

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

CONSTRUCTION OF ENGLISH-AMHARIC ELECTRONIC
SUBJECT DICTIONARY FOR SCIENCE AND TECHNOLOGY

TERMS: AN EXPERIMENT WITH MATHEMATICAL TERMS

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



BY

SEBSIBE HAILEMARIAM

JUNE 2001

ADDIS ABABA UNIVERSITY
LIBRARIES
P.O. BOX 1176
ADDIS ABABA ETHIOPIA

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION STUDIES FOR AFRICA

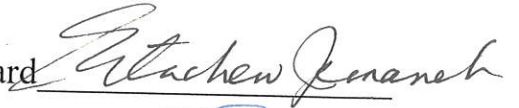
**CONSTRUCTION OF ENGLISH-AMHARIC ELECTRONIC
SUBJECT DICTIONARY FOR SCIENCE AND
TECHNOLOGY TERMS: AN EXPERIMENT WITH
MATHEMATICAL TERMS**

By

SEBSIBE HAILEMARIAM

Name and Signature of Members of the Examining Board

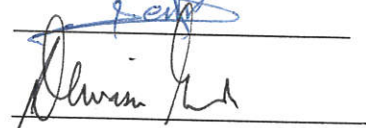
Ato Getachew Jemaneh, Chairman, Examining Board



Ato Dereje Teferi, Advisor



Dr. Demissu Gemedo, Advisor



Dr. Lars Asker, External Examiner



ACKNOWLEDGMENT

First and for most I want to forward my best admiration to my advisors Dr. Demissu Gemedu and Ato Dereje Teferi for their day to day follow up, comments and recommendation on the progress of this Research.

Next, I want to thank my best friends, Esete Arage, Yosef Adugna, Addisalem Negash, Worku Tamrie, Aderaw Workneh, and Nibret Hassen for their unforgettable contribution for the success of this research.

Finally, I want to dedicate this research to those that helped me to be a man of the day, my families.

Table of content

1	THE PROBLEM	1
1.1	Background	1
1.1.1	Introduction.....	1
1.1.2	Education in Ethiopia.....	3
1.1.3	The Amharic Language.....	5
1.1.4	Orthography of the Amharic Language	6
1.1.5	Language Issues.....	9
1.1.6	Issues and considerations in teaching using local language.....	11
1.1.7	Efforts of the STTTP.....	14
1.2	Statement of the problem	17
1.3	Scopes and Limitation	22
1.4	Objectives	23
1.4.1	General objective of the study.....	23
1.4.2	Specific objectives.....	23
1.5	Methods	24
1.6	Application of the result	26
1.6.1	Assist Document translation or production	27
1.6.2	Input to machine translation.....	27
1.6.3	Spelling and grammar checking.....	27
1.6.4	Development of Amharic language.....	28
2	THESAURUS CONSTRUCTION AND MACHINE TRANSLATION.....	29
2.1	Thesaurus	29

2.1.1	Introduction.....	29
2.1.2	Organization of terms in thesaurus construction	32
2.1.3	Terms form.....	36
2.1.4	Compound term factoring.....	37
2.1.5	USE and UF reference.....	38
2.2	Homography and scope notes	39
2.3	Multilingual aspect	40
2.4	Dictionary	42
2.5	<i>Machine Translation</i>	45
2.5.1	Human translation.....	46
2.5.2	Machine-aided human translation (MAHT)	46
2.5.3	Human-assisted machine translation (HAMT)	47
2.5.4	Fully automatic (automated) machine translation (FAMT)	47
2.5.5	The socio-political importance.....	48
2.5.6	Commercial importance.....	49
2.5.7	Scientific importance.....	50
2.5.8	Philosophical importance.....	50
2.6	Multilingual dictionary	53
3	THESAURUS AND DATABASE OF THE NEWLY TRANSLATED AMHARIC TERMS	54
3.1	Data source	54
3.2	Algorithm development	55
3.2.1	Method of thesaurus construction.....	55

3.3	Application of thesaurus in electronic dictionary ..	56
3.4	Database design	57
3.4.1	Term Code entity.....	66
3.4.2	Thierarchy entity.....	67
3.4.3	TRelation entity.....	68
3.4.4	Dictionary entity.....	68
3.4.5	Picture entity.....	69
3.4.6	ATSynonymous entity.....	70
3.4.7	ETSynonymous entity.....	70
3.4.8	SETCode entity.....	71
3.4.9	SESTCode entity.....	71
3.5	Entity relationship (E-R) diagram	72
3.6	Referential integrity constraints	74
3.7	Phrases matching and processing	74
3.8	Document processing	79
3.9	General logical design to document processing	80
3.10	General logical design of phrase processing	81
3.11	Detailed logical design of phrase processing	82
3.12	Amharic software	83
4	Program Development	84
4.1	System input	84
4.1.1	Phrasal input.....	84
4.2	Document file input	87
4.3	Record selection from term code table	88

4.4	Logical design to English input word processing to get the term code.....	91
4.5	Logical design of creating a link to Amharic document	96
4.6	The user interface	97
4.7	Sample test result	100
4.7.1	Document file processing result.....	100
4.7.2	Performance evaluation of the system for user query.....	101
4.7.3	Verification of the formula $\text{Sim}(\text{phrase}_i, \text{phrase}_j) = (CM) / ((NK) (N-K +1))$	105
5	Conclusions and recommendations.....	109
5.1	Conclusion	109
5.2	Recommendation	110
6	References	111
	Declaration.....	115

ABSTRACT

This thesis focuses on the construction of English-Amharic electronic subject dictionary for science and technology mathematical terms.

Development of the local languages in the third world countries becomes important issue for developing and keeping cultural heritage. The problem and solution to make Amharic a medium of instruction are investigated in the research. The existence of the electronic dictionary for science and technology terms will promote the performance of professionals to translate science and technology documents and to prepare document in the field.

An information system is developed that provide basic information and the procedures are described in this document. The performance of the system is also tested using document file and user queries as inputs. The result shows the output information accuracy is inversely proportional to the number of concept bearing words in the query.

