ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

Word Sequence Prediction for Amharic Language

Tigist Tensou Tessema

A Thesis Submitted to the school of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science

October 2014

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

Word Sequence Prediction for Amharic Language

Tigist Tensou Tessema

APPROVED BY:
EXAMINING BOARD:

1. Yaregal Assabie (PhD), Advisor _____

2. Mulugeta Libsie (PhD), Examiner _____

3. Fekade Getahun (PhD), Examiner _____

Acknowledgment

First and foremost, I am very thankful to the almighty God for entitling me to this opportunity.

Many Thanks to my advisor, Dr. Yaregal Assabie for his constructive comment, supervision, and patience till the completion of this study. Without your advice and assistance this work will be lacking.

My sincerely gratitude go to Micheal Gassar for his optimistic assistance while using Hornmorph morphological analyzer and generator program. I am also very grateful to Andualem Abate for his cheerful support to manually tag words with their POS in the testing data.

I am very thankful to my families specially my mother, and brothers for their valuable support throughout this study.

Last but not least, my heartfelt thanks go to my class mates, colleagues, and friends for their unlimited encouragement during my study.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

AAC	Augmentative and Alternative Communication
ASCII	American Standard Coding for Information Interchange
CMS	Case Marker Suffix
GMS	Gender Marker Suffix
HR	Hit Rate
IR	Information Retrieval
KE	Effective Number of Keystroke s
KSS	Keystroke Saving
KT	Total Number of Keystroke s
KUC	Keystroke Until Completion
MI	Mutual Information
NMS	Number Marker Suffix
POS	Parts-of- Speech
SMS	Short Message Service
SOV	Subject-Object-Verb
SVM	Support Vector Machine
SVO	Subject-Verb-Object
TC	Text Categorization
WIC	Walta Information Center
WP	Word Prediction
WTS	Word Type Saving

Abstract

The significance of computers and handheld devices are not deniable in the modern world of today. Texts are entered to these devices using word processing programs as well as other techniques. Text prediction is one of the techniques that facilitates data entry to computers and other devices. Predicting words a user intends to type based on context information is the task of word sequence prediction, and it is the main focus of this study. Word prediction can be used as a stepping stone for further researches as well as to support various linguistic applications like handwriting recognition, mobile phone or PDA texting, and assisting people with disabilities.

Even though Amharic is used by a large number of populations, no significant work is done on the topic of word sequence prediction. In this study, Amharic word sequence prediction model is developed using statistical methods and linguistic rules. Statistical models are constructed for root or stem, and morphological properties of words like aspect, voice, tense, and affixes using the training corpus. Consequently, morphological features like gender, number, and person are captured from a user‘s input to ensure grammatical agreements among words. Initially, root or stem words are suggested using root or stem statistical models. Then, morphological features for the suggested root or stem words are predicted using voice, tense, aspect, affixes statistical information and grammatical agreement rules of the language. Predicting morphological features is essential in Amharic because of its high morphological complexity, and this approach is not required in less inflected languages since there is a possibility of storing all word forms in a dictionary. Finally, surface words are generated based on the proposed root or stem words and morphological features.

Evaluation of the model is performed using developed prototype and keystroke savings (KSS) as a metrics. According to our experiment, prediction result using a hybrid of bi-gram and tri-gram model has higher KSS and it is better compared to bi-gram and tri-gram models. Therefore, statistical and linguistic rules have quite good potential on word sequence prediction for Amharic language.

Keywords: *Hornmorph, Keystroke Saving, Natural Language Processing, Parts-of-Speech, Word Prediction*

CHAPTER ONE

INTRODUCTION

1.1 Background

Amharic is a Semitic language of Afro-Asiatic Language Group that is related to Hebrew, Arabic and Syrian. It is a native language for people who live in north-central part of Ethiopia. It is spoken and written as a second language in many parts of the country, especially in urban areas and by significant number of Ethiopians living in the Middle East, Asia, Western Europe, and North America [1]. Next to Arabic it is the second most spoken Semitic language with around 27 million speakers [2, 3].

Ge‘ez alphabet is an ancient language used for liturgy of Ethiopian Orthodox Church and is used as a script for Amharic language. Amharic language has thirty-three basic characters with each having seven forms for each consonant-vowel combination. Among these, twenty-seven have unique sounds, being characterized in terms of their sound creation and their graphic symbols. It is unique to Ethiopia and written from left to right unlike Arabic, Hebrew or Syrian. Manuscripts in Amharic are known from 14^{th} century and the language has been used as a general medium for literature, journalism, education, national business and cross-communication. A wide variety of literature including religious writings, fiction, poetry, plays, and magazines are available in the language [1, 4].

Amharic is an under-resourced African language which has very complex inflectional and derivational verb morphology with four and five possible prefixes and suffixes respectively. It is morphologically complex and makes use of both prefixing and suffixing to create inflectional and derivational word forms which also requires some degree of infixing and vowel elision [1, 2, 4, 5].

So far some researches have been conducted on the language including design and development of Amharic word parser[6], automatic part of speech tagger [7], morphology based language modeling for Amharic [8], automatic morphological analyzer [9], automatic sentence parsing for

Amharic text [10], Amharic speech recognition [11], and stemming [1]. These researches help to obtain a crisp understanding about characteristics of Amharic language in order to incorporate them in this study. As the working language of the Federal Government and some regional governments of Ethiopia most documents in the country are produced in Amharic. There is also enormous production of electronic and online accessible Amharic documents [3]. Amharic texts are usually entered to computers with the assistance of software packages like Power Geez and Visual Geez.

Data entry is a core aspect of human computer interaction. Images, documents, music, and video data are entered to computers in order to get processed. Data entry can be through the use of keyboard, or other means. Text prediction provides better data entry performance by improving the writing mainly for people with disabilities [12, 13].

Text prediction is the task of estimating missing letter, word, or phrase that likely follow a given segment of text. Statistical information which is based on probabilities of isolated or more complex words, syntactic knowledge which considers POS and phrase structure, semantic knowledge which can be used through assignment of categories to words and finding a set of rules that constrain possible candidates for next word are few of the processes to make prediction. Word frequencies can be acquired from a corpus or from the user itself.

A research on Word Prediction for Amharic language using bi-gram model is conducted by Nesredin Suleiman and Solomon Atnafu [14]. The main focus of the work is to complete a word currently being typed by a user. Here, characters are suggested to complete the word using dictionary of words with their frequency. However, such approach has critical limitations for inflected languages [15]. For example, it is not possible to store all word forms in a dictionary, and doesn‘t use context information when predicting words. Due to this, it has high possibility of suggesting syntactically wrong output. In this work, Word Sequence Prediction implies predicting a word a user wants to type based on previous words. Word prediction, word completion, character prediction, letter prediction, text prediction are some of the terminologies used to express similar concepts. Text prediction is one of the most widely used techniques to enhance communication rate in augmentative and alternative communication. However, due to the absence of Word sequence prediction for Amharic language; it lacks core benefits of word sequence prediction.

1.2 Motivation

There are various word prediction software packages to assist users on their text entry. Swedish [26, 37], English [38], Italian [18, 19], Persian [20] are some of word prediction studies conducted lately. These studies contribute in reducing the time and effort to write a text for slow typists, or people who are not able to use a conventional keyboard.

In Ethiopia usage of computers and different handheld devices are growing from day to day. However, most software programs used with these devices are in English. On the contrary, a great number of people in Ethiopia communicate only using Amharic language. With this in mind, having alternative or assistive Amharic text entry system is useful to speed up text entry, and helps those needing alternative communication. Hence, in this study we will focus on word sequence prediction to address this issue. Morphological characteristics of Amharic language are a major challenge for most researches. In the case of non-inflected languages or less inflected languages, possible word forms can be stored in a lexicon since word forms are not vast like non-inflected languages. Hence, word sequence prediction program can use stored lexicon without any complications. However, languages like Amharic have enormous inflection possibilities and it is impossible to capture all word forms and store it in a lexicon.

For example: If we look a simple Amharic verb: ~~አንመጣም~~"/" 'anmeTam " which is equivalent to the English sentence ~~W~~e will not come", it is an aggregate of root or stem: ~~መጣ~~"/"meTa", prefix: ~~አን~~"/"an" and suffix: ~~ም~~"/"m" as shown in Figure 1.1. The affixes give additional meanings to root or stem of the word, which can be gender, number, case, person or other information.

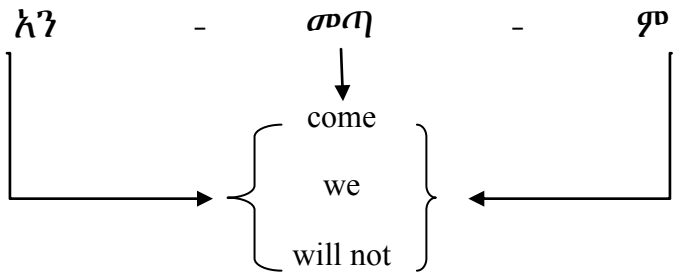

Figure 1.1: Morphemes of Amharic Verb

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

1.3 Statement of the Problem

In this work, word sequence prediction generally refers to the task of suggesting a word a user intends to type based on a given segment of text. In Amharic, a research has been done to complete a word a user is currently typing using dictionary of words with their frequency. One of the drawback 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. These problems are not addressed in the existing study and needs further research to support users of the language on their text entry techniques.

Implementation of word sequence prediction for one language has enormous advantages. This includes auto completion, mobile phone or PDA texting, handwriting recognition, speech recognition and communication aids. The purpose of this research is to design and develop Amharic word sequence prediction model with the aim of predicting appropriate word forms by considering context information. Furthermore, this study will be a stepping stone for further researches that can bring the aforementioned advantages for the language.

1.4 Objectives

The general objective of this research is to design and develop word sequence prediction model for Amharic language.

To achieve the above mentioned general objective the following specific objectives will be performed.

- Conduct literature review on word prediction, approaches to word prediction and structure of Amharic language.
- Review related works on word sequence prediction for other languages and supplementary researches conducted on Amharic language with the aim to find the best approach for this study.
- Collect representative corpus for training and testing the model.
- Morphologically analyze the training corpus.
- Construct root or stem word, affixes, aspect, tense and voice tagged corpus.
- Build language models of root or stem word sequences, root or stem with affixes, root or stem with aspect, root or stem with voice and root or stem with tense.
- Develop a prototype.
- Evaluate the performance of the word sequence prediction model using collected test data.

1.5 Methodology

1.5.1 Literature Review

Researches and related works will be thoroughly reviewed to grasp a firm knowledge with the intention of developing appropriate word sequence prediction model for Amharic. Word sequence prediction, word sequence prediction approaches, Amharic grammar and morphology, are some of the works that will be reviewed while conducting this research.

1.5.2 Document Collection

A training corpus containing 298,500 sentences which is equivalent to 125 MB will be used to train the Amharic word sequence predictor. In addition POS tagged corpus containing 8067 sentences will be used to extract representative sentences for testing by means of random sampling method. Simple random sampling method is chosen since every sentence has equal chance of being selected. These corpora are collected from Walta Information Center (WIC) in previous linguistic studies.

1.5.3 Tools

Hornmorph morphological analyzer and generator program will be used to analyze the corpus and to produce surface words. Moreover, Python programming language will be used to develop prototype for demonstration.

1.5.4 Prototype Development

To develop prototype, supporting tools are required. Hence, python programming language and Hornmorph morphological analyzer and generator program will be used. As previously stated, Hornmorph will be used to morphologically analyze collected training corpus. It will also be used to morphologically analyze user entered texts from the testing data, so that required features like gender, number, and person will be captured and used to generate proposed words in correct grammatical form.

Python programming language will be used to implement statistical language models (tri-gram, bi-gram, and hybrid). As part of the prototype development a user interface will be designed that allows users to type their text and choose from the list of suggested words.

1.5.5 Evaluation

Prototype development is one of the objectives of this study in order to demonstrate and evaluate the developed model. POS tagged test data will be used and the prediction activity is evaluated through calculation of keystroke savings. A Keystroke Saving (KSS) estimates saved effort percentage and is calculated through comparison of total number of keystrokes needed to type a text (KT) and effective number of keystrokes using word prediction (KE) [19, 21, 22]. Hence,

$$KSS = \frac{KT - KE}{KT} * 100 \qquad \text{(Eq.1)}$$

Therefore, the number of keystrokes to type texts taken from the test data with and without word sequence prediction program will be counted to calculate keystroke savings accordingly. The obtained KSS will be compared for tri-gram, bi-gram and hybrid models. The model that shows maximum keystroke saving is considered as better model.

1.6 Scope and Limitations

This research will be undertaken with the aim to model word sequence prediction for Amharic language based on statistical methods and grammatical agreement rules of the language. Statistical models of root or stem, affixes, aspect, tense, voice and rules of the language like subject-object-verb, adjective-noun, and adverb-verb agreement will be incorporated in this work. This research will not deal with errors either in the training corpus or the output of Hornmorph program while building the language model.

1.7 Application of Results

Word sequence prediction benefits people with severe motor and oral disabilities, on handwriting recognition, mobile phone or PDA texting, etc. Therefore, the model will be helpful to develop different applications and hence users of this language can gain the abovementioned benefits. Furthermore, it will support researchers to use important features of the developed word sequence prediction model for more NLP studies like speech recognition, handwriting recognition and more.

1.8 Organization of the Thesis

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

CHAPTER TWO

LITRATURE REVIEW

This Chapter discusses fundamental concepts of word sequence prediction and ideas associated with Amharic language. Prediction methods like statistical, knowledge based, and heuristics are presented in order to grasp clear overview of the topic. The main target of this study is to design and develop word sequence prediction model for Amharic language. Hence, morphological characteristics, grammatical properties, and parts-of-speech of the language are discussed in respective sections of this chapter.

2.1 Word Prediction

In humans‘ day to day life, massive amount of text and other documents are produced electronically and due to this, computers and technologies are becoming an essential part of day to day activities for most people. It has been more than a decade since people started processing Amharic documents using computers. Because of this, more and more documents, information and databases are being produced and are available in electronic form [4]. Texts can be entered to computers through the use of keyboard, or other techniques. Text prediction provides better data entry performance by improving the writing mainly for people with disabilities [12, 13]. Improving and enhancing text entry and interaction with computers for disabled users had been investigated for many years and many systems are proposed to facilitate and simplify text input process [23].

Dictionaries define prediction as an act of forecasting a thing with present or past experience. In natural language processing the task of prediction is to guess missing letter, word or phrase that likely follow a given segment of text. Different terminologies like text prediction, word prediction, and word completion have been used to express similar and related concepts. Predictors are those systems that display a list of most likely letters, words, or phrases for current position of a sentence [15, 24, 25]. Word prediction software is a writing support where at each keystroke, it suggests a list of meaningful predictions, amongst which a user can possibly identify a word he or she is willing to type. A user picks a preferred word from list of proposals,

and then the software will automatically complete a word being written, thus saving keystrokes [24].

In the context of assistive communication, a predictor is a system which tries to anticipate next block of characters a user wants to express based on previously produced blocks. These blocks can be letters, syllables, words, phrases, etc. and its core point is to reduce effort and message composition time. Reducing number of keystrokes needed for composing a message is a major issue to ease the effort. The number of characters included into text as a result of single prediction should be larger than the number of characters written by a single selection to reduce the needed time [13].

Word prediction is very helpful to enter utterances spoken in Augmentative and Alternative Communication (AAC) device to speed up text entry. It deals with the next word or words a user wants to write by offering a list of possible options and it is especially useful for movement impaired users who need help writing very common social phrases fast and often [15, 16, 22, 26]. The thought of automatic completion or auto completion has become increasingly pervasive. Based on current input of users, an auto completion mechanism discreetly prompts a user with set of suggestions, and this helps to avoid unnecessary typing, to save time and reduce user‘s cognitive burden [27].

The main purpose of word prediction software is to speed up text entry in different kinds of applications through minimum keystrokes. It can also be effectively used in language learning by means of suggesting correct words to non-native users and reducing misspellings for users having limited language proficiency. In augmentative and alternative communication, there is a need to apply different techniques to augment communication rate, and text prediction is one of the most widely used techniques [13].

The major issues in the development of word prediction systems include prediction methods and user interface issue. Prediction methods include decisions on prediction units (characters, words), information sources and structure (both lexical and statistical), levels of linguistic processing, size and type of corpora and learning methods [16]. Word prediction is facing a very ambitious challenge, as the inherent amount of arising ambiguities (lexical, structural, semantic, pragmatic,

cultural and phonetic ambiguities for speech) is complex problem to be solved by a computer [24].

