Adama Science and Technology University

School of Graduate Studies

School of Electrical Engineering and Computing

Department of Computing

Synonym Extraction for Afaan Oromoo Words

A Thesis Submitted to School of Graduate Studies of Adama Science and Technology University in Partial Fulfillment for the Degree of Masters of Science Degree in Software Engineering

By: Lamessa Garba

Main Advisor: Dr. Shimelis Aseffa (PhD)

Co-Advisor: Mr. Girma Debele (MSc)

June, 2018

Adama Science and Technology University

School of Electrical Engineering and Computing

Department of Computing

Synonym Extraction for Afaan Oromoo Words

A Thesis Submitted to School of Graduate Studies of Adama Science and Technology University in Partial Fulfillment for the Degree of Masters of Science in Software Engineering

By: Lamessa Garba

Main Advisor: Dr. Shimelis Aseffa (PhD)

Co-Advisor: Mr. Girma Debele (MSc)

Signature of the Board of Examiners for Approval

No.	Name	Signature
1.	Dr. Kula Kekeba (PhD), Examiner	_____
2.	Mr. Girma Debele (MSc), Advisor	_____
3.	_____	_____
4.	_____	_____
5.	_____	_____

Dedication

This thesis work is dedicated to my father, ***Gerba Debelo****, who taught me that the best kind of knowledge to have is that which is learned for its own sake, and to my mother,* ***Birknesh Abdisa****, who taught me that even the largest task can be accomplished if it is done one step at a time.*

Acknowledgment

First and foremost, I need to thank my almighty **God**, for giving me wisdoms, hopes and strength in every situation from the initial to the final progresses of the overall activities of my thesis work. Unless **God** helped me, I am not the person who am I today. However enormous challenges and struggles faced me in every moment in my study life in **Adama Science and Technology University**, **God** gave me a victory over all inconsistencies and irregularities and raises me to the completion of my thesis. Praises and thanks to my almighty **God**; ***"Yaa Waaqa koo galatni kee hin badiin!"***

Many people contributed to this research work either directly or indirectly; **Dr. Shimelis Aseffa** (my main advisor) and **Mr. Girma Debele** (my co-advisor), I thank you very much for your critical and timely comments on my thesis proposal. You are also my model in the field of academia. Dear all my close friends, especially, **Ibsa** and **Amex,** thank you for your inspirations and jocks during title selection and your continuous encouragement in providing me Afaan Oromoo words having numerous synonyms. Dear**, Hiwi**, thank you for your contribution in providing me the way I can get an internet access everywhere.

Dear my best friends and classmates such as **Frezer**, **Take**, **Wonde** and **Bire**, thank you for being my good friends as project/lunch/tea/lab partners during our graduate study period. I hope our friendship and intimacy will even get stronger and persists forever after the campus too. God bless you for your usual cooperation and team works in all the time we spent in **Adama Science and Technology University**. Dear **Frezer,** I never forgot your constructive advices in my life in Adama, especially when things gone ups and downs. Your unforgettable saying **"Nothing is yours on this earth!"** stays and rings in my mind forever. Dear **Malke**, thank you for your 11^{th} hour helps, morals and pray! Generally, to those **friends**, **classmates** and **staffs** I fail to call by your names, I cannot thank you enough.

And last, but not least, my especial thanks, gratitude, appreciations and blessings go to my **families**, for their efforts and contributions in several sides and directions, specifically to **Lali**, **Dad**, **Mom**, **Alo** and my beloved daughter, **Hawi**. Dear **Hawi**, this achievement is for you, even yours at all!

Table of Contents

List of Tables and Figures

List of Acronyms and Abbreviations

AOSE	Afaan Oromo Synonym Extraction
NLP	Natural Language Processing
ULIR	Unilingual Information Retrieval
MLIR	Multilingual Information Retrieval
WWW	World Wide Web
CSA	Central Statistical Agency
GQAO	Gumii Qorannoo Afaan Oromoo
POS	Parts of Speech
LSR	Lexical Semantic Relation
IR	Information Retrieval
BC	Before Christ
QAO	Qubee Afaan Oromoo
CV	Consonant Vowel
TOEFL	Test of English as a Foreign Language
LDOCE	Longman’s Dictionary of Contemporary English
WSD	Word Sense Disambiguation
MFS	Most Frequent Sense

Abstract

The aim of this study is to analyze and determine appropriate contexts of using Afaan Oromoo words and phrases in writing and readings. Even though many users can easily read and write Afaan Oromoo words and phrases, they feel uncomfortable with understanding them and formulating queries for the language. This is either because of their limited vocabulary in Afaan Oromoo, or because of the possible misusages of Afaan Oromoo words. Although Afaan Oromoo is spoken by large number of people (more than 34% of the total population of the country), and rich in words, many advances still have been lacking in a computational linguistics and natural language processing area. Since the numbers of word forms that can be generated from a single stem are so many, Afaan Oromoo language is morphologically very rich and complex when compared to other Latin-based languages. In addition to this, a single Afaan Oromoo word may have different Synonyms (a word or phrase which has the same or nearly the same meaning as another word or phrase in the same language). These and similar evidences made it not only morphologically, but also structurally complex than other languages. Hence, in this study, different analysis of Synonym extraction for Afaan Oromoo words and phrases are studied. Up to our knowledge, no study done before in the area of synonym extraction, and this work is the first of its kind to investigate synonym extraction for Afaan Oromoo language. The proposed system was evaluated using the dataset consisting of 123 words collected from different sources, through considerering all appropriates dialects of Afaan Oromoo. The result from our experiment shows that the lexicon size and rules in the knowledge base play a vital role to recognize the valid synonyms for valid input words. Algorithms and techniques used in this study obtained good performance when compared to the other resource-rich languages like English. The result obtained encourages the undertaking of further research in the area, especially with the aim of developing a full-fledged Afaan Oromoo synonym extraction.

Keywords: Afaan Oromoo, Synonym extraction, words, phrases, spelling error detection, spelling error correction, etc.

Chapter One

Introduction

1. Introduction

In this fast-growing age of technology, language is one of the fundamental aspects of human behavior and constitutes a crucial role in the overall progresses of technological improvements. In its written form, it serves as a means of recording information and knowledge on a long-term basis and transmitting what it records from one generation to the next. In its spoken or verbal form, it serves as a means of coordinating our day-to-day activities with others (Allen, 1996). Linguistics can be defined most simply as the scientific study of languages, particularly, of the natural languages.

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language texts and speeches to do useful tasks. NLP researchers aim to gather knowledge on how human beings understand and use languages, so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform desired tasks. The foundations of NLP lie in several disciplines, namely, Computer and Information Sciences, Linguistics, Mathematics, Software Engineering, Electrical and Electronic Engineering, Artificial Intelligence and Robotics, Psychology, etc. (Steven et al., 2009).

Applications of NLP also include several fields of study, such as machine translation, natural language text processing and summarization, unilingual and multilingual information retrieval (ULIR and MLIR), speech recognition, artificial intelligence and expert systems.

One important application area that is relatively new and has not been covered yet in NLP tends to the explosion of the unilingual text processing, namely synonym extraction. This study is therefore intended to analyze extraction of synonyms for Afaan Oromoo language words within the language (Afaan Oromoo). Several researchers have pointed out the need for appropriate research in this specific research area, through facilitating multilingual or cross-lingual information retrieval, which includes multilingual text processing and

multilingual user interface systems. For instance, to exploit the full benefit of different services such as World Wide Web (www) and digital libraries, applications of NLP play vital roles.

1.1. Background

In every language, whether it is spoken or written, the entire meaningful pattern has its own structure and format. Thus, the elements of the language should relate to each other in understandable manner (Abebe, 2013). These relationship requirements together with the existence of large number of classes in natural language make the processing tasks so complicated in NLP.

When speaking of a language as a general concept, definitions can be used which stress different aspects of the phenomenon. These definitions also entail different approaches and understandings of a language, and they inform different and often incompatible conservatories of linguistic theory. Debates about the nature and origin of language go back to the ancient world. Greek philosophers such as Gorgias and Plato debated the relation between words, concepts and reality (Banca, 2016). Gorgias argued that language could represent neither the objective experience nor human experience, and that communication and truth were therefore impossible. But, Plato maintained that communication is possible because language represents ideas and concepts that exist independently and prior to the language.

According to the 2007 Census, Ethiopia has approximately eighty-eight (88) different languages with up to two hundred (200) different dialects spoken over the country. The major and largest ethnic and linguistic groups among these languages are Afaan Oromoo, Amharic, Somali and Tigrigna respectively. Tigrigna and Amharigna (also called Amharic), are the major languages, which are derived from Ge'ez syllabuses (alphabets). Ge’ez is the ancient language and was introduced as an official written language during the first Aksumite kingdom, when the Sabeans sought refuge in Aksum. The Aksumites developed Ge’ez, a unique script derived from the Sabean alphabet, and it is still used by the Ethiopian Orthodox Tewahedo Church. Amharic language is the official national language of the country, Ethiopia. Languages such as English, Arabic, Italian and French are other languages widely spoken by many Ethiopians in the country.

Of the languages spoken in Ethiopia, about eighty-six (86) are living and two (2) are extinct. Forty-one (41) of the living languages are institutional, fourteen (14) are developing, eighteen (18) are vigorous, eight (8) are in danger of extinction, and five (5) are near extinction (Charles, 1976). According to the CSA 2007 report, the population size of Ethiopia is estimated to 73.9 million. The following table summarizes distribution of Ethiopian languages based on the CSA carried out in 2007:

Language Name	Percentage (%)
Afaan Oromoo	33.8%
Amharic	29.33%
Somali	6.25%
Tigrigna	5.86%
Sidamo	4.04%
Wolaytta	2.21%
Gurage	2.01%
Afar	1.74%
Hadiya	1.69%
Gamo	1.45%
Others	11.62%

Table 1.1: Distribution of Ethiopian languages as of 2007 GC

Ethiopian languages are also characterized by shared grammatical and phonological features (Charles, 1976). This language area includes the Afro-asiatic languages of Ethiopia, not the Nilo-Saharan languages. In our country, Ethiopia, English is the most widely spoken foreign language and is the medium of instruction in secondary schools and Universities. On the other hand, Amharic is the national language of the county and being used in the country's regions as a primary and secondary school's educational language instructions. But, nowadays, it has been replaced in many areas by own local languages such as Afaan Oromoo and Tigrigna.

After the fall of the Derg regime in 1991, the new constitution of the Federal Democratic Republic of Ethiopia granted all ethnic groups the right to develop their languages and to establish mother tongue languages as primary school education's languages. This is a marked change to the language policies of former governments in Ethiopia.

In terms of writing systems, Ethiopia's principal orthography is Ge'ez or Ethiopic syllabus that employed as an Abugida for several of the country's languages. It first came into usage in the 6th and 5th centuries BC to transcribe the Semitic Ge'ez language. Ge'ez now serves as the liturgical language of the Ethiopian and Eritrean Orthodox Churches. Other writing systems have also been used over the years by different Ethiopian communities. These include Arabic script for writing some Ethiopian languages spoken by Muslim populations and Sheikh Bakri Sapalo's script for Afaan Oromoo. Today, many Cushitic, Omotic, and Nilo-Saharan languages are written in Roman and Latin scripts.

Basically, the Ethiopian languages are divided into four major language groups or families. These are Semitic, Cushitic, Omotic, and Nilo-Saharan.

The Semitic languages are spoken in northern, central and eastern Ethiopia (mainly in Tigray, Amhara, Harar and northern part of the Southern Peoples' State regions). They use the Ge'ez script that is unique to the country, which consists of thirty-three (33) letters, each of which denotes seven (7) different characters, making a total of about two hundred thirty-one (231) different characters.

The Cushitic languages are mostly spoken in central, southern and eastern Ethiopia (mainly in Afar, Oromiyaa and Somali regions). The Cushitic languages use the Roman alphabets and Latin scripts. For example, Afaan Oromoo is written in the Latin script, whereas Somali is written in the Roman alphabet. The Omotic languages are predominantly spoken between the Lakes of southern Rift Valley and the Omo River, while Nilo-Saharan languages are largely spoken in the western part of the country along the border with Sudan (mainly in Gambella and Benshangul regions).

Afaan Oromoo (also known as Oromiffa) is one of the major languages that are widely used in Ethiopia. Currently, it is a working language of Oromiyaa national regional state (which is the largest region or state in Ethiopia). Unlike other languages such as Amharic, (another major language and currently the working language of the country, which belongs to Semitic family), Afaan Oromoo is part of the lowland east Cushitic group within the Cushitic family of the Afro Asiatic phylum (Kula, 2008). In this Cushitic branch of the Afro Asiatic language family, Afaan Oromoo is considered as one of the most extensively spoken language among the forty or so-called Cushitic languages.

According to the 2007 Census conducted by the Central Statistical Agency of Ethiopia (CSA), Afaan Oromoo is the largest and first language spoken with about twenty-five (25) million speakers (thirty-four percent (34%)) of the total population of the country. In addition to this, Afaan Oromoo is also spoken in neighboring countries other than Ethiopia, such as Somalia, Kenya, Eritrea, Sudan, Uganda, Djibouti, Tanzania, South Africa, etc.

About 95% of Afaan Oromoo speakers live in Ethiopia, mainly in Oromiyaa Region. In addition to this, in Somalia, there are also some speakers of the language. In Kenya, the Ethnologic study lists seven hundred twenty-two thousand (722,000) speakers of Borana and Orma (the two languages closely related to Afaan Oromoo).

Within Ethiopia, Afaan Oromoo is the language with the largest number of native speakers. Within Africa, it is the language with the fourth most speakers, after Arabic, Kiswahili and Hausa (Getachew et al., 2011). Besides the first language speakers, the number of members of other ethnicities who are in contact with Oromoo (for instance, the Omotic speaking Bambassi and the Nilo-Saharan speaking Kwamain northwestern Oromiyaa) speak it as a second language (Banca, 2016).

1.2. Statement of the Problem

Development of technology is highly related with the development of its corresponding language. The fact that initiated this study is also enabling development of Afaan Oromoo to grow with current Information Technology support systems.

Afaan Oromoo uses Latin based script called, Qubee and has 26 basic characters (G.Q.A. Oromoo, 1995 and Gamta, 1992). Currently, it is the official language of the Oromiyaa regional state and academic language for primary schools of the region. In addition to this, Afaan Oromoo language, literature and folklore delivered as a field of studies in many Universities located in Ethiopia and other countries. Nowadays journals, books, magazines, newspapers, online education, different Medias, videos and pictures are available in electronic format both on the internet and as offline resources. There is huge amount of information being produced with this language, since it is becoming the language of educational and research, language of administration and political welfares, and language of ritual activities and social interaction.

Among more than 80 languages existing in Ethiopia, as stated in the background, Afaan Oromoo is the language with largest number of speakers under the Cushitic family. However, Afaan Oromoo is being spoken by such large number of people in the country, few advances still have been made in computational linguistics and natural language processing in the language. Among these advancement, synonym extraction, voice recognition, machine translation, natural language text processing and summarization, user interfaces, multilingual and cross-language information retrieval systems, speech translation, artificial intelligence, and expert systems are some of the languages' critical issues that need immediate considerations.

Computational approaches to linguistic analysis of Afaan Oromoo so far have been hindered due to non-availability of well-studied linguistic resources. This scenario has recently started to change, primarily through research works and graduate theses that are being carried out in different Universities in the countries and abroad, especially at Addis Ababa University.

To this end, this study tries to answer the following research questions:

- What are the key features of Synonym Extraction in Afaan Oromoo words and phrases writing system?
- What are the challenges and inconsistencies of implementing Synonym Extraction for Afaan Oromoo words and phrases?
- To what extent is Synonym Extraction for Afaan Oromoo words and phrases satisfy the users?

1.3. Motivation behind the study

Since most computers work using English and other few languages, people who do not speak, listen, hear and write such language are either forced to access computers in those languages or miss not to use them at all. This has its own impact on the socio-economic development of the country. To increase the usability of computer devices and let people express their ideas using their native languages, these devices need to be localized into native languages. Hence, it is necessary to do researches to alleviate these problems. One of the related research areas which demand the researchers' attention is therefore the natural language processing (NLP) applications, that deals machine translation, natural language

text processing and summarization, unilingual and multilingual information retrieval (ULIR and MLIR), speech recognition, artificial intelligence and expert systems and other semantic relationships of words in the language, namely Synonym Extraction.

Synonyms and thesauruses was developed for languages like English, Chinese, Japanese, Arabic and Thai. Several commercial and non-commercial Synonyms and thesauruses, such as the ones embedded in Microsoft Word and other variants, have been extensively studied. However, their techniques still limited in their scope to a few languages. For example, if we take Microsoft Word it detects all Afaan Oromoo words as the spelling error and it tries to give a false suggestion based on the English dictionary. Since anyone who needs to process Afaan Oromoo using Microsoft Word uses English as the default language, every word that is written is underlined with red color, which makes it difficult to differentiate words which are wrongly spelled.

Afaan Oromoo is used in offices, schools and media. Hence, huge amount of electronic data was produced on each day. So, the development of NLP applications for this language is required to cope up with the current technologies of NLP. With the facts and view of having electronic Synonyms Extraction for the country's office workers, we motivated to do a research on developing Afaan Oromoo Synonyms Extraction (AOSE). Besides these, this study is inspired by the contribution of spell checker tasks to the language.

1.4. Objectives

1.4.1. General Objective

The general objective of the study constitutes extraction of synonymous words from/to a valid and existing Afaan Oromoo words and phrases.

1.4.2. Specific Objectives

To achieve the general objective, the following sets of specific tasks and activities are carried out:

- Review of relevant works in other Latin based natural languages such as English,
- Study of Afaan Oromoo existing tools for the language's data preprocessing, its writing systems and how it is represented in the computer,

- Exploration of the existing word-based predictive models, and corpus upon which the models can be built,
- Identification and adoption of one of the appropriate prediction models depending on the original word provided for the new word's Synonym Extraction,
- Development of the necessary designs and prototypes for Afaan Oromoo words Synonym extraction,
- Evaluation of the performance and effectiveness of Afaan Oromoo words Synonym Extraction, etc.

1.5. Scope and Limitation of the study

The scope of the study is highly limited to providing appropriate Synonyms for correctly spelled and valid Afaan Oromoo words. In English language, this feature is properly working in different ways, especially being integrated with different word processing applications and various versions of Microsoft office words.

In most languages, Synonym extraction for words and phrases of the language is often followed by checking the corresponding spelling of that word or phrase. Although the spell checker and Synonym extraction (Synonym provider) for a given word or phrase is codependent, the study mainly concentrated on accomplishment of Synonym extraction for the given word or phrase. But, the study did not deal with other features such as grammar checkers, stemmings, parts of speech tagging and other different computational linguistics and natural language processing applications, since the focus of the study is on the language's words and phrases.

1.6. Significance of the study

Development of a language is directly related with adaptation of emerging technology and making it suitable to the local languages. Speakers of a certain language are not more interested in speaking language of technology by themselves; rather they need technology to speak their own language. If the language managed to grow with technology, speakers of the specific language will be benefited from the development in many perispectives. That is why localizing works already done in developed languages, like English is important.

Generally, the study has the following significances:

- To allow learners to learn unknown Afaan Oromoo words that appears in readings, rather than just a restricted set of words in a list of target vocabulary.
- Using word's synonym feature can help us to add more varieties and alternatives to our writing and suggest words and phrases that our readers can better understand than the words we are uncertain of them.
- To open the ways for other researchers to focus on the area and work on new Synonym extraction from the existing words and phrases of Afaan Oromoo, and to push as stepping stone for further works in this specific research area.
- Enables Afaan Oromoo speakers retrieving text documents in Afaan Oromoo efficiently and effectively from different sources.
- Improves the development of the language through initiating oneself by providing the way they able to be familiar with Afaan Oromoo words and know new words either directly or contextually.
- Since the work or study is master's level thesis, it has also learning out comes for the researchers. It helps the researchers to investigate problems and solving it in scientific manner.
- Using word's synonym, in addition to adding more variety to our writing and word suggestions, can also help us to get more alternatives to the concepts that the readers cannot better understand than the words we are certain of them. This feature is very important concern especially in medical related tasks.

