ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

APPLICATION OF MULTILINGUAL THESAURI FOR CROSS LANGUAGE INFORMATION RETRIEVAL (CLIR)

[AMHARIC-ENGLISH CLIR FOR THE LEGAL ENVIRONMENT]

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

YOSEPH SHIFERAW

JUNE, 2005
ADDIS ABABA

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

APPLICATION OF MULTILINGUAL THESAURI FOR CROSS LANGUAGE INFORMATION RETRIEVAL (CLIR): AMHARIC - ENGLISH CROSS LANGUAGE INFORMATION RETRIEVAL FOR LEGAL ENVIRONMENT

BY
YOSEPH SHIFERAW

Name and Signature of Members of the Examining Board

Ato Getachew Jemaneh, Chairman, Examining Board	Getachew Jemaneh	
Dr. Nega Alemayhu, Advisor		
Ato Tesfaye Biru, Examiner		
Chairman, Faculty	Getachew Jemaneh	12/07/05
	Signature	Date
Chairman, Graduate Council	Signature	Date

DEDICATION

I would like to dedicate this research to

my Mom Ethiopia Semret and my Dad Shiferaw Asfaw

ACKNOWLEDGEMENT

I would like to thank quite a number of people that have been involved one way or the other while I went about doing my research as well as in graduate school. I would like to extend my respect and gratitude for my advisor Dr. Nega Alemayehu for his critical advice, encouragement and support. I would also like to thank my co-advisor Ato Wondwossen Demisse for his assistance in the area of law.

I would like to thank my beloved parents Shiferaw Asfaw and Ethiopia Semret. Thank you for all you have done for me all these years. Thanks for all the late night dinners, putting up with my complaints and most of all my phone bills. Thank you for never letting me run broke. I would also like to thank my brother Dr. Mahdere Shiferaw, my sisters Dr. Tirsit Shiferaw and Hudad Shiferaw (RN). You guys are an inspiration.

I would also like to thank my friends. Yared Ayalew, thank you for all your help in the .NET framework. I owe you big time. My classmates, Redu, Yaller and Tefe, I will never forget those nights at the lab trying to make assignments and papers on time. Very wonderful and painful memories as well. And of course, to the best lawyer/judge friends of mine, especially Selamawit Tesfaye and Selome Argaw, for their patience and dedication of their time in helping out with thesauri construction – we will have one on me – thank you very much.

In the department of Information Science, I would like to thank all our instructors and staff for all they have done. I would like to thank Ato Mesfin Getachew for his willingness to help out always and W/t. Atelach Alemu for her comments at the right time. Meseret Ayayno of our bibliographic library – you are the best, thank you. And of course, W/ro Ethiopia Tadesse, for doing all that you did to make things happen in such a short time and more. I am sure there are more I have missed to mention, I hope I will get to thank you on person.

Yoseph Shiferaw

ABSTRACT

It is crucial for cross language retrieval tools to be capable of translating queries and/or documents to make information accessible. In areas where such efficient systems are lacking for language pairs like Amharic and English, intermediate tools like the thesauri can be used to make translation of queries possible.

This research has developed parallel thesauri for business law to be used in testing the retrieval of thesauri based translation of queries across the two languages. The thesauri are developed by taking the commercial code of Ethiopia as a representative corpus and domain experts in law were involved to make the conceptual analysis and facet determination. This procedure was supplemented by machine assisted indexing.

An in-house retrieval system has been developed to test the application of the multilingual thesauri developed for this purpose. Thesauri based retrieval is tested where concept based translation of queries is attempted as compared to word based translations as in the case of machine translation. Queries have been collected from legal experts and a document collection in the area of law has been developed from research abstracts for this purpose. To test retrieval performance, queries are translated to their equivalent concepts using the thesauri and the equivalent concepts are used to query the collection in the target language.

Retrieval outputs measured in terms of precision and recall show promising results as they have managed to retrieve relevant documents across languages. However, the obtained performance can be made better if other resources are made available as indicated in the recommendation. Therefore, these results have made the recommendation of other tools as well as approaches as being important to arrive at better retrieval performance.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER ONE

1.1 INTRODUCTION

The rate at which information seems to be produced and made available via resources like the internet is tremendous. Though increased networking and communication technologies have made it easier to gain access to these resources, it remains a daunting task to actually locate the desired content – given the myriad resources. Thus, the availability of tools that can automatically locate and retrieve documents to users need is becoming more common. In light of this Oard (1996) presents, "automated systems capable of detecting useful documents are finding widespread application".

The popularity of these systems is however, not as prevalent for the other languages as seen in the availability of such tools for information retrieval for developed languages like English. Such systems are judged to be even more useful in case of monolingual users (Oard 1996), especially to users that are not fluent speakers of languages for which resources are abundantly available. This would also imply that other resources in languages that don't have large number of speakers would also face a similar problem in becoming accessible.

Given the above consideration for such languages, the need for resources available in other languages becomes an issue for the user. Hence, the question in this case would be; how to make these resources in different languages available to the potential user? Better yet, when asking this question, another one comes along. That is before making these resources available, there comes a question of locating these resources and issues of querying in a multilingual environment. Thus referring to what Oard (1996) regarded as "automated systems capable of detecting useful documents", as applying to a multilingual environment' there would be the need to develop translation resources either in search and/or retrieval.

Hence, from the user perspective, multilingual information retrieval can be defined as a selection of useful documents containing collections that may contain several languages (Oard, 1996). To be noted in this case, is the fact that the collection being searched may contain documents composed of more than one language. Now while this may foster ideal circumstances for a user who is multilingual, it presents another hurdle that has to be overcome for the monolingual one. That will be the translation of the retrieved documents. But at the same time, it is quite rare that we will find users who have fluent abilities in more than two languages, and therefore the need for language and retrieval resources in a multilingual environment will have to come into picture.

The use of such systems is particularly of great importance for countries like Ethiopia, where monolingual users are common in relation to the languages commonly available in relation to electronic resources. In such circumstances, given the resources in different languages, like English for example, performing retrieval after translation to a local language like Amharic cannot be justified considering the resources required. Hence a more economical approach in relation to the retrieval and translation would be to translate retrieved documents after retrieval. Then one would be able to justify the resources required for the translation of the retrieved documents that have been judged relevant.

All in all, retrieving relevant documents across languages is bound to present more interesting and challenging problems as compared to the monolingual environment. This is because we not only find the commonly seen problem of how to find documents in a monolingual environment, but also the additional translation problems of query and/or documents. Several approaches have flourished in the last few decades in light of CLIR. These mainly include controlled vocabulary approaches, dictionary based approaches and machine translation (MT).

In that regard, research in the field of information retrieval has shown that controlled vocabularies improve both precision and recall (ANSI/NISO Z 39.19-1993). One such tool exhibiting controlled vocabulary is the thesaurus. Thesaurus, according to the ANSI/NISO Z30.19-1993 standard, is defined as "A controlled vocabulary of terms in natural language that are designed for post-coordination". It further progresses to present the complexities of natural language necessitating the use of controlled vocabulary for indexing.

On the other hand, Soergel (1997) defines a thesaurus as "a structure that manages the complexities of terminology in a language and provides conceptual relationships, ideally through embedded classification and ontology". We can see further that a thesaurus provides relationships based on semantic implications of terminology. Hence, the above definitions indicate that a thesaurus is a controlled vocabulary that is formally organized showing prearranged relationships between concepts.

Soergel (1997) further explains the use of the thesaurus can be as a knowledge-based support for free text searching as well as for controlled vocabulary indexing and searching. Therefore, in whichever fashion we wish to employ a thesaurus, its obvious contributions can be seen from different perspectives. In one sense, it can enable grouping of terms into hierarchical classes thereby representing the context and content of the particular domain. At the same time, given the terms that are similar and used to express the same concept, a preferred term can be looked up or presented by the thesaurus to enhance the search outputs. Users may also broaden or narrow down their searches by navigating through the hierarchical structure of concepts presented. And of course, continuous maintenance of the thesauri will serve as a learning ground at the same time to recent collections or represent new or previously unfamiliar concepts.

The application of controlled vocabulary, or more precisely, thesaurus based indexing and searching is not limited to a monolingual environment. It has extended its application to Cross Language Information Retrieval (CLIR) as can be seen from

the track reviews continuously presented at the Text Retrieval Evaluation Conference (TREC)[1]. CLIR represents "...use of queries in one language to retrieve documents from a multilingual pools of documents". Other approaches, beside query translation, include document translation and a combination of both.

The thesaurus based CLIR gained popularity mainly through the inaccuracy of present stage automatic machine translation and probably because it is thus far one of the few experimental approaches implemented in multilingual environments like the EUROVOC thesauri in Europe. Given other services like the Information Service for Physics, Electronics and Computing (INSPEC), that produce English abstracts for documents in other languages, the application of such systems may be worth while the effort as well. And in response to this, more focused researches have been and are being undertaken like the European Multilingual Information Retrieval (EMIR), on specific domain multilingual thesauri like the Arts and Architecture Thesauri (AAT) and a variety of other researches by the Defense Advanced Research Projects Agency (DARPA).

[1] Available online: http://www.trec.nist.gov

1.2 STATEMENT OF THE PROBLEM

The world fosters people and cultures so vast and different that at times it is a wonder that such difference actually exists. One main difference among society is brought about by differences in language. Language differences have made communication among people a bit difficult. Looking at the level of development that we have reached today however, we see efforts aimed at overcoming this barrier as in the case of TREC (Text REtrieval Evaluation Conference) and CLEF (Cross Language Evaluation Forum).

In relation to the above, Information retrieval covers problems relating to the effective storage, processing and retrieval of documents in relation to user needs, given a query. Given the amount of information and information resources, nothing can be more frustrating than searching with respect to natural language uncertainties. To make matters worse, considering a large number of documents in a multilingual environment, those uncertainties and difficulties multiply.

These uncertainties and difficulties include problems of synonymy, ploysemy, and homonymy in the case of monolingual environment and translation problems in case of multilingual environment. Looking at the conditions of free text searching or searching in an environment where consistent indexing is not available, the use of queries that can ultimately bring about best results will be for the very few experts in the particular domain. At the same time, cross-cultural interactions and globalized activities make information storage and retrieval unlimited to a single monolingual context.

On the other hand, non-experts (in searching as well as to a specific domain) face problems of accessing information repositories with well-formulated queries. Some would like to get a better insight into issues surrounding their question and some

would like help in formulating their query. Hence, a need to clearly represent knowledge in the particular domain becomes an issue. Therefore, one question would be "How to guarantee effective control of vocabulary to represent concepts as well as terminological standard, provide a systematic presentation of knowledge, and at the same time enable better retrieval through well-formulated queries?"

Extending this problem to the multilingual environment, another logical question would follow. That is, "How can we provide translation facilities that can provide systematic retrieval of documents to the user in conditions where documents are available in a variety of languages?" These problems have generally prompted a great deal of research but remain still inconclusive and in some cases are not so well experimented for languages like Amharic.

In light of the above, the legal environment here in Ethiopia presents such a problem, where there exists a multilingual environment in academia, research and practice. We see a variety of publications, including proclamations, research papers as well as court house proceedings being implemented in two languages, namely Amharic and English.

These circumstances, as can be seen in the libraries of the Addis Ababa University and the Federal Supreme Court, have created problems for users to access information in either of the languages they are not fluent in. We can constantly see conditions where they face problems in addressing or explaining their information requirements when it comes to either one of the languages. Therefore, this research has tried to make attempts in providing an experimental indication to answer the two questions posed above in relation to vocabulary control and translation to enable better access to information.

1.3 OBJECTIVES OF THE STUDY

General Objectives

This research mainly focused on designing and developing multilingual thesauri to be tested in a multilingual environment. The thesauri to be developed are intended to offer a controlled vocabulary that can serve as a terminological standard as well as a source of organized knowledge reference. The initial thesaurus is developed in English and subsequently extended to Amharic through appropriate transformation. A search system is developed in-house to test the applicability of the tool.

Specific Objectives

The specific objectives of this research include major activities to enable the development of the thesauri and design of the retrieval system. They refer to tasks that need to be accomplished in the development of the translation tool as well as the retrieval of the documents from the collection. They are:

1. Investigation, selection, and review of available literature on thesaurus history, uses and of existing approaches to thesaurus construction.
2. Evaluation and selection of approaches for thesaurus construction.
3. Acquisition of concept bearing terms through semi automatic approach from a corpus obtained for this purpose.
4. Formulation and definition of concepts by use of facet analysis to determine concept spaces.
5. Organization of concepts into concept classes by establishing hierarchical relationships.
6. Design of appropriate access mechanisms to the thesauri.
7. Development of a document collection to be used in the retrieval.

8. Design and development of a retrieval system to be used in testing the performance of the developed thesauri.
9. Evaluation of results of thesauri based retrieval against a normally indexed environment using methods of precision and recall.
10. Recommend a way forward for the development of such system in information retrieval in the local context.

1.4 METHODOLOGY

Considering the extensive network environment, make available information across boundaries and cultures requires surmounting the linguistic barrier for the information to be accessed and understood in the fashion it was initially intended. This research mainly progressed to undertake the following major approaches in order to develop a system that is capable of surmounting the linguistic barrier.

I. Domain area for Thesauri

This study has initially progressed to identify such a domain where languages work together in academia, practice and day-to-day activities. One such domain is the legal environment here in Ethiopia. The academia, research and practice seem to employ two sets of languages, namely Amharic and English. We find proclamations coming out in the Federal Negarit Gazette containing direct translations in both languages along with the constitution, thus making up perfectly aligned parallel corpora.

Having identified the legal domain, a presentation of the historical evolution of thesaurus, its uses, applications, and design are given. Existing thesauri that have gained popularity and acceptance in the area of information retrieval are identified. It then gives an illustration of existing approaches towards thesauri construction starting form the early approaches to presently upheld ones. This evaluation also formed the basis for the selection of an approach used in this study.

