generalizations, such as grammar rules [37]. Generally, word prediction systems make use of statistical language models, though; some systems supplement these models with a knowledge-based source of information.

2.3.1 Statistical language modeling

Statistical language models have generally been the main source of information used in word prediction systems [24]. The statistics included in the systems varies from single word frequencies to part-of-speech tag n-grams.

2.3.1.1 Word Frequencies

Most of the earliest word prediction systems, developed in the beginning of the 1980s, are based on unigram statistics, taking only the frequency of single words into account when trying to predict the current or next word. A system that is solely based on unigram statistics will always come up with the same prediction suggestions after a particular sequence of letters or after a blank space, irrespective of the preceding context. The lack of context considered entails that the suggestions often will be syntactically and semantically inappropriate.

2.3.1.2 Word Sequence Frequencies

In n-grams (with $n > 1$) the probability of a certain word is computed based on the previous word(s). So an upcoming word is influenced by the previous context [24]. Incorporating an n-gram model, some syntactic and semantic dependencies are captured, since words that frequently co-occur are promoted. This is in particular true for languages like the European ones, with a rather specific word order.

Still, some syntactically and semantically inappropriate suggestions are likely to be displayed as well, since the n-grams are limited to a local context consisting of a specific number of words. Usually no more than two or three words are considered, as a great amount of text is needed to find more n-grams of a larger size.

2.3.1.3 N-gram Language Models

N-gram language models are the most well-known type of language models in speech and language processing. They are important tool in NLP tasks. Thus, N-gram language models have dominated the speech recognition area for years due to their simplicity, efficiency and robustness [38]. On the other hand, it plays vital role in Handwriting recognition for estimating confusable and unreadable letters and words [39]. In general, besides these sample areas, *N*-grams are also crucial in NLP foundations like part-of speech tagging, natural language generation, and word similarity, as well

as applications from authorship identification and sentiment extraction to predictive text input systems for cell phones.

N-gram approximate the probability of a word sequence as a product of conditional probabilities of the current word w_i given a history of the preceding n- 1 words. The order of an n-gram model refers to the value n, where n is the number of words used in the probability sequence.

2.3.1.4 Part-of-speech Sequence Frequencies

In order to generalize the n-gram model, some systems make use of n-grams of part-of-speech tags instead of word forms. Using this method the probability of the current tag given the previous tag(s) is computed [40]. Since part-of-speech tags capture a lot of different word forms in one single formula it is possible to represent contextual dependencies in a smaller set of n-grams. One major advantage of making use of part-of-speech frequencies is thus that a larger context may be taken into consideration. There is a loss in semantics though, since the tags only tell, for example, that a verb is likely to follow a noun and not what particular verbs are typically connected with what nouns.

2.3.1.5 Importance of Part of Speech Tagging

Part-of-speech tagging involves selecting the most likely sequence of syntactic categories for the words in a sentence [41]. These syntactic categories, or tags, generally consist of parts of speech, often with feature information included. Part- of speech tagging is useful for speeding up parsing systems and allowing the use of partial parsing. Many current systems make use of a Hidden Markov Model (HMM) for part-of-speech tagging [42]. Other methods include rule based systems, maximum entropy models and memory-based models. In an HMM tagger the Markov assumption is made so that the current word depends only on the current tag, and the current tag depends only on adjacent tags.

Part-of-speech tagging is normally treated as a classification task with the goal to assign lexical categories (word classes) to the words in a text [41]. Most work on tagging has concentrated on English and on using supervised methods, in the sense that the taggers have been trained on an available, tagged corpus. Both rule-based and statistical machine-learning based approaches have

been thoroughly investigated. Therefore, parts of speech tag help in prediction in order to know the words categories and context [9].

2.3.2 Knowledge-based language modeling

While n-gram based systems never consider contexts exceeding a limited number of words, a knowledge-based system may take an arbitrarily large sentence fragment into consideration. Thus a well-defined knowledge-based language model would have the advantage of being able to capture long-term dependencies not covered by any of the statistic language models described above [3]. If grammar rules are used, the number of syntactically inappropriate suggestions would be reduced, leaving more room for possibly intended word forms to appear in the prediction list. Further, a low rate of inappropriate prediction suggestions imposes a lower cognitive load, enhancing the comfort of the user [43].

2.3.3 Neural Language Models

Neural network approaches are achieving better results than classical methods both on standalone language models and when models are incorporated into larger models on challenging tasks like speech recognition and machine translation. A key reason for the leaps in improved performance may be the method's ability to generalize [44], [45]. Neural language models address the n-gram data scarcity issue through parameterization of words as vectors (word embedding) and using them as inputs to neural network. The parameters are learned as part of the training process. Word embedding obtained through NLMs exhibit the property whereby semantically close words are likewise close in the induced vector space [26].

More recently, recurrent neural networks and networks with a long-term memory like the LSTM [26], are used to allow the models to learn the relevant context over much longer input sequences than the simpler feed-forward networks. RNN language model provides further generalization: instead of considering just several preceding words, neurons with input from recurrent connections are assumed to represent short term memory. The model learns itself from the data how to represent memory [51].As Mingxuan Wang, Zhengdong Lu, Hang LiWenbin Jiang and Qun Liu stated in their paper called "Convolutional Architecture for Word Sequence Prediction" their model can effectively fuse the local correlation and global correlation in the word sequence, with a convolution-gating strategy specifically designed for the task making the use of Neural Language

Modeling. They argued that their model can give adequate representation of the history, and therefore could naturally exploit both the short and long range dependencies [26].

2.4 Algorithms

2.4.1 Hidden Markov Model

The Hidden Markov Model is a popular statistical tool for modeling a wide range of time series data. In the context of natural language processing, HMMs have been applied with great success to problems such as part-of-speech tagging and noun-phrase chunking [42]. Andrei Markov gave his name to the mathematical theory of Markov processes in the early twentieth century [42], [47].

2.4.1.1 Markov Processes

Given a sequence of observations, example: up-down-down we can easily verify that the state sequence that produced those observations and the probability of the sequence which is simply the product of the transitions [42], [48]. More formally, an HMM model is characterized by the following [49], [50].

- N, the number of states in the model. Generally, the states are interconnected in such a way that any state can be reached from any other state. The hidden states often represent important physical aspects. In part of speech tagging, the states represent the tags. We denote the individual states as T = t_1, t_2….. t_n, and the state at time t as q_t.
- V, the number of distinct observation symbols per state, i.e., the discrete alphabet size. The observation symbols correspond to the physical output of the system being modeled. For part of speech tagging, the observation symbols are the words of a given language. We denote the individual symbols as W = w_1; w_2…….w_m.
- The state transition probability distribution A = {a_{ij}} where
  $a_{ij} = P(q_{t+1} = \frac{t_j}{q_t} = t_i)$, $1 \leq i, j \geq N$, $a_{ij} \geq 0$
- When there are no connections between states, the corresponding a_{ij} transition probability is zero. For all other cases, it is greater than zero.
- The observation symbol probability distribution in state j, B = {$b_j(k)$} Where
  $b_k = P(\frac{w_k}{q_t} = t_j)$, $1 \leq i, j \geq N$, $1 \leq k \geq V$

- The initial state distribution $\pi = \{\pi_{i\}}$, where

  $\pi_I = P(q_1 = t_i),\ 1 \leq i \geq N,$

2.4.2 Recurrent Neural Network

The current state of the art models for the task are language models that use neural networks. Although they perform very well, they often require a lot of memory and computational power. The disadvantage is that it cannot take into account the context in which the prediction is being made, unlike some neural network methods such as RNN based model. [46]

The RNN-based models, including its variants like LSTM, enjoy more popularity, mainly due to their flexible structures for processing word sequences of arbitrary lengths, and their recent empirical success [44]. LSTM is well-suited to model long term dependencies and has been successfully applied to language modeling. RNN language model provides further generalization: instead of considering just several preceding words, neurons with input from recurrent connections are assumed to represent short term memory. The model learns itself from the data how to represent memory [51].

2.5 Related works

As mentioned in section 2.6, Amharic is morphologically rich language. Even though Amharic language was found to be morphologically rich and complex language for natural language processing few works has been done in word prediction. To mention those conducted in the area the first was by Nesredien Suleiman on "Word prediction for online hand writing recognition" and by Tigist Tenesou on "word sequence prediction for Amharic language" respectively [17], [18].

Nesredin Suleiman proposed a Word prediction for online handwriting recognition. He explained that it was challenging to develop both handwriting recognition with word sequence prediction as it is necessary to study the neighboring characters and the Structures or occurrences of words. However, he indicated its usage can highly improve the efficiency of online handwriting recognition. Thus, he tried to address this issue in order to explore the means for effective Amharic text writing. He used the dictionary approach and didn't consider context information while developing the word sequence prediction model. As we briefly discussed in the above section dictionary approach is not the best approach for word sequence prediction for morphologically rich

language since it is very difficult to store all word forms in the dictionary. However, this area needed further research to support users of the language on their text entry techniques and Tigist Tenesou's work was one such research. She tried to consider context information with some morphological feature and linguistic rules [18]. However, her work did not consider the two most important features of WSP which are speed and search space. Adding morphological features in prediction affects its response time as it is described in the paper and rather adds cognitive load.

Many researches have been done in other highly resourced languages. To mention some English [45], Japanese [13], Hebrew [35] and Urdu [4] text prediction systems are particularly reviewed. In Japanese word sequence prediction system, it is stated that Japanese morphological analysis takes an unsegmented string of Japanese text as input, and outputs a string of morphemes annotated with parts of speech. As morphological analysis is the first step in Japanese NLP, its accuracy directly affects the accuracy of NLP systems as a whole. The researchers used Adaptable approach. In adaptation, the word prediction technique is for the mobile SMS system that will work in mobile for a single user at a time. [2] Thus, the technique must adapt to the user using the mobile phone. The dictionary would be populated with words as a result of training, but users have their own vocabulary of words for daily use that must add to dictionary. Thus, an option given to the user to add user vocabulary (i.e., non-dictionary words) should be turned on automatically otherwise the system would prompt to ask whether the typed non-dictionary words should be added to the dictionary or not. Similarly, there is option to turn off dictionary predictions and switch to traditional multi tapping. Adapt and change according to user's preference means that words used by user will be preferred by the system. Words will sorted according to theirs frequencies and first five will be in prediction list and others all will be called as non-prediction list words.

In Hebrew word prediction system, the researchers explained as Modern Hebrew is characterized by rich morphology, with a high level of ambiguity, in which several words combine into a single token in both agglutinative and fusion ways. Hebrew lexemes are derived according to various processes, such as composition (*e.g.,* of root and pattern), linear derivation (affixation and compounding), and null derivation. Hebrew lexemes are inflected by gender, number, and person, tense and construct state [35]. These properties are applied to most content-carrying words. The definite article of Hebrew is attached pronominally to nominal: common nouns, adjectives, demonstratives, and numbers. A word in Hebrew can be attached, according to its lexical category,

to a sequence of formative letters indicating functions, such as coordination, preposition and subordination. In addition, nouns, adjectives, verbs, prepositions, and adverbs can be agglutinated with a pronominal pronoun suffix, indicating functions, such as possessive, accusative and nominative (with person/gender/number inflections) [3].They used the statistical approach which the approach that is widely and commonly used.