1.7. Organization of the study

The remaining content of the thesis is organized as follows:

Chapter 2 discusses literature review on different issues related to synonym extraction. In the chapter, issues related with synonym extraction, and the different techniques and approaches to develop synonym extraction are discussed. The chapter is also devoted to discuss related works done for different languages. Morphological properties of Afaan Oromoo and other language specific issues such as the writing systems, syllable structures, word formation processes like inflections, derivations and compounding have also been extensively presented in this second chapter.

Chapter 3 discusses design requirements, techniques, methods and approaches, the architectural and design issues of AOSE. This chapter also describes stem classification issues, signature building, algorithms and implementation of rules.

Chapter 4 is devoted to the the prototype and evaluation of the system, the last chapter, Chapter 5 concludes the thesis by outlining the benefits obtained from the research work and limitations of the system. It also shows some research directions and recommendations that can be accomplished in developing a fully fledged AOSE for future works.

Chapter Two

Literature Review

2. Literature Review

In this chapter, we concentrated on the general overview of Afaan Oromoo language features and components. The theoretical frameworks and fundamental aspects of the languages that initiate the work and links the study to its core objective is touched in detail.

In section 2.1 of this chapter, the fundamental concepts of synonymies and their applications are covered. In 2.2, the theoretical frameworks and overview of Afaan Oromoo that constitutes Afaan Oromoo writing system, syllables and dialects are roughly studied. In 2.3, parts of speeches (POS) of Afaan Oromoo words are studied. This section plays a vital role in the accomplishment of the overall study. Before proceeding working with words, it is better to have at least the fundamental knowledge and understandings about the classifications of words, namely their corresponding parts of speeches. Basically, five different parts of speeches of Afaan Oromoo words, namely nouns, adjectives, verbs, adverbs and prepositions are elaborated in this section. In 2.4, the comparatives and superlative forms of Afaan Oromoo words are studied in details.

The structural property of the language, namely the morphology of Afaan Oromoo words are also explained in 2.5. Lastly and importantly in this chapter, the related works, however there is no thesis done previously in Afaan Oromoo synonyms, are studied comparatively. As there is no previously done works in Afaan Oromoo synonym extraction for words, this work dealt with the development models and approaches established from the scratch.

2.1. Synonymy and its applications

Synonymy is one of the lexical semantic relations (LSRs) of languages, which are the relations between meanings of words. By the definition, synonyms are ‘one or two or more words or expressions of the same language that have the same or nearly the same meaning in some or all senses’ (Mish, 2003). Despite its importance in various NLP applications, the

task of synonym extraction remains challenging without satisfactory results in the NLP community.

In contrast to other lexical semantic relations (LSRs), synonymy closely associates different lexicalizations of the same concept, which is a unique and useful property in many NLP applications (Mohammad and Hirst 2006). As one of the many successful examples, conflate words into concepts (represented by thesaurus categories) when exploring their co-occurrence patterns. The resulting concept–concept matrix is compacted to approximately the smaller size of the lexical co-occurrence matrix and has proven very effective in measuring lexical semantic similarity. (Bikel and Castelli, 2008) applied synonyms in their event matching system, in which each lexical item is augmented by its synonyms from a thesaurus. If a surface-form match fails, the two-tier model will back off to a synonym match to improve coverage at a low cost in accuracy. In information retrieval and query expansion, using synonymy can help improve search results by substituting and expanding user queries with synonyms (Mandala et al., 1999), since the wording for the same topic varies greatly among users (e.g., car broker versus auto dealer).

Thesauri are obviously the most common sources for synonyms (e.g., Roget, 1911; Fellbaum, 1998). While such hand-crafted resources usually guarantee high quality of synonymy among entries, thesauri also exhibit various limitations when used in NLP applications, such as the amount of human effort involved in building thesaurus, fixed domain coverage, and limited availability. (Mandala et al., 1999) argued that when used in query expansion, manually constructed thesauri exhibit various problems, such as low convergence and domain incompatibility with the document collection in question. Their study showed that different types of thesauri, manually or automatically constructed, all have their own limitations, but Information Retrieval (IR) system performance is almost doubled when different sources of synonymy are combined, indicating the necessity for automated processes for synonym extraction.

Another application, related to synonym extraction is lexical substitution (McCarthy and Navigli 2009), which is useful for many NLP related tasks, such as question answering, summarization, and paraphrase acquisition. For given instances of words with contexts, synonyms to the senses of words in the given contexts are suggested and compared to answers elicited from humans. Thus, in addition to extracting synonyms for a given word,

lexical substitution system must also disambiguate senses of the word according to different contexts, which is a challenge beyond the scope of this study.

2.2. Theoretical frame works: A brief overview of Afaan Oromoo

Afaan Oromoo is a Cushitic language family that is spoken mostly in Ethiopia by about half of the country's overall population. It is also spoken in Kenya, Somalia and Djibouti and is the fourth largest language in rank in Africa after Arabic, Kiswahili, Hausa languages (Getachew et al., 2011). Afaan Oromoo, (meaning Oromoo language) also called Oromiffaa or simply Oromoo, most probably rates the second among the African indigenous languages. Even before two decades when Geda (Melbaa Gadaa, 1988) explored the history of this nation, Afaan Oromoo language was known as the mother tongue of about 30 million Oromoo people living in the Ethiopian Empire and neighboring countries. For this huge language, Afaan Oromoo, there has been a single attempt made so far as far as the information we have from the literature review on related works in Afaan Oromoo language. (Getachew et al., 2011) used Hidden Markov Model on a very few dataset and limited tag sets to develop part-of-speech tagging for Afaan Oromoo, which indirectly supports the application of Synonym extraction for words of the language.

The Oromoo people constitute a single largest ethnic group in Ethiopia, where the Oromiyaa region contains a third of Ethiopia's land area and population. The Oromoo language, namely Afaan Oromoo, is spoken as a first language by 87% of Oromiyaa's 27 million people (CSA 2007). In other regions of Ethiopia, more than 1.4 million people use it as their mother tongue, meaning native Oromoo speakers of about 25 million in Ethiopia. About more than 500,000 people live in Kenya and Somalia. Many others (yet not quantified) speak it as a second language. It is the most spoken language and has the largest ethnic group in the Cushitic family, which also includes Somali, Sidamo, and Afar languages.

Cushitic peoples were present on the central Ethiopian plateau as early as 5000 B.C. They spread to settle most of North-East Africa before the arrival of Semitic speakers around 1000 B.C. As Cushitic groups migrated south and east, five linguistic groups were formed: Northern (e.g., Beja), Central (e.g., Agaw), lower Eastern (e.g., Oromoo, Somali, Saho, Afar, Konso), highland Eastern (e.g., Kambatta, Haddiya, Walayitta), and Southern (in present-day Tanzania) (Bancaa, 2016). Other than this, little is known about the origins of the Oromoo people. There are slightly few written records of the Oromoo people until they

came into conflict with southern expanding Abyssinia in the sixteenth Century. Oromiyaa existed as several independent kingdoms and nations until the Abyssinian conquest of the 1880s, at which point Oromiyaa became part of the Ethiopian Empire. Many of the previous Oromoo kingdoms became provinces of Ethiopia, such as Jimma, Illubabor, Wollega, Shewa, Arsi, Bale, and Hararge. These areas had developed distinct dialects of Oromoo which are still present.

2.2.1. Afaan Oromoo writing system

Afaan Oromoo writing system is a modification to Latin writing system. Thus, the language shares a lot of features with English writing with some modification (Diriba, 2002). The writing system of the language, which is known as "Qubee Afaan Oromoo (QAO)", or shortly "Qubee", is straight forward which is designed based on the Latin script. Thus, letters in English language are also in Afaan Oromoo except the way it is written. As a result, the use of computers for processing information and storing huge number of repositories with this language in governmental and non-governmental organizational offices and institutions in the region has been growing extremely. So, it sounds to process the linguistic form of the language with computers to understand and use the language for communication and transformation of knowledge in this digitized world.

2.2.2. Syllables in Afaan Oromoo

Every language has its own structure of syllables. For example, in English more than two consonants can come consecutively in a single word as in 'syllable'. But, in Afaan Oromoo, more than two consonants cannot come together except in diagraphs (doubled characters such as ch, dh, ny, sh, ts). Hence, there are four types of syllable structures in the language (Abebe, 2013). These structures include CV, CVV, CVC and CVVC. All of these can be found at words initials, median or lands at the final positions.

A valid word can be composed from the combination of one or more type(s) of these structures. The words like **aadaa, eelaa, eegee,** etc seem to start with vowels, but linguists argue that there is hidden glottal stop called hudhaa (') in front of any word that seems to start with vowel. Considering this, we say that every syllable in Afaan Oromoo starts with consonant. The following are just few examples:

CV	Nama (man)
CVV	Hoolaa (ship)
CVC	Shan (five)
CVVC	Loon (cows)

2.2.3. Afaan Oromoo dialects

The main dialects of Oromoo are **Wollega** (spoken in the West Wollega, East Wollega, Illubabor, and Jima zones), **Tulama** (in the North, West, and East Shewa zones), **Wello** (in Northern Shewa and Southern Amhara), **Arsi** (in the Arsi and Bale zones), **Harar** (in the West and East Harerge Zones), and **Borena** (in the southern most zones by the same name, Borena) (Tilahun, 1992). This classification scheme is general, as there is no official division of Oromoo dialects, and many dialects go by multiple names (Wollega is also called Mecha). Additionally, more isolated dialects are spoken by small Oromoo communities that remain in eastern Amhara and southern Tigray. The major differences in dialects are in the form of pronouns, certain verb conjugations, and informal lexicon.

In this study, the Wollega dialect, which constitutes Illubabor and Jimma zone's dialects, is mostly applied. Certain features of Wollega Oromoo summarized below:

i. The Wollega dialect predominantly uses **"koo"** to mean **'my, mine',** whereas other dialects use **"kiyya"** and in some cases **"tiyya"** for the feminine (Wollega possessive pronouns do not distinguish by gender).

ii. **'She'** in Wollega Oromoo is **"isheen",** while other regions in Oromiyaa would use **"isiin", "ishiin",** or **"iseen".**

iii. **'We'** in Wollega Oromoo is most often **"nuti",** while other dialects prefer **"nu'i", "nuhi",** or normally **"nu".**

iv. The use of **"dha"** to mean **'is/am/are'** is much more common in Wollega Oromoo than in eastern dialects. In Harar Oromoo, for example, it is only used for emphasis.

v. Where Wollega Oromoo would use an apostrophe between vowels in a word, other dialects may use an **h, w,** or **y.** Thus, **"ta'e"** may in some places be spelled (and pronounced as) **"tahe", "haasa'uu"** as **"haasawuu",** and **"waa'ee"** as **"waayee".**

vi. The suffixes for present future tense verbs in the second person plural and third person plural are ***-tu*** and ***-u,*** respectively, while the eastern dialects use ***-tan(i)*** and ***-an(i).***

Other dialectal differences are noted in the text sentences. The dialects also differ using Amharic and Somali loan words, with the Wello dialect, being the most mixed of the Oromoo dialects with Amharic. Larger towns will also show a considerable amount of Amharic influence. In addition to synonym extraction, our effort in this study has been made to keep this text **"pure"** Oromoo, though the readers of this study should keep in mind that some Oromoo words have been mostly supplanted by their Amharic equivalents. The major examples include: **"ishi"** instead of **"tole"** for **'alright/ok', "baqqaa"** for **'enough', "guwaadenyaa"** for **'friend', "gobez"** for **'smart/clever',** and "**ayzo"** for **'be strong/take heart'.**

There has been some debate on the mutual intelligibility of Oromoo dialects, with many sources claiming that not all dialects can be understood by all Oromoo speakers. However, when (Tilahun, 1992) reflected on writing his Oromo-English dictionary, he recalled: "I visited Arsi, Bale, Gamu Goofaa, Goojjam, Harar, Kafaa, Shaggar, Sidaamo, Wallaggaa, and Wallo. I did not have to visit Ilu Abbaa Booraa, my birthplace. Due to my own reasons, I could not go to Tigray to interview Raayyaa, Azabo, and Waajiraat Oromos, either. However, I stayed in Waldiyaa, Wallo, overnight, where I had an opportunity to chat with an elderly Raayyaa Oromoo. Despite of some minor differences in the way of our pronunciation, **kaleesha/kaleecha/kaleessa,** meaning **"yesterday"**, for instance, we could understand each other very easily. After he told me, with a clear expression of concern on his handsome face, that the younger generation must be taught Afaan Oromoo and be urged to use it, and he said **nagaatti** (good bye) and left. In addition, when I was attending a conference in Nairobi in 1972, I had the opportunity to measure the situation in Kenya where about half a million Oromoos live. After these visits, I concluded that the pronunciation used by Oromoo in both Oromiyaa and Kenya is almost identical at the lexical level. The so called widespread and alarming rumor that there were wide regional variations in Afaan Oromoo, I became convinced, it was baseless."

2.3. Parts of Speech (POS) of Afaan Oromoo (Garee Jechootaa)

2.3.1. Nouns (Maqaa)

Nouns words in most languages are names given to a person, place, thing, idea, animal, quality, or activity (Abraham, 2014). Since lexical classes like nouns are defined functionally (morphologically and syntactically), some words for people, places, and things

may not be nouns, and conversely some nouns may not be words for people, places, or things. So, based on its contextual position, nouns are commonly found at the beginning of a sentence in Afaan Oromoo. For instance, words that are italic in the following sentences are nouns:

Afaan Oromo:	**English**
Abdiisaan qonnaan bulaadha.	***Abdisa*** is a farmer.
Eebbisaan qotiyyoo lama qaba.	***Ebisa*** has two oxen.

Words that are categorized as noun in a sentence could be a subject or an object. The subject mostly comes at the beginning of the sentence, whereas an object comes after the subject and in front of the verb in a sentence. Let's see the position of subjects and objects in the following sentences:

Afaan Oromoo:	**English**
Abdiisaan qonnaan bulaadha.	Abdisa is a farmer.
Eebbisaan qotiyyoo lama qaba.	Lelisa has two oxen.

In the above two sentences, Abdiisaan and Eebbisaan are subjects and qotiyyoo and qonnaan are objects of the sentences. The words qonnaan and qotiyyoo found in front of the verbs dha and qaba after the subject Abdiisaan and Eebbisaan.

In Afaan Oromoo, nouns are also characterized in singular and plural numbers. Singular nouns are nouns which are quantified and not add any suffixes in the language. For instance, *farda 'hourse', re'ee 'goat', nama 'man/women', mana 'house'* and so on. All of them indicate single entity or animal or person. Quantified plural nouns in Afaan Oromoo can be formed in two ways. The first way is, by adding numerals after singular nouns. For example, *farda lama 'two horses', re'ee sadii 'three goats', nama shan 'five men'* and so on. Plural nouns are also formed through the addition of suffixes in the language. The most common plural suffixes in Afaan Oromoo include: *–oota, -ota, -wwan, -een, -a(an), -lii,* and *-lee*.

Look the following table for examples:

Singular	Plural
Nama (man)	Nam**oota** (men)
Mucaa (child)	Mucoo**lii**(children)
Hojii (work)	Hojii**lee** (works)
Maatii (family)	Maati**wwan**(families)
Mana (house)	Mann**een** (houses)
…	…

Table 2.1: Plural forms of Nouns that are formed from suffixes

When the idea of Afaan Oromoo noun forms of words rose, pronouns (bakka maqaa or qooda maqaa) should have to be noted. Pronouns, not only in Afaan Oromoo, but also in some other languages such as English and Amharic, are words that are used instead of nouns or noun phrases. Afaan Oromoo pronouns are also used in place of nouns as pronouns of other languages. Like nouns, pronouns also characterized based on number and gender. For instance, *ishee 'her', isa 'him', isaan 'they'* are some of the characteristic features of pronouns.

There are different types of pronouns in Afaan Oromoo language based on their function and meaning in a sentence. These are personal (subjective) pronouns, possessive pronouns, demonstrative (objective) pronouns, and reflexive (reciprocal) pronouns. Personal pronouns are *ani 'I', ati* (singular) *'you', inni 'he', isiin/ishiin 'she', isin* (plural) *'you', nuy/nuti 'we'* and *isaan 'they'* are used in the subject position of nouns in the sentences. For instance, the following table summarizes the basic categories of pronouns in Afaan Oromoo:

Subjective Pronouns		Objective Pronouns		Possessive Pronouns	
English	**Oromoo**	**English**	**Oromoo**	**English**	**Oromoo**
I	Ani	Me	Na	My, Mine	Koo
We	Nuti	Us	Nu	Our, Ours	Keenya
You	Ati	You	Si	You, Yours	Kee
You	Isin	You	Isini	You, Yours	Keessan(i)
He, It	Inni	Him, It	Isa	His, Its	(i) Saa
She	Isheen	Her	Ishee	Her, Hers	(i) Shee
They	Isaan	Them	Isaani	Their, Theirs	(i) Saanii

Table 2.2: Major types and forms of Pronouns in both English and Afaan Oromoo

2.3.2. Adjectives (Ibsa Maqaa or Maqibsa)

Adjectives in Afaan Oromoo describes nouns to denote quality of a thing; that is, it specifies to what extent a thing is as distinct from something else (Noah, 1979). For example:

Tolaan *gurracha* 'Tola is black'
Inni *jaarsa* 'he is old'
Aliin qotee bulaa *jabaadha* 'Ali is strong farmer'etc.
In the above sentences, all words (*guracha 'black', jarsa 'old'*, and *jabadha 'strong'*) are adjectives. Adjectives in Afaan Oromoo sentences can be formed from original adjectives and compound words. For instance, as mentioned above, *gurracha 'black', diimaa 'red', dheeraa 'long'* and so on are original adjectives. *Humna dhabeessa 'weak', bifa dhabessa 'ugly', simboo qabeessa 'handsome'* are some of the adjectives formed from compound words. Adjectives also characterize number and gender in Afaan Oromoo like that of nouns and pronouns. The following examples indicate adjectives in numbers and gender:

Singular	**Plural**	**Masculine**	**Feminine**
Gabaabaa	Gaggabaaboo	Gabaabaa	Gabaabduu
Gurracha	Gugurraalee	Gurracha	Gurraattii
Dheeraa	Dhedheeroo	Dheeraa	Dheertuu
Jarjaraa	Jarjartoota	Jarjaraa	Jarjartuu
Ko'eessa	Ko'eeyyii	Ko'eessa	Ko'eettii

Table 2.3: Plural, Masculine and Feminine forms of sample Adjectives

2.3.3. Verbs (Xumura)

The verbs in natural languages are words or compound of words that expresses action, a state of being and or relationship between two things (Daniel and James, 2000). In its most powerful and normal position, it is found at the end of the sentence in Afaan Oromoo. Consider the following examples:

Eebbisaan kaleessa *dhufe***.** 'Eebisa *came* yesterday'.
Tolaan ni *dhukkubsate*. 'Tola was sick'.
Lammiin farda *bite*. 'Lami *bought* a horse'.

In the above examples, all italic words are verbs. They appear at the end of the sentences in Afaan Oromoo, while in English, most of the time, next to the subject of the sentence.

There are different categories of verbs depending on their needs and uses in sentences. Intransitive, transitive and auxiliaries are major subcategories of verbs in Afaan Oromoo (Diriba, 2002).

Intransitive verbs are verbs which do not take object or complement in a sentence. For example:

Konkolaatichi *cabe.* 'The car broke'.
Isaan kaleessa *dhufan.* 'They came yesterday'.
Aliin ni *fiiga.* 'Ali is running'.

In the above sentences, all italic words are intransitive verbs. They appear at the end of the given sentences and do not transfer message from subject to complement.

