# ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

## SCHOOL OF GRADUATE STUDIES

## DEPARTMENT OF COMPUTING

# AFAAN OROMO-ARABIC CROSS LANGUAGE INFORMATION RETRIEVAL

**HASSEN HUSSEN KASSIM**

September, 2017

ADAMA ETHIOPIA

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

# AFAAN OROMO ARABIC CROSS LANGUAGE INFORMATION RETRIEVALE

BY

HASSEN HUSSEN KASSIM

The Thesis Submitted to School of Graduate Studies of Adama Science and Technology University for Partial Fulfillment of the Requirements for the Degree of Master of Science in Information System

September, 2017

Adama Ethiopia

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

AFAAN OROMO ARABIC CROSS LANGUAGE INFORMATION RETRIEVAL

BY

HASSEN HUSSEN KASSIM

SIGNATURE PAGE (Approval sheet)

| | | |
|---|---|---|
| Advisor | Signature | Date |
| Chairperson | Signature | Date |
| External Examiner | Signature | Date |
| Internal Examiner | Signature | Date |
| Head of the Department | Signature | Date |
| School Dean | Signature | Date |

# Dedicated To:

# My baby Aphrahim

# (Fikir) and my family

ACKNOWLEDGMENTS

Allahamdulilah, for everything Allah has done for me and Thumal Allahamdulilah!

This thesis would not have become real without my advisor Wondwossen Mulugeta (PhD), who never gave up on helping me, from the beginning to the end of the study.

I would like to express my heartfelt thanks to Hussein Seid (PhD), Shimiles Assafa (PhD) and Kula Kekeba (PhD) for their comments and constructive ideas.

I would like to convey my gratitude to my beloved family for their encouragement and prayer.

Last but not least, to my friends, thanks for the entire moral support especially شخ/صاحبوالدين أدم, شخ/عبد الرحمان, Anwar Kelil, Wabal. A, Maaruf.K, Sultan.E, Mohammed.T, Koricho. A, Ebrahim. J, Tamirat.D and Yshaq.Y for that they have made some important resources available and especially my caring wife Zulfa Jemal and my sister Kerya Hussen for their endless devotion, enthusiasm, and patience. Without their unconditional supports and love. I could not have completed this study.

# Table of Contents

## List of Table



## List of Figures



## List of Algorithm

# ACRONYMS

| | |
|---|---|
| ACL | Association for Computational Linguistics |
| CLEF | Cross Lingual Evaluation Forum |
| CLIA | Cross Lingual Information Access |
| CLIR | Cross-lingual Information Retrieval |
| EM | Every-Match |
| FM | First-Match |
| GAP | Geometric Average Precision |
| GCLEF | Geographic Cross Lingual Evaluation Forum |
| IDF | Inverted Document Frequency |
| IR | Information Retrieval |
| IRS | Information Retrieval system's |
| MAP | Mean Average Precision |
| MRD | Machine-Readable Dictionary |
| MT | Machine Translation |
| NIST | National Institute of Science |
| OCR | Optical Character Recognition |
| PDF | Portable document format |
| POS | Parts-of-speech |
| SWB | Single-Word Based |
| TF | Term Frequency |
| TP | Two-Phase |
| TREC | Text Retrieval Conference |
| VSM | Vector Space Model |
| WBW | Word-by-word |

## Abstract

The objective of Cross Language Information Retrieval (CLIR) system is to put forward users with access to information that is in different language from their queries. It has the ability to issue a query in one language and retrieve documents in another language. In multi-lingual country like Ethiopia it is common to see language barriers while seeking information in language other than ones mother tongue. Afaan Oromo is one of the major languages that are widely spoken and used in Ethiopia. In spite of the fact that, Afaan Oromo has a large number of speakers little effort has been put in conducting research which aim at making Arabic documents available to Afaan Oromo speakers. Afaan Oromo speakers want to speak Arabic because of many fundamental facts. Among them are the diplomatic relation and the issue of public diplomacy is playing crucial role in the lives nations. This study is endeavor to develop Afaan Oromo-Arabic CLIR system which enables Afaan Oromo native speakers to access and retrieve the huge online information source that are available in Arabic by writing queries using their own language.

The approach selected and followed in this study is dictionary-based CLIR system. For this study a lot of documents like Qur'aana chapters, Bible chapters and other religious documents are included. The system is tested with 20 queries and 40 documents. Experiments were conducted by using one possible translation to Afaan Oromo query terms into Arabic. The retrieval effectiveness of the system is measured using recall and precision for both monolingual and bilingual runs.

The experiment returned a Maximum Average Precision of 0.706 and 0.675 for monolingual Afaan Oromo queries and also 0.617 and 0.7 bilingual translated Arabic queries runs respectively. From these outcomes, it can be concluded that, the performance of their trivial system could be increased with use of larger and clean corpora and by using different methods of like automatic alignment of words for translation.

***Keywords*:** Afaan Oromo, Arabic, Cross Language, Language, Information Retrieval

# Chapter one

## Introduction

This chapter gives general information about the thesis. It gives the general background of the study, the statement of the problem that motivated the research and also presents the methods employed to come up with the solution(s) of the problems. It also highlights the objectives, scope, limitations and significance of the study.

## 1.1 Background

Information Retrieval (IR) system solves the difficulty of identifying relevant documents from large amount of document collections. In the early of the 1990s, a fast growth of the Internet and the increasing multilingual contents of the web created a new challenge for language technology. The main challenge is the absence of a well-defined data model for the web, which implies that information definition and structure is frequently of low quality (Baeza R and Ribeiro, 1999). According to (Aljlayl, et al, 2002), with this fast growth of the Internet, the World Wide Web (WWW) is one of the most popular mediums for the dissemination of contents. As the result of the fast expansion of the Internet for communication and dissemination, online information resources are available in almost all major languages (Kula, 2008).

Web pages can be found in every popular non-English language including various European, Asian, Middle East and African languages (Aljlayl, et al, 2002). According to (Talvensaari T. , 2008), due to the existence of extremely large collection of documents, the ability of IR systems to retrieve relevant documents became critical. As extra documents written in numerous languages became available on the Internet, authors increasingly desire to explore documents that were written in either their native language or some other languages The increasing need for exploring more documents and the fact that authors can use any language to write their document, thus, became the main motivation for Cross-language Information Retrieval (CLIR) system.

CLIR systems provide users to retrieve documents written in one language by using a query written in another language (Ramanathan, 2003). Obviously, translation is needed in the CLIR process; either translating the documents into the query language (document

translation), or translating the query into the document languages (query translation). CLIR systems can help people who are able to read in a foreign language but are not capable enough to write a query in that language. It can also assist people who are not able to read in a foreign language but have access to translations resources. This situation increases the significance of CLIR systems which can make relevant document(s) from enormous collection accessible to the users. Its goal is to find the information a user needs even if it is written in a different language. This is achieved by designing a system where a query in one language can be compared with documents in another language. CLIR system make easy retrieval of relevant documents that written in one natural language with computerized the system that take queries articulated in other language(s). In monolingual IR, queries and documents are represented in the same language. It might occur, however, that a user is not talented to express his or her query in the document language, even if she or he talented to read documents. It is also possible if a user wants to retrieve documents in multiple languages by expressing a query in a single language.

CLIR is designed to solve the problem of these user situations. Cross-language information retrieval (CLIR) can momentarily be defined as a subfield of information retrieval system that deals with searching and retrieving information written in a language different from the language of the user's query. Facilitating the procedure of revenant documents that written in one language with computerized systems that takes queries in other language(s) is thus the major function of Cross Language Information Retrieval system (Hedlund, 2004). CLIR provides relevant documents to the user queries though the language of query and document is distinct. The language that the query uses is referred to as the source language (e.g. Afaan Oromo) and the language of the documents is the target language (e.g. Arabic).

Even though target language (document) translation into the source language (query) is comfortable for the user, it is expensive and hard to implement as it should obey complex rules of the natural language than query translation (Talvensaari T. , 2008). According to (Abusalah, M; Tait, , 2005) once it is known that relevant information exists, and retrieved documents can be translated to the user's native language either by human translator or automatically by machine translation system. So, the query translation approach is more common in CLIR and applied in present research as well. Cross

Language Information Retrieval has been studied widely in different languages, such as English, Chinese, Spanish, and Italy. Much research works have been reported and evaluation results have, in general, been satisfactory.

In the past, research in Cross Lingual Information Access (CLIA) has been powerfully practiced through, the Cross Language Evaluation Forum (CLEF), Question-answering Workshop and cross-language named entity extraction challenges by the Association for Computational Linguistics (ACL) and the Geographic Cross Language Evaluation Forum (GeoCLEF) track of CLEF (IJCNLP 2008). Arabic is dominant language in the web with which many document and online resources are produced with. As a result, access to non-Arabic Internet users has become an important challenge in recent times. When non Arabic users want to use search engines to retrieve, most of the time, they arrive at improper formulation of Arabic queries. Afaan Oromo users who use IR system to search the web and there is also huge amount of document written in Arabic. Researchers in the area of CLIR have been focused mainly on methods for query translation.

In particular, dictionary-based translation approach is a commonly used method because of its simplicity and the increasing availability of machine readable bilingual dictionaries (Oard D. , 1997). In any of such cases Cross Language Information Retrieval will be expected to support queries in one language with a collection in another language. As we move towards an increasingly globalized and knowledge-based economy, the ability to discover and share information across language and cultural boundaries will become essential. With the arrival and fast development of the Internet, the amount of information generated in different languages and disseminated by means of the World Wide Web and social media platforms is growing exponentially (Abusalah, M; Tait, , 2005). As the Internet becomes more ubiquitous and pervasive, users will become linguistically more diverse and culturally more heterogeneous. It is thus, the highest significance to ensure that online information resources and services are efficiently and fairly accessible to all users, regardless of their language and cultural backgrounds. Therefore, the focus of this research is to enable Afaan Oromo speakers to access and retrieve the vast online information resources that are available in Arabic using their own language queries by employing CLIR system that is based on Afaan Oromo –Arabic dictionary based on approach of query translation.

## 1.2. Statement of the Problem

Nowadays, the amount of digital data generated in diverse languages and made available through the World Wide Web and social media platforms are growing exponentially; tremendous amount of multilingual resource are found on the web. An automatic query translation tool will be very helpful to such users and query translation tool will also be helpful to retrieve relevant documents written in other language different from the query language. To achieve all these, Cross Language Information Retrieval will be a useful tool. Hence, translation of local queries into Arabic has been critical for making these useful online documents accessible for the local use as access to foreign Web pages is becoming an increasingly key problem. Afaan Oromo writing in Latin alphabets began only after 1991 (Tilahun, 1993). While Arabic writing in abjad script; when the two languages have different alphabets, is to translate the unknown words, that is, to render them in the orthography of the second language. As a result, most of the documents available on the internet, which are relevant to the Afaan Oromo speaking society, might also be available in other languages. Among these languages, Arabic is the most popular one. Afaan Oromo speakers want to speak Arabic because of many fundamental facts. Among them are the impact of Islamic teachings, the historical documents written about Ethiopia in general and of Oromo people in particular on Futuh al-Habesha, trade relationship on the east Oromia, diplomatic relation –for example since King Abajifar II of Jimma in west Oromia, religious pilgrimage to Mecca for Hajji can be listed. These days, the issue of public diplomacy is playing crucial role in the lives nations and this would also strengthen the reality there. We should not also forget that the wider globe is now becoming just a village, which would even encourage us to learn many more languages.

Languages are crucial tools for the easy accessibility of the enormous collection of documents on the internet. Usually, people have one language they favor to use. In general, while people may have the skill of speaking more than one language, it is not common for people to have equal competence in many. According to (Sisay, 2009), these resources of the documents available on Web site are inaccessible for most of the Afaan Oromo speakers due to lack of the knowledge of articulating queries in other languages.

To render the easiest accessibility of these documents to users, the barrier found between these two languages needs to be resolved. Thus, the basic problem of this study is to fill this gap by design and development of Afaan Oromo-Arabic Cross Language Information Retrieval system with a view to enable Afaan Oromo speakers to access and retrieve the huge online information sources that are obtainable in Arabic language by using Afaan Oromo query.

## 1.3. Research Question

In this study, to achieve the objectives of the study effective effort has been made to response the following basic research questions.

- What are the linguistic facial appearance of the two languages and their appropriate text process?
- To what coverage CLIR system can be put into practiced for a pair of languages using dictionary?
- How efficient does the CLIR system developed for the two languages suit information need of user?

## 1.4. Objectives of the Study

The general and specific objectives of the study are described as below.

### 1.4.1. General Objective

The general objective of this research is to explore and design Dictionary based Afaan Oromo-Arabic Cross Language Information Retrieval system.

### 1.4.2. Specific Objective

To achieve the general objective stated above, the following specific objectives accomplished throughout the study:

- To review related works on Afaan Oromo-Arabic CLIR.
- To set up and organize test documents and queries.
- To explore and identify an appropriate procedure for designing architecture for Afaan Oromo-Arabic CLIR.
- To translate Afaan Oromo queries by using bilingual dictionary.

- To design and develop CRIL prototype that uses Afaan Oromo queries and retrieve both Afaan Oromo and Arabic documents from the test compilation
- To evaluate the performance of the CLIR system; and
- To discuss and report experimentation results and recommend further research works in the area.

## 1.5. Significance of the Study

The main contribution of the study was the exploration of bilingual dictionary based Cross Language Information Retrieval system to benefit Afaan Oromo speakers to use Information Retrieval systems to satisfy their information need from a corpus by writing queries using their native language. Users enter their query in their native language (Afaan Oromo) and the system retrieves relevant documents in Afaan Oromo as well as Arabic. The retrieved documents that are relevant for the given query can promote those who can understand contents of Arabic documents that are return for the given query. That is based on dictionary-based query translation techniques with a view to enable Afaan Oromo speakers to access and retrieve the vast online information resources that are available in Arabic by using their own language queries. Besides, the attempt was contribute to future guide researcher in the area of information retrieval specially in developing multilingual information retrieval for other local languages. In addition, this study was helpful to develop Afaan Oromo corpus across the web as well as to establish relationship with the Arab World. As a final point, the outcome of this study gave benefit to individuals, groups and future researchers.

## 1.6. Scope and Limitation of the Study

### 1.6.1. Scope of the Study

The scope of this study has been restricted to the translation of Afaan Oromo queries into Arabic using Dictionary-based approach in order to retrieve Afaan Oromo from the documents. The translation of Afaan Oromo queries into Arabic has been done using bilingual dictionary built from a corpus. After query translation, the major information retrieval processes (indexing and searching) were performed.

### 1.6.2. Limitation of the Study

However the Dictionary-based approach requires quite a large number of documents for a better arrangement performance, the system uses only a small size of Afaan Oromo-Arabic documents. This is due to constraints such as time, resource and lack of computational linguistic resources and translation tools. Training of large corpus requires high processing speed and large memory of the computer. Bilingual dictionary has been built only translating single word at a time.

## 1.7. Methodology of the study

This research has used the following methods in order to figure out Afaan Oromo-Arabic cross language information retrieval (CLIR).

### 1.7.1. Literature Review

To achieve the objectives of this research pointed out above a numerous articles, books, journals and the internet were widely reviewed. Materials concerning Afaan Oromo language and Afaan Oromo-Arabic and other languages CLIR were also reviewed. Special notice has been given for both the local and global works for Afaan Oromo and Arabic in IR system were reviewed.

### 1.7.2. Data Collection

Dictionary-based approach of query translation requires the availability of a corpus, which contain translation of a given document in other languages. For this research some Afaan Oromo-Arabic documents that are publicly accessible were used. To perform the research some Qur'ans' and Bible chapters accessible in both languages were comprised. These documents were selected as they are already translated, have relatively standard translation and easy to acquire.

### 1.7.3. Data Preprocessing

To arrange collected documents for the bilingual word aligner tools (MatlabR2012b for this research) and some preprocessing such as tokenization, case normalization, and stop word removal was carried out. The data was also organized in such technique that it is appropriate for the software tool that was chosen for this research.

### 1.7.4. Query Preparation

To evaluate the effectiveness of the prototype, bench mark Afaan Oromo queries have been organized for the selected corpus for testing reason. A total of 20 Afaan Oromo queries were organized from the 40 selected Afaan Oromo documents. The chosen test documents were selected at random from the accessible documents. In view of the fact that, all the collected corpora for testing reason is infeasible random sampling was mandatory, because all documents were thought to be equally vital. The queries were prepared by native speaker of Afaan Oromo after reading the content of the documents for which they will prepare query.

### 1.7.5. Manual Alignment

The most important objective of this research is to retrieve both documents in Afaan Oromo and Arabic by using Afaan Oromo query. The translation of the Afaan Oromo query into its equivalent Arabic query done based on Afaan Oromo-Arabic bilingual dictionary which was constructed manually from corpora collected. Therefore, to build this bilingual dictionary for the reason of query translation an alignment has to be done between matching words in corpora based on manual alignment.

### 1.7.6. Experimentation and testing

The experimentation for evaluating the effectiveness of the system was done by using selected test documents and queries. To this end, the base line Afaan Oromo queries were presented for the CLIR system to retrieve Afaan Oromo documents judged to be relevant by the system for the given query. The experiment was also conducted by using the Arabic queries (translated from Afaan Oromo queries) for the retrieval of the Arabic documents that were judged to be relevant by the system for the specified query. The result obtained by the translated Arabic queries judged against the result of Afaan Oromo bench mark queries were used to evaluate the performance of the Afaan Oromo-Arabic CLIR system. The result obtained by the translation Arabic queries against the result of the bench mark is the judgments by the expert who selected the list of relevant documents for each query. Recall and precision techniques were used for measuring retrieval

effectiveness of the cross language information retrieval system; as they are frequently used and most basic measures of IR effectiveness.

### 1.7.7. Software Tool and Programming Language Used

MatlabR2012b was selected as a programming language for this research. MatlabR2012b is an interpreted, object-oriented, high level programming language. It enables there searcher to implement the functionality of the sought system without any hassle and allows writing programs that are clear and readable.