Word prediction and text input methods have been studied for diverse languages using different approaches like statistical as well as linguistic rules.

2.2 Approaches to Word Prediction

The methods for word prediction 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. Word frequency and word sequence frequency are the methods that are commonly used in prediction systems, especially for those developed commercially [15]. All prediction methods require lexical data that can be acquired from corpora along with word frequencies and lexical databases. Garey-Vitoria [13] presented a survey on text prediction techniques to provide systematic view of the topic.

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. 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 last n-1 word 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 written with the help of word prediction system is of a different style than the training data [16].

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. In the past, various studies are conducted to develop systems that consider previous history of words based on bi-gram or tri-gram model [15]. Although statistical techniques can be robust in

computing the suggestions in word prediction, machine learning can assist in re-ranking and reducing the number of suggestions [15, 23].

Statistical Word Prediction using Frequency

Building a dictionary containing words and their relative frequency of occurrence is the simplest word prediction method. It provides n most frequent words beginning by this string in the same way they are stored in the system. This method may need some correction by a user in order to adjust its concordance when applied to inflected words since context information are not considered. In other words this method uses unigram model with a fixed lexicon and it came up with the same suggestion for similar sequences of letters. To enhance accuracy of word prediction result, indication about recency of use of each word may be included in the lexicon. In this way, the prediction system is able to offer most recently used words among most probable words. Adaptation of each word to a user‘s vocabulary is possible by updating frequency and recency of each word used [15, 17].

Most probable words beginning with the same characters are offered when a user has written the beginning of a word. If the required word is not available among options offered by the system, a user may continue writing, else the required word is accepted from the given list and it may automatically adapt to user‘s lexicon by means of simply updating frequencies of words used and assigning an initial frequency for new words added to the system. In order to enhance the outcome of this approach, recency field is stored in a dictionary with each word and frequency information. Results obtained with recency and frequency based methods are better than the ones based on frequency alone. However, this method requires storage of more information and increases computational complexity [13, 17].

Statistical Word Prediction using Word Probability Tables

Prediction using word probability tables consider probability of appearance of each word after the one previously composed. This method builds a two dimensional table, where conditional probability of word Wj after word Wi is stored. Therefore, if the system has N words, there are N^2 entries in this table, where most of them are zero or nearly zero. By using this strategy, the system offers predictions before a user starts writing the initial character of a word and these

results may be improved via integration of recency. This method is based on restricted vocabulary size and one of its challenges is difficulty of adaptation to user‘s vocabulary [13, 17].

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 decrease in writing rate. Syntactic, semantic and pragmatic linguistic knowledge can be used in prediction systems.

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 [15].

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. Probability of appearance of the categories after each category is stored in two dimensional table stores. This table is much smaller than the one presented in frequency based approach and the number of probabilities which are nearly zero is also lower. The probabilities of table and frequencies in lexicon can be updated for adaptation of these systems [13, 15].

Syntactic prediction using grammars analyzes sentences using grammars either top-down or bottom-up, and natural language processing techniques are applied in order to obtain categories having highest probability of appearance. Each natural language has a set of syntactic rules which usually have right to left structure. The sequence that occurs in right category helps to decompose categories in left part of the rule. All categories are defined in the system if at least one category has to happen in right side of arrow. Among categories on right side of a rule, it is possible to define a number of morphological agreement constraints. So that, proposals offered by the predictor are in appropriate morphological characteristics. The dictionary requires inclusion of morphological information in order to enforce morphological agreement. These systems have a higher computational complexity than the previous ones, mainly due to the fact that they take the entire beginning of a sentence into account (while previous systems take, at most, last entirely composed word). Word probabilities and weights of syntactic rules can be updated to adapt these types of systems [13, 15, 17].

Semantic Knowledge for Word Prediction

Semantic prediction is to semantically analyze 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 [13, 17].

In semantic word prediction, Lexical source and Lexical chain are two methods that are used. The first method is lexical source, like WordNet in English, which 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 [15].

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 [15].

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 [15].

Comparison of word prediction approaches is presented in Table 2.1.

Table 2.1: Comparison of Word Prediction Approaches

Word Prediction Approaches		**Knowledge Representation**	**Strength**	**Weakness**
Statistical	Frequency Based	A dictionary containing words and their relative frequency.	✓ Simplicity ✓ Good for non-inflected languages.	✓ It doesn‘t consider context information. ✓ Cause extra load on a user in order to adjust concordance when applied to inflected languages.
	Probability Table	A dictionary containing probability of appearance of each word after the one previously composed.	✓ Offers a word before a user starts typing the first character of a word.	✓ Can offer a word syntactically, semantically or pragmatically wrong output.
Knowledge Based	Syntactic Knowledge	Probability of appearance of each word, sequence of syntactic categories and POS tags or grammatical rules.	✓ Considers relative probability of appearance and provide better result than pure frequency based.	✓ POS tagger and Lemmatizer are not available in all languages. ✓ High computational complexity when applying rule based since it considers the entire beginning of a sentence.
	Semantic Knowledge	Words with their associated semantic category. Lexical source, Lexical chain	✓ Provide better result than pure frequency based.	✓ Higher complexity and provides similar result with Syntactic knowledge

Word Prediction Approaches		**Knowledge Representation**	**Strength**	**Weakness**
Knowledge Based	Semantic Knowledge			✓ Difficult to implement in real time system. ✓ Slow in making predictions.
	Pragmatic Knowledge	Words tagged with their pragmatic knowledge.	✓ Increase in accuracy since it filters words that are wrong in discourse.	✓ Increase in complexity as pragmatic knowledge is added.
Heuristic		Recency , topic, trigger and target, and n-gram cache	✓ Considers user‘s preference and enhance prediction output. ✓ Reduces cognitive load.	

2.3 Evaluation of word prediction systems

Keystroke Saving (KSS) is primarily used evaluation means in word prediction. The common trend in research is to simulate a ―perfect" user that will never make typing mistakes and will select a word from the predictions as soon as it appears [18, 22]. A Keystroke Saving (KSS) estimates saved effort percentage in keys pressed compared to letter-by-letter text entry and it is calculated using (Eq.1) [19, 22].

Keystrokes Until Completion (KUC) is another metrics to evaluate word prediction systems where, $c_1 \ldots c_n$ being number of keystrokes for each of the n words before the desired suggestion appears in the prediction list [18]. It is the average number of keystrokes that a user enters for each word before it appears in the suggestion list [28]. Lower value of KUC shows better performance. KUC is computed using (Eq.2).

$$KUC = \frac{C_1 + C_2 \ldots + C_n}{n} X100\% \qquad \text{(Eq.2)}$$

Hit Rate (HR) is an additional word sequence prediction measuring metrics. It is the percentage of times that the intended word appears in the suggestion list and if its hit rate is high as the required number of selections decreases, the predictor is considered to have better performance [16, 28].

Accuracy is the percentage of words successfully completed by a word prediction system before a user reaches the end of a word. It is the ratio of words correctly guessed to total words guessed. A system that completes words in early stages of typing is considered to have better performance [28].

Perplexity is a means of measuring how well something is predicted and it computes average size of the word set over correctly recognized words. A model having low perplexity value is considered the best one and it is defined as 2 to the power of entropy, where entropy measures uncertainty of information content.

The existing word prediction work in Amharic is evaluated based on Accuracy. Furthermore, a number of researches on word prediction use Keystroke Saving (KSS) as a primary evaluation metrics [9, 11, 12, 13, 30]. Considering this fact, we have selected KSS to evaluate our word sequence prediction model. Other evaluation metrics are suggested to be incorporated in future works.

2.4 Structure of Amharic Language

Phoneme, morpheme, root and stem are word units of Amharic language where phoneme represents a basic sound or unit of sound. A phoneme is every glyph or consonant form and morpheme is the smallest meaningful unit in a word which is a phoneme or collection of phonemes. Morpheme can be free or bound, where a free morpheme can stand as a word on its own whereas a bound morpheme cannot. An Amharic root is a sequence of consonants and is the basis for the derivation of verbs. On the other hand, a stem is a consonant or consonant-vowel sequence which can be free or bound where a free stem can stand as a word on its own whereas a bound stem has a bound morpheme affixed to it. A word, which can be as simple as a single morpheme or can contain several of them is formed from a collection of phonemes or sounds [1].

2.4.1 Amharic Parts-of-Speech

Parts-of-speech are particular classes of a word in a text or corpus. POS tagging is one of the important applications of natural language processing. POS tagger is an application which helps to assign words to their appropriate word class like noun, adjective, verb, etc. In many word prediction studies [16, 18, 19, 20, 33, 38], POS tagging and POS n-gram models are used to optimize word prediction task.

In Amharic free morphemes and words are generally categorized in different word classes. The common word classes or part of speech (POS) are noun, pronoun, adjective, verb, adverb, conjunction, and preposition. Part of speech tagging is a task of assigning an appropriate word class for each token in a text.

Nouns

Nouns are free morphemes which are used to identify different objects, like names of a person, place and things. Amharic nouns have the possibility to have up to two prefix and four suffixes for each stem. Nouns can be identified based on their suffix, since they can be suffixed by bound morphemes like ‒ኤ"/"'E", ‒ኡ"/"'u", ‒ኦች"/"'oc", ‒ዎች"/"woc". Gender, number, and case marker suffixes are used in inflection of nouns, so that a noun is of the form stem+{GMS,NMS,CMS,DMS},where GMS, NMS, CMS, DMS are gender marker suffix, number marker suffix, case marker suffix, and definiteness marker suffix respectively [1, 29]. Table 2.2 shows examples of gender, number and case marker suffixes for Amharic nouns.

Table 2.2: Examples of gender, number, and case marker suffixes for Amharic nouns

	Gender marker		**Number marker**		**Case marker**	
Word	**Masculine**	**Feminine**	**Singular**	**Plural**	**Nominative**	**Accusative**
‒ልጅ"/ "lj"	‒ልጅ"/ "lj"	‒ልጅ-ኢት"/ "lj-'it"	‒ልጅ"/ "lj"	‒ልጅ-ኦች"/ "lj-_oc'	‒ልጅ"/ "lj"	‒ልጅ-ን"/ "lj-n"
‒ቢግ"/ "beg"	‒ቢግ"/ "beg"	‒ቢግ-ኢት"/ "beg-it"	‒ቢግ"/ "beg"	‒ቢግ-ኦች"/ "beg-_œ"	‒ቢግ"/ "beg"	በግ-ን/ "beg-n"

Pronouns

Pronouns are words or morphemes that can be used in place of nouns. Pronouns are limited in number and they can be categorized in different sub categories like personal, possessive, interrogative, demonstrative pronouns and the like. Personal pronoun represents speaker, listener and third party in any speech. It is classified based on gender, number, person and can be prefixed with prefix "እነ‒[29]. Table 2.3 shows list of representative pronouns based on person, number and gender.

Table 2.3: List of Representative Pronouns

Person	Gender	Singular	Plural
1st		‹እኔ›/’’’nE”	‹እኛ›/’’‘Na”
2nd	Masculine	‹አንተ›/’’‘ante”	‹አናንተ›/’’’nante”
	Feminine	‹አንቺ›/’’’anci”	
3rd	Maculine	‹እሱ›/’’’su”	‹እነሱ›/’’’nesu”
	Feminine	‹እሷ›/’’’sWa”	
	Polite	‹ርስዎ›/’’’rswo”, ‹አንቱ›/’’‘antu”	

Reflexive pronouns are words that are used combined with representative pronoun [29].

Examples:

Singular: ”እኔ ራሴ”/’’’nE rasE”, ‹አንተ ራስህ›/’’’ante rash”, ‹አንቺ ራስሽ›/’’’anci rasx”

Plural: ‹እናንተ ራሳችሁ›/’’’enante rasachu”, ‹እነሱ ራሳቸው›/’’’nesu rasacew”

Demonestrative pronouns indicate objects in reference to the place it is found. The indicated object can be found near or far from a person indicating the object or for the observant. Therefore this kind of pronouns are classified based on their distance as well as based on the indicated objects gender [29]. Table 2.4 shows examples of demonestrative pronouns.

Table 2.4: Examples of Amharic Demonstrative Pronouns

Number, Gender		Near	Far
Singular	Masculine	‹ይህ›/’’yh”	‹ያ›/’’ya”
	Feminine	‹ይቺ›/’’yci” ‹ይህች›/’’yhc”	‹ያቺ›/’’yaci”
Plural		‹እነዚህ›/’’’nezih”	‹እነዚያ›/’’’nziya”

Interrogative pronouns are used when we need to ask questions about something or some one. In English there are pronouns like who, when, what, where, whose and whom that are used to ask questions. Table 2.5 shows list of interrogative pronouns in Amharic which are classified based on the kind of questions to be asked [29].

Table 2.5: Examples of Amharic Interrogative Pronouns

	For person	**For things**	**For place**	**For time**	**For condition**	**For reasoning**
	~~ማን~~”/ ”man”	~~ምን~~”/ ”mn”	~~የት~~”/ ”yet”	~~መቼ~~”/ ”mecE”	~~እንዴት~~”/ ”‘ndEt”	~~ለምን~~”/”lemn ”
Singular	~~ማንን~~”/ ”mann”	~~ምንን~~”/ ”mnn”	~~ወዴት~~”/ ”wedEt”			
	~~የማን~~”/ ”yeman”	~~የምን~~”/ ”yemn”	~~ከየት~~”/ ”keyet”			
	~~እነ-ማን~~”/ ”‘ne-man”					
Plural	~~እነ-ማንን~~”/ ”‘ne-mann”					
	~~እነ-የማን~~”/ ”‘ne-yeman”					
Negation	~~ማንም~~”/ ”manm”	~~ምንም~~”/ ”mnm”	~~የትም~~”/ ”yetm”			
	~~የማንም~~”/ ”yemanm”					

Possesive pronouns are used to show possesion of something and they are formed with the addition of prefix ”የ”/ ~~y~~e” on personal pronouns [29].

Examples:

Singular**:** ~~የእኔ~~”/”ye‘nE”, ~~የአንተ~~”/”ye‘ante”, ~~የእሱ~~”/”ye’su”, etc.

Plural: ”የእኛ”/”ye‘Na”, “የእናንተ”/”ye‘nante”, etc.

Verb

Verb can be described as a word used to show that an action is taking place, a word to indicate the existence of a state or condition. Amharic verbs are very complex consisting of a stem and up to four prefixes and four suffixes and are inflected for person, gender, number, and time with the basic verb form being third person masculine singular. Verbs in passive voice are marked by suffixes that depend on person and number [3, 29, 30].

Adjectives

Adjective is a word that describes or qualifies a noun or pronoun and it appears before a word it modifies. It gives more information about noun or pronoun it modifies. Objects are differentiated from one another by different attributes like shape, behavior, color, etc. and this difference is described using adjective word class. Adjectives are inflected for gender, number and case in a similar fashion to nouns [1, 29].

Examples:

“ነጭ ዶሮ” /” neC doro” /”white hen”

“ጎበዝ ተማሪ”/ “gobez temari” / “clever student”

In the first example above, the word” ነጭ”/”neC”/”white” is an adjective that modifies the noun “ዶሮ”/”doro"/”hen”, it gives more information about color of the hen. In the second example, the word “ጎበዝ”/”gobez”/”clever” is an adjective that qualifies noun “ተማሪ”/”temari”/”student”, it gives more information about the student, which is clever.

Adverb

Similar to adjectives which qualify nouns, adverb is a word that modifies a verb. Adverbs can be classified based on time, place, circumstances, etc. [29].

Example:

In the sentence, ~~ልጁ~~ በፍጥነት መጣ"/"lju befTnet meTa"/"The boy came quickly", the word "በፍጥነት"/"befTnet"/" quickly" is an adverb that modifies the main verb "መጣ"/"meTa"/"came". It tells more about how the boy came, which is quickly.

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.

Example:"እና "/" 'na", ~~ስለሆነም~~"/"slehonem", ~~ነገር ግን~~"/"negergn", etc.

Preposition

Prepositions are words that are usually used before nouns to show their relation to another part of a clause and they are limited in number. The following are examples of prepositions, and list of prepositions with their probability of occurrence in the training corpus are extracted and shown in Annex 3.

Examples: ~~ለ~~"/"le", እንደ/"'nde", ~~ከ~~"/"ke", etc.

2.4.2 Amharic Morphology

Dictionaries define morphology as the structure of words in a language including patterns of inflections and derivations. Morpheme is the minimal unit of morphology which includes root or stem form and other meaningful parts of a word [1, 3, 29, 31].