On the other hand, transitive verbs are verbs which transfer message to complements (objects). For instance:

Baacaan burcuqqoo *cabse.* 'Bacha broke a glass'.
Caalaan muka *mure.* 'Chala cut the tree'.
In the above two sentences, *cabse 'broke'* and *mure 'cut'* are transitive verbs since they interrelated subjects and objects in the sentences. The other main subtype of verb categories is auxiliary verbs. Auxiliary verbs are verbs that support the main verb used in a sentence. Some auxiliary verbs in Afaan Oromoo are indicated in the following examples:

Boontuun barattuu *dha.* 'Bontu is a student'.

Lalisaan Injiinera *ta'e.* 'Lelisa becomes an engineer'.

Inni mana barumsaa deemee *ture.* 'He was going to school'.

Hojii kana hojjachuu *qabda.* 'You have to do this work'.

All words in the above example sentences that are italic are auxiliary verbs in Afaan Oromoo language.

2.3.4. Adverbs (Ibsa Xumuraa)

Adverbs are words which are used to modify and illustrate verbs. In Afaan Oromoo, adverbs often precede verbs that they modify. In both English and Afaan Oromoo languages, adverbs could be categorized as adverbial times, adverbial places and adverbial conditions.

Adverbs of Time		**Adverbs of Manner**	
English	**Afaan Oromoo**	**English**	**Afaan Oromoo**
Today	*har'a*	Very	*baay'ee*
Tomorrow	*boor*	Quite	*baay'isee*
Now	*amma*	Really	*dhugumaan*
Then	*gaafas*	Fast	*dafee*
Later	*eger*	Well	*gaarii*
Tonight	*eda*	Hard	*cimaa*
Rightnow	amma, *isa ammaa*	quickly	*dafee*
Last night	*eda darbe*	slowly	*suuta*
This morning	*ganamakana*	carefully	*qalbiidhaan*
Next week	*torban dhufu*	absolutely	*matuma*
Recently	*dhiyeenyakana*	together	*walii wajjin*
Soon	*dhiyootti*	Alone	*qophaa*
…	…	…	…
Adverbs of Place		**Adverbs of Frequency**	
English	**Afaan Oromoo**	**English**	**Afaan Oromoo**
Out	*ala*	always	*yeroo hunda*
Here	*as*	sometimes	*gaafgaaf*
There	*achi*	occasionally	*gaafgaaf*

Over there	*gara sana*	seldom	*darbeedarbee*
Everywhere	*iddoo hunda*	Rarely	*darbeedarbee*
Nowhere	*eessayyu*	Never	*yoomiyyuu*
Home	*mana*	Away	*fagoo*
…	…	…	…

Table 2.4: Major types of Adverbs in both English and Afaan Oromoo

Adverbial times include words like amma *'now', yoom 'when', har'a 'today', kaleessa 'yesterday', bor 'tomorrow', iftaan 'after tomorrow'* and the like. The following examples show words of adverbial time in the language:

Dheeressaan *kaleessa* dhufe. 'Deresa came yesterday.'
Inni *har'a* deeme. 'He goes today.'

In the above sentences, *kaleessa 'yesterday'* and *har'a 'today'* are adverbial times in Afaan Oromoo.

Adverbial places responses the question 'where?' in a sentence. Words includes *achi 'there', gubbaa 'above', jidduu 'middle', bira 'together', gadi 'below'* and the like are some of adverbial places in Afaan Oromoo. For example:

Dheeressaan *mana* jira. 'Deresa is at home.'
Dheeressaan *diidaa* hojjata. 'Deresa is working outside'.
Isheen *iddoo* hoteelaa jirti. 'She is at hotel'.

Mana 'home', diida 'outside' and iddoo 'at' are adverbs of place in the above example sentences. Adverbial manner is another sub category of the adverb in a sentence that responses to the question 'how?' The following examples illustrate use of some adverbial manners in sentences:

Suuta deemi. 'Go slowly.'
Inni *qajeelcheetu* hojjeta. 'He works well'.
Isheen barumsa sheetti *baay'ee* cimtuudha. 'She is very strong in her education'.
Suuta 'slowly', qajeelchee 'well' baay'ee 'verb' and *cimtuu 'strong'* are adverbial manners in above example sentences.

2.3.5. Prepositions (Firoomsee)

In essential structures, preposition's position may precede or follow the category with which they form a syntactic unit. On the bases of this, they may be referred to by the less general terms, prepositions or postpositions. Some preposition and postposition in the language includes *akka 'as', ergasii 'since', hamma 'until', gara 'to', gadi 'below', oli 'above', irra 'on'* and so on. For instance, let's see the following sentences:

Tolaan *gara* hospitaalaa deeme. 'Tola goes to hospital'.
Eebbisaan buna mana isaa *duuba* dhaabe. 'Ebisa planted coffee at the back of his house'.
In the first sentence, *gara 'to'* is a preposition that precedes *hospitaala 'hospital'* and in the second sentence, *duuba 'back'* is also the postposition. It comes after *mana 'house'* with which it forms a syntactic unit.

English	Afaan Oromoo	English	Afaan Oromoo
about	waa'ee	Since	ergii
above	gubbaa/gararraa	Than	caalaa
across	gama	through	gidduu
after	booddee/booda	Till	hamma
against	faallaa	To	tti
among	jara gidduu	toward	garas
around	naannoo	under	jala/gajjallaa
As	akka	unlike	faallaa
At	itti	Until	hamma
before	dura	Up	gubbaa
behind	dudduuba/dugda	Via	karaa
below	jala/gajjallaa	With	wajjin
beneath	gajjallaa	within	keessatti
beside	bira	without	malee
between	gidduu	two words	Jechoota lama
beyond	garas	According to	akka kanaatti
But	garuu	Because of	kanaaf
By	dhaan	close to	bira
despite	ta'uyyu	due to	kanaaf

down	lafa	Except for	kana malee
during	utuu	Far from	irraa siqee/irraa
except	malee	inside of	keessa isaa
For	f	Instead of	kana mannaa
from	irraa	Near to	Itti aanee
In	keessa	Next to	Itti aanee
inside	keessa	outside of	kanaa alatti
…	…	…	…
Into	keessatti	priorto	kanaan dura
near	bira	three words	Jechoota sadii
next	ittaanee	as faras	hamma
Of	kan	aswellas	fi
On	irra	in addition to	dabalatees
opposite	fuullee/faallaa	in front of	Fuullee isaa
Out	ala	despite	haata'uyyu malee
outside	ala	on behalf of	maqaa
over	irraan	on top of	kana irraan
Per	tti	demonstrative	agarsiisoo
plus	fi	This	kana/tana
round	naannoo	That	sana
…	…	…	…

Table 2.5: Prepositions in both English and Afaan Oromoo

2.4. Comparatives and Superlative forms of Afaan Oromoo words

There is no direct translation of the comparative and superlative forms in Afaan Oromoo, unlike that of words in English language. In English, the suffix**–er**is mostly used to indicate comparative form of the words, while **–est** is used to indicate the superlative forms respectively. In Afaan Oromoo, there is no such suffix added at the end of the word to show whether the word is either in its comparative or superlative form. Rather, when distinguishing between two objects, for instance as in "**the bigger one**", the Afaan Oromoo

phrase would simply be "**the big one**", roughly translated to Afaan Oromoo, "**isa guddaa**" or "**the very big one**" "**isa baay'ee guddaa**". The prefix "**baay'ee"**, in addition to meaning "**very**", can also be used to express the sense of comparativeness and superlativeness of the word.

The suffixes **–er** and **–est** are not the only common for all English words in expressing comparative and superlative forms of words. In some cases, the prefixes **more** and **most** as in "**more beautiful**" and "**most beautiful**", meaning "**baay'ee bareedduu**", are also used when dealing with most adjective forms of words.

The adjective "**caalaa**" can be used to mean **"better"** or **"more",** though most often it functions as an adverb and comes immediately before the verb, as in "**Isheen caalaa bareeddi**" meaning **"She is more beautiful"**. **Caalaa** comes from the verb **caaluu** meaning **"to be better leading",** something which is comparatively better. Some dialects in Afaan Oromoo may also use **daran** instead of **caalaa** as a comparative adjective or adverb. Here, even in this specific topic, **daran** and **caalaa** are synonymous words.

The preposition **irra**, meaning "**on**", can signify a comparison in a way that more literally means "**relative to**". For example, "**Isheen isa irra furdoo dha**", meaning **"She is fatter than him"** or **"She is fat relative to him"**. In many cases, **caalaa** can be added to **irra** for optional emphasis, as in **"Finfinneen Maqalee irra** (**caalaa**) **bareeddi"** meaning **"Finfine is more beautiful than Mekele"**. Either in Afaan Oromoo or in English language, even in most of the languages over the world, the names of cities are treated as feminine.

For equating two things, as in **"as good as"** or **"as <any adjective> as"**, **"akkuma"** can be used in Afaan Oromoo. "**Adaamaan akkuma Dirree Dawaa ho'iti**", meaning **"Adama is as hot as Dire Dawa"**. **Akka** can also be used to mean **"like"** or **"similar to",** as in **"Boontuun akka Bilisee barattuu dha",** meaning **"Bontu is a student like Bilise"**. Additionally, **hanga** (**haga** in some dialects) means **"as much as",** as in **"Bilisaan hanga Boonaa ulfaata"** meaning **"Bilisa weighs as much as Bona".**

Examples:

Afaan Oromoo:	**English**
Adaamaan jireenyaaf Finfinnee caalti.	Adama is better for living than Finfine.
Eenyutu irra (caalaa) bareeda?	Who is more beautiful?
Kopheen kun sanarra mi'aa dha.	This shoe is more expensive than that one.
Inni nu caalaa sirritti dubbisa.	He reads better than us.
Isheen akkuma koo sirritti dubbifti.	She reads as fluent as me.

Note that **akka** and **akkuma** come between the nouns being compared. When two things being compared are both objects (e.g., **"Inni sana irra kana jaallata"** meaning **"He likes this more than that"**), **irra** comes after the first object. When one item is the subject and the other is an object (e.g. **"Kun sana irra gaarii dha"** meaning **"This is better than that"**), **irra** comes after the object (after the second item being compared). But, **caalaa** can come either in between or after the nouns.

Examples:

Afaan Oromoo	**English**
Konkolaataan kee koorra guddaa dha.	Your car is bigger than mine.
Foon caalaa fuduraa nyachuun jaalladha.	I like to eat vegetables more than meats.

Other descriptors like **older** and **younger**; **taller** and **shorter**; **thinner** and **fatter** are somewhat special cases. The word **Hangafuu in** Afaan Oromoo is a verb meaning **"to be older",** while **quxisuu** is an adjective meaning **"to be younger".** They are used as in the examples below:

Afaan Oromoo	**English**
Obboleettiin koo waggaa lama na hangafti.	My sister is two years older than me.
Obboleettiin koo waggaa lama quxusuu kooti. **or** My sister is younger than me by two years.	My sister is two years younger than me

To speak of things being the same, one may use **tokkuma,** meaning **"same"**; **gosa tokkicha** meaning **"same kind",** or **wal fakkaataa** meaning **"similar".** Something that is **different** is "**adda"**, and things that **are different** from each other are **"adda adda"**.

Examples:

English	**Afaan Oromoo**
These two things are the same.	Waantootni kunniin lamaan tokkuma.
These two things are similar.	Waantoota kunniin lamaan wal fakkaatu.
These two things are different.	Waantoota kunniin lamaan adda-adda.
This one is different.	Inni kun adda.

The adverbs **ol(i)** meaning **"up, above"** and **gad(i**) meaning "down, below" may be used to compare things as **"higher" or "lower",** as in:

English	**Afaan Oromoo**
He is shorter than 1.8 meters.	Inni meetira 1.8 (tokko tuqaa saddeet) gadi dha.
He is taller than 1.8 meters.	Inni meetira 1.8 oli dha.

2.5. Afaan Oromoo Morphology

Morphology indicates the structural property of a certain language in general. It mainly focuses on the grammatical categories of the language components, since they are the most productive and widely used categories of language features.

Like many other African and Ethiopian languages, Afaan Oromoo has a very rich morphology. It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromoo, Amharic and Zulu, most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Both Afaan Oromoo nouns and adjectives are highly inflected for number and gender. For instance, according to (Kula, 2008), in comparison to the English plural marker s (es), there are more than 12 major and very common plural markers in Afaan Oromoo nouns (e.g. -oota, -ooli, -wwan, -lee, -an, -een, -oo, etc.).

Afaan Oromoo verbs are highly inflected for gender, person, number and tenses (Abebe, 2013). Moreover, possessions, cases and article markers are often indicated through affixes in Afaan Oromoo. Since Afaan Oromoo is morphologically very productive and derivative, word formations in the language that involves several different linguistic features including affixation, reduplication and compounding is obvious in the language. Although Afaan Oromoo words have several prefixes and suffixes, lots of predominant morphological features exist in the language.

There are two productive ways to form words from morphological process (morphemes), namely: inflectional and derivational ways.

2.5.1. Inflectional Morphology

Inflectional Morphology deals with the combination of words with grammatical morpheme, usually resulting in a word of the same class as the original stem, and serving some syntactic function, for example plurals of nouns. They do not change the part-of speech category, but the grammatical function (also called morph-syntactic information) is changed. Different forms of words are produced by this morphology. In English, for instance, inflectional forms for the word **'work'** can be **'works', 'working',** and **'worked'**. These inflections are produced by adding the third person singular maker **/-s/,** the present continuous marker **/-ing/** and the past perfective marker **/-ed/** respectively. These all word forms of **'work',** i.e. **'work', 'works', 'working',** and **'worked'** are all verbs and there is no change in the part-of-speech category due to the affixation.

2.5.2. Derivational Morphology

Derivational Morphology creates new words (i.e., words with a different part-of-speech category) by adding a bound morpheme to a stem. Derivation can be applied recursively, i.e., words that are already the product of one derivation process can undergo the process again. The following isanexamplefromEnglish word: large(adjective)=enlarge (en-+large) (v)=enlargement (enlarge +-ment) (noun), and from Afaan Oromoo: bar- 'to know'(v), barumsa- 'education' (bar-+-umsa) (noun).

On the other hand, compounding is defined as the process of forming new words by combining different lexical categories (Alemayehu, 2007). However, it is not the case that every two words combined to form a compound form. Rather, every language follows

certainrules by which it forms its compound word formation. In Afaan Oromoo, the combination of abbaa ‘father’ + lafa ‘land’ forms new word abbaa lafaa, meaning ‘landlord’. The rules of compound word formation in Afaan Oromoo are unpredictable, and thus needs further linguistic study in the language. Hence, it is out of the scope of this research.

Based on structural changes of the stem and other morphemes during affixation, morphological processes can also be classified as linear or nonlinear. In linear morphology, affixes are added to the stem without changing the internal structure of the stem, though some changes might take place at the boundary of stems and affixes. On the other hand, morphological systems where the internal structure of the morphemes changes during the addition of suffixes are classified as nonlinear morphology. English pluralization for instance, pertains to linear category whereas Semitic languages features pertains nonlinearity. Morphological processes in Afaan Oromoo are mostly linear in nature.

2.6. Related Works

Among the top and developed languages across the world, English is one of the well-researched languages in natural language processing area. Synonym extraction systems for English language was developed by many researchers and now have been integrated to so many search engines and commercials and businesses softwares like Word processing applications, mailing services and office.org applications for both Unix and Windows operating systems.

To accomplish the objective of this study mentioned above, several articles, books and literatures were reviewed. Literatures concerning Afaan Oromoo and English or other Latin-based languages were also reviewed, since there are several approaches needed in Synonym extraction of Afaan Oromoo words and phrases. Review of related literatures were also carried out and used in different approaches for the study. Moreover, developing Synonym extraction for Afaan Oromoo words and phrases needs thorough understanding of the language. The principles and rules of the language around phonology, lexical, morphology and parts of speech have been studied roughly. These all are for the sake of studying Synonym extraction of Afaan Oromoo words and phrases in general and that of different forms of words in the language.

Few works were done previously on Afaan Oromoo language; especially in machine translation, natural language text processing and summarization, user interfaces, unilingual and multilingual information retrieval (ULIR and MLIR) and speech recognition systems. However, there is no research done on the area of synonyms of Afaan Oromoo words, the following literatures are some of the works considered recent and helpful in our further study on Synonym extraction:

2.6.1. Morphological Synthesizer

Analysis of Rule Based Approach for Afaan Oromoo Automatic Morphological Synthesizer (Abebe, 2013), that is designed to evaluate the scientific study of the structures and forms (morphological synthesizer) for Afaan Oromoo verbs and nouns by applying a rule-based approach. The rule-based approach is one of the earliest approaches in developing morphological synthesis systems and uses rules created by linguists or by a computer program to generate appropriate word forms from stems. The rules may contain large number of morphological, lexical and/or syntactical information.

The researchers also briefly evaluated various morphological approaches and properties of the language's words, specifically of verbs and nouns. To analyze the performance of the synthesizer, the researchers have conducted an experiment. Besides their experiment, data have been collected from different sources like thesis, reference books, and the internet. After collecting the data, it has been coded in the database in the form suitable for computational scheme. The stems collected have been stored accordingly with their tags and subtypes. The suffixes were also compiled according to the type of stem to which they apply.

In their experiment, the error counting approach was adopted to evaluate the word generation algorithm. The number of correctly generated words and incorrectly generated ones are counted for analysis. By preparing questionnaires, the outputs from the synthesizer for verbs and nouns were then checked against the respective valid words by domain experts. The errors were then described in terms of correctness and incorrectness of the produced words. The produced word is said to be correct if it is accepted by speakers of the language in terms of actual and possible words, otherwise incorrect. Throughout their thesis, the statistics used to measure the performance of the system is accuracy. Accuracy refers to the closeness of agreement between a test result and the accepted reference value. The

accuracy of the system is calculated as the number of correctly generated words divided by the total number of words generated by the system multiplied by 100%.

The analysis of the total number of words generated from single stem, especially of the verbal stem shows that Afaan Oromoo is morphologically complex language. It has been indicated that verbs are the most productive classes of words in the language. From the experiment, it is possible to say that the performance of the prototype is acceptable. The study has indicated the possibility of developing the synthesizer for Afaan Oromoo in rule-based approach for verbs and nouns. It is easy to extend the system for other parts of speech with minimum effort.

Finally, the performance of the synthesizer was evaluated using accuracy as statistical measurement with an average performance of 96.28% for verbs and 97.46% for nouns.

2.6.2. Morphology based Spell Checker

A spell checker, a tool that enables to check the spellings of the words in a text file and validates and checks whether the corresponding words are rightly or wrongly spelled (Gaddisa, 2014) has been studied. In case the spell checker has doubts about the spelling of the word, it suggests possible alternatives. The two core functionalities have been provided by the spell checker application, namely spelling error detection and spelling error correction. Error Detection is to verify the validity of a word in the language, while Error Correction is to suggest corrections for the misspelled words.

From the researchers' point of view, spell checker may be stand-alone capable of operating on a block of text, or as part of a larger application, such as a word processor, email client, electronic dictionary, or search engine. Several researches have been done for other languages like English, Arabic, Chinese and few researches have been done for Amharic language, but none for Afaan Oromoo. However, Afaan Oromoo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia, it still lacks several technological features compared to other most frequently used languages. Afaan Oromoo belongs to the lowland East Cushitic sub-family of the Afro-asiatic super-phylum. Among the Cushitic language families to which it belongs, Afaan Oromoo ranks first by the number of its speakers as stated also in the introductory content of this study. Currently, it is an official language of

Oromiyaa regional state. Despite of its popularity and its status as a regional language, Afaan Oromoo language processing is still in its infancy.

The researchers did a simple study to analyze spelling error pattern of Afaan Oromoo before implementation. For this purpose, the module prepared for teaching Afaan Oromoo courses was selected. The researchers used text analysis data gathering technique. The finding of study depicted the existence of spelling errors. When analyzed, it was found that 1,342 words were misspelled. Out of these 1,287 words were in the category of non-word errors. Though a comprehensive study is required to come to a clear opinion, it was enough to realize that non-word error detection is the first step towards a truly professional spellchecker. The paper is organized in to the following sections.