## 1.7. Organization of the Thesis

This thesis is prepared in five chapters. The first chapter, contain: background, statement of problem, objective of study, research question with scope as well methodology and the contribution of study. Chapter two briefly describes the various works done related to CLIR system. It also gives some general background of the Afaan Oromo language with its typical characteristics such as alphabets, punctuation marks, and articles. The chapter also presents the general overview of CLIR and its relation with IR, some possible approaches of CLIR and challenges in manuals word alignment is occurred. Chapter three discussed the proposed architecture of the system, data collection, data preprocessing are presented. The major components involved in the architecture of the CLIR system are also explained in detail. In this chapter the description of Dictionary-based on Afaan Oromo Arabic cross language information retrieval described. In chapter four the experimentation and evaluation of the planned system are discussed. Analysis of the results acquired from the experiment is also presented in this chapter. As a final point, chapter five winds up what has been done in the research. It reviews the results gained and forwards recommendations for the future work.

# Chapter Two

## 2. Literature Review

## 2.1. Introduction

This chapter presents an outline of Information Retrieval and focus e on Cross Language Information Retrieval. Normally, the chapter can be seen as the following. The first part gives a conceptual overview of Information Retrieval and Cross Language Information Retrieval. Overview of Afaan Oromo and Arabic Language, Models of Information retrieval and also, this topics regarding Cross Language Information retrieval including CLIR Approaches, Translation Strategies has been discussed in the first part. The second part focuses on review of related works on Cross Language Information Retrieval. In this part different research works with their results have been analyzed. This part focuses both on global and local works done so far in the area of Information retrieval.

## 2.2. A Brief Overview of Afaan Oromo

Afaan Oromo is among the major languages that are widely spoken and used in Ethiopia (Abara, 1988). It is considered to be one of the five most widely spoken languages among one thousand languages of Africa (Grage, G. Kumsa, T., 1982). Afaan Oromo, although relatively widely distributed within Ethiopia and some neighboring countries like Kenya and Somalia, it is one of the most resource scarce languages (Tilahun.G, 1993) Afaan Oromo is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic phylum (Abara, 1988), unlike Amharic (an official language of Ethiopia) which belongs to Semitic family languages. Although it is difficult to identify the actual number of Afaan Oromo speaking society (as a mother tongue) , due to lack of appropriate and current information sources, according to census taken in 2007 it was estimated that 34.51 percent of Ethiopians are ethnic Oromo..

Afaan Oromo has a very rich morphology like other African and Ethiopian languages (Oromoo, 1995). Since 1991 Afaan_Oromo writing system (Qubee) has been adopted and became an official script (Tilahun.G, 1993). It is used for writing and spoken language in Ethiopia and as well as neighboring country like Somalia and Kenya (Kula, 2008). Now a day, Afaan Oromo is used as an instructional media for primary and

secondary school and also used as an official language of Oromia Regional state. It is also given as a subject starting from grade one throughout the schools of the region. Furthermore, few literature works, newspapers, magazines, educational resources, official documents and religious writings are written and published in this language. As far as the knowledge of the researcher is concerned there are no works done in the area Afaan Oromo –Arabic cross language information retrieval. Thus the purpose of this study is to provide Afaan Oromo –Arabic dictionary based approach.

### 2.2.1 Alphabet of Afaan Oromo and Arabic Alphabet (الحُرُوْفُ اَ لْهِجَاءِ)

Afaan Oromo uses Qubee (Latin based alphabet) that consists of twenty-nine basic letters of which five are vowels, twenty-four are consonants, out of which five are pair letters, In Afaan Oromo language, vowels are sound makers and are sound by themselves. In addition, vowels in Afaan Oromo are characterized as short and long vowels (Tilahun.G, 1993). While Arabic language uses abjad script Alphabet (Huruful-hijae(الحُرُوْفُ اَ لْهِجَا) ) that consists of twenty-eight basic letters of which twenty-five are consonants and three of them are vowels including this edition features (al-fathah , al-kasrah, al-damah, al-shaddah, al-sukun, al-madda, and al-tanween, respectively ( ْ) , المده(~ ) ,التنون ( ً , ٍ ,ٌ) ا (َ ) ي,مع (ِ ) و, (ُ ) الشدة, (ّ ) السكون,( ) . Afaan Oromo alphabet characterized by capital and small letters, unlike Arabic alphabet, which does not have a distinction between capital and small letters (Alduais, 2012).

The basic alphabet in Afaan Oromo does not contain ‘p’, ’v’ and ‘z’. This is because there are no native words in Afaan Oromo that are formed by these characters. However, in writing Afaan Oromo language they are used to refer to foreign words such as *“polisii” (“police”)*.

### 2.2.2 Dialects and Varieties

Depend on geographical area Oromo have different diversities that makes Afaan_Oromo is a sociolinguistic language: Harargy, Borana_Gujii, Arsii, Western central Oromo and Orma (Oromo in Kenya). Even though, there is a strong similarity, but also difference in dialect between them. Alike Afaan Oromo Arabic also has a sociolinguistic language which is used in different Arab World. (Adola, 2012.)

### 2.2.3. Punctuation Marks

Punctuation marks used in both Afaan Oromo and Arabic languages are the same and used for the same purpose with the exception of Apostrophe mark (‘). Afaan Oromo it is used in writing to represent a glitch (called *hudhaa*) sound. It plays an important role in the Afaan Oromo reading and writing system. For example, it is used to write the word in which most of the time two vowels are appeared together like *“du’a”* to means (“*di’e”*) and in Arabic ( “مَا ت”). with the exception of some words like *“har’a*” to mean *“today*” *and also in Arabic“اليوم”* which is identified from the sound created.

### 2.2.4. Conjunctions

Conjunctions are used to connect words, phrases or clauses. In Afaan Oromo there are different words that are used as conjunction. Some conjunction words in Afaan Oromo are listed in table

| Conjunction | English equivalent | Arabic equivalent |
|---|---|---|
| fi | and | وَ |
| yookaan(yookiin) | or | أَوْ |
| Waanta’eef | So, therefore | ولذا |
| yoo | unless | أو |
| waan | for | ل |

**Table2. 1 Some Afaan Oromo conjunctions (Hamiid, 1995.)**

### 2.2.5. Grammar

All languages have their own regulations and syntax, Afaan Oromo and Arabic also have their own grammar (_seer-luga’ in Afaan Oromo).

#### 2.2.5.1 Adjectives

Adjectives are very important in Afaan_Oromo and Arabic because their structure is used in every day conversation. Afaan_Oromo and Arabic Adjectives are words that describe another person or things in the sentence. (Mylanguage.org, 2011).

| Bifoota, Colors, |
|---|

| Afaan Oromo | English | Arabic |
|---|---|---|
| gurraacha | black | اسود |
| Baluu | blue | ارزق |
| daalacha | gray | اصفر |
| magariisa | green | اخضر |
| diimaa | red | احمر |
| adii | white | ابيض |
| Hamma , Size | | |
| guddaa | big | كبير |
| gadifagoo | deep | عمق |
| lafarradheerat | long | طو يل |
| dhiphaa | a narrow | ضيق |
| gabaabaa | short | قصير |

**Table2. 2 Some Afaan Oromo Adjectives (Hamiid, 1995.)**

### 2.2.5.2. Prepositions

Afaan Oromo preposition is used to links: the nouns, pronouns and phrases to other words in a sentence. The objective of prepositions is to introduce the word or the phrase. (Mylanguage.org, 2011). Here are list of some prepositions:

| Afaan Oromo Prepositions | English Prepositions | Arabic Prepositions |
|---|---|---|
| waa'ee | about | حولي |
| gubbaa / gararraa | above | فوق |
| Gama | across | مقابل |
| booddee / booda | after | بعد |
| Faallaa | against | عكس |
| jaragiddu | among | بينهم |
| naannoo | around | ناحية |
| Akka | As | مثل\شبه |
| Dura | Before | أمام\اعزاء |

**Table2. 3 Some Afaan Oromo Prepositions (Hamiid, 1995.)**

### 2.2.7. Word Segmentation

The word is the smallest unit of a language. There are different methods for separating words from each other. This method might vary from one language to another. In some languages, the written or textual script does not have whitespace characters between the words. However, in most Latin languages a word is separated from other words by white space character (Meyer, 2008). Afaan Oromo is one of Cushitic family that uses Latin script for textual purpose and Arabic is one of Semantic family and both are uses white space character to separate words from each other's. For example, "*Hassen Finfinnee deeme*" it means 'ذهب حسن الي فنفني'. In this sentence the word "*Hassen"," Finfinnee"* and*" deeme*" are separated from each other by white space character. Therefore, the task of taking an input sentence and inserting legitimate word boundaries, called word segmentation, is performed using the white space characters on what information is to be translated: topics, documents or both. The best results are obtained

## 2.3 Information Retrieval: An Overview

In the beginning of web in 1990s it became so straightforward for a web customer to push/ put their thoughts and documents on the web. This made the web as human knowledge and cultural which has been allowed unparalleled distribution of ideas and information in the regulation not seen before. In spite of, so much attainment the web has been introduce new problems of its self. One of the problems is being searching useful information on the web became complex task. These difficulties have been attracted and renewed over the researcher interest in IR and its techniques for promising solutions (Manning et al, 2008). Classical Information Retrieval is the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents (Abusalah, M; Tait, , 2005).

Data repossession in the perspective of an Information Retrieval system, consists essentially to settle on the collection of a documents for the user query by holding the keywords which, is not sufficient to satisfy the user need whereas user of an IR system is not concerned with retrieving data which satisfy a given query, but also concerned with

retrieving information about a subject only. (Talvensaari T. , 2008). Unfortunately, explanation of the customer information need is not a simple problem. As a replacement for the user information need must translate into a query to be processed by the search engine.

The translation yields a set of index terms which review the explanation of the user need. The key objective of an IR system is to retrieve information which might be useful or relevant to the user by taken a query from the client. The significance is on the retrieval of data as opposed to the retrieval of information. (Baeza R and Ribeiro, 1999)

### 2.3.1 Information Retrieval Models

We need information retrieval model because models can serve as a blueprint to implement an actual retrieval system in addition to their use in guiding research and providing means for academic discussion. One dominant problem regarding information retrieval systems is the issue of forecasting, which documents are relevant and which are not based on the ranking algorithm. A ranking algorithm works according to some specific premises regarding the concept of document relevance which is the retrieval model of the system (Manning et al, 2008). The IR model adopted controls the forecasts of what is relevant and what is not. An Information Retrieval system applies a retrieval model that comprises of the internal depiction of queries and documents, and the arrangement of a matching algorithm. The matching arrangement defines the way in which the document and query depictions are compared to measure the relevance of the document to the queries (Baeza R and Ribeiro, 1999).

1. Clearly characterized IR model as a quadruple {d, q, f, R ($q_i$, $d_j$)} where,
2. d is a set composed of logical representation for the documents in the collection.
3. q is a set composed of logical representation for the user information needs. Such representations are called queries.
4. f is a framework for modeling document representations, queries, and their relationships.

5. R ($q_i$, $d_j$) is a ranking junction which associates a real number with a query $q_i \in q$ and a document representations $d_j \in d$.

Such ranking defines an ordering among the documents with regard to the query $q_i$. There are three classic models in information retrieval: Boolean models, Vector Space models, and Probabilistic models (Baeza R and Ribeiro, 1999).

#### 2.3.1.1 Boolean Model

The Boolean model: simple retrieval model based on set theory and Boolean algebra which had great consideration in the past times. The Boolean model considers that keywords are present or absent in a document. As a result, the keywords weights are assumed to be all binary, i.e., Wi,j{0,1}. A query $q_i$ s composed of keywords linked by three connectives: NOT, AND, OR. The Boolean model forecasts that each document is relevant or non- relevant. There is no idea of a partial match to the query conditions. Although, the Boolean model has been cleaned formalism and simple it still suffers from major drawbacks. First, its retrieval strategy is based on a binary choice principle without any idea of a scoring rule, which avoid high-quality of retrieval performance. Another drawback is, since Boolean expressions have exact semantics, regularly it is not easy to translate an information need into a Boolean expression (Manning et al, 2008).

#### 2.3.1.2 Vector Space Model

The vector space model tries to improve weakness of Boolean model that came along with the use of binary weights which is moreover limited for partial matching. The vector space model intends a framework in which incomplete matching is possible by assigning non-binary weights to document and index terms in queries. (Ballesteros, L. and Croft, B., 1996).

The vector model takes the document into consideration by sorting the retrieved documents in decreasing order and calculating term weights of the degree of similarity between each documents and query, which match the query terms only partially (Manning et al, 2008). The vector space model intends to evaluate the degree of similarity of the query $q_i$ with regard to the document $d_j$ as the correlation between the vectors $q_i$and $d_j$. This correlation can be enumerated by the cosine of the angle between

these two vectors assign. Normally, there are three main procedures involves in the whole VSM; to represent the content bearing of terms in the document, indexing the document is the only solution. To enhance retrieval of relevant documents, weighting the indexed term is mandatory. Finally, to show the best matching between query and document, ranking all the documents in the corpus. The main advantage of VSM is:

(1)Retrieval performance is improves by term weighting.

(2) Retrieval of documents that approximate the query condition is rely on partial matching pals.

(3) The documents sorts according to their degree of similarity to the query by using cosine ranking formula Although, it is the strong and most popular ranking strategy now days, theoretically the VSM has the disadvantage that index terms are assumed to be mutually independent and also it is computationally expensive since it measures the similarity between each document and the query (Baeza R and Ribeiro, 1999).

**2.3.3.1 Probabilistic Model**

The probabilistic model tries to solve IR problem within a probabilistic framework. The fundamental idea is that, given a user query, the document collection can be divided into two sets; a set which hold exactly the relevant documents and a set. This holds the non-relevant documents. In other words it creates a perfect respond set. This means that the querying process is a process of identifying the properties of the perfect respond set which is not known exactly. However, since the properties of the relevant document set is not known in advance, there is always the need to forecast at the beginning the descriptions of this set (Manning et al, 2008). The user then takes a look at the retrieved documents for the first query and decides which ones are relevant and which ones are not. By repeating this process iteratively the probabilistic description of the relevant document set can be improved (Baeza R and Ribeiro, 1999). Absence of initial information on the relevant and non-relevant documents in the collection with admiration to specific query, its unawareness to the frequency with which an index term happens in a document, and the assumption of index terms independence are taken as its major faintness.

## 2.4. Possible Approach to Cross Lingual Information Retrieval

CLIR does not differ too much from IR and it only focuses on the techniques of translation process, to map a written word from one language-script pair to another language-script pair, to solve the language barrier. For example, Afaan Oromo word *"kitaaba"* corresponds to English word *"book" and also agrees Arabic word ' كتاب'*. Therefore, in CLIR to retrieve relevant documents the language of the queries and documents should not be necessarily the same. Approaches to CLIR can be categorized according to how they solve the problem of matching the query and documents across different languages. In CLIR, either the query or the document needs to be mapped into the common representation to retrieve the relevant documents. Translating all documents into the query language performs significantly better than query translation as results of some experiments identify (Oard D. W., 1998)but translating all documents into the query language requires huge storage space and it is computationally expensive (Hull, 1996). Due to this, query is usually translated into the language of the target collection of documents (Jagarlamudi, 2007). Thus, this research is focused on the accuracy of query translation. The basic approaches in translation of queries into a language of target documents can involve the following approaches: Machine Translation (MT), Machine-Readable Dictionary (MRD) and Corpus Based approaches (Abusalah, M; Tait, , 2005). In the following sections, the detail descriptions of these possible approaches are presented (Manning et al, 2008).

### 2.4.1 Machine Translation Approach

Machine translation (MT) is the translation of text from one human language into another human language by the use of a computer. In order to accomplish this it needs texts in one specific language as input and generates texts with a corresponding meaning in another language as output. Thus, machine translation is a decision problem where we have to decide on the best of target language text matching a source language text (Mathematik, V.at el, 2002). In CLIR MT can be implemented in two different ways (Abusalah, M; Tait, , 2005). The first way is to use MT system to translate foreign language documents in the corpora into the language of the user's query.

This approach is not applicable for large document collections, or for collections in which the documents are in multiple languages. Document translation approach can be accomplished on an 'as-and-when-needed' basis at query time, referred as on-the-fly translation, or in advance of any query processing (Ramanathan, 2003). The second method of using MT in CLIR is where the users query language is translated into the language of the documents in the stored collection. With both methods of MT, an ambiguity problem can exist since the translated query does not necessarily represent the sense of the original query. For instance, translating the Afaan Oromo query *"Jia'a"* to Arabic language could produce an inappropriate translation since it is not clear whether to mean *"شهر"* or *""قمر "Month" or "Moon".* Due to this, MT is more efficient in document translation when the context is unambiguous.

### 2.4.2. Dictionary query translation Approach –based

Dictionary based query translation is a CLIR approach in which the query words are translated to the target language manually (Ballesteros, L. and Croft, B., 1996). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination of both. It has been adopted in CLIR because bilingual dictionaries are widely available. Dictionary-based approaches are straightforward to implement, and the efficiency of word translation with a dictionary is high (Lu, C.et al ., 2008).

## 2.5 Related Works

Relay on reviewed literatures and knowledge of the researcher there is no work done exactly on this area. No developing and designing retrieval system is developed for Afaan Oromo-Arabic Cross Language Information Retrieval System. Nonetheless, some scholar made effort to enable Cross Lingual IR system for Afaan Oromo-English, Afaan Oromo-Amharic, Amharic-English and other global works. Some works has been reviewed so far on the related area as following.

