ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

AUTOMATIC THESAURUS CONSTRUCTION FROM WOLAYTTA TEXT

DEMEWOZ BELDADOS

JUNE 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

AUTOMATIC THESAURUS CONSTRUCTION FROM WOLAYTTA TEXT

A THESIS SUBMITED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIRMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFROMATION SCIENCE

BY
DEMEWOZ BELDADOS

JUNE 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

AUTOMATIC THESAURUS CONSTRUCTION FROM WOLAYTTA TEXT

BY

DEMEWOZ BELDADOS

Name and Signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____

DEDICATION

To University of Gondar for the sponsorship, creating conducive environment, and provision of the necessary facilities throughout the study time

ACKNOWLEDGEMENT

First and for most, my heartfelt gratitude goes to Ato Ermias Abebe for his patience, friendly approach and continued support for the proper compilation of the research work, regardless of my irregular appearance for progress follow up. I am very much grateful to Ato Wondemagegn Tomas for his valuable and relentless endeavor on consultation regarding what so ever enquiry I had on Wolaytta language. I also thank Ato Gebrieal Tilahun, W/o Emebet Kibreab, and W/o Rahel Asammnew for their concern on the progress and their time critical supports to the successful completion of the research work.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

IR	Information Retrieval
URL	Universal Resource Locator
WPXW	Wolaittatto Pitaliya Xaafiyo Wogga
WQHMK	Wolaytta Qaala Hiwote Maatteme Keettaa
WWW	World Wide Web

ABSTRACT

Thesaurus is a set of terms for documents classification during indexing and query expansion during the process of searching with the aim of enhancing retrieval effectiveness. The major problem associated with information retrieval system: in one hand, users are required to explicitly describe their information need to the system, on the other the system itself often retrieve irrelevant documents due to vocabulary mismatch between query term and index term. As information retrieval system compares query term and index term at a lexical level, the mismatch is so pronounced to affect the retrieval performance. Therefore thesaurus a means to the problem by providing precise and controlled vocabulary of terms for indexing and searching there by resolve vocabulary mismatch.

Wolaytta is an official language of literacy in Ethiopia. Since the introduction of the Latin script in the writing system in 1993, the language has evolved significantly from mere verbal communication to means of instruction then to source of information. To use the language as source of information, the retrieval system should be designed with enhanced capability in resolving what so ever mismatches that arise between query term and index term.

This research thesis develops an automatic association thesaurus from Wolaytta text for possible inception of enhanced retrieval system or to provide a frame work for the development of cross-language retrieval system. The developed system is based on term-to-term co-occurrence based automatically constructed association thesaurus from document corpora. In order to obtain a reasonably good performance the system incorporated manual approaches regarding stop words and suffix list compilation processes and achieved a better result in generating related concepts.

CHAPTER ONE

Introduction

1.1 Background

Thesaurus is a common tool in linguistic analysis but its application in information retrieval is basically different. The language thesaurus has a general purpose application as compared to the domain specificity of information retrieval thesaurus. Content bearing terms that are used for indexing documents in information retrieval often characterize to have a direct or hierarchical relationship with other terms, i.e. a term can be related to another term, can be broader or a narrow term [1].

Information retrieval system is a search engine designed to provide the information need of users by retrieving relevant documents based on users query. The explosive growths of document corpora in the World Wide Web often made the searching and locating process of specific documents (web) a challenging task and require effective retrieval systems. Effectiveness of Web search engine is the most important criteria of information retrieval research [2].

Thesaurus in information retrieval plays an important role by providing a precise and controlled vocabulary of terms categorization during the process of indexing and query expansion during searching there by enhance retrieval effectiveness [3].

Thesaurus can be constructed manually or automatically. Automatic thesaurus is important for information retrieval by providing a common, precise and controlled vocabulary of terms for indexing and query expansion for searching [4]. On the other hand, manual construction of thesaurus is a highly conceptual, knowledge-intensive, expensive to build and hard to update in

timely manner. Therefore, there is a great demand for thesaurus which can be constructed easily, and maintained with less human intervention [4].

Wolaytta is a language spoken by nearly 2,000,000 people of the Wolaytta zone [5]**.** The language plays a crucial role for the people of the zone in social, political, religious and economic activities. Education in primary schools is now being conducted in Wolaytta as a means of instruction and Wolaytta is also being offered as a subject in junior and secondary schools. As lemma indicated in his research work in [6], significant number of translation work written in other languages to Wolaytta has been implemented and there were about 10,000 hard copy documents available in the language by the time he did his research work in 2003. Following the role of the language in societial development and the escalating demand of in formation, 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 Wolaytta text for possible inception of enhanced retrieval system or provide a frame work for the development of cross-language retrieval system.

1.2 Statement of the Problem and Its Justification

Ethiopia has 83 different languages with up to 200 different dialects spoken. The Ethiopian languages are divided into four major language groups. These are Semitic, Cushitic, Omotic, and Nilo-Saharan.Wolaytta language is classified under Afro-Asiatic language of Omotic family [7].

As Lemma described in his research work in [6], Wolaytta language plays a crucial role in social, political, religious and economic activities for Wolaytta people. The language is spoken not only in Wolaytta zone, but also in the two neighbouring zone (Gamo Gofa and Dawro) with minor differences in dialect. Accordingly, the role of the language has stretched from day-to-day verbal public communication (marketing, shopping and religious teaching) to means of instruction in primary schools. Wolaytta is also being offered as a subject in junior and secondary schools. Consequently, text books and other referneneces are already prepared using the Latin script introduced since 1993. Further more, significant number of translation works of resources written in other languages to Wolaytta has also been implemented and there were about ten thousands hard copy documents available by the time the research was conducted.

Though that, Lemma didn't explicitly specified the source, most of the documents were identified to be religious and written before 15-25 years. In the research it was also indicated that the application of computer in offices and schools has increased significantly.

Therefore it is clear to see that the language has evolved through different developmental stages from verbal communication to means of instruction then to source of information for people of the native users. Following this remarkable evolution one can perceive that, in a very near future there would be a great demand for tapping the huge information resources available from any electronic sources through the inevitable retrieval system using the language as a tool or through cross language retrieval system.

Despite the remarkable development of the language and the complex nature of its morphology, very few works have been done on its lexicon for possible inception of retrieval system. The only work as a step towards the development of effective retrieval system is "Development of

Stemming Algorithm for Wolaytta Text" by Lemma Lessa. Unfortunately so far no one has tried to construct thesaurus from Wolaytta language.

As Wolaytta is morphologically complex language, there is a need for automated procedure to reduce the size of the lexicon to a manageable level and improve retrieval performance, and capture strong relationships that exist between word forms in the language [6].

Hence this research work attempts to construct automatic thesaurus from Wolaytta text for possible inception of enhanced retrieval system or as base line for the development of cross-language information retrieval ineffect addressing the following research questions;

- Which approach is effective for constructing automatic thesaurus from Wolaytta text?
- How the approach is implemented in constructing automatic thesaurus from Wolaytta text?
- What are the basic syllable structure, morphems, and phonetics in Wolaytta and how is the word formed?
- What is the level of performance archived through implementing the approach?

1.3 Research Objectives

1.3.1 General Objective

The main objective of the study is to construct automatic thesaurus from Wolaytta text as a supplementary step for the provision of information for the people of Wolaytta and the neighboring zone with similar dialec

1.3.2 Specific Objectives

To achieve the general objective this research work formulated the following specific objectives;

- To identify pertinent techniques important for constructing automatic thesaurus from Wolaytta text.
- To reduce the size of lexicon of the language under consideration for automatic thesaurus construction?
- To prepare data sets required for automatic thesaurus generation.
- To develop a prototype of automatic thesaurus from Wolaytta language.
- To evaluate the performance of the prototype.

1.4 Methodology

To answer the research questions and achieve the stated objectives this research work employ a quantitative experimental research methodology and follows the subsequent techniques as described below.

1.4.1 Literature review

Obtaining the best tool is one of the major tasks towards constructing automatic thesaurus from Wolaytta text. Hence, significant number of related works appearing in Journals, Proceeding papers, periodicals and books were be reviewed. Furthermore some graduate levels of locally researched thesises were also reviewed.

1.4.2 Tools and Techniques

In this research, Python programming language was employed 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 such as flexible arrays and dictionaries.

1.4.3 Datasets Preparation

To develop automatic thesaurus from Wolaytta text, a standard and representative document corpus should be selected. Therefore due attention was given on the preparation appropriate document corpus for training thesaurus. The test dataset should also appropriately be selected in order to perform exhaustive evaluation of the performance of thesaurus. To this end a person with comprehensive knowledge of the language was consulted.

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 data set employed. The evaluation also includes the performance of the stemming algorithm and stop words 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 peromed by taking a random sample of 20 terms from stemmed candidate terms (discriptors) and the result of the evaluation is indicated as a percentage of the sample test set.

1.5 Scope and Limitation of the Study

This research work encompasses text operations such as tokenization, stop words removal of Wolaytta text while the cardinal emphasis goes to stemming, as these processes are critical in obtaining terms with reduced lexicons. On constructing automatic thesaurus, the scope of the research was limited on co-occurrence based automatically constructed association thesaurus from document corpora. Furthermore due to the morphological complexity of the language, compound terms were disregarded from the automatic thesaurus construction. In view of the fact that there has been few research works on the language, unavailability of standard documents, time factors and budget constraints were the limitations on this research work.

1.6 Significance of the Study

This research work is a foundation for supplementing the ICT infrastructure expansion work of the government that aims to reach every village for information provision [8]. Automatic thesaurus construction from Wolaytta text is a framework setup for possible inception of enhanced Wolaytta 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, the case of Ethiopia, through which Wolaytta people and the neighboring zone with similar dialect could satisfy their information need.

The features that cross languge retrieval system provides is not solely applicable for language with similar dialects, but could handle a more diversified catagories of language groups (Semetic, Cushitic, Omotic, and Nili 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 languge can assume any number and

type of languges. However, query can be formulated by selecting from a menu presented in the source language [9].

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 retrieved due to word mismatch [10]. The problem is due to the fact that information retrieval system compare 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 Wolaytta language.

1.7 Organization of the Thesis

The research work in this thesis is organized in to five chapters. The first chapter is an introductory part of the thesis which, defines problem domain, justifies the relevance of the research, pinpoints the methodologies to be employed and discuses the possible outcome and its usefulness. Chapter two focuses on literature review where relevant research work will be discussed in detail so that techniques pertinent to this research work could be identified. Further more, the chapter presents some of the important works undertaken so far in the language in order to have the clear notion of the Wolaytta writing system and word formation. Chapter three contains the architectutre of the system and the basic techniques invoved for developing automatic association thesaurus. Chapter four has a discussion on the experiment employed and evaluation of the performance archived through experimentation. Chapter five presents the conclusion drawn from the research work and the recommendation of research areas for further research works.

CHAPTER TWO

2. Literature Review

2.1 Overview of thesaurus

Even though, there appear to be quite different definitions for thesaurus based on their area of application, their semantics coincide with their structural composition. Thesaurus is a set of terms consisting of both indexing language and lead-in vocabulary, where the indexing language is a descriptor that applies among terms in the lead-in vocabulary for identifying appropriate relationships [11].

From historical perspectives, Roget's thesaurus was the first, which had been designed and constructed to assist writers in creatively selecting vocabulary [12]. However, current research on classical thesaurus deals with the application of thesaurus for enhancing information retrieval effectiveness.

Thesaurus in information retrieval is designed to address vocabulary mismatch problem for effective information retrieval system. Information retrieval thesaurus is a list of candidate search terms and their relationships for broadening or narrowing the user's search topics by suggesting synonym or hyponym to retrieve potentially relevant documents from large document collection [13].

Thesaurus can be constructed manually or automatically. However, modern IR research on thesaurus contruction has given greater emphasis towards generating thesaurus automatically from huge document collections. Automatic supersede manual thesaurus construction on the fact

that, the later is costly, time consuming, labor intensive and difficult for keeping semantic recourses updated in a timely manner [3].

2.2 Application of thesaurus

Today many people are concerned with the limitation of effectiveness in finding the desired references; despite computer system aiding retrieval from bibliographic or full-text database have been availability for more than two decades [14].

Information retrieval system is a search engine designed to provide the information need of users by retrieving relevant documents based on users' query. Explosive growth of document corpus in World Wide Web often made the searching and locating process of specific document (Web) a challenging task and require effective retrieval system. Effectiveness of web search engine is the most important components of information retrieval research [2].

The main purposes to use a thesaurus are (1) to provide a standard vocabulary for indexing and searching, (2) to assist users with locating terms for proper query formulation, and (3) to provide classified hierarchies that allow the broadening and narrowing of the current request according to the needs of the users [15].

Thesaurus in information retrieval plays an important role by providing a precise and controlled vocabulary of terms categorization during the process of indexing and query expansion during searching thus enhancing retrieval effectiveness [16]. It is also a means to achieve consistency in indexing of written or otherwise recorded documents and other items, mainly for post-coordinates information storage and retrieval system [17]. A thesaurus 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. Most thesauri have been designed to facilitate access to information contained within one database or a group of specific database [17].

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 [18].

2.3 Features of Thesauri

In the following subsections some of the most important features of both manual and automatic thesaurus will be discussed. As discussed in [19], the features include: coordination level, term relationships, number of entries for each term, specificity of vocabulary, control on term frequency of class members and normalization of vocabulary.

2.3.1 Coordination level

Coordination is the construction of phrases from each available term. There are two types of coordination options; pre-coordination and post-coordination. A pre-coordinated thesaurus is a thesaurus that contains phrases and can be used for indexing and retrieval. However, in post-coordinated thesaurus phrases will be 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. Phrase combination can takes place during searching, as a result search precision drops.

2.3.2 Term Relationships

Term relationships describe the semantic association between terms in the vocabulary. There are three types of term relationships [19]; Equivalent relationships, hierarchical relationships, and non hierarchical relationships. Equivalence relations include both synonymous and quasi-synonymous terms. Quasi-synonymous are terms that posse's significant semantic overlap and are considered to be synonymous for retrieval performance. Consider "genetics" and "heredity" with significant overlap in meaning. A hierarchical relation can be described as genus-species, consider for example the malaria parasite with genus name Plasmodium has four species names vivax, malaria, falciparum and ovalae. Whereas, a non-hierarchical relationships can be described as thing-part relashionships, like car and its tyres.

2.3.3 Number of Entries for Each Term

Homograph-words 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 bonds. The semantics can contextually be deciphered as; bonds (chemical) and bonds (adhesive).

2.3.4 Specificity of Vocabulary

Specificity of a thesaurus vocabulary can be described as a function of the precision of individual terms in the vocabulary. The limitation with highly specific vocabulary is that, it requires regular maintenance as specific terms usually tend to change than more general terms.

2.3.5 Control on Term Frequency of Class Members

The basic idea here is that, statistical thesaurus constructed by partitioning the vocabulary in to a set of classes with equivalent terms. Terms with in the same class should be equally specific and 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 enhance retrieval precision. In order to achieve the best query-document match, it is necessary that terms in each thesaurus classeses should have equal frequency and the total frequency of each class should also roughly be equal [20]. Such feature will have a direct influence on the probability of document-query match to be uniform across classes.

2.3.6 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 stop word 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; term should be in noun form, noun phrases should avoid preposition unless commonly know, number of adjectives should be limited, singularity of terms, order of terms within a phrase, spelling, capitalization, transliteration, abbreviations, initials, acronyms and punctuation.

2.4 Approaches to Automatic Thesaurus Construction

During sixties, much of the emphasis was devoted to co-occurrence analysis as an approach to automatic thesaurus construction for enhancing retrieval performance. However, Peat and Willett [20] refuted the idea by demonstrating frequently co-occurring terms from document collection have no significance for enhancing retrieval performance, instead co-occurrence analysis should only be employed for identifying less frequent terms that co-occur in document collection.

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 [22].

The following subsections briefly discus some of the recent functional approaches to automatic thesaurus construction.

2.4.1 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 and 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 [23].

The concept space approach for automatic thesaurus generation was first developed by Chen, Ng, Martinez, and Schatz, [24]. 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 ivolved in the concept space approach includes; document and object list collection, object filtering and automatic indexing,co-occurrences analysis and associative retrieval.

2.4.2 Document clustering

Crouch and Yang [25], constructed automatic thesaurus by document clustering based on low frequency model of Salton et al [26]. 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.

2.4.3 Syntactic Analysis

Syntactical analysis is an alternative approach to address the limitation of co-occurrence analysis in that, co-occurrence analysis: (1) considers co-occurring terms to be similar despite the distance between them in each document. (2) Overlook similar terms from different documents. Syntactic analysis approach was first proposed by Grefenstette [27], 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.

2.4.4 Lexical Co-occurrence

Lexical co-occurrence is the co-occurrence of two terms in a limited window of k-size. Lexical co-occurrence first demonstrated by Schutze and Pederson[28], with the idea that if terms often co-occurred close to each other then they will have greater chance to be related than co-occurring in the same document as in the case of co-occurrence analysis.

2.4.5 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 [29].

2.5 Automatic Generation of an Association Thesaurus from Document

Thesauri can also be classified in to taxonomy based thesauri and association based thesauri. Though the applicability is emphasized in manual thesaurus construction, various researches has been undertaken to extract taxonomical information from document such as hyponyms based on linguistic patterns and synonyms based on similarity of co-occurring words [30],[31]. However, it is possible to automatically generate an association thesaurus, a collection of pairs of semantically associated terms, from document. As indicated in [19], both taxonomy and association based thesaurus are important for enhancing retrieval effectiveness. Church and Hanks [32], had conducted an experiment on word association norm based on co-occurrence analysis. The important steps for automatic generation of association thesaurus include; term extraction, co-occurrence data extraction and correlation analysis.

2.5.1 Term Extraction

Terms that would be incorporated to thesaurus first extracted from corpus. These terms are simple or compound noun that represent domain-specific concepts and frequently occurs in the document. Nouns would be extracted based on counting the frequency of occurrence that exceeds a given predetermined threshold. . Table 1 shows an example of terms with their occurrence frequencies.

Frequent term	T_1	T_2	T_3	----------	T_i
Occurrence frequency	F_1	F_2	F_3	----------	F_i

Table 1: Terms with their co-occurrence frequency

2.5.2 Co-occurrence Data Extraction

The semantic or conceptual meaning of terms lie on its context, i.e. meaning and context should be central in linguistics. The idea of collection on the lexical level first introduced by Firth in [33] and defined as the consistence co-occurrences of a word pairs within a given context.

After Firth's work the idea of a collection has gained a momentum of acceptance and a number of researches have been conducted. In line with lexical collection, the notion of grammatical collection has also been studied and defined as "a phrase consisting of dominant words (noun, adjectives, and verb) and a preposition or grammatical structure such as infinitive or a clause." [34] Lexical collection characterized to have lexical elements with strong dependencies between them and no interchangeability with in any of the elements.

The mathematical motivated notion of collection has also influenced modern computation of relation between words. The distributional hypothesis of Harris [35] stated that linguistic analysis should critically be viewed in terms of a statistical distribution of elements at different hierarchical level.

The new approach, interpretation of the meaning in linguistics from the computational perspectives i.e. an approach derived from psycholinguistic research in to word association in combination with method from information theory (mutual information) and computation (co-occurrence) was developed [36].

The prominent researchers Smadja, and Church have significantly contributed so that the notion of collection has undergone development beyond the description of lexical and grammatical collection method. They introduce the usage of term "word association" with its broader meaning and practical examples of automatically computed, strongly associated word pairs

which stand in relation such as metonymy, hyperonymy and polysemy and so forth far beyond Church's algorithm that compute just pairs of words that frequently appear together [34].

Co-occurrences data extraction initially mentioned by Church and also Schveneveldt [32] and the idea is to automatically extract pairs of "some how" related or similar words by statistically observing their co-occurrences patterns. The extracted pairs of words of significant but in a given special context could be considered as associated to each other in some ways.

Co-occurrence data has a number of application such as, word sense disambiguation [37], [38] Word sense discrimination [39], [40] Computation of thesaurus [41], keyword extraction [42], text summarization [43], extraction of terminology [44].

2.5.3 Correlation Analysis

After performing statistical significant tests on co-occurrences, it is important that all the parameters for the correlation analysis should be defined.

First assume that the co-occurrences are measured on a corpus of fixed size N (N can be the total number of running words, sentences, texts etc.) Secondly for each term T_i the number of occurrences N_{Ti} with in this corpus is F (T_i). Frequency can either be the number of times T_i occurred: N_{Ti} or occurrences of T_i per text (or per million of words) or in comparison with the corpus size N, of sentences. The latter case is known as the probability P (T_i) with which unit T_i is the next one to occur if the corpus to be extended by one word or one sentence.

$$P(T_i) = F(T_i)/N \quad \text{----------} \quad (1)$$

The third parameter determine whether two specific word form count as co-occurring or not. They are considered as co-occurring if they are near enough to each other and not co-occurring if

otherwise. However this explanation seems trivial, the exact interpretation can vary greatly because this parameter is not yet a standardize formulae used for co-occurrences measurements except in the works by [43]. The evaluation of the different measure will be performed on measuring co-occurrence in sentence. The fourth and final parameter is the interpretation of the result obtained from measuring co-occurrence. The semantic relation for each word, the strongest (in the measuring of highly significant value) co-occurrences were considered. These words, ranked by the significance of co-occurrences, are called word association [32].