Following was the evaluation and selection of appropriate approaches for thesaurus construction. This was not a simple issue of selecting the best approach, as there is no best approach but rather each approach depends on the purpose for which it was intended for. Other factors like the availability of the tools and applicability according

the study's context will also influence the selection. Issues relating to all design tools consideration will be done at this stage to enable a more practical approach.

II. Thesauri Construction

a) Creation of a term bank

In the actual stages of thesaurus construction the first step was to produce an index of the particular domain from which the acquisition of terms and concepts was carried out from. Here, the deductive method of compiling the terms and concepts is implemented based on a semi-automatic indexing and an independent compilation of candidate terms by domain experts. In this method, terms to be collected are primarily derived from the documents indexed (obtained from the Federal Negarit Gazette), as well as the domain experts, hence, the number of concepts that may remain unidentified were minimized.

b) Formation and definition of concepts

After terms and concepts of the legal domain were obtained, the formation and definition of concepts to mirror consistent use of terminology and at the same time display clearly understandable concepts to the user was done. Factors that were considered for control at this stage include grammatical forms, spelling, singular vs. plural representations as well as abbreviations and compound use of terms. Attempts were also made to control the meaning by selection through synonym sets that are identified at this stage.

c) Facet analysis

The next stage required the organization of concepts as stated in the objectives of this paper. The organization of concepts was done by adopting the process of facet analysis. Faceted analysis enabled to identify the different sets of subject isolates in the domain. This classification of the domain into non-overlapping sets of concepts was done with domain experts.

The next stage was to assign the various terms that have been identified in the facet analysis to their respective concept class/facet. Along with the assignment of concepts classes, the identification and assignment of three major types of relationships was carried out. These inter-term relationships related to equivalence, hierarchical and associative relationships. The equivalence relationship assigned terms corresponding to the same concept with a preferred instead of non-preferred term. Correspondingly, some terms were not necessarily referring to the same concept, but rather to parts of it or even bigger context. Such conditions have the hierarchical relationships assigned. On the other hand, some terms may display an overlapping reference to a concept though each term may at the same time represent another concept. Hence, the associative relationship of has been assigned for such terms.

This stage represented major portion of the overall project where the English monolingual thesaurus was finalized. This thesaurus needed to be converted to an equivalent Amharic thesaurus, to arrive at the bi-lingual thesauri, which is the aim of the study. The ISO 5964 (Guidelines for the establishment and development of multilingual thesauri) recognizes three major approaches to the construction of multilingual thesauri. The first one is the so-called "Ab initio" approach where establishment of a new multilingual vocabulary is done form scratch. The second approach is a translation of an existing monolingual thesaurus, and the third is the reconciliation and merging of existing monolingual thesauri in two or more working languages. Given the context of this study, since the corpora used in this case was the Federal Negarit Gazette, which represented a perfectly aligned parallel corpora, the second approach of translation of the English thesaurus was undertaken to arrive at the multilingual thesauri.

Given both languages, the thesauri were constructed by mapping the equivalence, hierarchical and associative relationships by choosing a source language per the ISO 5964 standard. The terms, as previously stated, came from the parallel corpora presented by the publications of the Federal Negarit Gazette.

III. Developing a document collection for testing retrieval performance of thesauri

Test documents were compiled from the Ethiopian Journal of Law as well as from the library of Social Science Research Network – SSRN Abstracts database[2]. The documents collected from the Ethiopian Journal of Law were compiled from the publications obtained from the year 1964-2000. Though the years seem long enough to develop a good document collection, publication came out only twice in the years 1974-1991. And the other years are highly irregular. Though the Journal represents documents that have been published in the general area of law, research abstracts are selected from the available publications in the law library of Addis Ababa University. These documents have been converted to soft copy to enable electronic processing.

IV. Developing queries for use in the retrieval evaluation

Queries have been collected from students in Faculty of Law, Addis Ababa University. The basis for the selection of these students was that they have to be in their final years of their study, either in undergraduate or postgraduate schools, and they also need to be researching in the general area of business law. Accordingly, 25 queries in all for both languages were collected and used to test the performance of the developed system.

V. Design of the Retrieval system

Given the thesauri, the next step was to design an architecture that enabled the storage of the two sets of thesauri in a fashion that will enable concept mapping and translation for retrieval purpose. For this a database tool capable of implementing the storage and cross referencing of concept classes and thesauri entries in both sets of

[2] Available online http://www.papaers.ssrn.com/sol3/displayabstractsearch.cfm

languages and a programming environment were selected. Hence, in this case Visual Basic.Net 2003 has been adopted and a backend of SQL Server 2000. This was done in light of availability and capabilities of the desired tools along with ease of implementation by the researcher given the time required to finalize the study.

VI. Evaluation of Results

Eventually the last stage of the study focused in evaluating the results by measuring the performance of the thesauri in a multilingual environment on a different set of test documents. Two sets of runs will be executed. Monolingual runs without the use of the thesauri and cross language runs with the use of the thesauri.

The measures of recall and precision are employed to measure the effectiveness of the retrieval process in a cross language environment for the documents that have been retrieved by the system.

1.5 APPLICATIONS SCOPE AND LIMITATIONS OF THE STUDY

This study mainly focused on developing prototype thesauri and retrieval system to be used in cross-language retrieval. Its major aim was to test the application of such popular system in overcoming the language barrier. It does not try to present this as the only solution for cross-language retrieval nor would the results be conclusive as the development is a prototype.

It presents a broader approach of indexing, in that it doesn't include all concepts presented in the domain. In this regard, performance limitations are seen due to the rather shallow indexing procedure followed. Similarly the application of an in-house developed system for retrieval and absence of morphological analyzers like stemmers and part-of-speech-tagger have also compromised the expected output. On top of this, though the study mainly focused on thesauri construction and design of retrieval system, the lack of other facilities (like test documents, available test beds, indexers, etc.) made it difficult to find sufficient time to do a more focused research.

On the other hand, the implementation of such a system can make access to information a lot easier through its efficient representation of knowledge and search facilities. Potential users of the system would include people in the academia, such as students, professors, judges, legal service customers, and local information users in general. Similarly, in the practicing environment this could be used to access information in relatively huge databases. And considering the recent publication of the Federal Negarit Gazette in a soft copy as a text collection, the implementation of such system in an integrated manner would greatly enable navigation through the system.

On top of this, the study sheds light into the possible incorporation of such systems to the information environment where languages seem to be barriers to the learning as well as practicing environment. Such thesauri may also be extended to further stages of development to cover larger dimensions of the languages thereby providing a huge bridge across cultures as well.

CHAPTER 2
LITERATURE REVIEW

2.1 *Cross Language Information Retrieval: An Overview*

The 21st century has brought with it its own revolution through overhauling communication with the extensive amount of information accessibility. This accessibility to information has, however, come along with its own barriers that need consideration. These issues, generally in the area of Information Storage and Retrieval, have become a major focus of research over the years. Aided with the rapid development of communication technologies, issues relating to information storage seem to have been abated more or less, whereas those relating to retrieval still remain with many more being discovered by the day and at the same time many more being identified.

Among the crucial areas of retrieval gaining attention is the issue of information access across languages. Much of the research has focused mainly on monolingual IR and according to Haddouti (1999), more focus has been given to English language, though English is a native language for about 6% of the world's population. Here a considerable gap can be seen growing where non-native speakers (or even non-speakers) of the English language would require access to information even though the information is not available in their own language. Accordingly, searchers of information with regard to language efficiencies can be broadly categorized into two (Gonzalo, 2002). The first category may represent a searcher that uses a strictly monolingual environment, as in the case of a native Amharic speaker having no idea of other languages including English. The other category represents searchers having a passive knowledge of a target language, where for example a searcher is a native Amharic speaker, has a working knowledge of English but may not be able to pose queries in English.

Both conditions mentioned above represent different demands in order to have their information needs satisfied. The first scenario would require full translation systems of query and document, whereas the second would require translation for appropriate search terms but not document, as the searcher might be able to grasp the general idea of the retrieved document having a passive knowledge of the other language. In this regard, Oard and Dorr (1996) point out three major motivations that could form the basis for Cross Language Information Retrieval (CLIR).

The First motivation is in consideration of document collections in quite a number of languages where query formulation for each language would become a cumbersome process. The second is in relation to documents containing text in more than one language. And the third is for users that are not capable in forming queries in other languages but still can make sense of retrieved documents, as fluency may not be a requirement then.

Haddouti (1999) further progresses to point out globalization as another drive for the desire to involve in multilingual information access. In this regard, he presents the continued integration of borders and cultures like the European Union, where large markets, international business and highly developed communication technologies are becoming the basis for interaction. Given these, he points out that though the technical barriers are solved, issues of multilingualism still remain.

2.1.1 Definitions of Cross-Language Information Retrieval

To get a more critical insight into the issues of Cross Language Information Retrieval, let's look at some definitions. Also known as "Multilingual Information Retrieval" (MLIR), it represents, according to Oard (1997) "selection of useful documents from collections that may contain several languages". This definition though not clearly specifying how the behind the scene activities work, it at least states out that the document collection as well as the retrieval is in more than one language.

A more detailed definition of Cross Language Information Retrieval can be obtained from the works of Hull and Grefenstette (1996), as quoted by Reitmeyer (2004), where they give five distinct definitions as follows.

1. IR in any language other than English.
2. IR on a parallel document collection or on a multilingual document collection where the search space is restricted to the query language.
3. IR on a monolingual document collection that can be queried in multiple languages.
4. IR on a multilingual document collection, where queries can retrieve documents in multiple languages.
5. IR on multilingual documents, i.e. more than one language can be present in the individual documents.

On the other hand Cross Language Information Retrieval is further extended by definition given in Soergel (1997). Soergel defines CLIR as "the retrieval of any type of object (text, images, products, etc.) composed or indexed in one language (the target language) with query formulated in another language (the source language)". As we can see here, the retrieval is not just limited to text as the traditional sense would imply but more so to any object represented in different language.

2.1.2 Basic Approaches to CLIR

In trying to cross the language barrier, cross language retrieval approaches could take on three major directions – through query translation, document translation, or both (Oard, 1997). The Selection of approaches from the above could depend mainly on the state of searcher's language proficiency as presented earlier. In that case, the conditions that would warrant one of Oard's approaches would then be easily determined.

Though the above approaches are broader, more technical consideration of how actually the translation takes place represents the major challenge. More popular approaches frequently experimented according to evaluations presented at TREC 7 and TREC 8 include Machine translation, controlled vocabulary and dictionary-based approaches as well as latent semantic indexing (LSI) and corpora-based approaches. We will progress to take a brief look at these approaches.[3]

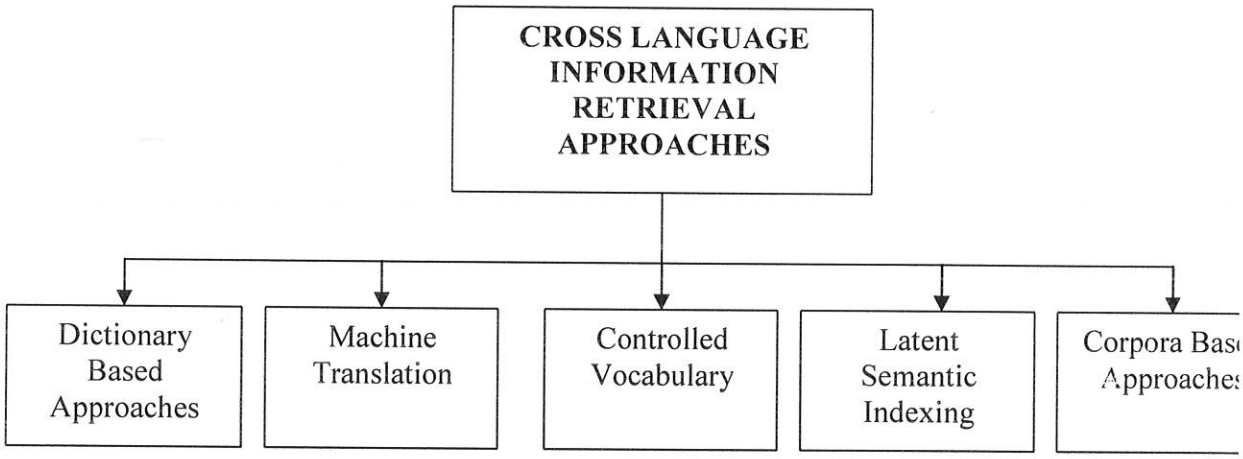


Figure 1. Cross Language Information Retrieval approaches.

2.1.2.1 Machine Translation

Most experiments take on Machine translation (MT) as an obvious choice for cross-language information retrieval systems. It also played a large role in the TREC-8 experiments of a number of groups. Machine translation can automatically translate queries or documents. However, CLIR is a difficult problem to solve on the basis of MT alone as queries that users typically enter into a retrieval system are rarely complete sentences and provide little context for sense disambiguation as noted by Braschler, Peters and Schäuble (2000).

[3] Available online http://www.trec.nist.gov

Though MT is envisaged with the idea of the query being translated from the language of the user to another language for search, and the results translated back into the user's language for viewing, other research also seem a bit skeptical in endorsing it as a solution for CLIR. Oard and Dorr (1996) continue to present this skeptism by stating that MT only produces high quality translations for specific domains, where technical terminology seems to be more dominant as semantic accuracy is not a major contending factor. This implies that MT is not that capable of addressing issues relating to semantic accuracy where issues of word sense disambiguation are predominant.

2.1.2.2 Controlled Vocabulary

The controlled vocabulary approach has long been the most dominant and effective for CLIR (Reitmeyer, 2004). Salton's initial experiments around 1970 are often quoted as the predominant ones in the use of controlled vocabulary in CLIR (Oard 1996, Reitmeyer 2004, Soergel 1997). In working with a controlled vocabulary a thesaurus is usually created with the notion of having words/descriptors compiled representing the concepts illustrated in the documents within the collection. Semantic relations would then have to be established across documents to enable concept level similarity for retrieval.