Another work is Urdu word sequence prediction system. Urdu is the national language of Pakistan and one of the state languages of India. It is spoken by around 200 million people across the globe [wiki.answers.com]. Computational aspects of Urdu language started in early 1980s. Natural language processing techniques are applied on Natural Language to obtain computable linguistic artifacts. They used the dictionary approach which still is a bit difficult for languages of morphologically rich [4].

We mentioned this works because they relate in a way with Amharic language which is the morphological characteristics of the languages. It is described that like other Semitic languages, Amharic exhibits a root-pattern morphological phenomenon. A root is a set of consonants (also called radicals) which has a basic lexical meaning. A pattern consists of a set of vowels which are inserted among the consonants of a root to form a stem. In addition to this non-concatenative morphological feature, Amharic uses different affixes to create inflectional and derivational word forms. Some adverbs can be derived from adjectives. Nouns are derived from other basic nouns, adjectives, stems, roots, and the infinitive form of a verb by affixation and intercalation. Case, number, definiteness, and gender marker affixes inflect nouns [10]. The related works are summarized in Table 2.1 below.

T
h
K
2
w
o

Author(s)	Year	Title	Statement of problem	Methodology and tools	
Nesredien Suleiman	July, 2008	Word Prediction for Amharic Online Handwriting Recognition	Importance of handwriting recognition in reducing the input burden and no work has been done. It is First research attempt in Amharic WSP	Dictionary based approach (using existing Amharic dictionary, Adoption of Character Recognition Engine, ***SuperWaba*** SDK Java-compatible platform	Pr
Yael Netzer, Meni Adler, and Micheal Elhadad,	2008	Word Prediction in Hebrew: Preliminary and Surprising Results	Previous work states that for languages with a rich inflectional morphology, a statistical method based on frequencies only is not sufficient, and a wide variety of syntactic knowledge is required	Statistical approach using morpho-syntactic knowledge, Hidden Markov Model	pr
Masood Ghayoomi and Ehsan Daroodi	2008	A POS-based word prediction system for the Persian language	As there was no previous work done in Persian language	Used statistics, syntactic categories of a Persian POS tagged corpus, and the third approach, main syntactic categories along with their Morphological, syntactic and semantic subcategories.	re sh

Fredrik Lindh,	2011	Japanese word prediction	The computer linguistic problems local to the Japanese language, the corpus used as the bigram data source and the architecture of the software in question	Statistical heuristics tagging with n-gram analysis. With .Net framework	

Ja

sy

in

co

a

s

A

c

4

2

T

h

ef

f

w

R

A

w

Author(s)	Year	Title	Statement of problem	Methodology and tools	
Sana Shahzadi, Beenish Fatima, Kamran Malik and Syed Mansoor Sarwar	2013	Urdu Word Prediction System for Mobile Phones	There was no previous work done to assist people in typing using Urdu language and to present a new technique that uses „bigram‟ model developed for Urdu language word prediction.	Dictionary based approach with adaptive modeling and using Bigram language model	be pr
Tigist Tensou Tessema,	October, 2014	Word Sequence Prediction for Amharic Language	Previous work doesn‟t consider context information and concept of morphology. This results in Syntactically wrong word proposal.	Statistical approach, Horn morph morphological analyzer, python programming language	ke in bi
Longkai Zhang, Li Houfeng, Wang XuSun	October, 2014	Predicting Chinese Abbreviations with Minimum Semantic Unit and Global Constraints	To solve the “character duplication” problem in Chinese abbreviation Prediction and Different to previous works new Chinese abbreviation prediction method was proposed which can incorporate rich local information while generating the abbreviation globally.	Statistical approach with n-gram language Model Integer Linear Programming (ILP) formulation, MSU (Minimum Semantic Unit) Based Tagging	co ba co

te
c

2
0

Th
m
e
in

m
e
A
s

h
o

5
=

2.
1
Th
60

Author(s)	Year	Title	Statement of problem	Methodology and tools	
Kenneth C. Arnold, Krzysztof Z. Gajos and Adam T. Kalai	October, 2016	On Suggesting Phrases vs. Predicting Words for Mobile Text Composition	A system capable of suggesting multi-word phrases while someone is writing could supply ideas about content and phrasing and allow those ideas to be inserted efficiently	Statistical approach	re he to co (1 th th
Wakshum Temesgen	June, 2017	Word Sequence prediction model for Afaan Oromoo	There was a dearth of research in Ethiopia on word prediction, particularly on Afaan Oromo language.	Statistical approach, python programming language	ke Bi of

o b

m o

Table 2-1 summary of related works

2.6 Amharic language

Amharic is a Semitic language which is the second-most spoken in the world, after Arabic [6], [52]. It is written (left-to-right) using Amharic Fidel, which grew out of the Ge'ez script (Ge'ez abugida) [6]. The Amharic language is written with the same letters as Ethiopic each letter varying in seven different forms in order to express different sounds and vowels [6], [53]. There are 33 consonant symbols that have seven variations that should be coupled with the consonants to form Amharic language.

2.6.1 Amharic Grammar

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 be put together to make sentences. Word order and morphological agreements are basic issues considered in Amharic grammar and are used as part of our word sequence prediction study. Sentences in Amharic language are formed from verb phrase and noun phrase. Amharic sentences are short in the number of words they contain because they have so many prefixes and suffixes [11], [53]. Let's take a simple example the sentence "Abebe ate his lunch." "አበበ ምሳውን በላ"። "ው" represents "his" and written as a suffix in the noun ምሳ which has a meaning of lunch. Therefore, we can say the number of words in Amharic sentence is short. However, due to data scarcity we changed the sentences in to phrases. Phrase is a small group of words that stands as a conceptual unit.

Formal Amharic texts follow subject-object-verb (SOV) word order unlike English language which follows subject-verb-object (SVO) sequence in a sentence. Example of a simple sentence in Amharic: - አበበ ምሳውን በላ ። Its transcription is Abebe msawn bela. Which means Abebe ate his lunch ended with a Verb "በላ" (ate).

Although in some Amharic texts, there can be OSV sequence like "ሌጁን አበበ መከረው ።" Its transcription it will be "ljun abebe mekerew" which means The boy is advised by Abebe, where in this case the object is suffixed by object marker ን/n, however this word order is not commonly used in formal Amharic texts [14].

2.6.2 Amharic Parts of Speech

Amharic language has different parts of speech as in English and other languages. Here in this thesis Part of Speech tag has a big role in giving syntactic information for the prediction model. In

most researches related to word sequence prediction POS tagger and POS sequence are used to give syntactic information in order to have better prediction [16]. POS tagging deals with assigning an appropriate word class for each word. Our Tag model gives information about Part of Speech comes after what and helps in the prediction model combined with the word model. Noun, Pronoun, Adjective, Adverb, preposition, could be some examples of the Amharic part of speech.

2.6.2.1 Nouns

Nouns are one part of speech that represents a place, name, different objects or things. [52] Nouns can be created depending on many variables such as gender (masculine/feminine), plural or singular, informal or formal, possession etc. To form the different combinations, suffixes are added to the root of the nouns. The suffix changes depending on gender [52]. For instance, if words end in a "h" sound and when speaking to a female, words end in a "sh". Let‟s take "how are you" in English. This sentence is the same for a male or female in English but in Amharic "dehna ne**h**" "ዶህና ነህ" and "dehna ne**sh**" ዶህና ነሽ? are used for male and second one is for female respectively. If they are speaking to a male or a female Plurals are formed adding "woč" or "oč" (if the word ends in a vowel or a consonant) to the noun. [53]

2.6.2.2 Pronouns

Words that are used in position of nouns are called Pronouns because pronouns can do everything that nouns can do [51]. A pronoun can act as a subject, direct object, indirect object, object or more. [53] The ten Amharic personal pronouns are shown in Table 2.2.

Singular		Plural		Formal/Respect	
I	እኔ	We	እኛ	You	እርስዎ
You(male)	አንተ	You	እናንተ	He/She	እርሳቸው
You(female)	አንቺ	They	እነሱ		
He	እሱ				
She	እሷ				

Table 2-2 Amharic personal pronoun

2.6.2.3 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. “የ”(ye) , “እንዲ” (ende) , “ከ”(ke) can be mentioned as an example of prepositions.

2.6.2.4 Verbs

Words that describe an action are called Verbs. An Amharic verb root consists of set of (usually three) consonants. A verb form normally has one or more suffixes and prefixes added to create verb forms and always agree with their subjects. Verbs are marked for person, number, gender and a verb form can also agree with the direct or indirect object of the verb. Amharic verbs often have additional morphology that indicates the person, number, and (second- and third-person singular) gender of the object of the verb. አሌማዘን አየኋት ('I saw Almaz'). [52] Below are some examples of Verbs in Amharic.

• “Mekeregn” “መከረኝ” which has a meaning in English „he advised me“

• “Mekeresh” “መከረሽ” which has a meaning in English 'he advised you: feminine“

• “Mekere” “መከረ”which has a meaning in English 'he advised“

2.6.2.5 Adjectives

Amharic has few primary adjectives. Adjectives can stand alone but are mostly added as a suffix to the noun. In general adjectives are words which describe or modify another person or object in a given sentence. For example: **a red flower (ቀይ አበባ)** the adjective is red(ቀይ) because it describes the noun flower(አበባ). The following examples use Amharic adjectives (the underlined ones) in different ways and places to demonstrate how they behave in a sentence [53]. **Examples** ቤቱ <u>ነጭ</u> ነው። which has a meaning in English My house is White. <u>ሶስት</u> ቡችሊዎች አሎት። which has a meaning in English She has three small dogs. <u>ትንሽ</u> <u>አረንጓዴ</u> ቤት አሇኝ። which has a meaning in English I have small green house.

2.6.2.6 Adverbs

An adverb is a part of speech that provides greater description to a verb, adjective, another adverb, a phrase, a clause, or a sentence [53]. በትንሹ (a little bit), ብዙ (a lot), ዳግሞ (also), ሁሌጊዜ (always) etc are some Amharic adverb examples.

2.6.2.7 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. እና (and), ስለዚህ (because), ስለሆነም (Therefore) are some examples of Amharic conjunctions [54].

2.7 Summary

In this chapter we assessed word sequence prediction approaches, techniques, potential algorithms. Then we reviewed previous studies both in local language, other languages, including what specific approach they followed, and results obtained from their experiment. This helped us to choose which approach and method of procedure to follow for Amharic word sequence prediction. We reviewed the Amharic language and its important features like Amharic grammar and Amharic part of speech tag. Therefore, considering the language nature and based on literature review we decided to use statistical approach with part of speech tag and keystroke savings as an evaluation metric for our model.

Chapter Three

3 METHODOLOGY AND ARCHITECTURE OF AMHARIC WSP

3.1 Methodology

Methodology is a way of understanding how research is done scientifically which has its own procedures and process to follow. It also includes Methods such as approaches that are used to the conduct the research, the technique used for collection of data, data preparation, analyzing the data, and training methods. The methodology that we particularly followed in this thesis is Design Science Research Methodology [19], [20].

Design Science is a Research methodology that creates and evaluates IT artifacts intended to solve identified social or organizational problems [21]. It incorporates principles, practices, and procedures required to carry out a research. The Design Science process includes six steps [21]. These are problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. Even though the process is structured in sequential order, there is no belief that researchers would always proceed in sequential order from activity 1 through activity 6 [19]. It can be started from any activity.

