2020-03-17

AUTOMATIC THESAURUS CONSTRUCTION FROM AFAN OROMO TEXT

Wakjira, Firomsa

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

FACULTY OF COMPUTING

AUTOMATIC THESAURUS CONSTRUCTION FROM

AFAN OROMO TEXT

Firomsa Wakjira Gemeda

Bahir Dar, Ethiopia

July 2018 G.C

AUTOMATIC THESAURUS CONSTRUCTION FROM AFAN OROMO TEXT

Firomsa Wakjira Gemeda

A Thesis submitted to the School of Research and Graduate Studies of Bahir Dar Institute of Technology, Bahir Dar University in partial fulfillment of the requirements for the degree of Master of Science in Information Technology in the Faculty of Computing

Supervised by: Bhabani Shankar Das Mohapatra

Bahir Dar, Ethiopia

July 2018 G.C

DECLARATION

I, the undersigned, declare that the thesis comprises my own work. In compliance with internationally accepted practices, I have acknowledged and refereed all materials used in this work. I understand that non-adherence to the principles of academic honesty and integrity, misrepresentation/fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the University and can also evoke penalty action from the sources which have not been properly cited or acknowledged.

Name of the student: Firomsa Wakjira Signature _____

Date of submission: July 2018 G.C

Place: Bahir Dar, Ethiopia

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

Advisor Name: Bhabani Shankar Das Mohapatra

Advisor's Signature: _____

Bahir Dar University

Bahir Dar Institute of Technology

School of Research and Graduate Studies

Faculty of Computing

THESIS APPROVAL SHEET

Student:
Firomsa Wakjira — Signature — 16/11/2010 E.C
Name — Signature — Date

The following graduate faculty members certify that this student has successfully presented the necessary written final thesis and oral presentation for partial fulfillment of thethesis requirements for the Degree of Master of Science in Information Technology.
Approved by:

Advisor:
Bhabani Shankar Das Mohapatra — Signature — 23/07/2018
Name — Signature — Date

External Examiner:
Dr. Million Meshesha — Signature — July 16, 2018
Name — Signature — Date

Internal examiner:
Dr. Gebeyehu Belay — signature — 23/07/2018
Name — signature — Date

Chair Holder:
Dr. Krishna Prasad — Signature — 23/07/2018
Name — Signature — Date

Faculty Dean:
Belete Biazen — Signature — 16/11/2010 E.C
Name — Signature — Date

DEDICATION

In memory of innocent Oromo people who are massacred during Irrecha celebration in 2016 G.C and who lost their lives and were displaced from Ethio-Somali in 2017 G.C!

ACKNOWLEDGEMENT

First and foremost, I express my deapest gratitude to the Almighty God with whose sanctification anything is feasible in this transitory world. "**Yaa rabbi galanni kee hin dhumu ati"!**

I would extend my sincere appreciation to my advisor Prof. Bhabani Shankar for his critical comments. With great pleasure and deep sense of gratitude, I express my indebtedness to Dr. Gebeyehu Belay for his invaluable guidance and constant encouragement at each and every step of my Thesis work regardless of my irregular appearance for progress follow up and always showed great interest in providing timely support and suitable suggestions.

My special thanks go to my family members for their moral support and encouragement during my entire study period. I wish God to sanctify them all.

I am also grateful to my colleagues and friends, especially Hailu Bashada, for their benevolence in sharing the ideas, directions, comments and encouragements.

Finally, I would like to extend my profound appreciation to Haileamlak Mulugeta, a tenth-grade student, for his contribution as a teenager. I wish him a great achievement and success in his future endeavor.

ABSTRACT

Thesaurus is a reference of words or of information about a particular field or set of concepts, especially, a tome of words and their synonyms or a list of subject-headings or descriptors usually with a cross-reference system for use in the organization of a collectionof documents for reference and retrieval. One of the major problems of modern information retrieval systems is the vocabulary problem that concerns with the discrepancies between terms used for describing documents and the terms used by the searcher to describe their information need which forms the information overload or information mismatch. One way of handling the vocabulary problem is using a thesaurus that shows the relationships between terms and query expansion which provides us the alternative terms for query to improve the effectiveness of retrieval. Since the manual thesaurus construction is a labor-intensive task and hence also expensive to build and hard to update in timely manner, Afan Oromo automatic thesaurus is implemented by using the term-clustering approach. In this research, 36869 selected words from the collected document are used and are suggested to improve the expansion process and to get more relevance documents for the user's query**.**

The performance of the experiment is very encouraging and promising as the accuracy of the system performance is 56.6% on Afan Oromo documents. And also 73.11% of the terms in the collection are registered to be similar. More challenge here is, the complexity of Afan Oromo which results in under or over stemmed and this is due to the non-proper preprocessing of the document. The performance and the accuracy of this system is improved if the document is properly preprocessed and more effective in large collections over multiple domains. The quality of the cluster is measured by intra-cluster and inter-clustering techniques and the result registers 1.33.

TABLE OF CONTENTS

LIST OF ACRONYMS

CV	Controlled Vocabulary
IDF	Inverse Document Frequency
IR	Information Retrieval
ISO	International Organization for Stnd
LSI	Latent Semantic Index
NLP	Natural Language processing
NT	Narrower Term
OBN	Oromia Broadcasting Network
OLF	Oromo Liberation Front
TF	Term Frequency
VOA	Voice of America
VSM	Vector Space Model

LIST OF ALGORITHMS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

A Thesaurus is a reference work that enlists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order. A thesaurus is a set of terms that are semantically related (Amirhosseini, 2008). It helps in improving the quality of retrieval by guiding indexers and searchers about which terms to use. It is a structure which supports the automatic indexing and retrieval, and it is a structured dictionary which is focused on the representation of a limited set of semantic relations between different concepts (Meusel & Stuckenschmidt, 2009). In most thesauri, user's encounter is manually constructed by domain experts and/or experts at document description. Manual thesaurus construction is a time-consuming and expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus makes choices that affect the structure of the thesaurus. There is a need for the methods of automatic construction of thesauri, which, besides the improvements in time and cost aspects, can result in more objective thesauri that are easier to update.

A fundamental problem for information retrieval (IR) is the mismatch of query (information need) of IR system users with the authors of the documents stored in the systems. This is due to the use of different words to refer the same concept (Xu, 1997). Thesaurus however is used to identify the concepts in a knowledge collection assisting as a unified navigation tools that enables users with effective and efficient search, i.e., it is a dynamic approach of satisfying users need. The different definitions of thesaurus vary from quite modest definitions that focus on the relations between words without stating which

kinds of relations that are meant, to such definitions that state more exactly about which relations that are concerned. An example of quite a modest definition is presented by (Schütze, 1998). They defined thesaurus as simply a mapping from words to other closely related words and must be specific enough to present the searcher synonyms of words in the corpus that is searched. Moreover, they stated that a thesaurus must cover all of the words found in queries (ibid). This is an interesting statement – "how can one know in advance that which words are to be used for searching?" Again, another researcher (MILLER, 1997) bequeathed a more sumptuous definition of a thesaurus as "a lexico-semantical model of a conceptual reality or its constituent, which is expressed in the form of a system of terms and their relations, offers access via multiple aspects and is used as a processing and searching tool of an information retrieval unit". Hence, it appears that a process of thesaurus construction is a process of simulation in a lexical form: of the whole universe of realities and concepts or its part of hierarchical and associative connections and relations between these realities and concepts. A thesaurus is a data structure that defines semantic relatedness between words and is typically used in IR to expand search terms with other closely related words (Schutze & Pedersen, 1997). A thesaurus provides a precise and controlled vocabulary that serves to enhance retrieval performance in an IR system both in document indexing and document searching (Mekonnen, 2009). With very large collection of documents, an effective thesaurus needs the reduction of set of representative index terms in order to reduce the computational costs. This can be accomplished through text operations. Text operations include five document processing procedures. These are lexical analysis, elimination of stopwords, stemming, selection of representative terms and term categorization through thesaurus (Ricardo Baeza-Yates, n.d.). This study is intended to examine the existing automatic thesaurus generation techniques and acclimatize them to Afan Oromo language. This is necessary because so far there is no research progress in English that details the construction of Afan Oromo thesauri. With the proposed thesaurus of this research, a set of experiments is performed to measure its applicability and quality.

1.2 Statement of the problem

The Oromo nation has a single common mother tongue with a basic common culture. The Oromo language, Afan Oromoo or Oromiffa, belongs to the eastern Kushitic group of languages and is the most widespread among almost forty Kushitic languages. Oromo language is very much closely related to Konso, with more than fifty percent of the words in common, closely related to Somali and distantly related to Afar and Saho.

Afan Oromo is considered as one of the five most widely spoken languages from among approximately 1000 languages of Africa, (Gragg, 1982). Taking into consideration the number of speakers and the geographic area it covers, Afan Oromo, most probably is rated as the second among the African indigenous languages and is the third most widely spoken language in Africa, after Arabic and Hausa. Afan Oromo is the mother tongue of around 40 million Oromo people living in the Ethiopian Territory and neighbouring countries. Perhaps not less than two million non-Oromo folks speak Afan Oromo as a second language. In fact, Afan Oromo is a lingua franca in the whole of Ethiopian Territory except for the northern part. It is a language spoken in common by several members of many of the nationalities like Harari, Anuak, Barta, Sidama, Gurage, etc., who are neighbors to Oromo and Afaan Oromo folks. Oromo and Afan Oromo are written with a Latin alphabet called Qubee. In spite of relatively large number of speakers and its power of expressiveness for things, it is a language for which very few computational linguistic resources have been developed for many years, and very little progress have been done in terms of making useful higher-level Internet or computer-based applications available to those who only speak the language. Following the role of the language in societal development and the escalating demand of information, there would be a critical time for tapping the available huge information resources through the inevitable retrieval system using the language as a tool or through cross language retrieval system. Hence this research attempts to construct an automatic thesaurus from Afan Oromo text for possible inception of enhanced retrieval system that delivers a framework for the development of cross-language retrieval system.

Thesauri are significant and valuable tools of IR systems for enhanced term classification, categorization, summarization and filtering during indexing, and query expansions during searching. One of the major problems in modern IR is the vocabulary problem that concerns the representativeness between terms used for representing documents and the terms used by the searchers to describe their information need (Khafajeh et al., 2002). Besides the information overload, another challenge in information processing is the vocabulary mismatch problem, referring to the fact that people tend to use different terms to describe a concept.

Due to their different backgrounds and expertise, the chance that two people describe a concept using the same term is pretty low. This even may happen with the same person. He/she may use different terms to describe the same concept at different times because of the learning process and the evolution of concepts (Chen, 2006). One way of handling the vocabulary problem is using a thesaurus that demonstrates the relationships among terms.

Therefore, the purpose of this research work is to conduct automatic thesaurus for Afan Oromo text retrieval to assist in the development of an effective Afaan Oromo IR systems. And the research questions are formulated as follows:

- ✓ What are different approachs used for constructing automatic thesaurus?
- ✓ Which approach is considered to be used for Afan OromoThesaurus construction?
- ✓ How the approach is implemented in constructing automatic thesaurus from Afan Oromo text?
- ✓ Which similarity computation should be used for this study?
- ✓ What is the basic syllable structure, morphemes, and phonetics in Afan Oromo and how is the word formed?

1.3 Objective of the study

1.3.1 General objective

The general objective of this research is to design automatic thesaurus for Afaan Oromo text from document collection.

1.3.2 Specific objectives

To achieve the general objective, this research work formulates the following specific objectives:

- ✓ To identify techniques important for constructing automatic thesaurus from Afaan Oromo text.
- ✓ To prepare data sets required for automatic thesaurus construction.
- ✓ To develop the blueprint or a prototype of automatic thesaurus construction from Afaan Oromo language.
- ✓ To evaluate the performance of the prototype.

1.4 Methodology

To answer the research questions and achieve the stated objectives, this research work used the design used to follow and subsequent techniques as described below.

1.4.1 Literature Review

Reviews of relevant literatures such as textbooks journals articles, reference papers and the Internet are conducted to investigate the principles/theories of the various approaches, techniques and tools that were employed in different areas pertinent to this research work. Furthermore, features of thesauri and approaches to automatic thesaurus construction are studied. Besides, some graduate levels of locally researched thesis are also reviewed.

1.4.2 Dataset preparation

To develop automatic thesaurus from Afaan Oromo text, a standard and representative document corpus aught to be selected. The dataset must be selected in order to perform exhaustive evaluation of the performance of thesaurus. Accordingly, the data is collected from different sources, from VOA 11000 words, from OBN (Oromia Broadcasting service) 16000 words and the rest is collected from Gospel Go, written in Afan Oromo. To this end, experts having comprehensive domain knowledge of the language are consulted.

1.4.3 Development Tools and Techniques

In this research, Python programming language is employed this work especially for data preprocessing and generating the result because it has the capability to easily perform a search and replace operation over a large number of text files with enhanced error checking mechanism. Python is also suitable for much larger application and problem domain with its high-level built-in data types like flexible arrays and dictionaries.

1.4.4 Evaluation

After developing the system for automatic thesaurus generation, the implementation phase evaluates the performance level of the system and the validity of dataset employed. The evaluation process also includes the performance of the stemming algorithm and stopwords compilation procedures, as both procedures have a relevant consequence on the overall performance of the system by providing terms with reduced lexicon for thesaurus term entry. The evaluation is performed by taking a random sample of 20 terms from stemmed candidate terms i'e content bearing terms (descriptors) and the result of the evaluation is indicated as a percentage of the sample testset.

1.5 Scope and Limitation of the Study

There are several types of thesaurus construction methods suggested by different researchers for different languages, and the most available thesaurus is domain dependent. However, the scope of this research is to construct an automatic similarity thesaurus for Afan Oromo documents based on term clustering approach from term co-occurrence value in the document collection. The thesaurus is aimed to include some stemming and stopwords elimination and be domain independent because the data covers different domain areas like politics, sport spiritual, and economics.

The lack of a freely available electronic corpora, standard stemmer and complexities of Afaan Oromo orthography were composed some of the limitation to this research. Furthermore, thesaurus construction needs a huge text documents, and no well-organized

text documents of Afaan Oromo which fits for this work is handy. The research needs to collect them from here and there from different sources, which is cumbersome. In this research, only affixations (specifically prefix and suffix) are going to be discussed, infix and other morphologies are out of the scope of this research.

1.6 Significance of the Study

Development of language is directly related with adaptation of ongoing technology and making it suitable to the local languages. Speakers of languages are of no more interest in speaking language of technology, rather they need technology speak their own language. If the language manages to grow with technology, speakers of the specific language will be benefited from the development in many ways. That is why it is important to localize the works already done in developed languages like English (Eggi, 2012).

Automatic thesaurus construction from Afaan Oromo text is a framework setup for possible inception of enhanced Afaan Oromo text retrieval system. It would also be a baseline for future research on cross-language retrieval system as thesaurus provides a means for handling multilingualism, in the case of Ethiopia, through which Oromo people and the neighboring zone with similar dialect could satisfy their information need.

The features that cross-language retrieval systems provide is not solely applicable for language with similar dialects but could handle more diversified categories of language groups (Semitic, Cushitic, Omotic, and Nilo Saharan). Cross language retrieval is a retrieval of any objects (text, image, products, etc.) that has indexing in target language while the query is formulated in source language. The source and the target language can assume any number and type of languages. However, query can be formulated by selecting from a menu presented in the source language (Soergel, 1995).

Thesaurus in information retrieval is a key component of the IR system and is a collection of concepts which are more or less important for the subject of the document collection. Most of the existing search engine in WWW works well but there are cases where irrelevant documents are retrieved due to word mismatch (Xu, 1997). The problem is due to the fact that information retrieval system compares query term and index term at lexical level. In

effect, this research work could be considered as an input for the inevitable retrieval system in resolving query-term mismatch with respect to Afaan Oromo language.

Generally, the study has the following contributions:

- ✓ To enhance Afaan Oromo Information Retrieval search engines.
- ✓ To ease the interface among Afaan Oromo language users with content developers
- ✓ To use for other researchers as a framework for further research

1.7 Organization of the Thesis

The research work in this thesis is organized into five chapters. The first chapter is an introductoin part of the thesis which, defines the problem domain and justifies the relevance of the research. In this part, the basic concept about Thesaurus Construction, what initiated the study-- the statement of the problem, objective of the study, scope of the study, research methodologies and study ‘s significance are discussed. And the rest of the chapters are organized as discussed. Chapter two focused on literature review in which the two main concepts, related works and conceptual reviews are discussed. Conceptual reviews are some highlighted concepts of thesaurus, automatic thesaurus construction and the details of Afaan Oromo and its related topics whereas related works deals with works done so far on the research related to the topic. Furthermore, the chapter presents some of the important works undertaken so far in the language in order to have the clear notion of Afaan Oromo writing system and word formation. In Chapter three, the major techniques and methods used in this study are discussed which also includes the architecture of the system involved for developing automatic association thesaurus. The fourth Chapter discusses the experiment and the result, as well as the evaluation of the performance achieved through experimentation of the study. Finally, Chapter five presents the conclusion drawn from the research work, findings, the challenge faced and the recommendation of the areas for further research works.

CHAPTER TWO

LITERATER REVIEW

Through the literate review, the subsequent issues are reviewed in detail including the detail of Afaan Oromo writing system, definition, purpose and use of thesaurus, features of automatic thesaurus, approaches to automatic thesaurus construction, vocabulary construction, stemming, conceptual models, and similarity computation.

2.1 Overview of Thesauri

According to (Lassi, 2002), a thesaurus (plural: thesauri) is a valuable tool in Information Retrieval (IR), both in the indexing and searching process, used as a controlled vocabulary for expanding or altering queries. Accordingly, a number of researchers have conducted thesaurus research on different languages and discussed on its uses, features and approaches of thesaurus. Scholars (Karen, S. and Peter, W. 1997) defined thesaurus in two ways – The first is in terms of its function, a thesaurus is a terminological control tool used in translating from the natural language of documents, indexers or users (searchers) into a more constrained system language (documentation language, information language). The second is in terms of structure- a thesaurus is a controlled and dynamic vocabulary of semantically related terms which covers a specific domain of knowledge. A thesaurus is a book that lists words in groups of synonyms and related concepts (Caplan, n.d.). This means that when we look up a word in a thesaurus, we don't see a definition, but lists of similar words. Words that have similar meanings are called synonyms. Sometimes, thesaurus also provides us antonyms, or words with opposite meanings.

A thesaurus is a controlled vocabulary that shows relations (e.g. semantic) between terms, which can aid searchers in finding related terms to expand queries. It is a structured collection of concepts and terms for the purpose of improving the retrieval of information.

A thesaurus should help the searcher to find preferable search terms, whether the terms be descriptors (keywords) from a controlled vocabulary or the multiple terms needed for a comprehensive free-text search all the different terms that are used in texts to express the search concept (Soergel, 1995).

Most thesauri establish a controlled vocabulary, a standardized terminology, in which each concept is represented by one term, a descriptor that is used in indexing and can thus be used with confidence in searching. In such a system the thesaurus must support the indexer in identifying all descriptors that should be assigned to a document or other object in light of the questions that are likely to be asked (Soergel, 1995).

A well-constructed thesaurus has been recognized as a valuable tool in the effective operation of an information retrieval system as expansion of queries with related terms using thesaurus can improve performance (Imran & Sharan, 2009). A good thesaurus provides guidance to the indexers, through its hierarchy better terms by associative relationships between concepts. It is a semantic road map and guidance for searchers, indexers and anybody interested in an orderly grasp of a domain specific field. User-oriented indexing (or user-oriented indexing), which is one approach of indexing, indicates that the concepts to be included in the thesaurus are collected from actual and expected search requests. These are then organized into an easily grasped structure that serves as a framework or checklist for the indexer in analyzing objects or documents. The users inform the thesaurus builder on what they are interested in and the thesaurus builder organizes these interests into a logical framework that communicates user interests to the indexer. The indexer can now consider these interests in analyzing documents, making sure that an object or document will be assigned to all descriptors under which a user may want to find them. Request-oriented indexing requires a well-structured thesaurus; it depends on the semantic road map provided by the thesaurus. Request-oriented indexing starts with a hierarchical display, using the alphabetical display only for augmentation (Soergel, 1995)

A thesaurus can be referred to as a networked collection of terms where all terms are connected to each other and not only assists user in finding information but also in understanding it. Thesauri are tools that allow both the indexer and the researcher to use the same terms to describe the same subjects or concepts, allowing for easier search and

retrieval of information about a particular domain (International Organization for Standardization 2011).

Thesaurus can be generated manually or automatically where; manual thesaurus construction is both an art and a science. Though, some form of manual thesaurus construction is mandatory due to the relational complexities, semantic ambiguities, and dynamics, inherent in languages, manual construction and maintenance is complex and time consuming(Schneider, 2005). It is prone to contain errors and is hardly ever consistent. Furthermore, it must be kept up-to-date continually if it is to be of any use. A carefully crafted thesaurus can improve the effectiveness of an information retrieval system considerably and is therefore an important component. It can aid the researcher in formulating his queries more effectively and provide disambiguation of problematic terms. A thesaurus constructed automatically from a document collection is useful for two reasons:The first is that it can include implicit knowledge about the domain contained in the documents; and the second is, it does not suffer from the problems of manually constructed thesauri (Kitsch, 2008).

Most thesauri that a user encounters were manually constructed by domain experts and/or experts at document description. Manual thesaurus construction is a time-consuming and quite expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus make choices that affect the structure of the thesaurus. There is a need for methods of automatically construct thesauri, which besides, from the improvements in time and cost aspects, can result in more objective thesauri that are easier to update (Lassi, 2002).

As an alternative to the cost of a manual thesaurus, we could attempt to generate a thesaurus automatically by analyzing a collection of documents. There are two main approaches. One is simply to exploit word co-occurrence. We say that words co-occurring in a document or paragraph are likely to be in some sense similar or related in meaning, and simply count text statistics to find the most similar words. The other approach is to use a shallow grammatical analysis of the text and to exploit grammatical relations or grammatical dependencies. In this case, simply using word co-occurrence is more robust (it cannot be misled by parser errors), but using grammatical relations is more accurate. However, the

major problem with the manual thesaurus is the highly conceptual, knowledge-intensive task and therefore also expensive to build and hard to update in the timely manner. Consequently, thesauruses that can be easily constructed and maintained with minimum human aids or constructed automatically are in great demand ("Chaiwat Ketsuwan , Nattakan Pengphon , Asanee Kawtrakul, n.d.). Automatic thesauruses were produced by processing corpora, with similarity between words measured (directly or indirectly) by clustering terms depending on their co-occurrence.

2.2 Types of Thesaurus

2.2.1 Global Vs. Local Thesaurus

Global: In this approach, the collected document is preprocessed and those terms in the document is going to be clustered to get the thesaurus class and thesaurus classes are constructed based on word co-occurrence and their relationship in the corpus as a whole and these classes are used to index both documents and queries thus it can be concluded that a global thesaurus is constructed prior to the indexing process whereas,

Local: Thesaurus uses information obtained from the top ranked documents retrieved in response to a particular query and it can be concluded that a local thesaurus is constructed dynamically during query processing and uses information retrieved in response to a specific query to modify only that query. Although the global analysis techniques are relatively robust, the corpus-wide statistical analysis consumes a considerable amount of computing resources. Moreover, since it focuses only on the document side and does not take in to account the query side, global analysis only provides a partial solution to the word mismatch problem.

2.2.2 Manual vs. Automatic thesaurus

The approaches for constructing thesaurus can be broadly classified into four groups: General purpose Hand-crafted thesaurus (Manual thesaurus), Co-occurrence based automatically constructed thesaurus, Similarity based automatically constructed thesaurus, Head Modifier based automatically constructed thesaurus. Construction of manual thesaurusisverylaborintensive work and it lacks domain specific terms as it generally

covers general terms. Handcrafted thesaurus describes the synonymous relationship between words. a manually constructed network of lexical relationships, and founds that expansion helps only for very short queries, high cost and long duration, restrictions of manual analysis of the large text corpus. Therefore, there is a need of automatically generated thesaurus. Automatically constructed thesaurus has shown improvements in retrieval performance. They are based on the co-occurrence information, linguistic information and relevance judgment information.