Expanding the concept of a thesaurus based retrieval in a multilingual environment the whole idea becomes translation of the entries made in the monolingual environment to the target language. Fluhr (1996) points out that a thesaurus, in addition to the translation of the thesaurus terms to each target language, addition of those terms can also be automatic, given the systems ability to learn from previous indexing which terms are likely to be important.

This addition of terms as they are found, and the availability of synonyms as well as narrow and broad level concepts can be used to facilitate post-translation after the translation has been done to a similar concept space. In this instance the expansion

of terms after the translation can be done in the same fashion queries are expanded by use of monolingual thesauri. In this case the narrow as well as synonym sets are identified and used to query the collection.

This approach has had its criticisms as well. Reitmeyer (2004) points out that the users query formulation is usually restricted to terms found in the thesaurus only. On top of this, retrieval performance provided by the system would greatly depend on the exhaustivity and specificity of representing all concepts classes involved in the area or language. To this end the number of terms used in the thesauri greatly enhance the output of the system, but then again in studies by Fluhr (1996) as well as Haddouti (1999), the amount of terms used in the thesauri are indicated as determinants for the retrieval performance. More discussion will follow as the core focus of this research is on the use of this approach for CLIR.

2.1.2.3 Dictionary-Based Approaches

This approach uses combinations of monolingual or bilingual dictionaries to provide something similar to a thesaurus (Oard and Dorr, 1996), which provides a platform for developing multilingual systems. The dictionary-based approach typically suffers from the problems of ambiguity and a limited scope, omitting technical terminology (Haddouti, 1999).

They can also be seen as extensions of applications of the multilingual thesaurus, where in this case the vocabulary will not be controlled to a set of concept representing words but extended to include all words included in the electronically available dictionary. To be noted in this case is the fact that many words don't have direct translation and at the same time disambiguation to the user's intention will be very difficult as the alternate translations refer to completely different semantics. We also have to keep in mind that the dictionaries may not in most cases contain a complete set of terms for all languages under consideration.

2.1.2.4 Latent Semantic Indexing

Another approach to CLIR is through latent semantic indexing (LSI), which makes comparisons between sets of semantically related words (Fluhr, 1996). "In LSI the principal components are thought to represent important conceptual distinctions" (Oard and Dorr, 1996, p.21). This allows for retrieval that works better with actual word concept relations. Because this approach orders documents by how closely related they are semantically and therefore clarifies which specific concept a term may represent, LSI can also help to limit or clear up ambiguity problems (Davis and Dunning, 1995).

Landauer and Littman (1991), as cited in Oard and Dorr (1996), were some of the first to work on CLIR using LSI, and their study depicts a basic approach to its use by evaluating passages from a training collection and identifying the principle components of them to make clusters of concepts. Berry and Young (1995) were able to use LSI with more success when they used more finely grained training data, such as the first paragraph of a passage from the Bible, instead of the whole passage.

2.1.2.5 Corpora-Based Approaches

According to Haddouti (1999), the corpora-based technique, "analyzes large collections of existing texts and automatically extracts the information needed." This is done by exploiting statistical information about term usage within the corpus, combined with linguistic constraints to avoid errors (Haddouti). Oard and Dorr (1996) describe this approach as a type of automatic thesaurus building where information about the relationships between terms is obtained, "from observed statistics of term usage." Lin and Chen (1996) have used this approach in their

research on machine learning and multilingual thesaurus construction. However, this technique ideally requires large collections of thousands of documents covering similar subjects to be made, and such collections are scarce.

It is important to note that different aspects of these approaches can be combined in an attempt to create superior CLIR systems. For example, the EMIR (European Multilingual Information Retrieval) project, which lasted from 1991 to 1994, combines MT and other IR methods, such as statistical models for weighting query-document intersections, as well as normalization of terms, grammatical tagging and a reformulation system aimed at disambiguation (Fluhr, 1996). Many approaches incorporate statistical and term vector translation techniques as well, mapping sets of TF*IDF term weights between languages (Oard and Dorr, 1996).

2.2 The Thesaurus: An Overview

2.2.1 Definition and natures of Thesauri

Even though the traditional sense of the thesaurus remains among most people, it has grown to become a tool more than Peter Mark Roget initially intended it for (Foskett, as quoted in Sparck Jones and Willett, 1997). During that time in 1852 Roget's thesaurus, "Thesaurus of English words and Phrases" was first published with the idea of helping out authors and writers in their literary work composition. This thesaurus was constructed to find better expressions in terms of word and phrases to better express thoughts and ideas. It has grown to become an information retrieval tool where current applications range in areas of indexing and searching.

The term thesaurus, according to Foskett as quoted in Sparck Jones and Willett (1997), comes from Greek and Latin words which mean "a treasury" where by the thesaurus has come to be understood as a treasury of word to enrich vocabulary.

More recent definitions however go beyond representing the thesaurus as a simple repository of words. Soergel (1997) describes the thesaurus as "a structure that manages the complexities of terminology in language and provides conceptual relationships, ideally through embedded classifications/ontology".

Aitchison, Gilchrist, and Bawden (2000) define the thesaurus as "a vocabulary of controlled indexing language, formally organized so that *a priori* relationships are made explicit; to be used in information retrieval systems, ranging from card catalogue to the internet". In this definition a more explicit reference is made to the retrieval dimension of thesaurus as compared to the previous definitions.

Similarly in relation to the last definition the use of thesauri in IR is seen as a combination of indexing and searching functions according to Aitchison, Gilchrist, and Bawden (2000). Here, principal applications of the thesauri are presented as a combination of the two possibilities – indexing and searching, where we would end up with:

1. Thesauri used in indexing and searching.
2. Thesauri used in indexing and not searching.
3. Thesauri used in searching and not indexing.
4. Thesauri used in neither case.

In the first instance, the thesauri used in both indexing and searching represents close mapping of the indexing and search processes, where the retrieval is usually done by the very people involved in the indexing process. This greatly simplifies the search process and makes the retrieval performance more effective as the users more or less are familiar with the kind of terms to be used since they have involved in indexing and hence know what words to use.

The indexing thesaurus however is much different according to Aitchison, Gilchrist, and Bawden (2000), where its principal value is "providing a rich set of terms, including especially synonyms and broader terms, to increase the chances of

successful retrieval". On the other hand the search thesaurus is mainly associated with providing "assistance in the searching of a free text database by suggesting additional search terms, especially synonyms and narrow terms". Here the contrast could be highlighted between the indexing and searching thesaurus, where the whole idea is towards better recall in case of the indexing thesauri and towards precision by the searching thesauri.

On the other hand what Aitchison, Gilchrist, and Bawden (2000) referred to the thesauri used in neither case, could be what Soergel (1997) referred to as thesaurus used in the "Knowledge-based support of free-text searching", whereby a conceptual framework would guide the user in formulating a better query through interactive systems of retrieval.

In a different approach Guidelines for the Construction, Format and Management of Monolingual Thesauri (ANSI/NISO Z39. 19-1993) presents four principal purposes of the thesauri. One is to provide a means to translate natural language expressions of users into a controlled vocabulary used by the system. Second is to enhance consistency in the assignment of index terms to documents. Third we have the role of the Thesauri expanded further to an application used to indicate relationships (semantic) among terms in use. And the last one is more closely in line with that of Aitchison, Gilchrist, and Bawden where it serves as a tool in aiding the retrieval process.

Though all point out thesauri in a different fashion and show its uses in their own perspectives, most details point out to same basic facts. For example, it is acknowledged that vocabulary is controlled and that applications focus on retrieval more so now than ever. Hence it is no wonder that in light of cross language retrieval issues, it is a worthy endeavor to implement thesauri for better results on top of monolingual environments.

2.2.2 Thesaurus Construction

So far construction of thesauri using the manual approach has been considered a conventional approach (Abuzir, 2002). In this regard, the main task of thesaurus construction process would fall in the hand of the domain experts who will be mainly in charge of creating the hierarchal as well as conceptual framework for the thesaurus. These experts will have to identify concepts that have to be expressed in relevant structure as well as enable lucid presentation of the thesaurus as knowledge based support.

Doing the above would enable searchers of information to use the thesaurus as the main tool to formulate search queries, which can be looked up from the thesaurus. Another approach of the thesaurus in this case would be to serve as an automatic tool for expanding queries, where by additional search terms would be easily incorporated to the searcher's initial query. Mostly however, the approach of using thesauri for automatic expansion is emphasized (Frei and Qui, 1995).

Though the results of constructing a thesaurus manually would represent very intellectual and professional presentation in both structure and content (Abuzir, 2002), the costs associated and the development and the time are major bottlenecks for this process. For this reason most thesaurus developers try and present automatic approaches towards thesaurus construction but few seem to show the extent of the performance of such developed systems (Salton (1983), Abuzir (2002), Marianne (2002), Beza-Yates, R., Robiero-Neto, B. (1999)). Major approaches of thesaurus construction can be presented briefly by figure 2. [4]

As can be seen from Figure 2 there are seven separate approaches towards thesaurus design. Looking at the manual approach, we can see that it requires an existing data source, which has to be a representative text on the area. Deciding to adopt this method sub-assumes that there exists a representative body of text that is

[4] Adapted with modification from Abuzir (2002).

readily available. As in the case of this research for example the Commercial Code of Ethiopia is assumed as a representative text. Proceeding in this fashion however doesn't necessarily imply total manual construction, as statistical as well as linguistic procedures are also employed and this could be done through machine assistance.

According to Abuzir (2002), we can also use a prearranged list of terms like table of content, index of books as well as other sources to enable identification of concept bearing words in constructing a thesaurus.

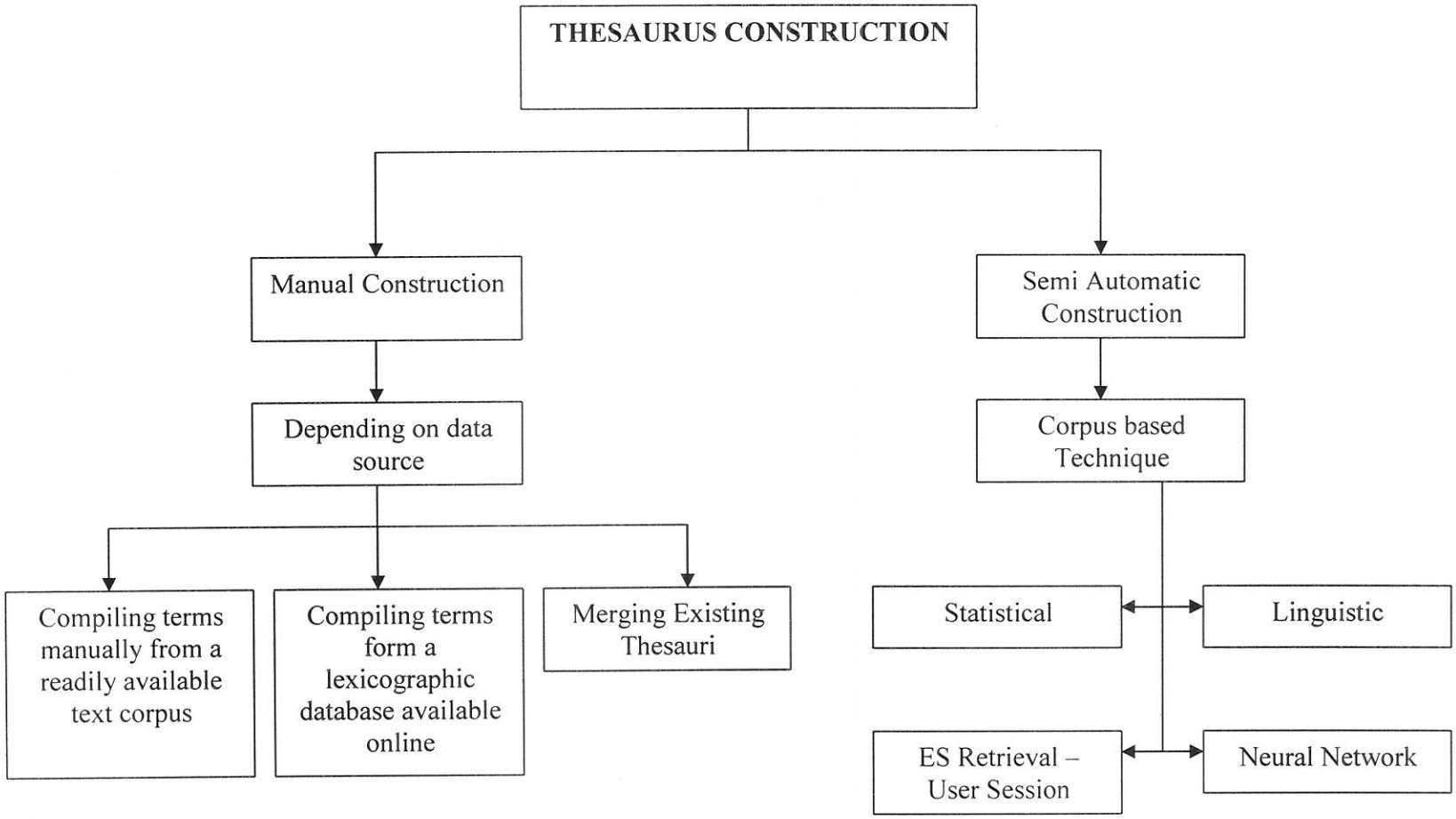


Figure 2. Thesaurus construction approaches.

Other methods in the manual construction include the use of online lexicographic databases as well as merging of existing thesauri. In the case of online lexicographic databases, we can look at subject codes or word lists that have semantic arrangements or to a similar extent any other notation that indicates concept based relationships. On the other hand merging of existing thesauri would require the existence of separate thesauri in the same area and, with minor modifications can be combined to form a thesaurus.