### 2.5.1. On Arabic-English Cross-Language Information Retrieval: A Machine Translation Approach

Mohammed Aljlayl, Ophir Frieder, & David Grossman Suggested that Arabic-English CLIR focuses on the retrieval of documents based on queries formulated by a user in the Arabic language and the documents are in the English language. Presently, no benchmark data are available for Arabic English CLIR. To provide a means to compare their efforts with future Arabic-English CLIR efforts, they used readily available English benchmark document collections and provide our Arabic queries, a translation of the National Institute of Science(NIST) and Technology, Text Retrieval Conference (TREC) (Aljlayl, et al, 2002). Briefly, TREC has three distinct parts: the documents, the topics, and the relevance judgments. They used two collections. The first, the 2.1 GB TIPSTER Disks 4 and 5, and the second, the 10 GB TREC-9 Web collections, consist of roughly 500,000 and 1,700,000 documents, respectively. For queries, they manually translated the TREC-7 (topics 351-400) and TREC-9 (topics 451-500) queries to Arabic. They used these 100 translated versions as their original Arabic queries issued against the TREC English collection. The Arabic queries were translated back to English using the ALKAFI MT system. Indexing is done using the Porter and K-stem algorithms after eliminating the stop-words. Similarity, querying is done after stemming and eliminating the stop words of the translated target English queries. The ALKAFI Arabic-English MT system is a commercial system developed by CIMOS Corporation and is the first Arabic to English machine translation system. The system attempts to analyze terms in context and then builds semantic links. Then, the English text is generated by the transfer technique according to English language grammar (Aljlayl, et al, 2002). They evaluate the effects of the MT system in Arabic English CLIR. As described earlier, they used both the TREC-7 and TREC-9 topics and collections. For TREC-7, the machine translation achieved 61.8%, 64.7%, and 60.2% for title, description, and narrative fields, respectively. While the average precision of TREC-9 topics. , the machine translation achieved 58.4%, 57.1%, and 53.4% for title, description, and narrative fields, respectively. They focus on Arabic-English machine translation, and in particular, on the evaluation of the ALKAFI Arabic-English MT system for Arabic-English CLIR. They investigated strictly the effectiveness (accuracy) and not its efficiency (speed of

processing) of the MT-based Arabic-English CLIR. High speed processing is meaningful only when high accuracy is obtained (Aljlayl, et al, 2002).

### 2.5.2 Effective Arabic-English Cross-Language Information Retrieval via Machine-Readable Dictionaries and Machine Translation

With the rapid growth of the Internet, the World Wide Web (WWW) has become one of the most popular mediums for the dissemination of multilingual information. This ability to disseminate multilingual information has increased the need to automatically mediate across multiple languages, and in the case of the WWW, access to "foreign language" Web pages. Their goal for Arabic Cross-Language Information Retrieval (CLIR) is to enable users to query in the Arabic language against an English collection. To achieve this goal, they investigate two techniques namely the use of Machine Readable Dictionaries (MRD) and Machine Translation (MT) based approaches. In the MRD realm, they consider three possible techniques: the Every-Match (EM), the First-Match (FM), and the Two-Phase (TP) methods. The Every-Match method considers all the translations found in a bilingual dictionary (Frieder, 2001). This leads to ambiguous translations because it introduces extraneous terms to the target query and yields relatively poor effectiveness. Another method is the First-Match method. Instead of considering all the target language equivalents in the bilingual dictionary, they use the first match in the bilingual dictionary as the candidate translation of the source query term. This approach takes advantage of the fact that dictionaries typically present the translations in the order of their common use. That is, the more common translations are listed first (Frieder, 2001). The FM method ignores some of the less common translations of the source language, and thus, potentially improves the retrieval effectiveness. Finally, the Two Phase method initially considers all the translations found in the bilingual dictionary as candidate terms then removes the translated candidate terms that do not return the original source query term. They found that the TP approach consistently outperforms the EM and FM methods. In addition to MRD, they also empirically evaluated the effectiveness of Arabic to English MT-based method. They summarize the average precision results for the TREC-7 collection and topics (351-400). The machine translation results are better than the Every-Match method (EM) in all runs. It yields to

61.3% of monolingual retrieval. As shown in both the FM and TP methods outperform the machine translation approach (Frieder, 2001). The reason behind the degraded effectiveness of the machine translation is that the used machine translation system is designed to perform best on well-formed sentences or at least any sequence of words that form a context. The titles of topics 351-400 are all three words or less. The effect of greater context is also apparent in the performance of machine translation using the description field of query topics 351-400 as shown in A comparison of the baseline, EM, FM, TP, and MT methods is represented in using the average precision and recall. As shown, the MT based technique outperforms the EM method, but again is lower than FM and TP methods the description field yields 64.6% and the narrative field yields 61.9% of the monolingual retrieval. According to these findings they conclude that the MT system performs best once a context is determined. That is, adding more terms to the full context query does not help the machine translation to disambiguate the term pulse. Since their results are only preliminary, no additional conclusions (Frieder, 2001)

### 2.5.3. Amharic – English Cross-lingual Information Retrieval: A Corpus Based Approach

As Aynalem Tesfaye, Kevin Scannell mentioned on Amharic – English Cross-lingual Information Retrieval: A Corpus Based Approach requires quite a large amount of parallel text in order to achieve a fairly good level of performance. However, the size of the parallel corpus that was used for conducting this research was not sufficiently large. In addition to the size, the quality of the parallel text greatly affected the performance of a corpus-based CLIR. Despite the fact that the research requires quite a large volume of parallel Amharic English text with good quality, the results found after conducting the second phase of the experimentation was a maximum precision value of 0.24 and 0.33 for Amharic and English respectively (Aynalem.T, Kevin.S, 2010).

### 2.5.4. Afaan Oromo–English Information Retrieval (CLIR): A Corpus Based Approach

Afaan Oromo–English Cross Lingual Information retrieval in the approach of corpus based was developed by Daniel Bekele for the first time (Ayan, 2011) as partial fulfillment of Master of Science in Information Science at Addis Ababa University. The

first a dictionary based Oromo- English CLIR has been developed by Kula Kekeba, VasudevaVarma and Prasad Pingali (Kula, 2008). The objective of the research by Daniel Bekele is to enable Afaan Oromo users to specify their information need in their language and to retrieve documents in English. The research work is based on a parallel corpus collected from legal, Bible chapters and some available religious documents for testing and training purpose. Since document translation is computationally expensive. The translation approach used in this work is word based query translation for the two language pairs, Afaan Oromo & English, (Hull, 1996). The study is conducted using 530 parallel documents and all the collected documents were used for the construction of the Afaan Oromo-English bilingual dictionary. The collected data has to go through different pre-processing tasks (data preparation, tokenization, and case 35 normalization) in the way appropriate for the word alignment tool and information retrieval task. The system is established using a statistical word alignment tool called GIZA ++. Vocabulary and bi text files are the two obligatory input files for GIZA++ tool for the formation of the word alignment (Och, 2003). These files were produced by the packages available in GIZA++ toolkit. Formerly, the statistical information of vocabulary and bitext file produced was used as input for the GIZA++ to generate word alignment. In this situation the researcher developed the bilingual dictionary and this dictionary served as translation knowledge source. Experimentation was carried out in two phases. In the first phase the un-normalized edit distance was used to relate difference of words between query and index terms and also the second phase of the experimentation by using normalized edit distance. Evaluation of the system is conducted for both monolingual and bilingual retrievals. In the bilingual run the Afaan Oromo queries are given to the system after being translated into English to retrieve English documents while in the monolingual run, Afaan Oromo queries are given to the system to retrieve Afaan Oromo documents. The performance of the system was measured using recall and precision. Evaluation was conducted for 60 queries and 55 randomly selected test documents. In the first phase of the experimentation, maximum average precision value of 0.421 and 0.304 are obtained for the Afaan Oromo and English documents respectively. In the second phase maximum average precision value of 0.468 and 0.316 are obtained for the Afaan Oromo and English documents correspondingly. The second phase of experimentation achieves

slightly better than the first. In the approach of, the experiment results, the researcher sum up with the use of huge and cleaned parallel Afaan Oromo-English document collections, there for developing CLIR for the language pairs is possible (Daniel, 2011).

### 2.5.5. Oromo-English Cross Language Information Retrieval

According to Kula et al. (2008) the bilingual information retrieval experiments as part of specific has been conducted by three different languages: Oromo-English, Telugu-English and Hindi English. The experimentation result of Oromo-English cross language information retrieval experiments at CLEF'06 is summarized as the following. The basic objective of the study was to design and develop an Oromo-English CLIR system with a view to enable Afaan Oromo speakers to access and retrieve the vast online information resources that are available in English by using their native language queries. In the paper, the dictionary based approach of CLIR was used to translate Oromo topics into a bag of words of English queries. In the experimentation three different official runs were yield to Oromo-English bilingual task. Afaan Oromo queries were translated into their equivalent English queries based on Machine Readable Dictionary (MRD) and submitted to the text retrieval engine. The English test collection was provided by CLEF (Cross Language Evaluation Forum) from two newspapers, the Glasgow Herald and the Los Angeles Times. The Oromo-English dictionary the researchers had adopted and developed from hard copies of human readable bilingual dictionaries by using Optical Character Recognition (OCR) technology. The developed dictionary was used to translate Oromo topics into a bag of words of English queries. In order to define Afaan Oromo stop words, the researchers first created a list of the top most frequent words found in Afaan Oromo text corpus collected. Then pronouns, conjunctions, prepositions and other similar functional words in Afaan Oromo were also added in the stop word list. After these less informative words were removed from Afaan Oromo text corpus, a light stemming algorithm had been applied in order to conflate word variants into the same stem or root. In lieu of indexing and retrieval of the documents Lucerne an open source text search engine was used. Once eliminating the stop words, the stemmed keywords of Oromo topics were automatically looked up for all possible translations in the bilingual dictionary. One of major problem of query translation using dictionary-based approach is

OOV due to the vocabulary limitation of dictionaries. Because of this limitation, some of the words in a query cannot be found in a dictionary. The researchers also faced this problem during query translation process since the match of those words was not found in the dictionary. To minimize this limitation of dictionary based query translation the researchers tried to handle manually which is not applicable for large collections of document.

The experimentation results of the three official runs, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN) for Oromo-English bilingual task in the specific track of CLEF'06 summarized. The Mean Average Precision (MAP) and Geometric Average Precision (GAP) scores of the three runs are also presented.

| Run-Label | Relevant-tot. | Rel. Ret. | MAP | R-Prec. | GAP |
|---|---|---|---|---|---|
| OMT | 1,258 | 870 | 22.00% | 24.33% | 7.50% |
| OMTD | 1,258 | 848 | 25.04% | 26.24% | 9.85% |
| OMTDN | 1,258 | 892 | 24.50% | 25.72% | 9.82% |

**Table2. 4 Summary of average results for the three runs (Kula, 2008)**

The evaluation of this experiment was also conducted by (Kula, 2008)and the purpose of the evaluation experiment was to evaluate the overall performance of the dictionary approach by using these different fields of Afaan Oromo topics. Based on the evaluation result it was concluded that limited language resources can be used in designing and implementation of a CLIR system for resource scarce languages like Afaan Oromo. The important average results for all official runs were achieved even though limited CLIR resources used in the experiments. The dictionary approach is easy to implement and good choice for CLIR as MRD can easily be found (Ballesteros, et al, 1998). Though, due to the vocabulary limitation of dictionaries, very often the translations of some words in a query cannot be found in a dictionary. According to David et al. (David, 1996),

automatic MRD query translation leads to a drop in effectiveness of 40-60% below that of mono-lingual retrieval. This is because of three primarily reasons. First, specialized vocabulary not contained in the dictionary will not be translated. Second, dictionary translations are inherently ambiguous and add extraneous terms to the query. Third, failure to translate multi-term concepts as phrases reduces effectiveness (Zens, 2002)also pointed out that one major disadvantage of the Single-Word Based (SWB) approach is that contextual information is not taken into account (Kula, 2008).

### 2.5.6. Afaan Oromo-Arabic Cross Language Information Retrieval

From the related works has been done so far mainly focal point on both local and global languages; this paragraph will introduce this research work. As we have seen some experimental efforts, if not so much when compared with international languages, have been made to develop monolingual and cross language information retrieval systems for these languages. So far Amharic-English, Oromo-English, Amharic-French, Arabic-English and Afaan Oromo-Arabic are among the local and global languages for which CLIR has been developed for research reason. To wrappings up of the results of the works done so far are reported as encouraging. But, language barriers may exist within two or more languages too and to the knowledge of the researcher there is no CLIR system developed for two languages so far. So, this work is intended to develop Afaan Oromo-Arabic CLIR system. As a final point, this work can be said as an original and first ever that have focused on breaking the language barrier for two languages (i.e. Afaan Oromo & Arabic), in fact Afaan Oromo also the most widely spoken languages in Africa alike to Arabic language. The table 2.5 below shows as result obtained from the system was good maximum average precision of 0.617 and 0.7 for Afaan Oromo and Arabic respectively.

| No | Author | Year | Language used | The Approach used | The result |
|---|---|---|---|---|---|
| 1 | Mohammed Aljlayl | 2002 | Arabic-English | A Machine Translation Approach | 58.4%, 57.1%, and 53.4% for title, description, and narrative |
| 2 | Frieder | 2001 | Effective Arabic-English | Machine-Readable Dictionaries and | 64.6% and the narrative |

| | | | | | |
|---|---|---|---|---|---|
| | | | | Machine Translation | field yields 61.9% |
| 3 | Aynalem Tesfaye | 2010 | Amharic – English | A Corpus Based Approach | 0.24 and 0.33 |
| 4 | Daniel Bekele | 2011 | Afaan Oromo– English | A Corpus Based Approach | 0.468 and 0.316 |
| 5 | Kula Kekeba | 2008 | Oromo-English | Dictionary Based Approach | 24.50% and 25.72% |

**Table2. 5 Summary of literature review in short.**

# Chapter Three

# Methods and Technique

## 3.1. Introduction

Information Retrieval system's (IRS) can retrieve exceedingly relevant documents from large collections such as WWW (Talvensaari, T. Juhola, M, 2007). These multilingual documents that are found on the internet those initiate the researcher in CLIR to break the language obstruction that exists now a day. Therefore, the growing numbers of documents in different languages are one strong initiation the researcher to CLIR. In CLIR, either query or documents are translated. This study attend on accuracy of the query translation in view of the fact, document translation is computationally expensive (Hull, 1996). In this study Afaan Oromo is referred as source language and Arabic referred as the targeted language. CLIR can support user who can read a foreign language but cannot write well a query in that language and also those users who have access to translation source. Dictionary-based CLIR is use in various statistical methods for multilingual text collection (Talvensaari, T. Juhola, M, 2007). It is one of the query translation approaches of CLIR corpus (Kishida, 2005) to set up a link between query and the documents. Nevertheless, (Talvensaari T. , 2008). This research, therefore, uses documents of Afaan Oromo and Arabic to study the application of dictionary-based query translation approach of CLIR. In this section the architecture of the proposed system, Data collection, Data pre-processing, Data preparation, the manual word alignment and the IR model used for retrieval purpose and the evaluation models used have been discussed.

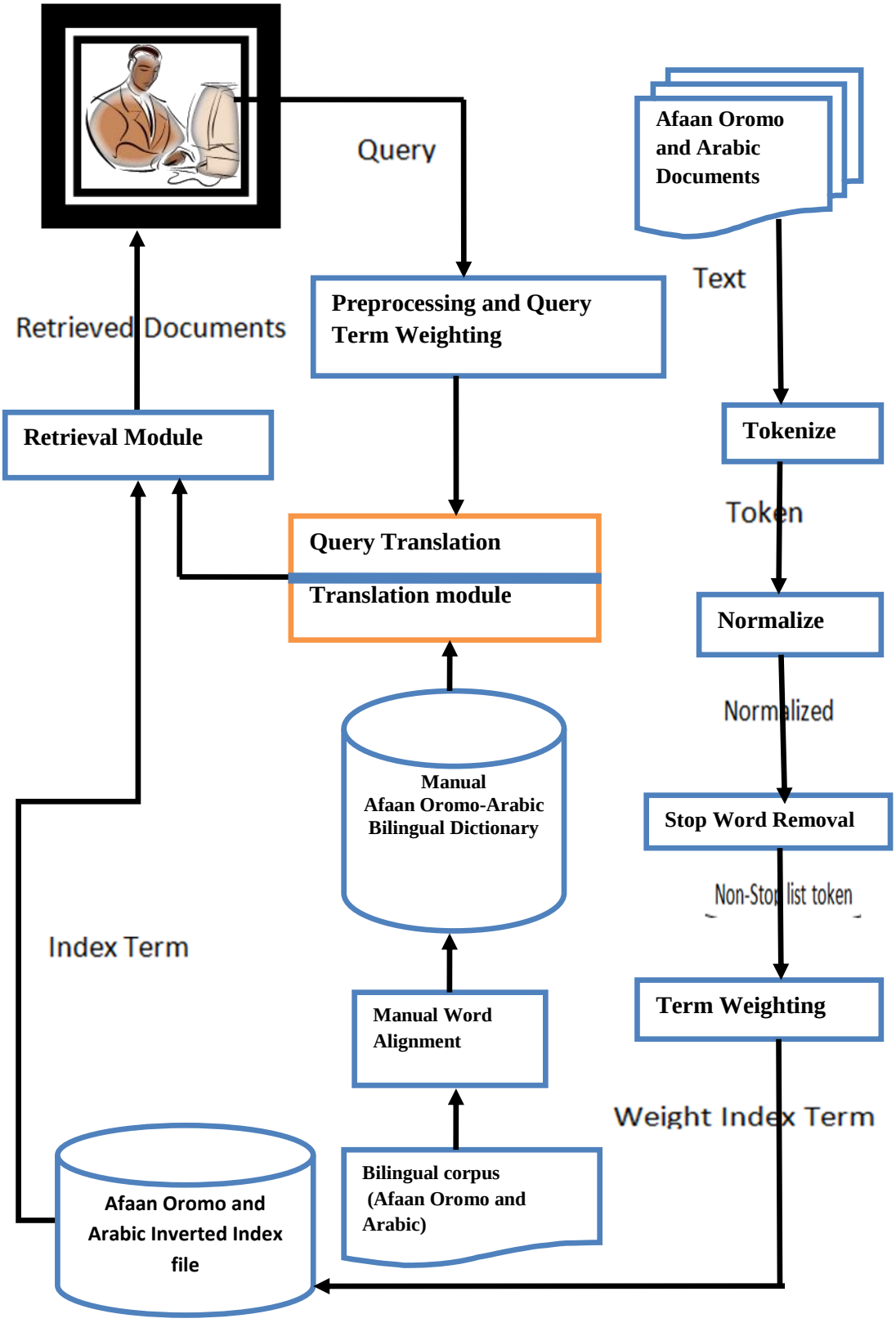
Query
Afaan Oromo and Arabic Documents
Text
Retrieved Documents
Preprocessing and Query Term Weighting
Tokenize
Retrieval Module
Token
Query Translation
Translation module
Normalize
Normalized
Manual Afaan Oromo-Arabic Bilingual Dictionary
Stop Word Removal
Non-Stop list token
Index Term
Manual Word Alignment
Term Weighting
Weight Index Term
Bilingual corpus (Afaan Oromo and Arabic)
Afaan Oromo and Arabic Inverted Index file

**Figure3. 1 Afaan Oromo-Arabic Cross Language Information**

Given Afaan Oromo-Arabic Cross Language Information Retrieve system, the CLIR system organize them using index file to enhance searching. The first step is tokenization of the text words to identify terms. This is followed by normalization in order to bring together similar word written with different punctuation marks. The normalized token is checked as it is not stop word. Content bearing terms (non-stop words) are remaining. For all terms tokens its respected weight calculated and then inverted index file is constructed. On the searching similar text pre-processing (tokenization, normalization and stop word removal) technique is followed as it was done in the indexing part. Then similarity is measurement a technique (cosine similarity) is used to retrieve and rank relevant documents.