For example, the word ~~ወሰደች~~"/"wesedec" has morphemes ~~ወሰደ~~"/"wesede~~ac~~d" -~~ች~~"/" -c", which stands for root or stem word and other meaningful piece of the word respectively.

Morphological analysis is one of the fundamental computational tasks for a language, where its goal is to derive root and grammatical properties of a word based on the internal structure. Morphological analysis, especially for complex languages like Amharic, is vital for the development and application of many practical natural language processing systems such as

machine readable dictionaries, machine translation, information retrieval, spell checkers, and speech recognition [1, 3].

The morphological analyzer takes a string of morphemes as an input and gives an output of lexical forms which is underlying morphemes and morph-syntactic categories. Amharic has a rich verb morphology which is based on tri-consonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. A significantly large part of vocabulary consists of verbs, which exhibit different morph-syntactic properties based on the arrangement of consonant-vowel patterns [32].

Morphological analysis can be performed by applying language specific rules, which may include a full scale morphological analysis with contextual disambiguation, or when such resources are not available, simple heuristic rules, such as regarding the last few characters of a word as its morphological suffix [33]. Inflectional and derivational affixes are removed to identify a word stem from full word. It is very helpful for various NLP applications like textual IR, text summarization, word prediction, etc.

Morphological analysis is segmentation of words in to their component morphemes and assignment of grammatical morphemes to grammatical categories and lexical morphemes to a lexical level, where as morphological generation is the reverse process. Both processes relate a surface level to a lexical level and relationship between the levels has traditionally been viewed with in linguistics in terms of an ordered series of phonological rules [34].

Inflectional Morphology

Nouns, verbs, and adjectives can be marked for person, gender, number, case, definiteness, and time. Gender, number and case marker suffixes are used in inflection of nouns. Verbs are inflected for person, gender, number, and time with the basic verb form being third person, masculine, and singular. The perfect tense normally expresses past tense. Prefixes are used for first, second, and third person future forms and suffixes are used to indicate masculine and feminine subjects, respectively. Adjectives are inflected for gender, number, and case in a similar fashion to nouns [1, 29].

Affixing is used to derive nouns by adding prefixes, infixes or suffixes to basic nouns, adjectives, verbs, stems and roots. In Amharic morphemes can be free or bound; where free morphemes can give complete meaning by themselves whereas bound morphemes need to be attached with free morphemes in order to be meaningful.

Examples:

Free	**Bound**	**Free + Bound**
–ላም”/”lam”	—ኤ”/”-_E’	–ላሜ”/”lamE”
–ወንድም”/”wendm”	—ህ”/”-h”	–ወንድምህ”/”wendmh”

Derivational Morphology

Nouns can be derived by adding prefixes, infixes or suffixes to basic nouns, adjectives, verbs, stems and roots. Adjectives are derived from verbs, nouns, verbal roots, and stems by adding suffixes. Infixing is used when deriving adjectives from verbal roots and unlike other word categories, the derivation of verbs from other POS is not common [1]. Nouns, verbs and adjectives can be marked for person, gender, number, case, definiteness, and time [29].

Amharic has a rich verb morphology which is based on tri-consonantal roots with vowel variants describing modifications to, or supplementary detail and variants of root form. A significantly large part of the vocabulary consists of verbs, which exhibit different morph-syntactic properties based on arrangement of consonant vowel patterns. Amharic nouns can be inflected for gender, number, definiteness, and case, although gender is usually neutral. Adjectives behave in the same way as nouns, taking similar inflections, whereas prepositions are mostly bound morphemes prefixed to nouns. The definite article in Amharic is also a bound morpheme, and attaches to the end of a noun [3, 35, 36].

There have been a lot of studies done on the topic of morphological analysis for Amharic language lately [1, 3, 31, 32]. Among them, Hornmorph is a set of Python programs for analyzing and generating words in Amharic, Tigrinya, and Oromo. It is a work in progress and users interact with the programs through Python interpreter. For each language, Hornmorph has a lexicon of verb roots and (except for Tigrinya) noun stems. It accepts a word to be analyzed and shows analysis result which includes root or stem form of the word, POS, usually grouped in

noun or verb word class, and grammatical structure. It marks words for person (singular, plural), gender (feminine, masculine), person (first, second, third), definiteness (indefinite, definite), etc. In addition to this, Hornmorph helps to generate words given root or stem and grammatical features like tense, voice, aspect, gender, number and the like [34].

Affixes in Amharic words

Affix is a morpheme fastened to a stem or base form of a word, and modifies its meaning or creates a new word. In Amharic affixes can be prefix, suffix, and infix. Prefix, is a morpheme added at the beginning of a word whereas suffixes are added at the end to form derivatives. Infixes are inserted in the body of a word causing a change in meaning, which can be easily observed in iterative and reciprocal aspect of a root word in Amharic language [1, 29, 30, 34].

Amharic verbs can have up to four prefix and up to four suffixes as shown in Figure 2.1.

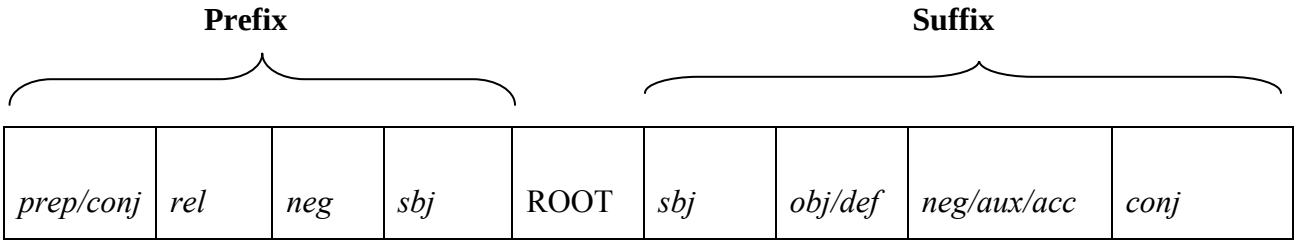

Figure 2.1: Placement of Affixes in Amharic Verbs

As shown in the Figure 2.1, prefix part has four options. First, second, third, and fourth options represent preposition or conjunction, relative, negation, and subject in terms of number, gender, person and definiteness respectively. List of conjunctions and prepositions with their probabilities are extracted from the training corpus and can be observed in Annex 2 and Annex 3 respectively. Relative verbs are marked using ”የ” /—ye-”, ”የማ”/”yemi-”, ”እማ”/”Imi-” and negation is marked with prefixes like ”አይ” /—ay-”, ”አል”/”al-”, etc. [3, 29, 34].

Similarly suffixes have four options, where the first and second option represents subject and object, in terms of gender, number, person, and definiteness respectively. The third option represents negation or auxiliary or accusation, where negation can be marked with ”-ም”/—m”, auxiliary is usually marked with morpheme ”አለ” /—l_e” and it usually appears with imperfective

and gerundive tenses, and accusative is marked with morpheme "ን"/—ን". The fourth option represents conjunction like "-ም"/ —ም", "-ስ"/"-s" etc. [29, 34].

Amharic nouns have up to two prefixes and up to four suffixes. Similarly the prefix and suffix slots have two and four sub-slots respectively. Figure 2.2 shows placement of affixes in Amharic nouns [29, 34].

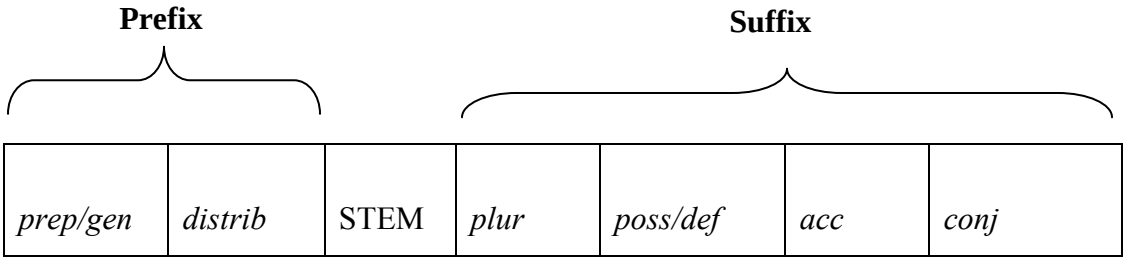

Figure 2.2: Placement of Affixes in Amharic Nouns

prep/gen option of the prefix represents preposition or genitive, where genitive is marked using morphemes —ye-—"የ-". In the second option of prefix, distributive (*distrib*) is marked using —ሀy_e-—"ሀየ-" morpheme. In case of suffix, option one, represents number information. Option two represents possessive or definiteness information. The third and fourth options represent accusative and conjunction respectively [34].

Aspect, Voice and Tense

Aspect is a grammatical category that expresses how status of an action or event is denoted by a verb. Aspect of a verb shows whether an action is completed or continuing and its relation with flow of time. Root words can be modified in two ways through introduction of vowel"ኦ-"/—ə" and, in Amharic, aspect is represented using infixes. Root words of Amharic language can have reciprocal, iterative, or simplex aspect. Simplex aspect is plain form where no vowel"-ኦ-"/—ə-" is inserted. Reciprocal aspect is obtained when vowel"-ኦ-"/—ə-" is inserted between third and second consonant from the end of a word. Reduplication of second consonant from end of a root word and inserting vowel"-ኦ-"/—ə-" between duplicated consonants produce iterative aspect [29, 34].

Examples:

Simplex: ”ሰደበ”/”sedebe”

Iterative: –ሰዳደበ”/”sedadebe”

Reciprocal: –ተሳደበ”/”tesadebe”

Voice is a form of a verb which expresses an action that a verb describes and its relation with a subject or other participants. Four voice values are possible in Amharic root which can be marked with”ተ-” /–te-–,”አስ-”/–as-–,”አ-”/—a prefixes. Simplex voice represents plain form with no prefix. Transitive, causative, and passive voice is marked with ”አ-”/—a–,”አስ-”/–as-” and ”ተ-”/ –te-”prefix respectively [29, 34].

Examples:

Simplex: –ጠቆረ”/”Tqore”

Causative: –አጠቆረ”/”’aTeqore”

Transitive: –አስጠቆረ”/”‘asTeqore”

Passive: –ተጠቆረ”/”teTeqore”

Tense is a verb form expressing different times at which an action takes place relative to the speaker or writer. Perfective, Imperfective, Gerundive, and Jussive/Imperative are the four possible values of tenses in Amharic language and are marked via prefixes and suffixes [29, 34].

Examples:

Perfective: –ወሰደ”/”wesede”

Imperfective: –ይወስዳል”/–y-wesd-al”

Gerundive: ”ይው·ሰድ”/”y-wsed”, –ው·ሰድ”/”wsed”

Jussive/Imperative: –ወስዶ”/”wesd-o”

2.4.3 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 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. A sentence is a group of words that express a complete thought. Sentences are formed from verb phrase and noun phrase and can be classified as simple and complex sentences. Phrase is a small group of words that stands as a conceptual unit. Simple sentences are formed from one verb phrase and one noun phrase whereas a complex sentence contains one or more subordinate verbs other than the main verb, where subordinate verbs are verbs that are integrated with conjunctions. A sentence is said complex because it has capability to contain other sentences within it [29]. Table 2.6 shows examples of simple and complex sentences in Amharic.

Table 2.6: Examples of Simple and Complex Sentences

Simple sentence	“አበበ ምሳውን በላ::”/”‘abebe msawn bela”/”Abebe ate his lunch”
Complex sentence	“አበበ ምሳውን እየበላ ስልክ ስለተደወለለት ሄደ::”/”‘abebe msawn _yebela slk sletedewelelet hEde”/”Abebe received a phone call while eating his lunch and he left”

Sentences are basic components of Amharic text and to give proper meanings for readers all the words with in it should be in proper order and also they should be in proper grammatical agreement. One of the basic task of word prediction software is to offer most likely word options with correct grammatical agreement based on past experience. Therefore gender, number, person, tense, etc. should be consistent throughout the sentence.

Order of Words

Formal Amharic texts follow subject-object-verb (SOV) word order unlike English language which follows subject-verb-object (SVO) sequence in a sentence. Although in some Amharic

texts, there can be OSV sequence like ልጁን አበበ መከረው”/”ljun ⁼abebe mekerew”/”The boy is advised by Abebe”, where in this case the object is suffixed by object marker ን ”/⁼ǐ, however this word order is not commonly used in formal Amharic texts. Table 2.7 shows example of word order in Amharic simple sentence.

Table 2.7: Order of words in Amharic simple sentence

	ውሻው ልጁን ነከሰው (SOV) / Wxaw ljun nekesew	The dog bite the boy(SVO)
Subject	ውሻው”/”wuxaw”	The dog”
Object	ልጁን”/”ljun”	bite”
Verb	ነከሰው”/”nekesew”	the boy”

Adjective and noun word order, Adverb and verb word order, main verb and sentence end are some of the common word sequences that should be considered in NLP studies. For example, adjectives should always appear before a noun it modifies even though other words can happen between them. Likewise an adverb always appears before a verb it qualifies[29, 34].

Subject and Verb Agreement

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

Example:

In a sentence,አበበ ልጁን ሰደበው”/”‘abebe ljun sedebew“/”Abebe insulted the boy”, the subject አበበ”/”‘abebe”/”Abebe” shows person, gender, number information which is third person, masculine, and singular respectively. This morphological properties are reflected on the verb, ሰደበው”/sedebew”/” insulted”. If one of this information is wrongly used on the verb, the

sentence cannot be in proper grammatical format and causes ambiguity to readers. For example if the above sentence is wrongly written as, “አበበ ልጁን ሰደበችው”/”‘abebe ljun sedebecw–/”‘Abebe insulted the boy”, gender information is wrongly reflected on the verb as feminine and it shows disagreement with subject. Disagreement in person and number can also cause a consistency problem in Amharic sentences. For example a sentence, –አበበ ልጁን ሰደቡት”/”’abebe ljun sedebut–/”Abebe insulted the boy”, shows disagreement in number since singular subject of the sentence is wrongly reflected on the verb as plural. Amharic verbs can have second or third person singular subject or third person plural subject to indicate politeness. For example, in a sentence –አቶ አበበ ልጁን ሰደቡት”/”‘ato ‘abebe ljun sedebut–/”Ato Abebe insulted the boy”, the verb ሰደቡት”/”sedebut–/”insulted” shows politeness and it is in agreement with the subject. However, politeness is not considered while checking subject verb agreement in this work. Therefore in order to predict words in proper morphological information, morphological properties of subject of a sentence should be captured and properly used on the verb while providing word suggestions.

Object and Verb Agreement

Object is a noun, pronoun or noun phrase denoting somebody or something that is acted on by a verb or affected by action of a verb. If a noun is used as an object in a sentence it can be suffixed by the suffix –ን”.The object of a sentence should be in accordance with the verb in gender, number, person, and case [29].

Example:

If we look this concept using the same sentence above,–አበበ ልጁን ሰደበው”/”‘abebe ljun sedebew“/”Abebe insulted the boy”, the object –ልጁን”/”ljun” shows person, gender, number and definiteness information which is third person, masculine, singular, and definite respectively and these morphological properties are reflected on the verb –ሰደበው”/sedebew”/” insulted”. If one of this information is wrongly used on the verb, the sentence cannot be in proper grammatical format. For example if the above sentence is written as, “አበበ ልጁን ሰደበቻት”/”‘abebe ljun sedebecat–/”Abebe insulted the boy”, the gender information is wrongly reflected on the verb as femininеand it shows disagreement with the object. Disagreement in person and number can also cause a consistency problem in Amharic sentences. For example the sentence, –አበበ ልጁን

ሰደባቸው"/"Abebe ljun sedebacew/"Abebe insulted the boy", shows disagreement in number. The object of the sentence is singular but on the verb it is reflected as plural. Therefore morphological properties of object of a sentence should be captured and properly reflected on the verb while providing word suggestions.

Adjective and Noun Agreement

Amharic adjectives should be in agreement in number and gender with the noun it modifies. Amharic adjectives may mark number (singular or plural) and gender (feminine or masculine) of a noun it qualifies and hence it should agree with number and gender of the noun [29].