In this work we are following a problem centered approach which starts from activity one to activity six since the research idea resulted from observation of the problem.

3.2 Problem identification and motivation

In our study we identified the problem through literature review. Different literatures and previous works done were reviewed in order to have background about the topic and introduce us with concepts and theories in word sequence prediction [54]. There are different researches that have been done in this study area. Some of them are researches that were done in other languages other than Amharic and English. Therefore, the literatures indicated that more work could be done in this area and motivated us to read more on the gaps on Amharic word sequence prediction.

3.3 Definition of the Objectives for a Solution

Following the first step it led us to find a better solution as well from the literatures we have reviewed. We proposed a solution from the works done in the area based on the statement of the problem. Hence, we developed word sequence prediction model for Amharic by adding POS as syntactic feature.

3.4 Design and development

Here we designed the Amharic Word Sequence prediction model and architecture using the concepts and the procedures as an input from step one and two. Below is the detailed description of the Amharic word sequence prediction model.

As explained in chapter one Prediction methods include decisions on prediction units (characters, words), information sources and structure (both lexical and statistical), levels of linguistic processing, size and types of corpora and learning methods. In this research word based statistical approach is used.

In chapter two we discussed three approaches, each having their own advantages and disadvantages. Knowledge based and heuristics approaches need syntactic, semantic or pragmatic linguistic knowledge as well as topic modeling. To make this approaches in use we need linguistic tools such as POS taggers, lemmatizes, and parser [34]. We also need lexical source like Word Net. Since these tools are not available for Amharic language and cannot be modeled due to time limitation with the scope of the research these approaches won't be functional.

Hence, in this research statistical word prediction approach was used to design Amharic word sequence prediction model. When using statistical word prediction, the selection of words is based on the probability that a string may appear in a text. Statistical word prediction is made based on Markov assumption in which only the last n-1 words of the history affects the succeeding word and hence it is named n-gram Markov model. The system trained using n-gram model that calculate probability for all n-grams in the corpus. Based on previously written characters or words N-gram characters or word prediction could be used to predict the next word (Nth) [38].

The statistics included in the systems varies from single word frequencies to part-of-speech tag n-grams. That means it includes the statistics of Word frequencies, Word sequence frequencies and part-of-speech sequence frequencies. N-gram language model approximates the probability of a

word sequence as a product of conditional probabilities of the current word *wi* given a history of the preceding *n*-1 words. The order of an *n*-gram model refers to the value *n*, where n is the number of words used in the probability sequence. In Part-of-speech tagging frequency it encompasses selecting the most likely sequence of syntactic categories, this is done using Hidden Markov Model (HMM) based on Markov assumption in which current word depends only on the current tag, and the current tag depends only on adjacent tags [32], [42], [48] .

Thus the general equation for this N-gram approximation to the conditional probability of the next word in a sequence, $w_1, w_2, ..., w_n$, is: (1) If N = 1, 2, 3 in (1), the model becomes unigram, bigram and trigram language model, respectively, and so on. For example, the probability of English sentence "I want to eat "using unigram and bigram is shown below [46].

- Unigram probability: - P (I want to eat) = P (I) × P (want) × P (to) × P (eat).
- Bigram probability:- P(I want to eat) = P(I) × P(want| I) × P(to | want) ×P(eat | to)

Now if any of these four words are not in the training corpus then the probability of the sentence will be zero for the cause of multiplication. So in this statistical method if we want to consider these words then we need a huge data corpus that must contain all the words of the language. So there may arise the problems like Many entries in corpus are with low frequency and Probability of a word sequence will be very low or zero. If an entry does not exist in corpus then the probability of the sentence will become zero because of multiplication. To solve the problem, we have applied the back-off. In the back off method, for N = 3 i.e. for a trigram model, the word sequences will follow trigram probabilities at first; if it could not match then word sequences will follow bigram model; if it also could not match then word sequence will follow unigram model and predict at least a word. As the system is for word prediction and it is playing with word sequence, so though any probability of word sequence is zero for multiplication, it predicts at least a word [25], [31], [55].

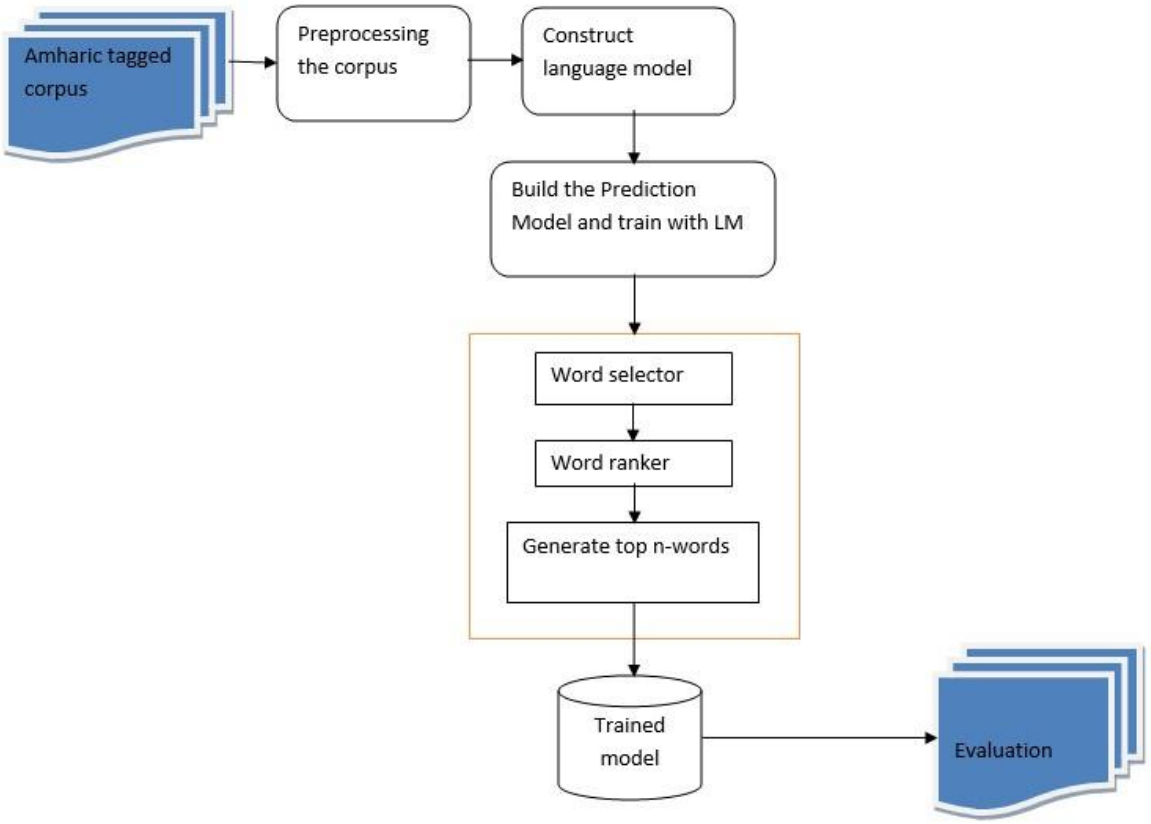


Fig 3.1 Adopted architecture of a Word Sequence Prediction system [16]

As shown in the figure 3.1 architecture of Amharic word sequence prediction the first activity is data collection or preparing corpus. A corpus is a large collection of written or spoken material in machine readable form which can be employed in linguistic analysis and is the main source of knowledge. Language models built from large corpora tend to perform better, particularly for words that are infrequent. Word sequence prediction task requires a large size of corpus in order to have sufficient statistical information for training the system. In this study, text collections are mainly gathered from Walta Information Center (WIC).

The WIC corpus is a medium-sized corpus of 210,000 tokens collected from 1,065 Amharic news documents [56]. The corpus is manually annotated for POS tags. The number of tag-sets, which is proposed based on the orthographic word, is 31 classes [56]. The tag-sets used are compound tags for words with prepositions and conjunctions. We couldn't find the whole corpus, so we trained the system with around 138,000 phrases.

The next step is Data preparation and cleaning, which is the most time taking and challenging task in statistical machine learning approach since this technique's efficiency depends on the data quality and amount of training. Data preparation and cleaning that was done here on the corpus are avoiding too much white space, removing punctuations, abbreviations and symbols. Since the corpus was a collection of manually POS tagged sentences there are many words that are tagged with wrong tag class [56]. We tried to correct those tags manually as much as possible. All the collected sentences were manually checked for spelling errors. The preprocessing also included preparing phrases so that we can avoid data sparsity. Lastly the corpus has been tagged in 31 tag classes but in order to see the detail tagging effect on the prediction result we have minimized it into the basic tag classes of seven as presented in reviewed literatures [57]. We used thirty tag classes excluding punctuation for the other experiment with detailed tag classes. After cleaning the data 90 percent of the corpus was used for training and 10 percent for testing purpose.

During the training, the first model generated is the language model. We trained the language model after generating n-gram count using SRILIM. This creates a file in the ARPA format. We used python code to read the ARPA file and calculate probability based on it. Thus, using the frequency distribution of the bigram count we were able to calculate the probability distribution by breaking up the phrases into bigram sequence of words incorporating part of speech words. In order to compute the probability we wrote python code that looks at each n gram in the corpus and if the n-gram is found in the table it reads the log probability by first building a vocabulary and scoring (using score method) using maximum likelihood estimation. The corpus is considered a stream of word-level tokens in a sliding window of size N, where N is the order of the Markovian assumption. Then the prediction algorithm takes the entered text, extracts the preceding n-1 and by backing off it build a list of probable next words. For each n, select the top matches based on n-gram frequency. Therefore we combine the results, sort by maximum likelihood high to low and select top n words after ranking the words.

3.5 POS Tagging

In Amharic language, words can be grouped within different tag classes based on different criteria's and the research area. For instance in WIC the linguists has tagged the words into 31 different tag sets [56]. One might group the tag classes depending on its work and considering how many tag sets needed. Here in data preprocessing step, we have removed tags that are not

going to be predicted or helpful in prediction. The best example for this is the punctuation <PUNC> tag. By adopting the basic tag classes from previous researches we grouped the words into seven basic tag sets such as nouns (N), pronouns (PRON), adjectives (ADJ), adverbs (ADV), verbs (V), prepositions (PREP), conjunction (CONJ), and added UNC which stands for unclassified and used for words which are difficult to place in any of the classes.[20] Some of these basic classes are further subdivided into a total of 30 to 33 POS tags in another work [58]. A noun "ቤት" ("bet") can be written along with a preposition "በ-ቤት", with a conjunction "ቤት-ና" or both with the preposition and conjunction "በ-ቤት-ና". Therefore the word "በ-ቤት-ና" can be tagged as a noun with preposition and conjunction (NPC). The same Amharic word can have different POS tag according to the context. For example, consider the following phrase in our corpus, "አብይ/ADJ ርእስ/N በሚካሄደው/VP" and አቶ/ADJ አብይ/N ታመነ/N እንዳስታወቀው/VP". In the first example context, the word "አብይ" is a name of a person i.e. noun and on the second example it is an adjective which modifies the noun "ርእስ" meaning main in English. Here in this research Amharic POS tagger was built based on 31 tag classes. But for the Amharic prediction model the tag sets did not include the punctuation (PUNC) tag classes since these tag sets are not that important in prediction. And to see the detailed tagging effect on the prediction we trained another experiment using seven basic tag sets and unclassified tag set.