2.5.4 Mutual Information

Most information extracted from text corpora comes in the form of co-occurrence counts. Co-occurrences of a word with another co-occurrence of a word, co-occurrence of a word with part of speech, co-occurrence of part of speech sequence with a syntactic structure, occurrence of a word in a certain type of text etc. If we consider co-occurrence of words, we can get different multi-word expression such as idioms (frozen figurative expressions), Collections (arbitrary choice of adjectives or verb to express a certain meaning in relation to noun), Lexical bundles (very frequent sequences that behave as a single function word for example as well as, in order), Compound (more or less lexicalized), Named entities (New York, League of Nation)

Mutual information is the oldest and most used association measure in computational linguistics [44].The common methods of measuring the significance of co-occurrences is using mutual information [19] for word associations, as well as for computing synonyms [45], [46]. Considering two terms T_i and T_j co-occurred in text corpora, in Information theory mutual information quantifies extra information about the possible occurrences of T_j when we know that first word is T_i. The basic idea here is to measure the mutual information between two random

terms, by computing the probability of observing T_i and T_j together with probability of observing T_i and T_j independently;

$$MI\ (T_i, T_j) = \log_2 P(T_i, T_j)/P(T_i)P(T_j) \quad = \text{Log}_2\ N.N_{TiTj}/N_{Ti}.N_{Tj} \qquad (2)$$

When the mutual information formula interpreted; Logarithm of (ratio of) empirically estimated probability of bigram, and a theoretical probability under independent product of empirical probability of unigram. In other words; logarithm of (ratio of) P ($T_j|T_i$) (probability of seeing second word if we see first word) to P (T_j) (probability of second word independent of context.)

Applying usual maximum likelihood estimates (C ( ) is a counting function; different strategy for what counts as Ti, Tj co-occurrences):

$$P(T_i,T_j)\ /P(T_i)P(T_j) = C(T_i,T_j)\ /\ N\ /\ (C(T_i)\ /\ N)\ (C(T_j)\ /\ N) \quad = C(T_i,T_j)N\ /\quad C(T_i)C(T_j) \qquad (3)$$

Depending on the task, N (The sample size) might be interpreted as the number of tokens in the whole corpus, or as a number of items in the bigram list. For example number of verb and noun pairs extracted with expression above (in case, unique frequencies should also be taken from list rather than from whole documents.)

Most often when the interest is only on ranking a list of pairs in which case N will not matter, being constant for all pairs. However for association measures that have statistical interpretation changing N will have an effect on the absolute value of score leading to the different probability value. In such case there is a need to take the logarithm of the equation in (3) so that;

$$MI\ (T_i, Tj) = \log2\ (C\ (T_i, Tj)) + \log2\ (N) - \log2(C\ (T_i)) - \log2\ (C(T_j)) \qquad (4)$$

The fact that using the formula;

$$C\ (T_i, Tj)\ /\ C\ (T_i)\ C\ (T_j)$$

Has a serious over estimate for low-frequency events, the value will increase with numerator (C (T_i, Tj)) and decrease with denominator (C (T_i) C(T_j)). However, since they are not independent quantities, C (T_i,T_j) can also be equal to C(T_i) and C(T_j). In the best case scenario:

$$C(T_i,T_j) = C(T_i) = C(T_j) = F \quad \text{the core formula becomes: } F/F^2 \tag{5}$$

Since F does not grow as fast as F2, the mutual information will decrease as F become larger and counter intuitively, highest possible theoretical mutual information for word that occurs once, and that time they occur together. The following table illustrates the trend in mutual information when the value of F increases.

F	F/F^2
1	1
2	0.5
3	0.33
10	0.1
100	0.01
1000	0.001

Table 2: The trend of mutual information value with increasing value of frquency.

Thus the empirical solution for such kind of behavior is to pick the minimum frequency cut-off. Frequency threshold often produce excellent results, however they are arbitrary, depending on corpus size, and might cause lose of important information. The more principled, yet not necessarily empirically better, approach is to use association measure that takes absolute

observed frequency in to account such as log-likelihood ratio [19]. This and related measurement have at their core a formula weighing absolute observed frequency by mutual information the formula also known as local mutual information [47].

$$C((T_i,T_j) \cdot \log P(T_i,T_j) / P(T_i)P(T_j) = C(T_i,T_j) \cdot \log P(T_i,T_j)N / P(T_i)P(T_j) \quad (6)$$

Table 3 illustrates weighing absolute observed frequency by mutual information.

F	F.log2(F. 100000/F2
1	16.6
2	31.2
3	132.9
10	196.6
100	996.6
1000	6643

Table 3 : Weighing absolute observed frequency by mutual information

2.5.5 Log Likelihood Test

Measuring co-occurrences should be modeled by statistical means assuming each sentence (window size) as an experiment in a row of repeated experiment with two outcomes: Either both words T_i and T_j contained in the sentence, or not [20].

The assumed properties of the experiments include;

- The probability of the occurrence of the observed words does not change (which is true by definition, since the probability is determined by fixed size corpus.)
- The experimental outcomes are independent on each other (or the dependence falls off very fast with distance between the experiments, so that it can be disregarded.)
- Both T_i and T_j occur in every sentence at most once, which is mostly true for higher frequency word.

Based on the three assumptions, we can describe the distribution of the outcomes using binomial distribution, which describe the probability that the co-occurrences of T_i and T_j will be observed k times if there are N experiments and independent assumption yields the probability P of T_i and T_j co-occurring:

$$P(k)= P^k(1-P)^{N-k} \binom{N}{K} \quad (7)$$

Where NP is the means and NP(1-P) is the variance. If NP(1-P) > 5, the distribution of the variable approach the normal distribution[52]. Nevertheless when measuring co-occurrences, N is very large, where as

$P= P(T_i).P(T_j) = (N_{Ti} \,.\, N_{Tj})/N^2$ (for the independent assumption) is usually very small, thus:

$$NP(1-P)=(N_{Ti} \,.\, N_{Tj} \,.\, (N^2-N_{Ti} \,.N_{Tj})/N^3 \quad (8)$$

The generalized likelihood ratio test of Dunning is the ratio λ between the maximum value likelihood function (probability function) under the constraints of the null hypothesis to the

maximum with that constraints relaxed. If the null hypothesis is true, then -2log λ called log-likelihood measure will be asymptotically X^2 distributed with degree of freedom equal to the difference in dimension of the hypothesis Θ and $Θ_0$. Since the likelihood ratio reject the null hypothesis if the value of these statistic is too small, taking the negative algorithm of λ yields a score for the significance which tells by how much the null hypothesis has been missed.

A null hypothesis $H(Θ|T_i)$: $Θ \in Θ_0$ it stated by saying that the parameter Θ is in specified subset of $Θ_o$ of the parameter space Θ. The likelihood function $H(Θ) = H(Θ| T_i)$ is a function of the parameter Θ with T_i held fixed at the value that was actually observed.

The statistical model to be used is the binomial distribution; the general form of the likelihood is then:

$$\lambda = (\max_{\theta \in \Theta_o} H(\Theta|T_i)) / (\max_{\theta \in \Theta} H(\Theta|T_i)) \quad (9)$$

Another way to formulate this as originally done by [19] it is to compare the hypothesis $H(Θ|T_i)$ with Θ as a point in the parameter space Θ and T_i a point in the space of observations.

For the occurrences the single parameter of the statistical model based on the binomial distribution is P1 where as the experimental outcomes can be described by N(the number of experiments) and k(the number of positive outcomes). Since there are two binomial distribution to be compared, the one representing the null hypothesis and the one representing the observed data, we have

$Θ_1 = P_1$ and $T_{i1} = k_1, N_1$ as well as

$Θ_2 = P_2$ and $T_{i2} = k_2, N_2$ thus;

$$H(P_1,P_2;k_1,N_1,k_2,N_2)=P_1^{k1}(1-P_1)^{N1-k1}(^{N1}_{k1})P_2^{k2}(1-P_2)^{N2-k2}(^{N2}_{k2}) \quad (10)$$

The hypothesis that the distribution have the same underlying parameter is represented by $P_1=P_2$. The likelihood ratio for this test is then;

$$\lambda = (\max_{P} H(P,P; k_1,N_1,k_2,N_2)) / (\max_{P1,P2} H(P_1,P_2;k_1,N_1,k_2,N_2) \quad (11)$$

The maxima of the likelihood function are achieved with $P_1 = k_1/N_1$, $P_2= k_2/N_2$ and $P = (k_1+k_2)/(N_1+N_2)$ which reduce the ratio to:

$$\lambda = (\max_{P} L(P;k_1,N_1)L(P;k_1N_1)) / (\max_{P1,P2} L(P_1;k_1,N_1)L(P_2;k_2,N_2)) \text{ with}$$

$$L(P;k,N) = P^k(1-P)^{N-k} \quad (12)$$

Taking the logarithm of the likelihood ratio gives;

$$-2\log\lambda=2[\log L(P_1;k_1,N_1)+\log L(P_2;k_2,N_2) - \log L((P;k_1,N_1) - \text{Log } L(P;k_2,N_2)] \quad (13)$$

From the contingency table it follows that $K_1= N_{TiTj}$, $N_1= N_{Tj}$, $k_2= N_{Ti} - N_{TiTj}$ and $N_2 = N - N_{Tj}$

Using these equations the final equation for the log-likelihood ratio:

$$-2\log\lambda=[N\log N-N_{Ti}\log N_{Ti}-N_{Tj}\log N_{Tj}+N_{TiTj}\log N_{TiTj}+(N-N_{Ti}-Tj+N_{TiTj})\log(N-N_{Ti}-N_{Tj}+N_{TiTj}+(N_{Ti}-N_{TiTj})\log(N_{Ti}-N_{TiTj})+(N_{Tj}-N_{TiTj})\log(N_{Tj}-N_{TiTj})-(N-N_{Ti})\log(N-N_{Ti})-(N-N_{Tj})\log(N-N_{Tj})] \quad (14)$$

This measure has got a momentum of acceptance and successfully employed for the general computation of word association [48].

2.6 Wolaytta Language

Wolaytta language is defined as a language spoken by the people of the wolaytta zone. Wolaytta language belongs to the Afro-Asiatic phylum under the subdivision of Omotic language family spoken by about 2,000,000 people [5]. Wolaytta is an evolving and a relatively large language based on the number of speakers. For instance from the 84 living language of Ethiopia, Only Amharic, Oromo, Somali, Tigrigna, Sidamo, and Gamo-Gofa-Dawro surpass Wolaytta based on the number of speakers[52].

2.6.1 Wolaytta Writing System

Wolaytta is an official language of literacy in Ethiopia, and its writing system was established in Latin script since 1970's [50]. The Latin alphabet is now adopted in mother tongue education system and various text books are being published using it [51]. Significant numbers of translation works from other language to wolaytta are also being implemented. Accordingly there were about 10,000 hard copy documents already available in the language (most of them are religious written before 15-25 years ago). The application of computer in offices and educational institutions is also significantly increased [6].

Most or almost all current publications basically use the writing system of Wolaytta with Latin alphabets called WPXW which literally mean "The custom in writing the Wolaytta letters" published in 1985 E.C by WQHMK [53]. Appendix 1; has the list of Wolaytta alphabet in WPXW.

WPXW's writing is as convenient as Qubee, which is a system based on Latin Alphabet for writing the Oromo language (one of the majors languages in Ethiopia that belongs to the Cushitic family of the Afro-asiatic phylum)

2.6.2 Syllable Structure, Morphemes and Phonetics

Wolaytta syllable can be formulated as CV (V) (C), where C and V stand for consonant and vowel respectively. Consonant and vowel in parenthesis signify optional syllable segment. Thus the general syllable formulation is applicable to any position of the word [55].

Morphemes are of two types; free and bound. Free morphemes are the one that stand as a word on its own whereas bound morphemes do not [52]. Adam indicated in [53] that, Wolaytta is characterized to have both types of morphemes.

Word in Wolaytta can be formed by the process of affixation and compounding. Affixes are morphemes that stand on its own as a word independently. From the three types of affixes Prefix, suffix and infix Wolaytta only uses suffixation as the process of word formation. A single word in Wolaytta can have a number of variations. This is the case because Wolaytta word is formed from successive addition of suffix that results in relatively long word [53].

Though that it is irregular, Wolaytta word can be formed by compounding [55]. According to Lamberti and Sottile, compounding is the combination of two linguistic forms with independent function.

2.6.3 Wolaytta Word Classes

According to Wakasa in [54] most of the common words in Wolaytta can be divided in to two major category, lexical stem and grammatical ending. However, some but not rare words are atomic, i.e. they do not inflect. Based on the presence or absence and nature of grammatical ending, Wolaytta words are classified in to three major classes; nominal, verb and indeclinable. Verb in Wolaytta are words that have suffix at the end and varies according to the person. Indeclinable are words that are atomic (they do not inflect).

Most adverbial words in Wolaytta are declinable, and they can be regarded as nominal although they may be fairly "defective", i.e. their inflection may be fairly incomplete. Some of them are actually indeclinable, but can be regarded as nominal based on their endings. Those words that resembles and which are not verb or indeclinable are grouped to nominal class. Accordingly nominal class has a number of subclasses that make the classifying scheme complicated and includes; noun, numerical expression, preposition, demonstrative expression, interrogative expression, and post preposition. As the purpose of this thesis is to construct automatic thesaurus for Wolayitta language and as nouns in most cases are content bearing terms the subsequent section will focus on Wolaytta nouns

2.6.4 Wolaytta Nouns

Based on the syllable formulation of Lamberti and Sottile in [55] Wolaytta nouns have the form C1V1C2V2, where C and V generally stand for consonant and a vowel respectively. However C1 represents a glottal stop while V1 seldom represents diphthong. In the same way C2 may represents simple or germinated constant and V2 either represents ending in the absolutive case or ending required by syntactic function associated with noun stems. Most Wolaytta words are bi-radical (the number of consonants in the word) however there are some pluriradical words.

Table 4: illustrates the general syllable formulation of a bi-radical Wolaytta nouns form.

Character Sequence	Example	Meaning
CVCV	Kaisoi	Theaf
CVC:V:	Shappa	River
CV:CV:	keettaa	House
CV:C:V:	Maattaa	Milk

Table 4: General syllable formulation of a bi-radical Wolaytta nouns form.

Most noun in Wolaytta are bi-radical [6], however, pluri-radical forms are also common in Wolaytta nouns.Table 5 illustrate the pluri-radical Wolaytta nouns form

Noun	Meaning
gisttiyaa	Wheat
gallassatuppe	Day

Table 5: Pluri-radical Wolaytta noun

As Adem discussed in [56], Wolaytta noun are formed from stems and suffixes in which the stems never change while suffixes exhibit changes. Table 6 is an illustration of his discussion.

Nouns	Stems	Suffixes	Meaning
keett-a	Keett	a	a house (absolutive)
keett-I	Keett	i	a house (nominative)
keett-eta	Keett	eta	a house (absolutive)
na?-a	Na?	a	a child (absolutive)
na?-eti	Na?	eti	children (nominative)

Table 6: Wolaytta noun formation from stems and suffix

According to the ending they take in inflection Wolaytta noun is classified in to four major classes [55]. First class nouns are nouns ending in –a in absolutive case with the stress on the last syllable. The second class nouns are noun that have an absolutive case ending in –iya and the stress is on the penultimate. The third and the fourth class nouns are nouns ending in uwa and –iyu respectively in the absolutive case. Table 7 shows the four classes of nouns with the ending they take in their inflection in absolutive case.

Noun Class	**Noun**	**Ending**	**Meaning**
First	awa	a	Sun-rise
Second	mezziya	iya	Tutor
Third	kawuwa	uwa	Government
Fourth	Kaniyu	iyu	Dog feminine

Table 7: Classes of nouns with the ending they take in their inflection in absolutive case.

All nouns that belong to the three classes, i.e. first, second and third are masculine and take the ending –a in absolutive case. However the fourth class feminine with ending –u in abssaultive case [55].

Masculine	**Feminine**
kana (male dog)	kaniyu (female dog)
kuttuwa(cock)	kuttiyu (hen)
aawa(father)	ayu(mother)
agg-a (his)	aar-o (her)

Table 8: Gender illustration in Wolaytta noums.

Wolaytta nouns also exhibit number as singular and plural. Singular noun is noun in its basic noun form while plural is formed by adding suffix to the basic noun form[55]. The plural from of the basic noun can be formed by morpheme –tv, where –v represents a final vowel that changes with respect to the case inflection of plural form. In absolutive case there is a direct correspondence between –v and –a as a result the plural marker takes –ta. In absolutive case first, second and third classes take –a -ta, -e- ta and -o –ta respectively. Nevertheless the fourth class feminine noun will form plural of the masculine counterpart. If it fails to have masculine counterpart, it exactly takes the form of second class.

Noun class	**Singular**	**Plural**
First	doorissiya(sheep feminine)	doors-a-ta (sheeps)
Second	biraxxiya(finger)	bir-e-ta (fingers)
Third	oosouwa (work)	oos-o-ta(works)
Fourth	dabbiyu (female relative)	dabb-a-ta(female relatives)

Table 9: Number illustration in Wolaytta nouns

2.7 Related Works

The subsequent section discuses some of the related works obtained from reviewing literature on automatic thesaurus construction both in local and foreign languages.

2.7.1 Amharic Thesaurus

Andargachew Mekonen [57], has constructed automatic thesaurus on Amharic Bible documents using word space model and evaluated its performance using indexing and searching system.

To generate thesaurus for each terms, the terms should be normalized for spelling inconsistency and stemmed. Normalization of spelling inconsistency of the stem is done in the same manner as done for index term. To stem the term Amharic stemmer implemented in 2007 in Java by Tessema was used.

After the term is normalized and stemmed the next step opens WORDSPACE model and to select a term vector that corresponds to the term for which thesaurus is required. To open the WORDSPACE model MMapDictionary.openinput ucene API was used.

To create and normalize a term vector for the given term getAdditiveQuery Vector method of CompoundVector builder semantic vectors' API was used. The method return a normalized term vector created by adding together vector retrieved from WORDSPACE model. Nearest neighbors of the term vector are computed by utilizing VectorSearchNeighbor method works as described below.

Cosine similarity based on the term vector and all the other term vectors were used in the model. The result provides a candidate list vectors from which the nearest neighbor was to be chosen. The criteria for choosing the nearest neighbor for the term are the highest cosine similarity value. That is a term has higher cosine similarity value with the term, for which the thesaurus is required, is more related to the term than another term with lower cosine similarity value.

To accept a query term and to display its thesaurus terms a graphical user interface was developed. The user interface was created using Swing GUI components, which are part of the Java programming platform.

In order to generate thesaurus terms that was used for query expansion, the WORDSPACE model was searched with the query term that was provided. The top five most related thesaurus terms for the query were used for query expansion. Finally the precision and recall value were computed and the evaluation result indicated that the average recall values were 37.29% and 73.34% before and after using thesaurus for query expansion, respectively. The corresponding average precision were 88% , 44%, and 31% after using thesaurus.

2.7.2 Tigrigna Thesaurus

Hagos Hiete [58], has developed automatic thesaurus for Tigrigna text retrieval from document collection based on co-occurrence approach and achieved a remarkable output from heterogeneous documents.

In the architecture he used to develop automatic Tigrigna thesaurus, the main tasks performed includes, preprocessing Tigrigna text, vocabulary construction, index term weighing, co-occurrence metrics formulation, and similarity computation.

The preprocessing stage includes: transliteration, tokenization, removal of stop words and punctuations, stemming, and normalization. The Tigrigna documents were first transliterated to Latin using the system for Ethiopic representation in ASCII(SERA) transliteration scheme. This is the case because the converted Latin scripts were convenient for stemming and internal

processing of documents. Then what followed was 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.

Stemming process was conducted after transliteration of the Tigrigna script using SERA transliteration scheme in to alphabetic. The stemmer developed removes prefixes, and suffixes only, other morphologies were not considered. As Tigrigna language is characterized by words of varies usages, there are no rules where to use those varying words. Through normalization of all words with varying forms but having the same meaning were converted to the common form.

When constructing vocabulary the main criteria should be sizable and representative of the subject area: and the determination of the required specificity for the thesaurus Kanyarat,L(cited by Hagos, 2011). Hogos finally identified the index term then develop 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 are generated, top five were selected for easy management. Otherwise top ten were selected. The result has showed that 75.28% of the output has related concepts to the respective terms. The remaining 24.72% identified not to have related terms.