List of Abbreviations

TCode	Term Code
Thierarchy	Term Hierarchy entity
TRelation	Term Relation entity
Tpicture	Term Picture entity
ATSynonymous	Amharic Term Synonymous entity
ETSynonymous	English Term Synonymous entity
SETCode	Stemmed English Terms Code entity
SESTCode	Stemmed English Synonymous Terms Code entity
STTTP	Science and Technology Term Translation Project
MTE	Mother Tongue Education
ECOSA	Ethiopian Computers Standard Association
IFSTD	Interim Fund for Science and Technology for Development
NT	Narrower Term
BT	Broader term (In thesaurus construction)
RT	Related Term (In thesaurus construction)
TCode	Term Code
PTerm	Preferred Term
ATerm	Amharic Term
ASTerm	Amharic Synonymous Term
Def	Definition
RTerm	Related Term (In database design)
NTerm	Narrower Term (In database design)
ETerm	English Term
ESTerm	English Synonymous Term

SETerm	Stemmed English Term
SESTerm	Stemmed English Synonymous Term
RTCode	Related Term Code
NTCode	Narrower Term Code
ETCode	English Term Code

List of Tables	page
Table 3.1	59
Table 3.2	60
Table 3.3	60
Table 3.4	60
Table 3.5	60
Table 3.6	61
Table 3.7	61
Table 3.8	61
Table 3.9	62
Table 3.10	62
Table 3.11	62
Table 3.12	63
Table 3.13	63
Table 3.14	63
Table 3.15	63
Table 4.1	101
Table 4.2	102
Table 4.3	104
Table 4.4	106

<u>List of figures</u>	<u>Page</u>
fig 3.1	73
Fig 4.1	98
Fig 4.2	99
Fig 4.2	99

CHAPTER ONE

1 THE PROBLEM

1.1 Background

1.1.1 Introduction

Rapid strengthening of national scientific and technological capabilities is becoming obligatory in all developing countries. One of the ways to realize this objective is through dissemination of scientific and technological concepts among the peoples (Girma Yilma, 1987). The most important means of disseminating science and technology concepts is the use of language. But only well-developed languages have the ability to transmit the concepts properly to the peoples. According to Girma (1987), the dissemination of science and technology concepts can be done in two ways. The first way is by modernizing the local language(s) so that it can be able to transmit the concepts to the society without problems. The second alternative is the use of one of the well-developed foreign languages, which require extensive training in the language by the society.

According to Bisrat (1987), some of the developing countries appear to be trying to solve such problems by adopting one of the developed European languages as a medium of conducting

science and technology in their respective countries. But language is not only a means of communication but also it is a storehouse of culture, tradition, history, etc.

When one society becomes exposed to another modern society they will encounter many science and technology concepts that are not part of their traditional world. In most cases, they will simply borrow words from another language and sometimes they will create their own terms by redefining existing terms or by resurrecting old terms that are no longer in use and give them new meanings in order to talk about these technologies.

The development of science and technology breaks the geographic barriers of societies for their contact and transfer of technology, culture, information and other human properties. As a result some communities fear that the use of loan words to transmit new science and technology concepts will erode their language and culture and so attempt to keep their language pure.

Even though the implementation of each of the above methods has their own cons and pros, development of ones language is the most realistic and rational choice for developing country like Ethiopia. It was with this view that the Science and Technology Terms Translation Project (STTTP) was launched.

1.1.2 Education in Ethiopia

Ethiopia is an ancient country with a very rich cultural heritage, possessing its own alphabet and literature.

Formal education started in Ethiopia during the fourth century when Christianity began as a well-organized religion in the country (Britannica 1996). The advancement of education has strong relation with the Orthodox Church in Ethiopia. According to Van Horn (1997), the Church had total control over education for about 1500 years. In addition to the schools of the Church, there was Qur'anic schools in the western part of Ethiopia.

Education in Ethiopia has changed tremendously since 1800's, because the government has made an attempt to improve children's education. Emperor Minilik II formed the European modeled school in 1908. The real development of the educational system came after World War I when Emperor Haileselassie was crowned (Van Horn, 1997).

The first college in Ethiopia was founded in 1950's with limited resources and staffs. A number of policies were formulated and implemented by the ruling class to increase the effectiveness of education as well as the enrolment rate of children in rural and urban areas of the country. However, the enrolment rate is even now very low which indicates that a lot

has to be done by the government as well as the public and community to achieve a better result (Nguyen, 1998).

The literacy rate of Ethiopia is about 32.9%, which is one of the lowest rates in Africa. This can be an indicator to the performance of educational service of the country (Weber, 1998 as cited by Van Horn 1997). The Ethiopian educational system is considered as poor. Ethiopian children are behind children in other African countries in acquiring basic academic skills such as reading, writing and mathematics.

In recent years, the involvement of the government and the private sector has made education opportunities wider and has resulted in a new structure than the education system implemented in the 1950's. In the 1950's formal education was composed of six years of elementary school, two years of junior elementary, and four year of senior secondary. Students then took the Ethiopian School Leaving Certificate Examination at the end of grade twelve.

"The current strategy adapted by the government is to widen the base of the system, to diversify the structure significantly with emphasis on vocational and technical training, and to make education relevant to the culture of Ethiopia" (Ofcansky 1993 as cited by Van Horn 1997).

In spite of the various problems that Ethiopia has, it is moving rapidly from policy making to implementation in the provision of mother tongue education (MTE) for its entire people. The transitional government of Ethiopia has done a lot in capacity building (language standardization, orthography development, training, material production, etc) for the success of the MTE for all ethnic classes.

1.1.3 The Amharic Language

Like any other sub-Saharan African country, Ethiopia has complex pattern of cultures and languages. One of the specifier of the Ethiopian culture is the availability of different languages. But, unlike the sub-Saharan African states, Ethiopia has old tradition of writing, which existed since the Axum civilization. This tradition or writing helped the transcription of the Gee'z language. From Gee'z, other languages like Amharic, Tigre, Tigrigna, Argoba, Harari etc were developed as modern Ethiopian language (Christine).