Example:

In noun phrase "ጥቁቁር ወፎች/"TqWaqur wefoc"/"Black birds", the word "ጥቁቁር"/" TqWaqur" is an adjective that modifies the noun "ወፎች"/ "wefoc"/ "birds". It is marked for plural number and is reflected on the noun. It is inappropriate to write the above phrase as "ጥቁቁር ወፍ/" TqWaqur wef"/"Black bird", since it shows number disagreement between the adjective and noun. To write this in correct grammatical format either the adjective should be marked with singular number "ጥቁር ወፍ/" Tqur wef"/"black bird" or the noun should be marked with plural number. Noun phrase, "ትልቁ በሬ"/"tlqu berE"/"The big ox" , the word "ትልቁ"/"tlqu"/"The big", is an adjective that modifies the noun "በሬ"/" berE"/" ox". It is marked with masculine gender and is in agreement with the noun. However if we take a phrase "ትልቂቷ በሬ"/"tlqitwa berE"/"The big ox", the adjective is marked with feminine gender while the noun it modifies is masculine. Therefore the adjective and noun are in disagreement and to avoid this kind of inconsistency either the adjective should be marked with masculine or the noun should be marked with feminine gender. For this particular example an appropriate phrase is either "ትልቁ በሬ"/"tlqu berE"/"The big ox" or "ትልቂቷ ላም"/"tlqitwa lam"/"The big cow", where there is agreement in number and gender between the adjective and noun.

Adverb and Verb Agreement

Amharic adverbs usually modify the first verb that comes next to it. Time adverbs describe the time a certain event or action occurred. Amharic verbs take certain tense form to indicate time. Time adverb should agree with the verb it modifies [29, 31].

Example:

In a sentence ‹ልጁ ነገ ይመጣል›/”lju nege ymeTal”/”The boy will come tomorrow”, the word ‹ነገ›/”nege”/” tomorrow” is an adverb that modifies the verb ‹ይመጣል›/”ymeTal”/” will come”. The adverb and verb are in agreement taking imperfective tense form.

2.5 Summary

In this Chapter we have reviewed the challenges and opportunities of word prediction. We have also discussed existing approaches to word prediction, their weakness and strength. Furthermore, evaluation methods for word prediction systems are discussed and KSS is selected to assess our word sequence prediction work. Finally, we reviewed concepts associated with Amharic language like Amharic Parts-of-Speech, Amharic Morphology, and Amharic Grammar.

CHAPTER THREE

RELATED WORK

This Chapter presents word or text prediction researches with their approaches and obtained results. Word prediction studies conducted for Western, Persian, Russian, and Hebrew languages are some of the works thoroughly reviewed to grasp satisfactory knowledge and to look for the finest approach for Amharic language.

3.1 Word Prediction for Western Languages

There are some researches conducted on word prediction for western languages like Italian, Swedish, English, German, French, and Dutch. Aliprandi *et al.* [18, 19], focuses on designing letter and word prediction system called FastType for Italian language. Italian has large dictionary of word forms, which go with a number of morphological features, produced from a root or lemma and a set of inflection rules. Statistical and lexical methods with robust open-domain language resources which have been refined to improve keystroke saving are used. The user interface, predictive engine and linguistic resource are main components of the system. The predictive engine is kernel of predictive module since it manages communication with the user interface keeping trace of prediction status and words already typed.

The morpho-syntactic agreement and lexicon coverage, efficiently accessing linguistic resources as language model and very large lexical resources are core functionalities of the predictive module. In addition, to improve morphological information available for prediction engine, POS n-grams and Tagged word (TW) n-grams are used. The prediction algorithm for Italian language is presented by extending combination of POS tri-grams and simple word bi-grams model. A large corpus prepared from newspapers, magazines, documents, commercial letters and emails are used to train Italian POS n-grams, approximated to n = 2 (bi-grams) and n = 3 (tri-grams) and tagged word n-grams, approximated to n = 1 (uni-grams) and n = 2 (bi-grams).

Keystroke saving (KS), Keystroke until completion (KUC) and Word Type Saving (WTS) are three parameters used to evaluate the system. The researchers indicate that 40 texts disjoint from

training set are used for testing. However, the size or number of words available in the testing data is not clearly specified. The result shows 51% keystroke saving, which is comparable to what was achieved by word prediction methods for non-inflected languages. Moreover, on average 29% WTS, meaning at standard speed without any cognitive load saving in time and 2.5 KUC is observed.

Moreover, Matiasek *et al*. [26] have done a multilingual text prediction study and a system named FASTY is developed. The aim of this work is to offer a communication support system to significantly increase typing speed, which adapts to users with different language and strongly varying needs. It follows a generic approach in order to be multilingual so that the concept can be used for most European languages. However, this study focused on Germen, French, Dutch and Swedish languages. The predictor and language specific resources are separated by the language independent prediction software, which helps the system with potential application to many European languages without sacrificing performance. Preliminary experiments with German as well as experiences with a Swedish system have shown that n-gram based methods still offer quite reasonable predictive power. N-gram statistics, morphological processing and backup lexicon, and abbreviation expansion are core components of this system. The frequency tables of word n-grams are easily constructed from text corpora irrespective of the target language and incorporating Part-of-Speech (POS) provides additional precision. The combination of different n-gram statistics constitutes the base of FASTY predictor providing a baseline performance for all target languages. Other modules interact with these results and improve on them.

Morphological analysis and synthesis are performed and morph-syntactic features needed by the components dealing with checking syntactic appropriateness are extracted since one of FASTY's goals is to be able to suggest only word forms appropriate for the current context. Also compound prediction needs morph-syntactic information of compound parts to correctly predict linking elements. Last but not least, if frequency based lexica run out of words with a given prefix, the morphological lexicon provided will serve as a backup lexicon and deliver additional solutions. Morphological processing is implemented via infinite state-transducers, which provide very fast, bi-directional processing and allow for a very compact representation of huge lexica. The grammar-based module is used to enhance the predictive power of FASTY and improve its

precision using syntactic processing in order to deliver only predictions that are not in conflict with the grammar.

Carlberger *et al*. [37] conducted a study on constructing a database for Swedish language called Profet via extension of available word prediction system which uses word frequency lexicon, word pair lexicon, and subject lexicon. Profet is a statistical based word prediction system that has been used for a number of years as a writing aid by persons with motoric disabilities and linguistic impairments. It gives one to nine word alternatives as a user starts spelling a word based on the selected settings. The main task of this work is to enhance the available prediction capability through extension of scope, addition of grammatical, phrasal and semantic information and using probability based system. This allows information from multiple sources to be weighted appropriately for each prediction. The predictor scope is extended considering preceding words in the prediction. Therefore, prediction is also based on previous words even after typing any letters of the new word. This leads the word suggestions to be grammatically more correct than those presently given. Since the available database lacks grammatical information as well as statistics for the occurrence of sequences longer than two contiguous words, a new database is built. Besides bi-grams (word and grammatical tag pairs with co-occurrence statistics), tri-grams as well as collocations (non-contiguous sequential word and grammatical tag bi-grams with 2-5 intervening words) are included. All information in the new database including collocations must be extracted from one single corpus in order to warrant implementation of a probabilistic prediction function. This work extends the previous version of profet which presents one word per line by displaying more than one word per line. It is briefed that choosing words from the word alternatives can result up to 26% keystroke savings (KSS) and up to 34% in letters when only one word is typed.

Agarwal and Arora [38] proposed a Context Based Word Prediction system for SMS messaging for English language in which context is used to predict the most appropriate word. The development of wireless technology has made available different ways of communications like short message service (SMS) and with its tremendous increase of use there comes a need to efficient text input methods. Various scholars came up with frequency based text prediction methods to attempt this problem. However, using only frequency based word prediction may not grant correct result most of the time. For example, considering a sentence *"give me a box of*

chocolate" and *"give of a box of chocolate"*, appropriate word after the word "give" is a word *"me"*. However, the system proposes the word *"of"* since it has higher frequency than the word *"me"*. Similarly the appropriate word after the word *"box"* is *"of"* than *"me"* and here frequency based is acceptable. Therefore incorporating context information is helpful to offer suitable word and this work models first order Marcov dependency between POS of consecutive words. A machine learning algorithm is used to predict the most probable word and POS pair, given its code and previous word's POS. Considering the fact that short emails resemble SMS messages closely, the algorithm is trained on 19,000 emails and testing is done on 1,900 emails which are collected from Enron email corpus. The results show 31% improvement compared to the traditional frequency based word estimation.

Al-Mubaid and Chen [23] conducted a research using machine learning method to address the problem of word prediction for English language. This work integrates supervised and adaptive learning to enhance text entry for physically disabled users and having minimized cognitive load. The process of browsing and reading the anticipated words imposes an extra cognitive load on a user especially when the number of suggestions is larger. This research focuses on minimizing cognitive load by offering, in most cases, only one suggestion, but no more than three suggestions in any case. In this research, two classes of learning methods, supervised and adaptive learning methods, are investigated, designed, and implemented. These two classes of methods are integrated into a comprehensive learning architecture that will be capable of acquiring reliable and relevant knowledge automatically and efficiently. The key objective is to allow the system to learn from prior training texts (supervised learning), and from the user (adaptive learning), so it can reliably predict words a user intends to input. The adaptive learning paradigm learns user's specific writing style and word usage to assist in word prediction. The proposed method allows for fast text entry and more accurate text communication with computers and reduces cognitive load due to less number of suggestions.

Trnka [21] made a research on a topic adapted language model for word prediction, which improves keystroke savings over a comparable baseline. This work is planned to develop and integrate style adaptations from the experience of topic models to dynamically adapt to both topically and stylistically. Topic models are language models that dynamically adapt to testing data, focusing on most related topics in training data. The first stage of this study is identifying

relevant topics and the second stage is tuning the language model based on relevant topics. Here a language model is adapted to most appropriate topics in training text and it is tuned to the most relevant portions. According to the evaluation, topic modeling can significantly increase keystroke savings for traditional testing as well as testing on a text from other domains. The problem of annotated topics is also addressed through fine-grained modeling and found a significant improvement over a baseline n-gram model.

Al-Mubaid [39] studied a learning classification based approach for word prediction. This study presents word prediction using highly discriminating context features and machine learning. Feature extraction method is adapted from Mutual Information (MI) and Chi-Square(x^2). These methods have been used successfully in Information Retrieval (IR) and Text categorization (TC). Thus, word prediction problem here is treated as a word classification task in which multiple candidate words are classified to determine the most correct one in a given context. First for a given occurrence of a word w, representation of w involves recording occurrence of certain word features extracted from the training corpus using new feature extraction technique adapted from MI and x^2. The encoding is used in the training phase to train word classifiers using SVM learner. The word classifiers are then employed by word predictors to determine the correct word given its context. One of the properties of this method is that it performs word prediction by utilizing very small contexts. As per the evaluation, best performance is obtained with context of size 3 using only the preceding 3 words. Additionally the best performance resulted when using 20 features (i.e., using the top 20 words having the highest 20 MI_2, or x^2). Thus, results reported here are generated using the preceding 3 words (context size = 3) and the top 20 MI_2, or x^2 words.

A word prediction study via a clustered optimal binary search tree is conducted by El-Qawasmeh [36]. Word prediction methodologies heavily depend on statistical approach that uses uni-gram, bi-gram, tri-gram, etc. However, construction of word n-grams requires large size of memory which is a challenging task for many existing computers. Therefore, this work intends to use cluster of computers to build an optimal binary search tree that will be used for statistical based word prediction. The suggested approach uses a cluster of computers connected to build frequencies. This system is evaluated based on keystroke saving and according to the experiment keystroke saving is improved.

Garay-Vitoria and Abascal [17] conducted a research on word prediction for inflected language, specifically Basque language, based on three approaches. Various word prediction techniques and their difficulties to apply to inflected language are briefed. The Basque language is mainly inflected using suffixes eventhough there is a possibility of infixes and prefixes. The first approach needs two dictionaries one for lemmas and the other for suffixes since it predicts lemmas and suffixes separately. The first dictionary stores lemmas of the language alphabetically ordered with their frequencies and some morphologic information in order to know which possible declensions are possible for a word. The second dictionary stores suffixes and their frequencies. The system starts prediction by providing lemma of next word and when accepted the system offers most probable suffixes, since the number of suffixes in Basque language is 62. Possibilities of recursively composed suffixes are some of the challenges in this approach even though hopeful results are obtained. In the second approach syntactic information is added to the dictionary of lemmas and some weighted grammatical rules on the system. The main idea is to parse a sentence while it is being composed and to propose most appropriate lemmas and suffixes, where parsing allows storing and extracting information that has influenced in forming a verb. The third approach treats beginning of sentences using statistical information while advancing in composition of a sentence, and uses this information to offer the most probable word including both lemma and suffix. Three tables are used; one with probabilities of syntactic categories of the lemmas to appear at the beginning of a sentence, probability of basic suffixes to appear after those words and probabilities of basic suffixes to appear after another basic suffix. Adaptation of the system would be made by updating the first table while suffixes would be added to a word and the other two tables are also updated. As the researchers state, to predict whole words it is necessary to determine syntactic role of the next word in a sentence, which can be done using syntactic analysis. However, the results are not good enough compared with results obtained in non-inflected languages.

3.2 Word Prediction for Hebrew Language

Netzer *et al.* [16] conducted a research on word prediction for Hebrew language as part of an effort for Hebrew AAC users. Modern Hebrew is characterized by rich morphology, with a high level of ambiguity. Morphological inflections like gender, number, person, tense and construct state can be shown in Hebrew lexemes. In addition, better predictions are achieved when

language model is trained on larger corpus size. In this work the hypothesis that additional morpho-syntactic knowledge is required to obtain high precision is evaluated. The language model is trained on uni-gram, bi-gram and tri-gram, and experiment is made on four sizes of selection menus: 1, 5, 7 and 9, each considered as one additional keystroke. According to the result, the researchers state that syntactic knowledge does not improve keystroke savings and even decreases them, as opposed to what was originally hypothesized. The result shows keystroke savings up to 29% with nine word proposals, 34% for seven word proposals and 54% for a single proposal. Contrary to other works, KSS is improved as the size of selection menu reduced. We believe that an increase in number of proposals affects search time. However, effect of selection menu‘s size on KSS is not clear and no justification is given by the researchers.

3.3 Word Prediction for Persian Language

Ghayoomi and Daroodi [20] studied word prediction for Persian language in three approaches. Persian language is a member of the Indo-European language family and has many features in common with them in terms of morphology, syntax, phonology and lexicon. This work is based on bi-gram, tri-gram, 4-gram models and it utilized around 10 million tokens in the collected corpus. The first approach uses word statistics, the second one includes main syntactic categories of a Persian POS tagged corpus, and the third uses main syntactic categories along with their morphological, syntactic and semantic subcategories. According to the researchers, evaluation shows 37%, 38.95%, and 42.45% KSS for the first second and third approaches respectively.

3.4 Word Prediction for Russian Language

Hunnicutt *et al*. [40] performed a research on Russian word prediction with morphological support as a co-operative project between two research groups in Tbilisi and Stockholm. This work is an extension of a word predictor developed by Swedish partner for other languages in order to make it suitable for Russian language. Inclusion of morphological component is found necessary since Russian language is much richer in morphological forms. In order to develop Russian language database, an extensive text corpora containing 2.3 million tokens is collected. It provides inflectional categories and resulting inflections for verbs, nouns and adjectives. With this, the correct word forms can be presented in a consistent manner, which allows a user to

easily choose the desired word form. The researchers introduced special operations for constructing word forms from a word‘s morphological components. Verbs are the most complex word class and algorithm for expanding root form of verbs to their inflectional form is done. This system suggests successful completion of verbs with the remaining inflectable words.

3.5 Word Prediction for Sindhi Language

Mahar and Memon [41] studied word prediction for Sindhi language based on bi-gram, tri-gram and 4-gram probabilistic models. Sindhi is morphologically rich and has great similarity with Arabic, Persian, and Urdu Languages. It is a highly homographic language and texts are written without diacritic symbols which makes word prediction task very difficult. The corpus of any language is very important for statistical language modeling. Hence, in this work, word frequencies are calculated using a corpus which approximately contains 3 million tokens and a tokenization algorithm is developed to segment words. Add one smoothing technique is used to assign non zero probabilities to all probabilities having zero probabilities. 15,000 sentences are randomly selected from the prepared corpora to evaluate developed models based on entropy and perplexity. According to the evaluation, 4-gram model is more suitable since it has lower perplexity than bi-gram and tri-gram models.

3.6 Word Prediction for Amharic Language

Nesredin Suleiman and Solomon Atnafu [14] conducted a research on word prediction for Amharic online handwriting recognition. As the researchers state, the study is motivated by the fact that speed of data entry can be enhanced with integration of online handwriting recognition and word prediction mainly for handheld devices. The main target of the work is to propose a word prediction model for Amharic online handwriting recognition using statistical information like frequency of occurrence of words. A corpus of 131,399 Amharic words and 17, 137 names of persons and places are prepared. The prepared corpus is used to extract statistical information like to determine value of n for the n-gram model, average word length of Amharic language, and the most frequently used Amharic word length. Hence, n is set to be 2 based on statistical information, and in retrospect to this, the research is done using bi-gram model, where the intended word is predicted by looking the first two characters. Finally, a prototype is developed

to evaluate performance of the proposed model and 81.39% prediction accuracy is obtained according to the experiment.