## 3.2. Data Collection

Dictionary-based approach requires the accessibility of corpus (Kishida, 2005). For this study, corpus of Afaan Oromo and Arabic documents are used. However, it is difficult to get corpus for diversity of domains and with good quality; it is good source for the translation knowledge (Talvensaari T. , 2008). Normally, the more data is used to guess the parameters of the translation model, the more it can approximate the true translation, which will obviously lead to a higher translation performance; however, such collections are difficult to get. Therefore, for conducting this study, corpus written in Afaan_Oromo and Arabic were collected from Qur'ans and Bible chapters and verses available in both (Afaan Oromo and Arabic) languages and also other religious documents are included.

## 3.3. Data Pre-processing

Data preprocessing is the preparation of raw corpus collected in suitable layout of document, as it is required for more processing. Data preparation, tokenization and normalization and stop word removal are some of text operation. This study is based on corpus; an alignment tool named MatlabR2012b is used for building the bilingual dictionary. The corpus must also be transformed into MatlabR2012b file format since it does not use the corpus collected as it is.

### 3.3.1. Data Preparation

To be appropriate study, the collected document need be preprocessing. The documents that are collected for this research need preprocessing tasks to be appropriate enough for this research. Most of the corpora collected are in portable document format (PDF) and thus needs to be converted into text format so as to make them appropriate for the alignment tool. MatlabR2012b is an alignment tool which is based on heuristics information of a given texts. Having large size corpus will result in better alignment. To prepare the collected corpora for training, it must be word aligned and empty lines must be removed. After the preprocessing tasks were completed, merging all the collected corpora into two, Afaan Oromo and Arabic, text documents were done manually.

### 3.3.2. Tokenization

The process of chopping down the character stream into token for potential index terms is called tokenization (Baeza R and Ribeiro, 1999). Certain characters such as punctuation mark, numbers and symbols are detached in the text operation (tokenization). Grammatical requirements of language are usually satisfy by using punctuation marks. MatlabR2012b code can be handled a tokenization process by using white space as separator. But, sometimes punctuation marks are attached immediately after a word in which case they will be treated as having different meaning by the alignment tool. Thus, removal of punctuation mark is also part of the tokenization. However, removal of some punctuation marks sometimes will alter the intended meaning of the word. . For example, the " ‘ "(Apostrophe) mark is used as a glitch sound (called "*hudhaa"*) lying between three consecutive similar vowels as in *"qe'ee"* which means place (مكا ن) or between two consecutive different vowels as in "ta'e" which means sit(جلس) in Afaan Oromo rather than being an apostrophe marker as in Arabic language not used. Since removing the apostrophe marker will change the meaning of word in Afaan_Oromo, removing is not allowed. Therefore, the collected corpus will be tokenized by chopping the text into words and removing the punctuation marks except " ' " So an exception list is created for such kinds of punctuation marker and task is accomplished using a MatlabR2012b script.

### 3.3.3. Case Normalization

Case normalization is simply converting the texts in the corpus and query into the same case for preserving meaning. This is because most of the time words may vary in their case regardless of their meaning and also, different users type their queries in different cases. For example, the Afaan Oromo word “laga” to mean river (الأَنْهَار) it is written as “Laga” at the beginning and, it is written as “laga” at the middle of a sentence while it has the same meaning. Normalization is one method for handling such differences. Finally, all the texts in the Afaan Oromo corpus will be converted into the most widely used case, lower case, except for the words in the exception list.

Arabic alphabets have no case variation rather some of the Arabic letters have different letters which looks the same sound but different sound for the native speaker. For example: the letters “أ and ع ”; “ح, and ه” ; “خ, and ك” ; “د, and ض” which like the same sound for none Arabic speakers but have different in sound, written and also in meaning. in addition; the letters “ث, س and ص ” and “ز,ذ and ظ” which like the same sound for none Arabic speakers but have different in sound, written and in meaning. So, for the Arabic corpus normalization to the most commonly used format is done for those letters and some common words. The task of case normalization was done by using MatlabR2012b for each of Afaan Oromo and Arabic documents.

## 3.5. Manual Word Alignment

The correspondence between a manual word alignment in a source language and their translations in a target language are represents a pair words (Brown, P. F et al, 1993).In this study, manual word alignment represents the mapping between a source language Afaan Oromo and a target language Arabic. At the present time, manual word aligned bilingual corpus being used as a significant source of knowledge. For the first time manual word alignment model was introduced in SMT by (Brown, P. F et al, 1993). MatlamR2012b uses a manual alignment which computes a translation similarity between the types of the two languages for each co-occurring word pair. A given word from the source language may appear as being aligned with only one translations candidates of target words for each one with a given similarity value. The source words and their

corresponding translations of the target word were constructed and stores in bilingual dictionary. By using these manually the natural language representation of the corpus will be converted to the suitable format. The most mandatory and minimum requirement input files for manual word alignment are the vocabulary file.

### 3.5.1 Vocabulary file

The vocabulary file input holds words /tokens from the corpus along with their frequency in the whole corpus. Also the vocabulary file contains a unique identifier for each of the words. The frequency value is used in similarity of aligning the word against its equivalent word in the other language while the Unique ID is used to identify the word uniquely after manual alignment since the alignment is done by their ID. Therefore both the Afaan Oromo and Arabic documents need to be converted into vocabulary file format before being aligned. The sample of Arabic and Afaan Oromo vocabulary file has been shown in Figure 3.2 and 3.3 respectively.

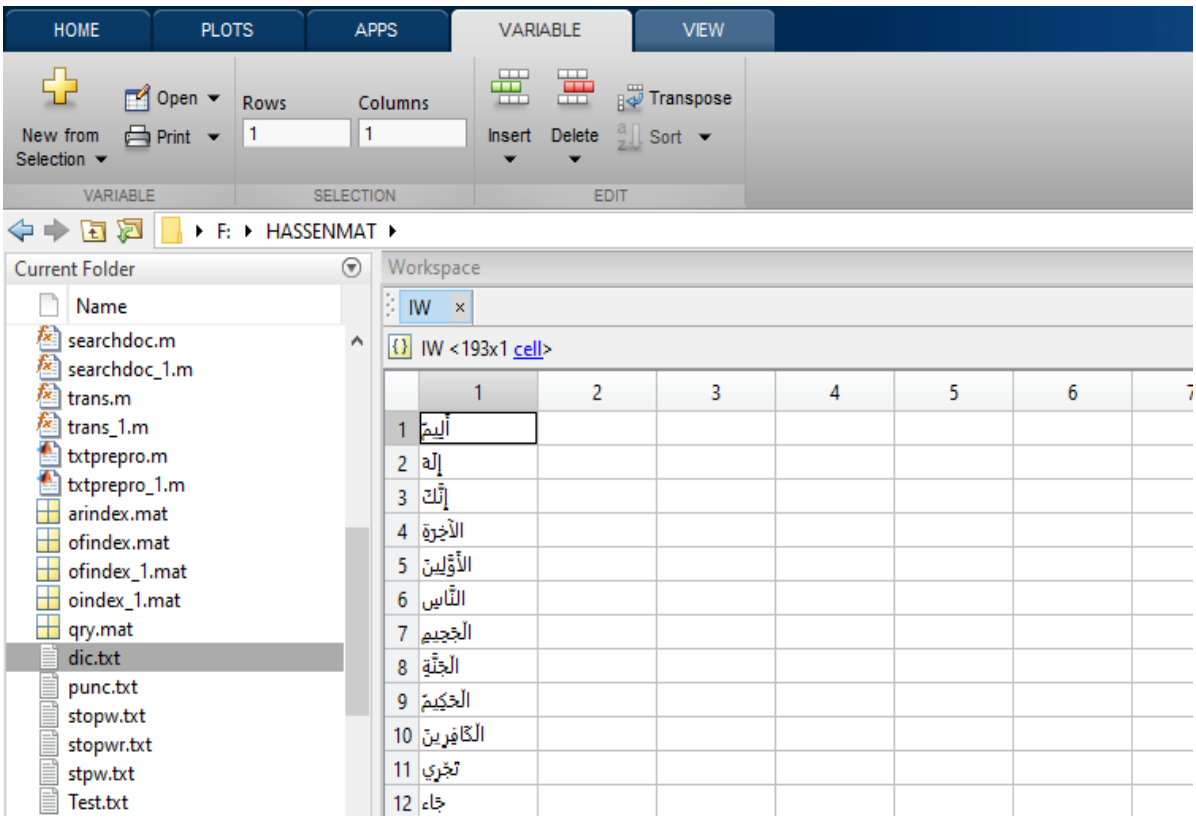


**Figure3. 2 Sample Arabic vocabulary file**

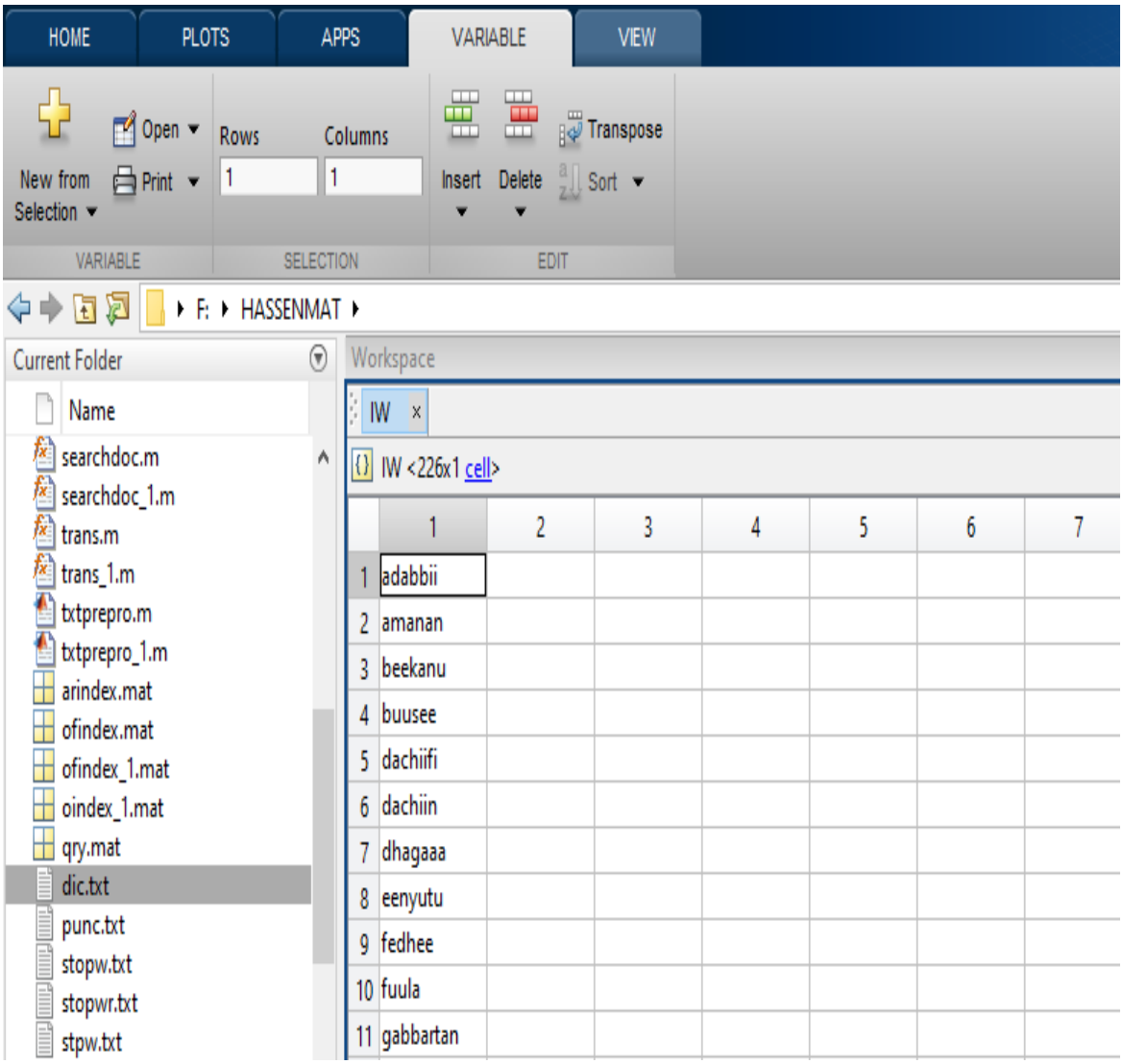


**Figure3. 3 Afaan Oromo vocabulary file**

### 3.5.2. Manual Alignment Models

There are two general approaches to calculating manual word alignments: heuristics and statistical alignment models. Heuristic Models: Considerably simpler methods for obtaining word alignments use a function of the similarity between the types of the two languages. Frequently, variations of the Dice coefficient are used as this similarity function. For each words pair, a matrix including the association scores between every word at every position is then obtained:

$$\mathbf{dic(i,j)} = \frac{\mathbf{2}.C(\mathbf{e_i f_j})}{C(\mathbf{e_i}).C(\mathbf{f_j})} \text{..... (3.1)}$$

$C(e, f)$denotes the co-occurrence count e and f in training corpus. $C(\mathbf{e})$and$C(f)$denote the count of f in the source words and respectively. From this association score matrix, the word alignment is then obtained by applying appropriate heuristics. One method is to alignment

**$a_j$ = argmax {dice (i, j)}…… (3.2)**

Statistical Models: In statistical machine translation, we try to model the translation probability **Pr ($f_1^J$| $e_1^I$),** which describes the relationship between a source language string $f_1^J$and a target language string $e_1^I$ In (statistical) alignment models $Pr\ (f_1^J\ ,\ a_1^J|\ e_1^I)$**,a** "hidden "alignment $a_1^J$is introduced that describes a mapping from a source position **j** to a target position $a_j$. The relationship between the translation model and the alignment model is given by as below:

$$Pr\ (f_1^J|\ e_1^I) = \sum_{a_1^J} Pr\ (f_1^J\ , a_1^J|\ e_1^I) \text{…….} \quad (3.3)$$

The alignment $a_1^J$may contain alignment $a_j = \mathbf{0}$with the empty $e_0$to account for source words that are not alignment with any target words. Therefore, in this research the heuristics model for word alignment has been chosen because it problem solving tools.

## 3.6. Bilingual Dictionary

The translation of Afaan Oromo queries into Arabic was based on the Afaan Oromo-Arabic bilingual dictionary which has been constructed manually from the Afaan Oromo Arabic corpora collected. The bilingual dictionary that is constructed stores both source words and their corresponding translation of the target words. The word alignment in the dictionary contains all probable translation of a word from the source text into a target word together with its similarity of alignment. This similarity value assigned for each possible translation of a word shows the degree to which Afaan Oromo word is most likely translated into its equivalent Arabic word. The highest the similarity value indicates the best translation among the candidates translations exist. The bilingual dictionary was, therefore, constructed by the help of this similarity value assigned to each translation. MatlabR2012b was developed to select the one that has the highest similarity

of alignment. The sample of Afaan Oromo-Arabic dictionary built has been shown in Figure 3.4 as following.