Gradually, Amharic became the working language of the Ethiopian court, which hasn't the capability of fully transcribing in to a written form. According to Levine as cited by Christine, in 1620's one can easily observe the development of Amharic language since any one that knew the

language can communicate to a large mass of the society. Through slow process, Amharic became both spoken and written language with a place in education. Until the end of the 19th century, Amharic was not a language of education. It was at the end of of the 19th century, it began to replace Gee'z as the language of secular education. Amharic was used as a medium of instruction for elementary students for a period of time. Currently, it is also used as a medium of instruction for junior secondary education in the regions that use Amharic as a first language.

1.1.4 Orthography of the Amharic Language

Orthography is the branch of linguistics that studies the elements of written language such as letters, punctuations and spellings. The written mode of the Amharic language has letters arranged in 7 columns (ግዕዝ፣ ካይም፣ ሳልስ፣ ራብዕ፣ ሐምስ፣ ሳድስ፣ ሳብዕ) and 35 variety of letters arranged in rows. In addition to this, there are more than 25 irregular forms of the Amharic letters.

Orthographically, some of the Amharic letters has identical pronunciation, which creates inconsistency on written documents. Some of these are:

❖ 0 & አ

❖ ሀ, ሐ, & ኀ

❖ ሰ & ሠ

❖ ጸ & ፀ

This inconsistency should be removed through standardization of the Amharic letters to simplify Amharic document processing. In addition to the orthographic problem that we have seen above, the character mapping of the available Amharic software's are poorly constructed. This creates a problem in order to use them to process the electronic version of the Amharic document. The most appropriate solution to this problem is standardization of the Amharic character. However little has been done in the area of standardization of the orthography of the Amharic language. Recently an association, namely the Ethiopian Computer Standard Association (ECoSA), that works on the standardization of the characters of the Amharic letters, the keyboard layout and on character mapping have been formed.

Communication is enhanced when standardized ways of writing and spelling and standard terms for technological phenomena are agreed upon and used within a speech community (Martin) [online 21].

In information retrieval (IR) systems hosting unrestricted text, the task of matching one string of characters to

another string of characters would be difficult unless there is a normalizing algorithm that processes both the document text and the query text. In the process of information retrieval that searches information from a database, normalization of documents can be attempted before, which shield searchers not to consider simple process but avoid relevant document from retrieval. For instance, problem of upper case and lower case letters can be avoided first. This normalization helps us to find common words of both the query and document (Brook, 1998).

The ideal aim of normalization is to produce the smallest distinctive unit of a language called lexemes or lexical words, which usually consists of a single word but some times multiple words (phrase). Due to the incompleteness of grammar or lexicons for a language, it is very difficult to obtain adequate parsers for a text in the process of normalizing a document based on the natural language parser (Hindle as cited by Brook (1998)).

According to Brook (1998), the identification of token or a word from a text is one of the fundamental problems of any mechanical process. He also classifies a word as graphic word, compound word, merged word and pseudo word.

According to his definition, graphic word is a string of letters demarketed with space that may contain only hyphen and/or apostrophes whereas compound word is string composed of other words, either appearing as continuous strings in hyphenated form, or linked with space. Merged words are words in which two words are merged to reflect a reduction in spoken language. And pseudo words are words that have been linked together by hyphen.

1.1.5 Language Issues

Language has variety of natures, function and properties for the society. According to Nagra (1976) the following are some of these functions and properties of language:

- ❖ Language is a human property
- ❖ Language is a system of sounds, structures, lexical items and meanings
- ❖ Language is a symbolism
- ❖ Language has written and spoken modes.

It is very important to create awareness and use the cultural value of oral literature to the citizen of a country. But in most third world countries, ❖ local language studies is very abstract in nature and often sound very esoteric and less familiar than the study of English or French whose terminology

is more commonly understood by average educated person❖
(Ngra, 1976). The Amharic language study is also at a very
infant stage; which is more of mechanical, theoretical and is
still underdeveloped.

The effort of STTTP, which was an attempt to replace the
foreign language loan words/expression by Amharic
word/expressions, is significant because it proved that
Amharic has adequate possibility of creating its own
expression for newly invented concepts.

Language development involves working with speakers of less
well-known languages to help them achieve their goals as a
community through language planning and inter cultural
development efforts. In some cases this means the development
and promotion of written forms of their language and the
production of literature (online document).

All languages have about the same level of structural
complexity and all are capable of allowing their speakers to
describe the world in which they live (online, 21) but not all
languages have words/phrases for all concepts of the world. It
doesn't mean that they are not capable of expressing that
concept. It is possible to provide new functions to any
language.

1.1.6 Issues and considerations in teaching using local language

Education in Ethiopia dated back long ago; but before 1900 education was limited to the use of Amharic and Ge'ez languages in religion. Emperor Menilik II established the first governmental school in 1908. The mediums of instruction for the modern education offered in Ethiopia were: English, French, Oromigna and Amharic. Modern Education became more organized after the establishment of Ministry of Education in 1930.

According to Assefa G/Mariam (1982), Amharic was the court language or "language of the king" in the 13th century and was the only language that was privileged. Amharic became the "national" or "official" language of Ethiopia in the revised constitution of 1955. It also became the medium of instruction in all elementary schools, in Ethiopia. Today Amharic is the medium of instruction in elementary schools in Amhara region, Addis Ababa and Diredawa and in junior secondary schools in Amhara region and Diredawa.

The contribution of a language to the development of a country is unquestionable. According to STTTP publication (Bisrat,

1987), there exist two different ways of language use for fostering modern science and technology development in third world countries.

These are:

1. Adopting one of well-developed European languages (like English and French) that can fully express concepts of science and technology and teach the adopted language to the citizens in order to spread the knowledge of the science and technology in the society.
2. Modernize the local language so that science and technology concepts can be transmitted to the society through these local language(s).

The second alternative is the best if possible, since learning using local language is easier than learning using foreign languages. Especially in third world countries, where the highest proportion of the population is illiterate, the second option is the only alternative to spread knowledge of science and technology among all citizens of the country (STTTP publication).

The second alternative presupposes that every language can be a language of science and technology if there is an effort in developing the language.