3.7 Summary

In this Chapter, we have discussed works related to word sequence prediction for different languages. A word completion study specifically targeted for online handwriting recognition of Amharic language and done using pure frequency based method is also presented. This approach is very challenging for inflected languages due to large possibility of word forms. In addition wrong morphological output will be offered since no context information is considered. Therefore this research aims to fill the unattained gap in the existing work so that words can be proposed in the correct morphological form by considering context information and linguistic rules. User interface, prediction module, and linguistic resources are main components of word prediction systems where the linguistic resource embraces statistical or other information depending on the target language. From the reviewed works, we also learnt that considering only frequency of words is not enough for inflected languages, root or stem words and morphological features can be treated separately, incorporating context information increases effectiveness of prediction output, and n-gram models have good capacity to capture context information.

CHAPTER FOUR

WORD SEQUENCE PREDICTION MODEL FOR AMHARIC LANGUAGE

This Chapter presents details of the Amharic Word Sequence Prediction Model. Architecture of the proposed Word Sequence Prediction Model and its components with their respective algorithms are described in this Chapter. N-gram statistical language model is applied to offer most expected root or stem words, and morphological features like aspect, tense, and voice. In addition grammatical rules of Amharic language, such as Subject-Object-Verb, Adjective-Noun and Adverb-Verb agreement are used to inflect the proposed root or stem words to appropriate word form. The Amharic Word Sequence Predictor accepts user‘s input, extract root or stem word and required features by analyzing a user‘s input, propose the most likely root or stem words with their most probable features and finally generates surface words using the proposed root or stem words and features.

4.1 Architecture of Amharic Word Sequence Prediction Model

The model shown in Figure 4.1 is designed to predict words a user intends to type by considering previous history of words. Constructing Language Model and Generation of Predicted Words are the two major parts. First the training corpus is morphologically analyzed using Hornmorph. Subsequently, using the morphologically analyzed corpus we built a tagged training corpus. Then, language models like root word sequences and root word with features are built based on the tagged training corpus. Morphological Analysis of User Input, Word Sequence Prediction, and Morphological Generation are key components of the Generation of Predicted Words part. Here, a user‘s input is accepted and analyzed using Hornmorph. Subsequently, root and morphological features of words are extracted so that the word prediction component uses this information to propose words by interacting with the language model. Finally the morphological generator produces surface words to the user given root and feature words proposal.

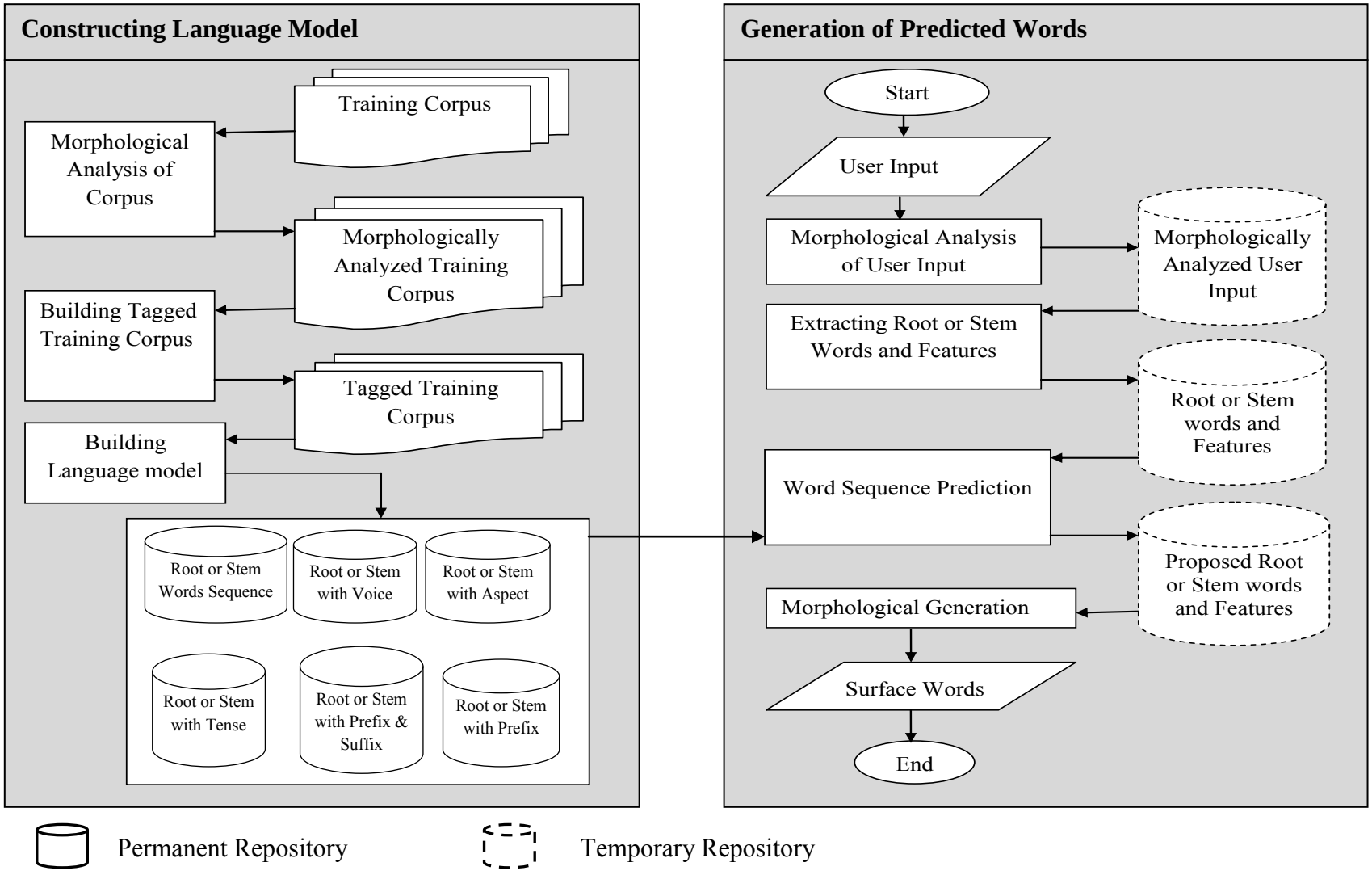

Figure 4.1: Architecture of Amharic Word Sequence Prediction Model

4.2 Morphological Analysis of Corpus

This module analyzes words in a given training data to identify root or stem form and component morphemes so that required features and root or stem word are extracted to build a tagged corpus. This tagged corpus is used to construct statistical language models. 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 knowledge-base. Language models built from large corpora tend to perform better, particularly for words that are infrequent. Word prediction task requires a large size of corpus inorder to have sufficient statistical information for training the system. In this study, text collection containing nearly 298,500 sentences which is gathered from Walta Information Center (WIC) is used.

Morphological analysis is the process of assigning each word found in a corpus to their morphemes which can be affix, root, stem, etc. It is useful to annotate words to their root form and other required morphological information. Morphological analyzer is a program used to analyze a separate word or words in a file to their component forms.

Amharic is a morphologically rich language as described in previous chapters. A verb lexeme can appear in more than 100,000 word forms [30], and it is impractical to store all forms of words in probabilistic models. For this reason, the training corpus is pre-processed to hold only the root or the stem and selected morphological features of words. Features are selected by studying structure of Amharic words and method of producing variety of words from the base word. As described in Section 2.4.2, Amharic verbs have four prefix and four suffix options. Similarly, Nouns in Amharic have two prefix and four suffix options. Subject, object, definiteness options can be handled using grammatical agreement rules of the language. However, other prefix and suffix options, and features like voice, tense and aspect are selected to be incorporated in the tagged training corpus since they have effect in inflection of Amharic words. Hornmorph is used in this study to analyze the training corpus.

Through morphologically analyzed training corpus, a tagged corpus consisting only root or stem form, affixes, aspect, voice and tense is constructed. However, words that cannot be analyzed using Hornmorph are taken as they are, to keep consistency of root or stem word sequences.

Hornmorph analyzes words as a verb or noun group and affixes have different characteristics for verbs and nouns. Therefore verbs and nouns are represented differently in the tagged corpus.

In this module prefix and suffix options of words except subject and object or definiteness, are handled statistically by embracing them in the tagged corpus. Here, we represented all words existing in the training corpus in six slots, where slot 1, slot 2, slot 3, slot 4, slot 5, and slot 6 stand for root or stem word, prefix, suffix, tense, aspect and voice respectively. The prefix and suffix slot have three and two sub-slots in that order for a verb and two and four sub-slots correspondingly for a noun. Figures 4.2 and 4.3 show how a verb and a noun are represented in the tagged training corpus respectively.

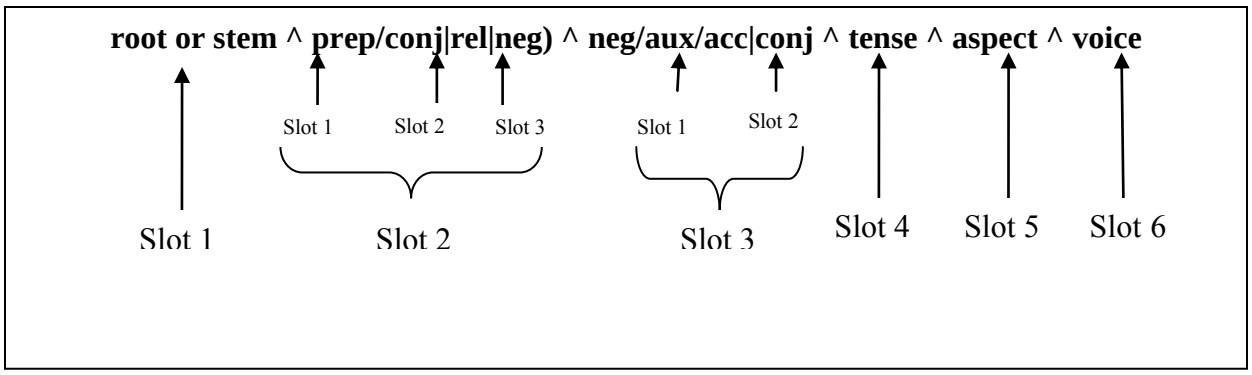

Figure 4.2: Representation of Amharic Verb in Tagged Corpus

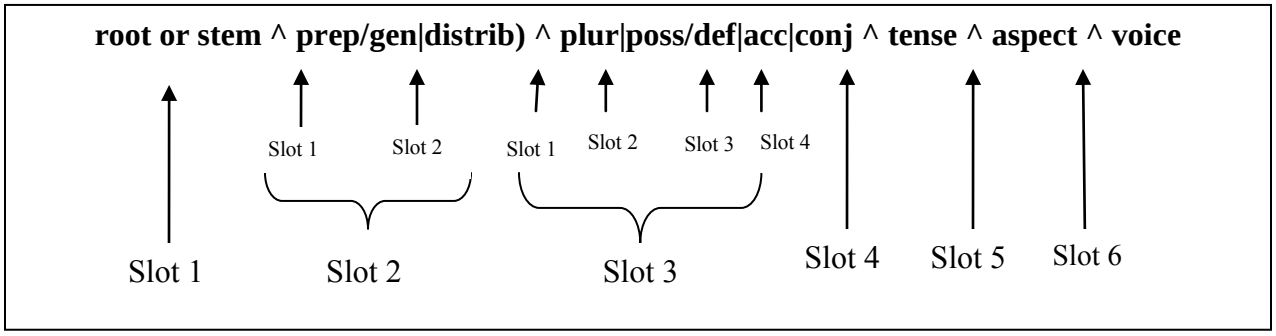

Figure 4.3: Representation of Amharic Noun in Tagged Corpus

Algorithm 4.1 describes an algorithm to construct tagged corpus.

```
INPUT training-corpus
ANALYZE training-corpus using Hornmorph and WRITE in analyzed-corpus
        INITIALIZE keywords for prefix, rootWord, suffix, aspect, tense, voice,newWord
        INITIALIZE prefix, root, suffix, aspect, tense, voice, value to 0,newWord,
newWord2 to FALSE
READ morphologically-analyzed-corpus
        FOR each line in morphologically-analyzed-corpus:
                ADD each word in the line to a list
                FOR each word in the list
                        IF word is in newWord key word and newWord2 is FALSE
                                SET newWord to TRUE
                        ELSE IF newWord is TRUE
                                newWord=FALSE
                                newWord2=TRUE
                                rootWord=word
                        ELSE IF newWord is TRUE and word is in prefix Keyword:
                                prefix=word
                        ELSE IF newWord is TRUE and word is in suffix Keyword:
                                suffix=word
                        ELSE IF newWord is TRUE and word is in aspect Keyword:
                                aspect=word
                        ELSE IF newWord is TRUE and word is in voice Keyword:
                                voice=word
                        ELSE IF newWord is TRUE and word is in tense Keyword:
                                tense=word
                        ELSE IF word in newWord key word and newWord2 is TRUE
                                WRITE(rootWord+'^'+prefix'^'+suffix+'^'+tense
                                        +'^'+aspect+'^'+voice) on tagged-training-corpus
                                SET newWord2 to FALSE and newWord to TRUE
OUTPUT tagged-training-corpus
END
```

Algorithm 4.1: Algorithm to Build a Tagged Corpus

The output of Algorithm 4.1 is a tagged training corpus, segment of the tagged corpus containing root or stem form of words and grammatical features is presented in Figure 4.4.

ሶማሌ^be|0|^0|0|0|0^0^0 ክልል^0|0|^0|0|0|0^0^0 w'b^gen|0|^0|0|0|0^0^0 ሸበል^0|0|^0|def|0|0^0^0
ወይብ^gen|0|^0|0|0|0^0^0 ወንዝ^0|0|^plr|0|0|0^0^0 mWl'^0|0|0^0|0^gerundive^0^0
T|lqlq^be|0|^0|def|0|0^0^transitive አራት^be|0|^0|0|0|0^0^0 ወረዳ^0|0|^plr|0|0|0^0^0
ህዝብ^0|0|^0|0|0|0^0^0 ንብረት^0|0|^0|0|0|0^0^0 ላይ^0|0|^0|0|0|0^0^0
gWd'^0|0|0^0|0^jussive/imperative^0^0 drs^0|0|^0|0|acc|0^0^transitive ክልል^gen|0|^0|0|0|0^0^0
አደጋ^0|0|^0|0|0|0^0^0 klkl^0|0|^0|0|0|0^reciprocal^passive ዝግጁነት^0|0|^0|0|0|0^0^0
ቢሮ^0|0|^0|0|0|0^0^0 'wq^0|0|0^0|0^perfective^0^transitive

Figure 4.4: Segment of Tagged Corpus

Representation of each tagged word is briefed in Table 4.1. Each word is denoted in six slots, which are root or stem, prefix, suffix, tense, aspect and voice. In addition to this, prefix and suffix slots consist sub slots. The value –θ" in each slot indicates null value for that particular slot, however, –θ" value for aspect and voice is equivalent to simplex value. A prefix, –θ|0|0" represents preposition or conjunction, relative, and negation slots having null value. The suffix –θ|0" represents negation or auxiliary or accusative and conjunction slots holding null values. Therefore this particular word does not have any prefix and suffix; it has Jussive or Imperative tense, simplex aspect, and simplex voice.

Table 4.1: Representation of Words in the Tagged Corpus

Tagged word	Root	Prefix	Suffix	Tense	Aspect	Voice
gWd'^0|0|0^0|0^jussive/imperative^0^0	gWd‘	0|0|0	0|0	Jussive /imperative	0	0

4.3 Building Language Models

Language model is a storage consisting of statistical information which serves as a knowledge base when predicting suitable words. The word sequence prediction task is accomplished in two

phases. In phase one, root or stem form of words are suggested using root or stem n-gram models. In the next phase, morphological features of proposed root or stem words are predicted using statistical methods as well as linguistic rules to ensure grammatical agreement among words. The proposed root or stem word and features are used later while generating appropriate surface words. Therefore building language model is one of the main components of our word sequence prediction model. Statistical models of root or stem word sequences and morphological features are constructed using the tagged corpus. A number of word prediction researches are conducted using bi-gram and tri-gram models [16, 18, 19, 41]. Accuracy of word predictor improves as n in the n-gram model increases due to its ability of suggesting words with more context information. However, its complexity and data size increases causing a reduction in response time. Therefore, based on related works experience and characteristics of n-gram models, we have decided to use bi-gram and tri-gram models.