As we can understand the process and components of Amharic prediction model from the architecture, tagged corpus was constructed with a normalized tag sets and detailed tag sets. Then we construct and trained language model so that word predictor module predict the word succeeding using statistical analysis of word frequencies and word sequence frequencies from list of all words in Amharic lexicon. Also the language model enabled the word predictor to produce the next word by first checking POS sequence providing basic information about what comes after what. Then selection of words took place by selecting the potential predictions. Finally the predicted words are ranked.

The Word Ranker gives orders by considering the frequency of each word and part of speech tag, in the list of found words and tags. Words and tags with utmost frequency got highest rank and those with minimum frequency got the least rank. In the case of two or more words having the equivalent frequency; the word rankers will decide the rank of the words by considering their alphabetical order and the most appropriate POS tag predict based on n-gram Prediction.

Literatures show that using the scaling approach, a window size of 2-10 improves keystroke savings by 0.2% to 0.4% and decrease savings for window size of 1 by 0.1%. We used this approach because most AAC systems recommend to have 5-7 window size [59].

3.6 Demonstration

In this step the algorithm developed and the approaches followed must come into practical ground and be demonstrated using different tools and techniques. Demonstrating the use of the artifact to solve one or more instances of the problem is important [21]. This could involve its use in experimentation, simulation, case study, proof, or other appropriate activity [19], [20]. So here in our study we conducted experiments with different parameters.

There are various programming languages that could be used in word sequence prediction such as C++, java, prolog and python. Even though these languages are potential choices, we preferred to use Python due to the advantages and characteristics of Python below.

Python programming language is easy to learn for novice programmers [60], [61] and because Python programs are typically 3-5 times shorter than equivalent Java programs [62]. Python has top the charts in the recent years over other programming languages like C, C++ and Java and is widely used by the programmers by its Improved Programmer's Productivity [60]. The language has extensive support libraries for Machine learning and Natural Language processing and clean object-oriented designs that increase two to tenfold of programmer's productivity compared to languages like Java, VB, Perl, C, C++ and C# [62].

The SRILIM was used to generate unigram, bigram and trigram count as an input for the language model [63]. SRILIM is a toolkit that is used to build and apply statistical language modeling [64]. Using the output of the SRILIM language model, the model is designed using python programming language after it is preprocessed. Linux Mint was used as the main operating system during the frame work design.

3.7 Evaluation

Evaluation is observing and measuring how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution to actual observed results from use of the artifact in the demonstration. It requires knowledge of relevant metrics and analysis

techniques. We evaluated our model using metrics called Keystroke savings. Keystroke savings (KS) measures the percentage reduction in keys pressed compared to letter-by-letter text entry [22].

3.8 Communication

The last step in design science research methodology is communication. Communicate the problem and its importance, the artifact, the rigor of its design, and its effectiveness to researchers and other relevant audiences such as practicing professionals, when appropriate takes place in this stage [20]. So we will present our work to other researchers and experts so that we can get feedback on the study. It is communicated through department of information science in Addis Ababa University. We also will present it in different NLP conferences and try to publish an article on it.

Chapter Four

4 EXPERIMENTS & EVALUATION

This chapter presents the experiments and evaluation conducted by the researcher on the Amharic word sequence prediction model developed. This chapter generally presents the training and testing data, the implementation, and the analysis of the experiment results.

4.1 Experiment

4.1.1 Experimental Setup

The experiment was done with Dell Latitude E7240 with processor of Intel CORE i5 4300U, 4GB RAM. We used Linux operating system for the developed Amharic word sequence prediction model.

The corpus we used for the Amharic word sequence prediction model is Walta Information News Corpus. As mentioned in chapter three preprocessing the corpus was the first task during the experiment. We have used 138,000 phrases (around 17.3 MB in size). There is a sample of corpus we used in appendix A

4.1.2 Experimental Procedure

We merged and applied the three techniques which are the word frequency, probability distribution for both word and POS and n-gram which in our case is a bi-gram model and trigram model. We used the word frequency to store proper words and help us rank the words according to the number of the occurrences in the corpus. Then the calculated probability was used to store the training data both word sequence and Part of Speech tag sequence. Finally the training result from bigram was stored into the probability table for use as a reference for the next word prediction. We only reported bigram for the next two reasons. One because of the data scarcity and second as the n-gram increases it makes the computing more complex and do not produce a significant difference.

Many researchers explained that data scarcity in statistical approach is a critical problem [14], [65]. So the best way to overcome this problem was to change the corpus into manageable phrases. We

removed punctuation marks since we don't predict punctuation marks, avoided too much white space and stop words. We used the tagged phrases to generate bigram to build the word model. Then we generated bigram word sequence with the POS sequence model in order to predict the next word as shown in appendix C and D.

The Basic approach in our research is the principle of a probabilistic language model. Thus a training text was used to extract information about the language in question. This information is used by the language model in order to predict successive words, given a sequence of words and their respective POS as shown in appendix B. Here the information to extract and how the information is used was denoted by transition probabilities that represent the conditional probabilities for the next word given the previous words. The bigram and the trigram were generated with their frequency and probability of occurring. Then HMM (Hidden Markov Model) algorithm, which considers the above context information, was used for the predictor model to make the word suggestions. Finally the word ranker module predicts the top n words from the right POS category by comparing their probabilities.

We have done four basic experiments and sub experiments. The First experiment is on pure word-based bigram model. Since we are testing the effect of parts of speech tag we need first to see how much keystroke is saved without the existence of POS. So we trained the predictor module with the word language model and evaluated with KSS.

The second experiment incorporates part of speech tag which was described in the previous section. We trained bigram for both word and POS sequence, generate conditional probability, compare the probability and suggested the top five words and finally evaluated the final result.

We conducted the third experiment using detailed part of speech tag that has thirty-one tag sets. We used all the tag classes in the WIC tagged corpus only removing the punctuation tag class ("PUNC") which is 30 tag classes. We followed the same procedure as of in second experiment only by changing the tag sets. We trained bigram for both word and POS sequence, generate conditional probability, compare the probability and suggested the top five words and finally evaluated the final result.

The fourth experiment conducted in bigram by differentiating POS parameters. This experiment assisted us to see how the POS Tag helped the prediction. So what we did was trying to capture

the output rank for both the word model and the POS model using the score calculation function like $S = w_1*P_r + w_2*W_r$ [65]. Where w_1 is weight 1 and w_2 is weight 2 that are weight values between 0 and 1 which are experimental parameters to see how the Part of Speech tag helps the prediction. P_r is rank for the POS probability and W_r is rank for word sequence probability. So using this function our model re-ranks the previous one by combining the two ranks.

The last sub experiments are done with window size two, five and seven. Following the scaling approach window size has an effect on keystroke savings of word prediction.

The difference between the second and the third experiment is that in the third experiment we analyzed the importance of Part of speech tagging in the prediction model where as in the fourth experiment we tested in how much detail tagging needed in the prediction model. We said this because as much as we minimize the search space there is a higher probability to predict the correct word needed by the user faster.

4.2 Evaluation and Testing

Evaluation and Testing includes testing every module or component of the Amharic word sequence prediction that is testing the language model, the Tagger, and finally the predictor module. The testing is done using Tagged Amharic news text having a total of 100 phrases. Furthermore, spelling errors, wrong POS tag information and some typographic errors found on the testing data are manually checked and corrected. We assumed a perfect user who doesn't make typing mistake and picks the appropriate word right away when it is displayed in the list of word suggestions.

4.2.1 Keystroke Savings

Since the major goal of word prediction systems is to reduce the number of keystrokes, the primary and well-known evaluation metrics for word prediction is keystroke savings which estimates the saved effort percentage [22]. Keystroke savings (KS) measures the percentage reduction in keys pressed compared to letter-by-letter text entry [14], [24]. To calculate keystroke saving (KS) we subtract keys with prediction from keys normal divided by keys normal. Then we take the output and multiply by 100% in order to get by how much percentage the key stroke was minimized.

$$KSS = \frac{keys\ normal - keys\ with\ prediction}{keys\ normal} * 100$$

A word prediction system that offers higher savings will benefit a user more in practice [4]. Even though calculating Keystroke saving is the best method to evaluate word prediction [22]. There are some factors that affect the result, Dictionary coverage is one dominant factor affecting KS especially for inflected languages. Typically, for inflected languages reasonable KS is obtained limiting the number of words. Keystroke savings gives us better results when there is more data or dictionary of words. For a significant KS above 30% can be offered and that, surprisingly, slightly better results have been proven for a dictionary of about 1,200,000 words than for a limited dictionary of 250,000 words [65]. However, we used a tagged corpus of 414,000 words to train the language model and a Part-of-Speech tagger that annotates on-the-fly words with their POS. This explains the keystroke percentage as the previous researches indicated and as we have witnessed from our experiment. Also the suggested words from the language model may be incorrect; as a result it limits the KS. Below in the test results section there are the results we found from each experiment.

4.2.2 Test Results

Experiment 1

The first experiment conducted was based on word model. We only took word frequency, word sequence frequency and the bigram word language model to calculate the conditional probability and finally compare the probability and suggested words that have highest frequencies. Calculating the keystroke saving we obtained 18% savings. We took 198 words for testing. The characters required to type the whole word is 962 characters. We took the variable 'N' to assign the window size two, five and seven. The demonstration is shown in appendix F, G and H.

What we mean by word based prediction model is generating the language model with only sequence of words. Then the predictor module compare the probability and suggest the top n words which in our case is two, five and seven. Let's take for example the phrase "ኮሚሽነሩ አስታውቀዋል". So in order to predict the word "አስታውቀዋል" the word model has to look for words that occurred with "ኮሚሽነሩ" as shown in appendix E. This is when the predictor module ranks and suggests the

words. The average Keystroke Savings for the word based prediction model is 18%. The results are shown in Table 4.1.

Model 1 (word-based prediction model**)**	**KSS**
N=2	14.7%
N=5	20
N=7	21.6

Table 4-1 Test result from experiment 1

Experiment 2

Since the major goal of this thesis is to show the importance of parts of speech tag in word sequence prediction we trained second bigram model with the tagged corpus. Word frequency, word sequence frequency, POS frequency were generated. The demonstration is shown in appendix I, J and K.This experiment incorporates parts of speech tag in to the language model. What the model do here is calculate the word sequence frequency, then the POS sequence frequency and outputs the suggested words by combining the two probabilities as shown in the appendix H. Let's take the same example we used in experiment 1 phrase "ኮሚሽነሩ/N አስታውቀዋል/V". So in order to predict the word "አስታውቀዋል" the word model has to look for words that occurred with "ኮሚሽነሩ" as it is shown in the picture. This is when the predictor module ranks and suggests the words from the word model. Then it will repeat the same process for the POS probability. Lastly it will take the product of the word and the POS probability giving a combined ranking. The average Keystroke Savings for the POS based prediction model is 35.8%. The results are shown in Table 4.2

Model 2 (POS based prediction model)	KSS
N=2	31%
N=5	37.5%
N=7	39%

Table 4-2 Test result from experiment2

Experiment 3

The other experiment is to see how the detailed tagging and general tag affects the prediction on both bigram and trigram. Here we called the Normalized tag with eight class tags briefly discussed in chapter 3. As it is shown in the appendix our normalized tag set contains basic eight tags and follows the same procedure as in experiment 2. We took the same test we used in other experiments. The characters required to type the whole word is 962 characters. The results are shown in table 4.3 below.