In Co-occurrence based automatically constructed thesaurus, similarity between terms are first calculated based on association hypothesis and then used to classify terms by selectingasimilarity threshold value. In this way the set of index term is divided into classes of similar term. The query is then expanded by adding the terms of classes that containqueryterm. Suchstrategies are based on local clustering of terms. A similarity based automatically constructed thesaurus is built considering the term to term relationship rather than simple co-occurrence data. Terms for expansion are selected based on similarity to the whole query rather than their similarity to individual term.

In Head-modifier based automatically constructed thesaurus term relations are gathered on the basis of linguistic relations. Specifically, manually constructed thesaurus structure is characterized as Hierarchy of thesaurus terms, high level of coordination, many types of relations between terms, complex normalization, and field limits are specified by the creator plus its goal is that to precisely define the vocabulary to be used in a technical field, due to this is its useful in document indexing, assistance in developing search strategy and the importance in retrieval through query expansion/construction. Finally, it is verified as soundness and coverage of concept classification. Whereas an automatically constructed thesaurus is specified as: structurally many different approaches can be considered, most of the time it is not always hierarchical, because of the difficulties of phrase selection, and the coordination level is lower, normalization rule is simple but it is hard to separate homographs (i'e two words those spelled in the same way but differ in meaning), and field limits are specified by the collection. Finally, its goal is that, depending on the level of coordination, it can be used for indexing, mainly used to assist in retrieval through

(possibly automated) query expansion/construction and is verified as ability to improve retrieval performance.

2.3 Application of Thesaurus

Improving a user's ability to find the information that they are looking for quickly and easily is the main goal of most thesauri and other controlled vocabularies and the ISO standard mentioned in its outline on how they do this. Thesauri are tools that allow both the indexer and the researcher to use the same terms to describe the same subjects or concepts, allowing for easier search and retrieval of information about a particular domain (International Organization for Standardization 2011). By doing so, they support the indexing, retrieval, organisation and navigation of information. The relationships in some thesaurus guides users to more general or more specific concepts by allowing them to navigate through the vocabulary and to choose the most suitable terms for their content. This navigability of the thesaurus makes it much more useful than a simple controlled list of terms as it allows a user to browse a subject domain or website the thesaurus can be displayed alphabetically by terms or instead be used as a systematic structure of hierarchical or classified relationships to act as a navigational tool and map of the domain. Associative relationships in a thesaurus can also direct the users towards related terms making connections that they may not have been previously considered. In cataloguing, a thesaurus can also be used as a source of metadata for subject cataloguing as it can connect different objects together and improve discovery of and access to materials by exploiting all of the above features.

Thesauruses also have their place in business. Significant time and money is lost when employees spend time searching for content on an intranet and cannot find it quickly or easily. Indeed, in some cases they cannot find the information they are looking for at all. It is a tool for vocabulary control. Usually it is designed for indexing and searching in a specific subject area. By guiding indexers and searchers about which terms to use, it can help to improve the quality of retrieval. Thus, the primary purpose of thesaurus is identified as promotion of consistency in the indexing of documents and facilitating searching.

The attractive aspect of automatically constructing or extending lexical resource rest clearly on its time efficiency and effectiveness in contrast to the time-consuming and outdated publication of manually compiled lexicons. Its application mainly includes constructing domain-oriented thesauri for automatic keyword indexing and document classification in information retrieval, Question Answering, Word Sense Disambiguation, and Word Sense Introduction (Padmini Srinivasan, University of Iowa," n.d.)

Thesauri are vital and valuable tools in content discovery, and in information organisation and retrieval, activities common to all fields, including cultural heritage and higher education as well as business and enterprise. Thesauri allow information professionals to represent content in a consistent manner and enable researchers, employees and the public to find this content easily and quickly.

2.4 Features of Thesauri

In the following subsections, some of the most important features of both manual and automatic thesaurus are discussed.

2.4.1 Coordination level

Coordination is the construction of phrases from each available term. There are two types of coordination options; pre-coordination and post-coordination(Srinivasan, n.d.). A pre-coordinated thesaurus, which is more common in manually constructed thesauri, is a thesaurus that contains phrases and can be used for indexing and retrieval. However, in post-coordinated thesaurus, phrases are generated at searching. The pre-coordinated and post-coordinated thesaures have their own advantage and disadvantage. Pre-coordination is advantageous because of its precise vocabulary, with reduced ambiguity in indexing and in searching. Besides, the accepted phrases become part of the vocabulary. Nevertheless, it has a limitation on the need for prior knowledge of phrases construction rules.

There is also an intermediate level coordination in manually constructed thesaurus that combines both phrase and a single term. However, in this coordination level, there appears to be a significant variety in terms of coordination level. Thesaurus may have two, three or large size phrases. As a result, it is impossible to assert that two thesauri are similar because

they are pre-coordinated and without coordination level information. The level of coordination indicates the level the precision of the vocabulary and increasing in the number of relationships to be encoded. But result in larger vocabulary size. Post-coordination is advantageous in that the ordering of the words in a phrase is not significant i'e users need not worry about the exact ordering of the words in phrase whereas its disadvantegous is that, since phrase combination can take place during searching, the serach precision may fall. From this it can be concluded as automatic thesaurus construction usually implies the postcoordination.

2.4.2 Term Relationships

Term relationships describe the semantic association between terms in the vocabulary. There are three types of term relationships (Aitchison, Gilchrist, & Bawden, n.d.): Equivalent relationships, hierarchical relationships, and non-hierarchical relationships. Equivalence relations include both synonymous and quasi-synonymous terms. Quasi-synonymous are terms that possess significant semantic overlap and are considered to be synonymous for retrieval performance. For example, consider "genetics" and "heredity" with significant overlap in meaning. A hierarchical relation can be described as genus-species such as "dog" and "german shepherd". Consider for example the malaria parasite with genus name Plasmodium has four species names vivax, malaria, falciparum and ovalae. Nonhierachical relationships also identify conceptually related terms and they can be described as: "Thing-Part "relationship such as "House" and "roof"; thing-attribute such as "Book" and "Author."

2.4.3 Number of Entries for Each Term

Having a single entry for each thesaurus terms is seldom achieved due to homograph-words, which are words with multiple meaning. It is usually important that a thesaurus term should have a single entry. Yet the presence of homograph will make automatic thesaurus construction more complex, as each semantic of a homograph requires a unique representation (entry). In manually constructed thesaurus, the problem of such type was resolved by using parenthetical qualifiers. Consider a homograph bond. The semantics can contextually be deciphered as; bonds (chemical) and bonds (adhesive).

2.4.4 Specificity of Vocabulary

Specificity of a thesaurus vocabulary can be described as a function of the precision of individual terms in the vocabulary. High specific vocabulary expresses the subject in great depth and detail and promotes precision in IR. The limitation with highly specific vocabulary is that, the size of the vocabulary grows since a large number of terms are required to cover the concepts in the domain and it requires regular maintenance as specific terms usually tend to change more rapidly than general terms.

2.4.5 Control on Term Frequency of Class Members

Salton and McGill (1983, 77-78) have stated that in order to maintain a good match between documents and queries, it is necessary to ensure that terms included in the same thesaurus class have roughly equal frequencies(Srinivasan, n.d.). The basic idea here is that, statistical thesaurus constructed by partitioning the vocabulary into a set of classes with equivalent terms in which the same class should be equally specific. The specificity across classes should also be the same. This means, if a term in a vocabulary completely expresses the subject then the vocabulary is said to be highly specific and enhances retrieval precision. In order to achieve the best query- document match, it is necessary that terms in each thesaurus classes should have equal frequency and the total frequency of each class should also roughly be equal (Grossman & Frieder, 1983). Such features have a direct influence on the probability of document-query match to be uniform across classes.

2.4.6 Normalization of Vocabulary

Normalization of vocabulary is a set of procedures applied on vocabulary terms to convert variant forms to their basic expression so as to achieve consistency to the vocabulary. Normalization in automatic thesaurus construction is simple as compared to manual one and involves stopword removal and stemming. Nevertheless, in case of manual thesaurus construction, normalization imposes a set of constant rules that are applied to the structure of individual terms thereby guide the form of thesaurus entry. The rules include that the term should be in noun form, noun phrases should avoid preposition unless commonly known, number of adjectives should be limited, singularity of terms, and order of terms

within a phrase, spelling, capitalization, transliteration, abbreviations, initials, acronyms and punctuations.

2.5 Approaches to Automatic Thesaurus Construction

Thesaurus can build all types of relationship that exist between words, such as hierarchic, synonym, and morphological relationships(Rahman, Bakar, & Sembok, 2010). It is a set of concepts in which each concept is represented with at least synonymous terms, broader concepts, narrower concepts, and related concepts (Abuzir & Vandamme, 2017). In other words, thesaurus is a kind of dictionary that defines semantic relationship among words. Automated thesaurus dictionary construction (machine-understandable) is one of the most difficult issues, though its effectiveness is widely proved by different research areas such as natural language processing (NLP) and information retrieval (IR). A number of contributors have spent much time to construct high quality thesaurus dictionaries in the past. However, since it is difficult to maintain such huge scale thesauri, they do not support new concepts in most cases. Therefore, a large number of studies have been made on automated thesaurus construction based on NLP. However, issues due to complexity of natural language, for instance the ambiguous/synonym term problems, still remain. There is still a need for an effective method to construct a high-quality thesaurus automatically avoiding these problems (Nakayama, Hara, & Nishio, 2007). In general, building manual thesauri requires a lot of human labor from linguists or domain experts and these are expensive to build and there is high demand of automatic thesaurus construction.

Automatic thesaurus construction is an extensively studied area in information retrieval. The original motivation behind an automatic thesaurus construction is to find an economic alternative to manual thesauri. Almost all automatic thesauri are based on the so-called association hypothesis, which states that words related in a corpus (collection of documents) tend to co-occur in a corpus (Xu, 1997). Researches in automatic thesaurus construction have swung from co-occurrence analysis approach to a number of other functional approaches that have significantly enhanced the experiments on retrieval effectiveness in early nineties(Grefenstette, 1993). The following subsections briefly discuss some of the recent functional approaches to automatic thesaurus construction.

2.5.1 Concept space

Concept space is a finite weighted unidirectional graph whose node represents concepts and whose weighted graph edge represents co-occurrence in some domain. The goal is to create a meaningful and understandable concept space (a network of terms and weighted association). Concept space represent the concepts (terms) and their association for underlying information space (Frank, n.d.).

The concept space approach for automatic thesaurus generation was first developed by Chen, Ng, Martinez, and Schatz, (Dorbin et al., 2017). Accordingly, a concept space is defined as a network of terms and weighed associations which can represent concepts (terms) and their association for underlying space which represent the documents in the database. The steps that are involved in the concept space approach include document and object list collection, object filtering and automatic indexing, co-occurrences analysis and associative retrieval.

2.5.2 Document Clustering

(Crouch & Yang, 1992) constructed automatic thesaurus by document clustering based on low frequency model of Salton et al (Grefenstette, 1993). Nevertheless, they indicated that the low frequent method is not worth employed for small documents collection. Crouch and Yang achieved significant improvement in retrieval effectiveness using complete link clustering algorithm to cluster documents collection for generating thesaurus automatically from low frequent terms obtained in the clustered documents.

2.5.3 Syntactic Analysis

Syntactical analysis is an alternative approach to address the limitation of co-occurrence analysis in that, co-occurrence analysis: (i) considers co-occurring terms to be similar despite the distance between them in each document and (ii) overlooks similar terms from different documents. Syntactic analysis approach was first proposed by Grefenstette (Schutze & Pedersen, 1997), with the idea that words found in the same contexts have

greater chance to be related semantically. Using syntactic analysis, word similarities are derived from the overlapping of all the contexts associated with words.(Srinivasan, n.d.)

2.5.4 Lexical Co-occurrence

Lexical co-occurrence is the co-occurrence of two terms in a limited window of size k. Lexical co-occurrence first demonstrated by Schutze and Pederson (Schütze, 1998), with the idea that if terms often co-occurred close to each other then they would have greater chance to be related than co-occurring in the same document as in the case of co-occurrence analysis. (Srinivasan, n.d.)

2.5.5 Bayesians Network

Bayesian network is made up of a graph and tables of conditional-probability distributions. The graph structure takes the form of a directed graph where a node represents a variable or a proposition domain. Each node produces an output value at any moment according to the conditional probability distribution maintained in the node. Usually, the human expert is responsible for the construction of the graph and the tables (Soergel, 1995).

2.6 Construction of Vocabulary

In Information Retrieval, often distinction is made between controlled and uncontrolled vocabularies. Uncontrolled vocabularies allow for every token in a document to be a potential index term, without considering word form and other linguistic features. Controlled vocabularies on the other hand, have rules that regulate the words that are allowed to be index terms, as well as the word forms and other specific features of those terms. A thesaurus is a controlled vocabulary that shows relations (e.g. semantic) between terms, which can aid searchers in finding related terms to expand queries (Lassi, 2002).

The aim of vocabulary construction is to identify the most representative terms (words and phrases) for the thesaurus vocabulary from document collections. The first step is to identify an appropriate document collection. Then, it is to determine the required specificity for the thesaurus regarding which vocabulary terms are now ready for normalization. The simplest and most common normalization procedure is to eliminate very trivial words such as prepositions and conjunctions. The next standard normalization

procedure is to stem the vocabulary. Stemming reduces each word into its root form. In vocabulary construction, the relevant aspect is term selection for the thesaurus from the document collections (Foo, Hui, Lim, & Hui, 2000).

Controlled Vocabulary (CV) construction by automatic or semi-automatic methods can be categorized into statistical and linguistic approaches. In the statistical approach, terms are extracted from a document by IDF (inverse document frequency). Adapted to the controlled vocabulary construction problem, the assumption is that frequently co-occurring words with a text window (sentence, paragraph or whole text) point to some semantic cohesiveness. The co-occurrence approach needs human intervention before terms can be used for controlled vocabulary creations. In the linguistic approach, terms and their relations are based on the distributional context of syntactic unit (subject and object) and the grammatical surrounding function these unit. Suppose we have two terms “Electronic business" and “Electronic industry". These two terms can be semantically mapped. The substituted words are semantically close (i.e. business and industry) (Morshed & Sini, 2009).

2.7 Stemming

Stemming is a computational process of removing inflectional and derivational affixes and returning a word base, not necessarily a real word. The main difference to morphological normalizing is that normalizing turns the word to lexical full base form. The negative effect of stemming and normalization as processes in information retrieval is that they may produce noise as unrelated word forms are sometimes conflated to a single form (Hedlund, 2003).

Natural language texts have variations in word forms. The reasons for such variants include requirements of grammar usage, antonyms, transliteration, abbreviation, and spelling errors. In natural language text, the main source of word variation is morphology, natural language text with suffixing and prefixing being the most common ways of creating a word variant. Both inflectional and derivational morphologies can result in very large numbers of variants for a single word depending on the morphological complexity of a language. Morphological complexity can have a strong impact on the effectiveness of information

retrieval (IR) systems. Therefore, there is a need for automated procedures that can reduce the size of a lexicon to a controllable level, and also capture the strong relationships between different word forms (morphology). Morphological analysis, the identification of a word-stem from a full word form, helps in conflating semantically related words to the same form (Alemayehu & Willett, 2003).

Stemming is used for reducing different morphological variants of a word into a common form. It is widely used in IR, with the assumption that morphological variants represent similar meaning. It is applied during indexing and is used to reduce the vocabulary size, and it is used during query processing in order to ensure similar representation as that of the document collection. For morphologically fewer complex languages like English or Swedish, this usually involves removal of suffixes. For languages like Amharic or Arabic, that have a much richer morphology, this process also involves dealing with prefixes, infixes and derivatives in addition to the suffixes (Argaw & Asker, 2007). Likewise, Afaan Oromo has a rich morphology but it deals with prefixes, suffixes, and derivatives.

2.8 Measures of association

Cluster is a set of entities which are alike, whereas entities from different cluster are not alike and clustering can be defined as grouping of a number of similar things. Organizing a data in to sensible is suggested to be the most fundamental modes of understanding and learning them because cluster is the combination of a number of similar objects collected or grouped together. There should be some means of quantifying the degree of association between clusters in order to cluster an item.

2.8.1 Similarity Measures

Similarity computation deals with deriving a relationship between pairs of terms to identify the statistical associations between terms. There are a number of similarity measures available in the literature. The Dice, Jaccard, Inner and cosine coefficients have the attractions of simplicity and normalization and have often been used in for clustering the documents. The first, Cosine calculates the number of documents associated with both terms divided by the square root of the product of the number of documents associated with the first term and the number of documents associated with the second. And the second

computes the number of documents associated with both terms divided by the sum of the number of documents associated with one term and the number associated with the other.

Further, comparison between any two documents (or between a query and a document) can subsequently be determined by the distance between vectors in a high dimensional space. Zero distance between words indicates similarity and the most common similarity measure is the cosine coefficient. Cosine coefficient defines the similarity between two documents by the cosine of the angle between their two vectors. It resembles the normalized inner product of the two vectors, means inner product divided by the products of the vector lengths (square root of the sums of squares*)* (Börner, Chen, & Boyack, 2005). The information structure that we call a similarity thesaurus is a matrix that consists of term-term similarities. A similarity thesaurus is based on how the terms of the probabilities of the documents representing the meaning of the terms (Lairungraung, 2003). Word meaning can be regarded as a function of word distribution within different contexts in the form of co-occurrence frequencies, where similar words share similar contexts. Word similarity depends on to what extent they are interchangeable across different context settings. The flexibility of one word or phrase substituting another indicates its extent to be synonymous provided that the alternation of meaning in discourse is acceptable (Yang & Powers, 2008). Many researchers have used term co-occurrence in IR to identify semantic relationships that exist among terms. To classify relevant and non-relevant documents query terms are valuable, and then their associated terms will also be useful, and can be added to the original query. A number of coefficients have been used to calculate the degree of relationship between two terms.

Dice coefficient:

Compute the number of documents associated with both terms divided by the sum of the number of documents associated with one term and the number of documents associated with the other (Hanandeh, 2013).

If the binary term weights are used, the Dice coefficient reduces to:

$S_{Di, Dj} = \frac{2C}{A+B}$ where C is the number of terms that D_i and D_j have in common, whereas, *A* and *B* are the number of terms in D_i and D_j.

Jaccard coefficient:

The Jaccard index, also known as the Intersection Over Union and the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets.

***Cosine coefficient*:**

Compute the number of documents associated with both terms divided by the square root of the product of the number of documents associated with the first term and the number of documents associated with the second term. The table 1 below shows the summery of the similarity Measures.

Table 1: Similarity Measures

Similarity Measure	Evaluation for Binary Term Vector	Evaluation for Weighted Term Vector
Cosine	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert^{1/2} \bullet \lvert q \rvert^{1/}}$	$sim\ (d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$
Dice	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert}$	$sim(d_j, q) = \frac{2\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2}$
Jaccard	$sim(d,q) = \frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert - \lvert d \cap q}$	$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2 - \sum_{i=1}^{t} w_{i,j} \times w_{i,}}$
Inner	$\lvert d_i \cap q_k \rvert$	$Sim = \sum_{k=1}^{t} (d_{ik} \bullet q_k)$

2.8.2 The similarity Matrix

It is usual that clustering methods depends on a pairwise coupling of the most similar documents or clusters,so,that the similarity between every pair of points must be known.This necessitates the calculation of the similarity matrix;when the similarity measure is symmetric (Sij = Sji),the lower triangular matrix is sufficient.It always assumed that the similarity measure function used to compute the similarity, returns a value zero (0), if term-term,document-document,term-document or document-cluster have no term in common and one (1) value, if they have terms in common i'e Only those documents/cluster

pairs sharing at least one common term will have their similarity calculated ;the remaining values in the similarity matrix are set to 0.

Figure 1: Similarity matrix

$$S=\begin{vmatrix} S_{21} & & & & & \\ S_{31} & S_{32} & & & & \\ S_{41} & S_{42} & S_{43} & & & \\ . & . & . & . & & \\ . & . & . & . & & \\ . & . & . & & . & \\ S_{N1} & S_{N2} & S_{N3}. & . & . & S_{N(N-1)} \end{vmatrix}$$

2.9 Related Works

Depending on the attempts made to review the available literatures, the researchers concluded that no work has been done on the specified title for Afaan Oromo, i. e., there is no automatic Afaan Oromo thesaurus construction. In the following section we present the works done so far in the area of Thesaurus which explore basic tasks that includes activities on different languages.

2.9.1 Amharic thesaurus

Andargachew, M. 2009, developed an automatic thesaurus generation for Amharic Text retrieval using word space model and evaluated its performance using indexing and searching system.

The endeavor was to develop thesaurus automatically for text retrieval to the development of effective and efficient Amharic retrieval system. The developed system for automatic thesaurus generation for Amharic text retrieval is domain independent. This means the system can be trained and used in different domains without any modification as long as the documents collection has a UTF-8 Unicode encoding. His experimentation of the system has made on Amharic Bible documents. In the research, WORD SPACE model is used for the automatic thesaurus generation system. The WORD SPACE model was derived from the inverted file index by using random projection algorithm for dimensionality reduction and nearest clustering algorithm to generate thesaurus from the

word space model developed. The Thesaurus developed by (Andargachew M.,2009) registers 58% accuracy and has an outstanding feature in its efficiency in terms of its response time. It takes only 12 minutes and 8 seconds to create 20,315 term vectors in the WORDSPACE. It was integrated to an IR system for query expansion in order to further investigate its applicability Amharic information retrieval. The two most frequent and basic measures for information retrieval effectiveness (precision and recall) were used before and after using the thesaurus for query expansion. Stemming and stop word elimination components of the system played vital role in the thesaurus generation and were evaluated.(Mekonnen, 2009) discussed related works to his work, features and approaches of thesaurus construction. In addition, he discussed Amharic scripts and the Ethiopic writing system with its Unicode representation in computer; ambiguities and inconsistencies in Amharic writing system. A corpus based automatic thesaurus generation approach was used to construct thesaurus for Amharic document collection. To develop the work, he used Java programming language for the programming task and the semantic vectors API for developing WORD SPACE model for document collections.

The researcher concluded that thesauri are among the components of IR systems and plays a significant role for enhancement of recall. Manual thesaurus construction is time consuming, labor intensive and suffers low coverage. It is costly, requires highly skilled experts in a subject domain, highly conceptual and knowledge intensive task. To solve these problems automatic, generation of thesaurus is required.

2.9.2 Tigrigna Thesaurus

An automatic thesaurus for Tigrigna text retrieval from document collection based on co-occurrence approach, which achieved a remarkable output from heterogeneous documents, was developed(Hiete, 2011). Here, the main tasks performed in the architecture he used to develop automatic Tigrigna thesaurus included, preprocessing of Tigrigna text, vocabulary construction, index term weighing, co-occurrence metrics formulation, and used the cosine similarity measure function for computating the similarity. He went through the preprocessing stages including transliteration scheme (the Tigrigna needed to be translated to Latin and the translation was taken place using the system for Ethiopic representation in

ASCII (SERA) because the converted Latin scripts were convenient for stemming and internal processing of documents.), tokenization, removal of stop words and punctuations, stemming, and normalization. Then what followed were tokenization, stop words removal and numbers removal. Since Tigrigna words exist in different format, the task of fully collecting and removing stop words were difficult.

Then after transliteration of the Tigrigna script into alphabetic using SERA transliteration scheme, stemming process was conducted. The developed stemmer removed prefixes and suffixes only. Other morphologies were not considered. As Tigrigna language is characterized by words of varies usages, there are no rules to use those varying words. Through normalization, all words with varying forms but having the same meaning were converted to the common form. Through constructing vocabulary, the main criteria was sizable and representative of the subject area and the determination of the required specificity for the thesaurus Kanyarat (cited by (Hiete, 2011). Finally, he identified the index term and then developed co-occurrence metrics which was useful for similarity construction.