Many African and Asian languages are now being modernized in order to serve for spreading of Science and Technology. For example, there was a big effort in Korea to change the medium of instruction of the country to local language. Egypt, Malagasy, and Tanzania in Africa, Israel in Asia are also countries that are good examples in modernizing their languages with great efforts by their governments, policy makers and professionals (Yanni, 1987; Anber, 1987; Rahaingoson, 1987; Massamba, 1987). Japan, India, and Turkey are also countries that use their language for spreading science and technology to their society (STTTP publication).

According to Abrham (1968), since the end of the 19th century, Ethiopia has witnessed uncontrolled influx of scientific and technological terms. It was also the time when Ethiopia opened her door to welcome everything that modern science and technology had to offer.

Therefore given that education is a tool for creating awareness among the society, Ethiopia should work a lot in modernization of the Amharic language and the vernaculars.

1.1.7 Efforts of the STTTP

The Science and Technology Term Translation Project (STTTP) funded by the United Nations Interim Fund for Science and Technology for Development (IFSTD) was implemented under the Academy of Ethiopian Languages.

The Academy of Ethiopian Languages has been set up by replacing the previous Amharic Language Academy, which was established in 1972 with a government order No. 79 to facilitate and standardize translation of words into Amharic (Amsalu, 1986).

Ermias & Demissu (1974 E.C.) explain the purpose of the STTTP as follows:

"The significance of developing Amharic as a medium of instruction in schools as well as improving the capability of this language in transmitting science and technological concepts through the mass-media and also facilitating absorption of science and technology concepts by ordinary people were underlined while formulating the terminology project."

In addition to this the major developmental objective of the project were:

" a) To provide a medium through which scientific and technological knowledge can be easily disseminated and absorbed at all levels of the society.

b) To facilitate the mass educational objectives of the government both in and out of school.

c) To harmonize the relationship and narrow the gap in scientific and technological understanding between lower, middle and upper echelons of national science and technology manpower, by the use of a common standardized Amharic science and technology terminology as far as possible.

d) To improve the delivery and assimilatory efficiency of technical based projects at the community level."

To implement the STTTP, 13 disciplines that deal with the basic sciences and technology were selected. These were Agriculture, Botany, Building Construction, Chemistry, Electro-Mechanics, Geography, Geology, Mathematics, Medicine, Nutrition, Physics, Statistics, and Zoology (Amsalu, 1986).

The project was subdivided into three phases: compilation of terms, translation of terms, and preparation of illustrated booklets and popular literature.

For the compilation of terms, English was taken as a source language to find out words/phrases. The reason for the

selection of the source language to be English has many folds. Some of these are:

- English is widely used as a medium of instruction in Ethiopia at junior secondary, secondary and tertiary level at the time.
- It is one of the developed languages in science and technology concepts.
- Materials are readily available in the language
- Professionals easily realize concepts of science and technology using this language. Etc.

During compilation process, words for which already coined equivalents exist in Amharic and names of units, laws named after individuals, names of most elements and compounds, internationally accepted scientific names are not included. As a result over 12,500 terms were compiled at the beginning. (Ermias & Demissu, 1974)

During the translation process, about 15,087 terms were translated from English to Amharic that includes newly coined terms, adopted terms and some terms that are already in use. Most of them are translated with the ultimate aim of offering secondary and early higher education in Amharic language (Ermias, 1986) in the country.

In the discipline of mathematics, 1038 word were translated (Ermias & Demissu, 1974 EC). One of the ultimate goal of the project was the development of a single Amharic-Amharic science and technology dictionary called “የሳይንስና ቴክኖሎጂ ስያሜ መዝገበ ቃላት”. In addition to the production of the dictionary, it was planned to illustrate the newly produced science and technology Amharic terms (Ermias, 1986). The illustration of the Amharic words cover only 50% of the total translated words in to their Amharic equivalence. Among the translated mathematical terms only 475 terms were illustrated.

1.2 Statement of the problem

Nowadays, with the rapid development of technologies, societies would like to keep their culture unaffected. When they borrow technology from other societies they adopt the technology to fit into their economical, cultural and other attributes of their own society.

Ethiopia, as in many developing countries of Asia and Africa (Abrham, 1968), faces the problem of achieving national and cultural integration by adopting the technology of industrial western countries within short period of time.

Language is the backbone for achieving this goal. English, as a medium of instruction, in universities, colleges, and secondary schools, as a second official language in various areas, as a language in broadcasting, and in some cases as a language for official correspondence, creates various problems in our society. According to Abrham (1968) and the Seminar of Ethiopian Language Academy and AAU Language Institute in Ministry of Culture and Sport (1972 EC), the following are some of the problems related to English as medium of instruction:

- Students face difficulty in mastering as their second and for some even the third language.
- Barely literate adults face constraint to understand science and technology because they need knowledge of the language in order to understand scientific disciplines.
- Students become immersed in a foreign culture without first appreciating their own culture.

The reasons for these problems are many folds. The external pressure as well as our condition forced us to know science and technology using one of the international languages (French and English) (Amsalu, 1983).

When students learn in their own language, they easily grasp the knowledge and can easily communicate with the society using their language. This helps them to improve their capacity of solving problems and results in development of the country in all aspects including science and technology. The adoption of foreign language affects the development of the local languages, effectiveness of education and research made on local languages.

Some of the current problems to provide education of science and technology and to conduct researches in Amharic language are:

- Shortage of teaching materials prepared in Amharic: This problem is manifestation of lack of trained manpower that prepares teaching and other reference materials in Amharic. So far efforts are made mainly to translate textbooks written for schools of advanced countries. Due to limited vocabularies of Amharic and absence of standardization, the translation also has problems.
- Shortage of qualified teachers that can teach in Amharic in secondary schools and higher level education.

- Inadequacy of Amharic vocabulary for newly born English terms.

One of the disciplines for which STTTP attempted to translate words from English to its equivalent in Amharic was Mathematics. Unlike the other disciplines of science, every branch of Mathematics is based on logical constructs and axiomatic logical systems. This means that, any mathematical term requires precise and unambiguous equivalent term in Amharic.

The STTTP have prepared two different booklets in the process of translation of terms for the discipline mathematics. The first one is Amharic dictionary of mathematics and the other is illustration for some of the concepts of the newly translated terms.