Type of tags used	Value of KSS
Normalized tags (8)	31.5%
Detailed tags (30)	34.7%

Table 4-3 Test result from experiment3

Experiment 4

Here in this experiment what we did was selecting three experimental parameters in order to explain the effect of the POS weight. We tested the model with all possible weights (between 0 and 1) and selected the weights below. So, we gave first the bigram word model a weight of 0.8 and POS model 0.2 and evaluated the Keystroke Saving, then a weight of 0.4 for the bigram word model and 0.6 for the POS model. Lastly we gave a weight of 0.1 for the bigram word model and 0.9 for the POS model which is a highest Weight. Even if we assign less weight for the POS model the existence of the POS by itself affected the result positively. The results are shown in table 4.4 below.

W1(word model)	W2(POS model)	Value of KSS
0.8	0.2	31%
0.4	0.6	32.3%
0.1	0.9	32.8%

Table 4-4 Test results obtained from experiment 4

Experiment 5

This experiment followed the scaling approach. Scaling approach states that window size of the prediction has a great factor in evaluation of word sequence prediction. So in order to see the effect we took 198 words for testing. The characters required to type the whole word is 962 characters. We took the variable 'N' to assign the window size two, five and seven.

When N = 7 the system saved 207 of characters which implies the system got a correct prediction of sequence words. Therefore, the evaluation of the prediction system saved 21.6% using keystroke saving.

When N=5 the keystroke saving scores 20% by correctly predicting 189 characters and when n=2, the system correctly predicts only around 14.7% by predicting 141 characters correctly. The results are shown in table 4.5.

Value of N	KSS(Word-based)	KSS(POS based)
2	14.7%	31%
5	20%	37.5%
7	21.6%	39%

Table 4-5 Test result from experiment5

It is clear that the higher the number window size the greater the chance of having the intended word among the suggestions. But larger value of N adds cognitive load on the user as they make the search time for the desired word longer. So we selected two, five and seven for N to measure how these parameters affect the performance.

When evaluating the model we took different test words that could explain the keystroke savings.as it is shown in appendix L and appendix M. Let's take the first word "የወፍጭ" as an example. The word that should be suggested based on the model is the word "ምንዛሪ". The model has correctly

guessed or predicted "ምንዛሪ" therefore it has saved 4 keystrokes.in the second case let's see the word "ቢቢሲን". Here the word suggested should be "ጠቅሶ" instead the model generated the word "ተቅሶ" so in this case it did not save any keystrokes. By summing up all these results from the test set we can calculate how many keystrokes the model saved. The second case also shows how much the spelling error affected the results.

As it is mentioned in the introductory part of this chapter every component should be evaluated module by module. Because the evaluation results of the components has a great influence on the overall system. For instance the language model and the tagger contribute for the prediction model. We calculated Perplexity, to measure the language model and must be computed with models that have no knowledge of the test set. In perplexity, the better model is the one that has a tighter fit to the test data or that better predicts the test data. The higher the conditional probability of the word sequence, the lower the perplexity. Therefore, low perplexity indicates the probability distribution is good at predicting the sample [59].

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

We used Trigram back-off and TnT tagger for the part of speech tagger and obtained accuracy of 79%. Means there is a 21% probability of the system tagging wrongly. The accuracy decreased because of poor quality corpus and size. This lowered the keystroke savings almost by half since our work depends on Part of Speech tags.

4.3 Discussion

The main objective of this thesis is to implement Amharic word prediction model that works with prediction speed less than 2 sec and with narrowed search space as much as possible. As it is mentioned in the previous chapters we introduced two models; tags and words and linear interpolation that uses part of speech tag information in addition to word n grams in order to

maximize the likelihood of syntactic appropriateness of the suggestions and we believe the results we found reflected this.

Word sequence prediction using bigram model with higher POS weight and detailed tagging gave better keystroke savings in all scenarios of our experiment. The chance of predicting accurately is higher when using larger training data. However, based on our experiments we found higher KSS in the bigram model. Finding higher n-gram sequence in the corpus is difficult due to the size limitation of the corpus. Word predictors for English [66], Swedish [14], Hebrew [35] and Urdu [4] shows 31%, 26%, 29%, and 38% KSS respectively. The approach used in these studies and complexity of the language varies from ours. Due to this, it is difficult to draw a firm conclusion based on their findings. However, we consider that the result in this work can greatly be improved with addition of more training data with different domain and POS tagger. The corpus we used for the training data is manually tagged data. This corpus produced errors due to the inconsistency of the tags by the linguists in process of manual tagging. There are many words that are tagged with wrong word class and this surely affected the efficiency of the POS tagger as well as the prediction. Also, there are too many words with spelling errors and this minimized the keystroke savings since corpus quality matters much in keystroke savings.

Chapter Five

5 CONCLUSIONS & RECOMMENDATIONS

5.1 Conclusions

Word prediction is a process of predicting a word that a user intends to write based on a lexicon and statistical information obtained from the corpus. This thesis presented Amharic word sequence prediction model using a statistical approach. We described a combined statistical and lexical word prediction system for handling inflected languages by making use of POS tags to build the language model.

We developed Amharic language models of bigram and trigram for the training purpose. The bigram and trigram model used can be extended to an n-gram (where n is greater than 3). Since prediction systems are used mostly for text messages, we normally use short sentences [25] in our case we trained phrases. This implies a higher N-gram model wouldn't be useful and complicates the prediction process. As the value of n increases so does the complexity but has no significant change on the result.

Based on the experiments carried out in this study and the results obtained, the following conclusions were made.

- Generally we can conclude that employing syntactic information in the form of Part-of Speech (POS) n-grams yields better predictions. Even if we started with a hypothesis of weighted part of speech tag can improve the keystroke savings but the difference was insignificant.
- One of the most important component of our model was the language model. The probability returned by the language model are mostly useful to compare the likelihood that finds right sequences and compare the probabilities of the phrase sequences. For instance given the word "ልየ" the model compare the following probabilities.

1. P(ወረዳ | ልየ) 0.9356
2. P(ዞን | ልየ) 0.9023

3. P(መልከተኛ | ልዩ) 0.8690
4. P(ግምት | ልዩ)) 0.8555
5. P(ነው | ልዩ)) 0.8555

The model compares the probability and suggests the word with highest probability of occurring. When the model observe a word which is not recognized as a known word as part of the sequence that means word does not occur in a list of known sequences. So using Hidden Markov Model we find the sequence closest including the candidate suggestions.

- Incorporating parts of speech tag information in the model helped the predictor module to discriminate words with equal word probabilities and assisted the model to narrow down the search space. For instance if we take probability of four and five they are equal but the POS sequence probability differ as "**ግምት**" a noun followed by noun whereas the later "**ነው**" is noun followed by auxiliary therefore here the probability of the POS sequence determines the final suggestion.
- We also can conclude data quantity and data quality highly affects the keystroke savings. Here in our study the tests were done on a small collection, so the effect of the prediction on bigger collection is not known. The smaller the training data the smaller keystroke savings. In addition to that prediction performance was decreased due to incorrect spellings. Meaning the model looks for exact match of the word in the corpus. But due to misspellings found in the corpus the predictor could not find the intended word.
- According to our evaluation better Keystroke saving (KSS) is achieved when using bigram model than the trigram models having a limited data. Since speed and search space are the basic issues in word sequence prediction the other high keystrokes with better speed were obtained using the experiment of detailed Part of speech tags. In previous studies the results were found to be 20.5%, 17.4% and 13.1%. we got significant change comparing our results to the previous results which is 35% on average.

5.2 Recommendations

Based on the findings of this study and the knowledge obtained from the literature, the following recommendations are forwarded for future work.

- Deficiency of accurate POS tagger and adequate size of POS tagged Amharic corpus makes our study very difficult. Thus, working on this area will greatly enhance the result in word sequence prediction. Amharic word sequence prediction can be optimized if good Amharic POS tagger is incorporated and if the model is enriched with POS.
- We tried to make our prediction model adaptive by using separate lexicon. However, because of poor performance of POS tagger and insufficient corpus, we didn't get the intended result and could not report it in the study. Therefore, we suggest taking corpus from different domains especially, short SMS texts, emails or chats then training it with POS tagger and develop the prediction model.
- The results clearly implied using high quality corpus and automatic spell checker in pre-processing the raw corpus will help to have more effective Amharic word sequence predictor. Since pre-processing corpus takes much of the time, it is great to work on the corpus for Amharic word sequence prediction as well as for other related works.
- We recommend researchers to conduct Amharic word sequence prediction using statistical approach by adding syntactic features like agreement and long-distance dependencies.
- We recommend addressing more accuracy in prediction by adding semantic knowledge to the system and conduct a study using a mixture of already proposed techniques for best results.

6. REFERENCES

[1] S. Agarwal, "Context based word prediction for texting language," *Large Scale Semant. Access to Content (Text,* pp. 360–368, 2007.

[2] K. Trnka, "Adaptive language modeling for word prediction," *HLT-SRWS '08 Proc. 46th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, no. June, pp. 61–66, 2008.

[3] G. R. Gendron and G. R. Gendron, "Natural Language Processing : A Model to Predict a Sequence of Words," *ModSim World 2015*, no. 13, pp. 1–10, 2015.

[4] S. Shahzadi, B. Fatima, and K. Malik, "Urdu word prediction system for mobiles," *World Appl. Sci. J.*, vol. 21, no. 9, pp. 1260–1265, 2013.

[5] V. Srinivasan, S. Moghaddam, A. Mukherji, K. K. Rachuri, C. Xu, and E. M. Tapia, "MobileMiner: Mining Your Frequent Patterns on Your Phone," *Proc. 2014 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. - UbiComp '14 Adjun.*, pp. 389–400, 2014.

[6] A. Wimsatt and R. Wynn, "Amharic Language and Culture Manual," 2011.

[7] D. Z. Zeleke and A. T. Submitted, "Amharic sentence to Ethiopian sign," 2014.

[8] W. M. Weldeyesus, "Case Marking Systems in Two Ethiopian Semitic Languages," *Color. Res. Linguist.*, vol. 17, no. 1, pp. 1–12, 2004.

[9] M. Gasser, "HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya," *Conf. Hum. Lang. Technol. Dev.*, no. May, pp. 94–99, 2011.

[10] Baye Yimam, "The interaction of tense, aspect, and agreement in Amharic syntax," *Proc. ACAL*, vol. 35, no. i, pp. 193–202, 2006.

[11] B. T. Ayalew, "The Submorphemic Structure of Amharic: Toward a Phonosemantic Analysis," 2013.

[12] S. Bickel, P. Haider, and T. Scheffer, “Learning to complete sentences,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3720 LNAI, pp. 497–504, 2005.