To evaluate the accuracy of thesaurus generation system, the evaluation scheme used was checking the thesaurus term whether properly stemmed or not. In this scenario, the top five and ten similarity generated terms per term was used as a threshold. If more terms were generated, the top five were selected for easy management. Otherwise top ten were selected. The result showed that 75.28% of the output related the concepts to the respective terms. The remaining 24.72% identified not to have related terms.

2.9.3 Wolaytta Thesaurus

(Beldados, 2013) developed an automatic thesaurus generation from Wolaytta Text, implementing on New Testament Bible corpora of the Wolaytta language. Here he first collected the original corpora collected from the web which was in XML format with different HTML tags and all numbers associated with each article were removed as a basic pre-processing step so as to convert the XML file in to a plain text format for further due pre-processing. Through Wolaytta thesaurus generation he went through some stages including the development of co-occurrence metrics which meant after text operation steps

were performed on the collected corpora, the most frequent terms, and descriptor, were generated for the thesaurus term entry. For each term in the entry, their co-occurring terms also were generated and the co-occurrence information obtained were kept in term-to-term co-occurrence matrices employed in the system.

To implement the system, he considered Python 3.1.2 programming language as language of choice by defining the cases and developed the system prototype, which was domain independent. Also, the developed system was based on a training dataset with extensive morphology and word distribution of Wolaytta language. Evaluation of Stemming and Stop words: He used the Lemma algorithm and conducted the evaluation and, in a way, that 20 terms from the stemmed test data set were selected by simple random sampling so that the performance of the stemming evaluated by checking whether terms were stemmed to their root form properly or not. Simultaneously, stemmed terms were inspected whether they belong to stop words list or not in order to access the accuracy and exhaustiveness of the stop words compilation process. Also, he concluded that 45 % of words were stop word that made 55% accuracy of the stop word compilation process and continuing, this evaluation is conducted in two phases.

Evaluation of Thesaurus Terms: Here in the prior, he considered 20 terms, accordingly, the result of the experiment indicated that 75% of terms were identified to have related concepts, on the other hand 25% terms did not relate to each other. Hence, the level of performance achieved in generating thesaurus term from Wolaytta text was 75%.

2.9.4 Arabic Monolingual thesaurus

In Arabic language, automatic thesauri had been designed and built using term-to-term similarity and association techniques (Khafajeh et al., 2002). The thesauri can be used in any special field or domain to improve the expansion process and to get more relevant documents for user's query using Arabic language. The researchers used the definition for thesaurus as in Merriam Webster Dictionary definition. According to this definition, a Thesaurus is a book of words or of information about a particular field or set of concepts, especially, a book of words and their synonyms. Here, a Thesaurus is also defined as a list of subject-headings or descriptors usually with a cross-reference system to be used in the

organization of a collection of documents for reference and retrieval. Thesaurus contains creative synonyms and related phrases that allow authors to enhance their vocabulary.

They used three phases in their thesaurus development. The first phase was about document preparation that meant removing stop words, tokenization (deleting punctuation marks, commas, special signs, and numbers), normalization and stemming. Elimination of stop words reduced the size of the indexing structure and thus increases the performance of the system and enabled it to retrieve more relevant documents. As reported the researchers used the stemming algorithm of (Smeaton, A.F., et al., 1983), with a little bit modification and they Use Vector Space Model to put the text of documents and the query in vectors.

In the second phase, they selected index terms. Terms selected for index term wefe terms which were repeated two to seven times in the text. This indicated that the researchers ignored the terms that appeared in most documents in the collection (i.e., have high frequencies), and the terms that appeared only once in a document (i.e., terms that have low frequencies). Their expectation was the use of a controlled vocabulary that could lead to an improvement in retrieval performance, creating the inverted file based on the stemmed words of each document. The stemmed words technique which they used was suffix-prefix removal. They used C# and an SQL database to implement their work, by computing term frequency (tf), inverted document frequency (idf), and tf*idf. In the third phase, to build the thesaurus they made comparison among the laws (Inner product, Cosine, Jaccard or Dice) used in finding 'Term Similarity'/'Term relationship' between the different terms. As a result, Cosine equation was used in building the similarity thesaurus because of its commonality. The computation of term frequency (tf), inverted document frequency (idf), and tf*idf was conducted. In addition, they calculated the weight or tf*idf of each term in a document by multiplying the normalized term frequencies with inverse document frequencies. After these steps, the inverted file that contained index terms (i.e., words) and terms frequency and the weight of each term in a document was retrieved.

The researchers used 242 documents that were presented in the Saudi Arabian National Computer Conference, for each run 59 queries entered automatically. They designed and built an automatic information retrieval system from scratch to handle Arabic text. To achieve this goal, they constructed an automatic stemmed words and full word index using

inverted file technique. Depending on these indexing words, the researchers built three information retrieval systems. Information retrieval systems were developed using a term frequency-inverse document frequency (tf*idf) for index term weights. In addition, the researchers used Similarity Thesaurus by using Vector Space Model with four similarity-measurements (Cosine, Dice, Inner product and Jaccard) and compared between the similarity measurements to find out the best that was used in building the Similarity thesaurus. At last, the researchers used an Association thesaurus by Applying Fuzzy model for index term weights. They implemented their system in C# language, and Runs on IBM/PCs and compatible microcomputer. The results they get were analyzed using the Recall and Precision criteria by applying 59 queries.

As the researchers' explanation, their experiment showed that the Jaccard and Dice similarity measurements are the same for the VSM model, while the Cosine and Inner similarity measurements are the same as well, but they are a little bit better than Jaccard and Dice measures. Their study pointed that the best case for development of information retrieval system was using association thesaurus with stemmed words.

2.9.5 English monolingual thesaurus

(Medelyan & Witten, 2006) and (Dongqiang, Y. and David, M. P. 2008) developed English thesaurus and (Medelyan & Witten, 2006) developed a domain specific thesaurus based automatic "Key phrase" Indexing. "Key phrases" represents a brief but precise summary of documents. The domain of the study used on UN Food and Agricultural Organization (FAO) by randomly downloading 200 full text documents from www.fao.org/documents/ for the training and evaluation material. There are key phrase extraction and term assigning existing approaches. In key phrase extraction, the phrases occurring in the document are analyzed to identify apparently significant terms, on the basis of properties such as frequency and length. In term assignment, term assignments are chosen from a controlled vocabulary of terms, and documents are classified according to their content into classes that correspond to units of the vocabulary.

As reported, (Olena, M. Ian, H. 2006), used key phrase indexing, an intermediate approach between key phrase extraction and term assignment that combines the advantages of both

and avoids their limitation. Documents in the collection were preprocessed. That means, white-space and punctuation were used to segment each document into individual tokens; Elimination of stop words, stemming remaining terms and sorting them into alphabetical order are conducted. For semantic term conflation, noninformative terms were replaced by their equivalent representative terms and for each training documents, candidate terms were identified and their feature value were calculated. Four features turned out to be useful in their experiments. These were the TF*IDF score, the position of the first occurrence of a phrase, the length of a candidate phrase in words and the node degree or the number of thesaurus links that connected the term to other candidate phrases. Index terms were assigned to the documents by professional indexers. Three semantic relations bi-directional links between related terms (RT) and Inverse links between broader terms (BT) and narrower terms (NT) were defined. (Yang & Powers, 2008) developed an automatic thesaurus for English language. As ground of their automatic thesaurus construction, distributional similarity was often calculated in the high dimensional vector space model (VSM). They proposed to first categorize contexts in terms of grammatical relations, and then overlapped the top n similar words yielded in each context to generate automatic thesauri. As they explained in their report, the hypothesis was that word meaning could be regarded as a function of word distribution within different contexts in the form of co-occurrence frequencies, where similar words shared similar contexts.

The researchers first employed an English syntactic parser based on Link Grammar to construct a syntactically constrained VSM to automate thesauri construction and the word space consisted of four major syntactic dependency sets that were widely adopted in the current English language research on distributional similarity. After parsing 100 million-words from British National Corpus (BNC) and filtering out non-content words and morphology analysis, they separately extracted the relationships to construct four parallel matrixes or co-occurrence sets. The dependency sets were RV: verbs with all verb-modifying adverbs and the head nouns in the prepositional phrases; AN: nouns with noun-modifiers including adjective use and pre/post-modification; SV: grammatical subjects and their predicates; VO: predicates and their objects. Following the reduction of dimensionality on the dependency sets, they created the latent semantic representation of

words through which distributional similarity could be measured so that thesaurus items could be retrieved. The cosine similarity to compute similarity of word vectors was employed by (Yang & Powers, 2008)

The cosine of the angle θ between vectors x and y in the *n*-dimensional space is computed using equation (2.1).

$$\text{Cos}\theta=\frac{\mathbf{x}.\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}=\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.1)}$$

, where the lengths of x and y are |x| and |y| respectively.

As the researcher's explanation, semantic similarity is often regarded as a special case of semantic relatedness, while semantic relatedness also contains word association.

Distributional similarity consists of both semantic similarity and word association between a seed word and candidate words in its thesaurus items, except for the 'noisy' words (due to the parsing or statistical errors) that hold no possible relationships with the seed.

2.9.6 Chinese monolingual thesaurus

(Foo et al., 2000) developed an automatic Chinese thesaurus that can be used to provide related terms to users' queries to enhance retrieval effectiveness. The thesaurus was developed by computing the co-occurrence values between domain specific terms found in a document collection. Term frequency and document frequency of terms wis used to compute the co-occurrence values. Since, Chinese texts has no word delimiters, an extra word processing called word segmentation was used. Because there were no papers in English that detailed the construction of Chinese thesauri, the automatic thesaurus was developed by examining existing automatic thesaurus generation techniques and adapted them to the Chinese language. The system was trained using an economics domain; this was not because the thesaurus was domain specific rather to derive a generic process for an automatic thesaurus generation that could be applied to another subject domain. The researchers used the Economics Terminology Dictionary, to identify and extract terms from the full text of each document found in their corpus. At the time of implementation, the researchers did not develope Chinese IR system from the scratch, rather they modified an existing English language-based IR system, called the mg (managing gigabytes) to

support the Chinese IR and automatic thesaurus generation. The corpus used in the research compromise three months Economics news (documents).

Finally, to measure the effectiveness of the system, set of queries were used. A total of 30 queries and the relevant documents for each query were derived and identified after manual reading through the complete corpus. Results obtained from the experiments conducted, ensured the automatic generated thesaurus was able to improve the retrieval effectiveness of a Chinese IR system.

2.10 Summery of reviewed Papers

Under this chapter we presented the revision of a number of papers which focuses on Automatic thesaurus construction which are done on different languages like Arabic, English, Chines, Amharic, Tigrigna, and Wolaytta languages. Our study is also related to these works specifically focused on Afan Oromo language. While reviewing these related works, we specially focused on the approaches used, corpus size, development languages and the performance of their works. From these works we have observed that most works done are depends on the co-occurrence approach. They used different size of corpus and the performance of their evaluation registered different results. This Automatic Thesaurus construction is new concept for Afan Oromo and the language is resource scarce and taking this problem in to account and considering some concepts what we obtained from reviewed of the related we proposed the Thesaurus construction based on the Bayesian network approach. The overall summery of each work is presented in the table 2.

Table 2: Riviewed papers summery

Paper	Title	Approach	Corpus
Andargachew M. 2009	Automatic thesaurus construction for Amharic text retrieval	word space model	21,312 words
Hiete, 2011	Automatic thesaurus construction for Tigrigna text retrieval	Co-occurrence	24,401 words
Demewoz B. 2013	Automatic thesaurus construction from Wolaytta text	Co-occurrence	25,232 words
Khafajeh et al., 2002	Building Arabic Automatic Thesaurus Using Co-occurrence Technique	Co-occurrence	242 documents
(Medelyan & Witten, 2006),Dongqiang, Y. and David, M. P. 2008	English Automatic Thesaurus	Co-occurrence	811 documents
Foo et al., 2000	Automatic thesaurus for enhanced Chinese text retrieval	Co-occurrence	

2.11 Over view of Afan Oromo

The Oromo folks have dominantly inhabited in Oromia region where Afan Oromo is the mother tongue. Afan Oromo is an Afroasiatic language and is the most widely spoken language in the family's Cushitic branch. Forms of Oromo are spoken as the first language by more than 30 million Oromo and neighbouring people in Ethiopia and by an additional half million in parts of northern and eastern Kenya. It is also spoken by smaller number of emigrants in other African countries such as:-SOUTH Africa, Libya, Egypt and Sudan. Oromo is a dialect continuum; not all varieties are mutually intelligible. Afan Oromo is the third largest language in Africa following Kiswahili and Hausa; 4th largest language, if Arabic is counted as an African language(Wikipedia, 2018). Taking into consideration the number of speakers and the geographic area it covers, Afaan Oromo probably rates second among the African indigenous languages. Perhaps, not less than two million non-Oromo speak Afan Oromo as a second language.

Currently Afan Oromo is the official working language of the regional state of Oromia (the largest regional state in Ethiopia) as it is being used as a working language in offices and

educational institutes for all non-language subjects in junior-secondary schools (1-8 grades). At the country level, in Ethiopia, out of the total 22 public universities, 8 are offering degree programs majoring in Afan Oromo. Addis Ababa University is offering Afan Oromo at Master's degree level (Kejela, 2010). Afan Oromo is also widely used as both written and spoken language in some neighbouring countries, including Kenya and Somalia.

2.11.1 Varieties

According to Ethnologies (2015), the Oromo macrolanguage was divided into four languages as listed below:

Southern Oromo Language:-Southern Oromo, or Borana (after one of its dialects), is a variety of Oromo spoken in southern Ethiopia and northern Kenya by the Borana people. *Günther Schlee* also notes that it is the native language of a number of related peoples, such as the Sakuye. Dialects are Boran proper (Borana, Borena), possibly Arsi (Arusi, Arusi) and Guji (Gujji, Jemjem) in Ethiopia and, in Kenya, these are Karayu, Salale (Selale), Gabra (Gabbra, Gebra) and possibly Orma and Waata. The language is locally and commonly known as *afaan borana* ("Borana language").

Eastern Oromo Language:- Eastern Oromo (also known as "Ittu Oromo") is a dialect of the Oromo language and is spoken in the Hararghe Zone, and northern Bale Zone of the Oromia Region of Ethiopia.

Orma Language: - Orma is a variety of Oromo spoken by the Orma people in Kenya and may be a dialect of Southern Oromo.

West-central Oromo: - Western Oromo and Central Oromo, including Mecha/Wollega, Raya, Wallo (Kemise), Tulema/Shewa. There are strong similarities among these four varieties but there are also differences among them and these four varieties depend on geographical area. There are many synonym words which make effectiveness of IR system lower (www.mylanguages.org/learn_oromo.php)

2.11.2 Writing System, Alphabets and Sounds of Afaan Oromo

Until the 1970s, Afan Oromo was written with either the Ge 'ez script or the Latin alphabet. Oromo is written with a Latin alphabet called *Qubee* which was formally adopted in 1991

and since 1991 Latin alphabet is used as official alphabet of Oromo Language(Wikipedia, 2018). Various versions of Latin-based orthography had been used previously, mostly by Oromos outside of Ethiopia and by the Oromo Liberation Front (OLF) by the late 1970s (Heine 1986). With the adoption of Qubee, it is believed that more texts were written in the Oromo language between 1991 and 1997 than in the previous 100 years. In Kenya, the Borana and Waata also use Roman letters but with different systems.

This writing system was adapted largely from the fact that its characters do explicitly represent the vowels and the consonants of a language (Gamta, 2000). The "Qubee" writing system has a total of 33 letters that consists of all the 26 English letters with an addition of 7 combined consonant letters. All the vowels in English are also vowels in "Qubee". Vowels have two natures in the language that can result in different meanings. The natures are short and long vowels. A vowel is said to be short if it is one. If it is two, which is the maximum, then it is called long vowel, e.g., the words lafa (ground) and laafaa (soft). The rest of "Qubee' are consonants. The combined consonant letters are known as "qubee dacha" and they are ch, DH, sh, NY, ts, ph and zy. Alphabets sound is also included with how it is pronounced in English words (Simon Ager, Oromo language…online). Audio based sound is also available online and generally includes upper case, lower case and their sounds as summarized in the table 3.

Table 3: Upper Case, lower case and their sounds

Alphabets	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***
A a	*[aa]* like **ask**	**B b**	*[baa]* like **bi**rd	**C c**	*[Caa]* like **ca**t	**D d**	*[daa]* like **da**m	**E e**	*[ee]* like **a**te	**F f**	*[ef]* like **fu**ngi
G g	*[gaa]* like **gu**n	**H h**	*[haa]* like **ha**t	**I i**	*[ie]* like **I**ndia	**J j**	*[jaa]* like **Ju**st	**K k**	*[kaa]* like **Ca**st	**L l**	*[la]* like **li**fe
M m	*[ma]* like **ma**n	**N n**	*[naa]* like **na**sty	**O o**	*[oo]* like **o**ld	**P p**	*[pee]* like **pa**st	**Q q**	*[quu]* like **quit**	**R r**	*[ra]* like **ra**t
S s	*[saa]* like **sa**lad	**T t**	*[taa]* like to**ta**l	**U u**	*[uu]* like **u**rge	**V v**	*[vau]* like **va**ry	**W w**	*[wee]* like **wa**nt	**X x**	*[taa]* like ___
Y y	*[y]* like **you**th	**Z z**	*[Zay]* like **Tha**t	**CH ch**	*[chaa]* like **cha**t	**DH dh**	*[dhaa]* like ___	**SH sh**	*[shaa]* like **sh**y	**NY ny**	*[nyaa]* like _____
PH ph	*[phaa]* like __										

A. Vowels-Dubbachiiftuu

Oromo has the typical Eastern Cushitic set of five short and five long (indicated in the orthography by doubling the five vowel letters) vowels **a, e, o, u, i** and **aa, ee, oo, uu** and **ii respectively**. The difference in length is contrastive; for example, *lafa* 'Earth', *laafaa* 'soft' and all vowels are pronounced in the similar way throughout every Afaan Oromo literature.

Table 4: Dubbachiistoota (Vowels)

	Front	***Central***	***Back***
Close	i /ɪ/, ii /iː/		u /ʊ/, uu /uː/
Mid	e /ɛ/, ee /eː/		o /ɔ/, oo /oː/
Open		a /ɐ/	aa /ɑː/

B. Consonants –Sagaleewwan (dubbifamtoota)

In afaan Oromo, consonant length can distinguish words from one another, for example, *steal* 'bad', *hattuu*' Thief'. The table below shows the summery of Afan Oromo consonants.

Consonants

Table 5: Dubbifamtoota (Consonants)

		Bilabial/ Labiodental	Alveolar/ Retroflex	Palato-alveolar/palatal	Velar	Glottal
Stops and Affricates	Voiceless	(P)	T	ch /tʃ/	K	' /ʔ/
	Voiced	(b)	D	j /dʒ/	g /g/	
	Ejective	Ph /P'/	X /t'/	c /tʃ'/	q /k'/	
	Implosive		dh /ɗ/			
Fricatives	Voiceless	F	S	sh /ʃ/		H
	Voiced	(V)	(Z)			
Nasals		M	N	ny /ɲ/		
Approximants		W	L	y /j/		
Rhotic			R			

2.11.3 Grammar

As other languages of the world, Afan Oromo have its own rules and syntax as well as grammar which is called Seer-Luga (in Afaan Oromo)

Gender: - Like other Afroasiatic languages, Oromo has two grammatical genders and unlike that of English there is no neutral gender in Afan Oromo. Except in some southern dialects, there is nothing in the form of most nouns that indicates their gender. A small number of nouns pairs for people, however, end in *-eessa* (m.) and *-eettii* (f.), as do adjectives, when they are used as noun: *obboleessa* 'brother', *obboleettii* 'sister', *dureessa* 'the rich one (m.)', *hiyyeettii* 'the poor one (f.)'. Grammatical gender normally agrees with biological gender for people and animals; thus, nouns such as *abbaa* 'father', *ilma* 'son', and *sangaa* 'ox' are masculine, while nouns such as *haadha* 'mother' and *intala* 'girl, daughter' are feminine. However, most

names for animals do not specify biological gender where as names of astronomical bodies are feminine: *aduu* 'sun', *urjii* 'star' and gender of other inanimate nouns varies somewhat among dialects(Wikipedia, 2018).

Number: Oromo cardinal numbers convey the "how many" as "counting numbers" because they show quantity. Accordingly, Oromo has singular and plural number, but nouns that refer to multiple entities are not obligatorily plural. That is, if the context is clear, a formally singular noun may refer to multiple entities: *Nama* "man" or "people", *Nama Shan* "five men" or "five people".

Noun plurals are formed through the addition of suffixes and when it is important to make the plurality of a referent clear, the plural form of a noun is used. The most commonplural suffix is *-oota*; a final vowel is dropped before the suffix, and in the westerndialects,the suffix becomes -ota following a syllable with a long vowel: *mana* 'house',*manoota* 'houses', *hiriyaa* 'friend', *hiriyoota* 'friends', *barsiisaa* 'teacher', *barsiiso(o)ta*'teachers'.Amongthe other common plural suffixes are *-(w)wan*, *-een*, and *-(a)an*;the latter two may cause a precedingconsonanttobedoubled:*waggaa*'year',*waggoota* 'years',*laga*'river',*laggeen*'rivers', *ilma* 'son', *ilmaan* 'sons'(Wikipedia, 2018).

Definiteness:- Unlike that of English, in Afan Oromo there is no indefinite articles (such as *a, an, some)* but (except in the southern dialects) it indicates definiteness (English *the*) with suffixes on the noun: *-(t)icha* for masculine nouns (the *ch* is geminated though this is not normally indicated in writing) and - *(t) ittii* for feminine nouns. Nouns ending with vowels are dropped before these suffixes: *karaa* 'road', *karicha* 'theroad', *nama* 'man', *namicha*/*namticha* 'theman', *haroo* 'lake', *harittii* 'the lake'. Note that for animate nouns that can take either gender, the definite suffixes may indicate the intended gender: *qaalluu* 'priest', *qaallicha* 'thepriest (m.)', *qallittii* 'the priest (f.)'. The definite suffixes appear to be used less often than *the*in English, and they seem not to co-occur with the plural suffixes(Wikipedia, 2018).

2.11.4 Pronouns

2.11.4.1 Personal pronouns

In most languages like Afan Oromo and English, there are a small number of basic distinctions of person; number, and often gender that play a role within the grammar of the

language. Oromo Pronouns include personal pronouns (refer to the persons speaking, the persons spoken to, or the persons or things spoken about), indefinite pronouns, relative pronouns (connect parts of sentences) and reciprocal or reflexive pronouns (in which the object of a verb is being acted on by verb's subject).

In all of these areas of the grammar—independent pronouns, possessive adjectives, possessive pronouns, and subject–verb agreement—Oromo distinguishes seven combinations of person, number, and gender. For first and second persons, there is a two-way distinction between singular ('I', ' your sg.') and plural ('we', 'you pl.'), whereas for third person, there is a two-way distinction in the singular ('he', 'she') and a single form for the plural ('they'). Because Oromo has only two genders, there is no pronoun corresponding to English *it*; the masculine or feminine pronoun is used according to the gender of the noun referred to.

The table 6 below gives forms of the personal pronouns in the different cases, as well as the possessive adjectives. For the first-person plural and third person singular feminine categories, there is considerable variation across dialects; only some of the possibilities are shown.