Automatic approaches generally fall in four different categories. The linguistic approaches according to Grefenstette (1992) and Ruge (1991) indicate the use of header-modifier relationships to determine semantic closeness. At the same time they refer to co-occurrence statistics as another approach to determine semantic closeness of words in a context. Statistical methods focus on clustering approaches where documents sharing a significant number of words are grouped together and these words are transposed into a term by document matrix, where they are defined by a similarity measure.

The remaining two approaches, expert system and neural network have a different perspective. Expert system based construction techniques tend to focus on user querying to establish relationships among terms. For example words used in a query by the Boolean operator OR would be regarded as synonyms. On the other hand according to Chen and Roussinov (1998), as quoted in Abuzir (2002), the neural network approach presents a method based on self organizing map (SOM), where a two layer unsupervised neural network is used to summarize high dimensional data.

Marianne (2002) presents the issues related to thesauri construction as falling in the headings: determination of purpose and scope, concept acquisition, formation and definition of concepts, and organization of concepts. Though each of these headings are very broad and contain elements that might overlap in some conditions they represent a good springboard to understand the bigger picture.

Soergel (1974) as quoted by Marianne (2002) states "...it is most important to see a thesaurus and its functions in the context of the whole ISAR system in which it is to be used". This would refer to understanding towards what objective the thesauri are being built, as this would put limits on the required output of the system. In light of this, a variety of approaches are evaluated further by Marianne (2002), whereby the purpose and scope of thesauri can be determined. Further stages of development would then ensue given the scope and purpose of the thesauri.

A clearer representation of thesaurus construction is presented in Soergel (1997), where he discusses issues in relation to monolingual and multilingual thesauri. There, illustrates a commonly adopted method whereby words are organized through a serious of activities into concept classes. Major activities in light of Soergel's approach include:

1. Collection of a list of words and phrases (could be from search requests, documents, free-term indexing, and other thesauri).
2. Sorting to increasingly fine-grained groups.
3. Homonym detection and disambiguation.
4. Selection of a preferred term to represent a concept.
5. Conceptual analysis at various levels.
6. Writing scope notes.

On a different note, thesaurus construction approach can be viewed from the method used in constructing the thesauri, that is, manual or automatic approaches, but still from a bit differing perspective from one discussed above. In automatic generation of a global thesaurus, co-occurrence analysis is carried out on the terms or the documents in the system to form thesaurus classes, using term weighting based on the Vector Space Model, devised by Salton, and a clustering algorithm (Sparck Jones, 1974; Salton & McGill, 1983; Lancaster, 1986; Aitchison, Gilchrist & Bawden, 2000).

The assumption behind is that the more frequently terms occur together in the documents within a particular subject field, the more likely it is that their meanings are related. The Vector Space Model determines the frequency (or weight) of a term in a document. Terms that appear frequently in a document indicate that a term is highly relevant to a document, and should be given a heavier weight. The inverse document frequency is the total number of documents in which the term appears, and terms that appear in fewer documents in the whole database (the more specific terms) are given higher weights. The combined weight of a term is then computed by multiplying term frequency by inverse term frequency Marianne (2000).

2.2.3 The Multilingual Thesauri

Given the monolingual thesaurus, which has terms form one language, the multilingual thesauri could simply be taken as one having a collection of terms in a number of languages where this collection of terms would fall in corresponding concept classes. In that line, therefore, the task of CLIR would be to map between these sets of words in identical concept classes to formulate the queries, presented in the simplest sense.

Other than that Aitchison, Gilchrist, and Bawden (2000) state that the construction of the multilingual thesauri is not that different or harder that the monolingual one given that there are linguists available. However, there comes the issue of a variety of morphology as compared to one only in the monolingual sense, hence the need for more linguists.

Soergel (1997), however, points out that there are other issues involved in the overall process of multilingual thesauri construction. The assumption of a thesauri mapping to a common set of concepts in all languages makes language matters seem easy given that there are identical classes of concepts in all languages. Though concept

class determination is a problem to both monolingual and multilingual thesauri, the mapping of classes is unique to the multilingual one only.

The issue here is that vocabularies of languages are a factor of a number of issues including the socio-cultural environment of that particular language. This would imply concepts in one language may or may not exist in the other. According to Keranen (2000), "the construction of multilingual thesauri demand operating in a multicultural environment", whereby it should be able to represent concepts lexicalized in all languages involved though all concepts are not always lexicalized in all languages. Hence, even though the logical sense would be to lexicalize and present concepts that exist in a language, the case is not true. As Soergel (1997) points out, "this is a misguided notion", implying that multilingual thesauri construction is more an issue of negotiation and arduous processing of linguistic resources to form a conceptual hierarchy that is capable of arranging terms and concepts to fit across languages.

It is essential to choose among the methods of construction in actually going about constructing the thesauri. This stage represents major portion of the overall project where the identification of the source and target thesauri would have to be done, depending on whether the approaches are going to be *Ab initio,* translation of an existing monolingual thesaurus, or reconciliation and merging of existing monolingual thesauri in two or more working languages. The source and target language though would sound as if to imply dominance of one language over the other, this should not be the case, as noted by Aitchison, Gilchrist, and Bawden (2000). That relationship simply connotes a starting point for the construction whereby one thesaurus would help as a springboard to continual feedback based concept representation through translation.

On the other hand Soergel (1997) presents a set of approaches with increasing complexities in his paper and the approach discussed above is presented in his arguments as one that suffers from a bias towards the initial language. Starting from monolingual thesaurus and translating does not capture concepts lexicalized only in

the other languages and is biased to the conceptual structure underlying the starting language. On top of that he argues that it may not produce all synonyms in the second language.

The next stage is mainly about the identification and assignment of three major types of relationships. These inter-term relationships relate to equivalence, hierarchical and associative relationships. The equivalence relationship will assign terms corresponding to the same concept with a preferred and non-preferred term relationship. However the issue of equivalence is not about just mapping the set of words or phrases, it is about doing so in a concepts basis to result in accurately formed queries. Correspondingly, some terms may not necessarily refer to the same concept, but rather to parts of it or even bigger context. Such conditions will have the hierarchical relationships assigned. On the other hand, some terms may display an overlapping reference to a concept though each term may at the same time represent another concept. Hence the associative relationship will be assigned for such terms.

2.2.4 Thesaurus Structure

Soergel (1997) tries to present the structure of the thesaurus as being made up of a two major components. A terminological structure that establishes synonym relationships and at the same time disambiguates homonyms on one hand and on the other hand a conceptual structure showing a faceted classification arranged in a hierarchical fashion. These major structural components are discussed hereafter.

2.2.4.1 The terminological Structure

The terminological structure as previously discussed represents the synonym and homonym relationships and a way of overcoming or sing such relationships in the retrieval process. Synonyms according to Aitchison, Gilchrist, and Bawden (2000)

are "terms whose meanings can be regarded as the same in the wide range of contexts, so that they are virtually interchangeable". Synonyms along with lexical variants (which represent the different word forms of an expression) as well as quasi-synonyms (near synonyms regarded as synonyms in indexing environment) are arranged in a structural format of equivalence relationship.

The equivalence relationship is "the relationship between preferred and non preferred terms where two or more terms are regarded, for indexing purposes, as referring to the same concept" Aitchison, Gilchrist, and Bawden (2000). Hence, structural relationships show in this case a UF (Use for) or Use (U) relationships to distinguish between preferred and non-preferred terms.

Another terminological relationship is defined by the hierarchical structure that presents "levels of superordination and subordination" Aitchison, Gilchrist, and Bawden (2000), representing structural relationships denoting broader concepts being a wider classes into which a variety of narrower concepts fall into. The levels of superordination and subordination will depend on the desired level of detail required and hence are not to be prescribed, however, this relationship has a great role to play in terms of presenting narrower or broader categories for indexing, searching or simple knowledge based support for either one or both.

The next terminological relationship falls in light of the associative relationship. This relationship defines the extent to which neither of the previous two relationships can define/explain the desired level of relationship existing between two or more concepts. This is to mean considering two concept bearing words, if for some reason it is found desirable to show that the words are related and if that relationship cannot be explained by the previously defined ones, it is usually left for this assumption to come up with a relationship in terms of "Related Terms".

2.2.4.2 Conceptual structure

A conceptual structure forms the basis of well designed thesauri, as the thesauri are supposed to present not just a word list but a semantic arrangement enabling knowledge support. Given that then the use of the thesauri in searching or indexing or both would definitely require well arranged structure that is able to define relationships among terms as well as dictate how they are supposed to be used. The two principles enabling a well formed conceptual structure are facet analysis and hierarchy.

A facet groups concepts that fall under the same aspect or feature in the definition of more complex concepts; it groups all concepts that can be answers to a given question. Using elemental concepts as building blocks for constructing compound concepts drastically reduces the number of concepts in the thesaurus and thus leads to conceptual economy (Soergel 1997). It uses the process of Semantic factoring or feature analysis where concepts are analyzed into their defining components (elemental concepts or features). This gives rise to a concept frame with facet slots. It also facilitates the search for general concepts, such as searching for the concept *dependence*, which occurs in the context of medicine, psychology, and social relations (Soergel 1997).

Hierarchies on the other hand enable the development of relationship among a variety of concepts in terms of part-of relationships or more specifically in whole-to-part relationships. This relationship opens the door for the representation of broader or narrower concepts in the final thesauri, where the expansion or narrowing of queries can be carried out by appropriate collapsing or expansion of the hierarchy tree. We can see that the queries will be expanded to related concepts and not just synonym sets. Therefore, through facet analysis and hierarchy building, the thesaurus developer can often discover concepts that are needed in searching or that enhance the logic of the concept hierarchy.

2.2.5 Review if existing thesauri in the CLIR environment

The approach of using multilingual thesauri is judged to achieve good results as it is judged to minimize the problem of polysemous words. However, it has major disadvantages which include difficulty in their development, deployment and maintenance. Initial experiments using thesauri for CLIR were done by Salton [Salton 1970], Gey and Jiang [Gey 1999], Ferber [Ferber 1997], and Gilarranz, Gonzalo and Verdejo [Gilarranz 1997] as presented in Soergel (1996).

The first types of resources created for cross-lingual information retrieval were multilingual information retrieval thesauri (Oard 1996). One example of such thesauri, thesaurus EuroVoc of European Community, is published on 9 languages of European Communities and nowadays used for retrieval of European documents (EUROVOC, 1995). The domain of the thesaurus is a broad domain of social relations including economic, political, military, cultural, sports and other problems, which are discussed in governmental documents, legislative acts, and newspaper articles.

Now the Thesaurus includes more than 27 thousand concepts, 64 thousands terms, and 105 thousand manually described relations. To compare, conventional information-retrieval thesauri for the same domain has the following quantitative characteristics: Legislative Indexing Vocabulary – 6800 descriptors (concepts), 9800 terms, about 15 thousand relations between descriptors (LI V, 1994), English part of EuroVoc has 5933 descriptors, about 17 thousand relations between descriptors (EuroVoc, 1995).

Another thesaurus used in the cross language environment is the Art and Architecture Thesaurus (AAT) initiated in 1979. At present, the AAT is maintained by the Getty Research Institute[5] and the thesaurus keeps growing. The AAT contains

[5] http://www.getty.edu/gri/

about 120,000 terms covering objects, textual materials, images, architecture and material culture from antiquity to the present. The AAT is based on a hierarchical structure. It comprises seven facets (categories), which are Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, and Objects.

The applicability of the AAT was tested at Rensselaer Polytechnic Institute using the Architecture Library's Slide Collection in 1985. Keefe (1990) summarized the experiences during the development of the Rensselaer Online Cataloguing System for Slides (ROCSS) Reitmeyer (2004). Keefe pointed that the cataloguers found the terminology to be very useful, especially the Styles and Periods hierarchy. The cataloguer (or indexer) needs to be trained initially for at least three months before reaching an acceptable level of competence and efficiency. However, indexing using the AAT was very time-consuming.

However such thesauri developed for manual indexing and retrieval (monolingual and multilingual) have properties which make it impossible to use them in automatic text processing of contemporary large electronic collections (Oard 1996). It is known that in monolingual information retrieval, the use of linguistic resources did not show improvements in retrieval performance that would be considerable enough to justify development costs in comparison to word-based models (Voorhees, 1999, as quoted in Reitmeyer (2004). It is argued that the unsuccessful attempts to enhance performance of information-retrieval systems with help of thesauri mean only that thesauri, to be used in automatic conceptual indexing and retrieval, have to be specially constructed, have specific features, and for their effective use it is necessary to develop special techniques of text processing.

CHAPTER THREE
THESAURUS CONSTRUCTION

This chapter will progress to give a detailed picture of the thesauri construction process. Major focus will be given to the construction of the English thesauri as it represents most of the arduous tasks starting from identification of facets through final development of the thesaurus.

3.1 Construction of the English thesaurus

Initially, the existence of thesauri in the field of commercial law for Ethiopia was sought for, or any other similar thesauri, but none can be found. Main reasons for this included the non-existence of such thesauri locally and also the extremely high cost of those available like the EUROVOC. At the same time a completely adaptable thesauri would be rare to find in this case, so no advances were made towards finding one to fit the purpose as none were available.

Given a domain area of commercial law, the next process was in proceeding to find a representative body of text where by a list of complete set of concepts can be identified. Though compiling a set of terms from book indexes and content lists presented a good starting point, it was obvious that concepts would not be exclusively presented in equal dimensions in these sources. Nor a standard index was able to be found, and therefore the representativeness of these lists of terms was not acceptable to the researcher. Hence, the need to identify all conceptual aspects of the domain was imperative and the available Commercial Code of Ethiopia (1960) was selected. The main reason this material was selected is because it is believed to contain all the possible scenarios and business related legal issues as all cases and conditions are judged according to what is written in this Code.