[13] F. Lindh, L. L. Larslarmostasluse, and A. H. Arthurholmerlingsuse, “Japanese word prediction.”, Japanese studies, Lund University, Sweden , 2011.

[14] E. Gustavii and E. Pettersson, “A Swedish Grammar for Word Prediction,” no. June, 2003.

[15] L. Zhang, L. Li, H. Wang, and X. Sun, “Predicting Chinese Abbreviations with Minimum Semantic Unit and Global Constraints,” *Emnlp 2014*, pp. 1405–1414, 2014.

[16] A. Fazly and G. Hirst, “Testing the efficacy of part-of-speech information in word completion,” *Eacl*, no. 1991. pp. 9–16, 2003.

[17] Nesredin Suleiman “Word Prediction for Amharic Online Handwriting Recognition,” *Word J. Int. Linguist. Assoc.*, 2008.

[18] Tigist Tessema, “Word Sequence Prediction for Amharic Language,” no. October, 2014.

[19] B. Kuechler and V. Vaishnavi, “Theory development in design science research: anatomy of a research project,” *European Journal of Information Systems*, vol. 17, no. 5. pp. 489–504, 2008.

[20] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[21] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design Science in Information Systems Research,” *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.

[22] K. Trnka and K. F. Mccoy, “Evaluating Word Prediction : Framing Keystroke Savings,” *Comput. Linguist.* No. June, pp. 261–264, 2008.

[23] K. C. Arnold, K. Z. Gajos, and A. T. Kalai, “On Suggesting Phrases vs. Predicting Words for Mobile Text Composition,” *Proc. 29th Annu. Symp. User Interface Softw. Technol. - UIST ’16*, pp. 603–608, 2016.

[24] M. Ghayoomi and S. M. Assi, “Word Prediction in a Running Text: A Statistical Language Modeling for the Persian Language,” *Proc. Australas. Lang. Technol. Work.*, no. December, pp. 57–63, 2005.

[25] N. Carmignani, “Predicting Words and Sentences using Statistical Models,” 2006.

[26] M. Wang, Z. Lu, H. Li, W. Jiang, and Q. Liu, “genCNN: A Convolutional Architecture for Word Sequence Prediction,” pp. 1567–1576, 2015.

[27] G. Lugosi, *Prediction, learning, and games*, vol. 9780521841, no. January. 2006.

[28] M. ANTOINE, M. DAILLE, and T. Wandmacher, “Adaptive word prediction and its application in an assistive communication system,” *Info.Univ-Tours.Fr*, p. 191, 2009.

[29] Alembante Mulu, “Amharic Text Predict System for Mobile Phone,” vol. 3, no. 4, pp. 113–118, 2015.

[30] S. Bickel, P. Haider, and T. Scheffer, “Predicting sentences using N-gram language models,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT ’05*, 2005.

[31] R. Rosenfeld, “Two decades of statistical language modeling: where do we go from here?” *Proc. IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[32] A. Ahmed, “Modeling and Predicting Sequences: HMM and CRF”, 2009.

[33] Nicola Carmignani , “Predicting Words and Sentences using Statistical N-gram language models” , July 5, 2006.

[34] R. Denaux and J. M. Gomez-Perez, “Towards a Vecsigrafo: Portable Semantics in Knowledge-based Text Analytics,” *Int. Work. Hybrid Stat. Semant. Underst. Emerg. Semant. @ISWC17*, 2017.

[35] “Word Prediction in Hebrew-Preliminary and Surprising Results”, 2007 no. Lm.

[36] Y. Ehara, Y. Miyao, H. Oiwa, I. Sato, and H. Nakagawa, “Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning,” *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pp. 1374–1384, 2014.

[37] Y. Even-Zohar and D. Roth, “A Classification Approach to Word Prediction *,” 2000.

[38] J. Boyd-graber, “N-gram models,” vol. 5, no. 2017, p. 9, 2009.

[39] D. Jurafsky and J. H. Martin, “N-Grams,” *Speech Lang. Process. An Introd. to Nat. Lang. Process. Speech recognition, Comput. Linguist.* 2008.

[40] A. Fazly and G. Hirst, “Testing the efficacy of part-of-speech information in word completion,” *Eacl*, no. 1991. pp. 9–16, 2003.

[41] G. G. A. Celano, G. Crane, and S. Majidi, “Part of Speech Tagging for Ancient Greek,” *Open Linguist.* vol. 2, no. 1, pp. 1–20, 2016.

[42] D. Jurafsky and J. Martin, “Hidden Markov Models,” *Speech Lang. Process.*, no. Chapter 20, p. 21, 2017.

[43] S. Wintner, “Natural Language Processing of Semitic Languages,” pp. 43–67, 2014.

[44] C. Republic and T. Mikolov, “Statistical Language Models Based on Neural Networks,” *Wall Str. J.*, no. April, pp. 1–129, 2012.

[45] M. Nakamura and K. Maruyama, “Neural network approach to word category prediction for English texts,” *Proc. 13th …*, pp. 213–218, 1990.

[46] P. Godard, “RNN-based sequence prediction as an alternative or complement to traditional recommender systems RNN-based sequence prediction as an alternative or complement to tra- ditional recommender systems,” 2017.

[47] R. Dugad and U. Desai, “A tutorial on hidden Markov models,” *… Electr. Eng. Indian Inst. …*, 1996.

[48] L. Rabiner and B. Juang, “Introduction to hidden Markov models.” *Ieee Assp Mag.*, pp. 1–15, 1986.

[49] M. Ward, “Hidden Markov Models,” no. February, pp. 1–7, 2012.

[50] E. Onegin, “Hidden markov and maximum entropy models,” *Language (Baltim).*, pp. 1–15, 2007.

[51] Y. Zhao, “Sequence Prediction Using Neural Network Classifiers,” pp. 1–6, 2016.

[52] “The Amharic Definite Marker and the Syntax-Morphology Interface ∗ Ruth Kramer University of California, Santa Cruz,” pp. 1–39, 2008.

[53] “Taye Assefa and Shiferaw Bekele, the Study-of-Amharic-literature: An Overview, JES vol 2, pp. 27-73, 2000”.

[54] K. E. N. Peffers *et al.*, “D Esign S Cience in I Nformation,” *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.

[55] P. Koehn, “Chapter 7 Language models Language models,” *Stat. Mach. Transl.*, 2009.

[56] A. M. Gezmu, B. E. Seyoum, and A. Nürnberger, “Contemporary Amharic Corpus : Automatically Morpho-Syntactically Tagged Amharic Corpus,” pp. 65–70, 2018.

[57] Binyam Gebre and L. Processing, “Part of speech tagging for amharic,” pp. 1–20.

[58] Martha Yifru Tachbelie, S. T. Abate, and L. Besacier, “Part-of-Speech Tagging for UnderResourced and Morphologically Rich Languages – The Case of Amharic,” no. May, pp.
2–5, 2011.

[59] M. G. Seyyed and M. Assi, “Word Prediction in a Running Text: A Statistical Language Modeling for the Persian Language,” 2005.

[60] C. R. Severance, “Python for Everybody,” *Creat. Commons*, p. 247, 2016.

[61] P. Python and J. Java, “A Comparison of the Basic Syntax of Python and Java,” pp. 1–6. [66] S. Nanz and C. A. Furia, “A comparative study of programming languages in rosetta code,” *Proc. - Int. Conf. Softw. Eng.*, vol. 1, pp. 778–788, 2015.

[67] R. Levy, “Working with n -grams in SRILM,” no. February, pp. 1–6, 2015.

[68] Y. M. Tachbelie and W. Menzel, “Amharic Part-of-Speech Tagger for Factored Language Modeling,” *Proc. Int. Conf. RANLP-2009*, pp. 428–433, 2009.

[69] M. M. Haque, M. T. Habib, and M. M. Rahman, “Automated Word Prediction in Bangla Language Using Stochastic Language Models,” *Int. J. Found. Comput. Sci. Technol.*, 2015.

[70] S. Agarwal and S. Arora, “Context Based Word Prediction for Texting 1 Language, 2007.”

7. APPENDICES

Appendix A: Sample corpus used for training

1. ልዩ/ADJ ዞኑ/N ያስገነባቸው/VREL ፕሮጀክቶች/N 50ሺ/NUMCR ነዋሪዎችን/N ተጠቃሚ/ADJ አደረጉ/V
2. በአዳማ/NP ሌዩ/ADJ ዞን/N በተጠናቀቀው/VP የበጀት/NP አመት/N በ6.8ሚሉየን/NUMP ብር/N ከተጀመሩት/VP 12/NUMCR ፕሮጀክቶች/N መካከሌ/PREP አብዛኞቹ/ADJ ተጠናቅቀው/V 50ሺ/NUMCR ነዋሪዎችን/N ተጠቃሚ/N ማዴረጋቸው/VN ተገሇጸ/V
3. በልዩ/ADJP ዞኑ/N የቴክኒክ/NP አገልግሎትና/NC የከተማ/NP ቦታ/N አስተዳደር/N ክፍል/N ሃላፊ/N አቶ/ADJ ግዛው/N ድንቁ/N ዛሬ/ADV ለዋልታ/NP እንፎርሜሽን/N ማእከል/N እንደገለጹት/VP በልዩ/ADJP ዞኑ/N 13/NUMCR ቀበላዎች/N ተገንብተው/V አገልግሎት/NP የበቁት/VREL የመንገድ/NP የግብርናና/NPC የመሰረታዊ/ADJP ልማት/N ፕሮጀክቶች/N ናቸው/AUX
4. ፕሮጀክቶቹም/NC 25ኪሎ/NUMCR ሜትር/N የውስጥ/NP ውስጥ/NP የጠጠር/NP መንገድ/N 1.5/NUMCR ኪሎ/N ሜትር/N የአስፋልት/NP ግንባታና/NC ትገና/N እንዱሁም/PRONP የአንድ/NUMP ዴሌዴይና/N የአንዴ/NP ስህፈትቤት/N ግንባታን/N ጨምሮ/V 3.8/NUMCR ኪልሜትር/N የጎርፍ/VREL መውረጃ/N ቦይ/N ግንባታና/NC 4.6/NUMCR ኪልሜትር/N የቦይ/NP ጠረጋ/N መሆናቸውን/VN ተናግረዋል/V
5. የተቀሩት/VREL ሶስት/NUMCR የገበያ/NP በረንዳዎች/N ፣/PUNC 12/NUMCR ቁጠባ/N ቤቶች/N እና/CONJ የ1.8/NUMP ኪልሜትር/N የአስፋልት/NP መንገድ/N ስራ/N ሶስትአራተኛ/NUMOR በሙቶ/NUMP የግንባታ/NP ዯረጃ/N ሊይ/PREP እንዯሚገኙና/VPC በተያዘው/VP የበጀት/NP አመት/N አገልግሎት/NP ይበቃል/V ተብለው/V እንደሚጠበቁ/V አስታውቀዋል/V
6. ከልዩ/ADJP ዞኑና/NC ከመንዴ/NP ፈንድ/N ፕሮጀክቶች/NP ማስፈጸሚያ/N ከተመደበው/NP 6ሚሉየን876ሺ/NUMCR ብር/N እስከ/PREP አሁን/N 80/NUMCR በሙቶው/NUMP ስራ/N ሊይ/PREP መዋሉን/NP ሃላፊው/N ገልጸዋል/V
7. አንዳንድ/PRON የ01/NP 02ና/NC 09/N ቀበሌ/N ነዋሪዎች/N በሰጡት/VP አስተያየት/N ልዩ/ADJ ዞኑ/N በተደጋጋሚ/NP የሚከሰተውን/VREL የጎርፍ/NP አዯጋ/N ሇመከላከል/NP ግንባታ/N በማካሄድ/VP ይደርስ/V የነበረው/VREL ችግር/N መቀረፉን/VN ተናግረዋል/VN
8. የአዳማ/NP ልዩ/ADJ ዞን/N በተሇይ/ADV በክረምት/NP ወራት/N በተደጋጋሚ/NP የጎርፍ/NP አዯጋ/N የምትጋሇጥ/VREL መሆኑ/VN ይታወሳል/V