Table 6: Afan Oromo personal pronouns

English	**Base**	**Subject**	**Dative**	**Instrumental**	**Locative**	**Ablative**	**Possessive adjectives**
I	*ana, na*	*ani, an*	*naa, naaf natti*	*Naan*	*Natti*	*Narraa*	*Koo, kiyya [too, tiyya(f)]*
You(sg)	*Si*	*Ati*	*Sii, siif Sitti*	*Siin*	*Sitti*	*Sirraa*	*Kee [tee (f.)]*
He	*Isa*	*Inni*	*isaa, isaa(tiif)f, isatti*	*Isaatiin*	*Isatti*	*Isarraa*	*(i)isaa*
She	*isii, ishii, isee, ishee*	*ishiin, etc.*	*ishii, ishiif ishiitti, etc*	*ishiin.etc*	*ishiitti etc.*	*ishiirraa etc.*	*(i)sii, (i)shii*
We	*Nu*	*nuti, nu'i nuyi, nu*	*nuu, nuuf, nutti*	*Nuun*	*Nutti*	*Nurraa*	*Keenna, keenya [teenna, teenya (f.)]*
You(pl)	*Isin*	*Isini*	*isnii, isinitti, isniif*	*Isiniin*	*Isinitti*	*Isinirraa*	*Keessan(i) [teessan (i) (f.)]*
They	*Isaan*	*Isaani*	*isaanii, isaaniif, isaanitti*	*Isaaniitiin*	*Isaanitti*	*Isaanirraa*	*(i)saani*

Adjectives

Afan Oromo adjectives (see table 7 below) are words that describe or modify another person or thing in the sentence and they are very important because its structure is used in every day conversation.

Table 7: Afan Oromo Adjectives

Colors-Halluuwan	**Size-hamma guddinaa.**	**Shape-Boca**	**Tests**	**Qualities**
English-Afaan Oromo	***English-Afaan Oromo***	***English-Afaan Oromo***	***English-Afaan Oromo***	***English-Afaan Oromo***
Black – Gurraacha *Blue - Cuquliisa* *Brown - Bifa Bunaa* *Gary - Daalacha* *Green - Magariisa* *Red - Diimaa* *White - Adii* *Eg. black men-nama gurraacha.*	*Big – guddaa* *Deep - gad-fagoo* *Long-dheeraa(lafa)* *Narrow - dhiphaa* *Short - gabaabaa* *Small - xinnaa* *Tall - dheeraa* *Thick - yabbuu* *Thin – haphii*	*Circular-geengoo* *Straight - sirrii* *Square - golarfee* *Triangular-rogsadee* *E.g. circle face-fuula geengoo.*	*Bitter - hadhaawaa* *Fresh - asheeta* *Salty - soodiddaawaa* *Sour - Ogonaawaa* *Spicy -kaurgoo baayatu* *Sweet - mi'aawaa* *E.g. salty food-nyaata hadhaawaa.*	*Bad - hamaa* *Clean - qulqulluu* *Dark - dukkanaawaa* *Difficult - ulfaataa* *Dirty - xuraawaa* *Dry - gogaa* *Easy - salphaa* *Empty - duwwaa* *Expensive - Qaalii*

2.11.5 Adverbs-Ibsa xumuraa

These are part of speech and generally these are words that modify any part of language other than a noun. Accordingly, adverbs can modify verbs, adjectives (including numbers), clauses, sentences and other adverbs.

Four categories of adverbs are known in Afan Oromo: Adverb of time, adverbs of place, adverbs of manner and adverbs of frequency and these were listed below in table 8 from left to right in English and Afan Oromo respectively with equivalent meaning.

Table 8: Adverbs in Afan Oromo

Adverbs of time			
English	**Afaan Oromo**	**English**	**Afaan Oromo**
Yesterday	*kaleessa*	**Adverbs of manner**	
today	*harr'a*	very	*Baayyee/hedduu*
tomorrow	*bor*	quite	*Baayyee*
Now	*amma*	really	*Dhugumaan*
then	*gaafas*	fast	*Dafee/ariitiin*
later	*eger /egere*	well	*Gaarii*
tonight	*Edans/galgala kana*	hard	*Cimaa*
right now,	amma *isa amma*	quickly	*Dafee*
last night	*edans*	slowly	*Suuta*
this morning	*ganam kana*	carefully	*Qalbiidhan*
next week	*torban dhufu*	absolutely	*Matuma*
recently	*dhiyeenya kana*	together	*walii wajjin*
soon	*dhiyootti*	alone	*Qophaa*
immediately	*hatattamaan*	**Adverbs of frequency**	
Adverbs of place		Always	*yeroo hunda*
here	*as*	sometimes	*gaaffii gaaf*
there	*achi*	occasionally	*gaaffii gaaf*
over there	*gara sana*	seldom	*Darbee*
everywhere	*iddoo hunda*	rarely	*darbee darbee*
nowhere	*Eessayyu*	never	*Yoomiyyuu*
home	*Mana*		
away	*Fagoo*		
out	*Ala*		

2.11.6 Prepositions-wal qabsiiftuu/tota

Like that of English, Afan Oromo prepositions links nouns, pronouns and phrases to other words in a sentence and are written separately from root word. So, it is easy to remove from content bearing terms easily, as stop words. In some cases, propositions are connected with root words. Below in table 9 are some Afan Oromo propositions with their translation to English.

Table 9: Afan Oromo Prepositions

English prepositions	**Afaan Oromo Preposition**		**English Prepositions**	**Afaan Oromo Preposition**
About	*Waa'ee*		Since	*Ergi*
Above	*Gubbaa/gararraa*		Then	*Mannaa*
Across	*Gama*		Through	*Gidduu*
After	*Booddee/booda*		Till	*Hamma*
Against	*Faallaa*		To	*Tti*
Among	*Jara gidduu*		Toward	*Garas*
Around	*Naannoo*		Under	*Jala//gajjallaa*
As	*Akka*		Unlike	*Faallaa*
At	*Itti*		Until	*Hamma*
Before	*Dura*		Up	*Gubbaa*
Behind	*Dudduuba/dugda duuba*		Via	*Karaa*
Below	*Jala/gajjallaa*		With	*Wajjin*
Beneath	*Gajjallaa*		Within	*Keessatti*
Beside	*Bira*		Without	*Malee*
Between	*Gidduu*		***Two words***	***Jechoota lamee***
Beyond	*Garas*		According to	*Akka kanaatti*
But	*Garuu*		Because of	*Sababa kanaaf/kanaaf*
By	*…dhaan/ttiin*		Close to	*Bira*
Despite	*Ta'uyyuu*		Due to	*Kanaaf*
Down	*Lafa/gad*		Except for	*Kana malee*
During	*Utuu*		Far from	*Irraa fagaatee/irraa siqee*
Except	*Malee*		Inside of	*Keessa isaa*
For	*F*		Instead of	*Kana mannaa*
From	*Irraa*		Near to	*Itti dhiyaatee/bira*
In	*Keessa*		Next to	*Itti aanee*
Inside	*Keessa*		Outside of	*Kanaa alatti*
Into	*Keessatti*		Prior to	*Kanaan dura*
Near	*Bira*		***Three words***	***Jechoota sadii***
Next	*Itti-aansee/aanee*		As far as	*Hamma*
Of	*Kan*		As well as	*Akkasumas*
On	*Irra*		In addition to	*Dabalataanis*
Opposite	***Faallaa/fuullee***		In front of	*Fuullee isaa*
Out	*Ala*		In spite of	*Haa ta'uyyuu malee*
Outside	*Alaan*		On behalf of	*Maqaa/bakka…*
Over	*Irraan*		On top of	*Kana irraan*
Per	*…. Tti*		**Demonstrative prepositions**	***Agarsiisoo***
Plus	*Fi*		This	*Kana/tana*
Round	*Naannoo*		That	*Sana*
			These	*Warra kana/jara kana*

2.11.7 Negations

Afan Oromo *negation* is the process that turns an affirmative statement (He is tall) into its opposite denial (he is short). In most cases negation is formed by adding prefix '*hin'* to the verb.

Examples:

We don't speak - hin dubbannu	I don't drink - hin dhugu
They don't fight - hin lolan	we don't forget – hin irraanfannu
She doesn't swim - hin daaktu	He doesn't move – hin socho'u

In addition to the 33 symbols, it is recommended to know the following principles associated with "Qubee" (Gamta, 2000):

1. Two vowels in succession indicate that the vowel is long, e.g. bitaa (left);

2. Gemination (a doubling of a consonant) is phonemic in Afaan Oromo, e.g. damee (branch), dammee (sweety);

3. h is not geminated at all;

4. The same word can have two or more forms depending on its context, e.g. Nama kadhu (ask person), namaa kadhu (ask for person);

5. When it occurs at the end of a word, the single "a" is pronounced schwa (inverted e) whereas it is pronounced (delta) elsewhere;

6. Understandably, instead of diacritic signs, the combined letters are used so as to align them with typewriter characters.

Through the literature review, the following issues are dealt: Definition, purposes and uses of thesauri, features of automatic thesauri, some approaches to automatic thesaurus construction, construction of vocabulary, stemming, conceptual models, similarity computation, and an Afaan Oromo writing system.

CHAPTER THREE

METHODS AND APPROACH

3.1 The Blue print

Development of an automatic thesaurus construction involves various techniques, processes and algorithms. Generally, from the following architecture we can observe two concepts, namely the Information retrieval concept and the machine learnig concepts.

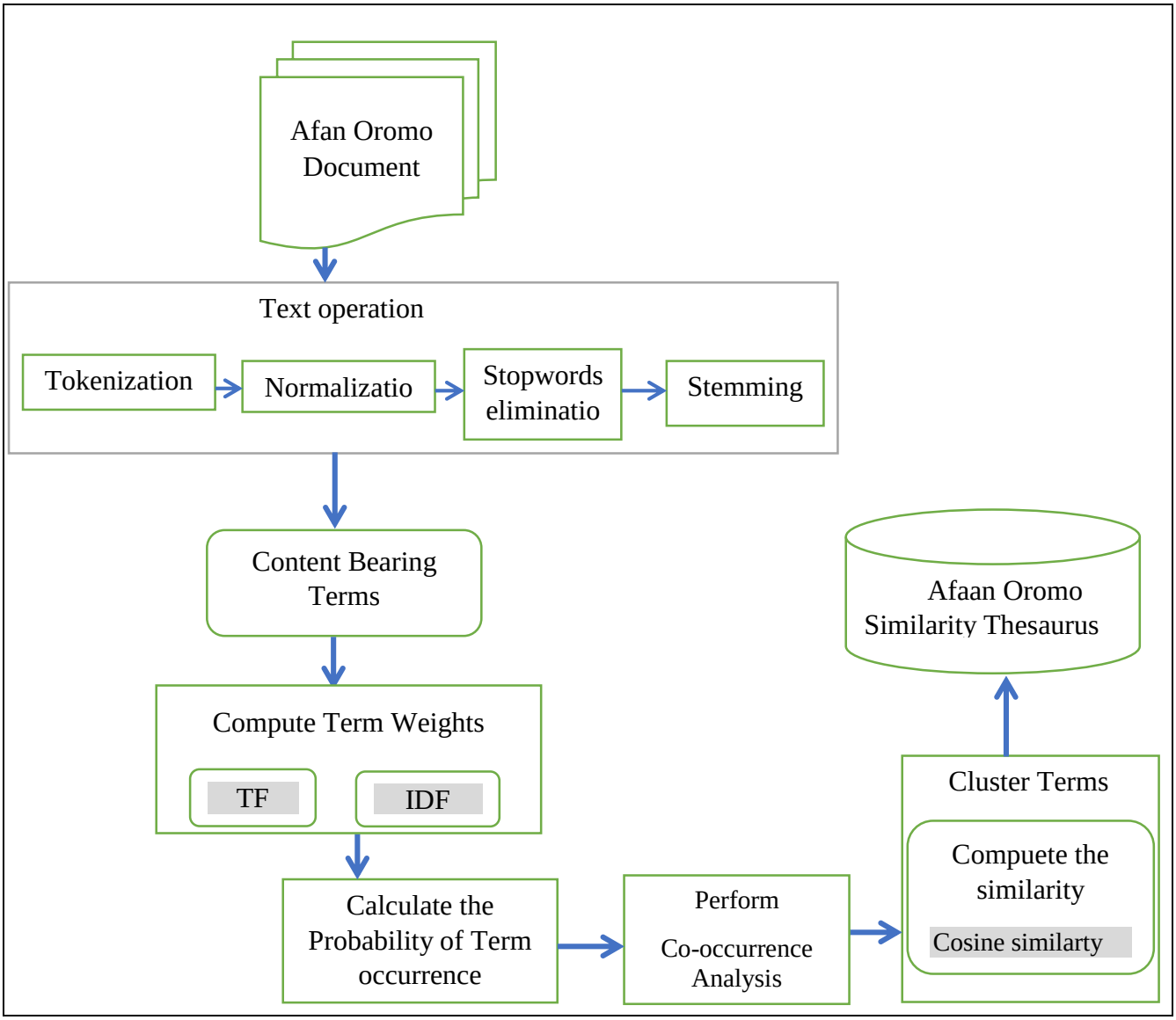


Figure 2: Automatic Afan Oromo Thesaurus construction System Architecture

In order to develop automatic thesaurus, the preprocessing action is takes place first and then the extraction of terms and their co-occurrences information should employ the flow model correlation analysis between the occurrence frequencies of terms in the document. Their co-occurrence frequencies are the co-occurrence matrix to identify pairs of associated terms that qualify for the entry of association thesaurus. lastly, after the co-occurrence analysis, terms are clustered i'e the class of the whole terms in the collection should be generated. Then after all, the degree of the similarity between each term should be calculated, incase the cosine simialarity is used. Figure 2 above showed that the flow and the processes take place to generate automatic Afaan Oromo thesaurus and the details for those procedures are discussed below.

3.2 Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another procedure and it includes case folding/normalization, stop word removal and stemming. It transforms the data into a format that will be more easily and effectively processed for the purpose of the user (Manning, Raghavan, & Schütze, 2009). Preprocessing raw input data is advantageous for saving processing time so that the full data after processing become logical representation of the data with the most representative tokens of data chosen whereas token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. Accordingly, since, preprocessing is the process of making the data ready for further processing and the collected documents are unstructured they are converted into appropriate way for thesaurus construction. It includes tokenization, removal of stopwords and punctuations, stemming, and normalization processes and this process helps to identify representative or descriptive terms which are manageable for thesaurus construction processes.

3.2.1 Tokenization

A simple strategy is to just split the text on all non-alphanumeric characters. As far as this step is language depending, some infprmation (special address, names etc) can be lost. Thus, advanced techniques are need for text tokenization, considering local habits.

Tokenization, in NLP can be defined as the task of splitting a stream of characters into words and it can be the first task in text operation step. Tokenization is the process of breaking a stream of text up in to words, phrases, symbols or other meaningful units. Or given a character sequence and a defined document unit, it is the task of chopping it up into meaningful units called tokens, perhaps at the same time throwing away certain characters, such as punctuations. In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics, where it is a form of text segmentation and in computational science where it forms part of lexical analysis (Adriani & Rijsbergen, 1999). In the tokenization phase of text operation, the major issue that requires due attention is what the correct token forms are to consider. Thus, the issue of tokenization is language dependent and requires prior knowledge of the language of the document. The commonest approach to tokenization is splitting text at its word borders. Hence the tokens refer to words of a language while the vocabulary refers to lexicon (morphemes). In this approach, two methods are paramount important. The first is splitting text using white space as a delimiter while removing special characters such as punctuation marks, hyphen, digits, letter case and parenthesis. The second and highly sophisticated method uses non-alphanumeric characters as a delimiter (Baroni & Bisi, 2004).

As depicted in figure above of system architecture, Afan Oromo documents are subjected to tokenization phase of test operations in which the stream of texts is splitted into sequence of meaningful unit called token. As the issue of cases should be resolved in tokenization phase, all words with upper cases of the alphabet are converted to lower cases. One challenge related to tokenization is--differentiating single word from compound word, e.g., *Hewlett-Packard* Versus *Hewlett and Packard, AddisAbaba* Versus *Addis Ababa, USA and*

USA. The other challenge is identifying numerical values like dates, phone numbers, and IP addresses. Additionally, Chinese and Japanese has no space between words, which makes difficult space and punctuation mark-based tokenization. Arabic and Hebrew are basically written right to left, but with certain items like numbers written left to right. The challenge here is thast it is not possible to use the same algorithm used for Latin based language for such languages too. That is why tokenization is called natural language dependent technique.

In case of Afan Oromo writing system, words are delimited by comma, white space and all the punctuation marks those are commonly used in English writing system. The delimiters such as full stop, comma, semicolon, single quotation, double quotation, exclamation marks, and question marks are also applicable in Afan Oromo writing system. For example, if the original document is "*Oromiyaan Jiituu fi Badhaatuu dha*", the tokens will be '*Oromiyaan', Jiituu', fi' badhaatuu', dha'*. So, in this work, tokenization is used for splitting documents into tokens. Detaching certain characters such as punctuation marks and the following algorithm enables us to implement on the given Afan Oromo texts documents.

```
#White space includes space, newlines and tab
# Punctuation includes comma, semicolon, question marks, and dots
  For file in corpus
    Define word delimiter to white_space
    Define punctuation marks
      Read files
       For file in read
     If there is white_space
    If there is punctuation_marks
  Put each term as separate token
  Remove from the corpus
```

Algorithm 1: Tokenization of Afan Oromo text document.

3.2.2 Normalization

In other words, the process of creating equivalence classes of terms and the goal is to map words with the same sense to the other class.

Normalization is the process by which we can perform certain transformations of text to make it reconcilable in a way which it may not have been before. Having broken up our documents (and also our query) into tokens, the easy case is that whether tokens in the query just match tokens in the token list of the document. However, there are many cases when two-character sequences are not quite the same but we would like a match to occur. Normalizing text means that converting it into convenient, standard form.

Normalization is the process by which terms are grouped in to equivalent classes. Each equivalent class is represented by the most frequently terms selected from each class and replace the occurrences of all other terms in their respective classes. Normalization is an important step of text operation through which best match between query and documents can be achieved, since it handles all the mapping of different tokens describing the same concepts to the same term. Thus, if normalization is performed both to the documents before indexing and to query before search, it yields increased number of relevant documents retrieved with enhanced recall of the system (Grefenstette, 1993).

There are different approaches to normalization and it is language dependent. However, this research has given the cardinal emphasis to stemming and elimination of stop words as an approach as well as part of preprocessing steps. Stemmer uses a set of rules to map terms to stem (Ricardo Baeza-Yates, n.d.). Even though, the stem from the stemmer does not necessarily correspond to a word, in most cases terms describing the same concepts might also be mapped to the same stem. Elimination of stopwords could be considered as a normalization approach, as it deletes frequent terms from the token stream while counting their frequency in the whole documents. As a result, the most frequent terms occur in all documents and are therefore not useful for discrimination between relevant and non-relevant documents.

The normalization is applicable when we need to convert characters with diacritical marks, change all letters case, decompose ligatures, or convert half-width katakana characters to full-width characters and so on. Also, normalization involves process of handling different writing system. Primarily every term in the document should be converted in to similar case format, in this study lowercase. For instance, ‗THESAURUS ‘, ‗Thesaurus ‘, ‗thesaurus’ are all normalized to understandable as lower case ‗thesaurus’ the system. Secondly punctuation marks in all the documents will be omitted. Punctuations marks in

Afan Oromo are almost similar to that of English except apostrophe (— _ ‖) which is considered as part of word in Afan Oromo. Some of punctuation marks are: ፡ “! | / ? @ # * ~ $ % ^ & ( ) { } < > [ ] _ + = - , ." ...\; - _ + £. Removal of punctuation marks helps to consider similar words with different punctuations mark, in similar way with no distinction of punctuation mark linked to it. For example, _information? ‘, _information! ‘, _information. ‘, _information ‘, and etc. all are represented as _information ‘. Numbers are also removed with punctuation marks. As the exceptional case the apostrophe (— _ ‖) is not included in as a part of punctuation marks to be removed. As normalization approach, both stemming and elimination of stop words are implemented on Afan Oromo text so that the rules of the stemming algorithm mappes terms into the same stem while elimination of stop words removes frequent terms from that training data set. The stemming algorithm mappes terms like, Barataa, Barattuu, Baratummaa, Barachuu, Barate, Baraticha, Barattoota, into Barat_. Afan Oromo language is morphologically rich language. This is the main reason to incorporate the stemming part in the thesaurus construction process of the system. In this process words are stemmed to the root word. In Afan Oromo writing system, affixes are added in many positions of the stem. The stemmer developed removes prefixes and suffixes only, other morphologies are not considered. Inconsistencies in alphabetical composition of words were also an issue that required a normalization step as some word pairs like; kitaaba with macaafa, kennuu with laachuu, etc. were written to mean the same thing. As a result, inconsistent words of such type with the same meaning manually are scanned and replaced with one of the two forms.

3.2.3 Elimination of Stop Words

Dropping common terms: stop-words- For every language we can easily identify most common words without any or only with small information value (and, is, be, in etc). By removing these words, we are able to significantly reduce the words’ space while in the most cases the information value of processed texts remains.

Some extremely common words that would appear to be of little value in helping selecting documents matching a user’s need are excluded from the vocabulary entirely. Stopwords are words that are among the most frequent words in a language and evenly distributed in

document and have been recognized since the earliest days of Information Retrieval (Evert, 2005). In computing, stopwords are words which are filtered out before or after processing of natural language data (text). Other search engines remove some of the most common words—including lexical words, such as "want"—from a query in order to improve performance. In any written text, there are frequent and more common words that have little or nothing to say about the text with no additional meaning to the actual context of the text. Such word groups are called stopwords and include conjunction, preposition, articles, and adverbs. Words of such type contained in a document are considered to be worthless as index terms.

Stopwords are non-content bearing words in a document. These are words which occur frequently and have less power in discriminating one document from another. According to Zipf's law, few terms occur frequently; a medium number of terms occurs with medium frequency and many terms with very low frequency. This shows that writers use limited vocabulary throughout the whole document, in which even fewer terms are used more frequently than others. Some are used for syntactic purpose; for instance, articles, prepositions, conjunctions, etc. (Manning et al., 2009). When searching for index terms that contain stopwords requires all the record contained in the database to be checked despite its irrelevancy, lots of processing time and working memory can be saved and does not damage the retrieval effectiveness if the words that do not contribute to the actual content of the document are removed. In information retrieval system, stopwords have little or no contribution for retrieval process. So, stopwords are normally removed from a document to facilitate effective representation of a document. Stopwords make up large portion of the text of most documents and if removed the percentage that the corpus is being decreased will be about 70-75% as compared to otherwise. Sopword elimination performed by filtering the term list using "stop-list" is also called as syncategorematic or non-content bearing words (Rapp, 2002). These terms should be removed. Since every language has its own list of stop words, list of stopwords for Afan Oromo are compiled from Afan Oromo literatures and some other sources. For example, (Kan, sun, ani, yookaan, akkasuma, booda, isaan, immoo) are some lists of Afan Oromo stopwords and they can occur in different formats.

The following algorithm 2 shows an algorithm used for removing stopwords.

```
#All stopwords should be included in stopwords list
#Stowords should be removed from the collection
    for stopw in stop list
        if stopw in collection then
            Remove stopw
        end if
    end for
```

Algorithm 2: Stopword removals

3.2.4 Stemming

Documents contain different words' forms and there are families of related words with similar meanings (car, cars, cars' etc). For the English language the most common stemmer is porter stemmer, wheras it can be complicated for other languages like Afan Oromo.

Stem is the portion of a word which is left after the removal of affixes (prefixes, infixes and suffixes) whereas stemming is the substitution of the words with their respective stems using different methods like Affix removal, Table lookup, Successor variety (determining the morpheme boundary), *N*-gram stemming are based on letters' bigram and trigram information.

The stopword list contains nonsignificant words that are removed from a document or a request before beginning the indexing process. The stemming procedure tries to remove inflectional and derivational suffixes in order to conflate word variants into the same stem or root. Stemming can be linguistic, automatic or mixed. Linguistic stemming uses a linguistic knowledge of the structure of the language, by providing manually compiled list of affixes, allotropy rule and so on. In such case, stemming is the process of reducing inflected or sometimes derived words to their stem, base and root form. However, automatic stemmer is built using algorithms and uses a statistical procedure such as frequency count, n-gram method and some combination of both. Even if there are different stemming algorithms, the most common one is that of Porter, called Porter Stemmer. A stemming algorithm is a process of linguistic normalization, in which variant forms of a

word are reduced to a common form with the purpose of improving the performance of information retrieval system. The basic difference is that the stem does not require being identical to the morphological root form of the word; instead it is usually only sufficient that related words are mapped to the same stem regardless of the validity of the stem. Hence key terms, query and documents are represented by stems instead of the original words. The fact that different variants of a stem can be conflated to a single number of distinct terms required for a set of documents to be represented thereby saves storage space and processing time (Jiang & Zhai, 2007).