3.1.1 Faceted Classification

Given a representative body of text, the first major task in going about the construction of the English monolingual thesauri was distinguishing the structural relationships between the terms contained in the Commercial Code of Ethiopia. To do so it was essential to discover the Macro level relationship between terms as they relate to the overall structure of the subject field. The process of classification to develop the macro level relationship was done using facet analysis.

This facet analysis is an intellectual operation that required the involvement of domain experts in order to classify the commercial code into non-overlapping subject isolates in the overall thesaurus structure. It required a thorough analysis of the contents of the Commercial Code and finally resulted in the identification of eight major facets that enabled the classification of ontology to proceed. In the analysis of the subject field into facets four domain experts, including the researcher, were involved in order to determine the facet classes of the thesauri. The facets are enumerated below.

Facet 1. Traders and Business
Facet 2. Business Organizations
Facet 3. Carriage
Facet 4. Banking
Facet 5. Insurance
Facet 6. Negotiable Instruments
Facet 7. Contracts
Facet 8. Sales

3.1.2 Thesaurus structure determination

Following the identification of facet structure, it was essential to determine the kind of structure that the thesauri would have. The options included a flat generic structure, hierarchical or graphic structures. A flat generic structure would not be able to integrate the desired level of relationships as the broader term and narrow term relationships will not be displayed. On the other hand, a graphic display is desired if the user is interacting with the display which in this case is not so. At the same time, since the major concern is the conceptual structure for mapping internally within the system, and not an alphabetical display for knowledge based support, focus is not intended on generating an alphabetical list of concepts, but rather on a hierarchical classification. Hence, the hierarchical structure is chosen as it would enable to cross reference concepts more clearly and tangibly in the case of mapping from one thesaurus to another.

Business organization	የንግድ ድርጅት
NT General Partnership	**NT** ተራ ሽርክና
Limited Partnership	ሁለት አይነት ሀላፊነት ሽርክና
Sole proprietorship	ሀላፊነቱ ተወሰነ ግል ማህበር

Example: Broad - Narrow term relationships

Limited Partnership	ሁለት አይነት ሀላፊነት ሽርክና
BT Partnership	**BT** ሽርክና
Business Organization	የንግድ ድርጅት

Example: Narrow - Broad term relationships

3.1.3 Collection of Terms and Phrases

In going about collecting the terms for the initial thesaurus, a combination of approaches were used whereby the empirical approach, or more precisely the deductive approach, was supplemented by machine assistance. This acquisition of terms and phrases for the thesauri was also done by a combination of two approaches from the ones suggested in Soergel (1997). These approaches are collection of words and phrases from documents themselves through manual means and use of free- term semi-automatic indexing by using an indexer developed for this purpose using Visual Basic 6.

Through the deductive approach terms were extracted by domain experts, where attempts were made to identify single-word terms or compound terms representing the broadest classes on the basis of logical relationships. Here, the major idea is to enable the establishment of hierarchies so that broader and narrower concept relationships and mapping can later be implemented with least effort.

Parallel to this, candidate terms were identified automatically from machine readable text (the Commercial Code of Ethiopia) available in .pdf format through an indexing procedure. This procedure was done according to the facet structure that was initially adopted. Therefore each book in the commercial code was indexed separately as the assumption was that each book had been identified as a subject isolate and hence it would be more logical if the indexing was done separately in the hopes of identifying concept bearing words in each book/facet.

The list of terms developed in this case was identified through generating a frequency distribution after they were sorted. The total set of terms was reduced by a stop list.

Following the above complementary processes, the terms extracted through the deductive approach by the domain experts and through machine assistance were

matched. This enabled to identify some terms that were overlooked in the deductive approach, and made possible the inclusion of previously unrecorded terms into the thesaurus. The reverse was also true as phrasal extractions (compound words) were not possible through the concept indexing process simply by using the indexer, and therefore manual involvement of the domain experts enabled a great complement in identifying most concepts, as some concepts were rather not represented though the use of single words.

The above two major activities resulted in the creation of a term bank, where individual terms were treated separately as descriptors of major concepts. These terms in the bank were then validated for concept representation by comparing their validity in relation to hierarchical capabilities to represent concepts. Again this was done through heavy involvement of domain experts.

Hence the faceted classification, in its analytico-synthetic classification was able to present a combination of terms to represent topics that were specifically enumerated in the code. And at the same time, class marks were established as they represented simple concepts, which later enabled organization into clearly defined categories of candidate terms during a rigorous process of further classification and assignment.

3.2 The Multilingual thesauri

As previously discussed the problems associated for the construction of multilingual thesauri are no worse, in kind, than those of the monolingual thesaurus construction (Aitchison, Gilchrist, and Bawden 2000). According to the methodology of this research paper, the multilingual thesauri are to be arrived at by direct translation of the monolingual English equivalent. However, the direct translation is not simply done by substituting equivalent meanings from intuition only. Nor are the approaches of handling the morphology as simple as the English equivalent.

Given the source language thesauri in English and the target language in Amharic (no implications of status are implied in this source-target language reference), the translation process started again having as a basis, the Commercial Code of Ethiopia. The best part of this body of text is that it perfectly represents an aligned corpus that is cross referenced by numbered system for the articles. As the area is tremendously sensitive to translation differences, the corresponding entries in the articles have been made in accordance with direct and impartial translation.

This parallel nature of the corpus presented a delightful opportunity to use it for the translation process. However, to clarify concepts and find their corresponding classes in terms of concept legal dictionaries were used. The first process was to establish the equivalence relationships among the concept classes of the thesauri. These equivalences usually come in three forms; Exact equivalence, Inexact equivalence, partial equivalence (Aitchison, Gilchrist, and Bawden 2000).

The first equivalence represents the true synonym of the concept representation across the languages. And in our construction, almost all the relationships have been identified on the basis of exact equivalence. No effort was made to split concepts or to create additional concept terms as the major rational of having a perfectly parallel corpus in such a sensitive ground would imply initial care in making the concepts match in the first instance.

Facet 1. Trader and Business	ነጋዴና ንግድ
Facet 2. Business Organization	የንግድ ድርጅት
Facet 3. Carriage	ማጓጓዝ
Facet 4. Banking	የባንክ ስራ
Facet 5. Insurance	የኢንሹራንስ ስራ
Facet 6. Negotiable Instrument	ተላላፊ የገንዘብ ሰነድ
Facet 7. Contracts	ዉሎች
Facet 8. Sales	ሽያጭ

3.3 The Database model for the Thesauri

There are quite a variety of database models that can be used to store the terms of the thesauri. These models may depend on the nature of the thesaurus, or the purpose to which it is intended for. It is also possible to adopt ones own structure so as to suit the required purpose. Soergel (1995) points out that there are basically two models that are used to store terms and relationships. They are the term based model and the Concept based model.

In the first case, the term based model stores all relationships between terms and is regarded as flexible but redundant. This is because even though all entries will be made accordingly as having terms for entry points, redundancies will occur as there will be cross relationships among terms that have already been presented. In the second case, the concept-based model on the other hand identifies concepts through the application of concept numbers and these numbers will be used to represent concept relationships.

For this research we will progress to employ a combination of the two approaches indicated by Soergel (1995). While we will maintain term based (in this case a term could also mean a compound term) entry to the thesaurus where all the entries will be considered individual, we will also adopt a concept based approach where each term will be designated with a number that will enable the cross referencing of the thesauri. Each term will act as a record, identified by a concept number.

In each record we will find all the descriptors associated with that term as designating a similar concept space. In this regard, all records will have specific fields that will be used to store all the related terms, broader terms as well as narrow terms. This will enable easy cross referencing for the concept based translation as well as query expansion to the similar concept space as all the records contain fields describing that particular term's concept space.

The thesauri database contains:

- A concept number (Field 1) designating a special entry of a term representing a concept that is unique. In this case, narrow and broad terms are considered as unique terms. Even though they represent the same concept space (or what we previously referred to a facet) the specificity of reference is different and hence will be regarded as different terms.
- A term (compound term) entry (Field 2)
- Broader term (Fields 3 – 6) entries going to three levels of concept space.

Broader term entries are evaluated on the basis of generic relationships and, partitive relationships. No term instance relationships are used in this case as term instance relationships would actually refer to existing examples than concepts.

Generic Relationships

Business organization	የንግድ ድርጅት
NT General Partnership	**NT** ተራ ሽርክና
Limited Partnership	ሀላፊነቱ ተወሰነ ግል
Sole proprietorship	ሁለት አይነት ሀላፊነት ሽርክና

Partitive Relationships

Insurance Policy	የኢንሹራንስ ፖሊሲ
NT Insurance Contract	**NT** የኢንሹራንስ ዋስትና
Premium	የኢንሹራንስ መግቢያ ዋጋ
Compensation	ካሳ

Term Instance Relationships

Passenger ships	የመንገደኛ መርከብ
NT Titanic	**NT** ታይታኒክ

- Narrow term (Fields 6 – 25) entries accommodating up to 20 levels of narrow concepts. One fact to note in this case is that even though 20 levels of narrow terms are contained, it doesn't refer to 20 hierarchical levels are included. It simply could be 20 disjoint and sometimes overlapping sets of terms could be used in indicating different specific concepts in a single concept space.

- Related term (Fields 26- 28) entries are generally used here to include synonym sets of words that are to be used in expanding the query terms. They also help in presenting preferred term entries in search as such terms are normally not preferred ones for search, but still maintain better performance for the system as they extend retrieval to documents that used terminology that is not so often used.

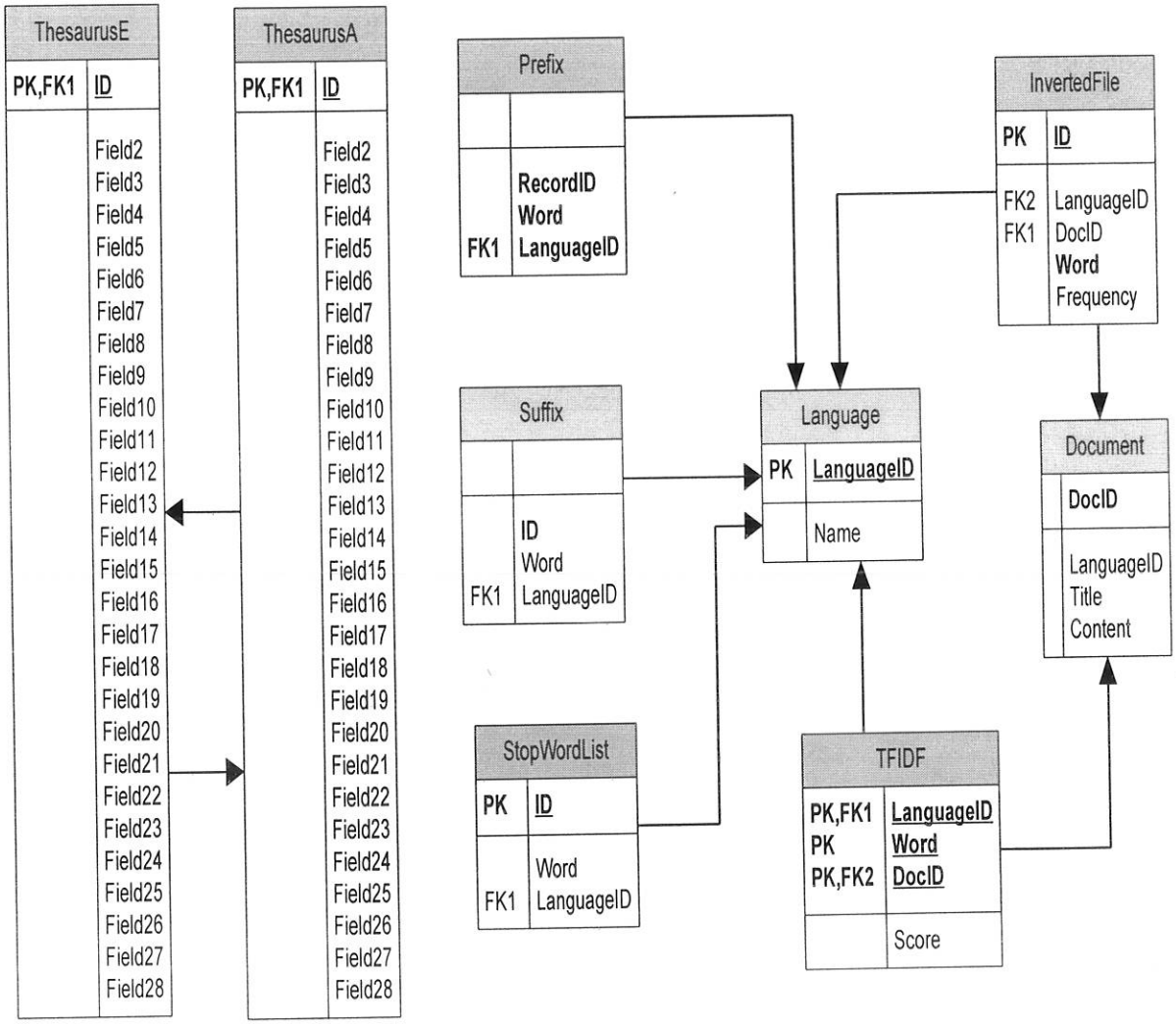


Figure 3: Database model of the system generated through Visio Enterprise 2000.

3.4 Thesauri Rules

Certain rules had to be developed or adopted with regards to the thesauri entries to be made for this paper to enable a more uniform entry mode and easier understanding of thesauri terms. To do so the following rules were developed and used to make the design process more focused and uniform, thereby enable better performance.