2.7.3 Automatic Thesaurus Generation for Chinese Documents

Yuen-Hsin [59], reported an approach to automatic thesaurus construction for Chinese documents. The approach includes an effective Chinese keyword extraction algorithm. The keyword extraction from each documents were further filtered for term association analysis. Association weighted larger than a threshold are accumulated to over all documents to yield the final term similarities. Consequently the method speeds up the thesaurus generation drastically.

Identifying words in a Chinese text was somewhat like identifying multiword's phrase in English documents, because both no lexical delimiters. Thus a term segmentation and selection process required before term co-occurrence can be analyzed. To segment Chinese word Yuen-Hsin proposed a hybrid approach that combines a single dictionary based segmentation method and a final keyword/key phrase extraction algorithm. The algorithm converts the input string into an ordered list with each elements bring an un segmented Chinese character or a segmented word. Occurrence frequency in accumulated frequency list elements produced in the algorithm.

Then in the next step element in the list were merged, dropped, accepted iteratively until no enough elements in the list were merged back in to a larger input string if both of their frequencies exceeded a predefined threshold. Otherwise, the first element of the adjacent pairs was dropped. If it was a low frequency term, or accepted as a keyword if it was a high frequency term and not merged with its preceding elements.

He found that a longest repeated string often was correct word (or phrase), because its repetition provides sufficient extraction to decide on the left and right boundaries word. Similarly, a repeated string that subsumes the other may also be likely to be legal term. The experiment showed that for each document an average of 33% keyword unknown to a lexicon of 123,226

terms could be identified by the algorithm employed. Of these unregistered words, only 8.3% of them are illegal. Keywords extracted from each document are further filtered for term association analysis. Association weight larger than a threshold is then accumulated over the documents to yield the final term pair similarities. The method is found to speed up the thesaurus generation process drastically. It also achieves a similar percentage level of term relatedness.

2.7.4 Automatic Thesaurus Construction for Thai Text Retrieval Enhancement

Chaiwak K. Nattakan P and Asannee K.[60] proposed a Thai association thesaurus, based on the statistical techniques and natural language processing techniques.

Automatic Thai thesaurus construction involves three steps: lexical preprocessing, thesaurus term association, relation extraction, and term/ relation organization. The lexical preprocessing is the first step of automatic Thai thesaurus construction. It is a well known fact that Thai text is a sequence of word with no delimiters. Thus unlike English and other European languages, identifying words root of the Thai text becomes a difficult task. The Thai text processing is concerned; the word segmentation process is required. The output of this step is the document the tagged corpus is word segmentation and part of speech tagging.

During the term acquisition, they extracted the term from tagged corpus that was received from the previous step. At this step all single noun and compound noun from tagged corpus that analyzed noun phrase extracted. During relation extraction the relation of thesaurus term was extracted. The relation in the thesaurus is broader, narrower, and related relation. Finally in the Thai term/relation organization steps terms were organized according to the relation received from the previous step.

The researcher concluded that the major objective of constructing Thai thesaurus automatically is useful for enhancing the performance of Thai text retrieval. However, for broader and narrower relation extraction the design was ready for implementation.

CHAPTER THREE

Development of an Association Thesaurus

As the objective of the research is to construct automatic thesaurus from Wolaytta text, the architecture of the system [61] is designed in a way as depicted in figure 1**.**

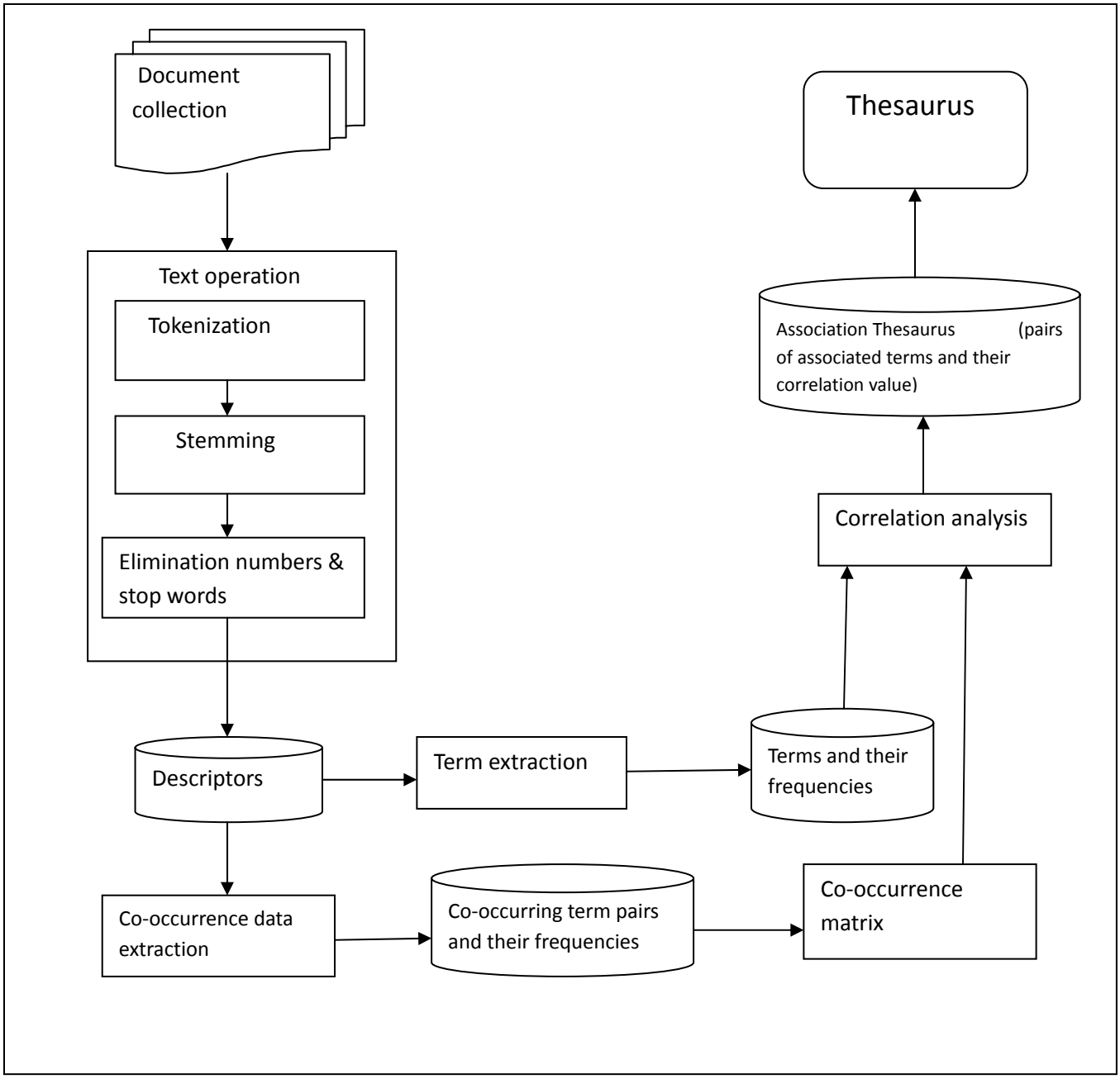

Figure 1: System Architecture for Automatic Generation of Association Thesaurus

In order to construct automatic thesaurus from Wolaytta document, the document should go through text operation steps in order to obtain representative terms that can be used as an entry for thesaurus terms. To develop automatic thesaurus, the extraction of terms and their co-occurrences information's should also be employed so that the correlation analysis between the occurrence frequencies of terms in the document and their co-occurrence frequencies in the co-occurrence matrix will identify pairs of associated terms that qualify for the entry of association thesaurus. Hence, the subsequent subsections discuses all the steps involved as depicted in the architecture in chronological order.

3.1 Preprocessing

Modern information retrieval system should be designed in a way to retrieve information of any nature. The system architecture requires that information should be organized into units with frequent occurrences in the data. Preprocessing basically has a number of text operation steps to optimize the list of terms for thesaurus term entry. The optimization process is important to reduce the number of representative terms required by eliminating those with poor information content.

Preprocessing takes one or more raw documents as an input and produces a set of tokens as an output. Tokens are specific occurrences of terms in document which represent the smallest unit of meaning [62]. 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.

Preprocessing has two basic text operation approaches; normalization and removal (cleansing) of elements from the original text, with insignificant information or introduce noise in the retrieval performance [31]. In this research, the New Testament bible corpora were preprocessed so that the removal or cleansing of elements can be performed using tokenization, elimination of stop words and stemming procedure and discussed as follows.

3.1.1 Tokenization

Tokenization is the first text operation step in preprocessing in information Retrieval. 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 in to meaningful units called tokens. A token is an instance of a sequence of characters in some particular document that are grouped as a useful semantic unit for further processing.

Tokenization is useful both in linguistics, where it is a form of text segmentation and in computational science where it froms part of lexical analysis [63]. In the tokenization phase of text operation the major issue that require due attention is what are the correct token form 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 [64].

In Wolaytta writting system words are delimeted by white space and all the punctuation marks that are commonly used in English writing system such as fullstop, comma, semicolon, single quotation, double quotation, exclamation marks, and question marks are also applicable in Wolaytta writing system. As depicted in figure 1 of system architecture Wolaytta documents are subjected to tokenization phase of test operations in which the stream of texts were splited into sequence of meaningful unit called token. The approach implemented in tokenization is chopping text at its word boarder at the same time removing all the punctuation marks with in the documents.As the issue of cases should be resolved in tokenization phase, all words with upper cases of the alphabet were converted to lower cases. Both removals of punctuation marks and cases resolution were performed by importing a built in class re from Python with a method split so that the stream of texts were splited in to stream of tokens.

3.1.2 Elimination of Numbers and Stop Words

Stop words 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 [45].

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 stop words and includes; conjunction, preposition, articles, and adverbs. Words of such type contained in a document are considered to be worthless as index terms.

When searching for index terms that contain stop words requires all the record contained in the database to be checked despite its irrelevancy. Lots of processing time, 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 [65].

Stop words make up large portion of the text of most documents and if removed the percentage that the corpus is being decrease will be about 70-75% as compared to otherwise. Sop word elimination is performed by filtering the term list using “stop-list” also called syncategorematic or non-content bearing words [66].

In the stop word elimination phase of text operation the New Testament Bible corpora written in Wolaytta were preprocessed in order to eliminate words with no additional meaning to the actual context of the text. In the first phase of the experiment, after tokenization was performed, all words with their occurrence frequencies were extraceted and sorted based on their frequencies in excel worksheet. As words which occur frequently have less power in discriminating one document from another, the top 127 words were removed and compiled as a stop word list. However, evaluating the performance of the stop word compilation procedure, it was identified that the compiled stop word list was not exhaustive enough to completely filter out stop words from the test data set. For example words like hegaa, wotti, azattiyooga, woigan, sinttau, etc, should have been removed as they belong to stop word list. Therefore in the second phase of the experiment, the entire training data set was scanned and stop words are compiled manually ineffect improved the performance of stop words compilation. Appendix 3 has the complete list of manually compiled stop words.

Number removal from text was a challenging task as Wolaytta word formation utilizes number 7 as a glottal stop in words like; de7ees, lo77o, de7uwaa, go7aara, poo7issiyaagaa, ha7iya etc. Since number removal is problematic in Wolaytta text, all the numbers associated with each articles in the documents manually scanned and removed for the subsequent phases of text operation.

3.1.2 Stemming

Most languages are characterized by words with different morphological variants, yet the same semantic interpretation with respect to information retrieval system. The attempt to reduce word variants to its stem or root form is the prime concern of stemming.

Stemming can be Linguistic, automatic or mixed. Linguistic stemming use 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, 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 [19].

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 [67].

After tokenization the stream of tokens were subjected to the stemming algorithm developed by Lemma Lessa so that varieant forms of a word were reduced to a common representative stem.The stemming algorithm developed by Lemma was both context sensitive and rule based that strips out suffixs one after the other iteratively from right to left until the root word with reduced radicals were obtained. The original work of the stemming algorithm developed by Lemma was further enhanced by adding some additional rules so as to resolve the problem of over stemmed and under stemmed words identified during the evaluation of direct introduction to the training data set. See table 14 for the performance of the direct introduction of the stemming algorithm.The following is the improved version of the stemming algorithm employed in the automatic thesaurus generation system.

```
function stemmer(token,radical)
  if radical is 1 then,
          if token ends with 'sh' then,
                  return token 'sh' replaced by 'y'
          else if token ends with 'ee' then,
                  return token with striped 'ee'
          else if token ends with 'aa' or 'a' then,
                  return token
 else if radical is 2  then
    for suffix in suffix list
          if token ends with suffix then
                token=token with striped suffix
                radical=radical-1
                return stemmer(token,radical)
 else if radical > 2   then,
      for suffix in suffix list
          if token ends with  suffix then
                token=token with striped suffix
                radical=radical-2
      return stemmer(token,radical)
 else
       return token
```

3.1.4Normalization

Normalization is the process by which terms will be grouped in to equivalent classes. Each equivalent classes is represented by the most frequently terms selected from each classes 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 could 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 amount of relevant documents retrieved with enhanced recall of the system [50].

There are different approaches to normalization and it is language dependent. However, this research thesis 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 rule to map terms to stem [68]. 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 stop words 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.

As normalization approach, both stemming and elimination of stop words were implemented on Wolaytta text so that the eprules of the stemming algorithm mapped terms into the same stem while elimination of stop words removed frequent terms from that training data set. The

stemming algorithm mapped terms like, dandayettenna;danddayaa, danddayaasa, danaaayai, danddaana,

danaddayees,danddayeetii,danddayekketa,danddaynnaagaa,danddayettees,danddayettennan,danddayibeenna,danddaibookkona,danddaayii,danddayikko,danddayiya,danddayiyai,

danddayiyo,danddayiyagaa,danddayokkona, danddayoos to dandday. While most of the stop words have been removed manually after they were stemmed as listed in appendix 3.

Inconsistencies in alphabetical composition of words were also an issue that required a normalization step as some word pairs like; yeessusa with yuusssa, maxaafa with mataafa, etc were written to mean the same thing. As a result inconsistent words of such type with the same meaning manually scanned and replaced with one of the two forms.

3.2 Term Extraction

Terms in most cases nouns that represent domain specific concepts will be extracted from documents. These terms are the most frequently occuring terms in the documents that exceeds the predefined frequency threshold. Stop word list should also be constructed as there are cases where most frequently occuring nouns might fail to be terms.

3.3 Co-occurrence Data Extraction

The Concern here is to extract semantically or conceptually associated terms regardless of the type of association. The entire sets of co-occurring term and frequent terms will be extracted. If a term appears frequently with a particular subset of terms, the term is likely to have important meaning. Every term that occurs together with in a sentence will be extracted. A sentence is a set

of terms separated by space. Two terms in a sentence in co-occurrence extraction can be considered to co-occur once considering a sentence as a bag of terms, regardless of the order of terms and the grammatical information[61].

3.3.1 Co-occurrence Matrics

Term co-occurrences analysis is one of the approaches used in information retrieval research for forming multi-phrase terms. Co-occurrence analysis on the other hand, is a statistical approach where the association of terms in documents, chapters or some other unit is computed. The basic idea is; the closer the words occur, the more significant in the co-occurrences. Many automatic indexing methods do not identify how close words occur, 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 [21].

Co-occurrence matrics is a matricx that displays the number of times word co-occurs with the other words in a document, chapter or in a window of a number of words [27]. They describe the matrics as a term to term co-occurrences matrics CF where elements $CF(T_i,T_j)$ records the number of times the term T_i and T_j co-occur in a window of size k. The association between the two words can then be computed using mutual information (MI) by taking the values of occurrence frequency count of a word and the co-occurrence frequency count from the co-occurrences matrics.

In co-occurrence extraction if there are N different terms in the document, NXN co-occurrence matrics will be constructed by counting the frequency of pair wise term co-occurrences [19]. Table 10 illustrates the NXN co-occurrence matrics pair wise term co-occurrences.

	T_1	T_2	T_3	T_4	T_5	-----	T_N
T_1	------	$CF(T_2, T_1)$	$CF(T_3, T_1)$	$CF(T_4, T_1)$	$CF(T_5, T_1)$	-----	$CF(T_N, T_1)$
T_2	$CF(T_1, T_2)$	------	$CF(T_3, T_2)$	$CF(T_4, T_2)$	$CF(T_5, T_2)$	-----	$CF(T_N, T_2)$
T_3	$CF(T_1, T_3)$	$CF(T_2, T_3)$	-----	$CF(T_4, T_3)$	$CF(T_5, T_3)$	-----	$CF(T_N, T_3)$
T_4	$CF(T_1, T_4)$	$CF(T_2, T_2)$	$CF(T_3, T_4)$	-----	$CF(T_5, T_4)$	-----	$CF(T_N, T_4)$
T_5	$CF(T_1, T_5)$	$CF(T_2, T_5)$	$CF(T_3, T_5)$	$CF(T_4, T_5)$	-----	-----	$CF(T_N, T_5)$
----	------	-------	------	------	------	-----	-----
T_N	$CF(T_1, T_N)$	$CF(T_2, T_N)$	$CF(T_3, T_N)$	$CF(T_4, T_N)$	$CF(T_5, T_N)$	-----	-----

Table 10: NXN co-occurrence matrics pairwise term co-occurrences.

In the above table term T_i and term T_j co-occur in co-occurrence frequency $CF(T_i,T_j)$ sentences in the document.

3.4 Correlation Analysis

Mutual information (MI) is a method for measuring association between terms [21]. The mutual Information between term T_i and term T_j is defined as follows;

$$MI(T_i,T_j)= \log_2 (CF(T_i,T_j)/\sum(T_i,T_j))/\{F(T_i)/\sum F(T_i)\}*\{F(T_j)/\sum F(T_j)\} \quad (15)$$

Where $F(T_i)$ is the occurrence frequency of term T_i and $CF(T_i,T_j)$ is the co-occurrence frequencies of term T_i and T_j. The threshold for mutual information of the maximum number of associated terms for each term should be predefined and associated terms are selected based on the descending order of mutual information. There is limitation in mutual information in that it overestimates low frequency terms [21], as a Log likelihood ratio will be employed to determine whether two terms are related. Pairs of terms that fail to pass the log likelihood ratio test will be filtered out.

CHAPTER FOUR

Experiment and Evaluation of the System Performance

Automatic thesaurus generation system from Wolaytta text is implemented on New Testament Bible corpora comprised of twenty seven books and 260 documents with Wolaytta language.The corpora are obtained from a website with URL : http://gospelgo.com distributed by Gospel Go under the project description of multilingual corpus of the bible.This chapter discuses the critical methods that have paramount importanct for conducting the experiment and evaluation of the experimental result.

4.1 Corpus collection

The search for standard corpus for automatic association thesaurus generation from Wolaytta text was not an easy task as there are hardly any documents in the language that can be used as as an ordinary corpus leave alone the standard one. Despite the availability of ten thousands or more documents [6], documents in electronic format is still in scarce. Most of the resources in the language available in WWW are either in audio or video file formats. However for this research work, The New Testament Bible corpora in Wolaytta language with twenty seven books 260 documents with a total words count of 126,883 where obtained from the website with URL:http://gospelgo.com

The fact that documents, in the New Testament written in the language, are found to be detailed in context and exhaustiveness of words distribution and morphology are the major characteristics taken in to consideration on selecting the corpus. Further more the New Testament Bible is the only compiled huge electronic documents obtained in the WWW except for audiovisual resources.

The original corpora collected from the web was in XML format with different HTML tags anchered with in the entire documents. Yet the HTML tags and all numbers associated with each articles have been 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..

|The 126,883words of the New Testament is further pre-processed so that stop words and rare words that account 80% of the document corpora are removed. The remaining 20% of the corpora with a total of 25,232 words are divided in to training and test datasets in which the training dataset composed of all the twenty six books with a total word count of 19,561while twelve documents of Romans with 5671 words were taken as a test dataset.

Based on the evidence obtained from one religious leader Romans Three was candidate for selection as a test dataset by considering the fact that most of the contextual contents in the entire documents are somewhat included in this document. However, in order to make the percentage of the test dataset optimal, documents from Roman three to fourteen are included in a test dataset.

Due to the contents and richness of the word distribution in the New Testaments written in the language, both the training and the test datasets of the corpora can be considered as representative to Wolaytta language. The training datasets with 77.52 percent is used for generating thesaurus terms whereas the test datasets that accounts for 22.48 percent of the corpora is an independent instance from the terms generation process [23].