Stemming is language-dependent process in similar way to other natural language processing techniques and often removes inflectional and derivational morphology, e.g., Oromiyaa, Oromiyaan, Oromiyaatti →*Oromiyaa*. Stemming has both advantages and disadvantages. The advantage is that it helps us to handle problems related to inflectional and derivational morphology which makes words with similar stem/root word to retrieve together and this increases effectiveness IR system, whereas the disadvantages are that some terms might be over stemmed which changes the meaning of the terms in the document and different terms might be reduced to the same stem which still enforce the system as to retrieve non-relevant documents (Eggi, 2012).

It is not possible to apply the stemmer developed for English or other languages like Porter's (Porter, 1980) to Afan Oromo due to differences in the patterns of word formations and differences in their morphologies. Some of the concepts from the Porter stemmer`s (Porter, 1980) are however adopted to develop a stemmer for Afan Oromo. Specifically, concepts about measure, arranging the rules in clusters, analyzing word formation based on the nature of their endings (for example words that attaches *–de* suffixes ends with b/g/d in Afan Oromo) are taken from Porter algorithm (Tesfaye, 2010).

According to (Tesfaye, 2010) suggestion the Afan Oromo stemmer is based on a series of steps that each removes a certain type of affix by the way of substitution rules. These rules only apply when certain conditions hold, for example, the resulting stem must have a certain minimal length. Most rules have a condition based on the so-called ***measure***. The measure is the number of vowel-consonant sequences (where consecutive vowels or

consonants are counted as one) which are present in the resulting stem. This condition must prevent that letters which look like a suffix but are just part of the stem will be removed. Other simple conditions on the stem are:

- ✓ *Does the stem end with a vowel?*
- ✓ *Does the stem end with a consonant?*
- ✓ *Does the stem end with specific character?*
- ✓ *Does the 1st syllabus of the stem repeated?*

Two versions of the algorithm were developed by (Tesfaye, 2010). The first version is totally rule based. As this stemmer is not exhaustive enough to include every rule, in the second version, statistics is used to complement the rule so as to handle cases that couldn`t be caught by the earlier. (Tesfaye, 2010) algorithm is rule-based Afaan Oromo Stemmer and it is porter stemmer-based algorithm for Afan Oromo text document. (Tesfaye, 2010)identified six stemmer rule clusters depending on the Afan Oromo language grammatical rule (Eggi, 2012).The first improved version of Afan Oromo stemming algorithm which was developed by (Tesfaye, 2010) was considered to be directly employed in the automatic thesaurus generation system. Following is the algorithm developed to conflate word variants for Afan Oromo text:

1. OPEN the collection
2. READ the next word to be stemmed
Read a word from the file until match occurs or End of File reached
3. If word matches with one of the rules
Remove the suffix and do the necessary adjustments
Go back to 3
ELSE
Go to 6
4. Return the word and RECORD it in stem dictionary
5. IF end of file not reached
Go to 1
ELSE
Stop processing
6. IF there is no applicable condition and action exist
Remove vowel and return the result
Go to 4

Algorithm 3: Stemming algorithm for Afan Oromo

3.3 The Content Bearing Terms

From their linguistic view of meaning, content bearing can be defined as words or expresions those are used to describe or identify something or about something else and these can be Texture, Colors, Smells, Testes, Tempratures, Sounds and Dimensins descriptors. Set of terms those are relevant to a certain domain of knowledge having different set of relationship between them are comprised in the thesaurus. Descriptors or the indexing terms are words or expressions which denote in unambiguous fashion the constituent concepts of the field covered by the thesaurus are the basic units of thesaurus not only this, also thesaurus consist and non-descriptors this are sometimes called the entry terms and they are words or expressions denoting the same or more or less equivalent concept as a descriptor in the language of the Thesaurus. Among different available semantic relations of thesaurus terms, three of them: Equivalent, Hierarchical and Associative are the most common.

Equivalent relationship covers different types of relationship such as genuine synonymy, near-synonymy, antonym and inclusion and this represented by abbreviations as "UF" (Used For) between the descriptor and the non-descriptor(s) it represents, and ''USE" between a non-descriptor and the descriptor which takes its place. And the *Hierarchical relationship* between descriptors is represented by the abbreviations: BT (Broader Term) and NT (Narrower Term) for specific and more generic descriptors whereas the ***associative relationship***, shown by the abbreviation ''RT" (Related Term), relates two descriptors that do not meet the criteria for equivalence nor a hierarchical relationship and this is suggesting another descriptor that would be helpful for the thesaurus user to search by. In this work we shall consider associative relationships.

3.4 Construction of Vocabulary

A thesaurus is a representation of keywords associated with a subject domain providing a precise and controlled vocabulary which serves to coordinate document indexing and document retrieval. In both indexing and retrieval, a thesaurus may be used to select the most appropriate terms. Additionally, the thesaurus can assist the searcher in reformulating search strategies if required. A thesaurus is a structured list of terms, usually related to a

particular domain of knowledge. The objective here is to identify the most informative terms (words and phrases) for the thesaurus vocabulary from document collections. The first step is to identify an appropriate document collection. The only loosely stated criteria are that the collection should be sizable and representative of the subject area. Terms can be selected from titles, abstracts, or even the full text of the documents if available. Therefore, it can be said that, vocabulary construction is the process of identifying the most representative or informative terms for the thesaurus vocabulary from document collection.

3.5 Index Term Weight

Indexing process, which can be performed manually (by humans-manual indexing, or by computer software-automatic indexing) is part of the document description process and concerns the content analysis and description. The objective of indexing is to assign such words to each document in a collection that the contents of the document are sufficiently disclosed by these words. Thus, these assigned words are called index terms since they are stored in the index and used in the matching process. All index terms are not equally important in representing and discriminating a document; the degree of the terms in representing a document is different from one to the other. It is thus, required to measure how important a term is with regard to representation and discrimination of a document. The most commonly used automatic indexing are based on weighting words according to their frequency of occurrence in documents, i.e., the better a term is considered to be at representing the content of a document, the higher weight it will be assigned.

Generally, Term Frequency (TF), Inverse Document Frequency (IDF) and a combination of the two, TF*IDF, are the most commonly used weighting techniques: Using TF, the optimal index terms are those that occur with a medium frequency in a document. The most frequent words are considered bad discriminators – they cannot discriminate one document from the rest of the collection. The least frequent words are considered too insignificant to be good content descriptors. While the TF technique considers each document by itself, the IDF technique takes the word occurrences in the entire collection into account. A word that occurs in all or many documents is not good at discriminating documents from each other. On the other hand, a word that occurs in few documents in the collection are considered

good index terms, since the word may clearly discriminate a few documents from the rest of the collection. When TF and IDF are combined into TF*IDF, both the occurrence of words in each document, and the occurrence of words in the document collection are considered. The result is high weights forwards that occur with medium frequency in an individual document and with low frequency in the collection(Lassi, 2002).

3.6 Term-term co-occurrence matrix for automatic thesaurus construction

Term co-occurrences analysis is one of the approaches used in information retrieval research for forming multi-phrase terms. Co-occurrence analysis is a statistical approach where the association of terms in documents, chapters or some other unit is computed. The basic idea is that the closer the words occur, the more significant in the co-occurrences. Many automatic indexing methods do not identify how close words occur, these just only consider if they co-occur in the same document.

Nevertheless, frequently co-occurring terms have little or no significance for enhancing retrieval performance; perhaps co-occurrences analysis should be performed to identify less frequent terms that co-occurred in a document (El-Hamdouchi, A., 2017). Thesaurus is able to disclose the conceptual knowledge of a document collection. The basic idea of the automatically constructed thesaurus is to calculate the similarities between two terms on the basis of their co-occurrence in a document collection (Song, Yang, Li, & Park, 2011). The similarity measure can be done based on different similarity measures, such as cosine, dice etc.

In automatic thesaurus construction, the entire document collection is analyzed and co-occurrence relationships between terms are used to build a matrix of term-term relationships. Usually, term-term matrices of this type contain weights which are a measure of how related one term is with another term. These matrices are large and computationally expensive to compute. The matrices are used to cluster terms based on their co-occurrence data in the hope that terms that are closer together in this term-space are synonymous. Conceptually, what is underlined is the role of documents and terms are interchanged in the retrieval model. In essence, documents become the features of the term. Thus, two terms that appear in the same document are indexed by a similar feature and are deemed to

have some type of synonymous relationship. Many formulae have been proposed to measure the association between two terms using co-occurrence data. The similarity between two terms k_i and k_i can be determined by evaluating the difference between the two-vector vector $k_i = (d_{i1}, d_{i2} ... d_{in})$ and vector $k_j = (d_{j1}, d_{j2} ... d_{jn})$ in the document vector space. A simple binary weighting on these document weights would lead to the following cosine formulation of similarity between two terms:

$$\text{Cos}\ (k_i,k_j)=\frac{df(k_i,k_j)}{\sqrt{df(k_i)df(k_j)}} \qquad \text{………………………………………….} (3.1)$$

where df (t_i, t_j) is the number of documents in which both k_i and k_j co-occur, df(k_i) is the number of documents in which k_i occurs and df (k_j) is the number of documents in which k_j occurs. There are many variations of such formulae which aim to accurately find the best synonyms for a term (Cummins & O'Riordan, 2005). While various techniques are in the literature, computing the co-occurrence values between domain-specific terms are found in a document collection appears to be the standard technique.

Automatic thesaurus from corpus is generated by computing the co-occurrence values between domain-specific terms found in a document collection. These co-occurrence values are derived from term frequency and document frequency of the terms (Angel F. Zazo, Carlos G. Figuerola, Jose L.A . Berrocal, Emilio Rodriguez, n.d.).

The key to automatic thesaurus generation represents each term as a vector. The terms are then compared using a similarity coefficient that measures the Euclidean distance, or angle between the two vectors. To form a thesaurus for a given term t, related terms for t are all those terms u such that sim (t, u) is above a given threshold (Grossman & Frieder, 1983).

Co-occurrence matrix is a matrix that describes how often or in how many documents terms occur together with other terms. This type of matrix may be called an association cluster. An association cluster is based on the frequency of co-occurrence of pairs of stems inside relevant documents that are retrieved. When simple association cluster of terms is generated from the document collection then the association matrix S (an association matrix) is said to be normalized. An alternative is to normalize the correlation factor.

$$S_{u,v}=\frac{C_{u,v}}{C_{u,u}+C_{v,v}-C_{u,v}} \quad \text{…….. (.3.2)}$$

For instance, if Cu,v is adopted, association matrix S is said to be normalized (Lairungraung, 2003). In co-occurrence extraction, if there are N different terms in the document, NXN co-occurrence matrices will be constructed by counting the frequency of pair wise term co-occurrences.

3.7 Documents and Term Clustering

Term clustering: - Term clustering sometimes corresponds to a statistical thesaurus i'e it is "dictionary" that provides for each word, not its definition, but its synonyms and antonyms. Term Clustering allows expanding searches with terms that are similar to terms mentioned by the query (increasing recall). Terms can be clustered manually or automatically: here the focus is given to an automatic term clustering.an automatic term clustering follows a principle like the more frequently two terms co-occur in the same documents, the more likely they are about the same concept.

Document clustering: -Here the main objective is that to minimize intra-cluster distance between documents, while maximizing inter-clustering distances (using appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of the document clustering. The large variety of documents makes almost impossible to create a general algorithm which can work best in case of all kinds of datasets. A document clustering allows expanding answers, by including *documents that are similar* to documents retrieved by a query (increasing recall)

We can consider a number of Common algorithms for term clustering, some of them are: Cliques, Connectedcomponents, Stars and Strings. In ***cluques*** algorithm, all terms in the clusters (thesaurus class) are required to be similar to *all* other terms and the ***Connected Components*** require all terms in a cluster (thesaurus class) to be similar to *at least* one other term. The following algorithm shaws that how the connected component algorithm works.

1. *Select a term not in a class and place it in a new class (If all terms are in classes, stop)*
2. *Place in that class all other terms that are similar to it*
3. *For each term placed in the class, repeat step 2*
4. *When no new terms are identified in Step 2, goto Step 1*

Algorithm 4: Connected component

There are several measures of cluster validity that have been used in some research for the purpose of accessing clustering methods. In cases where you have a dataset labeled with classes (supervised clustering) you can use precision and recall, or purity and entropy.

Purity of a cluster = the number of occurrences of the most frequent class / the size of the cluster (this should be high), ***Entropy of a cluster*** = a measure of how dispersed classes are with a cluster (this should be low); in cases where you don't have the class labels (unsupervised clustering), intra and inter similarity are good measures. ***Intra-cluster*** similarity for a single cluster = average cosine similarity of all pairs within a cluster (this should be high) and ***Inter-cluster*** similarity for a single cluster = average cosine sim of all items in one cluster compared to all items in every other cluster (this should be low).

CHAPTR FOUR

EXPERIMENT AND PERFORMANCE EVALUATION

4.1 Corpus collection

The problem behind this study is up-coming thesaurus generation by automatic approach from Afan Oromo text documents which satisfy in query expansion which means it provides the alternative term or word for searching. Corpus which is a collection of texts or speech can be stored in an electronic machine-readable format. Accordingly, Automatic thesaurus generation system from Afan Oromo text is mainly implemented on New Testament Bible corpora comprised of twenty-seven books and 260 documents with Afan Oromo those are obtained from a website with URL: http://gospelgo.com distributed by Gospel Go. And other documents were collected from different sources like VOA (Voice of America Afan Oromo service), OBN (Oromia Broadcasting Network) news and these articles involves different subjects like politics, education, culture, religion, history, sports, economy and other events. Since it covers concepts of various fields, the compiled corpus is said to be promising corpus for thesaurus construction. All of the documents are converted for so far pre-processing and saved as text files with a UTF-8encoding. A term is any preprocessed word within a document, and a document can be considered as a set of terms occurring in that document at least once. A total of 100 documents which have a total of 63434 words were collected and pre-processed so that stopwords and others like punctuations, numbers were removed from the corpora. The remaining documents with total of 36869 words are used to go through the work.

4.2 Thesaurus Generation

4.2.1 Calculating the probability of term occurrence

For the sake of ranking terms to identify the degree of importance of the temrs in the collection, the probability of these terms occurrence is calculated. Here the concept is that, the more the probability of the term occurred, the more important that term was.

The following figure 3 below Shaws the sample screenshot of the probability of the terms in the collection generated by the python program.

```
File  Edit  Shell  Debug  Options  Window  Help
Python 3.4.3 (v3.4.3:9b73f1c3e601, Feb 24 2015, 22:43:06) [MSC v.1600 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
>>> aangoo; 0.0011815675462780622
aare; 0.00019692792437967703
aarsaa; 0.0007877116975187081
aarus; 0.00019692792437967703
aasa; 0.00019692792437967703
aasaa; 0.00019692792437967703
abaabooww; 0.00019692792437967703
abba; 0.0027569909413154787
abbaa; 0.005513981882630957
abdii; 0.00019692792437967703
abiya; 0.00019692792437967703
abiyaa; 0.00019692792437967703
abiyuud; 0.00019692792437967703
abiyuudiin; 0.00019692792437967703
abjuudha; 0.0007877116975187081
abjuudhaanyoseefi; 0.00019692792437967703
abrahaam; 0.0005907837731390311
abrahaami; 0.00019692792437967703
abrahaamii; 0.00039385584875935406
achi; 0.0017723513194170934
achittis; 0.0005907837731390311
achuma; 0.0007877116975187081
adaba; 0.0007877116975187081
adda; 0.0005907837731390311
addaatiin; 0.00019692792437967703
addunyaa; 0.0005907837731390311
addunyaatii; 0.00039385584875935406
adeemanis; 0.00019692792437967703
adeeme; 0.00019692792437967703
adeemsa; 0.00019692792437967703
adeemu; 0.00019692792437967703
adeemus; 0.00019692792437967703
adii; 0.00039385584875935406
aduu; 0.00039385584875935406
aduun; 0.00019692792437967703
afuufamu; 0.00019692792437967703
afuufanii; 0.00019692792437967703
```

Figure 3: The probability of term occurrence

4.2.2 Term clustering

Virtually all techniques for automatic thesaurus generation are a process of term-classifying or term-clustering, which generate groups of related words by observing word co-occurrence patterns in the documents of a particular collection. here, the following terms are selected manually by considering: (i) terms those are properly stemmed (ii) their

correlation value i'e the more frequently two terms co-occur in the same items, the more likely they are about the same concept. (iii) The threshold value: choosen the threshold that determines if two terms are similar enough to be in the same class. Here the threshold values are 7 and 15, the terms having more co-occurrence than threshold value is represented as 1 and less than threshold value is represented as 0. The classes of terms are indicated by shading terms those belong to the same cluster.

	abbaa	dhugaa	gaarii	galiilaa	gooftaa	guddaa	guyyaa	harka	ilma	kiristoos	magaalaa	mana	muka	nama	samii	simoon	uffata	ulfina	waaqayyo	yesuus
abbaa	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
dhugaa	1	0	0	0	1	0	1	0	1	1	0	0	0	1	1	0	0	1	1	1
gaarii	1	0	0	0	1	0	1	1	1	0	0	1	0	1	1	0	0	0	0	1
galiilaa	1	0	0	0	0	0	1	1	1	1	0	1	0	1	0	0	0	0	0	1
gooftaa	1	1	1	0	0	1	1	1	1	0	0	1	0	1	1	0	1	1	1	1
guddaa	1	0	0	0	1	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1
guyyaa	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
harka	1	0	1	1	1	1	1	0	1	0	1	1	0	1	1	0	1	1	1	1
ilma	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
kiristoos	1	1	0	1	0	0	1	0	1	0	1	1	0	1	0	0	1	0	1	1
magaalaa	1	0	0	0	0	0	1	1	1	1	0	1	0	1	0	0	1	0	1	1
mana	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
muka	1	0	0	0	0	1	1	0	1	0	0	1	0	1	1	0	0	0	0	1
nama	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
samii	1	1	1	0	1	1	1	1	1	0	0	1	1		0	0	1	1	1	1
simoon	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	1
uffata	1	0	0	0	1	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1
ulfina	1	1	0	0	1	1	1	1	1	0	0	1	0	1	1	0	0	0	1	1
waaqayyo	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
yesuus	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Figure 4: The Cluster of terms

The various clustering techniques are easy to visualize using a graph view of the binary Term-Term matrix: The following figure 5 shows the network of terms considered to repreaent the clusters.

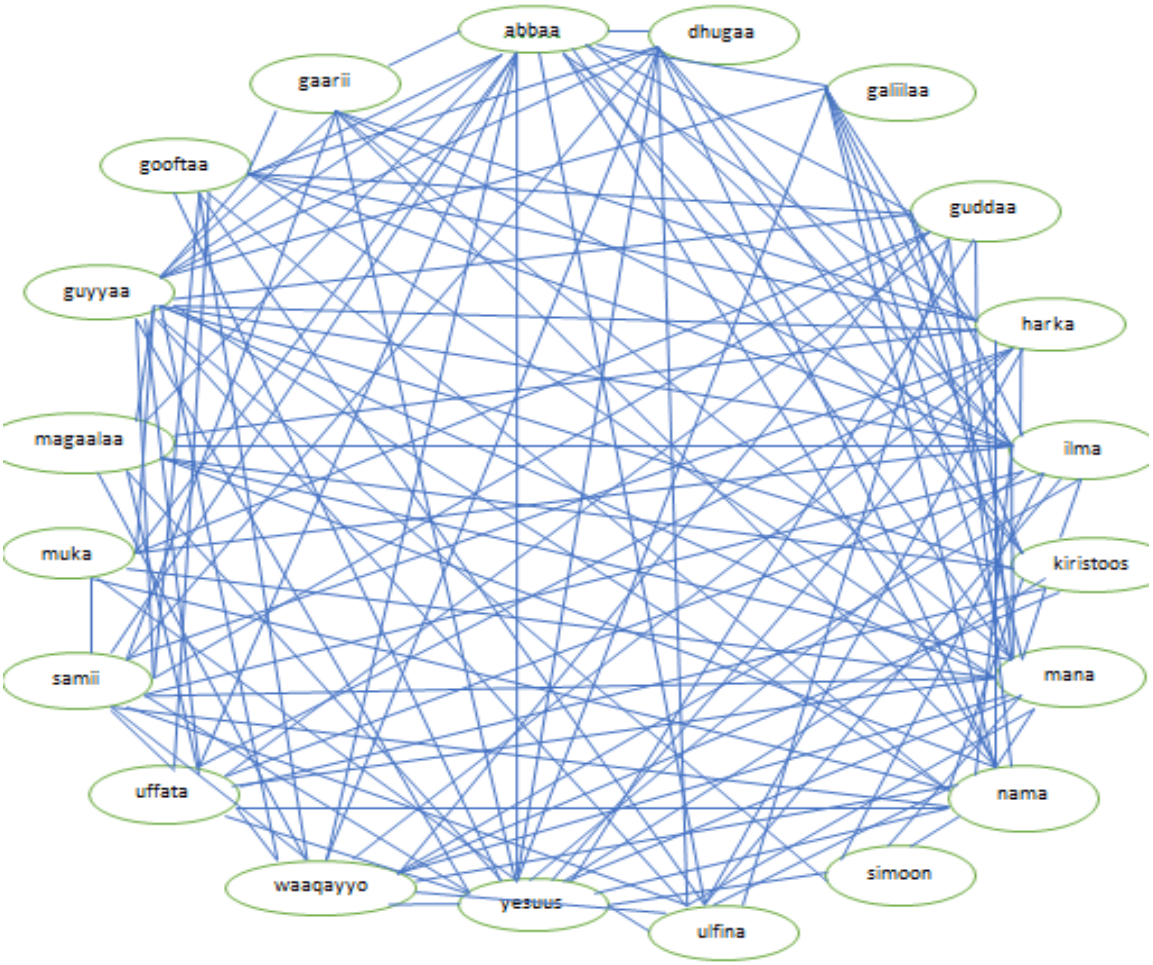


Figure 5: Network of clustered terms

4.2.3 Similarity computation

Given/feeded a term to the prepared prototype, the possible sysnonyms of the term and sometimes there may be possibility of antonyms to be generated, the following figure 6 shows the screenshot of some given terms and the generated corresponding sysnonyms and antonyms respectively.