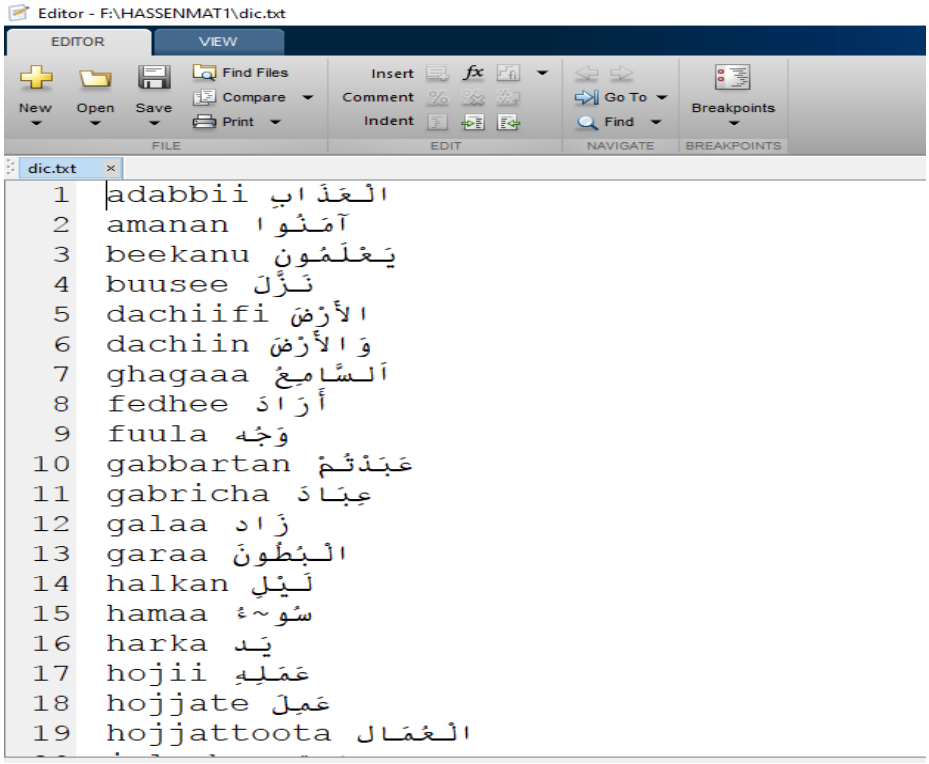


**Figure3. 4 Sample Afaan Oromo-Arabic bilingual dictionary constructed**

## 3.7. Translation

The main objective of cross language information retrieval is to retrieve the document rather than query language. In cross language information retrieval first, the queries want to be translated into the targeted language by using the translation knowledge source before targeting for the search purpose. In this study Afaan Oromo queries are used to retrieve Arabic documents besides Afaan Oromo documents. Therefore, the task of this module is to accept Afaan Oromo queries from the user and translate it to its equivalent Arabic query to retrieve Arabic documents. Translation of the query was done on a word-by-word basis for this study. MatlabR2012b was developed to accept Afaan Oromo queries and translate them into their Arabic equivalent query by using the Afaan Oromo Arabic bilingual dictionary constructed at earlier stage. Sample code of the translation is show in an Algorithm 3.1

```
%% Transilate if the quary language is Oromiifaa ('OF') into
output language Arebic ('AR') and vice versa
if IL=='AR'
    for i=1:numel(OL)
        for n=1:numel(lexi(:,1))
            if strcmpi(qry(i),lexi(n,2))
                OL(i)=lexi(n,1);
            end
        end
    end
elseif IL=='OF'
    for i=1:numel(OL)
        for n=1:numel(lexi(:,1))
            if strcmpi(qry(i),lexi(n,1))
                OL(i)=lexi(n,2);
            end
        end
    end
else
    disp(strcat(IL, ' is unknoun language! Please use ''OF''
for Afaan Oromo or ''AR'' for Arabic.'));
    return;
end
```

Continuo…….

**Algorithm3. 1 Sample code of the translation**

## 3.8. Retrieval

The document retrieval is tasked for both monolingual and bilingual case. The retrieval module is answerable for taking the query in the source (Afaan Oromo) language and retrieving the relevant documents from the target (Arabic) document collections. Using bilingual dictionary Afaan Oromo query translation changed into Arabic query. On the other hand, the same query will be given directly to the retrieve module to retrieve Afaan Oromo documents and this is known as monolingual run using base line query. Afaan Oromo documents are retrieved from monolingual retrieval by using base line query of Afaan Oromo without being sent to the translation module. Information is concerned with information users of IR system. The process of predicting relevant document from the whole corpus is a significant issue in information retrieval. An IR system applies a retrieval model that comprises of matching algorithm and internal representation of queries and the documents. To compare and measure the relevance of the document and

queries representation matching specification describes the integration of the document to the queries. Therefore, to differentiate between the relevant document and non-relevant document is based on the particular model adopted. Among the three major information retrieval models, Vector Space Model (VSM) has been implemented.

The vector space model includes three main parts. The first part is indexing of the document through content fashion terms represents the document. The second part is weighting the indexed terms to increase retrieval of the relevant documents from the corpus. The final step is ranking the document to indicate paramount matching with respect to the delivered query by the client. The vector space model was chosen for the following reasons. (1) it's term-weighting scheme improves retrieval performance; (2) its partial matching strategy allows retrieval of documents that approximate the query conditions; and (3) its cosine ranking formula sorts the documents according to their degree of similarity to the query (Baeza R and Ribeiro, 1999).

Normally, this module is accountable for representation of documents and retrieval of relevant documents based on user query. Mostly, the module involves of two major functions to handle this study; the indexing and searching which are define further down.

## 3.9. Indexing

The task of IR system is to process a request from the user for information and retrieve documents to fulfill the satisfaction and the information need of the user. According to (Salton, 1986)of the processes in information retrieval, document representation is the most vital task. Because of social media the volume of the text information is increase from time to time and stored in electronics media. In consequence, searching for full text is difficult for controlling and time consuming. One technique to overwhelmed representation problem coming with the ever increasing volume of text in electronic format is indexing. The content description are selected and assigned to documents to provide short-form description of the document using terms or keywords. The process of analyzing text and deriving short-form of description to sum up the message of documents is called indexing. Thus, the purpose of storing an index is to enhance speed and performance in finding relevant documents for a search query.

### 3.9.1. Index term selection

Using a MatlabR2012b code index terms are automatically generated from the collected corpus. From different representation on index terms, inverted file index term representation technique is used in this study. An inverted file is a data structure for efficiently indexing texts by their tokens. An inverted file contains of list of tokens where each token is followed by the identifier of every document that contains the word along with their number of existences in the document is represented. With this information inverted file allows an IR system to rapidly determine which documents hold a given set of words, and how often each word looks like in the document.

An IR system is organizes both Afaan Oromo and Arabic corpus by using index file to enhance searching. Tokenization of the corpus is the first step in the indexing in order to identify stream of token and stop words. A set of manually developed stop word lists have been used both for Afaan Oromo and Arabic languages. In this study Afaan Oromo stop words identified by Gezahegn Gutema (Eggi, 2012).and for Arabic language is stop words identified by El-Khair, Ibrahim Abu (El-Khair, 2006).

### 3.9.2. Index term weighting

From indexing tasks, term weighting is allocated to every single item term numeric values in order to determine the document representation importance. Term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. Furthermore, the consideration of the term weights can support in ranking document by sorting retrieved document in reducing order of their degree of similarity (Manning et al, 2008). There are numerous techniques of term weighting, for this specific study the tf*idf mechanism has been implemented. Based on term frequency, the tf*idf weighting technique is the number of occurrence of a term in a document and the inverted document frequency of term. The last weight of a term which is made of the term frequency and the inverse document frequency is calculated as follows (Baeza R and Ribeiro, 1999).

Let $frq_{ij}$ be the raw frequency of the term $k_i$ in a document $d_j$ (i.e the number of the term $k_i$ is declared in the text of document $d_j$). Though, raw term frequency suffers from a serious

problem- all terms are measured equally important when it comes to evaluating relevancy on a query. This activates a need to normalize term frequency. Then, the normalized frequency $f_{i,j}$ of term $k_i$ in document dj is given by:

$$f_{ij} = \frac{frq_{ij}}{max_i frq_{ij}} \text{.....} \quad (3.4)$$

From the text of the document $d_j$ the maximum is calculated over all the terms. If the term $k_i$ does not occurred in the document $d_j$ at that point $f_{i,j} = 0$.

Where;

i: a term

j: a document

tf: a frequency of a term i in documents

Let $d_f$ (document frequency) denote the number of documents in which the term $k_i$ performs. This will allocate higher value to terms that perform in many documents and lower value for terms that perform merely in few documents. Nevertheless, terms that perform in many different documents are less indicative of overall topic. At that point, we will use $idf_i$, (inversed document frequency) to compute the weight of a term and use logarithm to reduce the effect of the idf as follows.

$$idf_i = \log\left(\frac{N}{df_i}\right) \text{.....} \quad (3.5)$$

Where;

$idf_i$: the inverse document frequency of term i.

$df_i$: the document frequency of term i (total number of documents containing term i)

As a result the idf of an occasional term is performing high; whereas, the idf of frequent term is performing low. Currently the combination of both the term frequency and the inverse document frequency (equation 3.4 and equation 3.5) would reduce composite weight for every term in every document.

$$W_{i,j} = f_{i,j} * \log\left(\frac{N}{df_i}\right) \text{.....} \quad (3.6)$$

In this work the tf*idf technique is selected because of it a uses a composite result on the terms. The tf*idf weighting method denoted to term (t) a weight in document (d) that is:

Highest when term happens many times in small number of document (thus lending high discriminating power to those documents);

Lower when the term happens fewer times in a document (thus offering a less pronounced relevance signal);

Lowest when the term happens in virtually all document.

Where;

$w_{i,j}$: weight of term i in documents.

N: total number of documents.

## 3.10. Searching

Searching is a significant task in information retrieval module. In cross language information retrieval, search results are provided for the user in both source language and target language, in this case Afaan Oromo and Arabic respectively. The retrieval of Afaan Oromo documents is known as baseline run in which the Original queries of the user were directly used to retrieve Afaan Oromo documents whereas, the retrieval of Arabic documents is known as cross lingual run in which the queries that were run for Afaan Oromo document retrieval were translated by the translation module to retrieve Arabic documents. There are different procedures of measuring similarity between queries and every single document. Meanwhile this research is based on Vector Space Model of Information Retrieval, cosine similarity measure between query and document is used.

### 3.10.1 Cosine Similarity measure

The vector model denotes a non-binary weight to index terms both for user query and documents. Term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. Furthermore, the consideration of the term weights can support in ranking document by sorting retrieved document in reducing order of their degree of similarity, the vector space model takes into consideration documents which match the query terms only partially. The weight $W_{i,j}$ associated with a pair ($k_i$ , $d_j$ ) is positive and non-binary; and the vector for a document $d_j$ is represented by $d_j$ =($w_{1,j}$, $w_{2,j}$,…. $w_{t,j}$). Additional, the index terms in the query are also weighted. Let $W_{i,q}$ be the weight associated with the pair [$k_i$ ,$_q$], where $w_{i,q}$ >=0.

Then, the query vector is defined as q = ($w_{1,q}$, $w_{2,q}$,⋯, $w_{t,q}$) where (t) is the total number of index terms in the system. Thus, a document $d_j$ and a user query q are represented as t-dimensional vectors. In vector space model the degree of similarity of the document $d_j$ with regard to the query q is represented as the correlation between the vectors d and q. This correlation can be quantified by the cosine of the angle between these two vectors as follows (Manning et al, 2008) (Baeza R and Ribeiro, 1999).

Un-normalized Similarity

$$\mathbf{sim(d_i, q)} = \sum_{j=1}^{n} w_{q_j} * w_{d_{ij}} \text{.....} \quad \textbf{(3.7)}$$

Cosine Similarity is normalized

$$sim(d_i, q) = \frac{\vec{d_i}.\vec{q}}{|\vec{d_i}||\vec{q}|} = \frac{\sum_{j=1}^{n} w_{q_j} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{n} (w_{q_j})^2} * \sqrt{\sum_{j=1}^{n} (w_{d_{ij}})^2}} \text{.....} \quad (3.8)$$

### 3.10.2. Ranking

The relevant documents retrieved are given to the user as ranked document depend on their results from cosine similarity score. The document which is more similar with the query is ranked first and the document which is least similar with the query will be ranked last. For deciding the number of retrieved document, a threshold value is decided based on cosine similarity value of queries and human judgment after repeated evaluation of relevance of retrieved document. The sample code of the ranking is shown in an Algorithm 3.2

```
% trshold = cos(pi*30/180);
doc = cell(ND,1);
if indx == 'ofindex'
for i=1:ND
if wgt(i)>0
if i < 10
            tmp = strcat('00',int2str(i));
elseif i<100
            tmp = strcat('0',int2str(i));
else
            tmp = int2str(i);
end
         doc(i) = {strcat('doc',tmp,'.txt')};
else
         doc(i)={'NONE'};
end
end
end
if indx == 'arindex'
for i=1:ND
Continue ……
```

**Algorithm3. 2 Sample code of the ranking and threshold**

## 3.11. Performance Evaluation Model

Concept of ‘relevance’ Relevance is the degree of correspondence between retrieved documents and information need of users. Since individuals need varying relevance is subjective. Often peoples look relevance differently. What is relevant to somebody might be irrelevant to others. Nature of user query and document collection effects relevance. Additionally relevance depends on individual personal needs, performance subject, level of knowledge, specialization language. Even if relevance is subjective concept there is no possibility of ignoring it. One of the approaches to deal with subjectivity is by generating ‘user profiles’ user profile includes knowledge about user’s need, performance etc. this helps the IR system to give what ‘mean’ not what they asked for. An IR system aim is to give users access to items that provide information that is relevant to the users information need expressed as query. In fact the aim of an IR system is to retrieve a relevant document depends on the user query, even if retrieving as few none relevant document as possible (Baeza R and Ribeiro, 1999). In fact, a system designed for providing information retrieval, other metrics, besides space and times, are also of interest. In fact, since the user query request is inherently uncertain, the retrieved

documents are not directly response to the query, the ranking could be ranked according to the relevance of the query(information need). Such relevance ranking introduces a part which is not present in data retrieval systems and which plays a central role in information retrieval. Thus, information retrieval systems require the evaluation of how precise the reply set is (Manning et al, 2008). This type of evaluation is referred to as retrieval performance evaluation. Such an evaluation is usually based on a test reference collection and on an evaluation measure. The test reference collection consists of a collection of documents, a set of example information requests, and a set of relevant documents for each example information request. Given a retrieval strategy, the evaluation measure quantifies (for each example information request) the similarity between the set of relevant documents provided by the specialists and the set of documents retrieved. This provides an estimation of the goodness of the retrieval strategy (Abusalah, M; Tait, , 2005).

Two of the most widely used retrieval performance measures are Recall and Precision. Recall and Precision When a proposed IR algorithm is evaluated, it is applied to either of documents or query preprocessing, document-query matching, or all of these, depending on the algorithm. A baseline algorithm is also applied to the same part of the system. The query performance of each of the tested methods and the baseline is evaluated by matching the query results to the recall base.

Various performance metrics that are usually based on recall and precision are used in the evaluation (Talvensaari T. , 2008). Let R be the set of relevant documents for a test topic, and A the set of retrieved documents for the topic by some proposed algorithm. ***Recall:*** the fraction of the relevant documents that have been retrieved.

$$\boldsymbol{Recall = \frac{R \cap A}{|R|}} \text{..... (3.9)}$$

***Precision***: on the other hand is the fraction of retrieved documents that are relevant.

$$\boldsymbol{Precision = \frac{R \cap A}{|A|}} \text{..... (3.10)}$$

Recall and precision, well-defined above, suppose that, all the documents in the reply set A have been examined. Nevertheless, the user is not habitually presented with all the documents in the reply set A at once. As an alternative, the documents in A are first

sorted according to a degree of relevance. The user then examines this ranked list starting from the top document. In this condition, the recall and precision measures vary as the user proceeds with his examination of the answer set A. Thus, correct evaluation requires plotting a precision versus recall curve based on specialists‟ decision on relevance of a given document for the particular information request" (Abusalah, M; Tait, , 2005).

This technique calculates precision of the algorithm at different standard recall levels for each user query. Usually, however, retrieval algorithms are evaluated by running them for several distinct queries. In this case, for each query a distinct precision versus recall curve is generated. To evaluate the retrieval performance of an algorithm over all test queries, we average the precision figures at each recall level as follows.

$$\boldsymbol{P(r) = \sum_{i=1}^{N_q} Pi(r)/N_q} \quad \textbf{(3.11)}$$

Where p(r) is the average precision at the recall level r, $N_q$ is the number of queries used, and Pi(r)is the precision at recall level r for the $i^{th}$ query. Since the recall levels for each query might be distinct from the Mean Average Precision levels, utilization of an interpolation, which states that the interpolated precision at the $j^{th}$ standard recall level is the maximum known precision at any recall level between the $j^{th}$ recall level and the $(j + l)^{th}$ recall level, is necessary. Precision in competition with recall curve can also be used to compare the retrieval performance of distinct retrieval algorithms. Another retrieval performance measurement approach is to compute average precision at given document cut off values. Average precision in competition with recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza R and Ribeiro, 1999).

In addition to, the above techniques, other single valued retrieval performance measures can also be used to measure the performance of a system for single query. The single valued measures are used in situations in which we would like to compare the retrieval performance of our retrieval algorithms for the individual queries. The single valued measures include the harmonic mean (F-measure), E-measure, average precision at seen relevant documents and r-precision.

**Harmonic Mean** The harmonic mean combines the recall and precision values to a single value. The harmonic mean is calculated as,

$$\mathbf{F(j) = \frac{2}{\frac{1}{\alpha(pj)}+\frac{1}{(1-\alpha)(rj)}} = \frac{(\beta^2+1)pr(j)}{\beta^2 p(j)+r(j)} = \frac{2(pr(j))}{p(j)+r(j)}} \quad \textbf{(3.12)}$$

where r(j)is the recall for the j[th] document in the ranking, p(j)is the precision for the j[th] document in the ranking, and F(j)is the harmonic mean of r(j)and p(j)(thus, relative to the j[th] document in the ranking) (Talvensaari T. , 2008)..

The function F (j) assumes values in the interval [1, 0] and it is 1when all ranked documents are relevant and is 0 when no relevant documents have been retrieved. Further, the harmonic mean F assumes a high value only when both recall and precision are high. Therefore, determination of the maximum value for F can be interpreted as an attempt to find the best possible compromise between recall and precision (Manning et al, 2008).

**E-measure** The E-measure is another evaluating mechanism which combines recall and precision by allowing the user to identify whether he/she is interested in recall or precision. It is calculated as,

$$\boldsymbol{E(j) = 1 - \frac{1+b^2}{\frac{1}{p(j)}+\frac{b^2}{r(j)}}} \quad \textbf{(3.13)}$$

Where p(j) is the precision for the j[th] document in the ranking, r(j)is the recall for the j[th] document in the ranking, E(j)is the evaluation measure relative to p(j)and r(j),and b is a user specified parameter, which reflects the relative importance of recall and precision.

for b = 1, the E(j) measure works as the complement of the harmonic mean F(j).

Values of b greater than 1 indicate that the user is more paying attention in precision than in recall. While values of b smaller than 1 indicate that the user is more paying attention in recall than in precision (Talvensaari T. , 2008).

**Average Precision at seen relevant documents:** The idea here is to produce a single value outline of the ranking by averaging the precision figures gained after each new relevant document is observed in the ranking. This measure favors systems which retrieve relevant documents quickly.