The Amharic dictionary provides definition to newly constructed Amharic terms in addition to rendering equivalence. However, the Amharic definition of the Amharic term require recursive definition to obtain the concept coined to the new terms. For example the English term **cylinder** is translated in to Amharic as “በርሜሎ”. The Amharic definition becomes “አንድ መስመር ከአንድ የተወሰነ መስመር *ከልክልነቱን* እየጠበቀ አንድ የተሰጠ ኩርባ እየተከተለ ይህንኑ መስመር በመሸከርከር የሚሰጠው ገጽ የሚፈጥረው ጆሜትሪያዊ

ምስል ነው።". All the underlined words are also science and technology terms which require further processing. For example, ጀመትሪያዊ ምስል is defined as "ማንኛውም በጠለል ወይም በጠፈር ውስጥ ነጥቦችን፣ ጉማጅ መስመሮችን እና ጉማጅ ጠለሎችን በመጠቀም ማለት በማያያዝ፣ በማገጣጠምና በማቀነባበር የሚፈጠረው ጀመትሪያዊ ቅርፅ ወይም ውጤት ነው።".

In addition consider the statement "X is a point on the plane with coordinate (1,2)". There are science and technology concepts that have newly translated Amharic equivalence, when the above statement is translated to Amharic. It becomes "ሀ በጠለል ላይ የምትገኝ ነጥብ ሆና ፈርጆቿ (1,2) የሆነ።" The underlined terms are parts of the translated mathematical terms.

This complexity of understanding newly translated science and technology terms should be supported by the computer technology so that subject specialists and professionals can easily grasp the concepts and literature and documents can be easily compiled and distributed to the society.

1.3 Scopes and Limitation

The scope of this experiment is limited to only to the disciplines mathematics, which was one of the disciplines that was included in the STTTP. The STTTP had produced documents for the discipline.

Mathematics is full of symbols, graphs and pictures that transmit various ideas. Due to time constraint the sample prototype provides only textual information and some figurative illustration to limited concepts, making the output information partial.

1.4 Objectives

1.4.1 General objective of the study

The general objective of this study is to investigate the possibility of constructing Amharic-English machine-readable subject dictionary for mathematical science and technology terms that assist professionals of the discipline during translation of materials from English to Amharic or during preparation of science and technology literature in Amharic.

1.4.2 Specific objectives

The specific objectives of the study are to:

- Study the morphology of the newly translated science and technology Amharic terms;
- To illustrate the importance of local language to the society;
- Develop English-Amharic dictionary for mathematical terms;
- Develop English terms thesaurus for mathematical terms;
- Develop Amharic terms thesaurus for mathematical terms;
- Develop an algorithm for the prototype;
- Write the program for the algorithm;
- Test the program for selected terms.

1.5 Methods

Literature review

For the research work, literatures review will be done on machine-readable dictionaries, machine assisted machine translations, vocabulary control system, thesaurus constructions, language issues and modernization of languages that have been done for other languages. There are thesauruses constructed electronically for other languages on different subject area. The techniques that have been used for those systems are studied. For understanding the morphology of the Amharic language, works done on the language are reviewed. The sources for literature review are journal articles, books, and the Internet.

Data

For constructing the dictionary and the thesaurus, the materials prepared by the Science and Technology Term Translation Project for mathematics discipline is used. To test the system efficiency and usability, grade nine mathematics students' textbook, which is prepared in English by the new transitional government is taken. The terms will be selected from the different sub disciplines of mathematics.

Prototype development

For the development of the prototype, this study uses Visual Basic Programming language and Microsoft access application software for development of the database. The Visual basic programming language is used because:

- it is rich in routines to interact with database, which is one of the main components of the system;
- it helps to design flexible interfaces;
- method of interfacing Visual Basic Program to other software of databases is very simple.

1.6 Application of the result

This research is conducted by considering the potential problems of professionals for achieving the objective of the government current education policy and to push the efforts of the STTTP one step further for their success. According to the current education policy of the Amhara region, elementary and junior secondary levels of education of science fields are provided in Amharic.

Since Ethiopia is one of the countries with many vernaculars, similar and wider scope projects can be done in the languages of ethnic groups that provide education to their people in their mother tongue. In addition, this research also contributes for the development of the Amharic language in science and technology. The prototype developed in the process of this research will help professional to simplify their efforts for the modernization of the country.

In addition to the potential value of the research, this research will provide the following to the society:

1.6.1 Assist Document translation or production

The output prototype of this research will assist professionals during teaching material translation or preparation of teaching materials in local language (Amharic) for science and technology disciplines. If the prototype is developed fully for all science and technology terms in all disciplines, it will provide immense advantages for professionals, policy makers, and other governmental and non-governmental organizations. That means, the system developed by this research, machine-readable dictionary, can be used as machine assisted translation system.

1.6.2 Input to machine translation

Machine-Readable Dictionary is the primary source of machine translation system that translate document from one language to another language without human interference, natural language processing, and information storage and retrieval. Therefore the output of this result will have substantial input for further investigations in the areas mentioned above.

1.6.3 Spelling and grammar checking

The system will have considerable application in the area of spelling and grammar checking for science and technology terms.

1.6.4 Development of Amharic language

This study also pushes one step forward the modernization processes of Amharic language and provides smooth atmosphere for researchers who wish to work on related topics in various languages and disciplines.

CHAPTER TWO:

2 THESAURUS CONSTRUCTION AND MACHINE TRANSLATION

2.1 Thesaurus

2.1.1 Introduction

One of the most difficult problems of information retrieval is creating effective organization, which requires keeping related items relatively close to each other. Thesaurus is a tool for vocabulary control to guide indexers and searchers about which terms to use during indexing and searching process that can help to improve the quality of retrieval (Lancaster, 1986). A controlled vocabulary is a fixed list of terms used in IR systems to describe documents. It usually includes the relationships between terms and sometimes the rules for use of vocabulary. The rules may be the methods of combining word to express compound concepts and how to decompose words not found in the vocabulary (Livonen and Sonnenwald, 1998).

According to the research made on the selection of search term among the six discourses, (vocabulary control, document and domain, indexing practice, clients' search requests, databases, and searchers own previous search experiences),

vocabulary control appears to have strong position in the selection of search terms (Livonen and Sonnenwald 1998) .