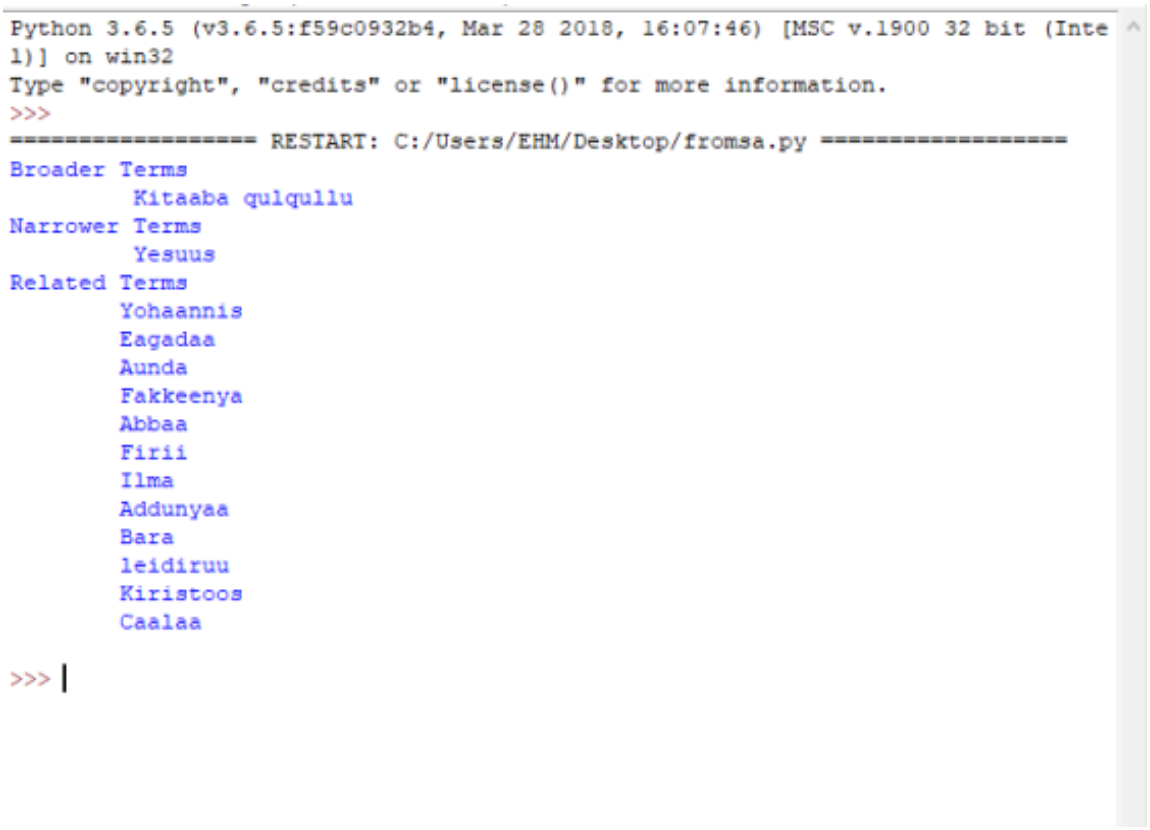


Figure 6: Synonyms and Antonyms of Terms

4.2.4 Cosine similarity

The similarity measure function, in our case the cosine similarity, returns a value one (1), if the term vectors are similar, and zero (0) value, if they are dissimilar. Accordingly, the following figure indicated the sample similarity thesaurus showing how the system is working.and olso the quality of the cluster is analysed/measured using the Intra-cluster technique from the computed similarity of terms.

abbaa																			
dhugaa	0.011																		
gaarii	0.011	0.008																	
galiilaa	0.012	0.012	0.012																
gooftaa	0.011	0.013	0.011	0.01															
guddaa	0.013	0.009	0.01	0.012	0.014														
guyyaa	0.011	0.012	0.01	0.013	0.01	0.011													
harka	0.01	0.007	0.014	0.013	0.011	0.012	0.009												
ilma	0.012	0.012	0.01	0.012	0.013	0.012	0.011												
kiristoos	0.015	0.015	0.011	0.018	0.009	0.009	0.012	0.014											
magaalaa	0.013	0.009	0.0075	0.014	0.008	0.01	0.011	0.012	0.016										
mana	0.012	0.01	0.013	0.011	0.012	0.012	0.011	0.011	0.007	0.009									
muka	0.012	0.012	0.009	0.011	0.011	0.015	0.014	0.011	0.014	0.012	0.011								
nama	0.011	0.01	0.012	0.012	0.012	0.011	0.012	0.011	0.012	0.01	0.012	0.013							
samii	0.18	0.147	0.19	0.15	0.2	0.2	0.16	0.185	0.16	0.135	0.194	0.23	0.173						
simoon	0.012	0.009	0.01	0.02	0.011	0.013	0.013	0.014	0.016	0.017	0.012	0.012	0.012	0.142					
uffata	0.009	0.009	0.011	0.011	0.012	0.01	0.01	0.013	0.013	0.012	0.012	0.014	0.011	0.145	0.017				
ulfina	0.015	0.013	0.011	0.012	0.012	0.013	0.011	0.01	0.012	0.008	0.012	0.012	0.01	0.19	0.01	0.009			
waaqayyo	0.5	0.5	0.36	0.33	0.51	0.44	0.43	0.5	0.55	0.46	0.46	0.42	0.43	0.99	0.38	0.42	0.57		
yesuus	0.012	0.011	0.011	0.012	0.012	0.012	0.011	0.012	0.012	0.01	0.012	0.012	0.011	0.19	0.013	0.012	0.012	0.47	

Figure 7: The cosine similarity of terms

In cases where we don't have the class labels (unsupervised clustering), intra and inter similarity are good measures. So, in our case we used Intra-cluster similarity. Intra-cluster similarity for a single cluster = average cosine similarity of all pairs within a cluster (this should be high) i'e the quality of our cluster is: **1.33585**

4.3 System implementation

The Python programming language 3.4.3 was chosen as the language for the implementation of the system as it is a clear, powerful, object-oriented, general- purpose programming language, that is versatile and can be used for just about anything. Comparable to other programming languages, python is best for newbie coders. Programmers and developers often recommended using it because it employs an elegant syntax making the program we write easier to read. It is an easy-to-use language that makes it simple to get our program working and this makes Python ideal for prototype development and other ad-hoc programming tasks, without compromising maintainability. Python also comes with a large standard library that supports many common programming tasks such as connecting to a web server, searching text with regular expressions, reading and modifying files. Its interactive mode makes it easy to test short snippets of code. There's also a bundled development environment called IDLE. And it is easily extended by adding new modules implemented in a compiled language such as C or C++. The language supports rising and catching exceptions, resulting in cleaner error handling. Code can be grouped into modules and packages and can also be embedded into an application to provide a programmable interface.

Python runs anywhere, including Mac OS X, Windows, Linux, and UNIX. Additionally, it is free software in two senses. The first is that it doesn't cost anything to download or use Python, or to include it in our application and the second is that Python can also be freely modified and re-distributed, because while the language is copyrighted it is available under an open source license.

Using Python 3.4.3 development environment, automatic association thesaurus is developed on a data set which are collected from different sources those written in Afan Oromo. The developed system for automatic thesaurus generation from Afan Oromo text is domain independent which implies that, the system can be employed in different domain areas. It can be applied to other domains without extra modification.

4.4 Discussion

4.4.1 Evaluation of Stemming and Stop words

In the developed system prototype, the preprocessing of data, like tokenizing, normalization, stop word removal, stemming, were involved as the basic steps towards generating automatic thesaurus which have some consequences to the performance of its generation.Accordingly, Afaan Oromo words are stemmed to their root words by depending on rule based stemming algorithm by (Tesfaye, 2010) which checks first measure of the token, then implements rules of stripping inflections of words. Here stemming is needed as a solution for words dimension reduction, speed of the system and memory management purpose.

For our analysis, the evaluation is conducted by considering 20 terms randomly selected from the stemmed training data set, so that the performance of the stemming algorithm employed is evaluated by checking whether terms are stemmed to their root form properly or not. Not only stemming, here simultaneously the stemmed words were identified as whether they belonged to stopwords or not which helped us for the evaluation of the accuracy and exhaustiveness of the stop words compilation process. Accordingly, Table 9 below depicts the evaluation of the performance of the stemming algorithm and the accuracy and exhaustiveness of stopwords compilation step.

Table 10:The evaluation of stemming and stop words of the experiment

NO.	**Stemmed Terms**	**Properly Stemmed**	**Stopword**
1.	Ilma	Yes	No
2.	Namoonni	No	No
3.	Qaama	Yes	No
4.	Obboleessa	No	No
5.	Jedh	Yes	Yes
6.	Wayyoo	Yes	No
7.	Sanyiin	No	No
8.	Aadaa	Yes	No
9.	Barattoonni	No	No
10.	Achi	Yes	Yes
11.	Iddoo	Yes	No
12.	Want	Yes	No
13.	Yesuusiin	No	No
14.	Sana	Yes	Yes
15.	Mana	No	No
16.	Yesuus	Yes	No
17.	Waaqayyoo	No	No
18.	Firii	Yes	No
19.	Yihudootaa	No	No
20.	Kana	Yes	Yes
Percentage		**60%**	**20%**

4.5 Interpretation of Evaluation

In this study, stopword elimination are conducted after the stemming process took place. Accordingly, for this research work, the stemming algorithm which was developed by (Tesfaye, 2010) was implemented to conflate Afan Oromo words into their proper root forms.

Table 10 above depicts the performance of the stemming algorithm on the collected dataset, which were from different source and which contained different concepts. The declination of the performance seen on this particular dataset was resulted due to the followings. (1) Afaan Oromo language was complex in which most Afan Oromo words were formed from continuous process of suffixation as there was no standard list of Afan Oromo suffixes which means the compiled suffixes during the system development were not sufficient enough and it required the complete list of compilation by the language experts. (2) In Afan Oromo, there was no known infixes, which means that as there were no infixes in case of Afan Oromo, rather they show redundancy as in the words like Sochii-Sosochii, Gabaaboo-Gaggabaaboo, Deebi'uu-Deddeebi'uu and qoode-qoqqoode whereas others suggested as there were infixes in Afaan Oromo. (3) The stemming algorithm developed by (Tesfaye, 2010) was rule-based. It could not be applied for prefixes rather it only worked for postfixes and this caused the algorithm not t o be complete enough to handle every word in the prepared corpus. The result on the table 10 above showed that 60% of the randomly selected terms were not properly stemmed, that these were whether overstemmed or understemmed. To minimize this problem and enhance the performance of the stemming algorithm on the collected dataset, as a measure, we redundantly scanned manually for additional suffixes from the entire training dataset and merged them in to the suffix list already collected. Here the knowledge of the background of the language was advantageous.

Table 10 above also showed that 20% of the selected terms were stopwords. This indicates that the accuracy of the stopwords compilation process was 80%. The fact behind this result was that there was no standardized list of stopwords by the language's experts and this indicates the compiled list of stopwords was not complete. Still, to enhance the

performance of the stopword removal, another stopwords were additionally scanned and merged to the existing list manually.

Another concept to be dealt here was that the stemmer reduces morphological variants of words into base or root form. Morphologies of a word, especially suffixes, can be composed of attached, derivational, and inflectional suffixes. Afan Oromo attached suffixes are particles or postpositions. Derivational suffixes are mainly used for the formation of new words in the language from stem or base form of a word. Inflectional suffixes of a word may indicate tense, case, plurality (number), and gender differences. The most common order/sequence of Afan Oromo suffixes (within a given word) is <stem> <derivational suffixes> <inflectional suffixes> <attached suffixes>. For example, the word ***barattootarratti*** (on the students) is composed of ***itti, irra*** (attached suffixes), ***oota*** (inflectional suffix), ***at*** (derivational suffix), and ***bar*** (the stem).

These are caused by some Afan Oromo characters that were used as both postfix and to create the complete word. For example –***n*** is used as postfix in Sanyii->Sanyii**n** whereas it is used just for complete word in Sha**n**. Again 20% of the selected terms still remaining as stopwords are removed, and the accuracy of the performance of stopwords compilation was increased to 80%. The result indicated that, still further work is needed to be done to improve the performance of the stemming algorithm and the compilation of stopwords and suffix list should be standardized by the domain/language/ experts.

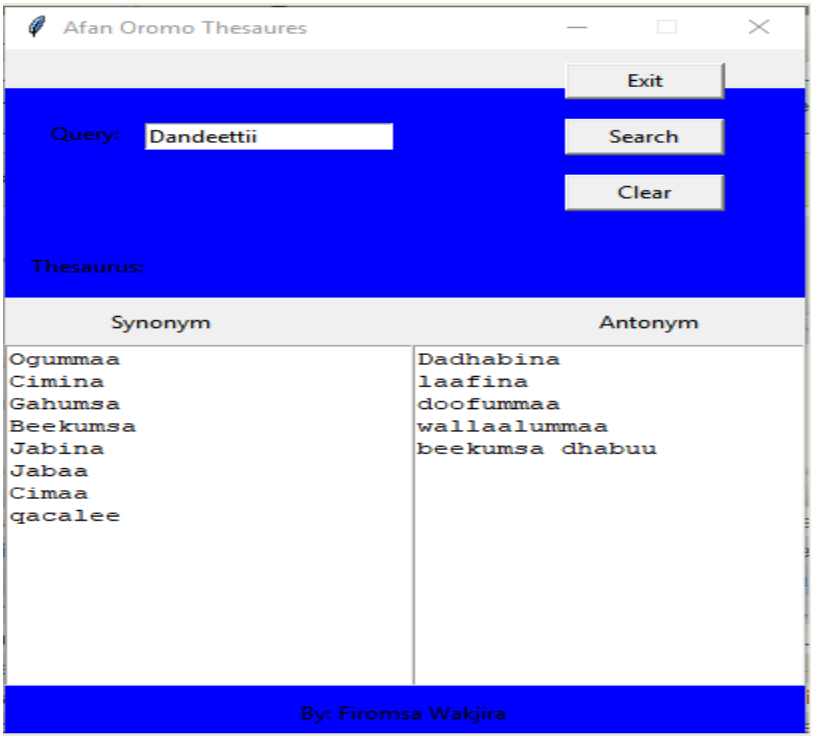
4.6 Evaluation of the Thesaurus Terms

For the purpose of evaluating thesaurus terms, the properly stemmed, non-stop word terms and most co-occurred terms are candidated to be shown here. Accordingly, 20 terms are selected, and the corresponding thesaurus terms were generated depending on the given threshold. In this way, the accuracy of the thesaurus generation system was evaluated based on the sample of properly stemmed terms by checking them in the system. Here, seven and fifteen terms per term was used as a threshold. The top seven and fifteen terms selected as per the number of similarity words was generated. If the generated words were many, top 15 were selected for manageability purpose. Then the terms were evaluated to see if they were similar or not. The result of the finding registers 73.11 % of the output had related concepts to the respective terms. Here this result can be considered as relative average

because it can be dynamic depending on the given threshold value. The result on the collection registered the accuracy of 56.6%. The improperly stemmed terms and the existence of some stopwords that were not detected in the document collection resulted in this irregularity. If the stemmer component was properly developed and the document was properly stemmed and all the stop words were removed from the document collection, a better output would have been expected. Generally, here, inorder to increase the above results, more attention should be given on the document preparation.

To accept a query terms and generate its synonyms and sometimes antonyms terms, a graphical user interface is developed using python 3.4. The following figure 8 Shaws the screenshot of the Graphical User Interface (GUI) of the prototype.

Figure 8: GUI of the prototype



The following shows the cluster of terms and it indicates terms in the same class have a similarity, generally the similarity of these terms is grouped by their class as indicated in the following figure 9 below.

class1: (abbaa, dhugaa, gaarii, galiilaa, gooftaa, guddaa, harka, ilma, kiristoos, magaalaa, mana, muka, samii, uffata, ulfina, waaqayyo, yesuus)

class2: (guyyaa, nama, simoon)

Figure 9: Evaluation of thesaurus

Set of terms are partitioned in to thesaurus classe, and possibly that terms that appear in two or more class are homographs, each of two or more words spelled the same but not necessarily pronounced the same and having different meanings and origins (e.g. bow[1] and bow[2]).

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Query formulation, specially the selection of terms, is the most crucial for search success. Searchers may have difficulties in expressing their information needs in query terms leading to short and simple queries with poor results. Thesaurus is one of the major means of providing searchers with terminological supports in the query formulation. A thesaurus consists typically of controlled vocabulary, which represents the semantic relations between the terms. Thesaurus is developed for supporting user by providing the synonym terms and providing alternative terms for searching purposes. Thesaurus can be constructed manually or automatically, whereas both of them have their own advantages and disadvantages. However, the major problem with the manual thesaurus construction is a labor-intensive task and therefore also expensive to build and hard to update in timely manner and it needs to have an expert on the language/area to do that and also the results are bound to be more or less subjective since the person creating the thesaurus make choice that affect the structure of the thesaurus. Consequently, this research proposed to construct automated Afan Oromo thesaurus using which, in addition to the improvements in time and cost aspects, can result in more objective thesauri that are easier to update. Here the Term-Clustering approach and the cosine similarity measure were used and the remarkable result/output was achieved. For this work, documents were collected from different areas. The developed system was domain independent having measurable response time. The system can be tested by adding extra documents also from difference sources.

For the generation of the thesaurus terms, the attention should be given to the preparation and preprocessing of the collected documents, which means, stopwords and suffixes should

be listed properly, the collected document should be properly stemmed and the non-content bearing words (stopwords) should be removed properly.

For this work, 100 documents containing 36869 are collected and and these collections is from different domains which covers different conceptual areas. For document preprocessing and system implementation, the python programming language, python 3.4.3 was used.

In stemming and stopword removal processes, the evaluation took place in two phases, first phase and the second phase. Accordingly, a random 20 terms were selected from stemmed dataset and evaluated as whether these were properly stemmed or not and whether as these were stopwords or not. Consequently, the evaluation indicated that 60% of them were properly stemmed and 20% of them were stopwords. The improperly stemmed terms and the existence of some stopwords those were not detected in the document collection caused for this irregularity.

The evaluation of thesaurus terms for automatic Afan Oromo thesaurus construction from Afan Oromo text was performed on the top 20 properly stemmed terms and the experimental result indicated that 73.11 % of terms were identified to be related concepts whereas 26.99 % of the remaining terms were not conceptually related. The result suggested that it could be increased more, if the developed stemming algorithm was well done and the stopwords and suffixes were collected and listed properly as the stopword removal and stemming plays the crucial role in generation of thesaurus.

5.2 Recommendation

Afan Oromo thesaurus construction has wide open place for a future work. The work is in its infancy. It needs the collaboration of researchers and the language experts. Accordingly, based on the finding of this research work, we would like to recommend the following points for the future work.

- Thesaurus construction needs standard corpus for testing and making experimentation, but in case of Afan Oromo, there is no standardized/well prepared corpus in the way it is possible to use. In this study, lack of standard or well-prepared

data set for NLP caused the irregularity of the result scored in the experiment. So, developing standard corpus should be given attention for the future work.

- The stemming algorithm used for this study/work doesn't work well for every inflated word in the language, it doesn't work for all affixes of the language and it works well only for suffixes, whereas also prefixes should be addressed by the stemming algorithm. So, there should be further work to develop the better stemming algorithm for Afan Oromo documents.
- For this language, this study was conducted for the first time by using the term-term co-occurrence approach. There are other different approaches, and the study was not compared with those approaches. Further study is needed to figure out the best approach that works for automatic Afan Oromo thesaurus construction.
- Even if it is a labor-intensive task and therefore also expensive to build and hard to update in timely manner, manually constructed thesaurus is more perfect to identify whether words are synonym, antonym or not. So, the immediate research work will be to develop the manual thesaurus and word net for Afan Oromo language.

REFERENCES

Abuzir, Y., & Vandamme, F. (2017). ThesWB : A Tool for Thesaurus Construction from HTML Documents ThesWB : A Tool for Thesaurus Construction from HTML Documents.

Adriani, M., & Rijsbergen, C. J. Van. (1999). Term Similarity-Based Query Expansion for Cross- Language Information Retrieval, 311–322.

Aitchison, J., Gilchrist, A., & Bawden, D. (n.d.). *Thesaurus construction and use: a practical manual* (4th Editio).

Alemayehu, N., & Willett, P. (2003). The effectiveness of stemming for information retrieval in Amharic. *Journal of Documentation Iss Journal of Documentation*, *37*(4), 254–259. https://doi.org/10.1108/00330330310500748

Amirhosseini, M. (2008). Dialectic schemes in thesaurus creation. *Library Philosophy and Practice, Annual Vol*. Retrieved from files/226/Amirhosseini et al. - Dialectic Schemes in Thesaurus Creation.pdf%5Cnfiles/225/summary.html

Angel F. Zazo, Carlos G. Figuerola, Jose L.A . Berrocal, Emilio Rodriguez, and R. G. (n.d.). Experiments in Term Expansion Using Thesauri in Spanish, *213*.

Argaw, A. A., & Asker, L. (2007). An {A}mharic Stemmer: Reducing Words to their Citation Forms. *Proceedings of the 45th Annual Meeting of the Association for {C}omputational {L}inguistics*, (June), 104–110.

Baroni, M., & Bisi, S. (2004). Using cooccurrence statistics and the web to discover synonyms in a technical language. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1725–1728. https://doi.org/10.1.1.58.4300

Beldados, D. (2013). *Automatic Thesaurus Construction From Wolaytta Text*. Addis Ababa University.

Börner, K., Chen, C., & Boyack, K. W. (2005). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255. https://doi.org/10.1002/aris.1440370106

Caplan, N. A. (n.d.). *Using a Thesaurus*.

Chaiwat Ketsuwan, Nattakan Pengphon, A. K. (n.d.). Automatic Thesaurus Extraction for Thai Text Retrieval Enhancement*.

Chen, L. (2006). Automatic Construction Of Domain-Specific Concept Structures. *Evaluation*.

Crouch, C. J., & Yang, B. (1992). Experiments In Automatic Statical Thesaurus

Construction, 77–88.

Cummins, R., & O'Riordan, C. (2005). Evolving Co-occurrence Based Query Expansion Schemes in Information Retrieval Using Genetic Programming. *The 16th Irish Conference on Artificial Intelligence and Cognitive Science (AICS05)*, 137–146. Retrieved from http://www.infc.ulst.ac.uk/~norman/aics05/AICS05_Proceedings_V3.pdf

Dorbin, T., Bruce, R., Chen, H., Ng, T. D., Martinez, J., & Schatz, B. R. (2017). A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval : An Experiment on the Worm Community System Link to item A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval :

Eggi, G. G. (2012). Gezehagn Gutema Eggi Advisor : Million Meshesha (PhD), (June).

El-Hamdouchi, A., and P. W. (2017). Comparison of hierarchical agglomerative clustering algorithm for document retrieval, (December).

Evert, S. (2005). The Statistics of Word Cooccurrences Word Pairs and Collocations. *Unpublished Doctoral Dissertation Institut Fur Maschinelle Sprachverarbeitung Universitat Stuttgart*, *98*(August 2004), 353. https://doi.org/10.1073/pnas.141413598

Foo, S., Hui, S. C., Lim, H. K., & Hui, L. (2000). Automatic thesaurus for enhanced Chinese text retrieval. *Library Review*, *49*(5 and 6), 230–239. https://doi.org/10.1108/00242530010331754

Frank, I. H. W. & E. (n.d.). *Data Mining Practical Machine Learning Tools and Techniques*.

Gamta, T. (2000). The Journal Of Oromo Studies, *7*.

Grefenstette, G. (1993). Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques. *Making Sense of Words, 9th Annual Conference of the UW Centre for the New OED and Text Research*.

Grossman, D. A., & Frieder, O. (1983). *Informatin Retrieval Algorithms and Heuristics*.

Hanandeh, E. (2013). Building an Automatic Thesaurus to Enhance, *10*(1), 676–686.

Hedlund, T. (2003). *Dictionary-Based Cross-Language Information Retrieval: Principles, System Design and Evaluation*.

Hiete, H. (2011). *Automatic Thesaurus Construction For Tigrigna Text Retrieval*. Addis Ababa University.

Imran, H., & Sharan, A. (2009). Thesaurus and Query Expansion. *International Journal of Computer Science & Information Technology (IJCSIT)*, *1*(2), 89–97.

Jiang, J., & Zhai, C. (2007). An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval*, *10*(4–5), 341–363. https://doi.org/10.1007/s10791-007-9027-7

Kejela, M. L. (2010). *Named Entity Recognition for Afan Oromo*.

Khafajeh, H., Technology, I., Private, Z., Yousef, N., Abdulaziz, K., Arabia, S., … Sciences, F. (2002). Automatic Query Expansion for Arabic Text Retrieval Based on Association and. *Information Retrieval*, (October).

Kitsch, S. M. (2008). Similarity Thesauri and Cross-Language Retrieval, 1–32. https://doi.org/10.1177/1097184X03255851

Lairungraung, K. (2003). Automatic Thesaurus Construction With Term Context And Syntactic Analysis For Thai Text Retrieval.

Lassi, M. (2002). Automatic thesaurus construction 2 . Information Retrieval, 1–10.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). An Introduction to Information Retrieva. *Information Retrieval*, (c), 1–18. https://doi.org/10.1109/LPT.2009.2020494

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '06*, 296. https://doi.org/10.1145/1141753.1141819

Mekonnen, A. (2009). *Automatic Thesaurus Construction For Amharic Text Retrieval*.