**R-precision** This measurement technique generates a single value summary of the ranking by computing the precision at the R-th position in the ranking. Where R is the total number of relevant document for the current query (i.e., number of documents in the set $R_q$). The R-precision measure is a useful parameter for watching the behavior of an algorithm for each individual query in experimentation. For this reason, in this study the retrieval effectiveness of the system is evaluated using Mean Average Precision recall curve

# Chapter Four

## Experimentation and Analysis

## 4.1 Introduction

As pointed out in chapter one, the general objective of this study is to experiment on the possibility of designing and developing a Dictionary based Afaan Oromo Arabic Cross Language Information Retrieval System. Whether the proposed objective is attained or not ought to be experimentally tested before concluding the possibility of developing such a system. And also the performance of the system ought to be tested vis-à-vis the experimentation environment and used data. Furthermore, it discusses how to select sample test documents and how queries are prepared for the experimentation. As a final point, the result of the experimentation (finding of the study) pursued a brief analysis of the result of the experimentation is also discussed.

## 4.2. Document and Query Selection for Experimentation

### 4.2.1. Document Selection

All the documents that were collected are not used for doing the experimentation. The reason is that, it is unachievable test for large sample size with the available commotional resources. On the other hand, all documents used have been used in the construction of Afaan Oromo Arabic dictionary based approach. Among the collected total data, 40 Afaan Oromo Arabic documents were selected randomly for the purpose of experimentation. Random sampling was employed for selecting test documents because all documents that were collected are equally essential for experimentation. Moreover, random sampling also avoids the unfairness that may occur in the selection is of test documents. As the experimentation is for cross language information retrieval in which ought to be conducted for both language so, in the documents selection process first 40 Afaan Oromo documents were selected randomly and their corresponding Arabic documents were added.

### 4.2.2. Query Selection

Stand on the selected Afaan Oromo corpus for testing at previous phase, an appropriate queries were organized with the help of language professionals. Then, these queries will be used to retrieve both Afaan Oromo and Arabic documents. Consequently, for the purpose of testing 20 queries were prepared in Afaan Oromo to retrieve relevant documents out of 40 test documents selected previous. These queries will be used with different techniques to measure performance and choose better combination. The combination of the queries with different translation techniques will be divided the testing process into different experimental setup.

## 4.3. Experimentation and Evaluation of the System

### 4.3.1. Experimentation

The two basic phases for each query are the monolingual and bilingual runs. In monolingual run the original Afaan Oromo queries will be to retrieve Afaan Oromo documents. While the bilingual runs the second phase of experiment to retrieve Arabic documents by using the previous Afaan Oromo queries and the bilingual dictionary too.

**Experimentation phase one**

In this step of the experiment the original queries developed will be used to retrieve Afaan Oromo documents in the test corpus. This task can be shown diagrammatically as in the figure 4.1

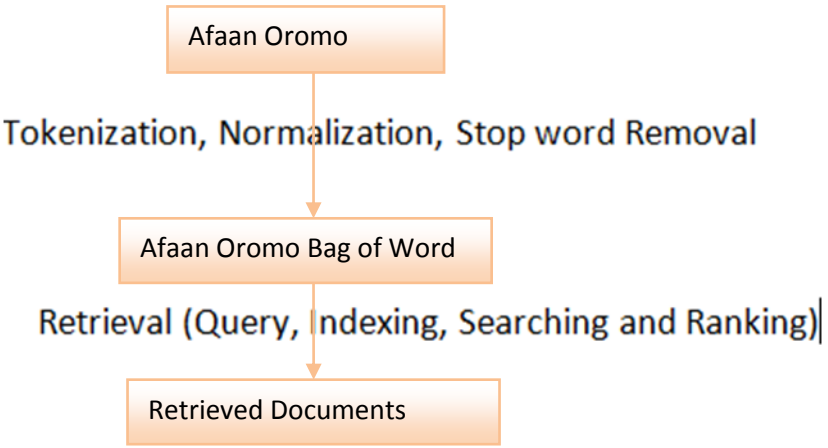

**Figure4. 1 Flowcharts showing the monolingual run.**

**Experimentation phase two**

In this stapes, first Afaan Oromo queries were translated into Arabic representation word by word and then the translated query terms will be used to retrieve Arabic documents. The experiment is performing by using one to one translation to a query term based on it is similarity value. The flowchart of this set up is shown as below in Figure 4.2

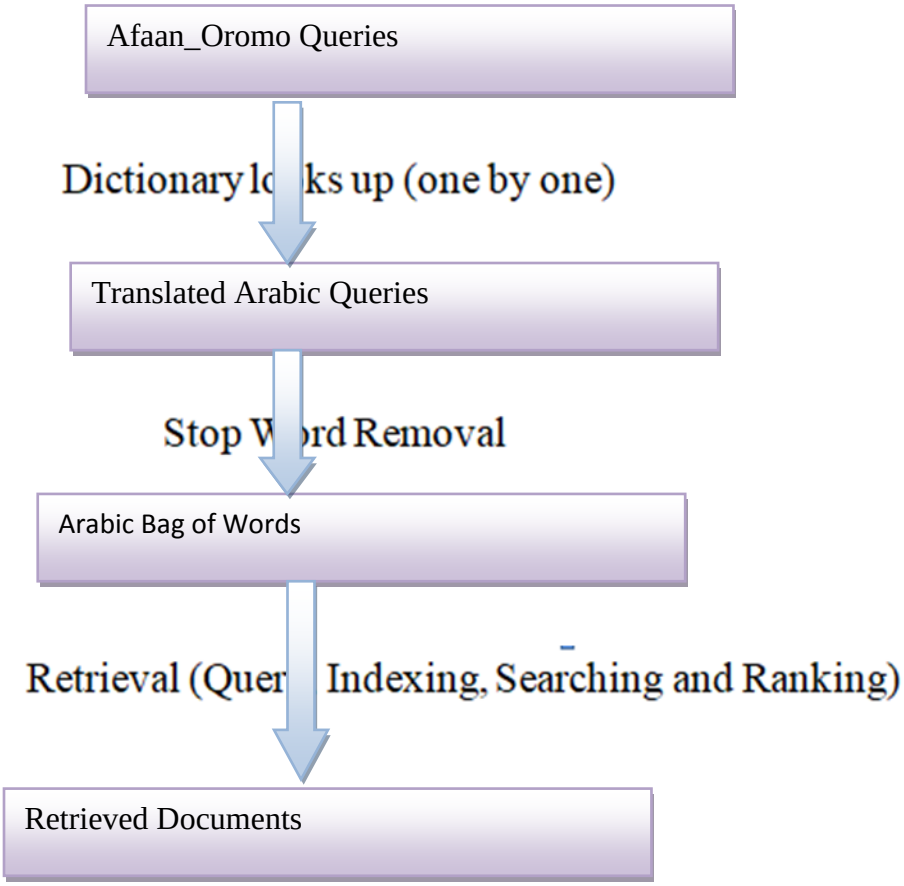


**Figure4. 2 Flowcharts showing the bilingual run with one to one translation**

## 4.4. System Evaluation

Performance of a system should be evaluated based on experiment results of the system. The system is evaluated for two different runs, monolingual and bilingual, in each experiment. In the monolingual run the base line Afaan Oromo queries will be evaluated for retrieving Afaan Oromo documents. In the bilingual run of the system, the same

queries used above will be used to retrieve Arabic documents after being translated to their equivalent Arabic queries by the bilingual dictionary built earlier. There are different techniques of evaluation of performance of a system based its aim. In fact, the primary aim of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza R and Ribeiro, 1999). The most widely used retrieval performance evaluation methods are recall and precision which are discussed in section 3.11. However, according to (Baeza R and Ribeiro, 1999), proper evaluation requires plotting a precision versus recall curve based on specialists‟ decision on relevance of a given document for the particular information request (query). Most of the time retrieval algorithms are evaluated by running them for several distinct queries. So, evaluation of retrieval performance of the system over all test queries is done by averaging the precision figures of each query at each recall level by using equation.

$$\boldsymbol{P(r) = \sum_{i=1}^{N_q} Pi(r)/N_q} \text{.....(4.1)}$$

*Where P(r)* is the average precision at the recall level r, *Nq* is the number of queries used, and *Pi(r)* is the precision at recall level r for the ith query. Mean Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza R and Ribeiro, 1999). Hence, in this study the retrieval effectiveness of the system is evaluated using Mean Average Precision versus recall curve.

## 4.4.1. Results

### 4.4.1.1 Retrieval effectiveness results

Since one of the aims of this research work is to experiment on the applicability of cross language information retrieval based on dictionary for two languages, Afaan Oromo and Arabic, evaluating the effectiveness of the retrieval of the system is enough to come up with a conclusion. Thus, in this research the evaluation is done from retrieval effectiveness perspective only. Accordingly, the result of the experiment conducted is discussed as follows.

**Experiment:**

An experiment is conducted with 20 Afaan Oromo queries where by each of the query terms is translated by using the manually built dictionary with a one to one translation to retrieve Arabic documents. This experiment, as stated earlier, can be seen as having two phases. The first phase being the monolingual run of the queries to retrieve Afaan Oromo documents while the second phase retrieves Arabic documents using the Afaan Oromo queries. The queries were given to four language experts who were asked to list the relevant documents (Afaan Oromo and Arabic) from the corpus. This is used as a reference for the evaluation. The results are given below.

**Result of phase one (Monolingual run)**

The system is tested with the 20 queries were given to four language experts who were asked to list the relevant document (Afaan Oromo) from the corpus. This is used as a reference to evaluate and to retrieve all the relevant documents for each of the given queries by leaving the non-relevant ones. The effectiveness of the retrieval of the system is measured using recall and precision for each query. Accordingly, the monolingual run returned an average recall value of 0.675 and average precision of 0.706 respectively. The sample of mean average precision and recall for monolingual, Afaan Oromo-Arabic monolingual ranking and Mean Average Precision has been shown in Table 4.1 and Figure 4.3 and 4.4 respectively.

| | **Query term** | **Precision** | **Recall** |
|---|---|---|---|
| **Q1** | Adabbii gadi buusee | 0.966 | 0.5 |
| **Q2** | ergamtoota fi duniya | 0.785 | 0.7 |
| **Q3** | Dilii adabbii qaba | 0.6237 | 0.7 |
| **Q4** | Ibidii adabbii cimaa dha | 0.551 | 0.7 |
| **Q5** | Hojii harka qaba | 0.816 | 0.8 |
| **Q6** | Jiruu duniyaa | 0.942 | 0.5 |
| **Q7** | Mallattolee rabbi keessa tokko dha | 0.604 | 0.6 |
| **Q8** | Aalama jinni fi duniyaa | 0.689 | 0.7 |
| **Q9** | Ibidii jahannama dhugaa dha | 0.693 | 0.8 |
| **Q10** | Namoota Jahannama hunda | 1 | 05 |

| | | | |
|---|---|---|---|
| **Q11** | gooftaan injifata dha | 0.576 | 0.8 |
| **Q12** | Badii adda addaa | 0.546 | 0.6 |
| **Q13** | Isaan jallina cimaa qabu | 0.708 | 0.7 |
| **Q14** | Ergamtoota gaggaarii ergine | 0.476 | 0.6 |
| **Q15** | kitaaba gaggaarii ergine | 0.645 | 0.8 |
| **Q16** | karaawwan haqaan | 0.723 | 0.8 |
| **Q17** | Allaahn laggeen uume | 0.683 | 0.7 |
| **Q18** | hojii adabbii hin qabne hojjate | 0.551 | 0.7 |
| **Q19** | Jannata fi jahannama sobsiisanii | 0.576 | 0.7 |
| **Q20** | amantii haqaan dhufee | 0.948 | 0.6 |
| **Average** | | **0.706** | **0.675** |

**Table4. 1 shows the mean average precision and recall for monolingual.**

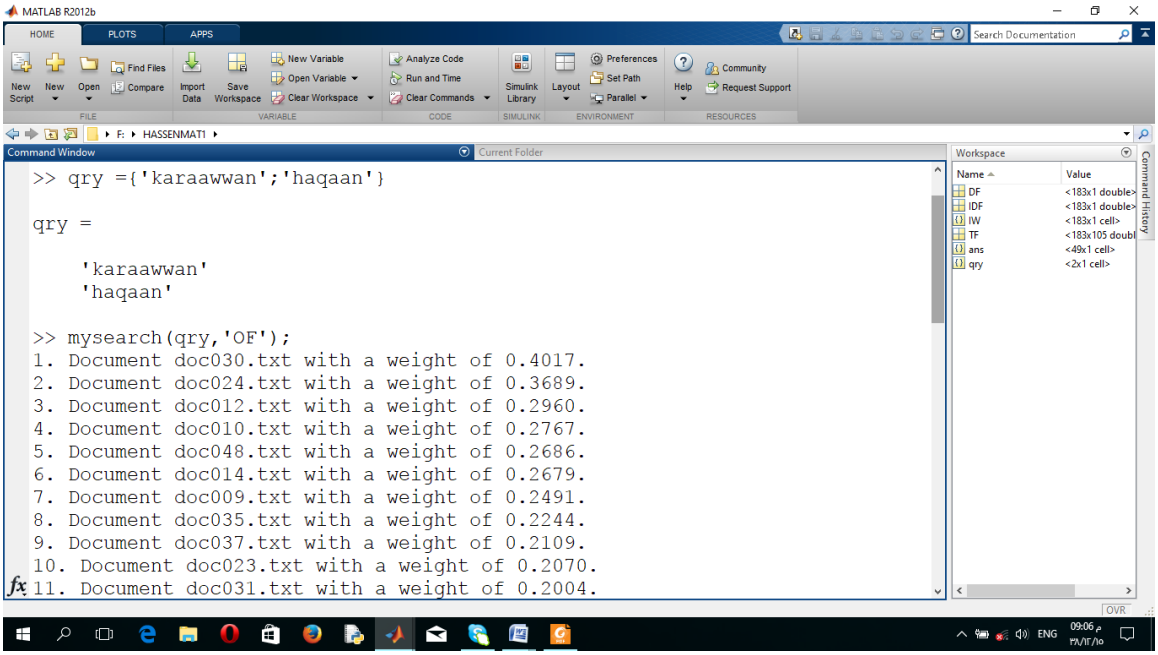


**Figure4. 3 The sample of Afaan Oromo-Arabic monolingual**

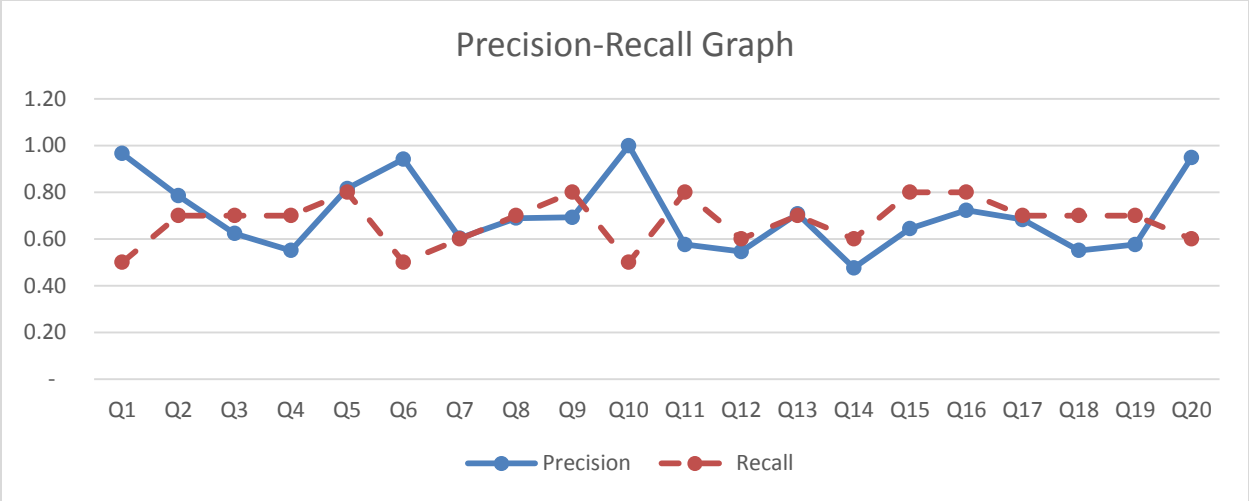


**Figure4. 4 Mean Average Precision for monolingual**

The Precision-Recall graph for the monolingual retrieval experiment (see Figure 4.5) illustrates that the precision and recall score for all the 20 queries shows stable or comparable result. This shows that the system behaviors in stable manner.