The test dataset was prepared to evaluate the number of valid words correctly accepted by the system and the number of invalid words correctly flagged by the system. Initially, the researchers randomly selected sentences (and paragraphs) from stories and papers belonging to several domains which produced 6521 words. To trim repeated words and select derivate, inflected and compounded words, Alchemist 2.0 has been used. This reduced the 6521 to 1464 unique words including compound words, words with derivational morphemes, inflected words, and functional words. The word lists are printed out and then manually spell checked by three postgraduate linguistic students.

The researchers found that, the data set consists of 1385 correctly spelled words and 79 misspelled words. In addition, they manually generated and added some inflection and derivation variants for some root words and the total of 347 unique words to the test data; making the total words of 1811. In this sample, out of 1,732 valid Afaan Oromoo words, 1,535 were accepted as a valid word and 197 words were flagged as misspelled words by the system. Out of 197 incorrectly flagged words, 186 words are due to absence of root word in the lexicon, whereas the remaining 11 words have root words in the lexicon, but the spelling checker did not recognize the inflection, derivation and compounding rules. On the other hand, all the 79 misspelled words were flagged.

2.6.3. Bilingual terminology extraction

The growing availability of comparable corpora, through the Interet or via distribution agencies providing newspapers articles in different languages, has led researchers to develop methods to extract bilingual lexicons from such corpora (Fatia Sadat), in order to

enrich existing bilingual dictionaries and thesauri, and help cross the language barrier for cross-language information retrieval. The paper presented several methods for exploiting multiple resources in bilingual lexicon extraction, either from parallel or comparable corpora. First, a special attention is given to the use of multilingual thesauri, and different search strategies based on such thesauri are investigated. Then, a method to optimally combine the different resources for bilingual lexicon extraction is presented. Their results showed that the combination of the resources significantly improves results, and that the use of hierarchical information contained in their thesaurus, UMLS/MeSH, is of primary importance. Lastly, methods for bilingual terminology extraction and thesaurus enrichment are discussed.

2.6.4. Supervised Word Sense Disambiguation using Phython

One of the fundamental tasks in natural language processing is Word Sense Disambiguation (WSD). WSD can be summarized as follows: given an ambiguous word, such as *bank*, determine which sense of the word (i.e. a financial institution, the side of a river, a type of basketball shot, etc.) is being used. There are a number of ways to approach this problem: one simple way is to determine which sense occurs most commonly (the Most Frequent Sense, or MFS), and always guess that sense. MFS is often accepted as a baseline for lexical sample tasks. There is not a well-defined upper bound on WSD performance: (Gale et al., 1992) argue that 95% should be regarded as an absolute upper bound because even human judges do not have 100% agreement on WSD tasks for a given language and sense inventory. The approaches to WSD discussed there used the *context*, or surrounding words and syntactic features, of the ambiguous word to try to determine the correct sense. In their paper, they discussed the problem of Word Sense Disambiguation (WSD) and one approach to solving the lexical sample problem, using training and test data from SENSEVAL-3 and implement methods based on the five different context-based classifiers, namely: Naive Bayes calculations, cosine comparison of word-frequency vectors, nearest neighbor cosine classifier, decision lists, and Latent Semantic Analysis. They also implemented a simple classifier combination system that combines these classifiers into one WSD module. They then proved the effectiveness of their WSD module by participating in the Multilingual Chinese-English Lexical Sample Task from SemEval-2007.

2.6.5. A metric to Assess the performance of MLIR Systems

The goal of IR is to retrieve documents that most directly relevant to the request of users (N. Moganarangan et al., 2014). With the speed of access and the large scale of the information sources available today, users often wish to reach beyond single information source in looking for relevant answers to their queries. MLIR systems have also to address the problem of documents content representation and the problem of relevance evaluation. This evaluation is more difficult than in monolingual IR. It is indeed difficult to build a correspondence function with different languages for the documents and the query. Traditional retrieval techniques create an illusion that it is presenting all relevant documents to the user but it is not so i.e. from the hit list, only one fourth of the documents are useful according to the user need. Thus relevance varies from one user to another, for example, for one user 20 documents may be relevant whereas for another user 40 documents may be relevant. Based on this binary relevance, all the documents in the resultant list are assessed and it is called continuous retrieval. An important consideration in evaluating MLIR systems is the need of the user and the knowledge of languages, since a multilingual system is most likely to be used in an interactive setting. It provides a means of measuring the quality of unranked information retrieval within a system. In this scheme, instead of simply employing the ranking produced by each SRE values, one can use the full potential of SRE values. This may be done by defining a measure that evaluates the utility or goodness of each SRE value. That is, F-measure uses the full potential of all the SRE values and it is the measure of performance that takes into accounts both recall and precision.

Information Retrieval plays a vital role in extraction of relevant information, that also leads to synonyms. Many researches have been working on for satisfying user needs, tough the problem arises when accessing multilingual information. This Multilingual environment provides a platform where a query can be formed in one language and the result can be in the same language and/or different languages. Performance evaluation of Information Retrieval for monolingual environments especially for English are developed and standardized from its inception. There is no specialized evaluation model available for evaluating the performance of services related to multilingual environments or systems. The unavailability of MLIR domain specific standards is a challenging task. The authors' paper presented and enhanced metric to assess the performance of MLIR systems over its counterpart IR metric. Their analysis also shows that the performance of the enhanced

metric is better than the conventional metric. And also these metric can facilitate the researchers and developers to improve the quality of the MLIR systems in the present and future scenarios.

2.6.6. Rule based Grammar checker

A rule-based Grammar checker that determines the syntactical correctness of a sentence, which is mostly used in word processors and compilers (Debela, 2011) has also been developed. However, the checker is evaluated and shown a promising result. For languages, such as Afaan Oromoo, advanced tools have been lacking and are still in the early stages. A rule-based grammar checker is entirely developed and presented depending on the morphology of the language.

The researchers used three methods that are widely used for grammar checking in a language, namely syntax-based checking, statistics-based checking and rule-based checking.

In syntax-based checking approach, a text is completely parsed, i.e. the sentences are analyzed and each sentence is assigned a tree structure. The text is considered incorrect if the parsing fails.

In the statistics-based approach, the system is trained on a corpus to learn what is correct. A POS annotated corpus is used to build a list of POS tag sequences in this approach. Some sequences will be very common (for example determiner, adjective, noun as in the old man), others will probably not occur at all (for example determiner, determiner, adjective). Sequences which occur often in the corpus can be considered correct, whereas uncommon sequences might be considered as errors. This method has a few disadvantages. One of these is that it can be difficult to understand the error given by the system as there is not a specific error message. This also makes it more difficult to realize when a false positive is given.

Using the rule-based approach to grammar checking involves manually constructing error detection rules for the language. These rules are then used to find errors in text that has already been analyzed, i.e. tagged with a part-of-speech tagger. These rules often contain suggestions on how to correct the error found in the text.

A lot of work has gone into developing grammar checkers for different languages. The most progress, by far, has been made for English. The earliest grammar checkers for English were

developed in the 1970s and have gradually been improving over the last decades. Although there is still room for improvement, their use is quite widespread as an English grammar checker is built into the most used word processor today, the Microsoft Office Word.

Easy English is a grammar checker developed at IBM, especially for non-native speakers. It is based on the English Slot Grammar. It finds errors by exploring the parse tree expressed as a network. The errors seem to be formalized as patterns that match the parse tree.

Chapter Three

Methodology

3. Methodology

In this chapter, different structures, approaches, methods, tools and techniques that help the fulfillment of the overall research study are stated. The initial methods followed to carry out the research study is data collection methods, that includes interviews, questionnaires, observations, document analysis and conversation with some language experts on Afaan Oromoo words and phrases, based on the study topic. To determine and analyze the data and information, that is being used in the language and makes it easier to come up with newly proposed study, these methods played vital roles. Hence, the following data sources and data collection methodologies were carried out to be used in the study.

3.1. Data Collection Methodologies

3.1.1. Questionnaire

We initially prepared the questionnaire to determine the acceptance of proposed study by the respondents. The respondents were needed to answer the questions provided mainly based on whether they will like the future study result that is going to be further researched or not. The questionnaire form of the study is provided for the respondents, aiming that the respondents can list down all the corresponding synonymy words for randomly selected Afaan Oromoo words (root words). Almost all the root words (words to which synonyms needed to be extracted) are taken from “Galmee Jechoota Afaan Oromoo: Wiirtuu Qo’annoo

fi Qorannoo Afaanota Itoophiyaa Yuunvarsiitii Finfinnee, Finfinnee" meaning "Afaan Oromoo's Words Store: Ethiopian Languages Research Center", by Tilahun Gemta.

These root words are selected purposefully, through considering some basic criterions such as dialects and PoS of words. Most of the respondent's educational level was bachelor of Degree and above. For instance, out of 100 respondents for 123 sample words **(**see **Appendix B**), 37 respondents were Afaan Oromoo department graduating class students at Adama Science and Technology University, those come from different corners of the Oromia region.

The overall data obtained and filled up from the questionnaire and its form is provided at the back end of this paper under the list of appendices.

3.1.2. Interview

Here in the interview, we interviewed sample Afaan Oromoo users, through considering the technical and professional people, officials, students and parents, by providing oral questions related to the topic area. The interview method was selected because it helped us to observe and verify an appropriate datum and users' interest.

This technique is used to enable the proponents to obtain additional knowledge and information that is somewhat more objective through conducting personal conversations. The interview data collection method is also used to list out currently existing problems, since it is directly related with what we discussed and observed.

3.1.3. Document analysis

We used some documents concerning the language such as books and dictionaries, different literatures, and some other related works done previously on the area of Afaan Oromoo words and phrases, that played a vital role for the progress of the study. For instance, "Galmee Jechoota Afaan Oromoo: Wiirtuu Qo'annoo fi Qorannoo Afaanota Itoophiyaa Yuunvarsiitii Finfinnee, Finfinnee" meaning "Afaan Oromoo's Words Store: Ethiopian Languages Research Center, Addis Ababa University, Ethiopia", by Tilahun Gemta is a typical book we used for the sake of data collection. During analysis of these documents, the proposed methodologies gave a special consideration to those documents which can bring more features to our further study.

3.2. Afaan Oromoo Alphabets and writing systems

To understand Afaan Oromoo language, starting at its basic data unit (Alphabet or Qubee in Afaan Oromoo) is mandatory. Afaan Oromoo language shares many features with English writing system except some differences. The Afaan Oromoo writing system is a modification of Latin Writing system. The writing system of the language is straightforward and designed based on the Latin script. All the letters in English language are also in Afaan Oromoo, but have different names, pronunciations and phonetic representations.

Afaan Oromoo has also "Qubee Dachaa", which is doubling of consonants from some other alphabets to represent different alphabet (Qubee) in the language. (Morka, 2001) described Afaan Oromoo writing system in more details. Learning Afaan Oromoo alphabet is very important because its structure is used in every day conversation. Without it, one will not be able to say words properly even if one knows how to write those words. The better you pronounce a letter in a word, the more you will understand speaking in the language. Specially, for someone who deals with text processing or NLP in general, understanding the characters is mandatory.

Afaan Oromoo alphabet has both vowels and consonants just like English language. Five of the letters *('a', 'e', 'i', 'o',* and *'u')* serve as vowels; whereas the rest of the alphabets including "Qubee Dachaa" are all consonants. In Afaan Oromoo language, the diacritic marker (Hudhaa) appears in between these vowels of words mostly, as these vowels are the one that have their own sounds and also give sound to the consonants in all syllable. In the subsequent sections, we will show how these Hudhaa is represented in words and the difficulties that it intrudes in differentiating from other quotation marks.

3.2.1. Diacritical Marker (Mallattoo Hudhaa)

The term diacritical marker is Hudhaa in Afaan Oromoo. Just like in many languages of the world, in Afaan Oromoo language also, there are places in words' syllable where it happens to have glottal sound. The main use of Hudhaa as diacritical marker in Afaan Oromoo is to change the sound value of letter to which they added. These show that the vowel with the dieresis mark is pronounced separately from the preceding vowel; the acute and grave accents, which can indicate that a final vowel is to be pronounced differently. Therefore, the definition of Hudhaa is the same as that of diacritical mark defined on English dictionary (Diriba, 2002), which is a mark placed over, under or through a letter in some languages, to

show that the letter should be pronounced differently from the same letter without the mark. In English language, a diacritical mark is also called diacritical point or diacritical sign. These marks are needed to accommodate sounds in a language with combination of those limited number of alphabet of the language. Hudhaa, serving as diacritical marker in Afaan Oromoo, is found in many words of the language. In computer technology, non-English languages seem to be unflavoured, regarding Hudhaa (diacritic) type of representation. Even though computer technology was developed mostly in the English-speaking countries, so data formats, keyboard layouts, etc. were developed with a bias favouring English (a language with an alphabet without diacritical marks), efforts have been made to handle them nowadays. Depending on the keyboard layout, which differs amongst countries, it is more or less easy to enter letters with diacritics on computers and type writers. Some have their own keys; some are created by first pressing the key with the diacritic mark followed by the letter to place it on. Such a key is sometimes referred to as a dead key, as it produces no output of its own but modifies the output of the key pressed after it. On computers, the availability of code pages determines whether one can use certain diacritics. Unicode solves this problem by assigning every known character its own code. By doing so, most modern computer systems provide a method to input it. With Unicode, it is also possible to combine diacritical marks with most characters. As stated earlier, Afaan Oromoo is one of the African languages that have many words that contain diacritical marks, Hudhaa. Of course, the variety of the marker is not too many in Afaan Oromoo, even though there are too many words. Moreover, a lot of unassimilated foreign loanwords are there in Afaan Oromoo.

3.2.2. Afaan Oromoo word formation

Word construction in Afaan Oromoo is straightforward and most convenient to represent any phoneme in the language and even to write foreign words. Words are formed from two or more sounds by combining vowels and consonants somehow like that of English. In Afaan Oromoo words, all the consonants have no their own sound unless a vowel is following them. But the vowels have their own sound as well as give sound to the consonants. In addition to this, there are many writing grammatical rules, just like in any other languages.

Besides the punctuations and quotations, there is another special character, usually apostrophe (') called “Hudhaa” for glottal sounding in some Afaan Oromoo words. This Hudhaa in the language appears between diphthongs (diphthong also known as a gliding

vowel, which refers to two adjacent vowel sounds occurring within the same syllable) of Afaan Oromoo word. Mostly, Hudhaa appears in between characters of words having consecutive vowels of different type (for example, *ta'e*, *ta'uu, bu'aa, ho'a, xaa'oo, mi'a, mi'aa, fe'uu, waa'ee*). Therefore, there is no instance in Afaan Oromoo words having consecutive vowels of different type without Hudhaa. Hudhaa sometimes appears between characters of words having consecutive vowels of same type also, depending on the phoneme of the syllable. Especially when the number of consecutive vowels is greater than two (for example, *re'ee, qe'ee, qu'uu*), though there are cases when it may rarely occur between just two of them (like in *sa'a).* On the other hand, *Hudhaa* may be inserted, between vowels and consonants, depending on the phoneme again (like in *har'a, hir'uu, mul'anne*, *etc*).

Here, our point is that the Hudhaa character selection found in many of Afaan Oromoo texts and the difficulty of differentiating it from other quotations. Some people just use apostrophe (') or use of the right-quote character (’) or left-quote character (‘) to indicate the glottal some wherein the middle of words or as component of words. This leads to serious confusion in text processing as the single quote character represents another thing and meaning in text. Look at the confusion of the apostrophe in the following text:

Yeroon nyaataa yommuu ga'e, 'Amma kottaa hundinuu qophaa'eeraa', jedhee warra waamaman haa yaadachiisuuf hojjetaa isaa itti erge.

As common practice, some people use symbols of vowels with hat from symbols dictionary (like *re'ee, bu'aa, qu'uu sa'aa, xaa'oo, etc).* One may find this kind of scriptures in Oromoo literature, which is, of course, not standardized way of writing the language. Because, the symbols are not alphabetic characters and intrude another complication in tokenization of Afaan Oromoo text processing activities in general. Still in other texts written by use of LATEX or any TEX typesetter, Hudhaa or glottal in Afaan Oromoo is represented by accenting of the vowels by control symbols.

3.2.3. Afaan Oromoo Punctuation Marks

Analysis of Afaan Oromoo texts reveals that different punctuation marks in the language follow the same punctuation pattern used in English, and other languages that follow Latin Writing System (Kula, 2008).The full stop (.) in statements, the question mark (?) in interrogatives and the exclamation mark (!) in command and exclamatory sentences marks

at the end of a sentence, and comma (,) which separates listing in a sentence and the semi colon (;) listing of ideas, concepts, names, items, etc. The other quotation marks also used just like in English language. The double quotations (“”) and single quotation (‘’) are used in similar rules as in English. We might as well use the correct dashes where we need them in writing Afaan Oromoo.

There are three types of dash like that of English language, namely the hyphen, the endash, and the emdash. Hyphens are typed as you would hope, just by typing a - at the point in the word, where you want a hyphen. Even though our typesetter or some conventional word processors takes care of hyphenation that is required to produce pretty line breaks, we type a hyphen when we explicitly want one to appear, as in a combination like “gar-tokko, irra-daddarbaa, qurxummi-qabduu, dhuga-baatuu”, and so on. An endash is the correct dash to use in indicating number ranges, as in Lk.1-3 meaning “No.1 to 3”.

3.2.4. Synonym Extraction Overviews

Terminologies that account for variations in languages are used by linking synonyms and definitions of words to their corresponding concepts and are important enablers of high quality information extraction (Tong Wang et al., 2009). Due to the use of specialized sub-languages in different domains, manual construction of semantic resources that accurately reflect language use is both costly and challenging, and often resulting in low coverage.

From the definition, Synonyms are words with similar or identical meanings. The synonymous relation is a typical lexical semantic relation, which is included in most lexical databases and ontology. These Synonymous relations are useful in several NLP and text mining applications, such as information retrieval, question answering, text summarization, language generations and recommendations, etc.

As stated earlier, the general goal of this study is to allow different users of Afaan Oromoo language to learn any unknown word that appears in a reading, rather than just a restricted set of words in a list of target vocabulary. Recent governments of the country promoted the use of local languages for official, administrative, judiciary and educational purposes. Because of such a language policy, regional states have chosen their respective official languages for various purposes. Today, in addition to Amharic, which is used as the working language of the federal government and the official working language, Afaan Oromoo is being used in the Oromiyaa regional state as a regional official language.

At the core of any NLP tasks, Synonym extraction is the important issue of natural language understanding. The process of building computer programs that understand natural language involves three major problems:

The first problem is related to thought processes; the second to the representation and meaning of the linguistic input; and the third one is to world knowledge. Thus, NLP system may begin at the word level to determine the morphological structure and nature (such as part-of-speech or meaning) of the word; and then may move on to the sentence level to determine the word order, grammar, and meaning of the entire sentence; and then to the context and the overall environment or domain. A given word or sentence may have a specific meaning or connotation in a given context or domain and may be related to many other words or sentences in the given context.

(Liddy, **1998**) suggested that to understand natural languages, it is important to be able to distinguish among the following seven interdependent levels that people use to extract meaning from text or spoken languages:

- Phonetic or phonological level, that deals with pronunciation,
- Morphological level, that deals with the smallest parts of words that carry out meanings, suffixes and prefixes,
- Lexical level, that deals with lexical meaning of words and parts of speech analyses,
- Syntactic level, that deals with grammar and structure of sentences,
- Semantic level, that deals with the meaning of words and sentences,
- Discourse level, that deals with the structure of different kinds of text using document structures, and
- Pragmatic level, that deals with the knowledge that comes from the outside world, i.e., from outside the content of the document.

A natural language processing system may involve all or some of these levels of analysis.

3.2.5. Properties of Dictionary Definitions

Some lexicographical properties of dictionary definitions as a basis for our synonym extraction methods has been established. We imitated to use some English language's lexicographical terminologies: the word being defined in a dictionary entry is called the definiendum, while the words that define it is called its definientia. This subsection

discusses special features of these elements in monolingual Afaan Oromoo dictionaries that facilitate synonym extraction. Within the definientia, there is usually a word or phrase that is more closely related to the definiendum than the rest of the definition: the genus term. Usually, genus terms are either synonyms or hypernyms of the definiendum, as in the example of automobile: a motorcar (synonym) and summer: the second and warmest season of the year (hypernym). Sometimes a genus term may include a quantifier. Such terms are known as empty heads; they prevent simple heuristics for identifying genus terms. These all holds true for Afaan Oromoo words and phrases.