1. Grammatical form: Thesauri terms were selected on the basis of being either one of the following; Nouns, noun phrases, gerunds. Since one of the basic rules, according to Aitchison, Gilchrist, and Bawden (2000), is that thesauri terms are usually preferred to be nouns, the same was done here.
2. Singular and Plural form: Singular entries were preferred for both sets of languages.
3. Spelling: Spelling did not pose a problem for Amharic, but for the English thesauri, the American standard spellings were adopted for ambiguous words. However, in some cases there are words spoken and used a bit differently. In such cases, the corpus was used to determine word forms to be used in this case.
4. Punctuation and hyphens: Punctuation were removed totally with the exception of hyphens as they are believed to represent terms that when combined with its use represent a different meaning than otherwise.
5. Use of direct entry: All entries in terms of compound entries are direct entries. That is for example, instead of presenting an indirect entry

   Organization, Business ድርጅት ፣ የንግድ

   A direct entry was made with:

   Business Organization. የንግድ ድርጅት

6. Use of compound terms: Since the presentation of the thesauri focuses on concept based mapping, the use of compound words was inevitable as concepts are rarely expressed in single terms.

CHAPTER FOUR
EXPERIMENTATION

Cross-language Retrieval

The cross language retrieval tasks come in two fold runs here. One is the English - Amharic retrieval and the other one is the Amharic-English retrieval. To accomplish this, the architecture of the system is designed in such a way as to have the following major components.

1. Document and query indexing
2. Thesaurus lookup for concept identification, translation of query and disambiguation.
3. Post-translation query expansion
4. Document retrieval.

These four components represent the major system parts and within them contain major functions that need to be implemented.

1. Document and Query Indexing

Both Amharic and English documents need to be indexed in for retrieval purposes. Initially, three major steps will be involved for this process, and they are tokenization, frequency determination, filtering (stop word removal for English) after which term extraction takes place.

First, however the procedure of lexical analysis was required in determining what would actually constitute words for the indexing process. In this regard punctuation marks were removed. Though these punctuation marks enabled better

understanding of sentences, for machine assisted indexing they bore no importance and were hence indiscriminately taken out.

The punctuation used in the two sets of languages differed greatly, and hence, it was essential to adopt a list of punctuation that was in standard usage for analysis. For this purpose all punctuation included in Unicode listing were identified and included for both sets of languages. On the other hand however, the punctuation "–"(hyphen) was purposely excluded as hyphenated words are generally regarded as single word entries and will be so in the index as well.

In the tokenization step, each word (or token) in the text is then identified by recognizing the delimiter separating it from another word - in this case space. Second, a list of frequency distribution was developed in case of the documents, and these documents were indexed against a subjectively determined threshold after reviewing the output. In the filtering step, a list of stop words was used to remove non-semantic-bearing expressions from English documents and queries. Since the Amharic collection represents a much smaller collection, the Amharic documents will be indexed fully. For the English language elimination of stop words, a stop list containing high frequency function words already developed and commonly used one from Rejisbergen (1993) is used.

The indexing will be word based indexing for both languages. Because word-based indexing did not capture phrases during our general indexing process for both sets of languages, we can expect some problems in relation to retrieval performance. However, thesauri entries are expected to significantly reduce this problem as single word indexes are jointly put together to describe a concept.

On the other hand, the query indexing procedure also followed the same approach. Although the indexed query did not require determination of frequency of terms, nor the generation of inverted file or vector space as the documents, it more or less adopted the same technique. After queries are submitted the process of tokenization

is applied, where by space was considered as a delimiter. Stop words (for English language) are then removed similar fashion. This finally resulted in a set of terms that form a reformulated query is used later for searching or concept translation through thesauri look-up.

2. Thesaurus lookup for concept identification, translation of query and disambiguation

Research in Cross-language information retrieval show that the performance of the any system, in terms of both recall and precision, suffers from the fact that there is no one-to-one correspondence between words in different languages, (Milestead, 1998). Recall would suffer in this circumstance because word based translations as in the case of MT systems for query translation may choose a wrong translation that does not occur in the target language document, and may even at times not find the translation. At the same time "precision would suffer if a translation of a query term is chosen that corresponds to an unintended reading of the query term, and/or if the translation has additional unintended readings" (Zhou, Qin, Michae, Chau, Hsinchun, Chen, 2004).

But given context where concepts can be mapped instead of direct word translations, the chances of reducing ambiguity could be enhanced. Therefore the thesauri in this case are mainly organized according to parallel concepts and hence will greatly enable better mapping of concepts as compared to word based translations. This is expected to be reflected in performance gains of the retrieval process though no MT system can be used for comparison, as no complete or prototype level systems were available for the languages under consideration.

On top of the above condition stated for disambiguation, the fact that the thesauri to be used here are designed for the purpose of legal information retrieval, there is to a large extent some level of disambiguation that occurs naturally. This is mainly

because of the fact that other sense of the words (if any) to be considered will automatically be ruled out by the user as the primary idea to refer to concepts represented in the legal environment. In addition to that, the fact that intensive trials have been implemented to minimize the use of a word from appearing in multiple facets or concept classes will enable more focused interpretation of the term under consideration.

Hence the translation component of the thesaurus on concept basis will still follow a word based approach; however, the words in this case are representative of one or more (rarely if at all) concept classes. The translation will proceed by finding a match for the set of query terms in the corresponding thesauri facet entries, then to sub-facets, then to more specific entries of narrow terms and related terms. This progress of identifying the thesaurus class is progressive in that the idea is to proceed from a possibly broad category to more refined and particular querying.

E.g. Given the following query, "**Business Organization**", it is found that this is the 23rd entry in the English thesaurus, the 23rd entry in the in the parallel Amharic thesaurus is looked up and is found to be "የንግድ ድርጅት", hence successfully mapping to the desired concept class.

3. Post-translation Query Expansion

The above process ends after having identified the concept class(es) in which the query terms of one language end up in. The next process will simply be to map to the corresponding thesauri entry in the other language by simple cross referencing of the parallel thesauri. But this only goes as far as identifying the possible concept class(es) for translation purposes. But as we all know the uses of thesauri don't end here and at the same time since we are mapping concepts than mere word based

translations, we need to expand the query to broader and/or narrower facet entries. This instance prompts two mutually exclusive scenarios.

In the first scenario, we might end up with the concept mapping ending up on the broad facet class or sub-classes, each having narrow entries. In this case the query expansion task will be mainly focused to expand the original indexed query and the translated version into sub-facets and narrow entries in case of mapping falling on to the main facet class. In conditions where the concept mapping falls on a sub-facet, the query expansion will be towards the broader facet class and at the same time to narrow entries with in that facet class.

On the other hand, there is a chance that the mapping of the query may fall on more specific entries of the thesauri. In this case, there is will be no chance to expand the query to more specific entries as no other specific entries exist. Therefore, the chances of query expansion will be, in such circumstances, bound to concept classes of broader concepts.

However, this does not mean that the concepts in equal/parallel entries within a concept class will be included along with the specific query. Here, by mapping to a broader class and not other equivalent narrow entries, the attempt is to include retrieval of documents that deal with broad discussions of the subject and not documents specifically relating to the equivalent narrow entries.
E.g. Given the above query, **"Business Organization"** and the eventual mapping into the equivalent concept seen to be successful in mapping to "የንግድ ድርጅት", the post-translation query expansion will retrieve terms in the record with the concept "የንግድ ድርጅት", which in this case include:

- ተራ ሽርክና
- ሽርክና
- አክሲዎን ማህበር
- ሀላፊነቱ ተወሰነ ግል ማህበር

4. Document Retrieval

The last major component of the system will be geared towards retrieval of relevant documents using the retrieval system developed in-house. After the target query has been built in both languages, it will be passed to the search module of both languages. The Document Retrieval component is responsible for taking the query in the target language and retrieving the relevant documents from the text collection. The search primarily proceeded in the following fashion:

1. Documents are submitted to the system from which an index is developed. Similarly, queries are submitted to the system, and processed. In this case, queries for the cross language retrieval are mapped and translated to the equivalent concept class identified through thesauri lookup.

2. After the documents and queries have been preprocessed, the search will find matches in the inverted file to retrieve relevant documents.

3. These documents are then ranked accordingly by use of a weighting mechanism where the weights are determined by respective Tf*idf weights of the terms used in queries.

   $$\textbf{Weight } (d_i , _J) = Tfi_j * (\log {}^{N} - \log^{n}) + 1$$

   Where:

   Tfi $_j$ *is frequency of term j in document i*
   N is the number of document in the collection and
   n *is the number of documents containing the term.*

4. Then RSV (Retrieval Status Value) is computed by summing the respective weights of the terms given in the document to arrive at a ranked order of documents.[6]

$$\sum_{k=1}^{t} term_{ik}$$

where:

I represents the i^{th} document

k represents the k^{th} term

term represents the weight of the term

[6] As given in Rejisbergen (1993)

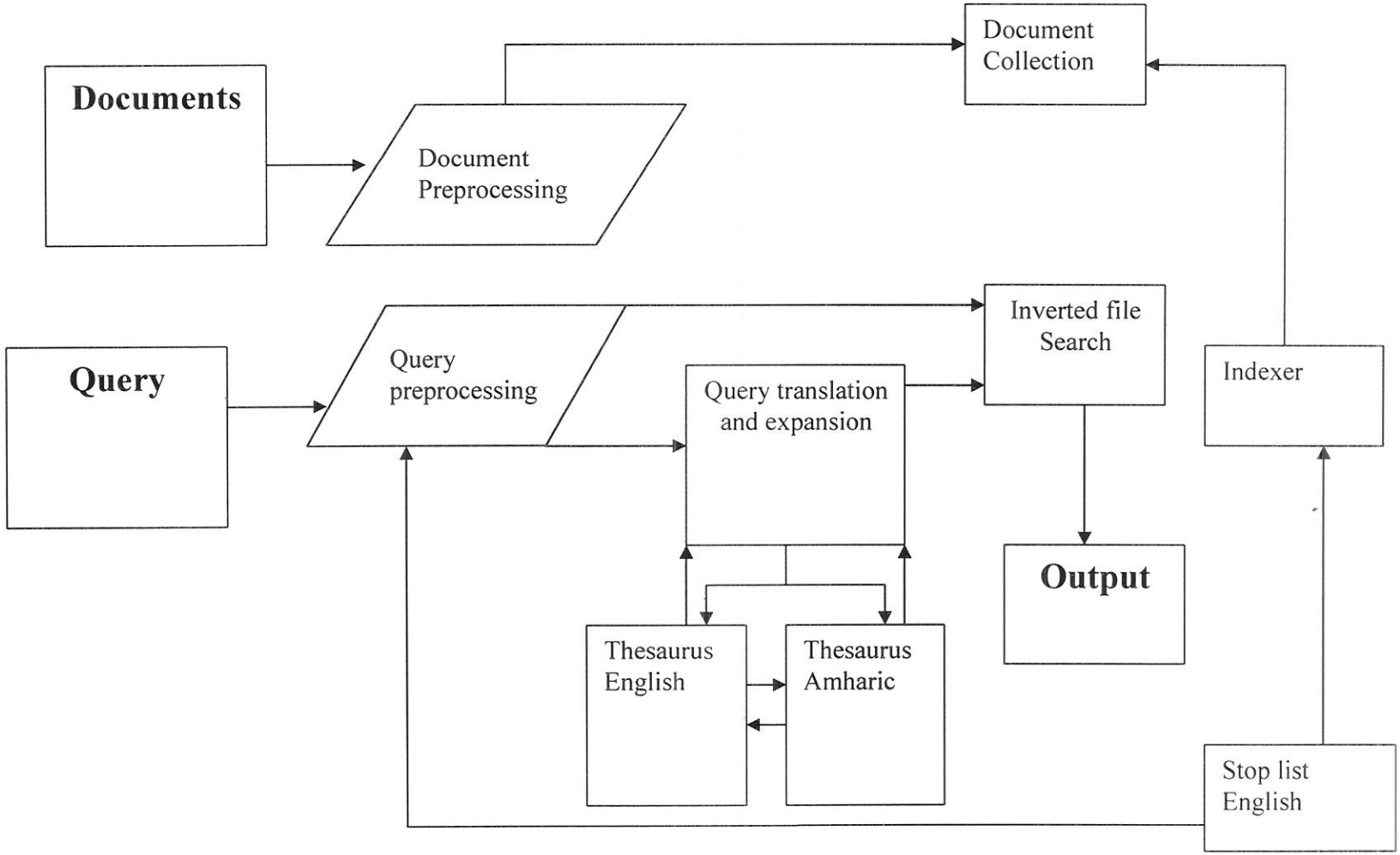


Figure 4. Proposed Architecture of the system.

4. Evaluation of Results

CLIR evaluation in our case will aim to test the effectiveness of the retrieval process by precision and recall. In this section, we will present two sets of runs whereby we will have a monolingual run without thesauri lookup and a cross-language run with thesauri lookup and expansion. The document collection is represented by close to 350 documents in both languages. These collections mainly come from the Journal of Ethiopian Law and other researches carried out by the Supreme Federal Court of Ethiopia.

We also have the retrieval outputs for monolingual experiments using the thesauri for query expansion only. Then a comparison of the retrieval performances using the two sets of languages querying an identical collection is presented by recall-precision graphs.

Test Results and Discussion

To determine the performance of the developed system, it was essential to have queries that have been judged for relevance. In this regard, three lawyers and two federal court judges were given the queries and sets of documents. The relevance judgments obtained from these experts[7] were then used to determine the performance of the system through measuring by recall and precision.

The queries used for the experimentation are listed in the annexed table along with the document numbers (DocID) that were used to identify the documents in the text collection. The results of four major experiments are presented below. The first two experiments focused on monolingual retrieval of documents without the application of the thesauri. The last two experiments were cross language experiments that used the thesaurus for translation as well as query expansion.

After the above data was obtained it was used to determine the recall and precision of all runs made. The following formulae were used to determine the respective performance levels.