4.3.1 Root or Stem Words Sequence

Bi-gram, tri-gram and hybrid of bi-gram and tri-gram statistical models are constructed for root or stem words sequence using the training corpus. Each n-gram model is separately kept in its own repository and they hold root or stem word sequences for each value of n with their probability of occurrence in the corpus.

Probabilities of all unique root or stem word sequences with this respective value of n is calculated by counting occurrence of n word sequences and n-1 word sequences in the corpus where n is 2 for bi-gram and 3 for tri-gram models, and then calculating their ratio. Bi-gram and tri-gram probabilities are computed using (Eq.3) and (Eq.4) respectively.

$$P(w_2|w_1) = \frac{C(w_2 w_1)}{C(w_1)} \quad \text{(Eq.3)}$$

where,w_1, w_2 are words , $P(w_2|w_1)$ is probability of a word w_2 given w_1, $c(w_2w_1)$ is frequency of word sequence w_2w_1 in a corpus, $c(w_1)$ is frequency of w1 in a corpus.

$$P(w_3|w_2w_1) = \frac{C(w_3 w_2 w_1)}{C(w_2 w_1)} \quad \text{(Eq.4)}$$

where,w_1, w_2, w_3 are words , $P(w_3|w_2w_1)$ is probability of a word w_3 given w_2w_1 previous words , $c(w_3w_2w_1)$ is frequency of word sequence $w_3w_2w_1$ in a corpus, $c(w_2w_1)$ is frequency of w_2w_1 in a corpus.

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

C (ጽህፈት ቤት ህላፊ) =6166

C (ጽህፈት ቤት) = 28765

$$\text{P (ህላፊ | ጽህፈት ቤት)} = \frac{C(\text{ጽህፈት ቤት ህላፊ})}{\text{C(ጽህፈት ቤት)}} = \frac{6166}{28765} = 0.2144$$

where, C(ጽህፈት ቤት ህላፊ) and C(ጽህፈት ቤት), are number of occurrences of words sequence ―ጽህፈት ቤት ህላፊ" and ―ጽህፈት ቤት" in a given corpus respectively, P(ህላፊ | ጽህፈት ቤት) is probability of word ―ህላፊ", given previous words sequence ―ጽህፈት ቤት" in a given corpus.

In a similar way, each unique word sequences probability is calculated. Along with this, bi-gram, tri-gram and hybrid probabilistic models are constructed and stored in a separate repository. Figure 4.5 shows sample of tri-gram root or stem probabilistic information. Consequently, using these prepared probabilistic models, fifteen most likely root or stem words are proposed as part of our word sequence prediction task. We set the number of suggestion to be fifteen empirically.

Word1	Word2	Word3	Probability
ኢትዮጵያ	ዜና	አገልግሎት	0.9541
ዜና	አገልግሎት	glS	0.5486
ጽህፈት	ቤት	'wq	0.1782
ጽህፈት	ቤት	glS	0.0957
ስራ	knawn	hwn	0.1536

Figure 4.5: Sample of the Tri-gram Root or Stem Probabilistic Information

Algorithm 4.2 describes the algorithm to construct n-gram root or stem words probabilistic model.

```
BEGIN
INPUT root-or-stem-word-file
READ value of N
FOR each sentence in a file:
        EXTRACT N sequences
        WRITE each sequence in root-stem-word-sequence file
READ root-stem-word-sequence file
FOR each unique sequence:
        COUNT number of its occurrence, and ASSIGN value to frequency
        WRITE the frequency with their respective sequences in a freq-sequence file
READ N and N-1 sequences with their frequencies from freq-sequence file
CALCULATE probability of N sequence of words by taking ratio of frequency of N
sequence words with N-1 sequence words
WRITE probability with their respective sequences in a file
OUTPUT root-or-stem sequence n-gram probabilistic model
END
```

Algorithm 4.2: Algorithm to Construct n-gram Probabilistic Models

4.3.2 Root or Stem Words with Aspect

Bi-gram model of root or stem words with their respective aspect is constructed by extracting and counting occurrence of unique root or stem word with its aspect sequence. This model stores frequency of each root word with its aspect. Aspect of a verb can be simplex, reciprocal, or iterative. The most frequent aspect for a particular root or stem word is used when producing surface words. Algorithm 4.3 describes an algorithm to construct root and aspect bi-gram model.

```
BEGIN
INPUT tagged-training-corpus
FOR each word in tagged-training-corpus:
        SPLIT each word by '^' and ADD each item to a list
        EXTRACT root and aspect using the item having '0' and '4' index from the list,
        WRITE root-aspect-sequence in a file
READ  root-aspect-sequence file
FOR each root-aspect-sequence in the file
        ASSIGN frequency=0
        IF root-aspect-sequence is new
                COUNT root-aspect-sequence and ASSIGN it to frequency
                WRITE root-aspect-sequence and frequency in a file
OUTPUT root-with-aspect n-gram model
END
```

Algorithm 4.3: Algorithm to Construct Root or Stem and Aspect bi-gram model

4.3.3 Root or Stem Words with Voice

Unique occurrence of root or stem words with their respective voice is counted from the training corpus to build root or stem word and voice bi-gram model. This model stores frequency of each root or stem word with its respective voice. The voice can be simplex, transitive, or passive. The most frequent voice for a particular root or stem word is used when suggesting most probable features for a given root or stem word. Algorithm 4.4 describes an algorithm to construct bi-gram model for root or stem and voice.

```
INPUT tagged-training-corpus
FOR each word in tagged-training-corpus:
        SPLIT each word by '^' and ADD each item to a list
        EXTRACT root and voice using the item having '0' and '5' index from the list,
        WRITE root-voice-sequence in a file
READ  root-voice-sequence file
FOR each root-voice-sequence in the file
```

```
        ASSIGN frequency=0
        IF root-voice-sequence is new
                COUNT root-voice-sequence and ASSIGN it to frequency
                WRITE root-voice-sequence and frequency in a file
OUTPUT root-with-voice n-gram model
END
```

Algorithm 4.4: Algorithm to construct Root or Stem and Voice bi-gram Model

4.3.4 Root or Stem Words with Prefix

Tri-gram statistical information is built for three consecutive root or stem word sequences where the last root or stem word is taken with its prefix. This model stores frequency of successive root or stem words with prefix. This information is used to predict the most probable prefix for suggested root or stem words so as to produce suitable surface words. Algorithm 4.5 shows the algorithm to construct root or stem and prefix tri-gram model.

```
BEGIN
INPUT tagged-training-corpus
FOR each sentence in  tagged-training-corpus
        ADD each word in tagged-training-corpus to a list,words
        FOR i in RANGE 0 to length of the list(words-2)
                WRITE(words[i][0],words[i+1][0],words[i+2][0],words[i+2][1]) in root-
                prefix-sequence//index '0' is for root word and index '1' is for prefix
        READ  root-prefix-sequence file
        FOR each root-prefix-sequence in the file
                ASSIGN frequency=0
                IF root-prefix-sequence is new
                        COUNT root-prefix-sequence and ASSIGN it to frequency
                        WRITE root-prefix-sequence and frequency in a file
OUTPUT root-with-prefix n-gram model
END
```

Algorithm 4.5: Algorithm to construct Root or Stem and Prefix tri-gram Model

4.3.5 Root or Stem Words with Prefix and Suffix

Frequencies of each root or stem word with its respective prefix and suffixes are identified and kept in its own repository. Based on this information, the most likely suffix for a given root or stem and prefix is predicted. The proposed suffix is used by Hornmorph morphological generator while producing surface words. Algorithm 4.6 describes an algorithm to construct this model.

```
BEGIN
INPUT tagged-training-corpus
FOR each word in tagged-training-corpus:
        SPLIT each word by '^' and ADD each item to a list
        EXTRACT root, prefix and suffix using the item having '0', '1' and '2' index
from the list,
        WRITE root-prefix-suffix-sequence in a file
READ  root-prefix-suffix-sequence file
FOR each root-prefix-suffix-sequence in the file
        ASSIGN frequency=0
        IF root-prefix-suffix-sequence is new
                COUNT root-prefix-suffix-sequence and ASSIGN it to frequency
                WRITE root-prefix-suffix-sequence and frequency in a file
OUTPUT root-with-prefix-and-suffix n-gram model
END
```

Algorithm 4.6: Algorithm to Construct Root or Stem, Prefix and Suffix Tri-gram Model

4.3.6 Root or Stem Words with Tense

Root or stem words with their respective tenses are extracted from the tagged training corpus and bi-gram model is constructed by counting each unique sequence. Here, frequency of each root word with its respective tense is constructed. Perfective, imperfective, gerundive, and imperative or jussive are possible tense categories. Based on this information, the most likely tense for a given root or stem is predicted. This statistical information is used when adverb-verb agreement

rule is not applicable. Algorithm 4.7 shows the algorithm to build root or stem words with tense bi-gram model.

```
BEGIN
INPUT tagged-training-corpus
FOR each word in tagged-training-corpus:
        SPLIT each word by '^' and ADD each item to a list
        EXTRACT root and tense using the item having '0' and '3' index from the list,
        WRITE root-tense-sequence in a file
READ  root-tense-sequence file
FOR each root-tense-sequence in the file
        ASSIGN frequency=0
        IF root-tense-sequence is new
                COUNT root-tense-sequence and ASSIGN it to frequency
                WRITE root-tense-sequence and frequency in a file
OUTPUT root-with-tense n-gram model
END
```

Algorithm 4.7: Algorithm to Construct Root or Stem and Tense bi-gram Model

4.4 Morphological Analysis of User Input

This module analyzes Amharic texts accepted from a user and extracts required morphological features. Context information like gender, number, person and definiteness is captured from a user's input to predict appropriate morphological features for the coming root or stem word. When a user enters a text, the system identifies the last phrase and morphologically analyzes each word found in it. Hornmorph is used to analyze user's entered text, so that words found in input text are tagged with their respective gender, number, person, definiteness, root or stem and POS information automatically, where POS is fetched from the user input in our case. This tagged information of a user's input is used further in word sequence prediction task to keep morpho-syntactic agreement. These words are tagged in five slots as shown in Figure 4.6, where Slot 1, Slot 2, Slot 3, Slot 4, and Slot 5 represent POS, gender, number, person and definiteness information respectively.

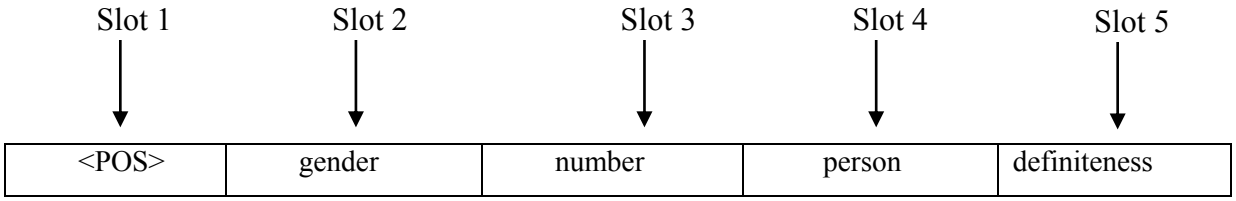

Figure 4.6: Placement of Captured Morphological Features from a user's Input

Slot1:

This slot stores word class of every analyzed word which can be <N>, <NC>, <V>, etc. The complete list of POS is listed in Annex 4. Here our main need is to look for adjectives, adverbs and nouns. This is because; the morphological generation component uses POS information to ensure subject-object-verb, adverb-verb and adjective-noun agreement among words.

Slot2:

This slot contains gender information of every analyzed word. It basically has three possible values, MASC, FEM, and UN, which stands for masculine, feminine and unknown respectively. The value of this slot is used when it is necessary to be reflected on next word based on adjective-noun, adverb-verb and subject-object-verb agreement rules.

Slot3:

This slot contains number information of words analyzed. Possible values for this slot are SING, PLR, and UN, which stand for singular, plural and unknown respectively. The value of this slot is used when it is necessary to be reflected on the coming word based on adjective-noun, adverb-verb agreement and subject-object-verb agreement rules.

Slot4:

Person information of analyzed words is stored in this slot. This slot basically has four possible values, P1, P2, P3, and UN, which stands for 1st person, 2nd person, 3rd person and unknown respectively. The value of this slot is used when it is necessary to be reflected on proposed word according to rules of adjective-noun, adverb-verb and subject-object-verb agreement.

Slot 5:

This slot contains definiteness information of analyzed words. It can have definite and UN values, which stands for definite and unknown respectively. The value of this slot is used when it is necessary to be imitated on the coming word according to subject-object-verb agreement rules.

Figure 4.7 illustrates how a word accepted from a user is represented in five slots and Algorithm 4.8 presents the algorithm to capture morphological features from a user input.

Example: The noun “ልጆች”/”lijochu” is tagged as:-

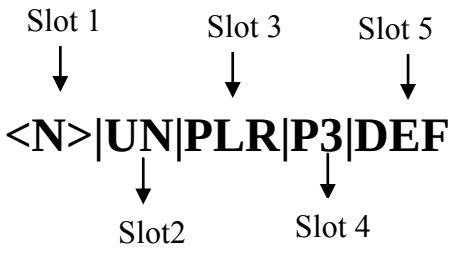

Figure 4.7: Placement of Morphological Features of a Noun “ልጆች”/”lijochu”

```
BEGIN
INPUT last- phrase-analyzed file and POS list
INITIALIZE keywords for root, person, gender, number, definiteness, new-word
INITIALIZE person, number, gender, definiteness, pos, new-word to “UN”
INITIALIZE feature-collection = []
FOR each word in last- phrase-analyzed file
        IF word is in new-word keyword
                ASSIGN word to new-word
        ELSE IF word is in person key word
                ASSIGN word to person keyword
        ELSE IF word is in number key word
                ASSIGN word to number keyword
        ELSE IF word is in gender key word
                ASSIGN word to gender keyword
        ELSE IF word is in definiteness key word
```

```
                ASSIGN word to definiteness keyword
        ASSIGN feature to new-
word+'|'+POS+'|'+person+'|'+number+'|'+gender+'|'+definitness
        ASSIGN person, number, gender, definiteness, pos, new-word to “UN”
        ADD feature to feature-collection
OUTPUT feature-collection for words in the last phrase of user-input
END
```

Algorithm 4.8: Algorithm to Capture Morphological Information from User Input

4.5 Word Sequence Prediction

This module predicts the most probable root or stem words and their morphological features using previously constructed language models. Bi-gram model predicts root or stem word based on previous single word from current position, whereas tri-gram predicts root or stem word based on preceding two words. Hybrid of bi-gram and tri-gram model predicts the future word by considering preceding one or two words.

4.5.1 Root or Stem Word Prediction

Morphologically analyzed user‘s input and previously constructed root or stem words bi-gram, tri-gram and hybrid probabilistic models are used to propose suitable root or stem words. Here, n last root or stem words are fetched from analyzed user‘s input, and then, top highly occurring 15 root or stem words following a given n root or stem words are extracted from the language model, where, n is 1 for bi-gram and 2 for tri-gram model. Algorithm 4.9 describes an algorithm to predict root or stem word.

```
INPUT root-stem-word-model and user-input// bi-gram or tri-gram and user-input
READ last n words from a user-input//n=1 for bi-gram and 2 for tri-gram
INITIALIZE root-stem-keyword
INITIALIZE root-word to “”
ANALYSE the last n word using Hornmorph and WRITE it last-n-analyzed-input file
READ last-n-analyzed-input file
FOR each word in the last-n-analyzed-input
        IF word is in root-stem-keyword
        CONCATINATE word with root-word
        READ root or stem words probability model
FOR each word-sequence in root-stem-word-model
        SPLIT the word-sequence to n
        IF root-word == (n-1)thword or CONCATINATE (n-2)th word with (n-1)th word
                        IF size of proposed-root-words list is <15
                            ADD the nth-word to proposed-root-words list
        OUTPUT proposed-root-or-stem-word list
END
```

Algorithm 4.9: Algorithm to Predict Root or Stem Form of a Word

4.5.2 Morphological Feature Prediction

Proposed root or stem words and previously constructed root or stem words with aspect, voice, tense, prefix and suffix n-gram models are used to propose the most probable morphological feature. Here, each proposed root or stem word is checked for the most frequent aspect, tense, voice, prefixes and suffixes in the language model. In addition the proposed prefix and suffixes needs to be represented in a way that the morphological generator can understand it. For this purpose suggested prefixes and suffixes are calculated in order to denote them in the required format. We have used similar algorithm to predict aspect, voice, tense, prefix and suffixes. Algorithm 4.10 and 4.11 show the algorithms to predict morphological features, specifically aspect, and an algorithm to calculate prefix and suffixes to the required representation respectively.

```
BEGIN
INPUT root-with-aspect n-gram  model and proposed-root-word list
FOR each proposed-root-word in the list
        FOR each root-word in root-with-aspect n-gram model
                IF proposed-root-word equals root-word in the aspect model
                        READ aspect that comes with proposed-root-word
                        ADD aspect to proposed-aspect list
OUTPUT proposed-aspect list
END
```

Algorithm 4.10: Algorithm to Predict Aspect for Expected Words

```
BEGIN
INPUT proposed-prefix list and proposed-suffix list
FOR each prefix in the proposed prefix list
        SPLIT each prefix by '|' and ADD it to pfx1
        FOR each value in pfx1
                ASSIGN value to affix-feature with appropriate representation of
preposition,
        conjunction, negation etc.
FOR each suffix in proposed suffix list
        SPLIT each suffix by '|' and ADD it to predSuffix1
        FOR each value in predSuffix1
                ASSIGN value to affix-feature with appropriate representation
ADD affix-feature to affix-feature list
OUTPUT affix-feature list

END
```

Algorithm 4.11: Algorithm to Calculate Affixes

4.6 Morphological Generation

This module produces surface form of words with a given root or stem and morphological features. Morphological Generator is a program used to produce required form of a word. In this work, Hornmorph morphological generator is used to produce correct words based on the proposed root or stem and morphological features. In addition this module employs Subject-Object-Verb, Adjective-Noun, and Adverb-Verb agreement grammatical rules of Amharic language. Here, to ensure morphological agreement among words, POS of words needs to be considered. Morphology and syntax interact considerably in many languages and languages with rich morphology need to pay attention to interaction of morphology and syntax in order to arrive at suitable computational models. In this study, interaction between syntax, particularly POS and morphology, is used to propose appropriate word form.