Training Dataset		Test Dataset	
Book	**Document Available**	**Book**	**Documents Available**
Matthew	**28**	**Romans**	**12**
Mark	**16**		
Luke	**24**		
John	**21**		
Acts	**28**		
Romans	**2**		
Corinthians 1	**16**		
Corinthians 2	**13**		
Galatiians	**6**		
Ephesians	**6**		
Philippians	**4**		
Colossians	**4**		
Thessalonians 1	**5**		
Thessalonians 2	**3**		
Timothy 1	**6**		
Timothy 2	**4**		
Titus	**3**		
Philemon	**1**		
Hebrews	**13**		
Jemes	**5**		
Peter 1	**5**		
Peter 2	**3**		
John 1	**5**		
John 2	**1**		
John 3	**1**		
Jude	**1**		
Revelation	**22**		
Total Documents	**246**	**Total Documents**	**12**
Total Words count	**19561**	**Total Words count**	**5671**
Percentage	**77.52**	**Percentage**	**22.48**

Table 11: Corpus statistics of the training and test dataset

4.2 Thesaurus Generation

Development of co-occurrence Metrics

Developing co-occurrence metrics is the most important step for thesaurus generation process. After text operation steps are performed on the collected corpora, the most frequent terms, descriptor, are generated for the thesaurus term entry. For each terms in the entry their co-occuring terms also generated and the co-occurrence information obtained are kept in term to term co-occurrence matrics employed in the system.

Terms	aihuda	ammana	nagara	haiqqo	maxaafa	asaa	yaaga	yaati	asa	goda	ha77i	xill	xoosa	gishsha
aihuda		5	1	0	6	7	6	4	11	3	0	6	8	6
ammana	5		7	3	9	7	15	10	12	10	3	14	17	8
nagara	1	7		10	4	5	11	7	17	10	4	15	17	17
haiqqo	0	3	10		1	0	3	3	6	6	1	4	7	11
maxaafa	6	9	4	1		7	15	6	11	9	1	6	16	6
asaa	7	7	5	0	7		11	4	14	4	0	10	8	4
yaaga	6	15	11	3	15	11		13	23	17	5	10	33	18
yaati	4	10	7	3	6	4	13		10	7	1	17	15	9
asa	11	12	23	6	11	14	23	10		4	3	20	29	19
goda	3	10	10	6	9	4	17	7	4		7	13	23	19
ha77i	0	3	4	1	1	0	5	1		7		3	7	8
xill	6	14	15	4	6	10	17	11	20	13	3		23	15
xoosa	8	17	17	7	16	8	33	15	29	23	7	23		27
gishsha	6	8	17	11	6	4	18	9	19	19	8	15	27	

Table 12: Term to term co-occurrence metrics

Based on the independent occurrence frequency of thesaurus terms and their dependent co-occurrence frequencies from co-occurrence matrices, a module in the system computes and returns the correlation values between pair wise terms. Based on the weight of correlation values obtained from the mutual information the system generate association thesaurus.

Table 13 : Correlation values between associated terms.

4.3 System Implementation

For system implementation Python 3.1.2 programming language is a language of choice. Python is a powerful, easy to learn, and has an efficient high-level data structure. Python with its elegant syntax, dynamic typing and object oriented paradigm is an ideal programming language for scripting and application development platform.

Python is an interpreted language, its interpreter is easily extended with a new function and data types implemented in C or C++, which can save considerable time during program development because no application and linking is necessary. Python enables programme to be written completely readable. Programme written in python typically are much shorter than equivalent C,C++, or Java programs.

Python makes things easier to perform a search and replace over a large number of text files, or rename and rearrange a bunch of files; with much more error checking mechanism than C. Python has a high -level data type built in such as flexible arrays and dictionary which makes it suitable for much larger problem domain.

Using Python 3.2.1 development environment automatic association thesaurus is developed on a training data set of the New Testament of bible written in Wolaytta. Nevertheless, the system developed is domain independent in a way that can be used in a range of environment. This is due to the fact that the developed system is based on a training dataset with extensive morphology and word distribution of Wolaytta language.

The performance of the system is also tested using a good percentage of the test dataset of Romans three to fourteen which is contextually included in a training dataset, yet independent

from it in the development process. After the programme is developed the system is run on a test dataset and its performance is evaluated.

4.4 Evaluation of Stemming and Stop words

In this thesis stemming and stop word evaluation is also performed as part of the overall evaluation of thesaurus generation system. As stemming process and stop words compilation are the basic steps towards generating automatic thesaurus, have a relevant consequences to the performance of thesaurus generation.

The evaluation is conducted in a way that 20 terms from the stemmed test data set are selected by simple random sampling so that the performance of the stemming algorithm employed evaluated by checking whether terms are stemmed to their root form properly or not. At the same time stemmed terms are inspected whether they belong to stop words list or not in order to access the accuracy and exhaustiveness of the stop words compilation process. Table 14 depicts the performance of the stemming algorithm and the accuracy and exhaustiveness of stop words compilation step.

NO	`Stemmed term	Properly stemmed	Stop word
1	hegaa	Yes	Yes
2	eraasa	No	No
3	waati	Yes	Yes
4	azattiyooga	Yes	Yes
5	Higgiyaa	No	No
6	Ooso	No	No
7	Xillotetta	No	No
8	Paxa	No	No
9	woigan	Yes	Yes
10	hearetta	No	No
11	huphen	No	No
12	wayyiyaa	No	No
13	ment	Yes	Yes
14	godai	No	No
15	yootyaab	No	No
16	sinttau	Yes	Yes
17	aibanne	Yes	Yes
18	ottidg	Yes	Yes
19	higga	No	No
20	siiqiyaa	No	No
Percentage		60	45

Table 14: Evaluation of stemming and stop words in the first phase of the experiments.

4.5 Interpretation of Evaluation

In this research, the stemming algorithm developed by Lemma is implemented to confilate Wolaytta words to their root form. Ttable 15 depictes the performance of the stemming algorithm on bible corpora of the test dataset. The declination of the erformance on this particular test dataset is due to (1) Wolaytta language is complex in which most Wolaytta words are formed from continues process of suffisation. The compiled suffix list during system development was not sufficient enough and requires a complete list of compilation by language experts.(2) inconsistency in alphabetical composition of words caused the same words to be written more than once that results in different word stem. For example the following pair of words are inconsistent ; asho with ashsha, benbba with bainnabaa, gooda with gaada,yeesusa with yuususa and maxaafa with mataafa (3) The stemming algorithm developed by Lemma is both context sensitive and rule based. Yet the rules that he applied in the algorithm were not complete enough to handle every word in the bible corpora of the test dataset.

As indicated in table 15, 45 % of words are stop word that made the accuracy of the stop word compilation process 55%. This is due to tha fact that, stop words compilation process is incomplete and requires standard stop word list compilation by language expert. Further more some words in Wolaytta are formed from derived nouns and verbs and when comilated will take forms of stop word.

In the first experiment, 40% of the words were not properly stemmed, that is, they are either overstemmed of understemmed. Hence to enhannce the performance of the stemming algorithm on a bible corpora test dataset, additional suffixes are manually scaned from the entire training dataset and incorporated to the suffix list already developed. Besides some additional rules and

exceptions are incorporated to the stemming algorithm so that it can handle more words to be confilated to their root froms.

In the first phase of the experiment 45% of the stemmed file consists of stop words which they should have completely been removed from stemmed file that has a list of candidate terms for thesaurus terms entry. Therefore, stop words are manually recompiled from the entire training dataset so as to incorporate significant number of words left over during the first phase of the experiment.

In Wolaytta, some similar words can have different suffixes depending on number, gender, case and class, but when confilated they will have the same root form. Therefore, the algorithm is modified to check for terms in stop word list after words are confilated before introducing them in to thesautus terms enry. Examples of such words include; aggaasu, aggada, agganna, aggidi, aggidosona, aggiis, aggikko, haa, ahai, aiba, aibakko, aibee, ashshiyaagee, baada and baala.

In the second phase of the experiment, after the improvement on the first phase, 25% of the words are not properly stemmed; whch makes the performance of the stemming algorithm to be 75%. 20% of the terms are stop words, which they should have been removed; consequently the accuracy of the stop word compilation is 80%. The result still mindicates that further work need to be done to improve the performance of the stemming algorithm and the compilation of of stop words and suffix list with responsible domain experts.

.NO	`Stemmed term	Properly stemmed	Stop word
1	gishsha	Yes	No
2	shaaramux	Yes	No
3	pirdda	Yes	No
4	xill	No	Yes
5	amm	No	No
6	yatti	Yes	No
7	nagar	Yes	No
8	asatett	Yes	No
9	qoppi	Yes	Yes
10	ashsha	Yes	Yes
11	waayett	Yes	No
12	mer	Yes	No
13	kawo	Yes	No
14	yeell	No	Yes
15	qaal	Yes	No
16	asa	Yes	No
17	giddab	No	No
18	gencc	No	No
19	der	Yes	No
20	galla	Yes	No
Percentage	75	70	20

Table 15: Evaluation of stemming and stop words in the second pahse of the experiment

4.5 Evaluation of Thesaurus Terms

For evaluating thesaurus terms, the threshold for term frequency were priorly set to 20 terms in order to resolve the problem of overestimation of mutual information for less frequent terms, keep the running time minimal, reduce the dimension of the co-occurance matrics and correlation value. The system is made to generate as much related terms as possible in desceding order of mutual inforrrmmmation (correlation values). However the top ten related terms are selected and the association of pairwise terms then evaluated whether they are related or not. The result of the experiment indicated that 75% of terms are identified to have related concepts, on the other hand 25% terms do not relate to each other.

Terms	Related Concepts	Non related concepts
maxaafa	go da ,asa, xossa , asaa	
Goda	Xillo, ammana, xoosa	
haiqqo	goda,asa	
gishsha		goda , asa ,ha77i, haiqqo
Asaa	asa, xillo, goda, ammana	
Asa	goda, aihda, xillo	
aihuda	asaa	maxaafa
nagar		Haiqqo, asa, xillo
ammana	Xillo, asa	
xoosa	asa	
Persentage	75	25

Table 16 Evaluation of thesaurus terms.

CHAPTER FIVE

Conclusion and Recommendations

5.1 Conclusion

Today, as people are striving to live across the flood of information explosion, there should be an organized means of managing the flood in order to tap it through an appropriate pipe with desired and relevant quality and quantity. Despite the abundance in the availability of information resources, lack of functional system with cross-language communicability deprived the information demand of the society ineffect hindered the desired development.

In addition to the scarceness of functional system that address the cross-language communicability, the major problem associated with information retrieval system is that, it requires users who are usually incapable of determining what they are looking for to explicitly describe their information need. Moreover, the retrieval system by itself often retrieves irrelevant documents due to vocabulary mismatch between query terms and index terms. As a result, corpus dependent automatic association thesaurus is a frame work for possible inception for enhanced system of Wolaytta text retrieval or a base line for future research on cross-languge retrieval system.

Wolaytta is a language that play crucial role in social, political and economic activities for the people of the region. The language is widely used in public service, as a means of instruction and religious events using the Latin scripts already employed. Consequently, this research work attemted to construct automatic association thesaurus in view of the fact that in a very near future there will be a great demand to access the huge information resources through the inevitable ritrieval system in the language or through cross-language retrieval system.

The notion behind choosing automatic thesaurus over the mannual approach is that, on one hand, the focus in modern information retrieval system research is to generate thesaurus automatically. On the other hand, manual construction of thesaurus is a labour intensive, highly conceptual, costly, time consuming, incapable to handle huge document collection, and the difficulty in updating sematic resource regularly.

Thus, this research work attempted to generate domain independent automatic association thesaurus with features of handling huge scalable document collection in a reasonable response time. Corpus independent association thesaurus from Wolaytta text is an automatic thesaurus based on term-to- term cooccurence data extracted from document corpora. In this model, first the most frequent terms are extraceted for thesaurus terms entry. Among the frequent terms, the cooccurences information are extracted and the information is kept in co-occurence matries for due computation of correlation values of mutual information between parwise terms for generating association thesaurus authomatically.

A new Testament bible with 248 documents in Wolaytta were used to train the automatic thesaurus while 12 documents of Roman are used to test it with 77.52% training dataset and 22.48% test dataset word compostion. The automatic association thesaurus system is implemented using python programming language and the evaluation of the system is also performed on thesaurus terms. As stemming and stop word compilations are the critical steps in thesaurus generation process, both are included in the evaluation scheme.

In the stemming and stop word evaluation process, a random sample of 20 terms is selected from stemmed test dataset and evaluated wether they are properly stemmed? And whether they are

stop word or not? The result of the evaluation has indicated that 75% of terms are properly stemmed and 20% of terms are stop word.

In the first phase of the experiments, the stemming algorithm, developed by Lemma, had 86.9% of stemmer performance. However, when directly implemented to the test corpus of the bible, the performance of the stemmer was observed to decline due to the fact that, the rules employed, the suffix list and stop word list compiled were not complete enough to exhaustively handle the test data set of the bible corpus.

The evaluation of thesaurus terms for automatic association thesaurus generation from Walaytta text is also performed on the top ten terms generated in decreasing order of the weight of mutual information values. The experimental result indicated that 75% of terms identified to be related concepts, whereas 25% of the remaing terms is not conceptually related.

Hence, the level of performance achieved in generating thesaurus term from Wolaytta text was 75%. This result would not be achieved if stop words and suffix list were not manually compiled. This emphasizes that, incorporating the manual approach to the automatic system yields a better performance than otherwise. The level of performance achieved so far in this research work could be further enhanced, if standard corpora, complete list of stop words and suffix list are compiled by experts' research in the language.

5.2 Recommendations

The following are some of the research area that requires further future research works.

- Construction of part of speech tagging for Wolaytta language.
- Development of standard corpus for Wolaytta language.
- Construction of morphological analyzer for Wolaytta language.
- Construction of automatic thesaurus using other similarity measures.
- Developing a standard and functional stemmer for Wolaytta language.
- Construction of automatic association thesaurus for cross language retrieval from parallel corpus.
- Corpus dependent authomatic association thesaurus based on cooccurrence data.

Bibliography

[1] Foo,s.,Hui,S.C,Lim,H.K.,Li,H.(2000)"Automatic Thesaurus for Enhanced Chinese Text Retrieval." School of Applied Science, Anayang Technology University, Singapore 639798.

[2] Blair, D.C. and M.E. Maron.(1985). "An Evaluation Of Retrieval Effectivness for a Full-Text Document-Retrieval System. " Commun.ACM,28(3):289-299.

[3] Chaiwat Ketsuwon, Nattaken Pengphon, and Asanee Kawtrakul. "Automatic Thesaurus Extraction for Thai Retrieval Enhancement." NLAP and IISTL, Department of Computer Engineering, faculty of Engineering, Kasetsart University, Bangkok, Thailand 10900.

[4] Harma Imran and Aditi Sharan.(2009,Nov.). "THESAURUS AND QUERY EXPANSION." Department of Computer Science, Jamia Hamdard, School of Computer Science and System Science, Jawaharlal Nehru University, New Delhi, India. ISCSIT, Vol 1, NO 2.

[5] Gordon, Rymond G., Jr.(ed.) (2005). "Ethnology: language of the world." Fifteenth edition. Dallas, Texas: SIL international. Online version. http://www.ethnologue.com/.

[6] Lemma Lessa.(2003,July)." Development Of Stemming Algorithm For Wolayitta Text,"

[7] Abebe, Alemayehu.(2002).Omotic Dialect Pilot Survey Report.

[8] Lishan Adem.(2010). "Ethiopian ICT Sector Performance Review 2009/2010 Towards Evidence Based ICT Policy and Regulation." Volume Two, Policy Paper 9.

[9] Dagbot Soerget."Multilingual Thesauri and Ontology In-Cross-Language retrieval." Colledeg of Library and Information Service, University of Maryland, Collage Park, Mo2074.

[10] Jixi Xu.(1997,May)."Solving The Word Mismatch Problem Through automatic text Analysis."Ph.D.thesis, Department of Computer science, University Of Massachusetts, MA, USA.

[11] Yuen-Hsien Tseng.(2002)." Automatic Thesaurus generation for Chinese Documents." Department of Library Science and Information Science, FuJen Catholic University, 510, chug Cheng Road, Hsinchuang, Taipie, Taiwan, Republic of Chain 242.

[12] Padmini Srinivasan.” THESAURUS CONSTRUCTION.” University of Iowa.

[13] Hearst, M.A.(1992).”Automatic Acquisition of Hyponyms from Large Text Corpora.” Proceeding COLING 92.yuki Yamaski. Corpus Dependent Association Thesaurus for Information Retrieval, Central research laboratory, Software Division, Hitachi Ltd, Tokyo 185-8601, Japan.

[14] Pum-Mo Ryu, et al.” Toward Domain Specific Thesaurus Construction: Divide-and-Conquer Method.",Computer Science Division, KAIST, KORTERM/BOLA 373-1 Guseong-dong Yuseong-gu Daejeon 305-307, Korea. Procedings, pp 69-83.

[15] Arselmo Penas, Felisa Verdejo and Julio Gonzalo.“ Terminology Retrieval: Toward a synergy between Thesaurus and Free text searching.” Dpto.Languajes Sistemas Informaaticos, UNED.

[16] Schubert Foo and Siu Cheungttui, Hong Koonlin, Lihui. “Automatic Thesaurus for Enhanced Chinese Text Retrieval.” School of Applied Science, Nenyang Technology University, Nenyang Avenu, Singapore 639798.

[17] Danggiang Yang.”Automatic Thesaurus Construction.”School of Informatics and Engineering ,Finland University of south Australia, South Australia.

[18] Bordag,Stefan.”Sentence co-occurrences as small-word-graphs: A solution to automatic lexical disambiguation,” In proceedings of CICling-03, LNCS 2588,2003, Pages 329-333.

[19] Schiffman,H.F.(1999). Content in Language.

[20]Luhn, H.P.(1957). “ A Statistical Approach to Mechanized Encoding and Searching of Library Information.” IBM Journal of Research and Development, Vol. 1, No. 4. October, pp. 309-315.

[21] El-Hamdouchi, A., and P.Willette.(1989).”Comparison of hierarchical agglomerative clustering algorithm for document retrieval,” The Computer Journal.

[22] Grefenstette, Gregory.(1994).Explorations in Automatic Thesaurus Discovery. Boston:Kluwer Academic Press.

[23] Witten, I. H., Frak,E 1999. Data Mining: Practical Machine Learning Tools and Technique, San Francisco, CA.

[24] Chen, H.,T.D. Ng,et al.(1997).” A concept Space Approach to addressing the Vocabulary problem in Scientific Information Retrieval: An Experiment on the worm Community System.” Journal of the American Society for Information Science 48(1): 17-31.

[25] Crouch, C. and Yang, B.(1992). “Experiments in automatic statistical thesaurus construction,” Proceeding of the Fifteen International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark.

[26] Grefenstette G.(1992). “Use of Syntactic Context to Produce Terms Association Lists for Text Retrieval.” Proceeding of the fifteenth International ACM SIGIR Conference On Research and Development in Information Retrieval, Copenhagen Denmark.

[27] Shutze, H. and Pederson, O. (1994). “Co-occurrence Based Thesaurus and Two Application to Information Retrieval.”Proceding RIAO 94 , Conference on Intelligent Image and Text Handling.

[28] Schutze, Hinrich.(1998).”Automatic word sense discrimination.”Computational Linguistics, 24:97-124.

[29] Dagobert Soerget. “Theoretical problems of Thesaurus Building with Particular Reference to Concept Formation.” Collage of Library and Information Science Services, University of Maryland.

[30] John A. Goldsmith, Derrick Higgins, and Svetlana Soglasnova. “Automatic Language-Specific stemming In Information Retrieval.” Department of Linguistics, University of Chicago, 1010 E.st. Chicago IL 60637 USA.

[31] Shutze, H. and Pederson, O. (1994). “Co-occurrence Based Thesaurus and Two Application to Information Retrieval.”Proceding RIAO 94 , Conference on Intelligent Image and Text Handling.

[32] Matsumura, Naohiro, Yukio Ohsawa, and Mitsuru Ishizuka.(2003).”Pai: automatic indexing for extracting asserted keywords from document.” New Generation Computing, 21(1):37-47. Feburary.

[33] Crouch, C. and Yang, B.(1992). “Experiments in automatic statistical thesaurus construction,” Proceeding of the Fifteen International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark.

[34] Church, Kenneth Ward, William A. Gale, Patrick Hanks, and Donald Hindle.(1989).”Parsing, word association and typical predicate-argument relations.” In national workshop on Parsing Technologies. MU.

[35] Firth, J.R.(1957). “A synopsis of linguistic theory 1930-1955,” Oxford: Philological Society. Repeated in f.r. Plamer(ed), selected papers of j.r. Firth 1952-1959, London: Long man, 1968 edition.

[36] Smadja, Frank. (1993).”Retrieving Collection from text: Xtract.”Computational Linguistics, 19(1):43-177.

[37] Harris, Zeeling S.(1968). Mathematical structure of language. Wiley New York.

[38] Church, Kenneth Ward, William A. Gale, Patrick Hanks, and Donald Hindle.(1989).”Parsing, word association and typical predicate-argument relations.” In national workshop on Parsing Technologies. MU.