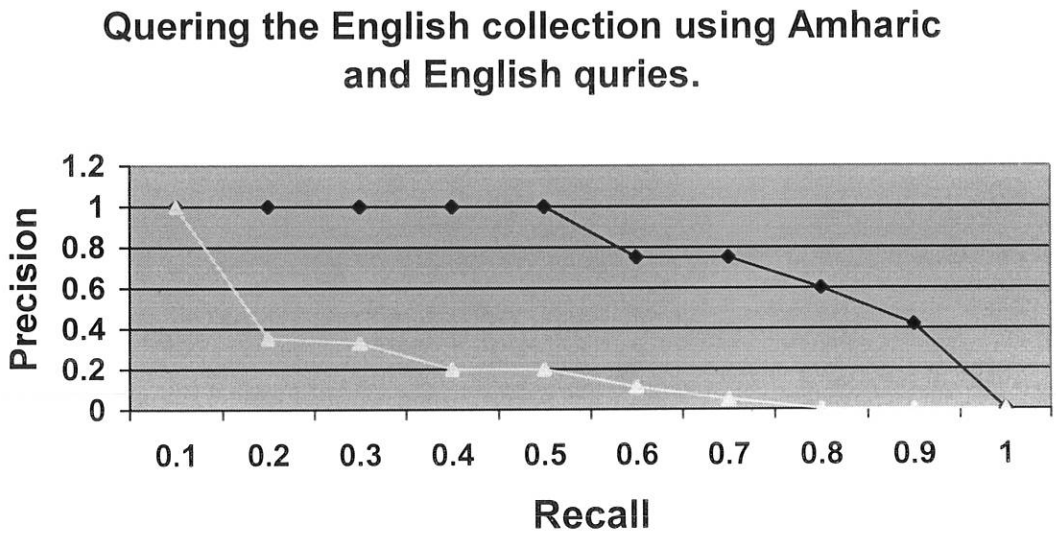
$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of total relevant documents in the collection}}$$

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}}$$

Then comes the performance as evaluated by precision and recall. Similarly the Keen (1972) smoothing algorithm, adopted by Saba (2001) is used to present the

[7] The relevance judgments are given at the end of the thesis as Annex 1.

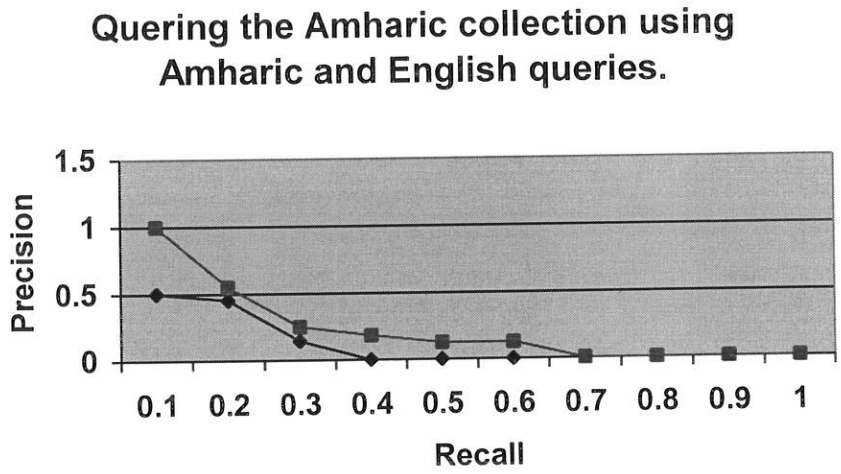
graph shown the recall and precision of the retrieval outputs. The recall-precision graphs smoothed for each retrieval run are presented below.



Upper bound line = English queries used on English collection.

Lower bound line = Amharic queries used on English collection.

Figure 5: Recall-precision graph for querying the English collection.

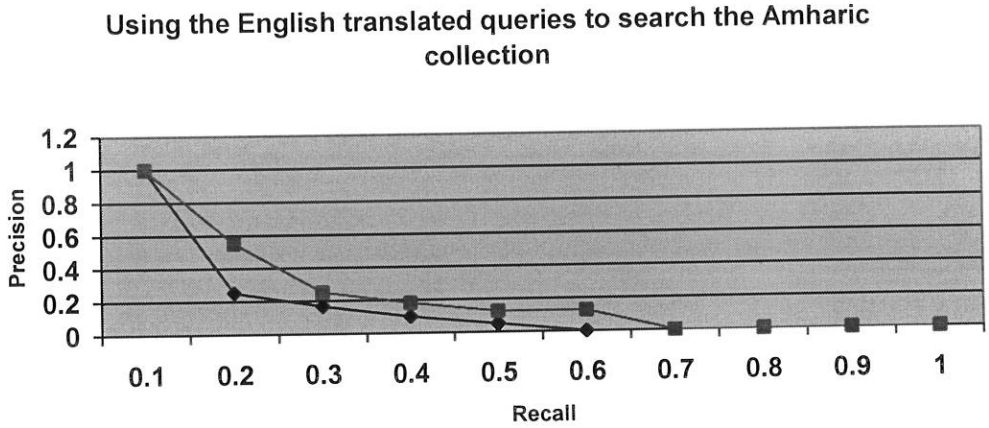


Upper bound line = Amharic queries used on Amharic collection.

Lower bound line = English queries used on Amharic collection.

Figure 6: Recall-Precision graph for querying the Amharic collection.

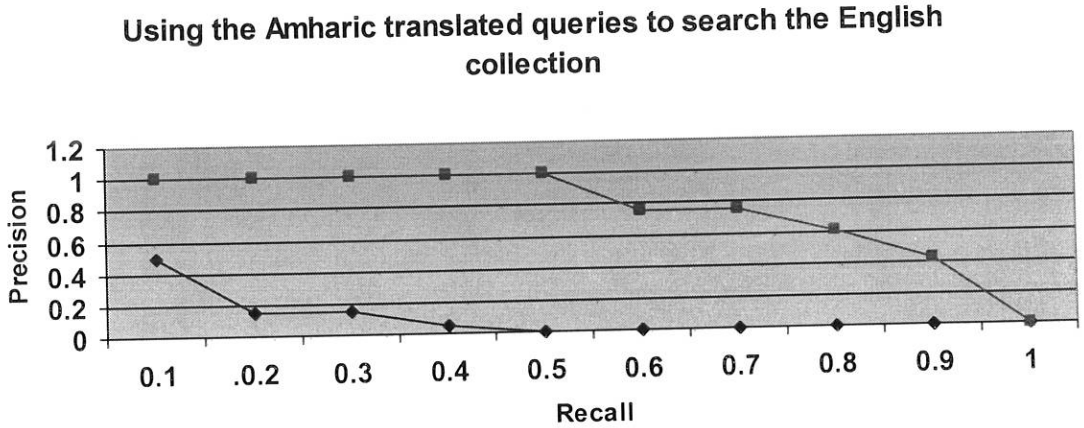
However, the above results did not provide a means to compare the actual capabilities of the thesauri as a query translation tool. Hence, additional experimentation was undertaken to clearly enable visualization of translation capabilities of the thesauri. To do so, the existing queries were translated by the initial developers themselves to yield what can be considered parallel queries. These queries were used to retrieve documents in the collection. The results are as follows.



Upper bound line = Amharic queries used on Amharic collection.

Lower bound line = English (translated versions) queries used on Amharic collection

Figure 7: Recall-Precision graph for querying the Amharic collection II



Upper bound line = English queries used on English collection.

Lower bound line = Amharic (translated versions) queries used on English collection

Figure 8: Recall-Precision graph for querying the Amharic collection II

The above illustration shows the recall and precision figures of the outputs obtained. As we can see, the graphs don't really show a pattern exhibiting high performance of the system. This is mainly because of a number of factors that influenced the retrieval process.

In the first instance (Figure 5) if we look at the recall precision graph for the English monolingual retrieval, we see a relatively better performance there. The main reason for this is the fact that there is no thesauri matching and the retrieval was done against a freely indexed environment.

In the same graph (Figure 5), we can see how the cross language retrieval performance turned out where the retrieval was using Amharic queries to search and retrieve from an English collection. It is relatively better as compared to English querying used to search Amharic collection. This is probably mainly due to the fact that the English collection is at least twice as large as the Amharic collection.

On the other hand, looking at querying using English queries to retrieve Amharic documents (Figure 6), we see that the results are very much below the monolingual performance standards. This is mainly because the document size obtained for the Amharic database is very much limited (101 research abstracts). On top of this, the translation efficiency of the system may not have performed that well when it came to matching query terms to thesaurus entries.

Looking at the monolingual Amharic retrieval we see a performance level similar to the English query-Amharic document retrieval, though a little bit better. Due to similar reasons the retrieval is as shown in the graph.

On the other hand the retrieval experimentation using the translated queries is as shown by the figures 7 and 8. In Figure 7 we can see a relatively closer mapping and translation among the two sets of languages by using the thesauri. This comparative output and initially high levels of precision indicate direct concept level matches.

Eventual decline in the retrieval performance is however mainly due lack of matches and incorrect mapping of the thesauri that results in retrieval of lesser relevant documents.

The same can be seen from the experiment shown on Figure 8. Here however we see a relatively higher gap in the performance levels shown by the thesauri. The same factors as in Figure 7 come into play only that they seem to be on a larger scale.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

6. Conclusion

Cross language Information retrieval presents opportunities to get access to documents in a variety of languages and make them available to users. In relation to that, any approach that can contribute to that effort is welcome. The thesaurus is such a tool that has the potentials to make such possibilities come true. Though it requires more study even into areas as far as culture and environment, it has created possibilities to overcome the language barrier.

The research has successfully managed to develop prototype thesauri for the legal domain in Ethiopia and similarly designed a test-bed to be used in retrieval evaluation. This study presented experiments carried out in a multilingual environment using thesauri developed for this purpose. Legal abstracts in both Amharic and English were used. However, the document collections were not large enough to properly enhance the performance of the retrieval as they had to be compiled for this research alone. The lack of research publications, to be used as test collections, especially in Amharic made the problem have a larger impact in evaluating the results.

The monolingual retrieval was not the major objective of this research but has somehow made it possible to see at least how far the multilingual retrieval could improve. The use of thesauri to translate and automatically expand the queries, at times resulted in increased recall, thus the concept-based expansion was seen to be fruitful.

In other instances, however, the thesauri showed that they needed careful redesign in that their application retrieved a set of documents that really had no significance to the query. In this regard, further disambiguation needs to be done in making sure

that concept classes remain totally isolate. However, on some conditions exact concept-based matches were seen with queries submitted. This made retrieval encouraging as the retrieved outputs generally were judged to be relevant.

Broadly speaking, this research - being the first done in this area involving Amharic language through application of the thesauri - has shed some light into the overall picture of cross language information retrieval. Its results may not be as robust as other research findings elsewhere, like the AAT and EUROVOC, but it is the belief of the researcher that this research has shown a progressive aspect to the development of such tools in our local context.

7. Recommendations

The results of the research could definitely be made better. The cross language retrieval system can be designed to incorporate a more efficient system in terms of the number of thesauri entries and morphological analyzers. Although each of the areas covered requires individual research, this research would like to point out ways forward to make such a system more robust. They include:

1. Morphological analyzers: Morphological analyzers can enhance the performance of the system. A light stemmer developed for suffix striping can greatly reduce word variations and thus enable greater matches and better retrieval. This can be seen from the results where direct matches of words managed to bring about higher precisions but variations were generally not detected and thus resulted in lower recall.

   Another important tool in this case would be a part-of–speech tagger. If a fully functional tagger is developed, query processing can be made a lot easier to identify concept-bearing terms likely to be used as thesauri entries. That would make constructing the thesaurus even a lot easier and enhance automatic approaches towards thesaurus construction as well. Therefore problems of identifying thesauri terms for search would be greatly enhanced.

2. Fuzzy matching for the thesaurus: If fuzzy matching techniques can be adopted for the thesauri, similar strings could easily be identified, though at times it may match against relatively unimportant terms that are not even in the thesauri. But it definitely is worth experimenting to enhance retrieval performance.

3. Indexing specificity of the thesauri: The thesauri developed for this research had to be indexed rather shallowly because of the lack of time and resources. This has also affected precision and recall by not retrieving documents that focused on more specific concepts not indexed by the thesauri. A more specific indexing would enable the inclusion of almost all concepts in the domain. This would in turn enable queries to be matched against more specific entries and hence, retrieval precision would have performance gains as compared to recall.

4. The availability of resources for the development of the thesaurus and search system is also another factor. In this case if other search systems (test-beds) that have been fully developed were used, more effort would have gone into thesauri construction. In a similar fashion the availability of thesauri construction experts to review the work would have enhanced the performance of the thesauri based retrieval.

5. Need to develop a representative corpus: Probably the most notable drawback for researches that involve Amharic retrieval is the lack of representative corpus that can be used for indexing as well as testing. The researcher, for example, was able to learn that there is only one research journal in law that is also sporadic in being published, even at times not coming out for years in a row. Similarly, most resources have not been converted to electronic format, or those that are, are not made accessible.

REFERENCES

Abdelali A., Cowie J., Farwell D., and William Ogden (2000). UCLIR: a multilingual information retrieval tool". Available online at:
www.nlp.uned.es/ia-mlia/iberamia2002/papers/mlia10.pdf

Abuzir, Y. (2002) "Deriving concepts hierarchy" Dissertation Work Alquds Open University, Jerusalem. Available online at:
http://www.cs.bham.ac.uk/~mgl/cluk/papers/abuzir.pdf

Abuzir, Y. (2002) "ThesConv: A tool for thesaurus construction from a prearranged list." A paper presented at the proceedings of the 2002 International Knowledge Engineering (IKE'02). Available online at:
http://www.cs.bham.ac.uk/~mgl/cluk/papers/abuzir.pdf

Aitchison, J. Gilchrist, A., Bawden, D. "Thesaurus Construction: A practical Manual 4th Ed. ASLIB. 2000.