Identifying genus terms or empty heads itself is a useful application in processing dictionary definitions. None the less the composition of definientia usually exhibits great regularity in terms of syntax, style, and sometimes, vocabulary. (Amsler,1980) showed that definitions of nouns and verbs in most dictionaries follow rigid stylistic patterns. In fact, this stylistic regularity goes beyond the definitions of nouns and verbs; definitions of adjectives exhibit comparable or even greater regularity than those of nouns and verbs.

In monolingual Afaan Oromoo dictionaries (like in monolingual English dictionaries), definition texts can often be decomposed into two parts: the interpretive part and the synonymous part. The former usually leads a definition in as a relatively lengthy description of the definiendum with simple vocabulary but complex syntax; many of these interpretive parts are followed by one or more synonymous parts, each consisting of a single word or phrase highly synonymous to the definiendum. When appearing together in the same definition, the two parts are usually separated by a special typographical format (e.g., capitalization) or delimiter (e.g., semicolons). Examples include the definition of look in (Mish, 2003) ‘to exercise the power of vision upon; examine’, and of looker-on: ‘one who looks on; a spectator’. Note that in the latter case, the synonymous part after the semicolon is, instead of being just one word, prefixed by an indefinite article. In terms of length and vocabulary, both examples conform to our observation on the differences between the interpretive and the synonymous parts. For the interpretive part of a given definition, real synonyms, if any, could be identified only through syntactic and semantic knowledge of the definientia; many attempts have been made to automate such deep analysis of the definition language. It is the simpler cases of the synonymous parts in definitions that are mostly left non-investigated. Semantically, as is shown in the previous examples, such parts of definitions are highly synonymous to the definiendum, whereas terms extracted from the

interpretive parts can often be hypernyms (even after successfully avoiding empty heads). Syntactically, synonymous parts can be identified by very simple typographical patterns. Although the patterns are dictionary-specific, their rigid nature usually necessitates minimal human intervention. The semantic relatedness between definiendum and definientia validates the hypothesis that synonymy does exist in dictionary definitions, while the regularities in the composition of definientia make it possible to develop algorithms for synonym extraction.

3.3. Different Approaches of Synonym Extraction

3.3.1. Distributional Approach

Unlike other LSRs, synonymous relations established more often by semantics than by syntax, and it is far from trivial to identify them from free texts. Hyponyms, for example, extracted accurately with the syntactic pattern '*A*, such as *B*' (Hearst, 1992).

Imagine if we plan to extend some heuristics for synonym extraction, for example, by taking *B* and *C* as synonyms according to pattern '*A*, such as *B* and *C*'. On a closer examination, the semantic closeness between *B* and *C* is determined by the semantic specificity of *A*, i.e., the more general *A* is in meaning, the more unlikely *B* and *C* are synonyms. For instance, for Afaan Oromoo word: "gaarii" we may have list of synonyms, namely: misha, dansa, bayeessa, etc. Alternatively, dansa and bayeessa can be synonyms for the word "misha", and vice versa.

Another intuitive approach to extracting synonyms from free texts is to find words sharing similar contexts, under the distributional hypothesis that 'similar words tend to have similar contexts' (Harris 1954). Such assumptions, however, are necessary but not sufficient for characterizing synonymy, since there is a fine but distinct line between being similar and being synonymous. In fact, words with similar contexts can represent many LSRs other than synonymy, even including antonymy (Mohammad, Dorr and Hirst, 2008).

In the work by (Lin, 1998), for example, the basic idea is that two words sharing more syntactic relations with respect to other words are more similar in meaning. Syntactic relations between word pairs were captured by the notion of dependency triples (e.g., (*w*1, *r*, *w*2), where *w*1 and *w*2 are two words and *r* is their syntactic relation). Semantic similarity

is well captured, but this is not equivalent to synonymy (it is shown in Lin's paper that antonyms are abundant in the list for the most similar word pairs).

To address the issue of false positives, Lin et al., (2003) devised two methods for identifying antonyms from the related words extracted using the algorithm by (Lin, 1998). The pattern-based method assumed X and Y to be antonyms if they appeared in patterns, such as from X to Y or either X or Y. The bilingual dictionary-based method was based on the observation that translations of the same word are usually synonyms. In comparing the classification between synonyms and antonyms to a randomly selected set of synonyms and antonyms from the Webster's Collegiate Thesaurus (Mish 2003), performance of the pattern-based method was generally good (F = 90.5%) but the dictionary-based method has very low recall (39.2%) due to the limited coverage of bilingual dictionaries used. In addition, the model can only identify antonyms among the various LSRs mixed in the extraction result.

Several later variants followed the work of (Lin, 1998). (Hagiwara, 2008), for example, also used the concept of dependency triples and extended it to syntactic paths in order to account for longer dependencies. Pointwise total correlation was used as the association ratio for building similarity measures, as opposed to the pointwise mutual information used by (Lin, 1998). (Wu and Zhou, 2003) used yet another measure of association ratio, i.e., weighted mutual information in the same distributional approach, claiming that weighted mutual information could correct the biased (lower) estimation of low-frequency word pairs in point wise mutual information.

Some studies also use the syntactic information in context indirectly in synonym extraction. (Curran, 2002), for example, compiled dependency relations into contextual vectors, which were used in a k-nearest neighbour model with ensembles to identify synonyms from raw texts.

Multilingual approaches can also be found in later studies (Barzilay and McKeown,2001; Shimohata and Sumita, 2002; Van der Plas and Tiedemann, 2006), hypothesizing that 'words that share translational contexts are semantically related'; the details of these approaches, however, differ in several important ways, such as the resource for computing translation probabilities and the number of languages involved. (Wuand Zhou, 2003), for example, proposed a semantic similarity model based on translation probability (which is calculated from a bilingual corpus) for synonym extraction.

Resulting synonym sets are compared to an existing thesaurus (EuroWordNet), asetting similar but not comparable to that of (Van der Plas and Tiedemann, 2006), since both the corpora and the gold standards are different in these two studies.

Another example of distributional approaches is that of (Freitag et al., 2005), where the notion of context is simply word tokens appearing within windows. Several probabilistic divergence scores were used to build similarity measures and the results were evaluated by solving simulated TOEFL synonym questions, the automatic generation of which itself is another contribution of the study.

3.3.2. Lexicon based Approach

Lexicons can be viewed as uniquely structured texts, which associate a word with other words that define it. Since lexicons are usually much smaller than text corpora, even of modest sizes, it is computationally less expensive to be used in graph-based methods. When compared to free texts, definition texts also exhibit stronger structural and syntactic regularity, which allows simple rule-based methods to achieve reasonably good performance.

Dictionaries are among the popular lexicons used for synonym extraction. In the early 1980s, extracting and processing information from machine-readable dictionary definitions was a topic of considerable interest, especially since the Longman's Dictionary of Contemporary English (LDOCE) (Procter, 1978) had become electronically available. Two special features were particularly helpful in promoting this dictionary's importance in many lexicon-based NLP studies. First, the dictionary uses a controlled vocabulary of only 2,178 words to define approximately 207,000 lexical entries. Although the lexicographers' original intention was to facilitate the use of the dictionary by learners of the language, this design later proved to be a valuable computational feature. Second, the subject code and box code label each lexical entry with additional semantic information, such as the domains of usage and selection preferences and restrictions.

It is debatable whether a learner's dictionary is indeed more suitable for the purpose of machine-based learning for NLP. A controlled vocabulary can also complicate the definition syntax, since there is usually a trade-off between the size of the defining vocabulary and the syntactic complexity of definitions (Barnbrook,2002). Nonetheless, with all the computationally friendly features, LDOCE soon attracted significant research interest.

(Boguraev and Briscoe, 1989) covered various topics in using this machine-readable dictionary, from rendering easier on-line access and browsing (which involved many engineering challenges under the computing environments of the time) to semantic analysis and utilization of the definition texts. The latter is of great relevance to the topics discussed in this paper.

(Alshawi, 1987) conducted a phrasal analysis of LDOCE definitions by applying a set of successively more specific phrasal patterns on the definition texts. The goal was to mine semantic information from definitions, which is believed to be helpful in 'learning' new words with the knowledge of the controlled vocabulary in LDOCE. (Guthrie et al.,1991) exploited both the controlled vocabulary and the subject code features. The controlled vocabulary was firstly grouped into 'neighbourhoods' according to their co-occurrence patterns; the subject codes were then imposed on the grouping, resulting in so-called subject dependent neighbourhoods. Such co-occurrence models were claimed to better resemble the polysemous nature of many English words, which, in turn, could help improve word sense disambiguation performance. Unfortunately, no evaluation has ever been published to support this claim.

The work of Chodorow, Byrd and Heidorn (1985) is an example of building a semantic hierarchy by identifying 'head words' (or genus terms; see Section 2.1) within definition texts. The basic idea is that genus terms are usually hypernyms of the words they define. If two words share the same head word in their definitions, they are likely to be synonymous siblings under the same parent in the lexical taxonomy. Thus, by grouping together words that share the same hypernyms, not only are synonyms extracted from the definition texts, but they are also, at the same time, organized into a semantic hierarchy.

Particularly, in recent years, one popular paradigm is to build a graph on a dictionary (a dictionary graph) according to the defining relationship between words. Vertices correspond to words, and edges point from the words being defined (definientia) to words defining them (definiens). This idea was first developed by (Reichert, Olney and Paris, 1969) and has ever since been extensively exploited by later studies as a basis for building dictionary graphs. Given such a dictionary graph, many results from graph theory can then be employed to explore synonym extraction. (Blondel and Senellart, 2002) applied an algorithm on a weighted graph similar to PageRank (Page et al., 1999); weights on the graph vertices would converge to numbers indicating the relatedness between two vertices (words), which are

subsequently used to define synonymy. (Muller, Hathout and Bruno, 2006) built a Markovian matrix on a dictionary graph to model random walks between vertices, which is capable of capturing the semantic relations between words that are not immediate neighbours in the graph. (Ho and C´edrick, 2004) employed concepts in information theory, computing similarity between words by their quantity of information exchanged through the graph.

Apart from synonym extraction, dictionary graphs have also been applied to other NLP-related tasks, such as word sense disambiguation. (Navigli, 2009), for example, proposed to disambiguate definiens in dictionary definitions by (quasi) circle-based scores computed from a dictionary graph. In a broader sense, the topic of a Wikipedia article can also be regarded as being 'defined' by the article content, motivating studies, such as (Gabrilovich and Markovitch, 2007), to develop semantic relatedness metrics based on 'dictionary graphs' using this type of defining relationship.

3.4. Development Tools and Algorithms

The prototype of Afaan Oromoo Synonym extraction is developed using the latest version of Microsoft Visual Studio (C#) and related platforms. C# language is used since it is a multi-paradigm programming language with strong typing, imperative, and declarative object-oriented programming discipline.

A database consisting of three important tables has been required as a knowledge base. The algorithms needed for the Synonym extraction have been designed from the scratch as there are no previously designed algorithms for Afaan Oromoo Synonym extraction. However, the morphological properties of the language are complex to generate words from an input stems (root words), some algorithms have been used and adapted from different languages particularly from those Latin-based languages.

The overall performances of the study finally evaluated and tested to verify its effectiveness. In the literature review section, some related works on the language that are done for evaluating different features are analyzed. Among these works, a work done by (Gaddisa, 2014) proposed spell checker tool that is the fundamental and initial requirement for the Synonym extraction. Hence, the spell checker algorithm, that identifies whether a word is valid or not, is also needed to be analyzed to come up with the suggestion adequacy for

evaluation of the study's result. Some of the measurements and parameters may be subjective and difficult to evaluate in this research work.

Input: Root Word

Output: List of Synonyms

Start

Get input Root Word

Check whether the Root Word is spelled correctly or not

Identify POS and Morphology of the inputted Root Word

Retrieve the Synonymy Words corresponding to the Root Word from wordlists

Makeup and Identify POS of the Root Word accordingly

Set an appropriate Synonym

End

3.5. Design and Implementation of AOSE

In this section, the developed Afaan Oromoo Synonym Extraction design requirements along with their subcomponents, and their interactions and applications will be presented.

3.5.1. Architecture of AOSE System

The architecture of Afaan Oromoo Synonym Extraction is represented in Figure 3.1. The architecture has nine components, namely: word tokenizer, stem presence checker, morphological analyzer, spelling checker, knowledge base, relevant synonyms generation, PoS identification, proving and ranking suggestions and word replacement. In this section, the system architecture for Afaan Oromoo Synonym Extraction is presented.

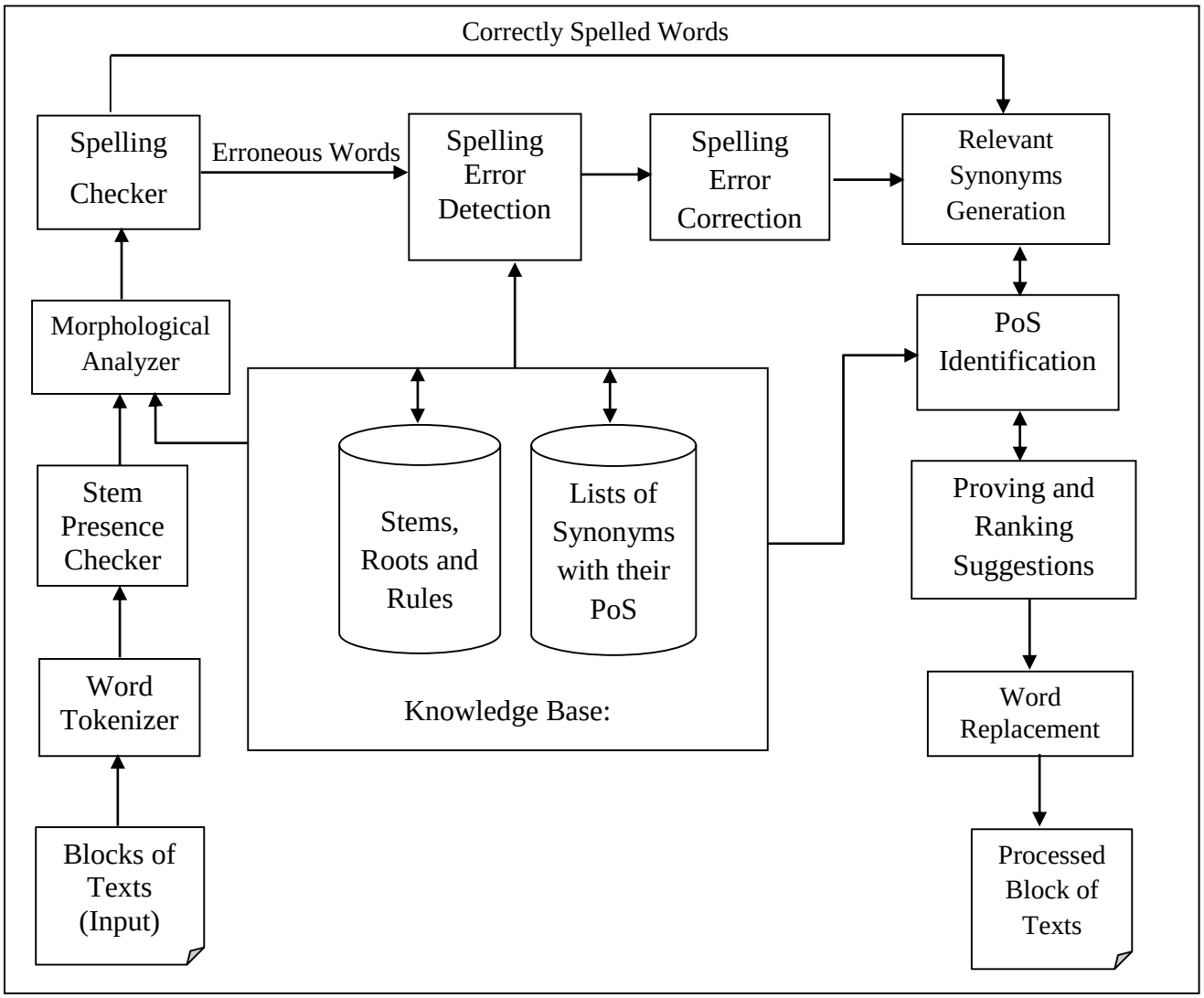

Figure 3.1: Architecture of Afaan Oromoo Synonym Extraction

Word tokenizer component:

This component split a block of text into individual words, digits and punctuation marks. In Afaan Oromo, like in other languages, the blank character (space) shows the end of one word. Moreover, parenthesis, brackets, quotes, etc are being used to show a word boundary.

Furthermore, sentence boundaries and punctuations are almost like English language (i.e. a sentence may end with a period (.), line break, a question mark (?), or an exclamation point). Thus, space marks are used as the explicit delimiters or token separator. Every time a space is encountered, the words after the space become a token. The output of this module (list of tokens) becomes input to spelling checker module.

Knowledge Base Module:

Testing data must be ready to acquire the necessary information in Synonym extraction process. The prototype is tested based on the information stored in lexicon to generate words. The Knowledge base is constructed by extracting necessary features and information from the stored data. Lexical information of the root words are the basic components to make the knowledge base. Lexical information is gained by considering a stem, given POS category and class. This enables us to distinguish between verbs and nouns, and different subtypes in each. The rules are also another essential element to the knowledge base.

Morphological Analyzer component

The task of this module is to accept a list of words from error detection component and then decompose each word into stem and affixes (based on their signature) and then pass them to the error detection component. Initially in the error detection component, the word is directly searched in the dictionary. If the word is found no further processing is needed. If it was not found in the dictionary, the word has to undergo morphological analysis to strip all the affixes to find all the possible morphemes of a word.

Error Detection component

Error detection component is responsible for whether the input word (either from the Tokenization component or Morphological analyzer component) is misspelled or not. Spelling error detection can be done through either n-gram analysis method or dictionary look up method. Dictionary lookup method has been used for our system.

The error detection component works first by looking the input word in the lexicon. When the input word exists in the lexicon, the spell checker does nothing. Otherwise, if this component returns ‘not exist’, the misspelled word will be sent to morphological analyzer component for further processing. Then, it checks whether the returned stem and affix exists

in the knowledge base. Consequently, the error detection component passes the erroneous word to the Error correction component.

The lookup process starts by affix stripping the word being searched for, and the word itself and the stem are looked up in the stem dictionary. If a match is found, a stop flag is attached to it. The word is also looked up in the original word list. If the word has a stop flag and appears in the original word list, it is assumed to be correct. Words that neither occur among nor are derivable from (by applying affix analysis) words in the word list are output as misspellings.

To illustrate how the error detection and morphological analyzer works, consider the unknown word ***bishaaniin*** 'by water'. The error detection module will first check if *bishaaniin* is found in the dictionary. Since it is not found in the dictionary, the system cannot automatically say that it is misspelled, for it may be inflected, derivated or compounded. Then the error detection component will call the morphological analyzer. To determine if this word is acceptable, affix stripping will be done. The affix stripping algorithm will first scan from the right to left and left to right to search for a valid affix. Since ***/-iin/*** is a valid suffix the morphological analyzer component strip /-iin/ from *bishaaniin* and returns suffix */-iin/* and unknown morpheme ***bishaan***. Now, the error detection component checks the unknown morpheme *bishaan* **'**water**'** for its presence in the stem dictionary. Since it is found in the stem dictionary, the system cannot automatically say *bishaaniin* is a valid word, for the stem *bishaan* may not be inflected or derivated for suffix */-iin/*. Finally, to determine if *bishaaniin* is an acceptable word, all the rules required to append suffix */-iin/* will be checked in the affix file (e.g. any nouns ending with consonant 'n' can append suffix */-iin/*). Now the error detection component will recognize *bishaaniin* as a valid word, and no further processing is needed. The same process will be done for a misspelled word.