[39] Agirre,Eneko and German Rigau.”A proposal for word sense disambiguation using conceptual distance.”In Tzigov Mark, proceeding of the first International Conference on Recent advances in NLP, 1995,Bulgaria.

[40] Pantel, P. and Dekang Lin 200. Word-for-word glossing with contextuall similar words. In Proc. Of the Ist Annual Meeting of the North American Chapter of Association for Computational Linguistics(NAAC200), page. 78-85, seattle, USA.

[41] Schutze, Hinrich.(1998).”Automatic word sense discrimination.”Computational Linguistics, 24:97-124.

[42] Salton, Gerard, Amit Singhal, Mandar Mitra,and Chris Buckley.(1997).” Automatic text structuring and summarization.” Information Proceedings and Management , 33(2): 193-207.

[43] Witschel, Hans Friedrich.(2004).” Terminilogy_Extraction: Moglichkeitender Kombination Statistischer und Musterbasierter verfahren.” In content and communication: terminology, Language resource and semantic interoperability. Ergon Verlag, Wurzburg.

[44] Holtsberg, A. and C. Willners. (2001).” Statistics of Sentential co-occurrence.” In working Paper 48, page 135-148.

[45] Stefen Evert.(2005).” The Statistics of word Co occurrences Word pairs and collections.” Institut fur maschinelle Sprachverarbeitung, Universitat Stuugart.

[46] Turney, Peter D.(2001).” Mining the for synonyms: Pmi-ir versus lsa on toefle.” In proceedings of ECML, page 491-502.

[47] Gracinda Carvalho, David Martins de Matos, Victor Rocio.(2007).” Document Retrieval For Question Answering: A qualitative Evaluation of Text Processing.” University of Aberta, Lisbon Portugal.

[48] Philipp Sorg.(2011). “Exploitation of Social Semantics for Multilingual Information Retrieval.”

[49] Christopher D. Manning, Prabhakaar Raghavan, and Hinrich Scutze.(2008) Introduction to Information Retrieval. Cambridge University Press.

[50] Grefenstette G.(1992). “Use of Syntactic Context to Produce Terms Association Lists for Text Retrieval.” Proceeding of the fifteenth International ACM SIGIR Conference On Research and Development in Information Retrieval, Copenhagen Denmark.

[51] United bible societies.(1991).” World translation progress reports and supplements.” London: United bible societies.

[52] Sim, R.J.(1994).” Wolaytta. In: R.E.” Asher(ed.in chief) The encyclopedia of Language and Linguistics, vol 9,4988-4989. Oxford, New York, Seoul, Tokeyo: Pergamon press.

[53] Adams, B.A.”A Taggmatic Analysis of the Wolaytta Language.” Ph.D thesis, University of London(Unpublished),1983.

[54] Motomiche WAKASA.(2008).” A Descriptive Study of The Modern Wolaytta Language.” , Doctoral Dissertation, The University of Tokyo.

[55] Lamberti, Marcollo, and Sottile,Roberto.(1997).”The Wolaytta Language.”Koln:Rudiger Koppe Verlag.

[56] Adams,B.A.(1990). “Name Nouns in Wolaytta .” University of London.

[57]Andargachew Mekonen.(2009). “AUTOMATIC THESAURUS CONSTRUCTION FOR AMHARIC TEXT RETRIEVAL.” Masters Thesis, Addis Ababa University, Ethiopia.

[58] Hogos Hiete.(2011). “AUTOMATIC THESAURUS CONSTRUCTION FOR TIGRIGNA TEXT RETRIEVAL.” Masters Thesis, Addis Ababa University, Ethiopa.

[59] Yuen-Hsien Tseng(2002).” Automatic thesaurus generation for Chinese documents.”Journal of the American Society for Information Science and Technology,Vol. 53,Issue 13.

[60] Chaiwat K.,Nattakan P., Asanee K.”Automatic thesaurus Extraction for Thai Text Retrieval Enhancement.” Natural language processing and Intelligent Information System Technology Laboratory, Department of Computer Enngineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand 10900.

[61] Hirouki Kaji, Yasutsugu Morimoto, Toshiko Aizono, and Nori.”Corpus-dependent Association Thesaurus for Information Retrieval.”Central research laboratory, Hitachi Ltd., Software Division, Tokyo 185-8601, Japan Kanagwa 2044-0801, Japan.

[62] Van Rijsbergen, C.J. (1979).” Information Retrieval.” London. Butter worths.

[63] Kauleh.k,.(2011,May.).” Information Retrieval: Tokenization.” Ioewyi.com.

[64] Baroni, marco and Sabrina Bisi.(2004).“Using co-occurrence statistics and the web to discover synonyms in technical language.”In Procedings of the ELRA 04.

[65] A.Henery. “Information extraction for automatic construction of thesauri and semantic networks from text corpora.” center for natural language processing.

[66] Rapp, Reinhard(2002).”The computation of word associations.” In Proceeding of COLING-02, Taipie, Taiwan.

[67]Jing Jiang, Chang Xiang Zhali.” An empirical study of tokenization for Biomedical Information Retrieval.”Department of Computer Science, University of Illinois at Urbana-Chapaign, Urbana ILC 180.

[68] Frakes, William B. and Baeza-Yates, Ricardo.(1992). Information Retrieval: Data Structure and Algorithms. Engle ward Cliffs, NJ: Prentice Hall, INC.

[69] O’Nell, CandPaice,C.D.’ “What is Stemming?” Avialable at www.comp.lence.ac.uk/computing/research/general/index.htm

Appendixes

Appendix 1 : Notation of Wolaytta with the Alphabet in WPXW [54].

Letters	**Sound value**	**Latter**	**Sound value**
a	/a/	r	/r/
b	/b/	s	/s/
c	/C/	t	/t/
d	/d/	u	/u/
e	/e/	v	[v]
f	[f]	w	/w/
g	/g/	x	/T/
h	/h/	y	/y/ [j]
i	/i/	z	/z/
j	/j/	ch	/c/
k	/k/	dh	/D/
l	/l/	nh	
m	/m/	ny	(palatalized n)

n	/n/	ph	/P/
o	/o/	sh	/sh/
p	/p/	ts	(glottalized s)
q	/K/	zh	/zh/
		7	/7/

Appendix 2: Suffix list

garsappe	bookkona	garsani
dhdhees	iisoona	dosona
doogaa	osoona	isoona
tuuppe	tettaa	tettai
niyaa	aanca	garsa
aanca	shine	diyee
tussi	tuyyo	tetta
tuura	nnaa	mati
gato	aasu	iisi
eese	toow	tura
kkoo	gati	tuni
ttaa	etee	uwaa
uwan	iddi	shin
assi	agaa	iyaa
idoo	iyoo	rkaa
sona	ggaa	iss
saa	too	idi
sha	ena	ppe
adi	nca	sha
iya	uwa	iyu
tti	tta	ibo
ppe	ssa	nna
nta	yoo	kko
ssi	uwa	oow
boo	okk	ada

eda	ida	ibe
okk	iss	kka
ana	nne	shi
iba	idd	yee
yyo	wee	ddi
daa	dan	ggi
gga	gaa	gee
kke	ani	asu
ara	mee	oos
baa	tta	tti
taa	ura	ido
ka	an	ua
os	ss	di
ne	ge	aa
ra	ga	sa
on	te	go
yu	tu	ni
ta	ti	su
na	is	ok
ow	ey	ay
yo	es	da
as	do	se
so	I	e
u	w	o
y	n	

Appendix 3 : Stop word list

a	aadhdhida	aani
aaro	aattidi	ai
ainne	akko	asai
attin	atto	ayyo
ba	baawa	baggaara

bainna	bana	bantta
be7idi	beeta	bessees
biidi	biis	bolli
danddayenna	daro	de7ees
de7iya	de7iyo	dumma
ekkidi	erais	erenna
erissiyo	eroos	eta
etaara	etappe	etassi
etau	eti	ga
gakkanaassi	gidana	giddon
giddooninne	gidenna	gidikko
gidiyo	gidiyoogaa	gido
giidi	giikko	gishshau
gita	giyo	giyoogaa
giyoogee	guutta	guyye
guyyiyan	ha	hagaa
hagaadan	hagan	hagee
hananabaa	hara	harai
he	heezzu	hegaadan
hegaappe	hegan	hegee
hosppun	I	immana
immiis	intte	inttekko
inttena	inttenaara	inttessi
intteyyo	issi	issippe
issuwaa	kase	keehi
kiyidi	koiro	laappun
merinaa	mulekka	naa77anttuwaa
naa77u	ne	neeni
neeyyo	nena	nu
nuna	nuuni	nuuyyo
oodais	ooninne	oottanau

oottidi	oottido	sinttan
siyidita	taani	taayyo
tammanne	tana	tau
ubba	ubbaa	ubbaappe
ubbabaa	ubbai	ubban
urai	wodiyaa	xoqqu
ya	yaagees	yaagidi
yaagidosona	yaagiis	yaagiis
yaagin	yaana	yaatin
aadhd	aadhdh	aadhdhi
aaho	aappu	aara
aatt	aay	aayye7
abaa	aayyi	aayyi
abbaa	abb	aggi
abrahaam	adu	ahai
agga	agge	addaa
ahaa	agg	aiba
aib	aibi	aifana
ainne	aissi	ajjuu
ajjuut	ak	akeek
aikko	aille	aill
akeek	alaafetett	alaaf
amar	amaridaag	ane
anjjettid	anjjettidaag	anjj
anjjo	anxx	anxxooki
aphil	appe	ar
araat	arshsh	ay
ashsh	ashshana	ashshi
att	au	aude
aug	auppe	ayy
ayyaan	a	ayyaa

ayyaana	azaz	azazett
azazi	azazidoog	b
baana	baa	baggaar
bagga	bagg	bai
baizz	bainnaag	bala
bal	banttar	bantt
bantta	baqat	barbb
bar	barnnaab	barnnaabaa
bashsh	bau	ba
be7	be7ana	be7ekk
beett	be7i	bessana
bessi	bi7eel	bir
biryyana	biryy	birshsh
bolla	bi	bochch
boll	bonchchett	bonchch
bonchchu	boopp	boo
caaqq	caaq	daafur
daafuranchch	dar	daro
de7	de7a	de7ana
de7ikk	de7i	de7iyaab
de7iyoog	de7o	de7uwa
demmab	demmana	demm
dendd	denddana	ekka
dent	denddid	denddi
dicc	diggana	digg
do7a	doom	doorett
door	efaana	dosi
dooy	dooyetti	doo
du	dumm	efi
eel	eelaas	eeno
eeq	eeqau	eessa

eessaa	eessai	eexx
eeyya	eta	efis
ekk	ekkana	gaana
eke	ekki	ekkid
elle	elssaabeex	eqqa
eqq	eqqana	eqqi
erissana	eqqid	eqqidaag
erai	erana	gai
eranchch	eratett	erekk
erekket	eri	gakka
erib	erii	ero
ess	gaittana	gakk
gakkana	gakkanaashi	gakkanaa
gakki	galaat	gal
galatana	galatett	galat
galiil	garamett	garam
garamissi	garcc	gar
geed	geell	geem
geetett	gel	gelana
gelissana	gelib	geli
gide	gidekk	gibxx
gib	gid	gide
gidennee	gididoog	gidi
gidikk	gidiyoog	gidokk
gidokkonaa	gidopp	giidoog
giid	gig	giigissana
giriik	giigi	giyoog
goinnana	gudu	gufa
gujj	gussi	guu
haa	haah	haakk
haar	haarana	haasa

haasayid	haasayana	haasayi
haay	hanana	haddir
hanaa	hageeh	hega
hanid	hanqqett	hefi
hann	hanqq	haruur
hassa	he	heezzantt
heezz	hefintt	higg
higgi	higgiya	hindd
hemett	hinkk	hirggopp
ho77	huuphe	ichchash
ikka	ikko	imma
immana	immi	immo
intteb	intte	ippe
issi	isxxifaa	issipp
issuwa	issu	ixxi
iyyo	kaall	kaalli
kaalliyaag	kaalliyaag	kare
kaset	kessana	kehatett
kiitt	kiitett	kiitanchch
kiittid	kiittidoog	kiitti
kiristt	kiristt	kiyidet
koshshi	koshsh	koir
koya	koyib	koyeet
koyidoog	koyi	koyiyoob
koyiyoog	kumi	kundd
lanqqi	lo77	lo77ob
maaddana	maana	maata
maati	maayana	magddal
mairaam	malaat	matatt
minttett	mokkana	moogett
na7a	naa77anttu	naa77a

naagana	naagiyoog	naagiyaag
nabbabibeekket	ne	nu
oda	odidoog	odett
odana	oichchana	oiqqi
oichchi	oiddu	oiqqana
oona	oiqqi	oiqqoppe
oitama	oitta	ooni
ooratta	oosetta	ooso
oosoi	oossi	ootta
oosuwa	ootte	oottana
ootti	oottiyaaga	oottiyooga
ootto	paaccana	ooyyoo
oottoppe	pattana	paxxa
polana	polettana	qanxxana
qaxxar	qifirinaahoom	qoncci
qophphir	qoppa	qoppeet
qoppana	saati	shaakkana
sarotetta	shammana	shiishshana
siiqana	simmana	sinttana
simmi	siyana	sohuwaara
tamaarissana	tamaarissiyaaga	tamaarissi
tama	tookkana	tumatetta
toomaa	ubbaa	ubbaappe
ubbaba	ubba	ubbai
utta	usuppuna	ubba
ubbau	uddufuna	uyana
waassa	wodi	wolqqaama
woggaa	woi	woiga
worana	worddanchch	wottana
wotti	wurssetta	wozana
wozannaam	xaafana	xaissana

xoossa

Appendix 5: Test dataset (avialable at http://www.gospelgo.com)

Hegaa gishshau, nenoo haratu bolli pirddiyaagoo, neeyyo gaasoyiyoobi baawa; aissi giikko, neeni haratu bolli pirddaidda, he neeni eta bolli pirddiyoogaa malaa oottikko, ne bolli pirddaasa.Nuuni hagaa mala oosuwaa oottiyaageetu bolli Xoossai xillo pirddaa pirddiyoogaa eroos.Nenoo, hagaa mala oosuwaa oottiyaageetu bolli pirddiyaagoo, qassi he neeni pirddiyoogaa malaa oottiyaagoo, neeni Xoossaa pirddaappe kessa ekkiyaabaa milatii? Woikko Xoossaa kehatettaa, genccaa, danddayaanne duretettaa karaishsha? Xoossai nena nagaraappe zaaranau koyiyo gishshau, i keha gidiyoogaa neeni eraasa. SHin intteyyo siyenna wozanainne nagaraappe simmenna wozanai de7ees; hegaa gishshau, Xoossaa hanqqoinne xillo pirddai qoncciyo gallassi, neeni ne bolli hanqquwaa kaseegaappe darissaasa. Xoossai asa ubbau a oosuwaa malaa immana. Issi issi asati lo77obaa danddayan oottoosona; qassi galataa, bonchchuwaanne xayenna de7uwaa koyoosona; etau Xoossai merinaa de7uwaa immana. SHin harati iitabaa oottanau yaaretoosonanne tumatettaa ixxoosona; Xoossai eta bolli bashshaanne hanqquwaa yeddana.Iitabaa oottiya asai ubbai tuggatananne waayettana; hegee kasetidi, Aihuda asaa gakkana; kaallidi qassi Aihuda gidenna asaa gakkana. SHin Xoossai lo77o ooso oottiya ubbatuyyo galataa, bonchchuwaanne sarotettaa immana. Kasetidi Aihudatuyyo immana; kaallidi qassi Aihuda gidennaageetuyyo immana. Aissi giikko, Xoossaa aifiyan asai ubbai issi mala. Muuse higgee bainnan nagara oottida ubbai higgee bainnan xayana. Muuse higgee de7ishin, nagara oottida ubbai higgiyan pirddettana. Aissi giikko, Xoossaa sinttan higgiyaa naagiyaageeti xilloosonappe attin, higgiyaa siyiyaageeti xillokkona. Aihuda gidenna asau higgee baawa; shin banttana zoriya wozanan higgee azaziyoogaa oottoosona; etau higgee xayikkokka, eti bantta huupheyyo higge. Higgee azaziyoogee eta wozanaa giddon xaafettidoogaa eta hanotettai bessees; issi issitoo eta zoriya wozanai etau markkattiyo gishshaunne qassi issi issitoo eta qofai eta mootiyo gishshau, woikko qassi etaara gaaddetiyo gishshau, hegee tuma gidiyoogaa erissees.Taani tamaarissido wonggeliyaadan, asaa wozanan geemmida qofaa Xoossai Yesuus Kiristtoosa baggaara hagaadan pirddana. SHin neeni nena Aihuda gaasa; higgiyan zemppada Xoossan ceeqettaasa. Neeni oottana mala, Xoossai koyiyoobai aibakko eraasa; qassi tuma gidiyaabaa dooranau higgiyaappe tamaaradasa. Neeni qooqeta kaalettiyaagaa, qassi xuman de7iyaageetuyyookka poo7issiyaagaa, eeyyaa erissiyaagaanne yelagaa tamaarissiyaagaa

gidiyoogaa eraas; higgiyaappe neeni kumetta eratettaanne kumetta tumatettaa demmiyoogaa eraasa. Yaatin, neeni harata tamaarissiyaagee nena ne huuphen aissi tamaarissikkii? Neeni, “Wuuqqoppa” yaagada qaalaa yootaasa. SHin neeni ne huuphen wuuqqaiyye? Neeni, “SHaaramuxoppa” yaagaasa. SHin neeni ne huuphen shaaramuxai? Neeni eeqaa shenetaasa. SHin eeqai de7iyo keettaa bonqqai? Xoossaa higgee neeyyo de7iyo gishshau, neeni ceeqettaasa; shin he higgiyaa menttada Xoossaa yeellayai? Aissi giikko, Xoossaa maxaafai, “Aihuda asatoo, intte gaasuwan Xoossaa sunttai Aihuda gidenna asatu giddon cayettees” yaagees. Neeni higgiyau azazettikko, ne qaxxarai go77aara de7ees; shin neeni higgiyaa menttikko, neeni qaxxarettidoogee qaxxarettibeennabaa mala gidana. Aihuda gidenna asi qaxxarettennan higgee azaziyoogaa naagikko, Xoossai a qaxxarettidabaadan qoodennee? Intte, Aihuda asati, intteyyo xaafettida higgee de7ikkokka qassi intte qaxxarettikkokka, higgiyaa intte menttiyo gishshau, Aihuda gidenna asai bollaa qaxxarettana xayikkokka, higgiyau azazettidi, intte bolli pirddana. Aissi giikko, Aihuda asa milatiya ubbai tumu aihuda asa gidenna; qassi asi ba bollaa qaxxarettido gishshau, tumu Aihuda asa gidenna. SHin tumu Aihuda asa giyo urai garssa baggaara Aihuda asa gididaagaa; qaxxarettiyoogee Xoossaa Ayyaanan oosettiya wozanaa qaxxarappe attin, xaafettida higgiyaana gidenna; he urai ba galataa Xoossaappe ekkiyoogaappe attin, asappe ekkenna.