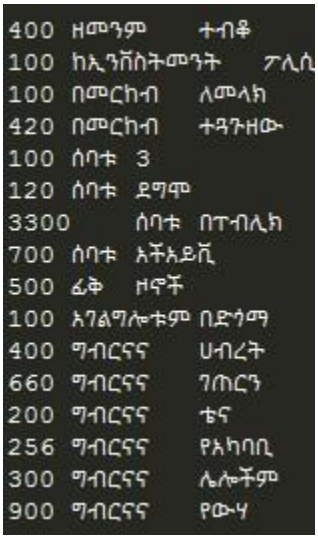
9. ፕሬዚዲንቱ/N ተቅላይ/ADJ ሚኒስትሩና/NC የውጪ/NP ጉዲይ/N ሚኒስትር/N አድጋው/N የኢትዮጵያን/NP ህዝብ/N እንዲሳዘነ/VP ገለጹ/V
10. በአሜሪካ/NP ኒውዮርክና/NP ዋሽንግተን/N ትናንት/ADV የደረሱትን/VREL አራት/NUMCR የአጥፍቶ/VREL መጥፋት/VN የአውሮፕሊን/NP አዩጋዎች/N የኢትዮጵያ/NP ህዝብ/N ያዲመጠው/VREL እጅግ/ADJ በበረታ/VP ሀዘን/N መሆኑን/VN ፕሬዚዲንት/ADJ ነጋሶ/N ጊዲዲ/N ፣/PUNC ተቅላይ/ADJ ሚኒስትር/ADJ መለስ/NC ዜናዊና/NC የውጪ/NP ጉዲይ/N ሚኒስትር/N ስዩም/N መስፍን/N ገለጹ/V
11. የኢትዮጵያ/NP ፌደራሊዊ/ADJ ዳሞክራሲያዊ/ADJ ሪፐብሉክ/N ፕሬዚዲንት/ADJ ነጋሶ/N ጊዲዲና/NC ተቅሊይሚኒስትር/N መሆስ/N ዜናዊ/N በአሜሪካ/NP አምባሲ/N በኩሌ/N ሆአሜሪካ/NP ፕሬዚዲንት/ADJ ጆርጅ/N ቡሽ/N በሊኩት/VP የሀዘን/NP መግሆጫ/N የአሸባሪዎቹን/NP ዴርጊት/N የኢትዮጵያ/NP መንግስትና/NC ህዝብ/N እጅግ/ADJ እንዲወገዘው/VP ገሌጸዋሌ/V ።/PUNC
12. በመሆኑም/NPC በአደጋው/NP ህይወታቸውን/N ላጡ/VP አሜሪካውያንና/NC ሆወዮመው/VP ንብረት/N ሀዘናቸውን/N በራሳቸውና/PRONPC በኢትዮያ/NP ህዝብ/N ስም/N የጉሆጹት/VREL ፕሬዚዲንቱና/ADJC ተቅሊይሚኒስትሩ/ADJ አሸባሪዎችን/N በመዋጋት/NP በኩሌ/N መንግስት/N ከአሜሪካ/NP ጎን/PREP እንዮሚቆም/VP አረጋግጠዋሌ/V

Appendix B: parts of speech tag used in Amharic WSP model

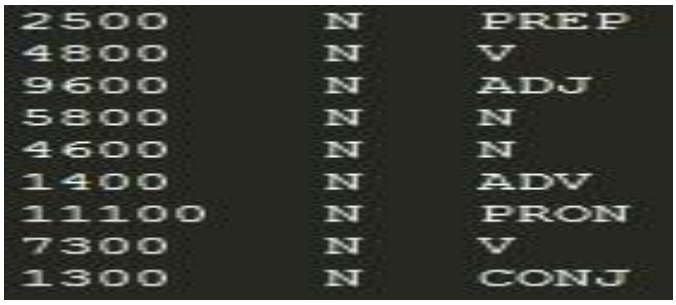
POS tags	Description
ADJ	Adjective
ADJC	Adjective with conjunction
ADJP	Adjectival phrase
ADJPC	Adjectival phrase with conjunction
ADV	Adverb
AUX	Auxiliary
CONJ	conjunction
INT	Interjection
N	Noun
NC	Noun with conjunction
NP	Noun phrase
NPC	Noun phrase with conjunction
NUMCR	Cardinal Number
NUMOR	Ordinal Number
NUMP	Number with phrase
NUMPC	Number phrase with conjunction

PREP	Preposition
PREPC	Preposition with conjunction
PRON	Pronoun
PRONC	Pronoun with conjunction
PRONP	Pronoun phrase
PRONPC	Pronoun phrase with conjunction
PROPNC	Proper Noun with conjunction
PUNC	Punctuation
UNC	Unknown or unclear
V	Verb
VC	Verb with conjunction
VN	Verbal noun
VP	Verb phrase
VPC	Verb phrase with conjunction
VREL	Relative verb

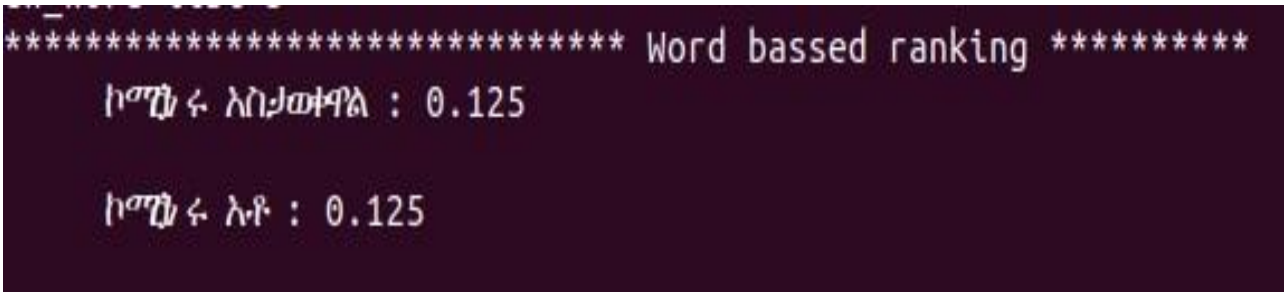
Appendix C: Representation of bigram word sequence frequency in the corpus



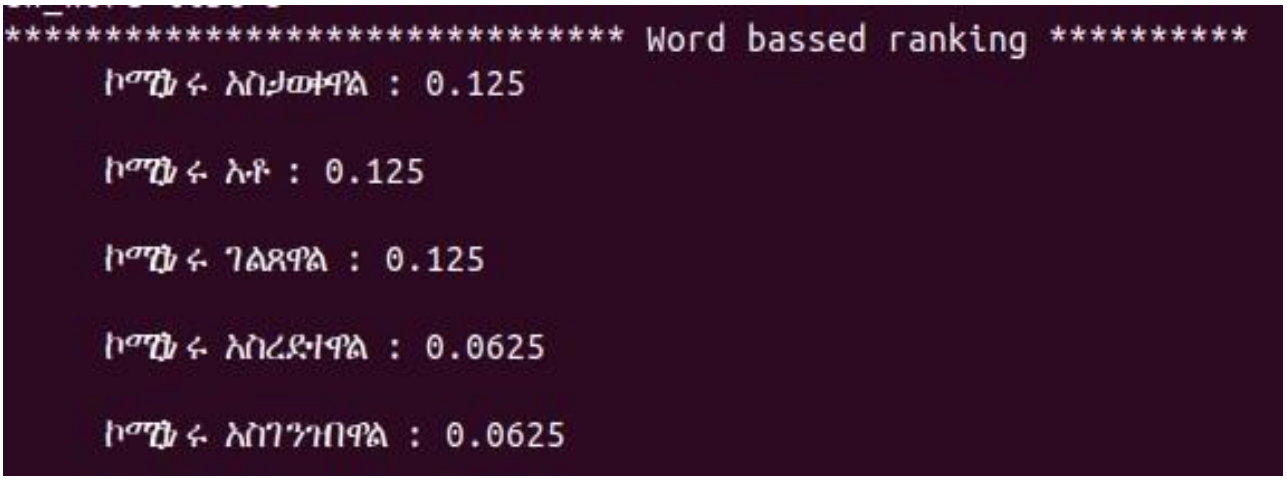
Appendix D: Representation of bigram POS sequence frequency in the corpus



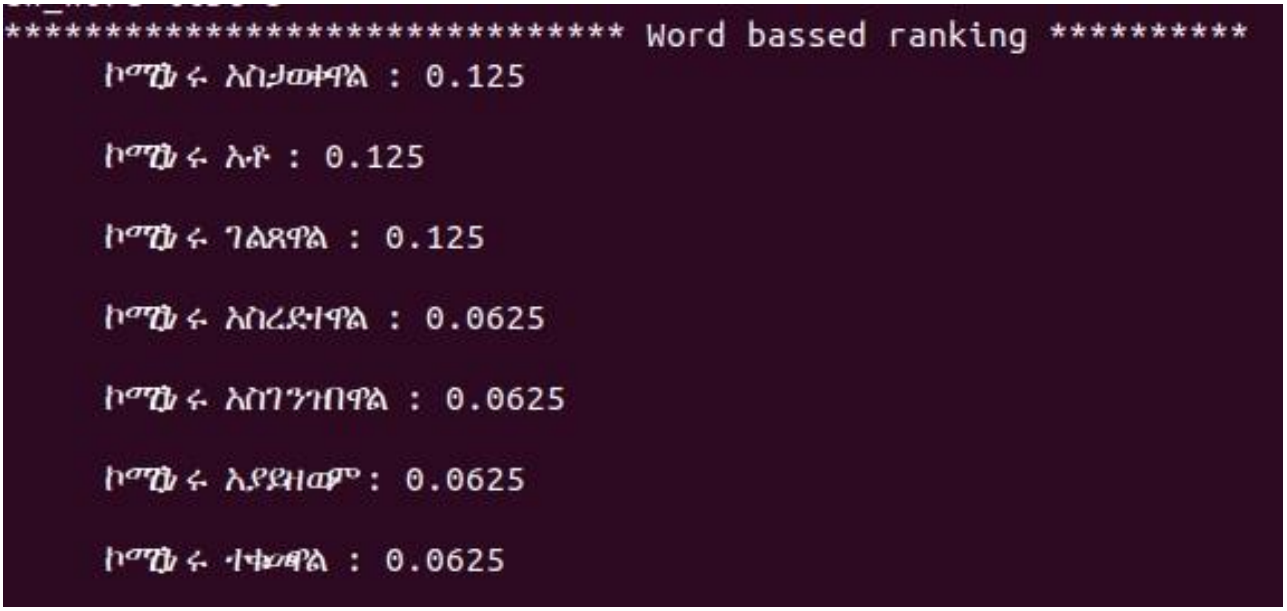
Appendix E: Demonstration of word based model with N=2