Vocabulary plays a major role in the effectiveness and efficiency of IR systems. Information retrieval is defined as "the process of searching a collection of documents, using the term document in its widest sense, in order to identify those documents that deal with specific subject" (Lancaster, 1979 as cited by Muddamalle, 1998). The success of information retrieval system can be determined by the accuracy of the data retrieved. Recall and precision are the two measures of the effectiveness of IR systems.

One of the advantages of the vocabulary control system is to avoid ambiguity by presenting standard terms. But in IR systems that are designed for science disciplines that rapidly advance with the technology, the thesaurus must be updated regularly (Muddamalle, 1998). Like wise science and technology terms of the English language are not fully translated into Amharic, which require parallel updating of the thesaurus of the Amharic language for science and technology terms that leads to new terms and usage of terms.

In the translation model of IR, the input request is translated in to a search term, which ideally represents

search concepts or search topics that can be input to the IR systems to identify relevant information.

A thesaurus provides preferred terms, non-preferred terms, semantic relations, guides to application, and rules for synthesis to indexers and searchers. Preferred terms are terms in which indexers and searchers are allowed to use whereas non-preferred terms are terms that are not allowed to them. The thesaurus shows, which preferred term should be used instead of the non-preferred one. A thesaurus also usually allows you to look up a preferred term and see its non-preferred terms (The University of Western Ontario 1998).

Looking at the hierarchy of the different preferred terms can easily show the semantic relations of terms. The semantic relation of terms provides a chance to make the meaning of terms clearer.

A good thesaurus should make it clear what a term is meant to cover. It can accomplish this to some extent by showing non-preferred terms and semantic relations. Other ways of guiding people in using a thesaurus include introductory matter and scope notes (The University of Western Ontario, 1998).

The thesaurus can be either enumerative (that lists all its preferred terms explicitly) or synthetic (that indicates some preferred terms indirectly: instead of listing all the preferred terms, they give rules for creating them). If the thesaurus is of type synthetic, then it contains the rules of synthesis of facets.

2.1.2 Organization of terms in thesaurus construction

Once terms of the domain are collected systematically, they have to also be organized based on their relationships. The objectives of organization of terms as explained by Lancaster are:

"1. To promote the consistent representation of subject matter by indexer and searchers, thereby avoiding the dispersion of related material. This is achieved through the control (merging) of synonymous and nearly synonymous expression and by distinguishing among homographs.

2. To facilitate the conduct of the comprehensive search on some topic by linking together terms whose meanings are related paradigmatically or syntagmatically."

Hierarchical and associative relationships are the two kinds of relationships during the organization process of terms.

Hierarchical relationships

In the process of developing coherent and cohesive hierarchical structure, first we should be able to have a classification scheme. The thesaurus entries derived from the hierarchical relationship development is a genus/species (BT/NT) relationship.

The hierarchical relationship represents the narrower term (NT) is a kind of the broader term (BT) i.e. BT/NT is a thing/kind relationship. There are different kinds of relationships between concepts that may confuse us with the thing/kind relationship (Lancaster, 1987). Some of these are:

- Thing/application
- Thing/derivative
- Thing/instance
- Whole/part

For example, if we consider the relationships between expression and operators (whole/part), or between set and elements of a set (whole/part), then one can easily identify the inconsistency of elements and operators as a narrower term for sets and expressions respectively. But the relation

between algebra and linear algebra is a think/kind relationship. Therefore linear algebra can be narrower term of algebra.

According to the UNESCO guidelines, the whole/part relationship is treated as genus/species relationship for some subject areas like anatomy and geography, for fields of discourse, and social structures. This shows that the narrower terms can be genetic (NTG) or partitive (NTP). But as much as possible this partitive relation should be removed by grouping the NTP in to another BT. This distinction of NTP and NTG should be at the BT level (Lancaster, 1986).

A well-constructed thesaurus will have only one broader term (BT) for each of the terms especially for specific scope.

In thesaurus, it is legitimate to group together as NTs a number of terms derived from the genus (BT) by the application of different principles of division. If there are larger hierarchies it is much convenient to arrange the NTs by principle of division(Lancaster, 1986).

Associative relationship

According to Lancaster (1986), the association relation of a thesaurus is expressed by related terms (RT). This relation

shouldn't be used like terms that appear in the same hierarchy. However it is very difficult to establish precise rules for associative relationships. It is syntagmatic or posteriori.

Any two terms whose meanings are related, but that appear in different thesaurus hierarchies are possible candidates for linking by the RT reference. The selection of related terms for a term is either by technique of co-occurrence of terms in a database if not possible by relying on sense and knowledge of the subject specialist.

Lancaster (1986), also specify ways to find strong candidates of inter term relationships. The following are the rules for the identification of candidate related terms during construction of a thesaurus.

- ❖ A thing can be a RT of its application and vise versa
- ❖ A cause can be a RT of its effect and vise versa
- ❖ A thing can be a RT of its strongly associated property and vise versa
- ❖ A raw material can be a RT of its product and vise versa
- ❖ A complementary activity can be a RT to each other
- ❖ A thing can be a RT of its opposite and vise versa
- ❖ An activity can be a RT of its property and vise versa
- ❖ An activity can be a RT of its agent and vise versa

- ❖ An activity can be a RT of its product and vise versa
- ❖ A thing can be a RT of its parent and vise versa

2.1.3 Terms form

Most thesaurus terms are nouns and noun phrases. Most of the noun phrases are adjectival phrases and some others are prepositional phrases. Verbs are not totally allowed but gerunds are permitted. Adjective appearing in the thesaurus are modifiers of nouns. However in some cases adjective may stand as a thesaurus term (The University of Western Ontario, 1998 and Lancaster, 1986).