Yaatin, Aihuda asati aibin Aihuda gidenna asaappe aadhdhiyoonaa? Woi qaxxarettiyoogan ai go77i de7ii? Hagee Aihudata ubbaban keehi maaddees; ubbabaappe kase Xoossai ba kiitaa hadaraa Aihudatuyyo immiis. Yaatin etappe issooti issooti ammanettennabaa gidikko, waanii? Hegaa gishshau, Xoossai ammanettenna yaagiyoogee? Mulekka hegaa gidenna; asai ubbai worddanchcha gidikkokka, Xoossai giyoogee tuma gidanau bessees. Xoossaa maxaafai, “Neeni haasayiyo wode, neeni xillo gidiyoogaa erissanau koshshees; neeni mootettiyo wode, shatimmanau bessees” yaagees. SHin nu mooro oosoi Xoossaa xillo oosuwaa erissiyaakko, nuuni woiganee? Nu bolli hanqquwaa ehiya Xoossai mooranchchee? Taani asi giyoogaadan gais. Hegee mulekka hanoppo. Aissi giikko, hegaadan gidikko, Xoossai ha sa7aa bolli waati pirddanee? SHin ta wordduwan Xoossaa tumatettai a bonchchuwau doorikko, yaatin, hanno gakkanaassi aissi ta bolli nagaranchchadan pirddii? Yaatin nuuni, “Lo77oi yaanaadan, ane iitabaa oottoos” aissi gookkonii? Issi issi asai hagaa tana mootiiddi cayidosona; eti Xoossaappe bessiya pirddaa ekkana. Hegaa gishshau, nuuni woiganee? Nuuni Aihudati, Aihuda gidenna asaappe gitatiyoonii? Mulekka gidenna; aissi giikko, Aihudatikka Aihuda gidenna asatikka ubbai issippe nagaraappe garssan de7iyoogaa taani kase intteyyo yootaas. Xoossaa maxaafai hagaadan

yaagees; “Xillo asi issoinne baawa. Akeekiya asikka baawa; woikko Xoossaa koyiyaabikka baawa. Asai ubbai Xoossaappe wora simmiis; ubbaikka muleera go77ennaagaa gidiis. Lo77iyaabaa oottiya ooninne baawa; harai atto issoinne baawa. Eta doonai dooyettida duufo mala. Bantta inxxarssan worddotidosona. Eta doonaappe asa woriya shooshshaa yiishshaa mala haasayai kiyees. Eta doonan hanttaara qanggettai kumiis. Eti suutta gussanau daro eesotoosona. Bashshainne qohoi eti biidosan awaaninne xayenna. Eti sarotettaa ogiyaakka erokkona.Qassi Xoossaassi yayyiyoobi eta qofan baawa” yaagees. Asa ubbaa haasayaa diggana malanne kumetta sa7aa Xoossaa pirddaappe garssau aattana mala, higgiyaa giddon odettida ubbabai higgiyaappe garssan de7iya asau haasayiyoogaa nuuni eroos. Xoossaa sinttan ooninne higgee azaziyoobaa oottiyoogan xillenna; higgee issi asi nagara oottidoogaa eranaadan oottees. SHin Xoossai asa xillissiyo ogee higgee bainnan ha77i qoncciis; higgeenne hananabaa yootiyaageeti he xillotettaabaa markkattidosona. Xoossai asaa Yesuus Kiristtoosa ammaniyoogan xillissees; Xoossai hegaa a ammaniya ubbau oottees; ai shaahotettinne baawa. Aissi giikko, asai ubbai nagaraa oottiis; ashshiya Xoossaappe haakkidosona. SHin eti Xoossai immiyo a aaro kehatettan, banttana woziya Kiristtoos Yesuusa baggaara coo xillidosona.Yesuusa suuttan asai a ammanidi, ba nagaraa maarotettaa demmanau, Xoossai a immiis; Xoossai hegaa ba xillotettaa bessanau oottiis; i beni wodiyan asaa nagaraa danddayidi, karidi aggiigiis; shin ha wodiyan ba xillotettaa bessanau asaa nagaraa aggenna; Xoossai hegaa oottiyoogan i ba huuphen xillo gidiyaagaanne Yesuusa ammaniya ubbaa xillissiyoogaa bessees. Yaatin, nuuni aibau ceeqettanee? Aibikka baawa. Ai gaasotaanee? Nuuni higgiyau azazettiyoogaanee? Akkai; ammanuwaa higgiyaanappe attin. Aissi giikko, asi ammanuwan xillanaappe attin, higgee azaziyoogaa oottiyoogaana gidenna yaagidi qoppoos.Woikko Xoossai Aihudatu xalaalau Xoossee? Aihuda gidenna asau Xoossaa gidennee? Tuma, Aihuda gidennaageetuyyookka i qassi Xoossaa.Issi Xoossaa xalaalai de7ees; Aihudata eta ammanuwan i xillissana; qassi Aihuda gidennaageetakka eta ammanuwan xillissana.Yaatin, nuuni higgiyaa ammanuwan gaxi sugiyoo? Gidenna; higgiyaa minttoosippe attin.

Yaatin, nuuni nuuyyo ashuwan zare gidiya Abrahaami ai demmiis gaanee? Aissi giikko, Abrahaami oosuwan xillidabaa gidiyaakko, ayyo ceeqettiyoobi de7ees; shin Xoossaa sinttan ceeqettana danddayenna. Aissi giikko, Xoossaa maxaafai woigii? “Abrahaami Xoossaa ammaniis; Xoossai hegaa ayyo xillotettan qoodiis” yaagees.Oottiya urai ekkiyo miishshaa bau bessin ekkeesippe attin, au imodan qoodettenna. SHin ooso oottennan, nagaranchchaa xillissiya

Xoossaa ammaniya uraa, Xoossai a ammanuwan a xillissees. Hegaadankka, ai oosoinne bainnan, Xoossai xillodan qoodido asa ufaissiyaabaa Daawiti haasayido wode giidoogee hagaa. “Xoossai eta mooruwaa atto giidoogeetinne eta nagaraa kammidoogeeti anjjettidaageeta. Godai a nagaraa qoodenna asi anjjettidaagaa” yaagees.Yaatin, ha Daawiti haasayido anjjoi qaxxarettida asaa xalaalaasseeyye? Woikko qassi qaxxarettibeenna asatussaaree? Aissi giikko nuuni, “Xoossaa maxaafai, 'Abrahaami Xoossaa ammaniis; Xoossai hegaa ayyo xillotettan qoodiis' yaagees” goos. Yaatin, hagee aude hanidee? Abrahaami qaxxarettidaashiineeyye? Woi qaxxarettennan de7ishiinee? Qaxxarettennan de7ishiinappe attin, qaxxarettiina gidenna. Guyyeppe i qaxxarettiis; Xoossai a ammanuwaa gishshau, i qaxxarettanaappe kase a xillo oottidi qoodidoogaa erissanau, a qaxxarai malaata gidiis. Hegeekka qassi qaxxarettennan Xoossaa ammaniya ubbata Xoossai xillo oottidi qoodanaadan, etau Abrahaami aawaa gidanaassa. I qaxxarettidaageetuyyookka aawaa; gidikkokka, eti qaxxarettido gishshassa gidenna; shin nu aawaa Abrahaami qaxxarettanaappe kase ammaniyo ammanuwaa malaa kaalliyaageetuyyookka aawaa. Abrahaamayyoonne a zaretuyyo Xoossai, “Ha sa7aa taani intteyyo immana” yaagidoogee, Abrahaami Xoossaa ammanidi xillido gishshassappe attin, higgiyau i azazettido gishshassa gidenna. Aissi giikko, higgiyau azazettiyaageeti Xoossai immana giidoogaa laattiyaageeta gidikko, asa ammanoi allaalle kessennabaa gidiis; Xoossai giidoogeekka hada gidiis; higgee Xoossaa boshaa ehees; shin higgee bainnasan higgiyaa menttiyoobi baawa. Hegaa gishshau, higgiyaa naagiyaageetu xalaalau gidennan, qassi Abrahaami ammanidoogaadan, ammaniya Abrahaama zaretu ubbau Xoossai immana giidoogee, woitodan tumattanaagee ammanuwaana. Abrahaami nuuyyo ubbaayyo aawaa. Xoossaa maxaafai, “Taani nena daro kawotettatuyyo aawa oottaas” giyoogaadan, haiqqidaageetuyyo de7uwaa immiya, ba azazuwan bainnabaa de7iyaabaadan oottidi xeesiya, i ammanido Xoossaa sinttan Abrahaami nuuyyo ubbau aawa. Xoossaa maxaafai Abrahaama, “Ne zerettai hagaa keena gidana” yaagidoogaadan, hagaa keena gidana giidi, ufaissan naaganau danddayenna wode, Abrahaami ammanidinne sinttappe demmana giidi, ufaissan naagidi, daro zaretuyyo aawa gidiis.He wode, Abrahaami yelettoosappe xeetu laitta kumana haniyo cima; shin ba daafurida asatettaanne Saara na7a yelanau danddayennaagaa qoppido wode, i ba ammanuwan daafuribeenna. Xoossaa i bonchchiiddi, au immana giida Xoossai qassi polanau danddayiyoogaa loittidi erees; eridi ba ammanuwan minnidoogaappe attin, ammanennan ixxidi, Xoossai immana giidoogaa siribeenna. Hegaa gishshau, Xoossai qassi a ammanuwaa xillotettan qoodiis.“Ayyo xillotettan qoodiis” giya qaalai a xalaalau

xaafettibeenna.He qaalai nu Godaa Yesuusa haiquwaappe denttida Xoossaa ammaniya nuuyyookka xillotettan qoodettana mala xaafettiis.Xoossai a nu nagaraa gishshau, haiquwau aattidi immin, i nuna xillissanau haiquwaappe denddiis.

Hegaa gishshau, nuuni ammanuwan xillidoogan, nu Godaa Yesuus Kiristtoosa baggaara nuuyyo Xoossaara sarotettai de7ees;qassi i nuna ha77i nuuni de7iyo Xoossaa aaro kehatettaa giddo ammanuwan ehiis; yaatin, nuuni Xoossaa bonchchuwaa shaakkanau nuuyyo de7iya ufaissan ceeqettoos.Hegaa xalaala gidenna; nuuni nu waayiyankka ufaittoos; aissi giikko, waayee danddayaa ehiyoogaa, qassi danddayai paaciyaa, paacee qassi demmanau ufaissan naagi uttidoogaa ehiyoogaa nuuni eroos.Xoossai nuuyyo immido Geeshsha Ayyaanaa baggaara ba siiquwaa nu wozanan gussido gishshau, he nuuni demmanau ufaissan naagi uttidoogee nuna yeellayenna. Aissi giikko, nuuni biron wolqqi bainnaageeta gididi de7ishin, Xoossai doorido wodee gakkin, Kiristtoosi nagaranchchatu gishshau haiqqiis.Xillo uraa gishshau, issi urai haiqqanaagee daro metobaa; shin keha asa gishshau haiqqanau, “Eeno” giya uri ooni erii beettennan aggenna.SHin nuuni biron nagaranchcha gididi de7ishin, Kiristtoosi nu gishshau haiqqiis; hegee Xoossai nuna ai keenaa siiqiyaakkonne a siiquwaa erissees.Simmi nuuni ha77i a suuttan xillidabaa gidikko, a baggaara Xoossaa hanqquwaappe attana.Aissi giikko, nuuni Xoossaara kase morkke; shin a Na7aa haiquwan aara sigettida. Hegaappe aattidi, ha77i nuuni Xoossaara sigettiichchidoogaappe guyyiyan, i paxa gidiyo gishshau nuna ashshana. SHin hegaa xalaala gidenna; ha77i nuna Xoossaara sigettida nu Godaa Yesuus Kiristtoosa baggaara nuna Xoossaabai ufaissees. Hegaa gishshau, nagarai ha sa7aa issi asa baggaara geliis; he nagarai banaara haiquwaa ehiis; hegaa gaasuwan, asai ubbai nagara oottido gishshau, haiqoi asa zare ubbaa gakkiis.Higgee imettanaappe kase, ha sa7an nagarai de7ees; shin higgee bainna wode nagari qoodettenna. SHin Addaame wodiyaappe, Muuse wodiyaa gakkanaassi, haiqoi asa ubbaa haariis; harai atto Addaamee Xoossaa higgiyau azazettennan ixxidoogaadan, nagara oottibeennaageetakka haiqoi haariis. Aissi giikko, Addaamee yaanaagau leemiso.SHin Xoossaa imoi Addaame nagaraa keena gidenna gishshau, eti issi mala gidokkona; issi bitaniyaa nagaraa gishshau, daro asai haiqqidoogee tuma; shin Xoossaa aaro kehatettainne a imoi issi asa aaro kehatettaa baggaara, Yesuus Kiristtoosa aaro kehatettaa baggaara daro asau dariis.Xoossaa imoi nagaraa oottida bitaniyaa baggaara yiidaagaa mala gidenna; aissi giikko, Addaamee issi nagaraa oottidoogaappe guyyiyan, Xoossai a mooranchchadan pirddiis; shin asai daro nagaraa oottidoogaappe guyyiyan Xoossai, “Intte mooribeekketa” giidi, ba aaro kehatettaa eta bessees.

Aissi giikko, issi bitaniyaa nagaraa gaasuwan, a nagaraa baggaara, haiqoi kawotidoogee tuma; shin issi urai, Yesuus Kiristtoosi oottidobai ai keena daro; Xoossaa daro aaro kehatettaanne xillotettaa imuwaa ekkiya ubbai Kiristtoosa baggaara xoonidi de7ana. Simmi issi urai nagara oottido gaasuwan asa ubbaa bolli mooranchchadan pirddai yiidoogaadan, hegaadan issi urai xillobaa oottido gaasuwan, Xoossai asa ubbaa, "Mooribeekketa" giidi etau de7uwaa immiis; aissi giikko, issi urai azazettennan ixxido gaasuwan, daro asai nagaranchcha gididoogaadan, hegaadankka qassi issoi eeno giidoogan daro asai xillana.Mooro oosuwaa darissanau, higgee yiis; shin nagarai darido wode Xoossaa aaro kehatettai keehi dariis; nagari haiquwaa wolqqan kawotidoogaadan, hegaadankka qassi Xoossaa aaro kehatettai xillotettaa wolqqan kawotidi, nu Godaa Yesuus Kiristtoosa baggaara, nuna merinaa de7uwau kaalettees.

Yaatin, nuuni woiganee? Xoossaa aaro kehatettai daranaadan, nagaran minnidi de7anee? Akkai; nuuni nagara ooso baggaara gidikko haiqqiichchida. Yaatin, nuuni hagaappe sinttau waati a giddon de7anee? Woikko nuuni ubbati Kiristtoos Yesuusaara issippe gidanau, xammaqettido wode, a haiquwankka aara issippe gidanau xammaqettidoogaa erekketii? Hegaa gishshau, Kiristtoosi Aawaa bonchcho wolqqan haiquwaappe denddidoogaadan, hegaadan nuunikka ooratta de7uwaa de7anau, a haiquwaa aara shaakkanau xinqqatiyan aara issippe moogettida. Aissi giikko, nuuni i haiqqidoogaadan haiqqidi, aara issuwaa gidikko, hegaadankka, haiquwaappe i denddidoogaadan nuunikka denddidi, aara issuwaa gidana. Nuuni hagaappe sinttanau nagarau aille gidenna mala, qassi nu nagaranchcha asatettaa wolqqai xayana mala, nu ceega asatettai Kiristtoosaara masqqaliyaa bolli misimaariyan xishetti kaqettidoogaa eroos. Aissi giikko, issi urai haiqqiyo wode nagaraa wolqqai a haarenna; shin nuuni Kiristtoosaara haiqqidabaa gidikko, aara qassi de7anaagaa ammanoos. Aissi giikko, Kiristtoosi haiquwaappe denddidoogaa, qassi naa77anttuwaa i haiqqennaagaanne haiqoi sinttappe a haarennaagaa nuuni eroos.I issitookka haiqqiis; nagarau a bolli wolqqi baawa; shin i ha de7iyo de7uwaa Xoossaayyo de7ees. Hegaadan qassi intte nagara oottennaadan, inttena haiqqidabaadan qoppite; shin Godaa Kiristtoos Yesuusa baggaara Xoossau intte paxa de7iyaabaadan, inttena qoppite.Hegaa gishshau, intte intte asatettaa amuwau azazettana mala, nagarai haiqqiya intte asatettaa haaroppo. Qassi intte asatettaappe issisaanne mooro ooso oottanau nagarau immoppite; shin haiquwaappe denddida asaadan inttena Xoossau immite; qassi intte kumetta asatettaakka xillo ooso oottanau Xoossau immite; aissi giikko, nagarai inttena haaranau bessenna; intte Xoossaa aaro kehatettaappe garssaara de7eetappe attin, higgiyaappe garssan de7ekketa.Yaatin, waananee?

Xoossaa aaro kehatettaappe garssaara de7iyoogaappe attin, higgiyaappe garssaara deenna gishshau, nagara oottanee? Akkai; intte inttena issi urau azazettanau, ailledan oottidi immiyo wode, he inttena azaziyaagau ailleta gidiyoogaa ereeta; woikko intte intteyyo haiquwaa ehiya nagarau haaretteeta; woikko intte inttena xillo oottiya azazuwau haaretteeta. SHin Xoossau galatai gido; aissi giikko, nagarau kase intte aille; shin guyyeppe intte ekkido timirttiyaa giddon de7iya tumatettau intte kumetta wozanaappe azazettideta.Intte nagaraa ailletettaappe kiyidi, xillotettau aille gidideta.Intte mereta daafuranchcha gidiyo gishshau, taani ooyyookka geliya haasayaa haasayais; intte kase intte kumetta asatettaa moorobaa oottanau, tunatettaunne moorobau aille oottidi immideta; ha77i hegaadankka intte kumetta asatettaa geeshshabaa oottanau, xillotettau aille oottidi immite. Intte nagarau ailletidi de7iyo wode, xillotettai intteyyo goda gidenna. Inttena ha77i yeellayiya yohota he wode oottidoogaappe ai go77aa demmidetii? Aissi giikko, he yohotu wurssettai haiqo. SHin nagaraa ailletettaappe intte kiyidi, Xoossau haarettideta; qassi intte oottiyo ubbaban inttena muleera au aattidi immideta. Hegee intteyyo wore; hegaa wurssettai merinaa de7uwaa. Aissi giikko, nagarai qanxxiyo miishshai haiquwaa. SHin Xoossai nu Godaa Kiristtoos Yesuusa baggaara immiyo imoi merinaa de7uwaa.

Ta ishatoo, intte ubbai higgiyaa eriyo gishshau, taani giyoogaa loittidi ereeta; higgee issi asi de7ido wode ubban a haarees Aissi giikko, azina gelida maccaasiyaa i azinai paxa de7ido keenan, higgiyan aara qashetta uttaasu; shin i azinai haiqqikko, aara maacettido higgee o haarenna. Hegaa gishshau, a ba azinai paxa de7ishin, hara attumaagaa gelikko, asai o, “SHaaramuxa” gaana. SHin i azinai haiqqikko, higgee o haarenna; a hara azina gelikkonne o ooninne, “SHaaramuxa” geennaTa ishatoo, hegaadan intte qassi Kiristtoosa asatettaappe issi bagga gidiyo gishshau, higgiyaa qoppiyaabaa gidikko, intte haiqqideta. Qassi nuuni Xoossaa go77iya ooso oottana mala, intte Xoossai haiquwaappe denttido KiristtoosabaaAissi giikko, nuuni nu asatettai koyidoogaadan de7iyo wode, higgee nuuni nagaraa oottanaadan, nuna denttettees; nu asatettaa giddon iita amoi de7iyo gishshau, nuuni haiquwau aife aifidaSHin ha77i nuuni nuna qachchidaagau haiqqido gishshau, Xoossai nuna higgiyaa qashuwaappe birshshiis. Hegaa gishshau, nuuni Geeshsha Ayyaanai giigissido ooratta ogiyan Xoossau azazettanaagaappe attin, xaafettida higgee nuussi giigissido ceega ogiyan azazettokkoYaatin, nuuni woiganee? Higgee ba huuphen nagaree? Mulekka gidenna; shin higgee nagari aibakko tana erissiis. Aissi giikko, taani higgee, “Asabaa amottoppa” geennaakko, asabaa amottiyoogee aibakko erikkeSHin nagari taani

ubbabaa amottana mala, azazuwaa baggaara tanan ogiyaa demmiis; aissi giikko, higgee bainnaakko, nagarau wolqqi baawaKase higgee bainna wode taani paxa; shin azazoi yiido wode nagarai paxin, taani qassi haiqqaas; tana paxa wottanau Xoossai qoppido azazoi, tau haiquwaa ehiis Aissi giikko, nagarai azazuwaa baggaara tana cimmiyo ogiyaa demmidi, he azazuwan tana woriis Hegaa gishshau, higgee geeshsha; azazoikka geeshsha, xillonne lo77o Yaatin, lo77obai tau haiqo ehiidee? Mulekka gidenna; shin tau hegaa nagarai ehiis; asai nagari aibakko shaakkidi erana mala, nagarai lo77oban go7ettidi, tau haiqo ehiis. Qassi azazoi, “Ooninne nagara oottanau bessenna” yaagiyo gishshau, nagarai kaseegaappe aadhdhi iitiis. Nuuni higgee Xoossaabaa gidiyoogaa eroos; shin taani tana nagarau baizzida haiqqiya asa; taani oottiyoobaa akeekikke; aissi giikko, taani ixxiyoobaa oottaisippe attin, taani oottanau dosiyoobaa oottikke. Taani ixxiyoogaa oottiyaabaa gidikko, higgee lo77o gidiyoogaa markkattais Hegaa gidikko, he yohuwaa tanan de7iya nagarai oottiyoogaappe attin, ha77i taani oottikke Tanan, hegeenne ta asatettan lo77obai de7ennaagaa taani erais; aissi giikko, lo77obaa oottiyo amoi tanan de7ikkokka, taani he lo77obaa oottanau danddayikke Taani ixxiyo iitabaa oottaisippe attin, taani dosiyo lo77obaa oottikke SHin taani ixxiyoogaa oottiyaabaa gidikko, hegaa tanan de7iya nagarai oottiyoogaappe attin, ha77i hegaa oottiyaagee tana gidikke Simmi taani lo77obaa oottanau dosishin, iitabaa oottiyoogee tau higge gidiyoogaa demmaas Aissi giikko, taani ta wozana giddon Xoossaa higgiyan ufaittais shin ta bollaa giddon de7iya ta wozanaa qofaa higgiyaara warettiya dumma higgiyaa be7ais. He higgee ta bollaa giddon de7iya nagaraa ootissiya higgiyan tana qashissees Taani aiba ufaissi bainna asee! Tana haiquwau efiya ha asatettaappe ashshanai oonee Nu Godaa Yesuus Kiristtoosa baggaara Xoossau galatai gido. Simmi taani ta wozanaa qofan Xoossaa higgiyau haarettais. SHin ta asatettan tana nagaraa ootissiya higgiyau haarettais Hegaa gishshau, Kiristtoos Yesuusaara de7iyaageetu bolli pirddi ha77i baawa aissi giikko, Kiristtoos Yesuusa baggaara nuuyyo de7uwaa ehiya Geeshsha Ayyaanaa higgee tana nagaraa higgiyaa ailletettaappenne haiquwaa higgiyaa ailletettaappe kessiis Asaa asatettai daafuranchcha gidiyo gishshau, higgee oottanau danddayibeennaagaa Xoossai oottiis; nagaranchchatu asatettaa mala asatettaara de7iya ba Na7aa, asai oottido nagaraayyo yarshsho gidanau Xoossai kiittidi, asaa asatettaa giddon de7iya nagaraa pirddiis Geeshsha Ayyaanai koyiyoogaadan de7iyaageetuppe attin, nu asatettai koyiyoogaadan de7ennaageetu bolli, nu bolli, higgiyaa xillo azazoi polettana