Meusel, R., & Stuckenschmidt, H. (2009). *Text-Mining for Semi-Automatic Thesaurus Enhancement*. *Master thesis*. Retrieved from http://ki.informatik.uni-mannheim.de/fileadmin/publication/meusel09da.pdf

MILLER, U. (1997). Thesaurus Construction: Problems And Their Roots, *33*(5), 489.

Morshed, A., & Sini, M. (2009). Creating and aligning controlled vocabularies. Retrieved from http://eprints.rclis.org/15830/1/controlled.pdf

Nakayama, K., Hara, T., & Nishio, S. (2007). Wikipedia Mining for an Association Web Thesaurus Construction. *Construction*, *4831*, 322–334. https://doi.org/10.1007/978-3-540-76993-4_27

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Rahman, N. A., Bakar, Z. A., & Sembok, T. M. T. (2010). Query expansion using thesaurus in improving Malay Hadith retrieval system. *Proceedings 2010 International Symposium on Information Technology - System Development and Application and Knowledge Society, ITSim'10*, *3*(October 2015), 1404–1409. https://doi.org/10.1109/ITSIM.2010.5561518

Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. *Proceedings of the 19th International Conference on Computational Linguistics-Volume 1*, (1992), 1–7. https://doi.org/10.3115/1072228.1072235

Ricardo Baeza-Yates. (n.d.). Modern Information Retrieval. *New York*, *9*.

Schneider, J. W. (2005). *Verification of bibliometric methods' applicability for thesaurus construction*. *ACM SIGIR Forum* (Vol. 39). https://doi.org/10.1145/1067268.1067293

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, *24*(1), 97–123.

Schutze, H., & Pedersen, J. O. (1997). A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval IPR2017-01039 Unified EX1023 Page 1 IPR2017-01039 Unified EX1023 Page 2, *33*(3), 1–9.

Soergel, D. (1995). The Art and Architecture Thesaurus (AAT): A Critical Appraisal1. *Visual Resources*, *10*(4), 369–400. https://doi.org/10.1080/01973762.1995.9658306

Song, W., Yang, J., Li, C., & Park, S. (2011). Intelligent information retrieval system using automatic thesaurus construction. *International Journal of General Systems*, *40*(4), 395–415. https://doi.org/10.1080/03081079.2010.530026

Srinivasan, P. (n.d.). Thesaurus Construction.

Tesfaye, D. (2010). *Designing a Stemmer for Afaan Oromo Text : A Hybrid Approach*. ADDIS ABABA UNIVERSITY.

Wikipedia. (2018). Oromo language (p. 2018).

Xu, J. (1997). Solving The Word Mismach Problem Through Automatic Text Analysis.

Yang, D., & Powers, D. M. (2008). Automatic thesaurus construction. *Conferences in Research and Practice in Information Technology Series*, *74*(January), 147–156.

APPENDICES

Appendix 1: Qubee afaan Oromoo fi Dubbachiiftuu (Qubees with their phones)

Sagalee gaggabaaboo (short sounds)				
A	E	I	O	U
Ba	Be	Bi	Bo	Bu
Ca	Ce	Ci	Co	Cu
Cha	Che	Chi	Cho	Chu
Da	De	Di	Do	Du
Dha	Dhe	Dhi	Dho	Dhu
Fa	Fe	Fi	Fo	Fu
Ga	Ge	Gi	Go	Gu
Ha	He	Hi	Ho	Hu
Ja	Je	Ji	Jo	Ju
Ka	Ke	Ki	Ko	Ku
La	Le	Li	Lo	Lu
Ma	Me	Mi	Mo	Mu
Na	Ne	Ni	No	Nu
Nya	Nye	Nyi	Nyo	Nyu
Pa	Pe	Pi	Po	Pu
Pha	Phe	Phi	Pho	Phu
Qa	Qe	Qi	Qo	Qu
Ra	Re	Ri	Ro	Ru
Sa	Se	Si	So	Su
Sha	She	Shi	Sho	Shu
Ta	Te	Ti	To	Tu
Va	Ve	Vi	Vo	Vu
Wa	We	Wi	Wo	Wu
Xa	Xe	Xi	Xo	Xu
Ya	Ye	Yi	Yo	Yu
Za	Ze	Zi	Zo	Zu

Sagalee Dheedheeroo (long sounds)				
‘aa	‘ee	Ii	oo	Uu
Baa	Bee	Bii	boo	Buu
Caa	Cee	Cii	coo	Cuu
Chaa	Chee	Chii	choo	Chuu
Daa	Dee	Dii	doo	Duu
Dhaa	Dhee	Dhii	dhoo	Dhuu
Faa	Fee	Fii	foo	Fuu
Gaa	Gee	Gii	goo	Guu
Haa	Hee	Hii	hoo	Huu
Jaa	Jee	Jii	joo	Juu
Kaa	Kee	Kii	koo	Kuu
Laa	Lee	Lii	loo	Luu
Maa	Mee	Mii	moo	Muu
Naa	Nee	Nii	noo	Nuu
Nyaa	Nyee	Nyii	nyoo	Nyuu
Paa	Pee	Pii	poo	Puu
Phaa	Phee	Phii	phoo	Phuu
Qaa	Qee	Qii	qoo	Quu
Raa	Ree	Rii	roo	Ruu
Saa	See	Sii	soo	Suu
Shaa	Shee	Shii	shoo	Shuu
Taa	Tee	Tii	too	Tuu
Vaa	Vee	Vii	voo	Vuu
Waa	Wee	Wii	woo	Wuu
Xaa	Xee	Xii	xoo	Xuu
Yaa	Yee	Yii	yoo	Yuu
Zaa	Zee	Zii	zoo	Zuu

The colored lines/alphabets indicate afaan Oromo double letters
A word in Oromo cannot begin or end with a double consonant. E.g. The word for “sport” is converted to ***isporti*****.**

Appendix 2: Afan Oromo Stop word list

aanee
f
isaanii
itti
agarsiiso
o
Faallaa
isaaniitii
n
ittiaanee
fagaatee
isaanirra
a
ittiaanse
e
fi
isaaniif
ittiaansu
un
akkasum
as
fuullee
isaanitti
ittiaansu
udhaan
akum
fuullee
isaatiin
ittumaak
kuma
gajjallaa
isarraa
itumaala
gama
isatti
itumallee
alatti
gararraa
isee
ituu
alla
iseen
ituulleea
mma
garuu
ishee
jalaamm
o
giddu
isheetti
jara
ammoo
gidduu
isheef
jechaana
n
gubbaa
isheen
jechoota
ana
ha
isheenis
jechuuan
i
hamma
ishii
jechuuna
ti
hanga
ishiin
jedhebira
henna
ishiitti
jettebood
a
hoggaa
ishiif
jedhanB
ooddee
hogguu
ishiidhaa
kandabal
atees
hoo
ishiidhaa
f
kanadha
an
illee
ishiillee
kanasDu
dduuba
immoo
ishiitti
kanaadu
gda
ini
ishiirraa
kanaafdu
ra
innaa
isii
kanaafid
uuba
inni
isiin
kanaafuu
eega
irra
isin
kanaanee

gana	isini
irraa	kana

eegasii	waa’ee	ofiiMaqa	koo
irraan	nuukiyya	a	narraa
isinii	otuullee	silaa	akka
karaaenn	dhimma	warra	tahullee
aa	nuufkoo	oggaamo	akkam
isa	saniif	o	yommii
isiniin	waan	simmoo	toonatti
keeerga	nuunkun	woo	tana
isaa	sana	oona	yoo
isiniif	waggaa	sinitti	ttinu
keenna	keenyala	yommuu	tanaaf
ergii	fa	osoonaa	yookaan
isaaf	si	siqee	keenyan
isinitti	hoggaa	yemmuu	u'i
keenyais	keessala	otoonaaf	tanaafi
aan	ma	sirraa	yookiin
isaani	siif	yeroo	teessann
isinirraa	hogguu	tiyyanaa	urraa
keessake	nuymale	n	tanaafuu
essan	e	sitti	yookiini
otumalle	siin	yommii	mmoo
e	wajjin	kiyya	nuttinuti
otuu	odooman	Naannoo	ta’ullee
ta’us	na	sun	yookiinis
keessatti	sila	garas	ta’uyyuu
otuu	waliin	yeros	yoom

Appendix 3: Afan Oromo punctuation marks

No.	Panc.list	Panc.Name
1.	‘	Hudhaa/apostrophes
2.	**?**	Mallattoo gaaffii/question mark
3.	“	Mallattoo waraabbii/Double quotation
4.	-	Tishoo/hyphens/dashes
5.	**.**	Tuqaa/periods
6.	**!**	Raajeffannoo/exclamation
7.	**….**	Itti fufaa/ellipses
8.	:	Tuq-lamee
9.	;	
10.	()	Mallattoo cuftuu

Appendix 4: Afan Oromo Affixes

No.	Affixes
1.	-oo
2.	-oota
3.	-wwaan
4.	-an
5.	-ummaa
6.	-lee
7.	-llee
8.	-een
9.	-n
10.	-ot
11.	-cha
12.	-oolii
13.	-tti
14.	-dhaan
15.	-yyii
15.	gar-
16.	Sab-
17.	al-

Three consonants cannot occur in a row in a word
Vowels cannot change without a break, either a consonant or apostrophe, between them.
What breaks are used can differ with spelling preferences and dialects
The apostrophe indicates that the vowels are produced independently and not as a diphthong.
The colored one identifies some list of prefixes

Appendix 5: Sample of representative terms and corresponding documents

Terms	Documents
dhalche	kitaabota seenaa yesuus ilma ilma abrahaam yisihaaqii yisihaaq yaaqoobiin yaaqoob yihudaa yihudaa taamaar faaresii zaaraa faares hesrooniin hesroon raamiin raam aminaadaabiin aminaadaab na'aasooniin na'aasoon salmooniin salmoon rahaab ruut yoobeediin yoobeed isseyiin issey daawit mooticha daawit haadha manaa uuriyaa solomooniin solomoon robi'aam abiyaa aasaa yoshaafaaxiin yoshaafaax yooraamiin yooraam ooziyaa yo'aataamiin yo'aataam akaaziin akaaz hisqiyaasiin hisqiyaas minaasee amooniin amoon yosiyaasiin yosiyaas bara yikooniyaanii biyya booji'amanii salaati'eel zarubaabeliin zarubaabel abiyuudiin abiyuud eliyaaqimiin eliyaaqim azooriin azoor saadoqiin saadoq akiimiin akiim eliyuudiin eliyuud alaazaariin alaazaar maataaniin yaaqoobiin yaaqoob yoseefiin abbaa manaa maariyaam maariyaam yesuus kiristoos deesse dhaloonni hundi abrahaamii daawitii boojuu boojuu baabiloniihanga dhalachuun yesuus kiristoos haati maariyaam kaadhimatamtee waliin jiraachuu jalqabuu hafuura ulfooftee manaa nama qajeelaa uummata qaanessuu hiikuu yaadee jiru yihowaa mul'atee ilma kaadhimni maariyaam ulfooftee hafuura mana fudhachuu ilma yesuus cubbuu hundi raajii dubbatee durbi ulfooftee ilma hiikni hirribaa dammaqee yihowaa ajaje kaadhima mana ilma saalqunnamtii yesuus kiristoos daawit abrahaam **dhalche** dhalche dhalche dhalche dhalche dhalche dhalche dhalche dhalche boaaziin dhalche dhalche dhalche dhalche dhalche robiaamiin dhalche dhalche dhalche dhalche dhalche dhalche dhalche dhalce dhalche dhalche dhalche dhalche dhalche salaatieeliin dhalche dhalche dhalche dhalche dhalche dhalche dhalche dhalche dhalche dhalche dhalche boaaz dhalche jirtuu argamtehaa barbaadne kunoo daawit sodaatinishiin deessi moggaasta raawwatamuufi deessi amaanuel jira waaqayyo raawwanne moggaase barbaadeta'us fayyisa moggaasu kunoo amaanueljechuun fudhate obbol yihuda abiya aasa ooziya minaas baabiloni obbol yikooniya baabiloni yikooniya maata yosee daawiti dhal baabiloni dhal kiristoosi dhal yoseefii qulqulluudha abba yosee dura dhoksaa gochuu maleeka abjuudha yosee qulqulluudha yihowa jechuudha yosee maleeka isaa deessu boojiamani turee jedh dhakanaa taee
	heroodis yesuus beetaliheem magaalaa dhalatee namoonni urjii biyya baha biiftuutii yerusaalem yihudootaa dhalate eessa baha biiftuu urjii arginee sagaduu **heroodis** mootichii namoonni yerusaalem hundi baay'ee buutota lubootaa barsiisota seeraa sabichaa hunda qabee kiristoos dhalatu raajiin tokko ‘yaa beetaliheem biyya bulchaa saba israa'eliin tiksu keessaa yihudaa matumaa

herodis	magaalaa tuffatamtu heroodis urjii urjichi mul'ate mucicha dhaqee deebi'aatii mootichi hime dhaggeeffatanii adeemsa urjichi baha biiftuu bakka mucichi jiru ga'ee bu'ee deema urjichi dhaabachuu baay'ee seenanis haadha maariyaamii arganii jilbeenfatanii saanduqa bananii kennaa ixaana adii qumbii deebine waaqayyo biraatiin biyya deemanii yihowaa mucichaa haadha fudhadhuutii heroodis mucicha ajjeesuu achuma mucichaa haadha mucichaa fudhatee halkaniin gibxii heroodis du'uttis achuma raajii ilma biyya gibxiitii dubbate heroodis namoonni urjii baay'ee urjii namoota dhiiraa umuriin lamaa beetaliheemii hunda ermiyaas raajichaa jedhamee dubbatame boo'ichaa wayyoo tokko magaalaa raamaa keessaa raahel dhumaniifis jajjabaachuu heroodis kunoo yihowaa lubbuu mucichaa balleessuu ka'iitii mucichaa haadha fudhadhuutii biyya israa'el mucichaa haadha mucichaa fudhatee biyya arkelaawos bakka heroodis abbaa biyya yihudaa bulchuu jalqabe achi dhaquu waaqayyo akeekkachiisa galiilaa jedhamee magaalaa naazireet jedhamtu dhufee achi jiraachuu bara dhufanii jira dhufneerra jeeqamanheroodis gaafate jedhaniin dhalata taus barreesseefi yihudaa taee mootiin beetaliheem waamee mirkaneeffate dhaqaatii barbaadaa himaa erge arganis mucicha warqii mul'atee baqadhu turi dhaqe waame raawwatamuufi aare fayyadamee ajjeesise raawwatame dhaga'ame boosse isaaniinis ka'iitii du'etti warri dhaqi deebi'ehaa sodaate malees deeme raajonni waamama raawwatamuufis jalqabe naaziricha dandeenye mulatee taatu yihudaa lakkaa' dhaga' wali eessa yihudaa argamtu bulchit bira iccitiidha sirrii eeggannoodha argatt sagaduu beetaliheemi dhaabatu manaa heroodisi abjuudha isaanii maleeka abjuudha yoseefi gibxii himu yosee yihowa gowwooms hubate lakkaa' sirrii mirkaneeffate erguudha ijool koonyaaww jiraat kana gudda ijool ishiitii maleeka gibxii abjuudhaanyoseefi barbaad yosee dhaga'ee abjuudha dubbat mootichaa ba'uu jedh lakkaa gammad sagad kenn akeekkachiisee deebi' barbaaduu lakkaa jedh sagal du'anii israaeli jedh kennee tauma
	yohaannis onaa yihudaa lallabaa samii yaada jedhu isaayyaas raajichi onaa keessaa ‘daandii yihowaa qopheessaa jedhu tokko dubbate yohaannis uffata rifeensa gaalaa hojjetame mudhii hidhata nyaatni awaannisaa damma bosonaa namoonni namoonni yihudaa hundii namoonni biyya yordaanos hundi dhufu cubbuu himachaa laga yordaanos harka cuuphamu fariisonnii saduuqonni hedduun bakka dhufaa isaaniin buutii dheekkamsa dhufaa jiru jalaa eenyutu firii yaada geddarachuu argisiisu jettanii waaqayyo keessaa ijoollee kaasuu iyyuu qottoon hidda mukeetii mukti firii gaarii godhanne hundi muramee yaada bishaaniin dhufu caalaa kophee ga'umsa hafuura qulqulluu laayidaa ittiin qulleessu harkaa oobdii qamadii habaqii ibidda dhaamuu yommus yesuus galiilaadhaa yordaanos **yohaannis** yohaannis harka keetiin cuuphamuu utuun yaalii qajeelaa hundumaa raawwachuun barbaachisaa yohaannis dhowwuu yesuus cuuphamee battaluma bishaanicha keessaa hafuurri waaqayyoos bu'ee qubatuu tokko samii ilma

yohaannis	jaallatamaa gammadu bara dhufelallabni geddaradhaa qajeelchaa dhaga'ame uffata yerusaalem argu jedhe nana akeekkachiise godhadhaa abrahaami yaadinaa hima danda'a qophaa'eera gatama cuupha jabaata qabu cuupha qaba qulqulleessa naqa dhufe qabuu yesuus dhowwin dhiise ba'e arge turanyohaannis guba dhuftaa jedhe cuupha jechuudha teepha sana cuuphaa ijool baqatt ‘abba dhagaaww abrahaamii muruu ibidda geddarattanii baasuu ibidda guutummaa gootaraa dandeeny cuuphamuu dhowwuu kana sana kana samii gugee sagal mootumma dhihaatee sagal sana deebisee banam
galiilaa	yesuus diyaabilosiin onaatti guyyaa soomee diyaabilos waaqayyoo dhagoonni daabboo yesuus dubbii yihowaa keessaa ba'u hundumaatiin nyaata jiraachuu jedhu diyaabilos magaalaa fiixee gamoo mana qulqullummaa maleekota miilli harka jedhamee ilma waaqayyoo taate yesuusis yihowaa jedhus diyaabilos gaara dheeraa tokko baasee addunyaa hundumaa ulfina tokko kuftee waaqeffatte hunda kanatti nana duraa ‘waaqa yihowaa tajaajila qulqulluu jedhu diyaabilos dhiisee maleekonni waaqayyoos dhufanii tajaajiluu hidhame magaalaa naazireetii qifirnaahom zaabiloonii niftaaleem qarqara argamtu dhaqee achi jiraachuu isaayyaas raajichaa jedhamee dubbatame biyyoonni zaabiloonii niftaalem warri yordaanos gamaa daandii geessu galiilaa yihudoota taane sabni dukkana jiraatu guddaa gaaddisa jiraataniifis ifni jalqabee samii yaada lallabuu galaana galiilaa deemaa obbolaa simoon pheexiros jedhamuu obboleessa indiriyaasiin kiyyoo darbatanii namoonni qurxummii duukaa isaaniin battaluma dhiisanii duukaa bakka obbolaa yaaqoob ilma zabdeewosii obboleessa yohaannisiin abbaa zabdeewosii bidiruu suphanii battaluma bidiruu abbaa dhiisanii duukaa sagadaa misiraachoo mootummichaa dhukkuba gosa hundumaa dhibee gosa hundumaa namoonni fayyisaa guutummaa biyya **galiilaa** oduun dubbatamus guutummaa sooriyaa dhukkubaa waraansa adda addaatiin gaggabdoo qabanii hunda namoonni yihudaa warri yordaanos hedduun duukaa geeffame dhihaatee taate ajaji ilma ‘namni barreeffameera baatu darbadhu deebisee qorin barreeffameera baayee argisiise ‘waaqa kenna yesuus deemi waaqeffadhu dhiheessi barreeffameera deeme dhagau seexana yesuusyohaannis baee jalqabe raawwatamuufi nana arge duaa yesuus geddaradhaa jalqabe jiruu arge kottaa godha jedhe adeemus arge taaanii waame barsiisaa lallabaa waaee fayyisekanaan galiilaa yerusaalem dikaappoolis beelae dandau dhaabachiisee‘waaqayyo deebie ajaja buaa naannae babalate kaes qoramuu hafuura halk qoruu qofaadha qulqulluu geessuudha ‘isa dhagaa rukutamnee irra mootumm want qofaadhaa galiilaa koonyaaww galaanaa galaanichaa irra dhihaatee lama galaana qabd sana kiyyoo lama kiyyooww sana mann jinniidha isaa gama argam jedh deebisee barreeffamee jedh jedh jalqab jedh jedh argamt mootumma dhiphat qabam lamshaa

gammadoo	hedduu arge taa'ee barattoonni barsiisuu hafuura gammadoo samii gammadoo jajjabina garraamiin **gammadoo** qajeelummaa gammadoo laafeyyii gammadoo laafina qulqulluu gammadoo waaqayyoon nagaa gammadoo waaqayyoo gammadoo samii namoonni hamaa hundumaa badhaasni samii guddaa raajota iyyuu akkanuma soogidda lafaa soogiddi mi'aa mi'aa soogiddummaa deebisee argachuu danda'a gatamee miilla namaatiin ejjetama iyyuu addunyaa gaara gubbaa jirtu tokko dhokachuu namoonni ibsaa mana baattuu ibsaa kaa'u guuboo haaluma namoonni hojii gaarii abbaa ulfina ifni seericha barumsa raajotaa diiguu dhufe ta'in seericha jiru hundi seericha keessaa qubee xinnoon tokko tokko raawwatamin hafuu samii lafti darbanii namni xixinnoo keessaa tokko cabsuu warri kaanis barsiisu mootummaa ga'umsa namni isaaniin hojii oolchuu warri kaanis hojii barsiisu mootummaa barsiisota seeraa fariisotaa caalu matumaa mootummaa bara ‘hin namni nama ajjeese hundi mana jedhame namni obboleessa dheekkamuu dheekkamsi qabbanoofne hundi mana namni obboleessa dubbii arrabsu hundi mana murtii walii namni obboleessa ‘gowwaa gatii qabne jedhu hundi ibidda gahaanamtu kennaa iddoo achi jirtuu obboleessi kennaa iddoo aarsaa dhiisii jalqaba obboleessa deebi'iitii kennaa himatee mana deemaa jirtuu dafiitii namni himate abbaa dabarsee murtii waardiyyaa mana dabarsee mana saantima dhumaa kaffaltee matumaa jedhame namni saalqunnamtii fedhiin qabu dubartii tokko ilaaluu fufu hundi garaa ejja mirgaa gufachiise keessaa baasii qaama keetii gatamuu kutaa qaama keessaa tokko dhabuu harki mirgaas gufachiise kutii qaama keetii gatamuu kutaa qaama keessaa tokko dhabuu ‘namni haadha manaa hiiku ragaa barreeffamaa hiikuu argisiisu namni halalummaa raawwatin haadha manaa hiiku ejja saaxilamtu namni dubartii hiikamte fuudhu hundis ejja bara durii turaniin ‘waadaa galte raawwadhu waadaa galte raawwachuu jedhame matumaa teessoo waaqayyoo lafattis ejjeta miilla magaalaa mooticha guddaa keetiinis rifeensa tokkittii adii gurraacha gochuu dubbiin eeyyee lakki hamaa ijaa bakka ilkaanii jedhame nama hamaa nama maddii mirgaa maddii bitaa tokko mana geessee uffata uffattu fudhachuu uffata uffattu namni aangoo qabu kiiloo meetira tokko deemtu kiiloo meetira duukaa nama nama liqeeffachuu barbaadus diina jedhame diinota jaallachuu kadhachuu abbaa samii aduu hamoo gaarii bokkaas qajeelotaa badhaasa maalii warri qaraxa iyyuu akkanuma godhu miti nagaa hojii kaanii caalu akkamii hojjechaa namoonni saba waaqayyoo taanes akkanuma godhu mitii abbaa samii mudaa qabne isinis mudaa qabne yesuus boodas jalqabe argatu dhaalu beelaanii quufu argatu argu ariatanii gammadaa ililchaas warri warri warri warri warri warri warri warri gammadaa dhabe fayyadu dandeessu qabsiisanii fakkaatuun arganii fakkaatin dhufe raawwatama mannaa salphata hundi qabaatu gaumsa turaniin ajjeesin gaafatama dhageessaniittu hima gaafatama gaafatama qabaata eega'a fiddu yaadatte deemi araarami dhiheessi araarami kenna kenna hima baatu ejjin dhageessaniittu hima raawwateera gati mannaa wayya gati mannaa wayya malees hundi hima hundi