POS n-gram models assist to filter proposed root or stem words, through selection of only those root or stem words expected to have required POS. Although there are few researches made in Amharic part of speech tagger, there is no commercial or any available POS tagger in order to incorporate it in our study. Hornmorph analyzes words and assigns their POS to verb group and noun group. Nouns and adjectives are treated the same way in this program. Due to this we used POS tagged corpus for demonstration, so that it helps to keep morphological agreement of our word sequence prediction task.

4.6.1 Subject-Object-Verb Agreement

Amharic subjects and verbs must agree in gender, number, person, and definiteness. If there is a disagreement in one or more of these features, the sentence becomes ambiguous and cannot give proper meaning. Subjects are usually noun, or pronoun, and in this work the most probable subject from user input is identified based on POS of words and order of words. Morphological information like gender, number, person and definiteness of the subject is extracted from morphologically analyzed user‘s input. Subsequently, if predicted word‘s POS is verb, features of the subject are used to inflect it to an appropriate word form.

For example: let‘s assume subject of a sentence is a noun, “አበበ”/”‘abebe”, and the most probable root words are estimated to be verbs “ብልዕ”/”bl‘”, and “ምጥ”/”mT” by the system. The word

ኣበበ"/"abebe" has morphological features of gender: masculine, number: singular and person: 3. Therefore this morphological information is used to generate words from the root word "ብልዕ"/"bl‘", and ምጥ"/"mT". Finally appropriate form of words በላ"/"bela", and መጣ"/"meTa" are offered to a user.

Objects in Amharic language must agree in number, gender, person and definiteness with their respective verbs. Objects are nouns and may or may not be suffixed by an object marker ን"/"n". In this work, object of a sentence is identified using POS and word order. Accordingly its grammatical features like gender, number, person and definiteness are extracted from a user input. Finally, a verb is inflected to appropriate word-form using captured morphological features and by maintaining its agreement with the object.

For example: let‘s assume subject and object of a sentence to be nouns ኣበበ"/"‘abebe", and "ልጁን"/"ljun" respectively. Along with this, we assumed the most possible word a user wants to type is a verb having root or stem form, "ምት"/"mt". The word ኣበበ"/" 'abebe " has morphological features, gender: masculine, number: singular and person: 3 and word ልጁን"/"ljun" has grammatical features of gender: masculine, number: singular, person: 3 and definite. Therefore this captured morphological information is used to generate words from root word, ምት"/"mt". Finally, the word"መታው"/"metaw" is proposed to a user, where the complete sentence becomes ኣበበ ልጁን መታው"/" ‘abebe ljun metaw" and there is object-verb agreement.

To propose words in appropriate word form, agreement among subject, verb, and object should be considered. Here, we analyze a given phrase from user‘s input to identify the most probable subject and object based on word order and POS. Moreover, if proposed word falls in verb category, its morphological features are predicted from the identified subject and object. Subsequently, the predicted feature is used when producing surface words. Algorithm 4.12 describes an algorithm to predict grammatical features of verbs in agreement with subject and object of a given user‘s input.

```
BEGIN
INPUT feature-collection, proposed-root-words
INITIALIZE subject, object to []
FOR each value in feature-collection
        IF value.
                IF length of subject == length of object
                        ADD feature in subject list
                ELSE
                        ADD feature in object list
IF subject list is not null and unknown
        SET subject-feature = last value from subject list
IF object list is not null and unknown
        SET object-feature = last value from object list
SET feature-for-next-word by concatenating subject-feature and object-feature
FOR each word in proposed-root-words list
        IF word has verb-POS-category
                RETURN feature-for-next-word
OUTPUT  feature-for-next-word
END
```

Algorithm 4.12: Algorithm to Propose Features Based on Subject-Verb-Object Agreement

4.6.2 Adjective-Noun Agreement

Adjectives are modifiers of a noun. In Amharic, adjectives should agree with their respective nouns in gender and number. In this work POS is used to identify adjectives from a user‘s input and features like gender and number information is extracted from morphologically analyzed user‘s input. This captured morphological information is used for declension of the first appearing noun after an adjective.

Example: Considering a phrase ”ትልቋ ላም“/“tlqWa lam”, an adjective ”ትልቋ”/” tlqWa” and a noun “lam”/”ላም“ has feminine gender. Therefore, agreement in gender is noticeable between the

adjective and noun. Similarly, there is an agreement in number since both the noun and adjective are in singular form.

To propose words in appropriate word form, agreement between adjective and noun should be considered. Here, we analyze the last phrase from a user‘s to identify if there is a word having an adjective POS. Along with this, if the proposed word falls in noun category and if it is the first noun which appears after the adjective, its morphological features are predicted using features of the adjective, and it is helpful when producing surface words. Algorithm 4.13 is an algorithm used to predict morphological features of a noun which is preceded by an adjective.

```
BEGIN
INPUT feature-collection
FOR each feature in feature-collection
        IF POS of last feature is in adjective group
                GET gender-number-feature
IF gender-number-feature is different from unknown
        ASSIGN feature-for-next-word gender-number-feature
OUTPUT feature-for-next-word
END
```

Algorithm 4.13: Algorithm to Propose Features Based on Adjective-Noun agreement Rule

4.6.3 Adverb-Verb Agreement

Adverbs are modifiers of a verb. In this study, list of time adverbs with their respective tense and probability is used from previous studies on ―Amharic Grammar Checker" [42]. Here, a word is checked if its POS is an adverb and if it is in time adverbs category. Next, mostly occurring tense for that adverb is used on the expected verb.

Example: In a sentence, ―ቆይ ቀድሞ _____/―ዙju qedmo____", ―ቀድሞ/―ɋedmo―is an adverb and it mostly appears with imperfective tense type. Therefore, if we assume the expected root word to be ―ጠT", ―ሃI-meTal"/"ይመጣል" is proposed to a user.

To propose words in appropriate word form, agreement among adverb and verb should be considered. Here, we analyze a given phrase to identify if the last word is time adverb. Moreover, if the proposed word falls in verb category, its most likely tense is predicted using time adverbs probability. Algorithm 4.14 is an algorithm used to predict morphological features of a word given its previous word as time adverb.

```
BEGIN
INPUT feature-collection list, time-adverb list, input-words, proposed-root-words
IF last-word from a user input is in time-adverb list and proposed-root-word has verb
POS
        FOR each word in time-adverb list
                IF last-word from a user input==word
                        ASSIGN tense with highest frequency from time-adverb list to
        tense-feature
OUTPUT tense-feature
END
```

Algorithm 4.14: Algorithm to Predict Tense of a Verb Given Previous Word to be a Time Adverb

4.6.4 Generation of Surface Words

Surface word is a morphologically suitable word that the user intends to type. Surface words are offered to a user using proposed root or stem words, aspect, voice, tense, prefix, suffix and features obtained from grammatical agreement rules described earlier. Algorithm 4.15 presents an algorithm to produce appropriate surface words.

```
BEGIN
INPUT proposed-root-words list
READ proposed-affix-features
READ proposed-aspect, proposed-voice, proposed-tense list
FOR each word in proposed-root-words list
        IF proposed root or stem word is in verb category
                CALCULATE features using subject-verb-object agreement checker,
                adverb-verb checker, proposed affix, aspect, voice and tense
                GENERATE surface-word given root-words and features
                ADD generated word to proposed-surface-words list
        ELSE proposed root or stem word is in noun category
                CALCULATE features using adjective-noun agreement checker,
                proposed affix, aspect, voice, and tense
                GENERATE surface-word given root-word and features
ADD generated-word to proposed-surface-words list
OUTPUT proposed-surface-words list
END
```

Algorithm 4.15: Algorithm to Generate Surface Form of Words

CHAPTER FIVE

EXPERMENT

Prototype development is one of the objectives of this study. Here, prototypes are designed and developed for bi-gram, tri-gram and hybrid of bi-gram and tri-gram models in order to demonstrate as well as evaluate the developed Amharic Word Sequence Prediction Model. This section presents testing data, the implementation, and experimental results.

5.1 Corpus

To evaluate the proposed model, there is a need to have POS tagged testing data since there is no any available POS tagger for Amharic language. The testing is done using Amharic news text having a total of 107 sentences. We couldn‘t conduct the experiment with more test data due to low response time of the predictor. However, we believe the sentences used are representative. Here, 87 sentences are taken from the collected POS tagged corpus and 20 sentences are taken from the training data. Two test cases are prepared, where test case one encloses texts within the training data and test case two contains texts disjoint from the training corpus. Test case one and two contains 20 and 87 sentences respectively. Besides, words found in test case one are manually tagged to their respective POS with the assistance of linguistic experts. Furthermore, spelling errors, wrong POS information and some typographic errors found on the testing data are manually checked and corrected.

5.2 Implementation

Prototype is developed using Python programming language. The main purpose of this prototype is to demonstrate and evaluate the developed word sequence prediction model. Figure 5.1 illustrates prediction result using hybrid model. Components available on these figure are described below.

1. Input area: It is used to accept texts from users.
2. List box: It is used to display list of most probable predicted words.

3 Reset button: It is used to reset entered values in the list box and text box.

Here, users type their text in the input area and when space bar is pressed, most frequently occurring fifteen words are displayed in a list box. Subsequently, a user clicks his or her preferred word from a given list of word options instead of typing each character. However, if the required word is not listed in a given option, then a user continues typing in normal way. In this work, statistical language model is used to predict grammatical features like aspect, voice, tense, prefix and suffixes, in addition to predicting mostly apparent root or stem words. Subject-verb-object, noun-adjective, adverb-verb agreement rules are incorporated while generating surface words. Finally the system offers list of possible surface words with appropriate grammatical features like gender, number, person, aspect, voice, etc.

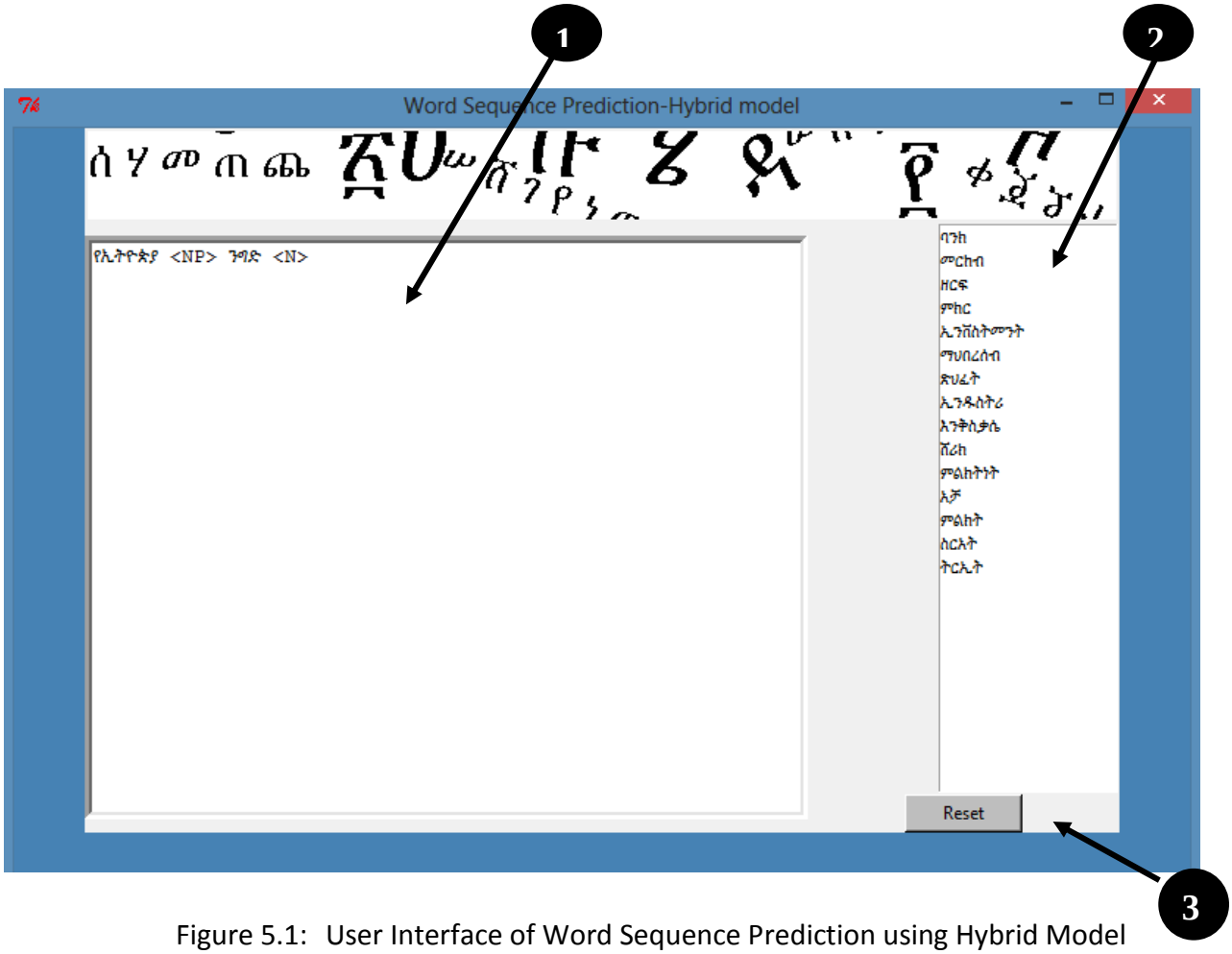

Figure 5.1: User Interface of Word Sequence Prediction using Hybrid Model

5.3 Test Results

The developed models are evaluated in two ways. In the first case, prediction is accepted as appropriate if the proposed words are exactly as needed by a user. In the second way, if root form of the proposed word is proper, then prediction is assumed to be appropriate, even though its word form is wrong. This is done to assess capability of the model to predict root words and morphological features separately. Here, 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 proposals.

The experiment conducted in this research exhibits obtained results based on keystroke savings (KSS) and using bi-gram, tri-gram, and hybrid models. Keystroke Saving (KSS) estimates saved effort percentage which is calculated based on (Eq.1) by comparing total number of keystrokes needed to type a text (KT) and effective number of keystrokes using word prediction (KE).

Table 5.1 shows test result when proposed words are exactly as needed by a user and Table 5.2 illustrate experimental result when root form of proposed word is correct even though its word form is not appropriate.