According to the online document by the University of Western Ontario (1998), during thesaurus construction, we should keep number consistency (ether plural or singular) and other issues. Therefore one should follow the following:

- ❖ Nouns of quantity (things that can be counted) should be in plural form
- ❖ Nouns of volume (mass noun) should be in singular form
- ❖ Singular for processes, properties, and conditions
- ❖ Avoid inversion
- ❖ Excluding prepositions
- ❖ Excluding punctuation marks, diacritics, special characters, and abbreviations

2.1.4 Compound term factoring

Compound term factoring is the process of identifying and analyzing word concepts so that a term/phrase that have two or more idea can be splitted into words of simple concepts. Compound term factoring can be semantic factoring or syntactic factoring. Semantic factoring involves the analysis of a word into largest number of ideas implied by the term. Whereas syntactic factoring involves the splitting of compound terms into components that could appear in the thesaurus by themselves (Lancaster 1986). But it doesn't mean that, all compound terms should be factored into components. According to UNESCO guidelines, cited by Lancaster (1986), there are three conditions of compound words that should be splitted. These are:

- *Compounds representing a transitive action modified by the name of the entity on which the action is performed;*
- *When a genus term (focus) represent a property or part and the species represents the possessor of the property of the sum of the parts*
- *compounds representing an intransitive action modified by the performer of the action*

Lancaster also observed the compound term that represent different principle of division should be factored.

2.1.5 USE and UF reference

The USE reference of a thesaurus directs from a used term (some times called non preferred term/non descriptors) to a used term. The use reference provides additional entry point into the vocabulary and, hence, are some times referred to as entry terms.

Lancaster states the two basic purposes of the USE reference as "to direct from a term not used in indexing to another term that is more or less synonymous and to direct from specific term, not used in indexing, to their appropriate more general terms."

The syntax of using USE reference on printed thesaurus becomes:

Non-preferred term USE preferred term

The reverse association is shown using the keyword UF (Used For) and its layout becomes:

Preferred term
UF Non preferred terms

UF stands for, the specified preferred term is used for the listed Non preferred terms

According to Lancaster (1986), mostly abbreviations are non-preferred terms and their acronym is the preferred one. But if the abbreviation got popularity over their acronym like RADAR, LASER and others it becomes the preferred term. The process of selecting preferred and non preferred terms take consideration of the users domain and becomes a difficult task in organization of terms.

2.2 Homography and scope notes

Mathematics as a subject matter uses limited term set and structured comparative to the natural language and every concept are more or less expressed by unique concept bearing words/terms. In the discipline of mathematics, homographs (term with more than one meaning), homonym (term with only one pronunciation and more than one meaning) and homophones (different terms with identical pronunciation) are not difficulties in IR.

A scope note of a thesaurus is an explanation of a term in a limited subject area. It needn't be true definition. It is

merely a guide to how a particular term should or shouldn't be used.

According to Barhydt and Schmidt (1968) cited by Lancaster, there are four different uses of scope notes of a term. These are:

- ❖ Positive limitation in the scope note of a term;
- ❖ Negative limitation in the scope note of a term;
- ❖ True definition
- ❖ Combination of definition and Positive limitation

2.3 Multilingual aspect

Nowadays various people and scholars want to handle publications in many languages. The design of information system that can accommodate several languages is the issue of current researches.

According to Lancaster (1986), there are various means for solving the language barrier problem: expansion and improvement of language teaching, research on terminology, development of terminology banks, publication of multilingual glossaries and thesaurus, refinement of MT procedures and so on.

In the third world countries, like Ethiopia, there are many science and technology publication published in the language

of one of the developed language like English which is widely used by local professionals. This process as we have discussed before will affect the cultural heritage of the country. To minimize the effect and to modernize the local language based on the developed language, construction of multilingual thesaurus is very important.

UNISIST provide a guideline for the establishment and development of multilingual thesauri and recognizes three possible approaches to the construction.

- 1. Begin from scratch and collect terms from all the languages involved, seeking equivalencies after the terms are collected*
- 2. Take an existing thesaurus in one language and translate it in to the other language*
- 3. Take two or more existing thesauri, in different languages, and merge them.*

Shuck (1977) as cited by Lancaster concluded that the first alternative is easier and more efficient guideline. The most difficult activity of construction of multilingual dictionary is finding a term equivalence of one language from another language.

2.4 Dictionary

Dictionaries have various purposes and formats. The most familiar form of dictionaries is its printed form; but due to the increasing needs of dictionaries, large numbers of them exist in electronic form. The Electronic dictionary, also called Machine-Readable Dictionary (MRD), is usually the typesetting computer tape of an existing paper based dictionary. In addition, electronic dictionaries may contain some lexicographic information like, information about how to prepare paper based dictionaries and the fonts and other special symbols and formats, that doesn't appear in the printed version.

Dictionaries will differ from one another in the type of information they contain, and in the way they describe the contents; but within a single dictionary consistency of formats for the information they hold is very important. Nowadays, there are a number of MRD constructed as an electronic version of the printed dictionaries. For example, the Oxford Advanced Learner's Dictionary (OALD), Long man Dictionary of Contemporary English (LDOCE), and the Collins Cobuild English Language Dictionary (COBUILD) have corresponding electronic versions (Pocock, 2000).

From the electronic version of dictionaries, various forms of display can be generated and will have various functions. According to Pocock, the range and variety of the information provided by any particular dictionary will depend on a number of factors:

- *Degree of coverage it is aimed at providing.*
- *Intended readership e.g. general, foreign learner, specialized.*
- *Specific purpose.*
- *Monolingual or bilingual.*

Both typographically and structurally, dictionaries are extremely complex. In addition, dictionaries interest many communities with different and sometimes conflicting goals.

According to Still (1992) as cited by Pocock, a reasonably sophisticated monolingual dictionary typically includes

Form of the headword (and spelling variants)

Part of speech: noun, verb, adjective etc.

Further grammatical information: countability, transitivity etc.

Inflectional variants: plurals, past tense, irregular forms etc.

Phonological: pronunciation and stress patterns.

Sense indication(s), by meaning definitions or a cross-reference to a synonym.

Examples of the different ways a word may be used.

Expressions and collocations in which the headword commonly occurs

Derivations, compounds etc

Usage notes, special uses etc.

According to the on line document information¹, during the process of developing EDR dictionary for translating Japanese language into English and vice versa, five different kinds of dictionaries were constructed. These are: Word dictionary, Bilingual dictionary, Concept dictionary, Co-occurrence dictionary, and, Technical Terminology dictionary including the EDR Corpus.