mala, Xoossai hagaa oottiis Bantta asatettai koyiyoogaadan de7iyaageeti bantta asatettaabaa qoppoosona; shin Geeshsha Ayyaanai koyiyoogaadan de7iyaageeti Geeshsha Ayyaanaabaa qoppoosona satettaabaa qoppiyoogee haiquwaa ehees; shin Geeshsha Ayyaanaabaa qoppiyoogee de7uwaanne sarotettaa ehees. Asatettaabaa qoppiyoogee Xoossau ixo; aissi giikko, asatettai Xoossaa higgiyau azazettenna; qassi azazettanaukka au danddayettenna. Bantta asatettau azazettiyaageeti Xoossaa ufaissanau danddayokkona SHin Xoossaa Ayyaanai inttenan de7ikko, Geeshsha Ayyaanai giyoogaadan de7eetappe attin, intte asatettai koyiyoogaadan de7ekketa. SHin ooninne Kiristtoosa Ayyaanai bainnaagaa gidikko, he urai Kiristtoosabaa gidenna. SHin Kiristtoosi inttenan de7ikko, intte asatettai nagaraa gaasuwan haiqqidaagaa; shin intte ayyaanai xillotettaa gaasuwan inttenan paxa de7ees. SHin Yesuusa haiquwaappe denttida Xoossaa Ayyaanai intte giddon de7ikko, haiquwaappe Kiristtoosa denttida Xoossai qassi inttenan de7iya ba Geeshsha Ayyaanan, haiqqiya intte asatettaukka de7uwaa immana Simmi ta ishatoo, nuuni acuwaara de7oos; nu asatettai koyiyoogaadan de7anau nuuyyo bessenna Aissi giikko, intte intte asatettai giyoogaadan de7ikko, intte haiqqana; shin intte intte nagara oosuwaa Geeshsha Ayyaanan worikko, paxa de7ana. Xoossaa Ayyaanai kaalettiyo ubbati Xoossaa naata Aissi giikko, nuuni nu qaalaa xoqqu oottidi, “Abbaa! Ta Aawau” yaagidi xeesiyo, nuna Xoossaa naata kessiya Geeshsha Ayyaanaa Xoossaappe ekkidoogaappe attin, naa77anttuwaa yayyanau aille kessiya ayyaanaa ekkibookko. Xoossaa Ayyaanai ba huuphen nu ayyaanaara gididi, nuuni Xoossaa naata gidiyoogaa markkattees. Nuuni a naata gidikko, i ba asau minjjido anjjuwaa laattana; qassi nuuni Xoossai Kiristtoosau minjjidoogaakka aara laattana; nuuni Kiristtoosa waayiyaa aara shaakkikko, a bonchchuwaakka aara shaakkana. Aissi giikko, nuuni ha77i waayettiyo waayiyaa nuuyyo sinttappe qonccana bonchchuwaara giigissanau, mulekka danddayettenna gaada taani qoppais. Xoossai medhdhido mereta ubbai Xoossaa naati qonccanaagaa daro laamoti uttiis. Aissi giikko, mereta ubbai go77i bainnabaa gidanaadan, Xoossai a bolli pirddiis; sinttappe demmana giidi ufaissan naagi uttidoogan haarissida Xoossai koyido gishshau hanidoogaappe attin, mereta ubbai ba huuphen koyidoogaana gidenna. Mereta ubbai ba huuphen bashshaa ailletettaappe issi gallassi kiyana; qassi he meretati Xoossaa naati ailletettaappe kiyido bonchcho kessaa shaakkana. Aissi giikko, mereta ubbai issippe hanno gakkanaassi, maccaasa leessoi oiqqin ooliyoogaadan, sahuwan ooliyoogaa nuuni eroos. Qassi hegaa xalaala gidenna; nuuni,

Geeshsha Ayyaanaa Xoossaa baira imodan oottidi ekkidaageeti, nu huuphen qassi Xoossai nuna ba naata oottanaadaaninne nu kumetta asatettaa wozanaadan naagiiddi, nu giddon ooloos.Aissi giikko, nuuni Xoossai immana giidoogan attida. SHin nuuni Xoossai immana giidoogaa be7ikko, yaatin, hegee ufaissan naagiyoobaa gidenna; aissi giikko, ba be7iyoobaa demmana giiddi, ufaissan naagiyai oonee? SHin nuuni be7ennaagaa demmana giidi, ufaissan naagikko, hegaa danddayidi naagoos. Hegaadan qassi Geeshsha Ayyaanai nu daafuraa maaddanau yees; aissi giikko, nuuni Xoossaa waati woossanau bessiyaakko erokko; shin Geeshsha Ayyaanai ba huuphen asau haasayanau danddayettenna ooliyan nuuyyo gaannatees. Qassi asa wozanaa qoriya Xoossai, Geeshsha Ayyaanai i koyiyoogaadan, geeshshatuyyo gaannatiyo gishshau, Geeshsha Ayyaanaa qofai aibakko erees. Xoossai bana siiqiyaageetuyyo, ba qoppidoogaa oottanau xeesidoogeetuyyo, yoho ubbaa lo77obau oottiyoogaa nuuni eroos.Xoossai ba Na7ai daro ishanttu giddon baira gidanaadan, i kase dooridoogeeta qassi ba Na7aa milatanaadan, kasetidi shaakki wottiis. Xoossai kasetidi shaakki wottidoogeeta, eta qassi xeesiis; he xeesidoogeeta qassi xillissiis; he xillissidoogeeta qassi bonchchiis. Yaatin, nuuni hagaa ubbaa woiganee? Xoossai nu bagga gidikko, nuna oonee qohanai? I harai atto ba Na7aakka waayiyaappe guyye ashshibeenna; shin a nuuyyo ubbaayyo aattidi immi aggiis. I ba Na7aa immidaagee qassi nuuyyo ubbabaa coo immennee? Xoossai dooridoogeeta oonee mootanai? Xoossai ba huuphen eta xillissiis.Yaatin, eta bolli pirddanau danddayiyai oonee? Oonee giikko, haiquwaappe denddidi, Xoossaappe ushachcha baggaara uttida, Kiristtoos Yesuusa; i qassi nuuyyo gaannatees. Simmi nuna Kiristtoosa siiquwaappe shaakkanai oonee? Waayee woi tuggai woi yedetai woi koshai woi kalloi woi yashshai woi olai nuna shaakkanee?Xoossaa maxaafai, “Nuuni ne gishshau, gallassaa muumiyaa olettida; goora7anau efiyo dorssadan qoodettida” yaagees. SHin nuuni nuna siiqida aani ha yoho ubban poli xoonida. Aissi giikko, haiqo gidin, de7o gidin, kiitanchchata gidin, haariyaageeta gidin, ha77i de7iyaagaa gidin, sinttappe yaanaagaa gidin, wolqqaamata gidin, bolla saluwaa gidin, garssa sa7aa gidin, ai mereta gidinkka, Kiristtoos Yesuusan de7iya Xoossaa siiquwaappe nuna shaakkanau danddayennaagaa taani eraas.Taani tumaa haasayais; taani Kiristtoosabaa; qassi worddotikke. Geeshsha Ayyaanai haariyo tana zoriya wozanai taani worddotennaagaa taayyo markkattees.Taayyo ashuwan zare gidiya ta ishanttuyyo, ta wozanan daro qaretainne aggenna qofai de7ees. Aissi giikko, taani ta

huuphen eta gishshau, Xoossaa qanggettaappe garssan de7arkkinaashsha! Qassi Kiristtoosappekka shaahettarkkinaashsha! Eti Israa7eela asata; Xoossai eta ba naata oottidi, ba bonchchuwaa etaara shaakkiis. Etaara maacettidi, higgiyaa etau immiis. Eti Xoossau bessiya ogiyan goinnoosona; qassi Xoossai immana giidoogaakka ekkidosona. Eti aawatu zare; qassi Kiristtoosikka ashuwan eta zare. Xoossai, ubbaappe bolla gididi haariyaagee, merinau galatetto. Amin77i. SHin taani, "Xoossai immana giidoogee attana" yaagikke. Aissi giikko, Israa7eela asa ubbati Xoossai dooridoogeeta gidokkona. Qassi Abrahaama zare ubbati Xoossaa naata gidokkona. Xoossai Abrahaamayyo, "Ne zaree Yisaaqa baggaara xeesettana" yaagiis; hegaa giyoogee eti Xoossai immana giido ogiyan yelettida naatu zaredan qoodettana; shin asa qofaadan yelettidaageeti Xoossaa naata gidokkona yaagiyoogaa. Aissi giikko, Xoossai immana giidi haasayidoogee hagaa; "Taani laitti hannoode yaana; attuma na7a Saara yelana" yaagiis. Hegaa xalaalakka gidenna; Irbbiqa yelido mentte naatu aawai issuwaa; i nu aawaa Yisaaqa. SHin issi na7aa dooroi Xoossai ba huuphen qoppidoogaa gidanaadan, Xoossai Irbbiqiyyo, "Bairai kaaluwau azazettana" yaagiis; i hegaa eti yelettanaappenne iitabaa woi lo77obaa ainne oottanaappe kase giis. Hegaa gishshau, Xoossaa dooroi a xeesidoogaadaanippe attin, eti oottidoogaana gidenna. Xoossaa maxaafai, "Yaaqooba siiqaas; shin Eesawa ixxaas" yaagees.Yaatin, nuuni woiganee? Xoossai moori pirddii? Mulekka gidenna. Aissi giikko, Muuseyyo i, "Taani maaranau koyiyoogeeta maarana; qassi qarettanau koyiyoogeetuyyookka qarettana" yaagiis. Simmi Xoossai asa maariyoogee i maaranau koyiyo gishshassappe attin, asi koyiyo woi oottiyo gishshassa gidenna. Aissi giikko, Xoossaa maxaafai Gibxxe kawuwaayyo, "Taani ta wolqqaa nenan bessanaunne ta sunttai sa7an ubban odettanau, hegau taani nena kawoyaas" yaagees. Simmi Xoossai maarana koyiyoogaa maarees; qassi wozanaa gorddana koyiyoogaa wozanaa gorddees. Simmi intte tana, "Hegee tuma gidikko, Xoossai asaa mooruwaassi aibissi walassii? Xoossai giidoogaa ixxaas gaanai oonee?" yaaganau danddayeeta. SHin ta laggiyau, Xoossaa bolli zaaranau neeni oonee? Urqqaappe merettida miishshai guyye zaaridi bana medhdhidaagaa, "Tana hagaadan aissi medhdhadii?" giidi oichchanau danddayii? Woikko urqqaa medhdhiyaagee ba medhdhiyo urqqaa ba koyiyoogaadan oottanaunne issi mala urqqaappe issuwaa bonchcho miishsha oottidi, qassi issuwaa tooshe ooso oottiyo miishsha oottidi medhdhanau au maati baawee? Xoossai oottidoogee hagaara issuwaa. I ba hanqquwaa bessanaunne ba wolqqaa erissanau koyiis; i

ba hanqquwan pirddanaageeta, bashshanau giigi uttidoogeeta, daro genccan danddayiis. Qassi i maaridoogeetuyyo, ba bonchchuwaa ekkanau giigissi wottidoogeetuyyo i ai keenaa bonchchettiyaakkonne bessanau koyiis. Nuuni i xeesidoogeeta; shin Aihuda asaa giddo xalaalaappe gidennan, Aihuda gidenna asaa giddoppekka xeesiis. Hoose7a maxaafan i giyoogee hagaa; “Ta asaa gidennaageeta, 'Ta asa' gaada xeesana. Taani siiqabeenna zariyaakka, 'Ta siiqiyoogeetoo' gaada xeesana. 'Intte ta asa gidekketa' giyoogaa eti siyido he sohuwan, eti, 'De7o Xoossaa naata' geetettidi xeesettana” yaagees. Isiyaasi Israa7eela asaabaa yaagidi haasayiis; “Israa7eela asai abbaa matan de7iya shafiyaa keena gidiyaakkokka, etappe amaridaageetu xalaalai attana. Aissi giikko, Godai sa7aa bollan de7iya asa ubbaa bolli eesuwan pirddana” yaagiis. Isiyaasi kasetidi, “Ubbaappe Wolqqaama Godai nuuyyo amarida zareta ashshennaakko, nuuni Sadooma katamaadaaninne Gamoora katamaadan hanidashin” yaagidi haasayiis. Yaatin, nuuni woiganee? Xillotettaa koyibeenna Aihuda gidenna asai ammanuwan xilliis; shin Israa7eela asai banttana xillissiya higgiyaa koyidaageeti he higgiyaa demmibookkona.Aissi demmibookkonaa? Eti oosuwaanappe attin, ammanuwan xillotettaa koyibeenna gishshassa. Eti Xoossaa maxaafai, “Akeekite; taani Xiyoonen asaa xubbiya shuchchaa, eta ooggiya zaallaa wottana. SHin aani ammaniya ooninne yeellatenna” yaagidi haasayiyo xubbiya shuchchan xubettidosona.Ta ishatoo, taani Israa7eela asaayyo ta kumetta wozanaappe amottiyoogeenne Xoossaa woossiyoogee eti attanaassa; aissi giikko, eti Xoossau oottanau keehi koyiyoogaa taani markkattais; shin eti Xoossau oottanau keehi koyiyo koshshai tumu eratettaappe gidenna. Aissi giikko, Xoossai asaa xillissiyo ogiyaa eti erennan, bantta huuphe xillotettaa essanau koyidosona; qassi bantta huuphen Xoossai asaa xillissiyo ogiyau haarettibookkona. Aissi giikko, ammanida ubbai xillanaadan, Kiristtoosi higgiyau wolqqaa xaissiis. Asi higgiyau azazettidi xilliyoobau Muusee, “Higgee azaziyoogaa oottiya ooninne he higgiyan paxa de7ana” yaagidi xaafiis; shin ammanuwan xilliyoobau geetettidaagee hagaa; “Ne wozanan, 'Ooni saluwaa pude baanee?' yaagoppa. Hegee Kiristtoosa duge wottanaassa; woikko,'Duge sa7aa garssi wodhdhanai oonee?' gooppa. Hegee Kiristtoosa haiquwaappe denttanaassa. SHin i giyoogee hagaa; 'Xoossaa qaalai ne mataana; ne doonaaninne ne wozanan de7ees' geetettiis” yaagees. Hegeekka nuuni intteyyo yootiyo ammanuwaa qaalaa. Neeni, “Yesuusi Godaa” gaada ne doonan markkattikko, qassi Xoossai haiquwaappe a denttidoogaa ne wozanan ammanikko, neeni

attana. Aissi giikko, asi ba wozanan ammanidi xillees; ba doonankka markkattidi attees. Xoossaa maxaafai, “Aani ammaniya ooninne yeellatenna” yaagees. Aihuda asaanne Aihuda gidenna asaa giddon dummatetti baawa. Issi Xoossai ubbaa Godaa; i bana xeesiya ubbatuyyo daro keha. Aissi giikko, Xoossaa maxaafai, “Godaa sunttaa xeesiya ooninne attana” yaagees.Yaatin, eti a ammanibeennaageeta gidikko, waatidi a xeesanau danddayiyoonaa? Qassi qaalaa siyibeennaageeti waani ammananau danddayiyoonaa? Qaalaa yootiyaabi xayikko, eti waani siyanau danddayiyoonaa? Qassi qaalaa yootiyaageeti kiitettennan aggikko, kiitai waanidi odettanau danddayii? Aissi giikko, Xoossaa maxaafai, “Ufaissiya mishiraachchuwaa yootiyaageetu tohoti aiba lo77iyoonaa!” yaagees. SHin eti ubbai wonggeliyau azazettibookkona; aissi giikko Isiyaasi, “Ta Godau, nuuni markkattidoogaa ooni ammanidee?” yaagiis.Simmi asi qaalaa siyidi ammanees; qassi issi urai Kiristtoosa qaalaa yootin siyees.SHin eti qaalaa siyibeennaagee tumee gaada taani oichchais. Tuma eti siyidosona; Xoossaa maxaafai, “Eta haasayaa cenggurssai sa7a ubban laalettiis; eta qaalaikka sa7aa gaxaa gakkiis” yaagees. SHin Israa7eela asati eribookkonaayye gaada taani oichchais. Muusee kase, “Taani inttena asa gidenna asatun mishissana; qassi eeyya asatun taani inttena hanqqetissana” yaagiis.Qassi Isiyaasi daro xalidi, “Tana koyibeennaageetuyyo beettaas; tana oichchibeennaageetuyyookka qonccaas” yaagees. SHin Israa7eelatubaa i haasayiiddi, “Gallassaa ubbaa azazettennaageetuyyoonne tabaa ixxiyaageetuyyo ta kushiyaa miccaas” yaagees.

Intte ta eeyyatettaa guuttaa gencciyaakko lo77o. Hayyanintta genccite. Xoossai qanaatiyoogaadan, taanikka intteban qanaatais; aissi giikko, taani inttena issi attumaagau, Kiristtoosau immana gaada, geela7otettaara de7iya geeshsha geela7eedan giigissa wottaas. SHin shooshshai ba geniyaara Hewaano balettidoogaadan, Kiristtoosayyo intte immido intte suure qofai moorettibeennaagee wora biichchanaakkonne gaada hirggais. Aissi giikko, issi urai yiidi, nuuni intteyyo yootido Yesuusappe hara Yesuusa intteyyo yootikko, woi intte ekkido Ayyaanaappenne wonggeliyaappe dumma ayyaanaanne wonggeliyaa intte ekkikko, loittidi gencceeta. Yesuusi kiittido ha waannatuppe taani aibinkka laafiyoobaa taayyo milatenna.Taani haasayiyo haasayan tamaarabeenna asa gidikkokka, eratettan gidikke; shin nuuni intteyyo ha yohuwaa asa ubbaa giddon ubba baggaara qonccissida.Woi taani inttena miishsha oichchennan,

Xoossaa wonggeliyaa mishiraachchuwaa intteyyo coo yootidoogeenne inttena xoqqu oottanau ta huuphiyaa toochchidoogee taayyo nagara gidanddeeshsha!Taani intteyyo oottanau hara woosa keettatun de7iya asaappe miishsha ekka ekkada, he woosa keettata bonqqaas. Qassi taani inttenaara de7ishin, tana metin, intteppe issi uraakka waissabeikke. Aissi giikko, Maqidooniyaappe yiida ishantti tana koshshida ubbabaa taayyo immidosona; taani inttena waissenna mala, kase ta huuphen naagettidoogaadankka sinttaukka naagettana. Kiristtoosa tumatettai tanan de7ishin, ha ta ceeqoi Akaayiyaa biittan awaaninne attenna. Taani hagaa aissi giyaanaa? Inttena siiqenna gishshassee? Taani inttena siiqiyoogaa Xoossai erees. SHin he bantta ceeqettiyooban nuuni oottiyoogaadan oottanau gaasottiyaageetu gaasuwaa digganau, ha77i taani oottiyo oosuwaa gujjada oottana. Aissi giikko, eti Yesuusi kiittibeenna worddanchcha asata; hegeetu mala asati Kiristtoosi kiittidoogeeta milatanau bantta meraa laamerettiya kiitettibeenna worddanchchatanne genanchcha oosanchchata. Hegee garamissiyaabaa gidenna; aissi giikko, Seexaanai ba huuphen poo7o kiitanchcha milatanau ba meraa laammees. Hegaa gishshau, Seexaanaa ashkkarati xillotettaa ashkkarata milatanau bantta meraa laamerettikko, hegee gitabaa gidenna; eti wurssettan bantta oosuwaa malaa ekkana.Taani naa77anttuwaakka gais; tana ooninne eeyya giidi qoppoppo; shin intte tana, i eeyya giikkokka, taani guuttaa ceeqettanaadan, tana eeyya giite.Taani hagaadan xalada ceeqettaidda haasayiyoogee, eeyyatetta haasayappe attin, tana Godai azaziina gidenna.Daro asai ha sa7aaban ceeqettiyo gishshau, taanikka ceeqettana.Intte aadhdhida eranchcha gidiyo gishshau, eeyya asaabaa ufaissan gencceeta. Aissi giikko, ooninne inttena ailleyikko, woi inttebaa miikko, woi inttebaa bonqqikko, intte bolli otorettikko, woi intte qinxxalliyaa baqqikko, gencceeta. Nuuni hegaa oottanau daro daafuranchcha gididoogaa taani yeellataidda yootais. SHin ooninne issiban xalidi ceeqettikko, taanikka ceeqettanau xalais; taani eeyya asadan haasayais. Eti Ibraawetee? Taanikka Ibraawe. Eti Israa7eela asee? Taanikka Israa7eela asa. Eti Abrahaama zaree? Taanikka qassi Abrahaama zare. Eti Kiristtoosa ashkkaratee? Taanikka etappe aadhdhiya ashkkara; taani gooyyiya asadan haasayais; taani daro ootta daafuraas; daro wodiyaa qashettaasinne lissuwan garafettaas; qassi daro wodiyaa haiqqaichchada attaas. Aihudati tana ichchashutoo, hasttamanne uddufunaa lissuwan garafidosona. Roome asai tana heezzutoo dullan shociis; issi gallassi shuchchan caddii caddidi aggiis; markkabee heezzutoo me77in, abba giddon issi qammaanne issi gallassaa aqada pe7aas. Taani daro ogiyaa hemettaas; tana kixa haattinne panggi daafirggidi ashshiis; qassi ta biittaa asainne Aihuda gidenna asai daafirggidi ashshidosona. Qassikka katamai, bazzoi, abbai,

tana yashissiis; worddanchcha dabbotikka tana daafirggidi ashshidosona. Taani daafuraasinne metootaas. Darotoo xiskkuwaa xayaas; namisettaas; saamettaas. Darotoo xoomaas; qassi meegettaasinne kallottaas. Harabaa ubbaappekka ubba gallassi tau tooho gididabai, woosa keetta ubbau hirggiyoogaa. Daafuriyai oonee? Taani daafurikkinaayye? Nagaran geliyai oonee? Taani un77ettada ayyo qarettikkinaayye?Taani ceeqettanau koshshiyaabaa gidikko, taani daafuranchcha gidiyoogaa erissiyaaban ceeqettana.Merinau anjjettida Xoossai, nu Godaa Yesuus Kiristtoosa Aawai, taani worddotennaagaa erees.Damasqqo kataman Aretaasa giyo kawuwaappe garssaara deriyaa haariyaagee, tana oittanau Damasqqo katamaa naagissees. SHin katamaa dirssaa maskkootiyaara tana keeshiyan duge wottin, appe kessa ekkaas.