	godha raawwata dhiisin dhageessaniittu hima kakatinaa kakatinaa kakatinaa kakatinaa timataa kakatin dandeessu ‘eeyyee nama ‘hin jedhameera ‘lakki dhageessaniittu morminaa mannaa garagalchinamni barbaade dirqisiise deemi kenni dhowwatin jaalladhu jibbi dhageessaniittu fufaa argisiistu baasa roobsa argattu jirtu ‘michuu ariataniiru nana ti‘bakka taaa gaaraa want dheebot mootumma gadd dheebot gara ijool qajeelummaadhaa mootumma soba irra dubbat irra argatt akkami tokkoo magaala hunda ifuu samiitii kennanii dura diiguu dhugum tuqa ajajaww kann cabs samii galuu oolch samii galuu dhugum qajeelumma samii durii murtii isaa murtii qaanessaadha galaa aarsaa dura murtii baanna murtii abba murtii hidhaa fixxu raawwachuu isaa guutumma gahaannami guutumma gahaannami raawwachuu yihowaadhaa samii yerusaalemi kabaluu murtii jala irra kadhatuu goot ilma irra irra jaallat jaallatt sassaab obbol ariatam arrabs ta'ee taana galt galfamtadhugum kakau kennuu ilka kennii ariatanii gaafatt raawwachuu
samii	jettanii hojii qajeelummaa hojjenne abbaa **samii** gatii kennitu fakkeessitoonni ulfina mana sagadaa malakanni afuufamu gatii harki mirgaa godhu harki bitaa samii ilaalus deebisee fakkeessitootaa warri mana sagadaa qarqara guguddaa dhaabatanii kadhachuu gatii kutaa balbala kees cufadhuutii abbaa iddoo dhoksaa jiru ilaalus deebisee saba waaqayyoo waanuma fakkaatu deddeebitanii baay'ee deddeebi'anii dhaga'amu gaafatin iyyuu barbaachisu jedhaa abbaa samii fedhiin samii akkasuma irrattis nyaata barbaachisu yakka hamaa oolchi namoonni dhiifama samii irraas dhiifama cubbuu namoonni dhiifama keessanis cubbuu fuula fuula gatiiisaanii fuula mataa kees goote soomtu fakkaattee soomaa jirtu abbaa bakka dhoksaa jiru qofatu samii ilaalus gatii deebisee biliin nyaatuu ligidni bakka hattuun cabsee hatuu qabeenya kuufachuu bakka biliin nyaachuu ligidni hattuun cabsee hatuu samii qabeenya sababiin bakkuma qabeenyi jiru kees achuma qaamaa ilaalu hinaafu qaama dukkana ifni jiru dhuguma dukkana dukkanni hammam guddaa gooftolii garbicha bitamuu tokko jibbee maxxanee waaqayyoo bitamuu jireenya ‘maal ‘maal dhagna keessanii ‘maal jireenyi nyaata qaamnis uffata samii sanyii samii maarree keessaa namni yaadda'ee umurii irratti dhundhuma tokko dabaluu maaliif bakkee hubadhaa solomoon ulfina qabu keessaa tokkoo waaqayyo bakkee argamanii akkanatti uffisa amantii xinnoo ‘maal ‘maal jettanii matumaa saboonni iyyuu hunda hawwii guddaa samii wantoonni hundi hundumaa mootummichaa qajeelummaa barbaaduu wantoonni hundis matumaa sababiin borii yaaddoo mataa hundi rakkina mataa eeggadhaa godhin argataniiru kennitu beekin kenni kenna kadhattanis ta'inaa jaallatu hima argataniiru kadhattu seeni kadhadhu kenna dubbatinaa fakkaata beeka kadhadhaa jiraattuu ‘yaa dhufu taaa jiru kenni dhiisne dhiisi galchin cubbuu godha dhiisu gurraachessinaa daalacheessu argataniiru soomtu dhiqadhu dibadhu mul'attu mannaa arga kenna dhiisaa mannaa balleessuu kuufadhaa bakka ta'a namni taee dandau jiru jaallata tuffata hima keessanii laata

	waaee waaee yaaddauu dhiisaa ilaalaa soora caaltanii taus maleeswanta yaaddoftu ilaalaa fo'anis miidhagne hara nana caalaahinuffisuuree qabu beeka fufaa dabalamu waaee qaba gaaa qaba tainaa qulqullaau egaa nyaannayookiin dhugnayookiin uffannajettanii dandaujiraa taee nyaanna dhugna yaaddainaa yaaddainaa ‘maal'uffanna guutummaanqaamakeeifaataa argisiifachuu dura goot hiyyeeyyiidhaa dura argachuu irra dhugum guutummaa hiyyeeyyiidhaa dhoksaa hiyyeeyyiidhaa abba dhoksaa arganii dhugum guutummaa abba dhoksaa kadhatt taanee dubbachuudha kadhanna isaanii abba maqa mootumma irra yakkanii qorumsa irra hojjetanii isaanii abba irra hojjetanii isaanii goot taana abba fakkeessitootaa soom argisiifachuu guutummaa dura abba dhoksaa keessanii dandeenye keessanii irra gara sirrii guutumma lamaa tokkoo qabeenyaa xiyyeeffannaadha facaas haam gootaraa abba irra uffattanii abaabooww akkami guddat biqilt ibiddaa gatam qabd want argachuu abba barbaachis boriitii guyya guyya argatt godh namoota haraa goot soomt balleessu dandau dhugumanisiniinjedhaisa ibsa dandeess caaluusimbirr galch dhama'
yesuus	gaara namoonni hedduun duukaa namni dhukkuba nadaayiitiin qabame tokko dhufee sagadee fedha fayyisuu **yesuus** harka hiixatee namichis battaluma dhukkuba nadaayii himne mannaa dhaqiitii lubootaaf ragaa ta'uttis kennaa dhibbaa tokko dhufee lamshaa'ee mana malees dhiphachaa dhibbaa baaxii mana seenuun malu tokko koos aangoo nama jiru tokkoon yommuun jedhu tokkoon yommuun jedhu garbicha jedhuni yesuus dinqisiifatee jiraniin israa'el nama amantii guddaa akkanaa qabu tokko namoonni hedduun baha biiftuu lixa biiftuutii mootummaa samii yisihaaqii yaaqoobii mootummichaa dukkana achittis isaaniis yesuus ajajaa dhibbaa amantii yesuus mana pheexiros amaatiin pheexiros dhukkuba dhaqna gubaatiin qabamtee ciiftuu harka qaqqabu dhaqna ishiinis kaatee tajaajiluu dhihee namoonni hedduu jechuma tokkoon hafuurota keessaa dhukkubsachaa hundas isaayyaas dhibee dhukkuba keenyas dubbate yesuus namoonni hedduun barattoonni seeraa tokko bakka dhaqxu duukaa yesuus boolla simbirroonni samiis mana ilmi namaa bakka mataa hirkifatu keessaa tokko jalqaba dhaqee abbaa awwaalladhu duukaa yesuusii barattoonni bidiruu yaabbatanii obomboleettii tokko galaanicha bishaanichi yesuus rafaa barattoonni dhaqanii dammaqsanii gooftaa amantii xinnoo isaaniin qilleensichaa galaanicha tasgabbii guddaanis namoonni nama qilleensii galaanni biyya gadaareensi namoonni bakka awwaalaa gidduudhaa namoonni nama eenyu jabaatee darbu ilma maal yeroon murteeffame xinnuma booyyee hedduun tokko dheeda baastu hoomaa booyyee kadhachuu isaaniin booyyota hoomaa booyyee qarqara hallayyaa didichanii tiksitoonni booyyotaa baqatanii magaalaa irratti dabalatee hundumaa namoonni magaalattii hundi yesuusiin arganis biyya gadhiisee adeemu cimsanii yesuus buee gooftaa dandeessa qaqqabee fedha fayye yesuus eeggadhu argisiisi dhiheessi seenus kadhate gooftaa

	ciiseera yesuusis fayyisa gooftaa miti dubbadhu fayya qaba deema dhufa dhagau jedhe hima arganne hima dhufanii abrahaam dhihaatu gatamu baay'ee saaatiidhuma fayye dhufu arge gadhiise jalqabde boodas baase fayyise raajichi fuudhe baate raawwatamuufi argu ajaje dhufee barsiisaa bu'a qabu qabu qabu dhufee gooftaa heyyami yesuusis buuu fufi jeedaloonni jedheen duaa jedheen namuu jenna warri jalqabe jedhaniin nana jedhe ifate baayee akkamiiti ajajamu ceee dinqisiifatanii jedhangalaanicha baanii baayee waaqayyoo qabda gain ergi jedhe baanii namanii jira ‘deemi ‘kottu kootiinis‘kanagodhiyommuun boou qaru oolchi dhuftee fayyi godha dhaqi dhaqaa kana sana lubicha ajaje qifirnaahomi ajaja tajaajila ajaja anaa fayyuu tajaajila lolt anaa ajajam maaddii ilma alaa ilka tajaajila sana guba jinniidha qabam isaa gama barsiisa hundumaa baratt anaa kana cima irra bidiricha dhumuu qabd maalii gama jinniidha qabam dhiphisuu hooma jinniiww sanaa kana guutumma galaanicha jinniidha qabam kana argachuu jedh jedh eenyu deebisee jedh dhugum sana awwaallat deem jedh jedh jedh jedh sodaachisanii jalqab dhaq dhum dhaquudha odeess kadhat kaee sodaatt
mana	yesuus bidiruu yaabbatee galaanicha magaalaa qifirnaahom namoonni nama siree ciise tokko yesuus amantii namicha ilma cubbuun seeraa tokko waaqayyoon arrabsaa garaa yesuusis yaadda hubatee hamaa ‘cubbuun jechuutu ilmi namaa dhiifama cubbuu aangoo qabaachuu nama siree baadhuu **mana** namichis mana namoonni sodaatanii waaqayyo aangoo akkanaa ulfina yesuus bakka deemaa namicha maatewos bakka qaraxni sassaabamu buutuu battaluma duukaa mana dhihaatee namoonni qaraxa sassaabanii namoonni hedduun dhufanii nyachuu fariisonni qaraxa sassaabanii cubbamootaa yesuus ogeessi fayyaa ‘ani aarsaa laafina maal sababiin qajeeloo barattoonni yohaannis fariisonni hunda barattoonni soomne yesuus isaaniin misirrichaa misirrichi jiruu sababii yeroon misirrichi fudhatamu eenyu uffata moofaa erbee haaraa miicamne sababiin shushuntuuree uffata namoonni daadhii wayinii haaraa qalqalloo qalqallichi daadhichi qalqalli faayidaa qabne namoonni daadhii wayinii haaraa qalqalloo lamaanuu dheeraa himaa mana sagadaa tokko dhufee sagadee yoona kottuutii harka ishiinis deebitee yesuus namicha duukaa dubartiin dhiigni dhiphachaa turte tokko dhuftee fiixee uffata garaa uffata yoon tuqe yaadaa yesuus garagalee amantiin dubartiin kaasee mana namicha geggeessaa mana sagadaa gaddaa afuufanii hedduu intalattiin rafaa jirti kolfuu namoonni mana seenee harka intalattiinis oduun dubbatamus yesuus deemaa namoonni qaro dhabeeyyii ilma garaa iyyaa duukaa mana tokko seenee namoonni qaro dhabeeyyiin banuu amantii keessanii yommus dhimma beekne cimsee manaa tokko nama arrabni hidhamee qabame tokko jinniin keessaa namichi arrabni hidhamee dubbachuu namoonni hedduun akkanaa matumaa argamee fariisonni mootii jedhu sagadaa misiraachoo mootummichaa dhukkuba gosa hundumaa dhibee gosa hundumaa fayyisaa magaalotaa hunda hedduu hoolota

	tiksee dagatamanii isaaniif sassaabamu hojjettoonni makara ergu ceuu dhaqe lamshaaee lamshaae jabaadhu dhiifameera tokkos jira jedhe yaaddu dhiifameera deemi salphata lamshaae jedheenbarsiisonni namni garaa ‘kaiitii kaee dhaqe kaee jiruu taauu arge duuka baayee kaee boodas jiruu dhagaee barbaachisu dhaqaa tain barbaada namoota hubadhaa soomnu jedhe qabuu dhufa soomu erbu tarsaasa dhoee michoonni mannaa naqu jiruu kaa'i kaee bu'e intalli foaa tuqxe fayya turte fayyiseera saaatii fayyite intala arge jabaadhu yesuus duune baafamaniin qabe kaate waaee jedhe kaee jiruu daawit laafi boodas dandau yesuusis gaafate gooftaa qaqqabee jedhe baname eenyu yesuusis akeekkachiise baanii deemanis baee jalqabe baayee israael beeku jinniiwwaniitiin baasa barsiisaa lallabaa waaee dinqisiifatanii jedhan gargaarsa yesuus argu cunqurfamanii baayee gadde iyyuu baayee muraasa makarichaa kadhadhaa dhuguma jedhe dhaqi nyaata tain dhufe maaliifi dhangalaa turu baaa babalate amantuu eeyyee eeggadhaa naannae dhufaniinutii gama isaa isaanii keessani maalii irra gochuu sana namootaa kennee sana maatewosi maaddii cubbam baratt baratt maalii jedh dhukkubsatanii fayyaadhaa cubbam kana gaddanii irra haara caalaa moofaa godh haaraa geggeessa baratt kana ulul kana tuffiidha hunda hunda jarr jinniidha isaa jinniiww mann gand qabnee turanii baratt midha goofta hojjett isaa kana ilaa yaad hubattanii sana jedh kenn jedh jalqab barsiisa jedh gaafat duutee jiraa dhangalauudha jedh ilaa jedh jalqab deebisanii odeess jedh gausnam waamuu
	hafuurota aangoo kanas hafuurota dhukkuba gosa hundumaa dhibee gosa hundumaa ergamoota yesuus fufee duraa simoon pheexiros jedhamee obboleessa yaaqoob ilma zabdeewosii obboleessa yaaqoob filiphoosii toomaasii maatewos qaraxa yaaqoob ilma simoon yihudaa asqorotuu yesuusiin dabarsee yesuus qajeelfama jedhu kennee yihudoota geessu **magaalaa** samaariyaa iyyuu hoolota mana keessaa badanii samii jedhaa dhukkuba nadaayiitiin tola tola warqii meetii sibiila diimaa sabbata korojoo nyaataa uffata kophee namni hojjetu nyaatu argachuu ganda iyyuu nama simatee ergaa fudhatu bakka deemtanittis achuma mana tokko mana nagaa manichi malu manichi malle namni tokko simachuu baate dubbii baate mana magaalaa awwaara miilla keessanii guyyaa adaba magaalaa adaba sodoomii gomoraa hoolota yeeyyii bofaa gugees garraamii namoonni mana dabarsanii sagadaa keessattis anaafjettanii bulchitootaa mootota isaaniifis dhugaa dabarsanii maaldubbanna jettanii dubbatu hafuura abbaa keessaniitu dubbata obboleessi obboleessa mucaa dabarsee ijoolleenis jettanii nama hundumaa namni jabaatee dhaabatu magaalaa tokko magaalaa ilmi namaa matumaa magaalotaa waliin geessanii tokko barsiisaa garbichi tokkos gooftaa tokko barsiisaa garbichi tokkos gooftaa abbaa manichaatiin miseensota maatii caalaa sababiin haguugamee iccitii beekamin hafu dukkana hime dhageessanis bantii manaa dhaabadhaatii foon ajjeesuu lubbuuajjeesuu dandeenye lubbuus foon gahaannam balleessuu simbirroo

magaalaa	dimbiixiin saantima gatii xinnoo qabdu tokkoon gurguramti keessaa tokko beekin rifeensi mataa keessanii tokko hafin egaa dimbiixota hedduu caalaa gatii guddaa nama dhugaa ba'u abbaa samii dhugaa ganu hunda abbaa samii irratti nagaa dhufe goraadee nagaa ilma abbaa intala haadha haadha manaa ilmaas amaatii gargar dhuguma namni tokko miseensonni maatii diinota namni caalaa abbaa haadha jaallatu namni caalaa ilma intala jaallatus muka dhiphinaa baatee duukaa fedhii qabne namni lubbuu oolfachuu barbaadu hundi namni lubbuu dhabu hundi deebisee simatu anaanis namni simatu erge raajii tokko raajii simatu gatii raajii tokkoo nama qajeelaa tokko qajeelaa simatus gatii nama qajeelaa tokkoo namni xixinnoo keessaa sababii barataa kubbaayyaa tokko matumaa gatii yesuus waamee xuraaoo kenne xuraaoo ibsameera indiriyaas yohaannis bartolomewos sassaabu alfewoos taadewos hinaafticha kenne erge deeminaa seeninaa mannaa israael dhaqaa dhaqxanis daandii lallabaa fayyisaa kaasaa fayyisaa baasaa kennaa gudunfatinaa qabatinaa qaba barbaadaa turaa gaafadhaa duaa magaalaa dhagauu gautu salphata eeggataa kennu reebu eeggadhaa dhiheeffamtu waaee saaatii miti malees kenna kaanii ajjeesisu taatu fayya ariatama baqadhaa israael caalu bieelzebul sodaatinaa gaaa taee jiru dubbadhaawanta lallabaa sodaatinaa mannaa dandau mitii kuftu namoonni sanasodaadhaa sodaatinaa qabdu gana fakkaatin dhufne dhufe iyyuu malu lubbuu dhaba garuulubbuu argata simata simata argatanamni argata hima qabbanaaaa kennu namni namni dhihaateera deebiu dhadhaadhaa kunoo taaa bauudandeessu yaaddainaa mulifamin lakkaaameera taus baaafnama dhabu malunamni baratt irra isaanii akkafayyisanii maqa kana taane kami ‘mootumma dhukkubsat qabam jinnii keessani karaadhaa kami naga hawwit naga baat dhugum murtii jiranii murtii mann dura addunyaatii dubbatt sana kenn akkami abba irra kootii bira jibbamt dhumaa biraa dhugum dhufu gand barata hasaasa abba irra dura anaa dura dura dura fiduu fiduu fiduu anaa anaa anaa anaa dhugum bisha baas fudhattanii seent seent kennamuu duaa geessis fixx jedh barata hundumaatii buuu taee baasuu
guyyaa	tokko yesuus **guyyaa** maasii midhaanii darbaa barattoonni asheeta ciranii nyaachuu fariisonni barattoonni guyyaa dhorkame hojjechaa isaaniin innii namoonni maal mana waaqayyoo seenee daabboo waaqayyo dhihaatu namoonni nyaachuun heyyamamne nyaatanii luboonni guyyaa mana qulqullummaa hojii seera ilaalamne seericha mana qulqullummaa caalu ‘ani aarsaa laafina maal hubattaniittu cubbuu sababiin ilmi namaa gooftaa sanbataa yesuus bakka deemee mana sagadaa namni harki tokko achi jira ittiin fayyisuun keessaa namni hoolaa tokko qabu jiraatee guyyaa boolla buute harkisee achi keessaa baasne maarree namni tokko hoolaa guddaa guyyaa gaarii gochuun namicha namichis harka harki harka tokkoo fayyaa fariisonni yesuusis iddoo namoonni hedduunis duukaa hunda yesuus eenyummaa beeksisne cimsee ajaja kanas isaayyaas raajichaa dubbatame gammadu hafuura

	haqni maal falmu caraanu eenyu daandii guguddaa haqni mo'u shomboqqoo buruqfame fo'aa ibsaa aarus saboonni abdii nama qaro dhabeessa arrabni hidhame tokko namichi arrabni hidhame dubbachuu ilaaluu namoonni achi hundi tarii ilma daawit fariisonni mootii jinniiwwanii yesuusis yaada beekee isaaniin qoqqoodu hundinuu manni qoqqoodu hundinuus jabaatee haaluma seexanni seexana baasu qoqqoodeera maarree jabaatee dhaabachuu baasu humna eenyuutiin baasu hafuura waaqayyootiin waaqayyoo beekin dhufeera namni nama humna guddaa qabu tokko jalqaba hidhin mana seenee qabeenya saamuu mana namichaa saamuu namni cinaa dhaabanne hundi namni anaa qabne hundi cubbuun namoonni hojjetanii arrabsoon hundi namni hafuura qulqulluu arrabsu dhiifama namni ilma namaa dubbii hamaa dubbatu hundi dhiifama namni hafuura qulqulluu dubbii hamaa dubbatu sirna ammaa kanattis sirna dhufu dhiifama tokko firii muka gaarii firii gaarii argattu muka gadhee firii gadhee buutii taatanii jirtanii gaarii dubbachuu garaa guute namni gaariin gaarii kuufate keessaa gaggaarii namni hamaanis gadhee kuufate keessaa gaggadhee namoonni gadhee guyyaa sababiin dubbii keetiin qajeelaa dubbii keetiin barsiisonni seeraa fariisonni tokko tokko mallattoo argisiistu deebisee isaaniin hamaanii ejjaa hunda mallattoo ilaaluu mallattoo yoonaas raajichaa maleemallattoon yoonaas garaa qurxummii guddaa guyyaa ilmi namaas garaa lafaa guyyaa namoonni nanawwee lallaba yoonaas yaada guyyaa murtii kunoo yoonaas caalu mootuun kibbaa ogummaa solomoon andaara lafaatii kaatee guyyaa murtii kaatee kunoo solomoon caalu nama tokko keessaa bakka boqotu barbaacha qabne iddoo boqotu ‘gara mana keessaa dhufus manichi qulqulluu miidhagaa dhaqee hafuurota caalaa hamoo torba fudhatee galanii achi haalli namicha boodaa duraa caalaa gadhee hamaa irras waanuma akkanaatu dubbachaa haati obboloonni barbaadanii dhaaba namni haati obboloonni barbaadanii dhaabatanii yesuus deebisee obboloonni eenyu harka diriirsee haati obboloonni namni fedhii abbaa samii raawwatu obboleessa obboleettii haadha jiru jedhe dubbisnee turree jedhaniin dubbisnee hima jirahaa tain barbaada jedhu taee seenu lamshaae guyyaa jedhe jiraa caaluu lamshaae harka diriirsi diriirse baanii hubatee deeme fayyise raawwachuufi filadhe jaalladhu ifagodhainni cabsu dhaamsudhuguma iyyuu qabame taee fayyise dinqisiifatanii namni namni bieelzebuliin jedhe bada dhaabatu fakkaatuun maleesani bieelzebuliin baasu tokko dandau morma bittinneessa hima dhiifama argatu argata argatu argattu nana dandeessu dubbata baasa baasa hima gaafatamu jedhamta mukti deebisanii barsiisaa barbaadna jedhaniin kennamu tura dhagaanii kaanii murteessu jira dhufteef murteessiti jira hafuurri jooragaruu argatu deebi'a deebiee duwwaa taee dhufa jiraatu argata jiruu haasauu tokko haasauu jiru eenyu haati jedhe jedhe hundi taus daawit heyyamameeraa heyyamameera kaaa dhagau qabaatu dandae ilaa kaes dandaa dandaa murteeffama jedhedhaloonni barbaadataus kunoo kunoo gaautuu sanbataa sanbataa seera dura isaanii sanbataa hojjet cabsani qabne murteessit himat

	sanbataa seera hoola sanbataa sanbataa akkami ajjees jechuudha sabootaa irra saga jinniidha magaala mootumma akkami jinniiww ilma murteess jinnii mootumma akkami wali dubbat isaanii irra irra dhaabd dhaabd ijool akkami want want dubbat hundumaatii murtii kanbiraanisaa halk halk sana dhal sana dhal bisha dhal baratt isaa anaa beelaanii jalqab beela turt argachuu gaafat sana jedh tajaajila godhu maleejinniiww mootumma beekamuu fakkeenyaa geddaratanii dhagauu xuraaa namicha jedh mariat dhaga baasujedh tainnaalaatajedh