Error Correction component

This module performs two tasks. First it accepts the pairs obtained after affix stripping in morphological analyzer from error detection module to classify the errors into one of the following classes:

- Valid prefix, invalid root, and invalid suffix
- Valid prefix, invalid root, and valid suffix
- Valid prefix, valid root, and invalid suffix

- Invalid prefix, valid root, and valid suffix
- Invalid prefix, valid root, and invalid suffix
- Invalid prefix, invalid root, and valid suffix
- Valid prefix, valid root, and valid suffix (incorrect combination)

After classifying, it will make a probable correction to the misspelled morphemes based on their error classes. A single erroneous word may be classified under two or more classes and each of the error is handled separately. For instance, in the case of a valid affixes (prefix and suffix) and invalid root, the valid affix will give us the specific category of the possible root word. Then using the replacement rule or Levenshtein edit distance (Levenshtein, 1966), the most likely word is searched in the dictionary. Similarly, on finding that the suffix is invalid, the word is scanned from the beginning i.e. from left to right in the dictionary. Upon finding a valid stem, the given word is stripped into a valid stem and a possible (or unknown) suffix. This process of stripping provides a list of pairs. For each pair, the suffix is searched in the affix file corresponding to the stem. If no suffix match is returned, then the pair is considered to be invalid and hence it is dropped.

However, if a suffix match is found then obtained suffix is combined with the stem to form a suggestion. Further in this category the stripped part is checked for a word match. Then the words obtained are combined with the suffix to form a suggestion. The same process is repeated for all possible correction. In addition, the word split error and run-on error (typing two words as one by removing space between them) is handled by this module.

Proving and Ranking Suggestions

Once the process of error correction and morphological generation was done, the next step is to provide and rank suggestions for the detected error. Hence upon detecting the error, the user will be provided with a list of probable correct words which the user can select to update the misspelled word. It may also be possible that the correct word expected by the user may not be listed in the suggestions hence the tool also provides an option for the user to type in the correct word and update it in the place of misspelled word. Here we used keyboard layout (or character distance), replacement rule in the affix file to rank the suggestion.

3.5.2. Design Requirements

A good Synonym Extraction must fetch all or most of appropriate synonyms and at the same time able to minimize wrong or incorrect synonyms. It must offer either the single right suggestion or a small set of suggestions for correction in which the right suggestion is included. The suggestions offered must be ranked and the right suggestion must occur in the top of the list in majority of cases.

The problems encountered in the design of programming structures that determine whether a word's synonyms are correctly fetched or not, will differ from language to language in some interests, as well with a more complex morphology. These languages often require additional morphological processing during the word validation phase, since increase in the size of the lexicon is not only time-consuming, but also error-prone. Specific and accurate analysis at word level is therefore a necessity (Aduriz et al., 2000). Having efficient and effective lexicon look up mechanisms is also one of the requirements in designing every Synonym Extraction.

There are several other issues that are closely related to Synonym Extraction and correcting processes: dictionary partitioning schemes, stem classification, coverage of the language of interest in terms of both lexical coverage and coverage of morphological and orthographic phenomena, availability for the research community and word boundary issues. Dictionary partitioning schemes were motivated by memory constraints in many early spell checkers. Word boundaries in most spelling error detection and correction techniques are often defined by inter-word separation, such as spaces and punctuation marks. However, it is largely dependent on the language under consideration. For instance, in languages like Chinese and Japanese, word boundaries are not defined. In Chinese, there is no space between words to act as word boundary delimiter (Sproat et al., 1996). But this is not an issue in languages like Amharic, English and Afaan Oromoo since word boundaries are defined. All the other issues are considered during the design of Afaan Oromoo Synonym Extraction.

3.5.3. Design Approaches and Techniques

Synonym Extraction can be done either manually or automatically. The manual aspect is done by domain experts or linguists. In this research, standalone Synonym Extraction that automatically fetches list of appropriate synonyms for the given word is considered. Both

word's error detection and correction can be done using dictionaries, morphologies and other linguistic tools using dictionary lookup techniques. In a dictionary-based approach, a word that is not found in the dictionary is either mistyped or invalid word and leads to the spelling error indication and other words from the dictionary which are close to the input word in terms of spelling are given out as suggestions for corrections. In a simple dictionary lookup approach, the probability of an alphabet sequence is the critical issue.

A morphology-based Synonym Extraction has more advantages such as its ability to reduce the dictionary size drastically and the ability to recognize new words that are not included in the dictionary. Morphological rules address word categories and their possible inflections, derivation and compounding. Further, the approach can be drawn upon in building grammar checkers.

As stated before, knowledge of the source language plays an important role to design Synonym Extraction. Such knowledge can be obtained in various ways. In this study, the knowledge base for morphological rule and lexicon design was obtained from "Galmee Jechoota Afaan Oromoo: Wiirtuu Qo'annoo fi Qorannoo Afaanota Itoophiyaa Yuunvarsiitii Finfinnee, Finfinnee" meaning "Afaan Oromoo's Words Store: Ethiopian Languages Research Center", by Tilahun Gemta, and discussions with linguistic professionals. However, an effort has been made to minimize the number of grammar rules of the language, the knowledge of the grammar plays central role in developing an efficient morphological analyzer (for Synonym Extraction) and morphological generation (for generating suggestions).

3.5.4. Lexicon Design

To build Synonym Extraction for Afaan Oromoo words, one would require a powerful dictionary for reference in the word error detection and suggestions prediction phases. Creating a dictionary having all the words of the language is a laborious task and infeasible, considering the large variation of affix combinations in Afaan Oromoo. To handle this problem, we used a morphological analyzer and a dictionary word stems. For example, for a single verb stem ***deem***- 'go', over 60 forms that are either adjectives or verbs (Gaddisa, 2014) can be generated (see **Appendix C** too).

In this section, stem classification and the design of lexicon needed to develop Synonym Extraction for Afaan Oromoo words will be discussed. The design process is based on the morphological properties of the language and assumptions of different approaches.

3.5.4.1. Stem classification

To define rules for permitted combinations of affixes and stems, it is necessary to observe the conditions under which affixes accept certain stems (or vice versa) while they don't produce a meaningful combination with others. If we succeed in understanding this system, it is possible to create a synonym which would not only recognize all words found in the source it was based on, but also valid compounds and combinations of stems and affixes which may be untypical, but still perfectly correct and usable in a context which the author of the dictionary simply did not think about. Implementing such a system of stem classification could result in a significantly shorter word list file.

The big challenge to classify a word stem is the grouping of nouns and verbs in such a way that the members of the same group have similar inflections and derivations. The inflections and derivations are not the same for all nouns and verbs. A traditional way to represent the morphology of inflectional languages is through classes of similar stems. The main features used for the classification is the stem ending and/or syllable structure.

Accordingly, the stems in the language are divided into classes. In our manual classification-based approach, the regularities in the way the stems take person makers for verbs, and plural makers and syllable structure for nouns are used to define classes by refining the observations in (Abebe, 2013). By this categorization, stems with the same morphological phenomena are grouped together into classes. Besides morphological features, syllable structure phenomena are also used in fine tuning of the classification. The stem endings in verbs of the same class have similar morphological behavior which enforces the stems toward affixing the same group of suffixes at the end slot.

For every stem, a signature is constructed which matches all word forms belonging to the class of this stem. When an arbitrary word needs to be processed, the stem checker looks up a matching stem in the lexicon. If such a stem has been found using the rules, inflectional type of word form is generated.

Classification of stems and its usefulness for implementing rules for Afaan Oromoo morphology has already been touched by Abebe in his morphological synthesizer (Abebe,

2013). He was found 17 different inflectional types which are divided into parts of speech, namely 7 for the nouns and 10 for the verbs. Every Afaan Oromoo inflecting stem can be classified as a member of some of these types. Two stems are in the same class if their paradigms are generated in the same way. The paradigm is described as a list of word forms with concrete grammatical features for each of them.

3.5.4.2. Building Signatures

After listing of stems and affixes, for each stem class there should be corresponding set of affixes with which it can combine to form variants of words. This association is called signature. Signatures are used to organize the stems and suffixes in such a way that the stems in the lexicon or new stems can be organized with appropriate suffixes. The stem classification discussions and examples in this subsection are based on the discussion and analysis from (Abebe, 2013).

3.5.4.2.1. Noun Signatures

According to (Abebe, 2013), the groupings of nouns are guided by the number of vowels after the last consonant, the last consonant itself and some specific endings. Hence, we have the following noun classes:

Class N1: Nouns ending in long vowels.
These nouns form plural and affix case makers without boundary change of the noun stem. But, the last vowels are deleted when definiteness suffixes are concatenated. If the last consonant of the noun is '*t*' and the sound preceding this letter is short, like in the case of ***barataa*** 'student', the last consonant doubles before common suffixes such as ***-oota***, ***-oolee***, and ***-oolii*** are added. Besides, these suffixes become ***-ota***, ***-olee***, ***-olii***, when the syllable preceding the last has long sound. For instance, nouns such as *aadaa*, *aarsaa*, *amantii*, *daandii*, and *haroo* are grouped under this category. The suffixes of this class include: *-wwan, -lee, -n, -tiin, -dhaa, -dhaan, -dhaaf,* and *–tii.*

Class N2: Nouns ending in short vowel '***a*'** in the last syllable and consonants like ***l***, ***r*** and ***m*** are of this category. The stem in this class may have two or more syllables, but the syllable preceding the last syllable must have long vowels. The group is characterized by the doubling of last consonant of the stem and deletion of last vowel when plural maker suffix is added. During case formation, they double last consonant and add case maker '**i**' in the nominative case. In addition, the short form of plural makers ***-ota***, ***-olee***, and ***-olii*** is suffixed

as common suffix instead of ***oota***, ***oolee***, ***olii***. For instance, ***wasiila*** 'uncle' and ***daa'ima*** 'child' are nouns of this class. The suffixes of this class include: *-an, -i*.

Class N3: Two syllable nouns each ending in short vowels and having end consonants like ***k***, ***g***, ***d***, ***r***, ***b*** and ***n*** are classified here. They take plural making suffix after doubling the last consonant and deleting the last vowel. This group shares similarity with class **N2** in that both double the last consonant. For instance, ***laga*** 'river', and ***muka*** 'tree' are nouns of this class. The suffixes of this class include: *-een, -ni, -a*

Class N4: Nouns ending in **"*eessa, eeysa, eecha, eettii, eeytii, essa, eensa*"** take the plural suffix by deleting these endings. This class deletes the last vowel before suffixing '**i**' in nominative case formation. Examples are ***jaldeessa*** 'monkey', ***hiyyeessa*** 'poor', ***dureettii*** 'rich woman', and ***bineensa*** 'wild animal'. The suffixes of this class include: *-eeyyii, -i.*

Class N5: This class comprises nouns that should be handled as special cases. Though most nouns end in vowels, there are a few nouns that end in consonants, particularly the letter '**n**'. They are characterized by having no plural suffixes. They only have case makers. The nouns such as ***aannan*** 'milk' and ***ilkaan*** 'teeth' are instances of this class. The associated case suffixes are: *-ii, -iin -iifi, -iis, iif, -iinis, -iifuu, -iifis, -uma, -umaa, -umaanuu, -umaaf, - umaanis, -umatti, -umattuu, -umattis, -umaratti, -umarattis, itti, ittis, ittuu, -irraa, -irraahuu, -irraahis, -irraan, -irraanuu, -irraanis, -irratti, -irrattis, -irrattuu, illee, -iinillee, - umallee, -umaafillee, -umaanillee, -ittillee, -umattillee, -irraahillee, -irrattillee, -irraanillee.*

Class N6: Nouns that end in the same or different two consonants and nouns having three or more short syllables are categorized under this group. For instance, ***dargaggeessa*** 'adult' is noun of this class. The main suffix of this class is *–oota*.

Class N7: This class includes proper nouns that have only case suffixes. This noun doesn't have plural makers. Either they should be stored in the lexicon or identified by capitalization of the first consonant provided that typing error is avoided. The corresponding nominative suffixes are: *-n, -tu, -dhaan, -dha, -rra, -dhaaf, -tumoo, -rratti*.

The suffixes that are not listed as the member of plural or case makers for all the above classes of nouns are common for all noun stems. They include ***-oota, - oolii, -oolee, -tti, -rraa, -ummaa, - ooma, -ittii, -icha***. We should also note that though common to all classes,

the principal usage of the suffix '**-oota**' is with stems having double consonant endings or nouns which have short sound in each syllable.

3.5.4.2.2. Verb Signatures

There is a link between stem endings of verbs and a set of suffixes. Observations confirm that stems belonging to the same category behave in the same way under morphological and phonological properties. According to [69] and [75], based on their stem finals Afaan Oromo verbs are classified into the following 10 classes with some adjustments made by linguists. The groupings are mostly done by person maker suffixes based on various aspects/tenses/.

Class V1: Stems ending in letters ***f***, ***k***, ***m***, ***n***

These stems are nasals in the process of sound formation. For instance, ***daf***, ***adeem*** and ***akeek*** are verb stems of this class. The associated suffixes for this class is *-na, -ne* and *-nu*.

Class V2: Stems ending in ***dh***, ***h***, *apostrophe* (**'**).
These are glottal and have two forms. The first form is those stems having short vowel before the ending '***dh'*** or '***h***'. In this form, the vowel before the ending is doubled and the ending itself is deleted prior to the affixation of subject makers in different tenses. The second form is when the vowel before the ending is long. In this case, no doubling of the vowel before the ending is necessary as three consecutive vowels are not allowed in the language. But the rest processes are similar. The suffixes corresponding to this class are: *-atini, -achise, -achisa, -achisan, -achiste, -achisna, -achistan, -aniiru, -anna, -anne, -annu, -ata, -atani, -ate, -atine, -atte, -attu, - attan, -atu,* etc.

Class V3: Stems ending in ***b***, ***d***, ***g***
These phonemes have similar phonological characteristics in that they all are voiced groups, and hence they behave similarly in suffixation process [69, 75]. The associated suffixes set for this class is *-da, -di, -dani, -de, -du, - eenya,* etc.

Class V4: Stems ending in **t**
This ending is called palatal, which is active participant in the process of assimilation. For instance, ***argat*** and ***baaft*** are verb stems of this class. The corresponding suffixes of this class are: *-nna, -nnu, -dha, -dhe, -dhu, -chisiise, -chisiisa, - chisisa, -chisiisan, -chisiste, -chisisna, -chisistan, -chiise, -chiisan, -chiisna, -chiistan, - chiisne, -iinsa, -iisa, -noo*

Class V5: Stems ending in ***s***, ***ch***, ***c***

This group is fricatives in the category of sound formation. The corresponding suffixes of this class are: *-ita, -iti, -itani, -ite, - itu, -ina, -inu, -ine, -ise, -isa, -isan, -iste, -isna, -ifte, -ifna, -isise, - isisa, -isisan, -isiste, -isisna, -isista, -isistan*

Class V6: Stems ending in ***l***

This ending is called approximants. In this case, the suffix '***n'*** hides itself to take the sound of the stem ending letter 'l' in the process called regressive assimilation. Hence, the corresponding inflections are *-la, -lu, -le, -lani*

Class 7: Stems ending in ***r***

This class has similar characteristics with the previous class, but they differ only in the form of person maker suffixes that fill the slot after a stem. The associated suffixes for this class are: *-ra, -ru, -re.*

Class V8: Stems ending in ***x***, ***q***, ***ph***

This class is similar with the third class and called ejectives. The only difference is that assimilation process gives the suffix '***-te***' the form of ***-xa, -xi, -xani, -xe, -xu*** which are considered as inflections of this class as indicators of persons in various aspects.

Class V9: Stems ending in "***aw***"

The associated inflectional suffixes are: *-oofte, -oofti, -oofna, -ooftan, -oofta*. The ending is replaced by these suffixes. For instance, ***qaanaw*** is a verb stems of this class.

Class V10: Stems ending in ***j***

The suffixes associated to this class are: *-ja, -ju, -je, -jani.* For instance, ***ajaj*** and ***jej*** are verb stems of this class.

The following suffixes are common suffixes for all classes of verbal stems: *-a,-achuu,-achuuf, -adha, -adhe, -adhu, -ama, -amaa, -aman, -amne, -amoo, -amta, -amtan, -amte, -amti, -amtuu, -amuuf, -ani, -e, -eera, -I,-uu, -uuf, -neerra, -aaf, -aas, -aat, -aatu, -uuttan, -uutti, -aa, -naan, -u, -eet, -eeti,- ees, -is, -uufan, -uufi, -uufii, -adhuu, -adhee, -amani, -amanii, -ame, -amni, -amu, -amtu, -amtani, -amuu, - amuudhaa, -amuudhaaf, -amuun, -ullee, -aatii, -umsa, -iisa, -iinsa.*

In addition to the suffixes compiled privately for every class, there is a significant overlap among some classes on the set of suffixes they take i.e. there are suffixes that are shared among two or more classes, but not among all classes. For instance, the suffixes *-ta, -tani, -taniittu, -te, - teetta, -teetti, -ti, -tu, -tuu* are shared among class V1, V2, V4, V6 and V7, and the suffixes *-sisa, -sisan, -sise, -siste, -sisna, -sistu, -sisne* are shared among the classes V1, V3, V6, V7 and V8. There are also suffixes that apply for all classes as mentioned before. The suffixes private to one class cannot be used by the other, while the common suffixes or the combination of non-private suffixes can be used by some or all the other classes. The surface forms of the verbal stem can be synthesized only after the generation of the signature corresponding to it.

Chapter Four

Results and Discussions

4. Results and Discussions

Due to the use of specialized sub-languages in different domains, such as medical, educational, administrational and communications, manual construction of semantic resources that accurately reflect the language use is both costly and challenging, and often resulting in low coverage. As stated earlier, since the number of word forms that can be generated from a single stem are so many, Afaan Oromoo is morphologically very rich and complex when compared to other Latin languages. In addition to this, a single Afaan Oromoo word may have multi Synonyms. Out of these multiple synonyms for a single Afaan Oromoo word, predicting and deciding the exact and appropriate synonym that fit the sense of the original root word may again need another expert system that recognizes different approaches of defining words in sentences and paragraphs. For instance, in common English vocabulary rules, there are two forms of word meanings, namely direct and contextual meanings.

4.1. Data Preprocessing

Most of the time, natural language processing is a data-intensive field. The success or failures of most NLP applications depend on the quality and availability of appropriate data. The data used in computational linguistic tasks generally takes the form of corpora. Corpora can be divided into two categories: annotated and unannotated corpora. Unannotated corpora are simply large collection of raw text, where as annotated corpora add additional information to the text, such as phonetic transcription, part-of-speech tags, parse trees, etc. (Abebe, 2013).

Annotated data in the form of lexicon is useful for synonym extraction. In this experiment, data have been collected from different sources like dictionaries, books, and the Internet. To evaluate the synonym extraction algorithm, a test data of 123 words from different sources have been collected. Out of total words collected, 72 are nouns, 18 are adjectives, 22 are verbs, 6 are adverbs and the remaining 5 are prepositions. The stems in the lexicon are not

exhaustive as there are infinitely many stems in a given language. But the number of suffixes is finite and listed as much as possible, though not exhaustively compiled. After collecting the data, it has been coded in the database in the form suitable for computational scheme. Normally, the collected stems have been stored with their tags and subtypes. The suffixes were also compiled according to the type of stem to which they apply. At the same time, the data collected have been checked by linguists for validity.

According to (Mandala et al., 1999), corpus is expected to have two important characteristics: sampling and representativeness, and machine readable. Stems are sampled and used being the representative of other stems as languages have large number of stems. The sample has represented all stems considering the morphological and structural variation of the stems. In this regard, sample stems have been taken from each class of stems equally in number and are believed to represent the rest of stems in the respective class for our synonym extraction. We have prepared lexicon for the system to use it as a source of stems and suffixes.