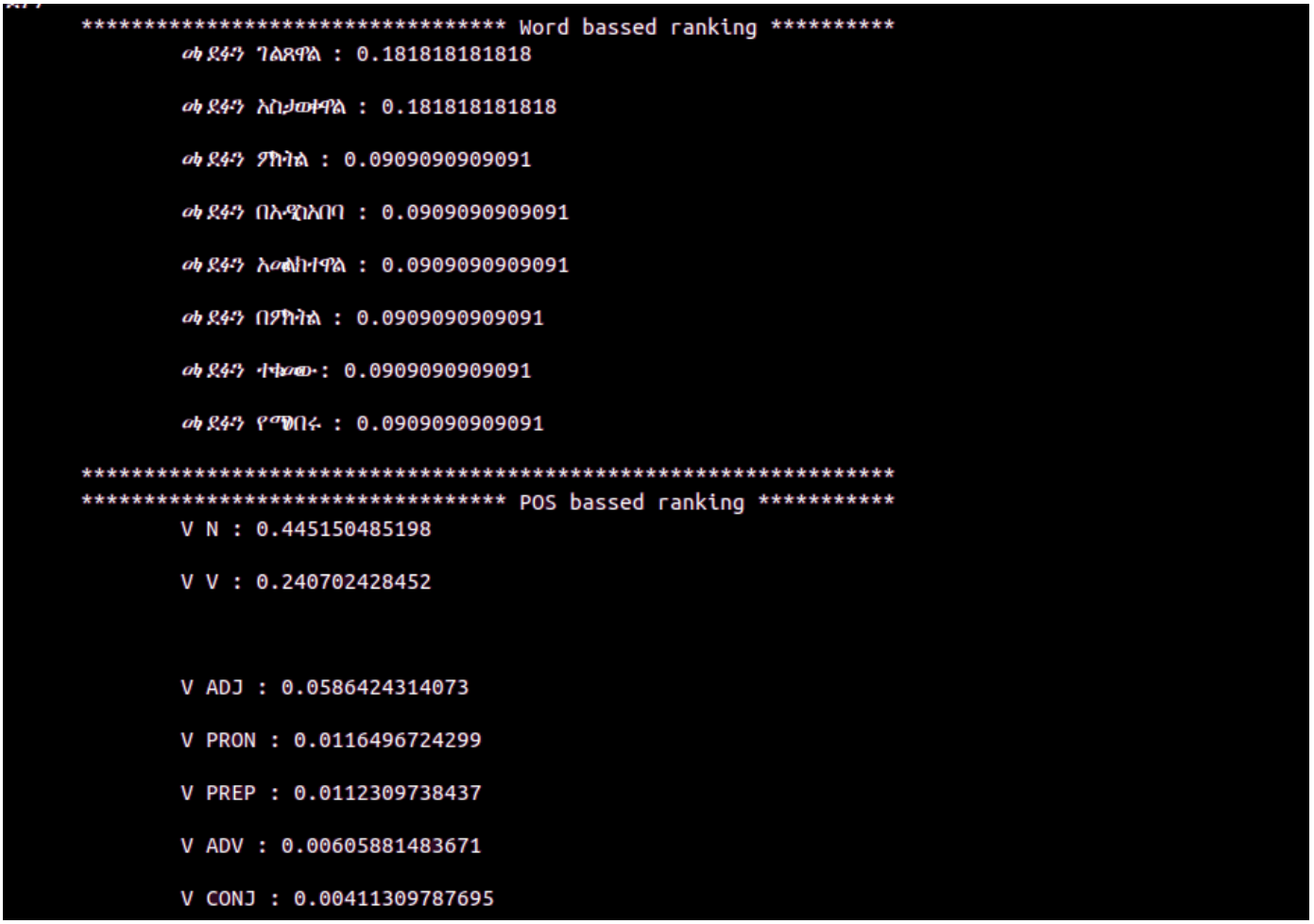
Appendix F: Demonstration of word based model with N=5



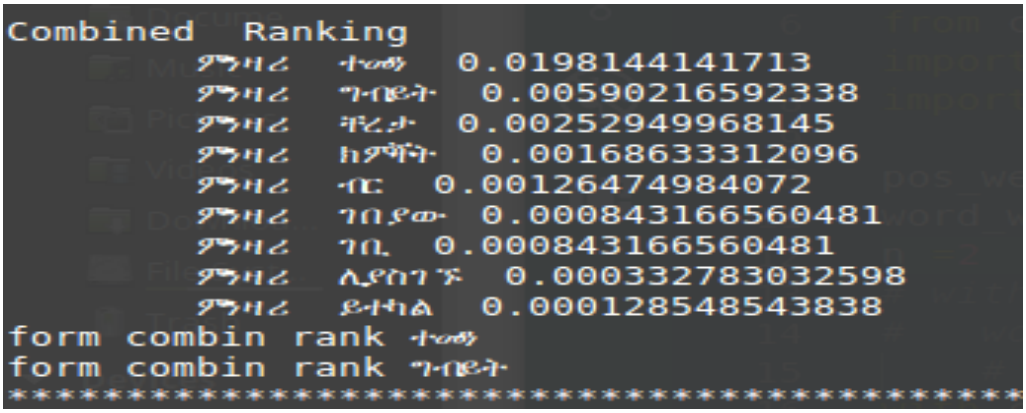
Appendix G: Demonstration of word based model with N=7



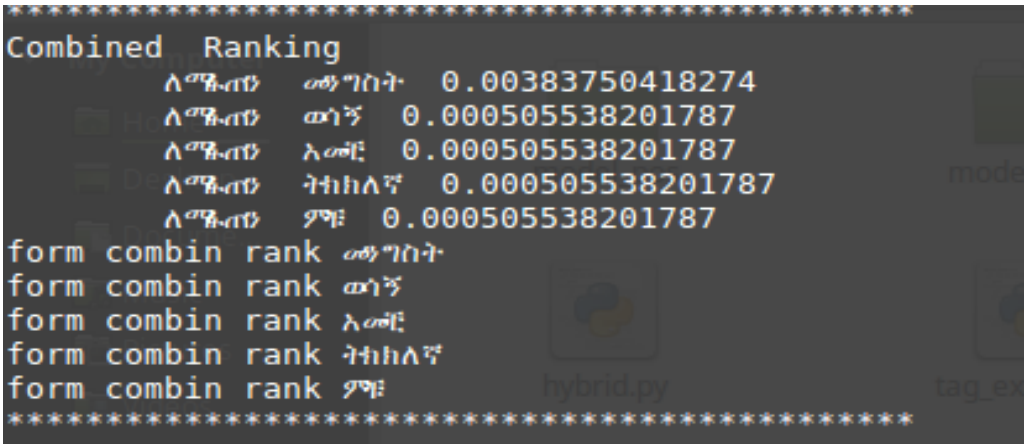
Appendix H: Demonstration of the word and POS model



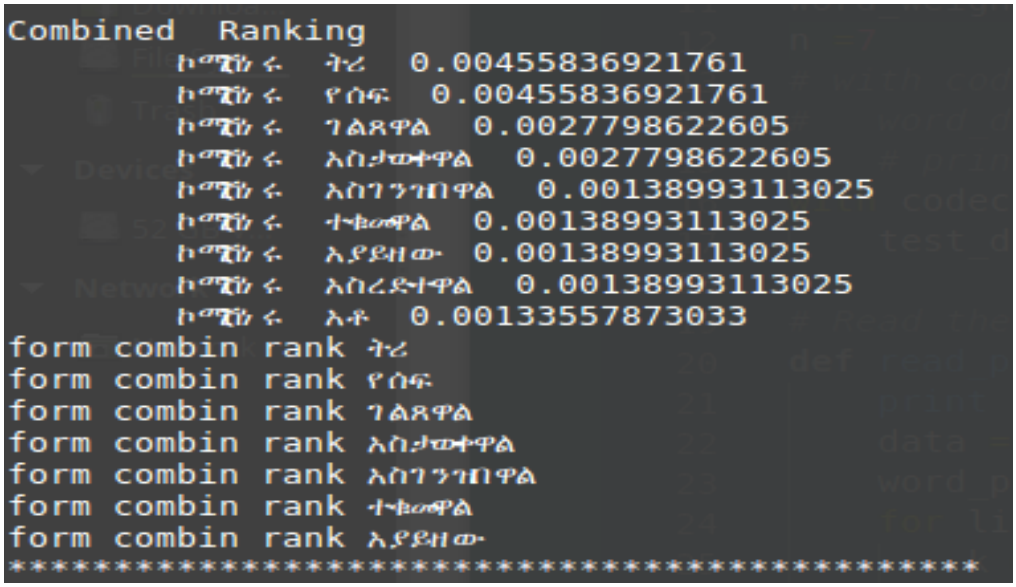
Appendix I: Combined ranking model for N=2



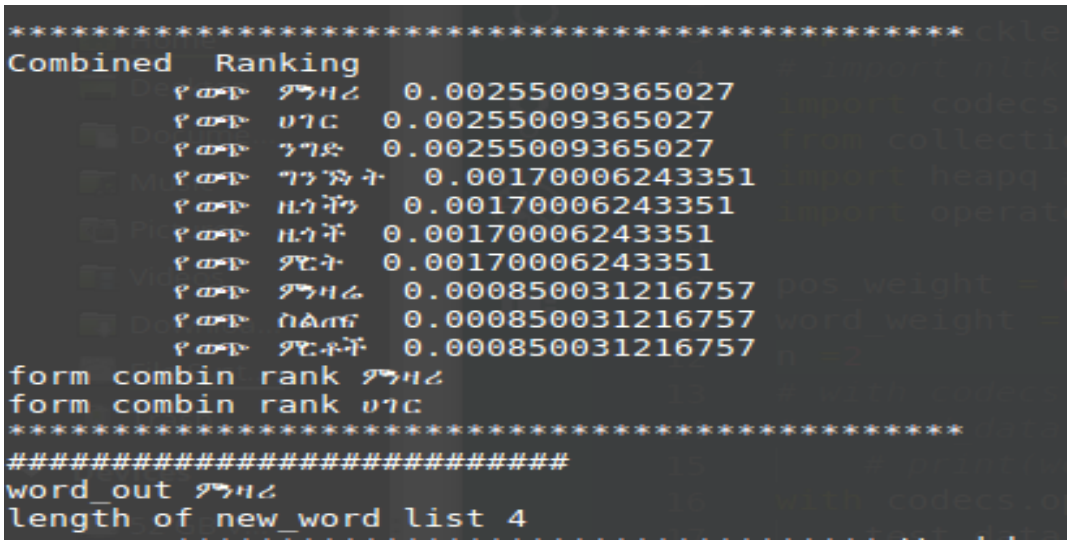
Appendix J: Combined ranking model for N=5



Appendix K: Combined ranking model for N=7



Appendix L: Keystrokes Saved when N=2



Appendix M: Keystrokes not Saved when N=2