**Result of Phase two**

The system is tested with the 20 queries were given to four language experts who were asked to list the relevant documents (Afaan Oromo and Arabic) from the corpus. This is used as a reference to evaluate and to retrieve all the relevant documents for each of the given queries by leaving the non-relevant ones and the result is given in Table 4.2. In this run the queries are allowed to be translated to a single word and the translation word is selected based on similarity value of translation and the term with higher similarity value is taken. The result of this run returned an average recall value of 0.7 and an average precision of 0.617. Since all queries may not exactly have the standard similarity levels, we used Mean Average Precision to calculate the overall performance of the system across queries. The sample of Mean Average precision and recall for bilingual, Mean Average Precision for bilingual and Afaan Oromo-Arabic bilingual ranking has been shown in Table4.2 and Figure 4.5 and 4.6 respectively.

| | Query term | Precision | Recall |
|---|---|---|---|
| Q1 | Adabbii gadi buusee | 0.688 | 0.6 |
| Q2 | ergamtoota fi duniya | 0.617 | 0.7 |

| | | | |
|---|---|---|---|
| Q3 | Dilii adabbii qaba | 0.69 | 0.8 |
| Q4 | Ibidii adabbii cimaa dha | 0.551 | 0.7 |
| Q5 | Hojii harka qaba | 0.765 | 0.8 |
| Q6 | Jiruu duniyaa | 0.733 | 0.8 |
| Q7 | Mallattolee rabbi keessa tokko dha | 0.536 | 0.7 |
| Q8 | Aalama jinni fi duniyaa | 0.559 | 0.5 |
| Q9 | Ibidii jahannama dhugaa dha | 0.650 | 0.8 |
| Q10 | Namoota Jahannama hunda | 0.536 | 0.7 |
| Q11 | gooftaan injifata dha | 0.827 | 0.5 |
| Q12 | Badii adda addaa | 0.543 | 0.6 |
| Q13 | Isaan jallina cimaa qabu | 0.647 | 0.8 |
| Q14 | Ergamtoota gaggaarii ergine | 0.412 | 0.8 |
| Q15 | kitaaba gaggaarii ergine | 0.668 | 0.7 |
| Q16 | karaawwan haqaan | 0.419 | 0.7 |
| Q17 | Allaahn laggeen uume | 0.6 | 0.6 |
| Q18 | hojii adabbii hin qabne hojjate | 0.562 | 0.7 |
| Q19 | Jannata fi jahannama sobsiisanii | 0.593 | 0.7 |
| Q20 | amantii haqaan dhufee | 0.746 | 0.8 |
| **Average** | | **0.617** | **0.7** |

**Table4. 2 Mean Average precision and recall for bilingual.**

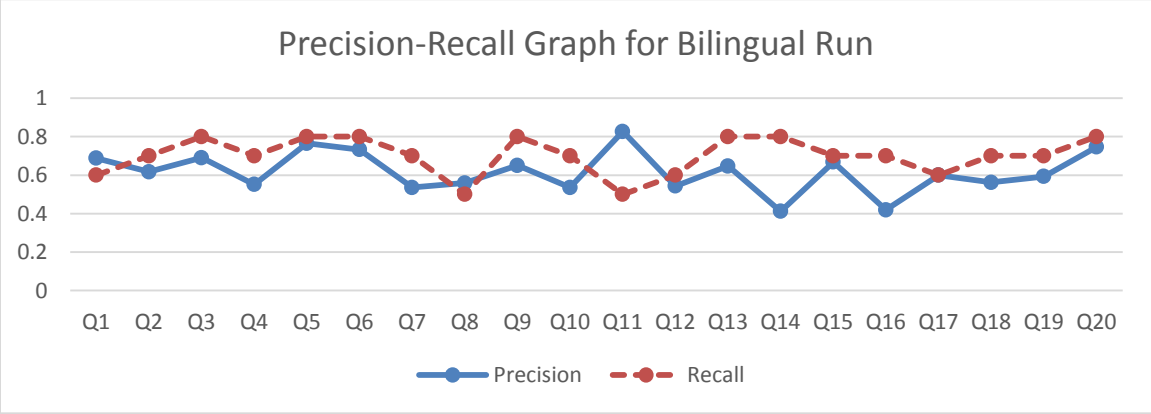

**Figure4. 5 Mean Average Precision for bilingual**

The Precision-Recall graph for the bilingual retrieval experiment (see Figure 4.5) just like the monolingual experiment, illustrates that the precision and recall score for all the 20 queries shows stable or comparable result. This shows that the system behaviors in stable manner.

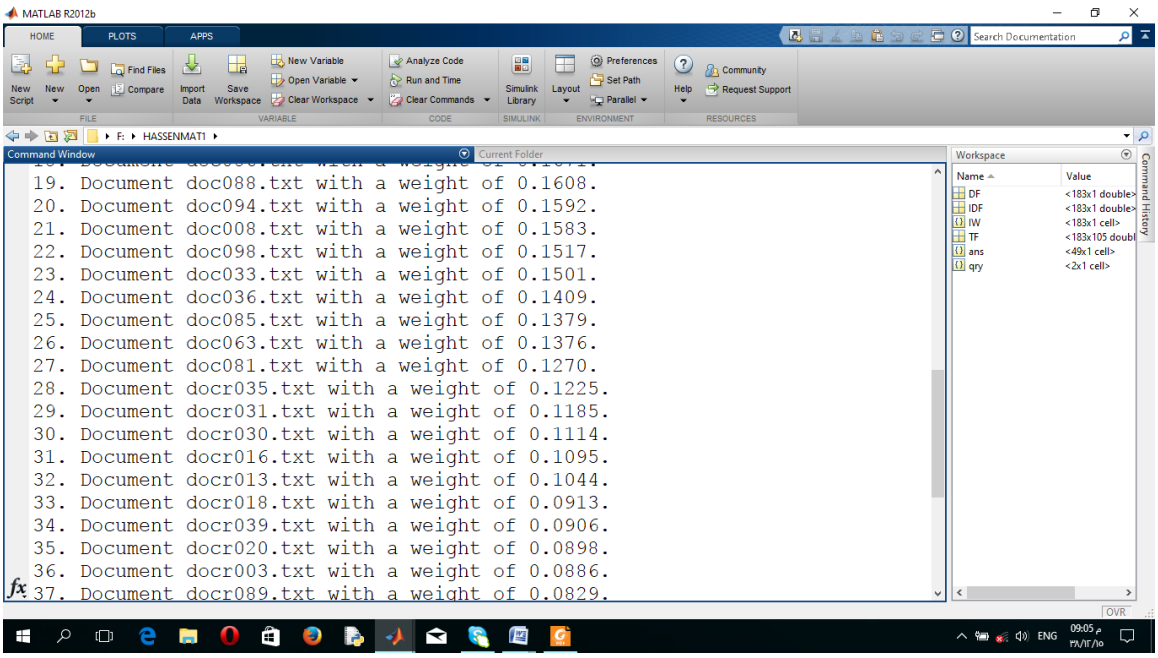


**Figure4. 6 Afaan Oromo-Arabic bilingual**

In this experiment relevant documents were not returned for 2 queries in the monolingual run and for 4 queries in bilingual run out of the total 20 queries. These queries were excluded while calculating the mean average precision at recall levels as it affects the overall performance. Table 4.3 shows the number of queries for which relevant documents were retrieved and not retrieved

| | **Relevant document retrieved** | | **Relevant document not retrieved** | |
|---|---|---|---|---|
| | Afaan Oromo | Arabic | Afaan Oromo | Arabic |
| Number of query | 18 | 17 | 2 | 3 |

| | | | | |
|---|---|---|---|---|
| Percentage | 90 | 85 | 10 | 15 |

**Table4. 3 The ratio of relevant document returned and not returned to queries.**

## 4.5. Analysis

As result from the first experiments show, the monolingual run is comparable with the bilingual runs in experiments. However, the results found in the second experiment indicate a great improvement for the bilingual run with even a better recall value than the monolingual. In general, the results found are encouraging. In this experiment the number of queries for which relevant documents were not returned for the bilingual run was more than that of the monolingual run. This is mainly because the dictionary that was constructed from the corpus has low alignment quality. One reason for the low performance of the dictionary is the size of the corpus used. In addition to the limited size of the corpora that was employed for the bilingual dictionary construction, there are some spells errors that affect accuracy of the alignment result. This is a case where the content is expressed using multiple words in Afaan Oromo and a single word in Arabic and vice_versa.

For example, the Arabic equivalent for the Afaan Oromo word "Ayyaana" is "يَوْم عُطْلَة". In this case a single Afaan Oromo word is translated into two Arabic words. There is also a condition in which a single Arabic word is aligned with more than one Afaan Oromo word. For example, the Arabic word "مَهْمَ" is equivalent with two Afaan Oromo words "Baay'ee Barbaachisaa". These situations are found in the corpora collected but such kinds of alignment was not carried out in this research as the scope is limited to word based alignment only. Another reason for the low performance of the bilingual run (Arabic documents retrieval using Afaan Oromo queries) was the incorrect translation of the Afaan Oromo queries into Arabic. A given test document can be represented by either single word or multiple words of query. For the multiple words of query the translation could be completely correct (i.e. all words of query are correctly translated), partly correct (i.e. not all words of query are correctly translated) or all words in a query wrongly translated. This may happen as a result of reasons indicated. Moreover, the corpora used for this research was not well translated and reliable which result in awful alignment of the bilingual dictionary which also has an effect on the performance of the

bilingual run. As we have observed in the thesis on Oromo-English represent on dictionary based approach techniques and the result obtained precision of 24.50% and recall of 25.72% respectively. According to Aynalem Tesfaye, we have viewed in the thesis on Amharic-English represent on corpus based approach procedures and the result gained precision of 0.24 and recall of 0.33 in that order. And also we have observed in the thesis on Arabic-English represent on A Machine Translation Approach techniques and the result obtained precision of 57.1% and recall of 53.4% respectively. Finally, from Table 2.5 and above literature reviewed we performed a better average precision of 0.617 and recall of 0.7 respectively.

# CHAPTER FIVE
# CONCLUSION AND RECOMMENDATIONS

## 5.1. Conclusion

Cross Lingual Information Retrieval (CLIR) system assists the users to present their query in one language and retrieves documents in another language. In this study, Afaan Oromo-Arabic CLIR, which is based on dictionary-based approach, was developed for Afaan Oromo users to specify their information need in their inhabitant language and to retrieve documents in Arabic as well as Afaan Oromo language. Performance of CLIR systems using dictionary based approach is highly affected by the size, reliability and correctness of the corpus used for the study (Ballesteros, L. and Croft, B., 1996). However, the size of the documents used for this research was limited in size and not quite reliable and clear which affected the level of performance to be achieved. Furthermore, the domains of documents used to carry out the research were very limited. Even though the dictionary-based approach is influenced by size and quality of the corpus, the result obtained is encouraging to develop Afaan Oromo-Arabic CLIR by using this approach. A maximum average precision of 0.617 and 0.7 for Afaan Oromo and Arabic was obtained respectively after conducting the second phase of experimentation. The low performance achieved for the bilingual run (Arabic document retrieval by using Afaan Oromo queries) was because the depth of bilingual dictionary which affects the accuracy of the query translation. This incorrect dictionary was caused due to the limited data size and the low quality of the corpora used for this research. Furthermore, there was a situation in which the number of words in a given Afaan Oromo query was not the same when translated to the equivalent Arabic words. This also affected the accuracy of the Arabic documents retrieved. Thus, it can be concluded that the performance of the system obtained was encouraging given the insufficient amount of resources used.

## 5.2. Recommendations

It is considered that there is much room for improvement of the performance of Afaan Oromo-Arabic CLIR system developed in this research. As a result, the following recommendations should be looked at in the future so that effective Afaan Oromo-Arabic dictionary-based CLIR system can be developed to assist users in their information need:

- The corpora used for this research was from limited domain. This will minimize the similarity of having different variety of words in a bilingual dictionary which affects the performance of the system for other domains. Therefore, some work that incorporates variety of the domains should be tested to see the effect.
- The size of the documents used for this research was limited. These limitations might affect the accuracy of word alignment of the bilingual dictionary because if resources used are small, we may not be able to find all possible translations. Therefore, some work should be done with large and high quality corpora to minimize these problems.
- The current system developed implemented with word-based query translation which reduces the effectiveness of query translation. Use of bilingual phrases instead of single words significantly improves translation quality. The study conducted by Bian et al. (1998) indicates that effectiveness of phrasal translation enhances performance by 14 ~ 31 % than the word level translation. Therefore, it is recommended to test the achievement of phrasal translation techniques.

Reference

Abara, N. (1988). Long Vowels in Afaan Oromo: A Generative Approach. *Master thesis* , 50-51.

Abusalah, M; Tait, . (2005). Literature review of cross-language information retrieval. *In Transaction on engineering, Computing and Technology* , 4-6.

Adola, A. R. (2012.). Homonymy as a barrier to mutual intelligibility. *Journal of Language and* , pp. 32-43,.

Ahmad M. Hasnah Jihad M. Jaam. (2002). Thesaurus-Based Query Disambiguation Method for Cross-Language Information Retrieval. *JICIS,* , 58-68.

Alduais, A. M. (2012). Simple Sentence Structure of Standard Arabic Language. *Department of English, King Saud University (KSU)* , 500-504.

Aljlayl, et al. (2002). On Arabic-English Cross-Language Information Retrieval:. *Information Retrieval Laboratory* , 34-37.

Ari Pirkola, et al. (2002). Dictionary-Based Cross Language Information Retrieval: Problem, method and research Finding. 53-56.

Ayan, D. B. (2011). Afaan Oromo-English Cross-Lingual Information retrieval. *Corpus base approach MSc, Addis Ababa University* , 34-37.

Aynalem.T, Kevin.S. (2010). Amharic – English Cross-lingual Information Retrieval: A Corpus Based Approach. *M.Sc. Thesis, Addis Ababa University, Addis Ababa,* , 23.

B.N.V Narasimha Raju, M S V S Bhadri Raju. (2015). Dictionary Based Translation Approaches in Cross Language Information Retrieval. *International Journal of Scientific & Engineering Research,* , 324-330.

Baeza R and Ribeiro. (1999). *Modern information retrieval,.* chicago, united s: Adventure works press.

Ballesteros, et al. (1998). Resolving ambiguity for cross-language retrieval. *In: Preceedings of the 21st annual international. ACM SIGIR conforence on rReseach and development in information retrieval* , pp.64-71.

Ballesteros, L. a. (1996). Dictionary-based methods for cross-lingual. *In Proceedings of the 7th International DEXA* , pp. 79-80.

Ballesteros, L. and Croft, B. (1996). Dictionary methods for cross-lingual information retrievall. *In Proceedings of the 7th International DEXA* , pp. 791-801.

Ballesteros, L. and Croft, B. (1997). Phrasal Translation and Query Expansion. *In: Proceeding of the* , pp 84-91.

Bian, G. a. (1998). New Hybrid Approach for Chinese-English Query. pp. 156-.

Brown, P. F et al. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics, 19,* , 263-311.

Daniel. (2011). Afaan Oromo-English Cross-Language Information Retrieval: Corpus Based Approach. *Master thesis , Addis Ababa University* , 43.

David, A. a. (1996). Querying across languages: A dictionary-based. *In Proceedings of the 19th* , pp. 49-57.

Deng, Y. a. (2006). MTTK: An Alignment Toolkit for Statistical Machine. *Association for Computational Linguistics , Companion* , pp.265-268.

Eggi, G. G. (2012). Afaan Oromo Text Retrieval System. *MSc, Addis Ababa University* , 23-25.

Elghazaly, T. A. (2009). English/Arabic Cross Language Information Retrieval. *Communications of the IBIMA* , 208-217.

El-Khair, I. A. (2006). Effects of stop words elimination for Arabic information retrieval. *a comparative study." International Journal of Computing & Information Sciences 4.3* , 119-133.

Fraser, A. & Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational linguistics 33,* , 293-303.

Frieder, M. A. (2001). Effective Arabic-English Cross-Language Information. *Information Retrieval Laboratory* , 33.

Gale, W. a. (1991). A program for aligning sentences in bilingual corpora. pp. 177-184.

Grage, G. Kumsa, T. (1982). Oromo Dictionary. African Studies Center, Michigan. *Masters thesis* , 30.

Hamiid, M. (1995.). Hamiid Muudee's English-Oromo Dictionary. *ol.1: Atlanta.* , 23-30.

Hedlund. (2004). Dictionary-Based Cross-Language Information. *Learning Experiences from CLEF 20002.* , 45-46.

Hull, D. A. (1996). Querying across languages: a dictionary-based approach to. *Proceedings of the 19th annual international ACM* , ACM, 49-57.

Hull, et al. (1996). Querying across languages: a dictionary-based. *ACM,* , 49-57.

Jagarlamudi, J. a. (2007). Cross-Lingual Information Retrieval System. *Bangalore, INDIA* , 40.

Kishida, K. (2005). Technical issues of cross-language information retrieval:. *: a review. Information* , 433-455.

Kishida, K. (2005). Technical Issues of Cross-Language Information retrieval:. *A review* , 433-455.

Kula. (2008). Evaluation of Oromo-English CrossLanguage Technologies Research Center. *Information Retrieval IIIT,* , 59-61.

Leah S. Larkey and Margaret E. (2005). Structured Queries, Language Modeling, and Relevance Modeling in Cross-Language Information Retrieval. *Amherst* , 1-20.

Lu, C.et al . (2008). Web-Based Query Translation for English-Chinese. *Vol. 13,*, pp. 61-90.

Manning et al. (2008). *Introduction to information retrieval.* Cambridge: Cambridge university press,.

Mathematik, V.at el. (2002). Statistical Machine Translation:.

Mohammed Aljlayl, Ophir Frieder, & David Grossman. (2002). On Arabic-English Cross-Language Information Retrieval:. *Information Retrieval Laboratory* , 45-47.

Mohd Amin Mohd Yunus, Roziati Zainuddin, and Noorhidawati Abdullah. (2013). Semantic Method for Query Translation Cross language information retrieval (CLIR). *The International Arab Journal of Information Technology* , 253-259.

Mustafa Abusalah, John Tait, and Michael Oakes. (2009). Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base. 13-16.

Mylanguage.org. (2011, May 21). —*Oromo Numbers,||.* Retrieved -April- 3, 2016, from [Online]. Available:: http://www.mylanguages.org/learn_oromo.php

Nasreen AbdulJaleel and Leah S. Larkey. (2003). Statistical Transliteration for English-Arabic Cross Language Information Retrieval. *Center for Intelligent Information Retrieval* .

Oard, D. (1997). Alternative Approaches for Cross-Language Text Retrieval. *AAI* , pp. 154-162.

Oard, D. W. (1998). Cross-language information retrieval. . *Annual* , 33.

Och, F. a. (2003). A Systematic Comparison of Various Statistical Alignment. 29(1):19-51.

Oromoo. (1995). Gumii Qormaata Afaan Oromoo. *Caasluga Afaan Oromoo, Jildi I, Komishinii* , 45-45.

Pirkola, A. et, al. (2001). Dictionary-based crosslanguage Information Retrieval: Problems, Methods and Research Findings.,. 209-230.

Ramanathan, A. (2003). State of the Art in Cross-Lingual Information Retrieval, National.

Ramis, G. (2006). Introducing Linguistic Knowledge into Statistical Machine.

Salton, G. &. (1986). Introduction to modern information retrieval. *McGraw-Hill, New York.* , 43.

Saralegi,et al,. (2009). Comparing different approaches to treat. *DEXA'09.* , 398-404.