4.2. Prototype of the Study

The prototype of Afaan Oromoo synonym extraction is developed with visual studio 2012 (C#) update 3, with Microsoft Structured Query Language (SQL) 2014 as a back end. Thescreen snapshot of Afaan Oromoo Synonym Extraction (AOSE) is represented in the figure below:

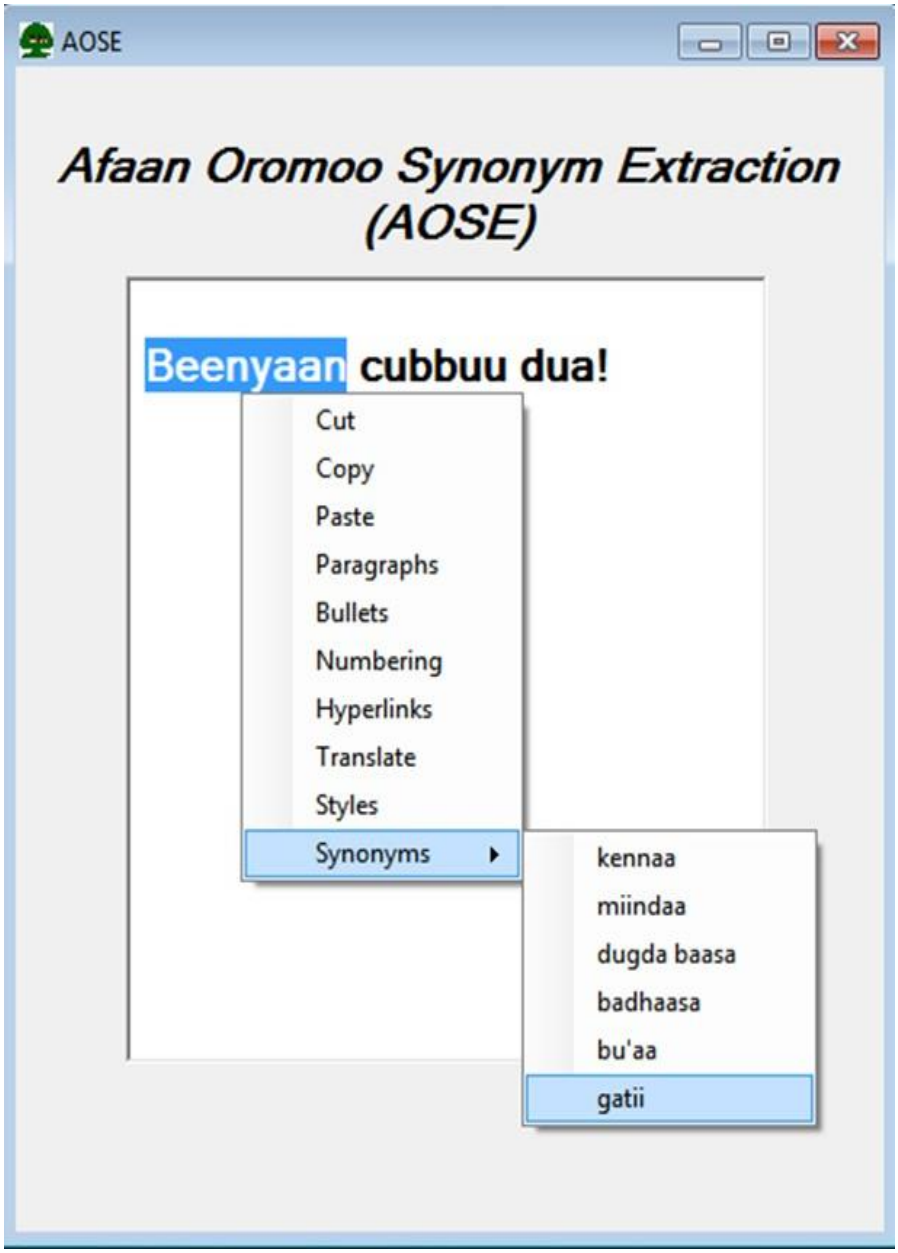

Figure 4.1: Prototype of Afaan Oromoo Synonym Extraction

The input for the prototype is text files or documents. The text files can be either copied or typed directly on the blank interface. The system expected to check the spelling automatically after the space bar is pressed or automatically after the copied text file was pasted (spell checker is not the part of our work, another researched investigated before). A word that the system believes to be misspelled is flagged with red zigzag line underlined and suggested corrections are displayed in a popup menu after right clicking on the erroneous word. Once after the spelling error of the word is verified and the given root word is identified as a valid Afaan Oromoo word, the system automatically fetches all appropriate

synonyms for the corresponding valid Afaan Oromoo words by again right clicking and selecting the Synonym menu.

As stated earlier, the dictionary lookup method is used in this work. List of 123 root words are needed to be stored in the initial table of the database, namely in the root table. Appropriate and proper synonyms for these root words are also needed to be provided in the synonym table following their level of similarities and relativeness to the roots. More synonymous words to the root words ranked first and less synonymous words to the root words ranked last respectively.

Basically, Afaan Oromoo words, like English words have two forms of meanings, namely direct (hiika kallattii) and contextual (hiika akka galumsaa) meanings. Direct meanings of words are entertained and determined independently on single standalone words, while contextual meanings are determined based on the perspectives of the word in the concerned sentence. Contextual meanings on the other hands, as their name indicates, are entertained contextually. In this work, we considered the direct meanings of words on separate sample root words randomly selected. In our synonym extraction, that used the direct meanings of words, the spelling of words has been a big issue. The word to which we need to provide a synonym may seem to be unavailable or invalid, unless spelled properly and accordingly. For instance, in Afaan Oromoo language, unlike in some other languages such as Amharic, which is currently the official and working language of our country, Ethiopia, words like gara, mala, kute, bira, fala, dhufe, balaa, karaa, etc., when spelled in incorrectly, gave another meaning instead of the expected meanings as follows:

Gara	Gaara	Garaa	Garraa	ጋራ
Mala	Maala	Malaa		ማላ
Bira	Biiraa	Birraa		ብራ/ቢራ
Fala	Faala	Falaa	Faallaa	ፋላ
Kute	Kutee	Kutte	Kuttee	ኩተ/ኩቴ
Dhufe	Dhuufe	Dhuufee		ዱፈ/ዱፌ
Balaa	Baala	Ballaa		ባላ
Karaa	Karra	Kaarraa		ካራ
Guba	Gubaa	Guuba	Gubbaa	ጉባ

4.3. Performance evaluation of the system

To evaluate the synonym extraction system, a test data of 123 words from different sources have been collected. Out of the total words collected, 72 are nouns, 18 are adjectives, 22 are verbs, 6 are adverbs and the remaining 5 are prepositions. Collected data has been coded in the database in the form suitable for computational scheme. Normally, the collected stems have been stored with their tags and subtypes. The suffixes were also compiled according to the type of stem to which they apply. At the same time, the data collected have been checked by linguists for validity.

The performance and effectiveness of synonym extraction is evaluated and tested manually. Although human layman knowledge can be used to measure the performance of the system, we considered six language experts and late them to rate how much they satisfied to the Afaan Oromoo synonym system. Three options provided for the experts, namely: Low (L), Medium (M) and High (H), and four of them rated as H, two of them rated as M and none of them rated as L. The average performance of the system is therefore 66.67%. However, the system is evaluated and shown a promising result, advanced improvements still lacking and we are in the early stages!

Chapter Five

Conclusions and Recommendations

5. Conclusions and Recommendations

5.1. Conclusions

In this work, we analyzed Synonym Extraction for Afaan Oromoo words called AOSE. We have studied some morphology and spell checker systems developed before for Afaan Oromoo and other Latin languages. Different approaches to develop a spell checker have also been studied, since it is the backbone of Synonym Extraction in identifying whether the word is valid Afaan Oromoo word or not. Therefore, AOSE is expected to be provided for valid words of the language only. As studied earlied under different sections of this thesis, the validity of the Afaan Oromoo word is checked through considering defferent syntactic and semantic rules. The nature, structure and pattern of Afaan Oromoo language has also been studied, since it is the crucial point in the development of Synonym Extraction for the language.

Synonym extraction needs deep understanding of the features and rules of languages under consideration. Hence, there are several holes for improvement and modification for Synonym extraction of Afaan Oromoo. Below are some of the recommendations we propose for future work.

5.2. Contributions of the work

Among the major contributions of this thesis work, some are:

- Proposing a new architecture for Afaan Oromoo synonym extraction system with the state-of-the art approach.
- Identifying the nature and pattern of Afaan Oromoo word formation and how they can be processed to be understood by a machine.
- Identifying and proposing the procedures, techniques, algorithms and tools used for the development of fully fledged AOSE.
- Preparing and encouraging environment for the development of other Afaan Oromoo NLP studies that need synonym extraction as a component in their work.

5.3. Recommendations

Synonymous words tend to have similar contexts. Predicting and optimizing words from their contexts, either from their sentential form or from their surrounding words needs special techniques through using applications of AI and expert systems.

The application of natural language processing techniques to Afaan Oromoo, as a language of education policy, official, administrative and judiciary is a promising step toward automatically producing vocabulary practices and assessments materials for the language learners.

Because of such a language policy, regional states have chosen their respective official languages for various purposes. Today, in addition to Amharic, which is used as the working language of the federal government and the official working language of the four regional states and the two federal cities, Addis Ababa and Dire Dawa, Afaan Oromoo in Oromia, Tigrinya in Tigray, Harari (Aderi) in Harari, Afar in Afar, and Somali in Somali serve as regional languages.

Synonym Extraction has become a part of everyday task of current generations in different fields and organizations. Whether it be working with text editing tools, such as Microsoft Soft Office Word, or typing text messages on one's cellular/mobile phone, after spell checking and correcting, confirmation of words Synonyms (if uncertain of direct word's meaning) is an inevitable part of the process.

Developing fully fledged and efficient Synonym Extraction needs a group of experts from linguistic and other coherent computational perspectives. Additional features can be added or the existing components of our Synonym Extraction can be modified to increase the performance to the better improvements. Synonym Extractions are vital components for any word processing applications and search engines. They are also helpful in detecting and correcting mismatchments of words meanings. The developed AOSE has portions that require further improvements that we want to recommend them as future works. Hence, the following recommendations are made for further researches and improvements:

- The AOSE developed involves many tasks starting from linguistic study to lexicon analysis for the language to work on user interface and prototype developments. As the ambition of this research is to develop a computational

framework for sample words through considering their PoS, the program does not aspire to account for all the grammatical categories in the language under discussion. Enhancing the system to include all the other parts of speech to make full fledged system is also an interest in future works.

- The morphological processes considered in this thesis are inflections and derivations. Further linguistic study should be considered to include simple and complex compoundings as well.

- The lexicon coverage of our knowledge base is small. One can increase the lexicon size and add more word formation (i.e. inflection, derivation, and compounding) rules to the knowledge base to obtain an improved performance. Afaan Oromo has six major dialects. Though it is difficult to handle all dialects. The more dialects considered, the more number of words generated from a single root word. Therefore, dialect consideration is crucial to have more word formation rules.

- Our synonym extraction mostly focuses on the noun PoS, but Afaan Oromoo documents also exhibit real-words in different PoS. Hence, there is a need of providing synonyms for real-words occuring in Afaan Oromoo documents.

- There are a lot of holes in the linguistic study of the language in general, especially in morphology and phonology. Linguists should give due consideration to intensively study the language structure and make it available for use in developing computational models.

- The prefixes and suffixes are almost exhaustive. There may be some prefixes, suffixes and rules missing during manual compilation, and they can be added to the knowledge base rules.

- Although we proposed a dictionary look-up with morphological rule-based techniques, interested researchers can use other approaches such as n-gram based approach or noisy channel equation techniques to develop a synonym extraction.

- The system developed in this research work is just a prototype. Any interested person can do a project to develop a full-fledged Afaan Oromo synonym

extraction that can be easily integrated into different Afaan Oromoo NLP works like:

- Grammar checkers
- Search engines
- Text to speech synthesis
- Machine translation
- Dictionaries and the like

- It will be very helpful to integrate this work with text processing application software like Microsoft Office due to its popularity as well.

Appendices

Appendix A: Questionnaire for key respondents:

List of synonymous words for the sample of 123 selected words from Afaan Oromoo monolingual dictionary to evaluate the validities of words for the analysis of synonym extraction for the Master's Thesis at Adama Science and Technology University.

Dear respondents,

Synonym extraction is a process of fetching one or more words or expressions of the same language that have the same or nearly the same meanings. The synonym extraction in this case will enable users to learn unknown Afaan Oromoo words that appear in a reading and writing list of target vocabulary.

Analysis of synonym extraction for Afaan Oromoo words is currently being explored at the Computing Department of Adama Science and Technology University as a Master's thesis. This research is expected to provide valuable inputs for Afaan Oromoo as a sub component of natural language processing systems in applications like machine translation, dictionary (lexicon) development, speech recognition, parts of speech tagging, automatic sentence construction, spelling and grammar checking, etc.

To this end, we kindly request you to provide us all the necessary and genuine information by responding to the questions about the validity of the words for synonym extraction for Afaan Oromoo words.

We would like to inform you that your responses will remain anonymous and will be used only for the above stated purpose.

Thanks in advance for your cooperation.

Kind regards,

The researchers

Section I: Respondent Identification:

Position/Job Title: _____

Region: _____ **Zone:** _____

Gender: ☐ Male ☐ Female

Age: ☐ <20 ☐ 20-30 ☐ 30-40 ☐ 40-50 ☐ 50-60 ☐ >60

Education Level: ☐ Diploma ☐ BSc/BA ☐ MSc/MA ☐ PhD ☐ Other

Section II: For the following 123 root words listed in the below table, give their corresponding POS and synonymous words as much as you can.

1. From the words listed in the following table, to which POS do you think you can list more synonymous words? (Rank them by giving 1, 2, 3, 4 or 5 in the box in front of the POS categories based on the order and priorities of providing synonyms)

   POS Categories:

   - ☐ Maqaa
   - ☐ Ibsa Maqaa
   - ☐ Xumura
   - ☐ Ibsa Xumuraa
   - ☐ Firoomsee

2. Do you think the POS for the root words and their synonyms will be different? (Yes or No) If your answer is "Yes", please indicate some sample root words in which their root's POS and their synonym's POS differs.
3. Do you think more POS categories exist or root words that cannot be classified to the above provided POS categories available? If yes, mention the appropriate POS you thought accordingly.
4. Provide the general comment (if any):

   _____

   _____

   _____

   _____

   _____.

SN	Root word	Gloss	POS	Syn 1	Syn 2	Syn 3	Syn 4	Syn 5	Syn 6
1	Araddaa	Farming land	Maqaa	Kaloo	Maasii	Ooyruu	Lafa Qonnaa		
2	Baatii	Moon	Maqaa	Ji'a	Bakkalcha				
3	Baay'ee	Many	Ibsa Xumuraa	Danuu	Hedduu				
4	Bakka	Place	Maqaa	Iddoo	Fuula				
5	Barbaaduu	Searching	Xumura	Soquu	Sakatta'uu				
6	Bareedaa	Handsome	Maqaa	Miidhagaa	Fuula Toleessa				
7	Barreessuu	Writing	Xumura	Katabuu	Qubeessuu				
8	Beekaa	Expert	Maqaa	Ogeessa	Hayyuu				
9	Beenyaa	Earning	Maqaa	Gatii	Dugda Baasa	Bu'aa	Miindaa	badhaasa	qabeenya
10	Bidiruu	Ship	Maqaa	Doonii	Oboloo				
11	Bohaartii	Recreation	Maqaa	Bashannana	Aara Galfii	Mijuu			
12	Buqushaa	Vagina	Maqaa	Shooshoo	Foshee	Muxxee	Fuchii	buqaa	
13	Caalaa	Better	Firoomsee	Wayyaa	Daran	Filatamaa			
14	Cinaa	Near/Behind	Firoomsee	Bira	Bukkee	Maddii			
15	Daawitii	Mirror	Maqaa	Ofilaalee	Fuullee				
16	Dagachuu	Forgetting	Ibsa Xumuraa	Irraanfachuu	Dedhuu				
17	Dargageessa	Young Man	Maqaa	Qeerroo	Saafela	Konfolleessa			
18	Deegaa	Poor	Maqaa	Iyyeessa	Dhabaa				
19	Deeggarsa	Support/Help	Maqaa	Gargaarsa	Hirpha	Tumsa			
20	Deemuu	Going	Ibsa Xumuraa	Imaluu	Sokkuu	Dhaquu			
21	Dhamaatee	Effort	Maqaa	Ifaaja	Dadhabbii				
22	Dhokachuu	Hidden	Xumura	Miliquu	Dheessuu	Baduu			
23	Dhukkuba	Disease	Maqaa	Dhibee	Michii	Miidhama			
24	Diida	Outside	Firoomsee	Bakkee	Ala	Halaala			
25	Diina	Enemy	Maqaa	Seexana	Nyaapha	Jinnii			
26	Dildila	Bridge	Maqaa	Riqicha	Ce'umsa				
27	Duwwaa	Empty	Ibsa Maqaa	Qullaa	Adiyyoo				

28	Eelaa		Maqaa	Ba'aa	Miidhama	Caba	Godaannisa		
29	Faaya	Beauty	Maqaa	Bareedina	Miidhagina				
30	Fiiguu	Running	Xumura	Kaachuu	Fulla'uu	Arreeduu			
31	Filaa	Comb	Maqaa	Meedoo	Mushuxii				
32	Fixuu	Finishing	Xumura	Xumuruu	Goolabuu				
33	Fokkisaa	Ugly	Ibsa Maqaa	Bushaa'aa	Jibbisiisaa				
34	Gaarii	Good	Ibsa Maqaa	Bayeessa	Dansa	Misha	Osee	tolaa	
35	Gaggeessaa	Leader	Maqaa	Hoogganaa	Dura Bu'aa	Bulchaa			
36	Galgala	Night	Maqaa	Waarii	Halkan				
37	Gamna		Ibsa Maqaa	Abshaala	Qaruuxee				
38	Gannee		Maqaa	Durba Duudaa	Bantii	Waa'elka			
39	Garraamii		Ibsa Maqaa	Arjaa	Gara Laafessa	Oo'a Qabeessa			
40	Goolaba	Finish	Ibsa Maqaa	Xumura	Dhuma				
41	Gowwaa		Ibsa Maqaa	Daallee	Doofaa				
42	Haadha	Mother	Maqaa	Harmee	Deessee				
43	Hallayyaa		Maqaa	Qilee	Hurufa				
44	Hanqaaquu	Egg	Maqaa	Killee	Buuphaa				
45	Hayyuu	Expert	Maqaa	Ogeessa	Beekaa	Ilaaltuu			
46	Hortee	Species	Maqaa	Gosa	Sanyii				
47	Ija	Eye	Maqaa	Agartuu	Qaroo				
48	Ilaaluu	Seeing	Ibsa Xumuraa	Mil'achuu	Arguu				
49	Iyyeessa	Poor	Maqaa	Dhabeessa	Harka Qalleessa	Deegaa			
50	Karaa	Road	Maqaa	Daandii	Deemsa	Imala			
51	Kolaa		Maqaa	Cinaan	Gumguma	Miciraan			
52	Koodee	Brother/Sister	Maqaa	Obboleessa	Obboleettii				
53	Liqii		Maqaa	Ceegoo	Gulbata				
54	Miilana		Ibsa Xumuraa	Dhiyoo	Alana	Booda			
55	Milkaa'uu	Success	Xumura	Galma Ga'uu	Injifachuu	Luba Ba'uu	Xumuruu		