ANSI/NISO Standard Z39.19 – 1993: Guidelines for the Construction, Format and Management of Monolingual Thesauri; American National Standards Institute; 1993.Available online at:
http://www.niso.org/obtains.html

Ballestros L., Sanderson M. (2000). Addressing the lack of direct translation resources for cross-language retrieval". Available online at:
http://www.dis.shef.ac.uk/mark/cv/publications/papers/my_papers/CIKM03.pdf

Baye Y. *የአማሪኛ ሰዋሰው::* አዲስ አበባ ዩኒቨርስቲ:አዲስ አበባ 1987EC

Beza-Yates, R., Robiero-Neto, B., "Modern Information Retrieval" Addison Wesley, 1999.

Brashler, M., Krause, J., Peters, C., and Schauble, P. (1999). Cross Language Information retrieval (CLIR) Track overview. In Proceedings of the Seventh Text REtrieval Conference (TREC 7). Available online at:
http://www.citeseer.ist.psu.edu/3475.html

Brashler, M., Kan, M-Y, Schauble, P and Klavans, J. (2000). The Eurospider Retrieval Systems and the TREC – 8 Cross Language Track. In Proceedings of the Seventh Text REtrieval Conference (TREC 8). Available online at:
http://www.citeseer.ist.psu.edu/3475.html

Cheng PuJ., Teng Jei., Chen R.C., and Wang J.H. (2002). Translating unknown queries with web corpora for cross-language information retrieval". Available online at:
http://www.iis.sinica.edu.tw/~pjcheng/p238-cheng.pdf

(CLIR) track overview". Proceedings of the Eighth Text REtrieval Conference (TREC 8).
http://trec.nist.gov/pubs/trec8/papers/trec8ov.pdf

Davis M.W., Ogden (1995). Implementing cross-language text retrieval systems for large scale text collections and the world wide web". Available online at::
http://www.ee.umd.edu/medlab/filter/sss/papers/

Foskett D.J. Thesaurus, As edited and presented in *Readings in Information Retrieval*. In Sparck Jones, K., & Willet, P. (Eds.), San Francisco, CA: Morgan Kaufmann Publishers. (1997)

Fluhr, C. (1996). "Multilingual information retrieval". In A. Zaenen (Ed.), Survey of the State of the Art in Human Language Technology.
http://cslu.cse.ogi.edu/HLTsurvey/ch8node7.html#SECTION85

Gaussier H., Sadat F. (2003). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction". Available online at:
http://www.acl.ldc.upenn.edu/C/C02/C02-1166.pdf

Gillam R. *Unicode Demystified: A Practical Programmers Guide to the Encoding Standards*. Boston: Addison-Wesley. 2003

Grefenstettte, G (1992). "Use of syntactic context to produce term association lists for text retrieval". In SGIR'92. Available online at:
http://www.citeseer.ist.psu.edu/context/182322/0

Haddouti, H. (1999). "Survey: multilingual text retrieval and access."
http:zSzzSzwww.forwiss.tumuenchen.dezSz~haddoutizSzsurvey.pdf/haddouti99survey.pdf

Hansah A., Evens M. (1999). "Arabic/English Cross Language Information Retrieval using a bilingual dictionary"Available online at:
http://www.elsnet.org/arabic2001/hasnah.ppt

Hull, D. A., & Grefenstette, G. (1996). "Querying across languages: a dictionary-based Approach to multilingual information retrieval". In Sparck Jones, K., & Willet, P. (Eds.), *Readings in Information Retrieval*. San Francisco, CA: Morgan Kaufmann Publishers.

International Standard 5964: Documentation Guidelines for the Establishment and Development of Multilingual Thesauri. First Edition; 61 p; American National Standards Institute; 1985.

Keranen, S (2004). "Content Management – Concept Indexing and term Equivalence in multilingual thesauri." PhD dissertation. Available online: http://abo.fi/-skernen/research.html

Milestead, J-L. (1998) "NISO Z39.19: Standard for the Structure and Organization of Information Retrieval Thesauri" Paper presented at the Taxonomic Authority Files Workshop. Available online at:
http://www.bayside-indexing.com/Milstead/z39.htm

Nega A. and Willet P (2002). Stemming of Amharic Words for Information Retrieval. *In Literary and Linguistic Computing.* Oxford: Oxford University Press. Vol. 17, No 1. pp 1-17

Oard, D. W. (1997). Cross-language text retrieval research in the USA." Available online at:
http://www.citeseer.ist.psu.edu/cache/papers/cs/1674/http:zSzzSzwwwir.inf.ethz.chzSzDELOSzSzOardzSzoard.pdf/oard97crosslanguage.pdf

Oard, D. W. (1997). Serving users in many languages: cross-language information Retrieval for digital libraries. D-Lib Magazine." Available online at:
http://www.dlib.org/dlib/december97/oard/12oard.html

Oard, D. W., & Dorr, B. J. (1996). A survey of multilingual text retrieval.
http://citeseer.ist.psu.edu/oard96survey.html

Ruge, G (1991), "Experiments on linguistically based term associations". In RIAO'91, pp528-545. Available online at:
http://www.citeseer.ist.psu.edu/context/176640/0

Reitmeyer B. (2004). "Cross Language Information Retrieval" Available online at:
Http://www.ischool.utexas.edu-i385df04.html

Saba A. (2001). The Application of Information Retrieval Techniques to Amharic Documents on the Web. Master Thesis . Addis Ababa, Addis Ababa University.

Shalgren M., Hansen P., Karlgren J. (2003). English-Japanese cross-lingual query expansion using random indexing". Available online at:
http://www.research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-PATENT-SahlgrenM.pdf

Salton, G, Michael, J. (1983) "Introduction to Modern Information Retrieval". New York, Mc-Graw Hill.

Salton, G. (1971) "The Smart Retrieval System". Englewood, Prentice Hall.

Soergel, D. (1997) "Multilingual Thesauri and onthologies in Cross Language Retrieval" Paper presented at the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval." Available online at:
http://www.dsoergel.com/cv/B60.html

Soergel, D. (1995) "Data Models for an Integrated thesaurus database". Paper presented at the Research Seminar on Compatibility and Integration of Order Systems, Warsaw, Poland. Avavilable online at:
http://www.dsoergel.com/cv/B60.html

Sormunen E., Laaksonen J. Keskustalo H and Kekalainen J. (2002). The IR game - a tool for rapid query analysis in cross language IR experiments". Available online at:
http://www.info.uta.fi/qpa/julkaisuja/irgame.pdf

Spark. J.K, William. P (1997) "Readings in Information Retrieval". San Francisco. Morgan, Kauffmann.

Yonggang Qui and Hans-Peter Frei (1995). Concept Based Query Expansion. In proceedings of SGIR-93, 16th ACM international Conference on Research and Development in Information Retrieval. Available online:
http://citeseer.ist.psu.edu/qui93concept.html

Zhou Y., Qin J., Chau M., and Chen H., (2003). Experiments on Chinese-English cross-language retrieval at NTCIR-4". Available online at:
http://www.business.hku.hk/~mchau/papers/NTCIR4.pdf

ANNEX

No	queries	DocID's Relevant Same language	Partially relevant	DocID's Relevant Cross language	Partially relevant Cross
1	የኢንሹራንስ ፖሊሲ	93	19	187,209,285,286, 287,290	291,293
2	የወኪል ኃላፊነት እና ተጠያቂነት	35		159,222	
3	የውል አፈጻጸም ተጠያቂነት	29		279,295,297,127, 223,294,295,297	
4	አሸሙር ማህበራትና አደረጃጀታቸዉ			102,231,233,234	208
5	ተላላፊ የገንዘብ ሰነድ ዋጋ	12,19		260,261,262,264	
6	የባንክ እንቅስቃሴዎች አይነትና ተግባር	68,69,70,71,72,7 3,74,75,77,78,79, 80,81	83,84,87		
7	የንግድ ምልክት	21		170,211,218,219, 225,227	
8	የንግድ ድርጅት አወቃቀር	26			

9	የዋስትና አሰጣጥና ሀላፊነት	4,			
10	የሶስትዮሽ ስምምነት				
11	የዉል ማፍረስና የሚያመጣዉ ተጠያቂነት			279,295,296,127, 279	223
12	የንብረት ባለቤትነትን መብት ማስከበር	17,47			140
13	property insurance	293,		93	19
14	contractual liability of agents	279,297,296, 127,223,279, 296,297	294,295,302, 303	29	
15	tort	209,214,296, 297			
16	commercial laws in Ethiopia	185		56,68	

17	major insurance and bank services	287,288,290		16,18,69,70,71,72, 73,74,75,77,78,79, 80,81	83,84,87
18	property rights	140		47	76
19	bank account			18,69,70,71,72,73,	74,75,77
20	dissolution	101			
21	capacity of minors in contracting	223		29	
22	cheque fraud			12,16,71	
23	agent liabilities	159,276,298, 300,301			
24	copyright issues in Ethiopia		200,203,210, 226		
25	Bankruptcy	105,162,191, 192,213,216, 217,278		25,86	

Table 1: The relevance judgments of documents by domain experts for queries used.

No	Queries	Total relevant retrieved	Total relevant in collection	Total retrieved
1	የኢንሹራንስ ፖሊሲ	2	2	2
2	የወኪል ኃላፊነት እና ተጠያቂነት	0	1	0
3	የውል አፈጻጸም ተጠያቂነት	1	1	8
4	አሽሙር ማህበራትና አደረጃጀታቸዉ	0	0	0
5	ተላላፊ የግንዘብ ሰነድ ዋጋ	2	2	11
6	የባንክ እንቅስቃሴዎች አይነትና ተግባር	5	16	9
7	የንግድ ምልክት	1	1	4
8	የንግድ ድርጅት አወቃቀር	1	1	8
9	የዋስትና አሰጣጥና ሀላፊነት	1	1	2
10	የሶስትዮሽ ስምምነት	0	0	0
11	የንብረት ባለቤትነትን መብት ማስከበር	0	2	3

12	property insurance	1	1	9
13	contractual liability of agents	5	12	20
14	tort	4	4	4
15	commercial laws in Ethiopia	1	1	20
16	major insurance and bank services	3	3	16
17	property rights	1	1	11
18	bank account	0	0	0
19	dissolution	1	1	2
20	capacity of minors in contracting	0	1	7
21	cheque fraud	0	0	1
22	agent liabilities	3	5	12
23	copyright issues in Ethiopia	3	4	9
24	Bankruptcy	6	8	12

Table 2: Results of the monolingual runs.

No	Queries	Total relevant retrieved	Total relevant in collection	Total retrieved
1	የኢንሹራንስ ፖሊሲ	6	8	17
2	የወኪል ኃላፊነት እና ተጠያቂነት	1	2	20
3	የውል አፈጻጸም ተጠያቂነት	4	6	20
4	አሽሙር ማህበራትና አደረጃጀታቸዉ	0	0	4
5	ተላላፊ የግንዘብ ሰነድ ዋጋ	1	4	3
6	የባንክ እንቅስቃሴዎች አይነትና ተግባር	0	0	13
7	የንግድ ምልክት	6	6	6
8	የንግድ ድርጅት አወቃቀር	0	0	9
9	የዋስትና አሰጣጥና ሀላፊነት	0	0	5
10	የሶስትዮሽ ስምምነት	0	0	0
11	የዉል ማፍረስና የሚያመጣዉ ተጠያቂነት	4	5	20
12	የንብረት ባለቤትነትን መብት ማስከበር	1	1	9

13	property insurance	1	2	7
14	contractual liability of agents	0	1	7
15	tort	0	0	0
16	commercial laws in Ethiopia	0	2	0
17	major insurance and bank services	5	17	11
18	property rights	0	2	4
19	bank account	0	0	0
20	dissolution	0	0	4
21	capacity of minors in contracting	0	1	0
22	cheque fraud	1	3	2
23	agent liabilities	0	0	8
24	copyright issues in Ethiopia	0	0	0
25	Bankruptcy	0	2	0

Table 3: Results of the Cross language runs.

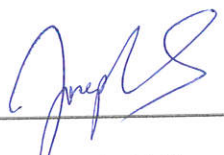
No	Queries	Monolingual		Cross lingual	
		Recall	Precision	Recall	Precision
1	የኢንሹራንስ ፖሊሲ	1	1	0.75	0.352
2	የወኪል ኃላፊነት እና ተጠያቂነት	0	0	0.5	0.05
3	የውል አፈጻጸም ተጠያቂነት	1	0.125	0.66	0.2
4	አሸሙር ማህበራትና አደረጃጀታቸዉ	0	0	0	0
5	ተላላፊ የገንዘብ ሰነድ ዋጋ	1	0.182	0.25	0.33
6	የባንክ እንቅስቃሴዎች አይነትና ተግባር	0.3125	0.55	0	0
7	የንግድ ምልክት	1	0.25	1	1
8	የንግድ ድርጅት አወቃቀር	1	0.125	0	0
9	የዋስትና አሰጣጥና ሀላፊነት	1	0.25	0	0
10	የሶስትዮሽ ስምምነት	0	0	0	0
11	የዉል ማፍረስና የሚያመጣዉ ተጠያቂነት	0	0	0.8	0.2
12	የንብረት ባለቤትነትን መብት ማስከበር	0	0	1	0.11

13	property insurance	1	0.11	0.5	0.142
14	contractual liability of agents	0.42	0.25	0	0
15	tort	1	1	0	0
16	commercial laws in Ethiopia	1	0.05	0	0
17	major insurance and bank services	1	0.19	0.294	0.45
18	property rights	1	0.1	0	0
19	bank account	0	0	0	0
20	dissolution	1	0.5	0	0
21	capacity of minors in contracting	0	0	0	0
22	cheque fraud	0	0	0.33	0.5
23	agent liabilities	0.6	0.25	0	0
24	copyright issues in Ethiopia	0.75	0.33	0	0
25	Bankruptcy	0.75	0.5	0	0

Table 4: Recall-precision table for all queries.

DECLARATION

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



Yoseph Shiferaw

June, 2005

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

Nega Alemayehu [PhD]

June, 2005