Sisay, A. (2009). English-Oromo Machine Translation: An Experiment Using a. *An Experiment Using a M.Sc Thesis, Addis Ababa University, Addis Ababa,* , 21-23.

Soucy, P. &. (2005). Year. Beyond Tfidf weighting for text categorization in. *Lawrence Erlbaum Associates ltd, 1130.*

Talvensaari, T. (2008). Comparable corpora in cross-language information Retrieval. *Department of Computer Sciences.*

Talvensaari, T. (2008). Comparable corpora in cross-language information Retrieval. *University of Tampere, Department of Computer Sciences.*

Talvensaari, T. Juhola, M. (2007). Corpus-Based Cross Cross-Language information retrieval in retrieval of hily relevant documents. *journal of the American Society for Information Science and Tecnology* , 322-334.

Tilahun, G. (1993). Qube Afaan Orom: Reasons for Choosing the Latin Script for. *The Journal of Oromo Studies* , 1(1).

Tilahun.G. (1993). Qube Afaan Orom: Reasons for Choosing the Latin Script for. *The Journal of Oromo Studies* .

Tune, K. K. (2015). Development of Cross-Language Information Retrieval for Resource-Scarce African Languages. *International Institute of Information Technology, Hyderabad* , 80-92.

Vogel, S., Hermann, N. and Christophe, T. (1993). *Hmm-based word alignment in statistical translation. In Proceedings of the 16th Conference on Computational Linguistics,.* Copenhagen, Denmark: Association for Computational Linguistics.

Zens, R. J. (2002). Phrase-Based Statistical Machine Translation,. *KI 2002: Advances in* , pp35-56.

# Appendix A: Afaan Oromo stop words

| | | | | |
|---|---|---|---|---|
| **aanee** | **agarsiisoo** | **akka** | **akkam** | **akkasuma** |
| **akkum** | **akkuma** | **ala** | **alatti** | **alla** |
| **amma** | **ammo** | **ammo** | **an** | **ana** |
| **ani** | **ati** | **bira** | **booda** | **booddee** |
| **dabalatees** | **dhaan** | **dudduuba** | **dugda** | **dura** |
| **duuba** | **eega** | **eegana** | **eegasii** | **ennaa** |
| **erga** | **ergii** | **faallaa** | **fagaatee** | **fi** |
| **fullee** | **fuullee** | **gajjallaa** | **gama** | **gararraa** |
| **garas** | **garuu** | **giddu** | **gidduu** | **gubbaa** |
| **haa** | **hamma** | **hanga** | **henna** | **hoggaa** |
| **hogguu** | **hoo** | **illee** | **immoo** | **ini** |
| **innaa** | **inni** | **irra** | **irraa** | **irraan** |
| **isa** | **isaa** | **isaaf** | **isaan** | **isaani** |
| **isaanii** | **isaaniitiin** | **isaanirraa** | **isaanitti** | **isaatiin** |
| **isarraa** | **isatti** | **isee** | **iseen** | **ishee** |
| **ishii** | **ishiif** | **ishiin** | **ishiirraa** | **ishiitti** |
| **ishiitti** | **isii** | **isiin** | **isin** | **isini** |
| **isinii** | **isiniif** | **isiniin** | **isinirraa** | **isinitti** |
| **ittaanee** | **itti** | **itumallee** | **ituu** | **ituullee** |
| **jala** | **jara** | **jechaan** | **jechoota** | **jechuu** |

| | | | | |
|---|---|---|---|---|
| **jechuun** | **kan** | **kana** | **kanaa** | **kanaaf** |
| **kanaafi** | **kanaafi** | **kanaafuu** | **kanaan** | **kanaatti** |
| **karaa** | **kee** | **keenna** | **keenya** | **keessa** |
| **keessan** | **keessatti** | **kiyya** | **koo** | **kun** |
| **lafa** | **lama** | **malee** | **manna** | **maqaa** |
| **moo** | **na** | **naa** | **naaf** | **naan** |
| **naannoo** | **narraa** | **natty** | **nu** | **nu'i** |
| **nurraa** | **nuti** | **nutty** | **nuu** | **nuuf** |
| **nuun** | **nuy** | **odoo** | **ofii** | **oggaa** |
| **oo** | **osoo** | **otoo** | **otumallee** | **otuu** |
| **otuullee** | **saaniif** | **sadii** | **sana** | **saniif** |
| **si** | **sii** | **siif** | **siin** | **silaa** |
| **silaa** | **simmoo** | **sinitti** | **siqee** | **sirraa** |
| **sitti** | **sun** | **tahullee** | **tana** | **tanaaf** |
| **tanaafi** | **tanaafuu** | **ta'ullee** | **ta'uyyu** | **ta'uyyuu** |
| **tawullee** | **teenya** | **teessan** | **tiyya** | **too** |
| **tti** | **utuu** | **waa'ee** | **waan** | **waggaa** |
| **wajjin** | **warra** | **woo** | **yammuu** | **yemmuu** |
| **yeroo** | **yommii** | **yommuu** | **yoo** | **yookaan** |
| **yookiin** | **yoolinimoo** | **yoom** | | |

# Appendix B: Arabic stop words

| | | | | | |
|---|---|---|---|---|---|
| من | له | لن | لم | كل | فى في |
| منذ | لها | كما | قوة | هي | هو |
| هناك | مقابل | لقاء | نفسه | ولا | وقد |
| للامم | وكانت | وقالت | نهاية | وكان | وقال |
| ولم | وقف | وفي | لكن | كلم | فيه |
| مليار | فيها | يوم | وهي | وهو | ومن |
| حيث | مليون | يمكن | يكون | لوكالة | منها |
| | السابق | امس | اما | الا | اكد |
| ايام | امام | الان | الذي | الذى | التى |
| ذلك | بين | الاولى | الاول | الذين | خلال |
| اول | انه | الى | الف | حين | دون |
| المقبل | الوقت | الماضي | جميع | انها | ضمن |
| ما | لا | قد | و | و - | اليوم |
| فان | واضافت | واضاف | واحد | هذا | مع |
| وان | هذه | نحو | لدى | كان | قبل |
| أ | ا | ب | مايو | واوضح | واكد |
| عام | عدم | عشرة | عدة | عدد | ، |
| عليها | عليه | على | عندما | عند | عشر |
| بعض | بعد | ضد | تم | سنوات | زيارة |
| اثر | احد | اذا | حتى | بسبب | اعادة |

| | | | | | |
|---|---|---|---|---|---|
| اربعة | اطار | صباح | شخصا | غدا | برس |
| بن | حاليا | بشكل | غير | اجل | اخرى |
| ف | المساء | باسم | سنة | اعلنت | به |
| ان | اف | بان | عن | كانت | حول |
| عاما | صفر | بها | اي | او | ثم |

# Appendix C: MatlabworkR2012b code for Indexing.

```
function [IW,TF,DF,IDF]=indexer(refdir,stpw)
%function [TBL,IW,TF,DF,IDF]=indexer(refdir,stpw)
disp('Please wait ...')
files = dir(refdir); % list the content of the reference
directory
files = files(3:end); % remove current (.) and privous (..)
directory from the list
ND = numel(files); % number of documnents
term = cell(0,0); % term container
for i=1:ND
fid = fopen(strcat(refdir,'\',files(i).name));% open file
wrd=textscan(fid,'%s'); %read words
term = [term; wrd{:,:}]; % concatinate words to the terms
fclose(fid); % close the file
end
term = unique(term); % remove redendent words
%% read stop words
fid = fopen(stpw);% open file
wrd=textscan(fid,'%s'); %read words
stw = wrd{:,:}; % concatinate words to the terms
fclose(fid); % close the file

%% remove stop words from the terms
s = numel(term); % size of trems
u = cell(s,1);
ff=0;k=1;
for i=1:s
for n=1:numel(stw)
if strcmpi(stw{n,1},term{i,1})
ff=1;
end
end
if ff==0
u{k,1}=term{i,1};
    k=k+1;
else
ff=0;
end
end
term = cell(k-1,1);
for i=1:k-1
term{i,1}=u{i,1};
end
%% calculate term frequency
NT = numel(term); % number of terms
```

```
TF = zeros(NT,ND); % Term frequency matrix
for n=1:ND % open each document
fid = fopen(strcat(refdir,'\',files(n).name));% open file
wrd=textscan(fid,'%s'); %read words
fclose(fid);
wrd = wrd{:,:};
for i = 1:numel(wrd)
for j=1:NT
if strcmpi(term{j,1},wrd{i,1})
TF(j,n)= TF(j,n)+1; % count frequency of term j in documnet
n
end
end
end
end

%% Remove special characters
if refdir == 'afcorpus'
for i=1:numel(term)
term{i,1}(regexp(term{i,1},'[^A-Za-z'']'))=[];
end
else
for i=1:numel(term)
term{i,1}(regexp(term{i,1},'[0-9()]'))=[];
end
end
% for i=1:numel(term)
%     term{i,1}(regexp(term{i,1},'[^A-Za-z'']'))=[];
% end
k=0;
for i=1:numel(term)
if numel(term{i,1})>1
        k = k + 1;
term{k,1}=term{i,1};
TF(k,:) = TF(i,:);
end
end
[term,ia,ic] = unique(lower(term)); % remove redendent
words
%% Recalculate TF acourding to the sorted unique terms
tmp = TF(ia,:);
for i=1:numel(ia)
tt=0;
for n=1:numel(ic)
if ia(i)==ic(n)
tt = tt+1;
if tt>1
```

```
tmp(i,:)=TF(i,:)+TF(n,:);
end
end
end
end
TF = tmp;
%% Calculate document frequency
DF = sum(TF>0,2);
[DF,I] = sort(DF);
TFt = TF;
IW = term;
NT = numel(DF);
for i=1:numel(DF)
TFt(i,:) = TF(I(i),:);
IW{i} = term{I(i)};
end
%% Remove terms with DF <= 5% and DF >= 40%
LL = ceil(.05*ND);
UL = ceil(.4*ND);
% KK = DF>LL; KK = DF<UL;  % one will overide the other
find another mechanithm
% NNT = sum(KK);
term = cell(NT,1);
TF = zeros(NT,ND);
DDF = zeros(NT,1);
n=0;
for i=1:NT
if DF(i)>LL && DF(i)<UL
        n = n +1;
term{n} = IW{i,1};
TF(n,:) = TFt(i,:);
DDF(n) = DF(i);
end
end
DF = DDF(1:n);
TF = TF(1:n,:);
IW = cell(n,1);
for i=1:n
IW{i} = term{i};
end
%% calculate IDF
IDF = -log10(DF/ND);
%% export to excell worksheet
[r,c] = size(TF);
R = r+1; C = c+2;
xel = cell(R,C);
hdr = cell(1,C);
```

```
hdr{1,1} = 'Terms';
hdr{1,2} = 'Raw DF';
if refdir == 'afcorpus'
for i=3:C
if i < 10
tmp = strcat('00',int2str(i-2));
elseif i<100
tmp = strcat('0',int2str(i-2));
else
tmp = int2str(i-2);
end
hdr{i} = strcat('doc',tmp);
end
end
if refdir == 'arcorpus'
for i=3:C
if i < 10
tmp = strcat('00',int2str(i-2));
elseif i<100
tmp = strcat('0',int2str(i-2));
else
tmp = int2str(i-2);
end
hdr{i} = strcat('docr',tmp);
end
end
for c=1:C
xel{1,c} = hdr{1,c};
end
for r=2:R
xel{r,1} = term{r-1,1};
xel{r,2} = DF(r-1,1);
end
for r = 2:R
for c=3:C
xel{r,c} = TF(r-1,c-2);
end
end
if strcmpi(refdir,'arcorpus')
xlswrite('ArabicRawTermFreq.xlsx',xel);
else
xlswrite('OromifaaRawTermFreq.xlsx',xel);
end
%% Draw IDF~TF graph to select index terms
X=sum(TF,2);
[X,I] = sort(X);
Y = IDF(I);
```

```
plot(X,Y)
gridon;
ylabel('Inverse document frequency');
xlabel('Term frequency over all documnets')
%% calculate Document-term wight vector
for i = 1:ND
    TF(:,i) = TF(:,i)/max(TF(:,i));
end
%% export to excell worksheet
for r = 2:R
for c=3:C
xel{r,c} = TF(r-1,c-2);
end
end
if strcmpi(refdir,'arcorpus')
xlswrite('ArabicNormTermFreq.xlsx',xel);
else
xlswrite('OromicNormTermFreq.xlsx',xel);
end
%% Calculate Normalized document-term wight vector
for i = 1:ND
TF(:,i) = TF(:,i).*IDF;
end
%% Export to excell worksheet
xel{1,2} = 'IDF';
for r=2:R
xel{r,2} = IDF(r-1,1);
end
for r = 2:R
for c=3:C
xel{r,c} = TF(r-1,c-2);
end
end
if strcmpi(refdir,'arcorpus')
xlswrite('ArabicWeightVector.xlsx',xel);
else
xlswrite('OromicWeightVector.xlsx',xel);
end
% % xlswrite('WeightVector.xlsx',xel);
%% Return output  variables
% IW = term;
% %%total TF
% TTF=sum(TF,2);
% TTF = TTF/max(TTF);
% plot(TTF.*IDF);
% %% Normalize Document-term wight vector
% for i = 1:ND
```

```
%     TF(:,i) = TF(:,i)/sqrt(sum(TF(:,i).^2));
% end
% Note: Here after TF is not Term frequency rather it is
Normalized document-term wight vector
if refdir == 'afcorpus'
saveofindexIWIDFTF
elseif refdir == 'arcorpus'
savearindexIWIDFTF
else
```

**Continue……**

# Appendix D: MatlabworkR2012b code for Searching.

```
function [Rdoc,wgt] = searchdoc(qry,indx)
if indx == 'ofindex'
loadofindex;
else
loadarindex;
end
qtf = zeros(numel(IDF),1);
qdf = zeros(numel(IDF),1);
for i=1:numel(qry)
for n=1:numel(IW)
if strcmpi(qry(i),IW(n))
qdf(n) = IDF(n);
qtf(n) = qtf(n)+1;
end
end
end
I = find(qtf==0);
qtf = 1+log10(qtf);
qtf(I) = 0;
qwt = qdf.*qtf;
qwt = qwt/sum(qwt);
[~,ND] = size(TF);
wgt = zeros(ND,1);
for i=1:ND
    wgt(i) =
dot(qwt,TF(:,i))/(sqrt(sum(qwt.^2))*sqrt(sum(TF(:,i).^2)));
end
% trshold = cos(pi*30/180);
doc = cell(ND,1);
if indx == 'ofindex'
for i=1:ND
if wgt(i)>0
if i < 10
tmp = strcat('00',int2str(i));
elseif i<100
tmp = strcat('0',int2str(i));
else
tmp = int2str(i);
end
doc(i) = {strcat('doc',tmp,'.txt')};
else
doc(i)={'NONE'};
end
end
end
```

```
if indx == 'arindex'
for i=1:ND
if wgt(i)>0
if i < 10
tmp = strcat('00',int2str(i));
elseif i<100
tmp = strcat('0',int2str(i));
else
tmp = int2str(i);
end
doc(i) = {strcat('docr',tmp,'.txt')};
else
doc(i)={'NONE'};
end
end
end
[wgt,I] = sort(wgt);
doc = doc(I);
D = find(wgt>0);
Rdoc = cell(numel(D),1);
for i=1:numel(D)
Rdoc{i,1} = doc{D(i),1};
end
% Rdoc = doc{i:end,1}=[];
wgt = wgt(D);
function [rdoc,wgt]=mysearch(qry1,IL)
qry2=trans('dic.txt',qry1,IL);
if IL == 'OF'
    [doc1,wgt1]=searchdoc(qry1,'ofindex');
    [doc2,wgt2]=searchdoc(qry2,'arindex');
elseif IL == 'AR'
    [doc1,wgt1]=searchdoc(qry2,'ofindex');
    [doc2,wgt2]=searchdoc(qry1,'arindex');
else
disp(strcat(IL, ' is unknoun language! Please use ''OF''
for Afaan Oromo or ''AR'' for Arabic.'));
return;
end
doc = [doc1; doc2];
rdoc = doc;
wgt = [wgt1;wgt2];
[wgt, I] = sort(wgt,'descend');
for i = 1:numel(I)
rdoc{i,1} = doc{i,1};
fprintf('%d. Document %s with a weight of
%6.4f.\n',i,rdoc{i,1},wgt(i));
end
```

```
function [OL]=trans(dicfile,qry,IL)
%%function [OL]=trans(dicfile,qry,IL) transilates quary
(qry) language IL into
%%output language (OL) using dictionary file (dicfile)
OL=cell(qry); % define OL
%%open dictionary file
fid = fopen(dicfile);% open file
wrd=textscan(fid,'%s'); %read words
wrd = wrd{:,1};
fclose(fid); % close the file
%% copy lexicals of the dictionary to a variable lexi
lexi = cell(round(numel(wrd)/2),2);
%read column 1
o=1;
for i=1:2:numel(wrd)
lexi{o,1}=wrd{i,1};
    o=o+1;
end
%read column 2
e=1;
for i=2:2:numel(wrd)
lexi{e,2}=wrd{i,1};
    e=e+1;
end
%% Transilate if the quary language is Oromiifaa ('OF')
into output language Arebic ('AR') and vice versa
if IL=='AR'
for i=1:numel(OL)
for n=1:numel(lexi(:,1))
if strcmpi(qry(i),lexi(n,2))
OL(i)=lexi(n,1);
end
end
end
elseif IL=='OF'
for i=1:numel(OL)
for n=1:numel(lexi(:,1))
if strcmpi(qry(i),lexi(n,1))
OL(i)=lexi(n,2);
end
end
end
else
disp(strcat(IL, ' is unknoun language! Please use ''OF''
for Afaan Oromo or ''AR'' for Arabic.'));
return;
```

```
end
end continuo ……
```

# Declaration

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

---

**Hassen Hussen Kassim**

**September 2017**

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

---

**Wondwossen Mulugeta (PhD)**

**September 2017**