56	Mulluu		Maqaa	Shummoo	Affeellii				
57	Mulquu		Maqaa	Baasuu	Gatuu	Dhisuu	Dhaanuu		
58	Muraasa	Few	Ibsa Xumuraa	Xiqqoo	Bicuu				
59	Niiraa	Cent	Maqaa	Taamunii	Dumbulloo				
60	Qaama	Body	Maqaa	Nafa	Dhaqna				
61	Qoosee	Shoulder	Maqaa	Saggoo	Qaceera				
62	Qorra	Cold	Maqaa	Dilalla	Qabbana				
63	Rakkoo	Problem	Maqaa	Dhibee	Miidhama	Quuqama			
64	Reebuu	Hit/Beat	Xumura	Rukkutuu	Dhaanuu				
65	Rifeensa	Hair	Maqaa	Dabbassaa	Qajjisa	Shurrubbaa			
66	Sakaale		Xumura	Xaxe	Gudunfe				
67	Sarbaa		Maqaa	Tafa	Gudeeda				
68	Seene	Enter	Xumura	Lixe	Gale				
69	Soba	Lie	Maqaa	Dhara	Kijiba				
70	Sokke	Gone	Xumura	Kute	Deeme	Qajeele			
71	Sooressa	Rich	Ibsa Maqaa	Dureessa	Badhaadhaa				
72	Tokkummaa	Unity	Maqaa	Marabbaa	Gamtaa				
73	Uffata	Cloth	Maqaa	Wayyaa	Huccuu	Kafana			
74	Uummata	People	Maqaa	Saba	Hawaasa				
75	Wayyaba	More	Maqaa	Caalmaa	Irra Jireessa				
76	Wiixata	Monday	Maqaa	Dafinoo	Hojja Duree				
77	Xiinxaluu	Thinking	Xumura	Qorachuu	Mirkaneessuu	Madaaluu	Yaaduu		
78	Xiqqaa	Small	Ibsa Maqaa	Bicuu	Xiqishuu				
79	Dhaggeeffate	Listen	Xumura	Caqase	Hordofe				
80	Daakuu	Swimming	Xumura	Bulleessuu	Dhukkeessuu				
81	Barii	Morning	Maqaa	Ganama	Subii	Barraaqa	Obboroo		
82	Bukkee		Maqaa	Middisa	Luuxii				
83	Barreessuu	Writing	Maqaa	Katabuu	Quruu				

84	Laphee		Ibsa Maqaa	Onnee	Qoma				
85	Qaroo		Xumura	Abshaala	Haxxee				
86	Nadheen		Maqaa	Dubarttii	Elellee	Dhalaa			
87	Shakkee		Xumura	Mame	Irra Galgale				
88	Xomboree		Maqaa	Gubbitii	Tissee				
89	Kiyyoo		Maqaa	Xuruuraa	Iskillaa	Shaaraa			
90	Baangaa		Maqaa	Sholee	Qankooraa	Mancaa			
91	Dhibaa'aa		Ibsa Maqaa	Yaraa	Seqqeeqa				
92	Dabeessa		Ibsa Maqaa	Daafanaa	Sodataa				
93	Kukkute		Xumura	Mummuree	Ciccire				
94	Sanyoo		Maqaa	Jaaltoo	Gaarayyuu	Toltuu			
95	Jaala		Maqaa	Hirriyyaa	Shirkaa	Michuu			
96	Dubara		Maqaa	Shamarra	Intala				
97	Boone		Xumura	Koore	Shoobe	Forree			
98	Rifeensa		Maqaa	Mataa	Dabbasaa				
99	Bokkaa		Maqaa	Rooba	Sora				
100	Ejje		Xumura	Dhaabbate	Sagaagale	Yaabe			
101	Imoo		Maqaa	Dhalee	Basuu				
102	Soquu		Maqaa	Barbaaduu	Sakatta'uu				
103	Abbeeraa		Maqaa	Wasiila	Eessuma	Abbaa Xqqaa			
104	Manguddoo		Maqaa	Jarsa	Buleessa				
105	Balleesse		Xumura	Barbadeesse	Dhabamsisee	Duuge	Dhidhimsise		
106	Gugee		Maqaa	Saphalisa	Makoodii	Bulaalla			
107	Jaallate		Xumura	Fedhe	Hawwe	Barbaade	Dharra'e		
108	Boosettii		Ibsa Maqaa	Of Gattuu	Kodheettii	Addayyuu			
109	Irra		Firoomsee	Gubbaa	Gararraa	Tabba Irra	Ol		
110	Hunda		Ibsa Maqaa	Mara	Danuu	Cufa	Duucha		
111	Badhaadhe		Xumura	Hore	Duroome	Soorome	Deebane		

112	Boqqolloo		Maqaa	Midhaan	Badala	Alquuqa			
113	Surree		Maqaa	Hidhaa	Kofoo	Lukee			
114	Godoo		Maqaa	Gad Ijjattii	Harii				
115	Fokkistuu		Ibsa Maqaa	Godeetti	Hamtuu				
116	Soreessa		Ibsa Maqaa	Duressa	Badhaadhee				
117	Balleessuu		Xumura	Haquu	Haxawuu				
118	Booda		Firoomsee	Gulana	Maayii				
119	Makoodii		Maqaa	Gugee	Bulaalla	Sabaliisa			
120	Qomoo		Maqaa	Sanyii	Gosa				
121	Dubartii		Maqaa	Nadheen	Elellee				
122	Ariifataa		Ibsa Maqaa	Jarjaraa	Qirqiraa	Daddafaa			
123	Albaadhessa		Ibsa Maqaa	Agarii	Hafuura				

Appendix B: Data Set for Synonym Extraction

1	Deema	16	deemte	31	deemneerra	46	hindeemta
2	deemaa	17	deemtee	32	deemneetu	47	hindeemtaa
3	Deeman	18	deemteef	33	deemteettaa	48	hindeemtan
4	Deemanii	19	deemti	34	hindeema	49	hindeemtan
5	Deeme	20	deemtii	35	hindeemaa	50	hindeemtee
6	Deeme	21	deemtiif	36	hindeeman	51	hindeemteef
7	deemna	22	deemtu	37	hindeemanii	52	hindeemti
8	deemnaa	23	deemtuu	38	hindeeme	53	hindeemti
9	deemnaaf	24	deemtuuf	39	hindeemee	54	hindeemtiif
10	deemneef	25	deemu	40	hindeemna	55	hindeemtu
11	Deemnu	26	deemuu	41	hindeemnaa	56	hideemtuu
12	deemnuu	27	deemuuf	42	hindeemnaaf	57	hindeemtuuf
13	Deemta	28	deemeera	43	hindeemneef	58	hindeemu
14	Deemtaa	29	deemeeraa	44	hindeemnu	59	hindeemuu
15	deemtan	30	deemeeraaf	45	hindeemnuu	60	hindeemuuf

Appendix C: Sample Verb Inflection for the stem: deem- ‘go’

No.	Capital	Small	Type	Long	Short
1	A	a	Vowel	Aa	A
2	B	b	Consonant	-	-
3	C	c	Consonant	-	-
4	D	d	Consonant	-	-
5	E	e	Vowel	Ee	E
6	F	f	Consonant	-	-
7	G	g	Consonant	-	-
8	H	h	Consonant	-	-
9	I	i	Vowel	Ii	I
10	J	j	Consonant	-	-
11	K	k	Consonant	-	-
12	L	l	Consonant	-	-
13	M	m	Consonant	-	-
14	N	n	Consonant	-	-
15	O	o	Vowel	Oo	O
16	P	p	Consonant	-	-
17	Q	q	Consonant	-	-
18	R	r	Consonant	-	-
19	S	s	Consonant	-	-
20	T	t	Consonant	-	-
21	U	u	Vowel	Uu	U
22	V	v	Consonant	-	-
23	W	w	Consonant	-	-
24	X	x	Consonant	-	-
25	Y	y	Consonant	-	-
26	Z	z	Consonant	-	-
27	CH	ch	Doubleconsonant	-	-
28	DH	dh	Doubleconsonant	-	-
29	NY	ny	Doubleconsonant	-	-
30	PH	ph	Doubleconsonant	-	-
31	SH	sh	Doubleconsonant	-	-

Appendix D: Afaan Oromoo Alphabets

Appendix E: Sample C# codes for Afaan Oromoo Synonym Extraction

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using WordEnties;
namespace DAL
{
  public partial class SynonymWordDAL
    {

        public bool AddSynonymWordDAL(SynonymWord_Entities sywEntities)
        {
            Synonym_DBEntities db = new Synonym_DBEntities();
            tblSynonymsWord tbl = new tblSynonymsWord();

            bool found = false;
            tbl.RID = sywEntities.RID;
            tbl.Synonym = sywEntities.Synonym;
            db.tblSynonymsWords.Add(tbl);
            int result = db.SaveChanges();

            if (result == 1)
            {
                found = true;
            }

            return found;
        }
        public List<SynonymWord_Entities> SearchRootWordDAL()
        {
            Synonym_DBEntities db = new Synonym_DBEntities();
            var result = (from rw in db.tblRootWords
                          select new RootWord_Entities
                          {
                              RID = rw.RID,
                              Word = rw.Word
                          }
                         ).ToList();
            return result.ToList<SynonymWord_Entities>();
        }
        public List<SynonymWord_Entities> SearchRootSelectedWordDAL(string
selectedText)
        {
            Synonym_DBEntities db = new Synonym_DBEntities();
            var result = (from rw in db.tblRootWords
                          join s in db.tblSynonyms on rw.RID equals s.RID

                          where rw.Word == selectedText

                          select new RootWord_Entities
                          {
                              RID = rw.RID,
                              Word = s.Synonym
                          }
                         ).ToList();
            return result.ToList<RootWord_Entities>();
        }
    }
}
```

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Data.Entity;
using System.Data.EntityClient;
using WordEnties;
namespace DAL
{
   public partial class RootWordDAL
    {
       public bool AddWordDAL(RootWord_Entities rootWord_Entities)
       {
             Synonym_DBEntities db = new Synonym_DBEntities();
           tblRootWord tbl = new tblRootWord();

           bool found = false;
           tbl.Word = rootWord_Entities.Word;
                db.tblRootWords.Add(tbl);
                int result = db.SaveChanges();

                if (result == 1)
                {
                    found= true;
                }

              return found;
       }
       public List<RootWord_Entities> SearchRootWordDAL()
       {
           Synonym_DBEntities db = new Synonym_DBEntities();
           var result = (from rw in db.tblRootWords
                         select new RootWord_Entities
                         {
                             RID=rw.RID,
                             Word=rw.Word
                         }
                       ).ToList();
           return result.ToList<RootWord_Entities>();

       }
       public List<RootWord_Entities> SearchRootSelectedWordDAL(string
selectedText)
       {
           Synonym_DBEntities db = new Synonym_DBEntities();
           var result = (from rw in db.tblRootWords
                         join s in db.tblSynonymsWords on rw.RID equals s.RID

                         where rw.Word==selectedText

                         select new RootWord_Entities
                         {
                             RID = rw.RID,
                             Word = s.Synonym
                         }
                       ).ToList();
           return result.ToList<RootWord_Entities>();
       }
    }
}
```

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using WordEnties;
using DAL;
namespace BLL
{
  public class SynonymWordBLL
    {
      public bool AddSynonymWordBLL(SynonymWord_Entities entities)
        {
            SynonymWordDAL da = new SynonymWordDAL();
            da.AddSynonymWordDAL(entities);
            return true;
        }
           }
}

using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Data.SqlClient;
using WordEnties;
using DAL;
namespace BLL
{
  public  class RootWordBL
    {
      public bool AddWordBLL(RootWord_Entities entities)
      {
          RootWordDAL da = new RootWordDAL();
           da.AddWordDAL(entities);
           return true;
      }
      public List<RootWord_Entities> SearchRootWordBLL()
      {
          RootWordDAL da = new RootWordDAL();
          return da.SearchRootWordDAL();
      }
      public List<RootWord_Entities> SearchRootSelectedWordBLL(string selectedText)
      {
          RootWordDAL da = new RootWordDAL();
          return da.SearchRootSelectedWordDAL(selectedText);
      }
    }
}

using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.IO;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;
using WordEnties;
```

```
using BLL;
namespace WordProject
{
    public partial class Form1 : Form
    {
        string str;
       ToolStripMenuItem toolstripmenu = new ToolStripMenuItem();
        public Form1()
        {
            InitializeComponent();
            textBox2.ContextMenuStrip = contextMenuStrip2;
            toolstripmenu.Text = "Synonym";
            ToolStripMenuItem toolstripmenuCut = new ToolStripMenuItem();
            toolstripmenuCut.Text = "Cut";
            toolstripmenuCut.Click += new EventHandler(CutAction);
            contextMenuStrip2.Items.Add(toolstripmenuCut);
            ToolStripMenuItem toolstripmenuCopy = new ToolStripMenuItem();
            toolstripmenuCopy.Text = "Copy";
            toolstripmenuCopy.Click += new EventHandler(CopyAction);
            contextMenuStrip2.Items.Add(toolstripmenuCopy);
            textBox2.ContextMenuStrip = contextMenuStrip2;
        }
        private void textBox2_TextChanged(object sender, EventArgs e)
        {
            RootWordBL rootWordBL = new RootWordBL();
            List<RootWord_Entities> listRootWord = rootWordBL.SearchRootWordBLL();
            List<RootWord_Entities> d = new List<RootWord_Entities>();
            foreach (RootWord_Entities r in listRootWord)
            {
                toolstripmenu.DropDownItems.Add(r.Word);
                d.Add(r);
                }
            contextMenuStrip2.Items.Add(toolstripmenu);
        }
        private void textBox2_MouseUp(object sender, MouseEventArgs e)
        {
             RootWordBL rootWordBL = new RootWordBL();
            List<RootWord_Entities> listRootWord =
rootWordBL.SearchRootSelectedWordBLL(textBox2.SelectedText);
            List<RootWord_Entities> d = new List<RootWord_Entities>();
            toolstripmenu.DropDownItems.Clear();
            foreach (RootWord_Entities r in listRootWord)
            {
               toolstripmenu.DropDownItems.Add(r.Word);
                d.Add(r);
            }
            contextMenuStrip2.Items.Add(toolstripmenu);
        }
        void CutAction(object sender, EventArgs e)
        {
            textBox2.Cut();
        }

        void CopyAction(object sender, EventArgs e)
        {
            Graphics objGraphics;
            Clipboard.SetData(DataFormats.Rtf, textBox2.SelectedText);
            Clipboard.Clear();
        }
        void PasteAction(object sender, EventArgs e)
        {
            if (Clipboard.ContainsText(TextDataFormat.Rtf))
```

```
            {
                textBox2.SelectedText =
Clipboard.GetData(DataFormats.Rtf).ToString();
            }
        }
        private void addToolStripMenuItem_Click(object sender, EventArgs e)
        {
            AddForm af = new AddForm();
            af.Show();
        }

        private void contextMenuStrip2_Opening(object sender, CancelEventArgs e)
        {

        }
        private void menuStrip1_ItemClicked(object sender,
ToolStripItemClickedEventArgs e)
        {
        }
    }
}
```

Appendix E: Sample C# codes for Afaan Oromoo Synonym Extraction

Declaration

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

Declared by:

Name: Lamessa Garba

Signature: _____

Date: 04/06/2018

Confirmed by advisor:

Name: Dr. Shimelis Aseffa (PhD)

Signature: _____

Date: _____

References

G.Q.A. Oromoo, Caasluuga Afaan Oromoo Jiildii I, Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia

Allen J., Natural Language Understanding, 2nd Edition. California: Redwood, Benjamin/Cummings Publishing Company, Inc, 1996

Steven Bird, Ewan Klein, and Edward Loper (June 2009). "Natural Language Processing with Python", First Edition

Gaddisa Olani Ganfure, Design and Implementation of Morphology Based Spell Checker, International Journal of Scientific and Technology Reasearch (IJSTR), 2(12), 118-125, 2014

Rajashekara Murthy, Vadiraj Madi, Sachin D, Ramakanth Kumar, A non-word Kannada Spell Checker Using Morphological Analyzer and Dictionary Lookup Method, IJESET, 2(2), 43-52, 2012

Tesfaye Tolessa, Early History of Written Oromo Language up to 1900, Star Journal, 1(2):76-80, 2012

Taye Girma, Human Language Technologies and Afan Oromo, May 2014, International Journal of Advanced Research in Engineering and Applied Sciences (IJAREAS)

Debela Tesfaye, A rule-based Afan Oromo Grammar Checker, 2011, International Journal of Advanced Computer Science and Applications (IJACSA)

Abebe Abeshu, Analysis of Rule Based Approach for Afan Oromo Automatic Morphological Synthesizer, October 2013, STAR Journal

Abraham Teso, Automatic Part-of-speech Tagging for Oromo Language Using Maximum Entropy Markov Model (MEMM), July 2014, Journal of Information and Computational Science

Girma Debele Dinegde, Afan Oromo News Text Summarizer, October 2014, International Journal of Computer Applications

Kula K. T., Vasudeva Varma and Prasad Pingali, Evaluation of Oromo-English Cross-Language Information Retrieval, IIT, Hyderabad (India), 2008

Diriba Megersa, Automatic Sentence Parser for Oromo Language Using Supervised Learning Technique, Addis Ababa University, 2002

Aduriz, E., and Agirre, I., A Word-Grammar based morphological analyzer for agglutinative languages, In the proceeding of the 18th International Conference on Computational Linguistics (COLING), Saarbrucken, Germany. 2000.

R. Sproat, W. Gale, C. Shih, and N. Chang, A stochastic finite-state word-segmentation algorithm for Chinese, Computational Linguistics, 22(3):377-404, 1996.

Steven Abney, Semi supervised Learning for Computational Linguistics, Chapman and Hall, 2008

Bancaa, A. (Ed.), (2016, February 7). Oromo language. In Wikipedia, the free encyclopedia. Retrieved from https://en.wikipedia.org/

Tilahun Gamta, (1992), Afaan Oromo: Retrieved on February 10, 2016, from http://www.africa.upenn.edu/Hornet/Afaan_Oromo_19777.html

Ferguson, Charles, (1976), The Ethiopian Language Area: Language in Ethiopia, ed. by M. Lionel Bender, J. Donald Bowen, R.L. Cooper, Charles A. Ferguson, pp. 63–76. Oxford: Oxford University Press

Mish, F. ed. 2003. Merriam-Webster's Collegiate Dictionary, 11th ed. Springfield, MA: Merriam-Webster

Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Technical report, Soviet Physics Doklady, 1966

Mohammad, S., and Hirst, G. 2006. Distributional measures of concept-distance: a task oriented evaluation in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 35–43, Sydney, Australia

Bikel, D., and Castelli, V. 2008. Event matching using the transitive closure of dependency relations in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 145–148, Columbus, Ohio, USA.

Mandala, R., Tokunaga, T., and Tanaka, H., Combining multiple evidence from different types of thesaurus for query expansion in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 191–197, Berkeley, California, USA.

McCarthy, D., and Navigli, R. 2009. The English lexical substitution task, Language Resources and Evaluation 43(2): 139–159.

Melbaa Gadaa, Oromiyaa: An Introduction to the History of the Oromo People, Khartoum, Sudan, 1988

Getachew M. Wagari, Million Meshesha, Parts of speech tagging for Afaan Oromo, International Journal of Advanced Computer Science and Application, Special Issue on Artificial Intelligence, www.ijacsa.thesai.org, Available online, 2011

Noah Webster, Webster's new twentieth century dictionary of English language. 2nd edition, New York, USA, 1979.

Daniel Jurafsky and James H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice-Hall, Inc., 2000

Alemayehu Dumessa, Word Formation in Diddessa Mao, 2007

Tong Wang and Graeme Hirst, Extracting Synonyms from Dictionary Definitions, International Conference RANLP 2009 - Borovets, Bulgaria, pages 471–477

Wakshum Mekonnen, Development of Stemming Algorithm for Oromo Texts, Master's Thesis, Addis Ababa University, 2000

Morka Mekonnen, Text to Speech System for Afaan Oromo, Master's Thesis, Addis Ababa University, 2001

Fatia Sadat, Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora, Nara Institute of Science and Technology, Nara, Japan

Phil Katz, Supervised Word Sense Disambiguation using Python, Swarthmore College Swarthmore, PA

N. Moganarangan, Pothula Sujatha, Shailesh Khapre, P. Dhavachelvan, A Metric to Assess the Performance of MLIR Systems: Web Service based Assessment Approach, International Journal of Engineering and Technology (IJET), ISSN: 0975-4024, Vol 6 No 1 Feb-Mar, 2014