Ceeqettiyoogee taayyo ainne allaalle kessennabaa gidikkokka, taani ceeqettanau bessees. SHin taani ajjuutaabaanne Godaappe qonccidabaa haasayana.Tammanne oiddu laittappe kase, heezzantto saluwaa pude ekettida issi bitaniyaa, Kiristtoosa ammaniyaagaa, taani erais; he bitanee ba asatettaara biidaakkonne woikko ajjuutan biidaakko, taani erikke; Xoossai erees. He bitaniyaa, heezzantto saluwaa pude ekettidaagaa, erais; i ba asatettaara biidaakkonne woikko ajjuutan biidaakko, taani erikke; Xoossai erees. Hegan he bitanee asi haasayanau danddayennabaa, asa inxxarssikka haasayanau danddayennabaa siyiis.Taani he bitaniyaaban ceeqettana; shin taani daafuranchcha gidiyoogaa erissanaappe attin, ta huupheban ceeqettikke.Taani ceeqettanau koyiyaabaa gidikkokka eeyyikke; aissi giikko, taani tumaa haasayais. SHin asi tabaa tanan be7iyoogaappenne taappe siyiyoogaappe aattidi qoppenna mala ceeqettikke. Qassi Xoossai qoncciyo aadhdhida gitatettan taani otorettiyoogaa digganau tana caddi oiqqiyaagee, Seexaanaa kiitanchchai, tana dechchanaunne taani otorettenna mala, tana naaganau taayyo imettiis. taappe shaakkana mala, taani Godaa heezzutoo woossaas. Godai tana, “Ta aaro kehatettai neeyyo gidana; aissi giikko, ta wolqqai daafuran polettees” yaagiis; hegaa gishshau, Kiristtoosa wolqqai tanan de7ana mala, ta daafuran ceeqettanau taani daro ufaittais.Taani Kiristtoosa gishshau daafuriyoogan, asan cayettiyoogan, waayettiyoogan, yedetettiyoogan, metootiyoogan ufaittais; aissi giikko, taani daafuriyo wode, he wode minnais.Taani ceeqettidoogan eeyyaas; shin hegaa intte tana wolqqan ootissideta; aissi giikko, intte tana galatanau bessees; taani aibanne gidana xayikkokka, Yesuusi kiittido ubbaappe aadhdhiyaageetuppe laafikke.Yesuusi kiittidoogee tuma gididoogaa erissiya malaatainne oorattabainne Xoossai oottiyo malaatai intte giddon daro genccan oosettiis. Taani ta huuphen inttena waissabeennaagaappe attin, hara woosa keettatuppe intte aibin laafeetii? Hayyanintta

taani inttena naaqqidoogaa taayyo atto giite.Taani inttekko baanau giigettishin, hagee taayyo heezzantto wode; taani inttena waissikke. Taani inttena koyiyoogaappe attin, intte miishsha koyikke; aissi giikko, aayyiyaanne aawai bantta naatuyyo miishsha minjjiyoogaappe attin, naati bantta aayeeyyoonne aawaayyo minjjokkona. Taani intte shemppuwaayyo taayyo de7iyaabaa tanaara gattada ufaissan immana; yaatin, taani inttena keehi siiqiyo gishshau, intte tana ashshi siiquuteetiiyye? Gidikkokka, taani inttena waissabeikke; shin intte tana, “Neeni nuna cimmada wordduwan oiqqadasa” yaagana.Taani harai atto, inttekko kiittido issi asa baggaarakka inttena cimmidanaa? Tiitu inttekko baana mala, taani woossaas; qassi ammaniya ishaakka aara yeddaas. Yeddin Tiitu inttena cimmidee? Nuuni nu oosuwaa issi qofan oottibookkonii? Qassi nu hanoikka issuwaa gidenneeyye? Nuuni ubba wodekka intte sinttan nu huupheyyo mootettiyoobaa intteyyo milatii? Nuuni Kiristtoosan de7iyaageetudan, Xoossaa sinttan ha77i gakkanaassikka haasayoos. Nu siiqotoo, nuuni oottiyo ubbabaikka inttena maaddanaassa. Aissi giikko, taani inttekko biyo wode, ooni erii taani koyiyoogaa mala intte gidennan agganaakkonne; qassi taanikka intte koyiyoogaa mala gidennan agganaakkonne. Qassi ooni erii hegan palamai, qanaatee, hanqqoi, yiiqetiyoogee, zigirssai, saasukkiyoogee, otoroi, metoi de7ennan waayi aggana; tana yashshees. Taani inttekko naa77anttuwaa biyo wode, ta Xoossai intte sinttan tana kaushshanaakkonne yayyais; taani nagara kase oottidi, bantta oottido tunatettaappe, shaaramuxatettaappenne amuwaappe simmibeenna daro asatuyyo yeekkennan aggikke gaada yayyais.

Asai ubbai deriyaa heemmiyaageetuyyo haarettanau bessees; aissi giikko, Xoossai geennan de7ishin, dere heemmiyaabi de7anau danddayenna. Ha77i de7iya heemmiyaageetikka Xoossai wottidoogeeta. Hegaa gishshau, ha77i de7iya heemmiyaageeta ixxiya ooninne Xoossaa azazuwaa ixxees; ixxiya ooninne ba bolli pirddaa ehees. Aissi giikko, heemmiyaageeti iitabaa oottiyaageeta yashshiyoogaappe attin, lo77obaa oottiyaageeta yashshokkona. Heemmiyaagaayyo yayyennan agganau koyai? Yaatikko lo77obaa ootta; i nena galatana.Aissi giikko, i neeyyo lo77obaa oottanau, Xoossau ashkkara. SHin neeni iitabaa oottikko, iitabaa oottidaagaa i seeriyo gishshau, ayyo yayya; i Xoossaa ashkkara. Qassi iitabaa oottiyaageetu bolli Xoossaa hanqquwaa i ehees.Hegaa gishshau, deriyaa heemmiyaageetuyyo azazettiyoogee Xoossaa hanqquwaa xalaalaassa gidennan, inttena zoriya wozanaa gishshaukka haarettanau bessees. Aissi giikko, deriyaa heemmiyaageeti bantta oosuwaa oottiyo wode, Xoossau oottiyo gishshau, intte giiraa

giiriyoogee hegaassa. Ubbau bessiyaagaa immite. Giiraa giiranau bessiyoogau giirite; qaraxaa qanxxanau bessiyoogau qanxxite; yayyanau bessiyoogau yayyite; bonchchoi bessiyoogaa bonchchite. Intte issoi issuwaara siiqettanaappe attin, o acoinne intte bolli de7oppo. Aissi giikko, ba shooruwaa siiqiya ooninne higgee giyoogaa poliis.“SHaaramuxoppa; woroppa; wuuqqoppa; asabaa amottoppa” giya azazotinne hara azazo ubbati, “Ne shooruwaa ne huuphedan siiqa” giya azazuwan kuuyettoosona. Ba shooruwaa siiqiya ooninne ayyo iitabaa oottenna; hegaa gishshau, shooruwaa siiqiyoogee higge ubbau azazettiyoogaa. Xiskkuwaappe intte beegottiyo saatee ha77i gakkido gishshau, ha wodiyaa akeekite. Aissi giikko, nuuni koiro ammaniyo wodeppe, nuuni attiyo wodee ha77i nuukko matattiis. Qammai aadhdhanau hanees; gallassai matattiis. Simmi nuuni xumaa oosuwaa aggidi, poo7uwan olettanau, tooraa gonddalliyaa oiqqoos; gallassi poo7uwan de7iya asadan, ane maara de7oos. Yettaaninne mattuwan gidoppo; shaaramuxiyoogaaninne shaaramuxanau qaaqqatiyoogan gidoppo; palamaaninne qanaatiyan gidoppo.SHin Godaa Yesuus Kiristtoosa tooraa gonddalliyaa oiqqite; intte nagaranchcha asatettai ba wozanaa amuwaa polanaadan, au qoppoppite.

Ammanuwan azalliya uraa shiishshi ekkite. SHin i ba huuphen qoppiyooban aara palamettoppite. Issi ura ammanoi i ai miininne diggenna. SHin ammanuwan azalliya urai gaden mokkidabaa xalaalaa mees.Ubbabaa miya urai ubbabaa meenna uraa karoppo; ubbabaa meenna uraikka ubbabaa miya uraa bolli pirddoppo. Aissi giikko, Xoossai a shiishshi ekkiis.Hara asa ashkkaraa bolli pirddiyaagee neeni oonee? I kunddin woi denddin ba godaassa; shin Godai a denttanau danddayiyo gishshau, i eqqana. Issi urai, “Hiqqa gallassai hara gallassatuppe aadhdhees” yaagidi qoppiyo wode, harai, “Ubba gallassatikka issi mala” giidi qoppees. Asai ubbai huuphiyan huuphiyan hegaadan aissi qoppiyaakko akeeko.Issi gallassaa hara gallassatuppe aattidi bonchchiya urai Godaa bonchchuwau bonchchees; quma miya uraikka Godaa bonchchuwau mees. Aissi giikko, i he qumaa gishshau, Xoossaa galatees. Qumaa meennaageekka Godaa bonchchuwau meennan aggees; qassi Xoossaa galatees. Aissi giikko, nuuppe ooninne ba huuphe xalaalau de7enna; qassi ooninne ba huuphe xalaalau haiqqenna. Nuuni paxa de7ikko, Godau de7oos; qassi haiqqikkokka, Godau haiqqoos. Simmi nuuni paxa de7in woi haiqqin Godaassa. Hegaa gishshau, Kiristtoosi paxa de7iyaageetuyyoonne haiqqidaageetuyyo Goda gidanau haiqqidi, haiquwaappe denddiis.Yaatin, neeni ne ishaa bolli aissi pirddai? Woikko neeni ne ishaa

aissi karai? Nuuni ubbaikka Xoossaa pirddaa araataa sinttan eqqana. Aissi giikko, Xoossaa maxaafai, “'Taani Godai ta de7uwan caaqqais; asai ubbai ta sinttan gulbbatana. Qassi taani Xoossaa gidiyoogaa asai ubbai markkattana' yaagais” yaagees. simmi nuuni huuphiyan huuphiyan nubaa Xoossau yootana.Hegaa gishshau, nuuni issoi issuwaa bolli pirddiyoogaa hagaappe sinttau ane aggoos. SHin ubbaappe aattidi, intte ishaa xubbiyaabaa woikko nagaran onggiyaabaa oottennaadan qofaa qachchite.Taani Godaa Yesuus Kiristtoosaara de7iyoogan, ba huuphen tuna gididabi bainnaagaa eraisinne akeekais. SHin issi urai issibaa tuna giidi qoppikko, hegee au tuna gidees. Neeni miyoobaa gishshau, ne ishaa qohikko, simmi neeni siiquwan de7akka. Kiristtoosi a gishshau haiqqido asa ne miyo quman bashshoppa. Simmi neeni lo77obaa giyoobaa asi cayoppo. Aissi giikko, Xoossaa kawotettai Geeshsha Ayyaanai immiyo xillotetta, sarotettanne ufaissa gidiyoogaappe attin, muussanne ushsha gidenna. Aissi giikko, Kiristtoosayyo hagaadan oottiya urai ooninne Xoossaa ufaissees; asaikka a galatees.Yaatikko, ane nuuni sarotetta ehiyaabaanne nuna issuwaa issuwaara minttettiyaabaa ubbaa koyoos. Xoossaa oosuwaa quma gishshau bashshoppa. Quma ubbaa min aikko baawa; shin hara ura nagaran yeggiyaabaa miya urau iita.Asho meennan woi ushsha uyennan aggiyoogee woikko ne ishaa xubbiyaabaa aibanne oottennan aggiyoogee lo77o. Ha yohuwan neeni ammaniyoobaa neeppenne Xoossaappe gidduwan naaga. Issi urai lo77o giidi ba qoppidobaa oottiyo wode, “Taani mooraas” geennan aggikko, he urai tumu anjjettidaagaa.SHin i ba miyoobaa sirikko, a meettai ammanuwaana gidenna gishshau, Xoossai a bolli pirddiis; qassi ammanoi bainnan oottidobi aibinne nagara.

Appendix 6: Sample code of the system

```
import re

import os

import math

import sys

from operator import itemgetter
```

```
stwList=open('SWList.txt','r')
stopWstr=stwList.read()
stwList.close()
sfxList=open('suffixList.txt','r')
suffixString=sfxList.read()
sfxList.close()
os.chdir('test_corpus')
suffixList=re.split(r'\W+',suffixString)
stopWList=re.split(r'\W+',stopWstr)
string=''
stemmedStr=''
stemmedList=[]
def radical(tkn):
    consonant=['b','c','d','f','g','h','j','k','l','m','n','p','q','r','s','t','v','w','x','y','z','c','h','7']
    twoChar=['dh','nh','ny','ph','sh','ts','zh']
    count=0
    prevC=''
    currC=''
    tk=''
    F=0
    newTkn=''
```

```
    for twC in twoChar:
        if tkn.count(twC)!=0:
            count=tkn.count(twC)
            F=1
            tk=tkn.replace(twC,'')
    if F==1:
      newTkn=tk
    else:
        newTkn=tkn
    for t in newTkn:
            currC=t
            if t in consonant:
                if prevC!=currC:
                    count=count+1
                    prevC=currC
            else:
                    prevC=''
    return count
def stemmer(token,rad):
    if rad==1:
        if token.find('s')+1 == token.find('h'):
```

```
        return token.replace('sh','y')
    elif token.endswith('ee'):
        return token.rstrip('ee')
    elif token.endswith('aa')or token.endswith('a'):
        return token
    elif rad==2:
        for ds in suffixList:
            if token.endswith(ds):
                token=token.rstrip(ds)
                rad=rad-1
                return stemmer(token,rad)
        else:
            return token
    elif rad>2:
        for sl in suffixList:
            if token.endswith(sl):
                token=token.rstrip(sl)
                rad=rad-2
        return stemmer(token,rad)
    else:
        return token
```

```
for files in os.listdir("."):
    if  files.startswith('test_corpus'):
        txt=open(files,'r')
        read=txt.read()
        string=read
        string=string.lower()                    #Normalization
        token=re.split(r'\W+',string)              #Tokenization
        for t in token:
                radic=radical(t)
                stemmed=stemmer(t,radic)
                if stemmed not in stopWList:            #Elimination of  stop words
                     stemmedList.append(stemmed)
                else:
                   continue
        line=["","","","","","","","","","","","","","","","","","","","","","",""]
        lInd=0
        line[20]='\n'
        for stem in stemmedList:
           if line[lInd]=='':
              line[lInd]=stem
              lInd=lInd+1
```

```
        else:
          for word in line:
            stemStr=str(word)                        #convert list to string
            stemmedStr=stemmedStr+stemStr + ' '
          fil=open('stemmed_test_corpus.txt','a')
          fil.write(stemmedStr)
          line=['','','','','','','','','','','','','','','','','','','','','']
          lInd=0
          line[20]='\n'
          stemmedStr=''
      fil.close()
stemmedFile=''
for file in os.listdir("."):
    if file.startswith('stemmed'):
      f=open(file,'r')
      stemmedFile=f.read()
      f.close()
termAndFrq={}
frqTerm=[]
stemmedFile=re.split(r'\W+',stemmedFile)
for word in stemmedFile:
```

```
    tF=stemmedFile.count(word)
    if tF>16:
        termAndFrq[word]=tF
        frqTerm.append(word)
print (termAndFrq)
print ('----------------------------------------------------------------------------------------------------')
SumOf_tF=0
for t,F in termAndFrq.items():
   SumOf_tF=SumOf_tF+F
coMatrix={}
coTerm=''
coTerm=''
coMatrix={}
for t in frqTerm:
   for coT in frqTerm:
        if t== coT:
            pass
        else:
            cF=0
            coTerm=coTerm+t+' '+coT
            with open('stemmed_test_corpus.txt','r')as f:
```

```
            for line in f:
                if t in line:
                    if coT in line:
                        cF=cF+1
            f.close()
        coMatrix[coTerm]=cF
        coTerm=''


print (coMatrix)
print('----------------------------------------------------------------------------------------------------------')
Filtered={}
trmL=[]
for tms,cv in coMatrix.items():
    trmStr=''
    trmL=[]
    trmL=re.split(r'\W+',tms)
    trmL.reverse()
    trmStr=trmStr+trmL[0]+' '+trmL[1]
    if trmStr not in Filtered:
        Filtered[tms]=cv
    else:
```

```
        continue
print (Filtered)
print('----------------------------------------------------------------------------------------------------')
SumOf_cF=0
for t,cF in coMatrix.items():
    SumOf_cF=SumOf_cF+ cF
SumOf_cF=SumOf_cF/2
MI={}
coT=[]
for trms,cf in Filtered.items():
        coT=re.split(r'\W+',trms)
        if cf==0:
            MI[trms]=0.0
        else:
MI[trms]=math.log(float((cf/SumOf_cF)/((termAndFrq[coT[0]]/SumOf_tF)*(termAndFrq[coT[1]]/SumOf_tF))),2)
print (MI)
print('----------------------------------------------------------------------------------------------------')
sortedMI={}
maxKey=''
maxValue=0.0
valueList=[]
```

```
keyList=[]
for tm,cv in MI.items():
    if tm in keyList:
        pass
    else:
        maxKey=tm
        maxValue=cv
        for t,v in MI.items():
            if t in keyList:
                pass
            else:
                if maxKey==t:
                    continue
                elif maxValue < v :
                    maxValue=v
                    maxKey=t
                elif maxValue >= v:
                    continue
        keyList.append(maxKey)
        valueList.append(maxValue)
print('-----------------------------------------------------------------------------------------------------')
```

```
print (keyList)

print (valueList)

print('----------------------------------------------------------------------------------------------------------')

term1=''

term2=''

splited=[]

print ('Term'+'\t\t'+ 'Related term')

print ('-----------------------------------------------')

for k in keyList:

    splited=re.split(r'\W+',k)

    print (str(splited[0])+'\t\t'+str(splited[1]))
```