Table 5.1: Test Result when Proposed Words are exactly as needed by a User

Testing data	**Model**	**KT**	**KE**	**KSS**
Evaluation based on test case1	Bi-gram	2118	1804	14.8%
	Tri-gram	2118	1564	26.1%
	Hybrid of bi-gram and tri-gram	2118	1546	27.0%
Evaluation based on test case 2	Bi-gram	9214	8007	13.1%
	Tri-gram	9214	7608	17.4%
	Hybrid of bi-gram and tri-gram	9214	7322	20.5%

Table 5.2: Test Result When Correct Root Word is Proposed though the Surface Word may not be Appropriate

Test data	Model	KT	KE	KSS
Evaluation based on test case1	Bi-gram	2118	1533	27.6%
	Tri-gram	2118	1124	46.9%
	Hybrid of bi-gram and tri-gram	2118	1099	48.1%
Evaluation based on test case2	Bi-gram	9214	6924	24.8%
	Tri-gram	9214	6568	28.7%
	Hybrid of bi-gram and tri-gram	9214	6281	31.8%

Figure 5.2 shows sample text written with the assistance of word sequence prediction model. The text found in Figure 5.2 contains words in italic and underlined. Underlined words are predicted words exactly as desired by a user and words in italic are words having correct root or stem proposal but wrong word form.

ፌዴሬሽኑ <N> ***ዛሬ*** <ADV> በሰጠው <VP> ***ጋዜጣዊ*** <ADJ> ***መግለጫ*** <N> ***ላይ*** <PREP> የቴክኒክ <NP> ክፍል <N> ዋና <ADJ> ሃላፊ <N> ሻምበል <N> ***ዱቤ*** <N> ***ጅሎ*** <N> እንዳስታወቁት <VP> ከ***አሥራ*** <NUMCR> ***አምስት*** <NUM> ***ቀና***ት <N> በ***ኋላ*** <PREP> ***በሚካሄደው*** <VP> በዚህ <ADJ> ሻምፒዮና <N> ተካፋይ <VP> ***የሚሆኑ*** <VP> 32 <NUM> ***አትሌ***ቶች <N> ***የታቀፉ*** <VP> ***ሲሆን*** <VP> ግማሾቹ <ADJ> ***ሴ***ቶች <N> ***ና***ቸው <V>

Figure 5.2: Sample Text Written with Assistance of Hybrid Model

5.4 Discussion

Word sequence prediction using a hybrid of bi-gram and tri-gram model offers better keystroke savings in all scenarios of our experiment. For instance, when using test data disjoint from the training corpus, 20.5%, 17.4% and 13.1% keystroke savings is obtained in hybrid, tri-gram and bi-gram models respectively. The chance of predicting appropriate root or stem is higher even though it is in wrong word form as shown in Table 5.2. In all cases, KSS is greater when using test data within the training corpus. However, speed of prediction is not considered in this experiment due to the fact that Hornmorph takes much time while analyzing a user‘s text and generating surface words. In cases where Hornmorph couldn‘t generate words that are analyzed using the same tool, we made an assumption to consider root or stem word itself as the right suggestion. Word predictors for English [38], Swedish [37], Hebrew [16] and Persian [20] 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 believe that the result in this work is promising and can be enhanced with addition of more linguistic resources in the language model. In this work, the testing result highly depends on the training data, and due to this the outcome can differ when tested on other training corpus. Rooms for improvement and extension of this work are presented in Section 6.2.

CHAPTER SIX

CONCLUSION AND FUTURE WORK

6.1 Conclusion

In this study, Amharic word sequence prediction model is developed using statistical methods and linguistic rules. Word sequence prediction assists people in their text input means, and there have been a number of researches done on the topic for various languages as briefly stated in Chapter-3. Even though there are diverse linguistic researches in Amharic language, there is no work on the topic of word sequence prediction that considers context information.

This study is set out to suggest the next word to be typed by a user, based on previous history of words. This is done using n-gram statistical models which are developed using Amharic news corpus, and grammatical rules of the language. For this purpose, we built n-gram statistical models of root or stem words, and morphological features like aspect, voice, tense and affixes. In addition, rules of Amharic grammar like subject-verb-object, adjective-noun, and adverb-verb agreement rules are incorporated to predict words in appropriate morphological form. Root or stem words and their respective features are predicted first and then surface words are generated accordingly.

According to our evaluation better Keystroke saving (KSS) is achieved when using a hybrid of bi-gram and tri-gram models. In conclusion, the developed model has potential advantages since an effective word prediction can be carried out using very large corpus size, statistically based techniques, and linguistic rules. We believe that application of this technology is ample, and among them, it has capability to bring benefits of fast text typing to virtual keyboards, portable devices like Smartphone‘s or PDAs, and in assisting people with disabilities.

6.2 Future work

This work can be extended in many ways to optimize the task of Amharic word sequence prediction. The following points are suggested for future work.

- Hornmorph program is a work in progress and it has some limitations. For example, there are words that cannot be analyzed, wrongly processed or cannot be generated at all. Due to this, training done with wrong morphological analysis result brings erroneous prediction output. Therefore using high performance morphological analyzer and generator is recommended to upgrade Amharic word sequence prediction work. This can help to come up with a reasonable speed word sequence predictor and speed of text entry can be considered as an evaluation metrics.
- Lack of POS tagger or sufficient size of POS tagged Amharic corpus makes this task challenging to keep morph-syntactic agreement complete. We used POS tagged test data to evaluate our proposed model. However, Amharic word sequence prediction can be optimized if good Amharic POS tagger is incorporated and if the model is enriched with POS.
- Word sequence prediction requires quite good quality and quantity of training data. A model trained with a corpus having errors offers wrong prediction output. Even though we have used massive size, 125MB, of Amharic text for training, it contains misspelled words and typographic errors which are almost impossible to correct all of them manually.We believe using high quality corpora or automatic spell checker while pre-processing the raw corpus will help to have more proficient Amharic word sequence predictor.
- In this work, when predicting features like aspect, voice, tense and affixes for a given root or stem word, the first highly frequent feature is used, but it is not necessarily correct proposal. Therefore we recommend considering other methods along with highest frequency to make more precise feature prediction in future studies of this topic.
- Keystroke saving is used to evaluate the developed word sequence model in this work. However, other evaluation metrics can also be used and we suggest considering other evaluation metrics in future studies.

REFERENCES

[1] Nega Alemayehu and Peter Willett, –Stemming of Amharic Words for Information Retrieval", Literary and Linguistic computing, 17(1): 1-17,2002.

[2] Atelach Alemu, Lars Asker, Rickard Cöster, Jussi Karlgren, and Magnus Sahlgren, –Dictionary-based amharic-french information retrieval", Springer Berlin Heidelberg, 2006.

[3] Wondwossen Mulugeta, and Michael Gasser, –Learning morphological rules for Amharic verbs using inductive logic programming", *Language Technology for Normalisation of Less-Resourced Languages,* 7, 2012.

[4] Atelach Alemu, Lars Asker, and Mesfin Getachew, –Natural language processing for Amharic: Overview and suggestions for a way forward", In *Proceedings of the 10th Conference on Traitement Automatique des Langues Naturelles*, 2003.

[5] Atelach Alemu, Lars Asker, Rickard Cöster, and Jussi Karlgren, –Dictionary-based Amharic–English information retrieval", In *Multilingual Information Access for Text, Speech and Images*, pp. 143-149, Springer Berlin Heidelberg, 2005.

[6] Abyot Bayou, –Design and development of word parser for Amharic language*",* Master's Thesis, Addis Ababa Univeristy, 2000.

[7] Sisay Fisseha, –Part of speech tagging for Amharic using conditional random fields*"*, In *Proceedings of the ACL workshop on computational approaches to semitic languages*. 2005: Association for Computational Linguistics.

[8] Martha Yifiru, –Morphology-based language modeling for Amharic", PhD diss., Hamburg, Univ., Diss., 2010.

[9] Tesfaye Bayu, –Automatic morphological analyzer for Amharic: An experiment employing unsupervised learning and autosegmental analysis approaches*"*, Master's Thesis, Addis Ababa University, 2002

[10] Atelach Alemu, –Automatic Sentence Parsing for Amharic Text: An Experment using Probabilistic Context Free grammars", Unpublished Master's Thesis, School of Graduate Studies, Addis Ababa University, 2002.

[11] Solomon Teferra, and Wolfgang Menzel, –Automatic speech recognition for an under-resourced language-amharic", *INTER SPEECH*. 2007.

[12] Nicola Carmignani, –Predicting words and sentences using statistical models”, 2006.

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

[14] Nesredin Suleiman and Solomon Atnafu, –Word Prediction for Amharic Online Handwriting Recognition”, Master‘s Thesis, Addis Ababa Univeristy, 2008.

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

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

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

[18] Carlo Aliprandi, Nicola Carmignani, Nedjma Deha, Paolo Mancarella, and Michele Rubino, –Advances in NLP applied to Word Prediction”, 2008

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

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

[21] Keith Trnka, –Adaptive language modeling for word prediction”, In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop*. 2008.

[22] Keith Trnka and Kathleen McCoy, –Evaluating word prediction: framing keystroke savings”, In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 261-264, Association for Computational Linguistics, 2008.

[23] Hisham Al-Mubaid and Ping Chen, ―Application of word prediction and disambiguation to improve text entry for people with physical disabilities (assistive technology)", *International Journal of Social and Humanistic Computing* 1(1):10-27, 2008.

[24] Peter Brown, Peter Desouza, Robert Mercer, Vincent Della-Pietra, and Jenifer Lai, ―Class-based n-gram models of natural language", *Computational linguistics* 18(4): 467-479,1992.

[25] Fredrik Lindh, ―Japanese word prediction", 2011.

[26] Johannes Matiasek, Marco Baroni, and Harald Trost, ―FASTY—A multi-lingual approach to text prediction", In Computers Helping People with Special Needs, pp. 243-250. Springer Berlin Heidelberg, 2002.

[27] Arnab Nandi and HV Jagadish,―Effective phrase prediction", In *Proceedings of the 33rd international conference on Very large data bases*, pp. 219-230, VLDB Endowment, 2007.

[28] Afsaneh Fazly, and Graeme Hirst, ―Testing the efficacy of part-of-speech information in word completion", In *Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*, pp. 9-16, Association for Computational Linguistics, 2003.

[29] Baye Yimam, *Yamarigna Sewasiw (Amharic Grammar),* Addis Ababa, Ethiopia: EMPDA Publications, 1995.

[30] Michael Gasser, ―A dependency grammar for Amharic", In *Proceedings of the Workshop on Language Resources and Human Language Technologies for Semitic Languages, Valletta, Malta*. 2010.

[31] Michael Gasser, ―HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya", In *Conference on Human Language Technology for Development, Alexandria, Egypt*. 2011.

[32] Atelach Alemu and Lars Asker, ―An Amharic stemmer: Reducing words to their citation forms", In *Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources*, 2007

[33] Einat Minkov, Kristina Toutanova, and Hisami Suzuki, ―Generating complex morphology for machine translation", In *ACL*, vol. 7, pp. 128-135, 2007.

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

[35] Atelach Alemu, ―Amharic-English information retrieval with pseudo relevance feedback*"*, In *Advances in Multilingual and Multimodal Information Retrieval*,119-126, Springer Berlin Heidelberg, 2008.

[36] Eyas El-Qawasmeh, ―Word Prediction via a Clustered Optimal Binary Search Tree", *Int. Arab J. Inf. Technol.* 1(1)2004.

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

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

[39] Hisham Al-Mubaid, ―A Learning-Classification Based Approach for Word Prediction", *Int. Arab J. Inf. Technol.* 4(3): 264-271,2007.

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

[41] Javed Ahmed Mahar, and Ghulam Qadir Memon, ―Probabilistic Analysis of Sindhi Word Prediction using N-Grams", *Australian Journal of Basic and Applied Sciences* 5(5): 1137-1143,2011.

[42] Aynadis Temesgen and Yaregal Assabie, ―Development of Amharic Grammar Checker Using Morphological Features of Words and N-Gram Based Probabilistic Methods", IWPT-2013, 2013: p. 106.

ANNEXES

Annex 1: List of Conjunction Suffixes with their Probability

N	Suffix		Probability
0	na	ና	0.6513
1	m	ም	0.2593
2	s	ስ	0.0744
3	nI	ን	0.0083
4	ma	ማ	0.0043
5	sa	ሳ	0.0023

Annex 2: List of Conjunction Prefix with their Probability

N	Prefix		Probability
0	IndI	እንዲ	0.4705
1	lI	ሊ	0.1249
2	sI	ሲ	0.3337
3	IskI	እስኪ	0.0032
4	bI	ቢ	0.0677

Annex 3: List of Preposition with their Probability

N	Preposition		Probability
0	be	በ	0.4998
1	le	ለ	0.1689
2	Inde	እንደ	0.1024
3	ke	ከ	0.1078
4	Iyye	እየ	0.0305
5	I	እ	0.0135
6	wede	ወደ	0.001
7	Iske	እስከ	0.0031
8	sIle	ስለ	0.005

Annex 4: List of POS Tags with their Description

No	POS tag	Description
1	<ADJ>	Adjective
2	<ADJC>	Adjective attached with conjunction
3	<ADJP>	Adjective attached with preposition
4	<ADJPC>	Adjective attached with conjunction and preposition
5	<ADV>	Adverb
6	<AUX>	Auxiliary verbs
7	<CONJ>	Conjunction
8	<ENDPUNC>	Sentence end punctuation
9	<INT>	Interjection
10	<N>	Noun
11	<NC>	Noun attached with conjunction
12	<NP>	Noun attached with preposition
13	<NPC>	Noun attached with conjunction
14	<NUMC>	Number attached with conjunction
15	<NUMCR>	Number cardinal
16	<NUMOR>	Number ordinal
17	<NUMP>	Number attached with preposition
18	<NUMPC>	Number attached with preposition and conjunction
19	<PREP>	Preposition
20	<PRON>	Pronoun
21	<PRONC>	Pronoun attached with conjunction
22	<PRONP>	Pronoun attached with preposition
23	<PRONPC>	Pronoun attached with preposition and conjunction
24	<PUNC>	Punctuation
25	<UNC>	Unclassified
26	<V>	Verb
27	<VC>	Verbs attached with conjunction
28	<VN>	Noun formed from any verb form
29	<VP>	Verb attached with preposition
30	<VPC>	Verb attached with preposition and conjunction
31	<VREL>	Relative verb

Annex 5: SERA Transcription System to Romanize Amharic Language using ASCII

ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
ha	hu	hi	ha	hE	h	ho
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
le	lu	li	la	lE	l	lo
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
Ha	Hu	Hi	Ha	HE	H	Ho
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
me	mu	mi	ma	mE	m	mo
ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
^se	^su	^si	^sa	^sE	^s	^so
ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
re	ru	ri	ra	rE	r	ro
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
se	su	si	sa	sE	s	so
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
xe	xu	xi	xa	xE	x	xo
ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
qe	qu	qi	qa	qE	q	Qo
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
be	bu	bi	ba	bE	b	bo
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
te	tu	ti	ta	tE	t	to
ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
ce	cu	ci	ca	cE	c	co
ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
^ha	^hu	^hi	^ha	^hE	^h	^ho
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ne	nu	ni	na	nE	n	no
ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
Ne	Nu	Ni	Na	NE	N	No
አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
_a	_u	_i	_a	_E	_	_o
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
ke	ku	ki	ka	kE	k	ko
ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
He	Hu	Hi	Ha	HE	H	Ho

ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
we	wu	wi	wa	wE	w	wo
ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
`a	`u	`i	`a	`E	`	`o
ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
ze	zu	zi	za	ze	z	zo
ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
Ze	Zu	Zi	Za	ZE	Z	Zo
የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
ye	yu	yi	ya	yE	y	yo
ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
de	du	di	da	dE	d	do
ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
je	ju	ji	ja	jE	j	jo
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ge	gu	gi	ga	gE	g	go
ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
Te	Tu	Ti	Ta	TE	T	To
ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
Ce	Cu	Ci	Ca	CE	C	Co
ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
Pe	Pu	Pi	Pa	PE	P	Po
ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
Se	Su	Si	Sa	SE	S	So
ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
^Se	^Su	^Si	^Sa	^SE	^S	^So
ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
fe	fu	fi	fa	fE	F	fo
ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
pe	pu	pi	pa	pE	P	po
ሏ	ሗ	ሟ	ሧ	ሯ	ሷ	ቧ
lWa	HWa	mWa	sWa	rWa	sWa	bWa
ቷ	ቿ	ኋ	ኗ	ኟ	ኳ	ኸ
tWa	cWa	hWa	nWa	NWa	kWa	KWa
ዟ	ዧ	ዷ	ጇ	ጓ	ጧ	ቋ
zWa	ZWa	dWa	jWa	gWa	TWa	CWa
ፏ	ፗ	ቈ	ቌ	ቒ		
fWa	pWa	qWa	PWa	SWa		

Declaration

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university.

_____

Tigist Tensou Tessema

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

_____

Dr. Yaregal Assabie