The word dictionary² consists of entry information, grammatical information, semantic information, and pragmatic and supplementary information. Entry information consists of the headword, the portion of the word that is invariable during inflection, the adjacency attribute pair, pronunciation, and kana notation for the Japanese Word Dictionary and the English Word Dictionary, syllable division. The headword describes the root form of a word

when it inflects regularly or the inflected form if it has an irregular inflection. Words that don't inflect are described as they are.

The invariable portion of the headword is the string of characters that is common to all the inflected forms of the word, and is not necessarily the word stem. This term is used synonymously with 'morpheme' in this document. Adjacency attributes are symbols that indicate the morphological restrictions on the adjacent words or morphemes in a sentence.

2.5 Machine Translation

Machine translation is the process of automatically translating a word, phrase, sentence or a document from one language into another language using machines (computers). The resulting output of the translation may or may not be equivalent in concepts. But it will provide basic information to the second parties that are not able to communicate with the others language.

According to Hedden (2000), there are four basic types of translation, three of which are types of machine translation or machine-assisted (-aided) translation:

¹ http://www.ijnet.or.jp/edr/E_Struct.html

2.5.1 Human translation

A human translator performs all the steps in the translation process, using a computer only as a word processor, if at all.

2.5.2 Machine-aided human translation (MAHT)

A human translator performs the translation of source language concept into target language, but he/she uses the computer as a tool to improve or speed up the translation process. This is called computer-aided translation (CAT) by people in the field of translation as opposed to the field of MT. This approach places the human translator at the center of the translation process, while attempting to provide him or her with tools that make the translation process more efficient and accurate.

Some of machine-assisted translation tools which are available nowadays range from word processors to translation editors, comparison programs, and terminology management programs. Translation editors identify selected terms and suggest translations. Comparison programs speed the updating of existing translations by finding changes between versions of source documents. Terminology management programs allow

² http://www.ijjnet.or.jp/edr/E_Struct-WD.html

translators to maintain flexible databases of terms in their subject area (Diego, 1996)

2.5.3 Human-assisted machine translation (HAMT)

The source language (SL) text is modified by a human translator either before, during, or after it is translated by the computer.

2.5.4 Fully automatic (automated) machine translation (FAMT)

The SL text is fed into the computer as a file, and the computer produces a translation automatically without any human intervention. This is sometimes referred to as batch mode. There are two types of fully automatic machine translation: fully automatic high-quality machine translation (FAHQMT) and low-quality machine translation.

According to Arnold (1995), there are two approaches that dominate machine translation research over most of the past twenty years. The first is inter lingual approach, where translation proceeds in two stages, by analyzing input sentences in some abstract and ideally languages independent meaning representation, from which translations in several different languages can potentially be produced.

The second is the so called transfer approach, where

translation proceeds in three stages, analyzing input sentences into a representation which still retains characteristics of the original source language text. This is then input to a special component called a transfer component that produces a representation that has characteristics of the target (output) language, and from which a target sentence can be produced.

One of the most important components of MT is the dictionary. Dictionary construction requires the sort of issues like designing, constructing, or modifying the dictionary.

According to Arnold (1995) [online document], MT has got social, political, commercial, scientific and intellectual or philosophical importance. Arnold describe these importance as follows:

2.5.5 The socio-political importance

Many countries are well known by having communities with multiple languages. The social or political importance of machine translation, according to Arnold (1995), arises from the importance of translation of concepts from one language to another and to keep the social and political stability of the country. In many countries the use and modernization of a

single common 'lingua franca' of the language is widely adopted, which is not nice-looking to the country community since it initiate the weakening of the other languages which ultimately leads to disappearance of the second language, creates disadvantage to the speakers of the second language for the success of communication with the lingua franca and other social and political problems.

The other issue is the vanishing of unique culture associated to the language speakers and the language itself, and the way of thinking will matter the society.

Therefore translation is the remedy to avoid these problems by facilitating ordinary human interaction and for gathering the information one needs to play a full part in society. Being allowed to express yourself in your own language, and to receive information that directly affects you in the same medium is very important to the society.

2.5.6 Commercial importance

The commercial importance of machine translation is a result of factors that are directly related to its social and political importance.

To justify some of the commercial importance of MT, the following can be considered as fact to the idea:

One can select product without language constraint (for example, if we need material to specific topic in the language of Amharic and if the best material is available in English, we can take it and it can be translated to the language of our interest).

Secondly, it will simplify the expensive activity of manual translation.

2.5.7 Scientific importance

According to Arnold(1995), *"it is an obvious application and testing ground for many ideas in Computer Science, Artificial Intelligence, and Linguistics, and some of the most important developments in these fields have begun in MT"*. He illustrate this using the first widely available logic programming language called prolog. Prolog was formed a key part of the Japanese 'Fifth Generation' program of research in the late 1980s, can be found in the 'Q-Systems' language, originally developed for MT (Arnold, 1995).

2.5.8 Philosophical importance

Philosophically (according to Arnold 1995), MT *"represents an attempt to automate an activity that can require the full range of human knowledge that is, for any piece of human knowledge, it is possible to think of a context where the*

knowledge is required". The extent to which one can automate translation is an indication of the extent to which one can automate 'thinking'.

Now a day there are several MT systems that are used in day-to-day use around the world. Some of these are: METEO (since 1977 used at the Canadian Meteorological Center in Dorval, Montreal), SYSTRAN (in use at the CEC, and elsewhere), LOGOS, ALPS, ENGSPAN (and SPANAM), METAL , GLOBALINK (Arnold, 1995).

Users consider many of MT system as useless based on the quality of the output of the MT system. Of course, it is true that the quality of many MT systems is low, and probably no existing system can produce really perfect translations. However, this does not make MT useless. First, not every translation has to be perfect since very rough translation of materials that doesn't give any sense to the user can provide much relevant information. Second, a human translator normally does not immediately produce a perfect translation. Therefore getting rough translation will facilitate human translation process.

The process of producing draft translations, along with the often tedious business of looking up unknown words in dictionaries, and ensuring terminological consistency, will

become automated, leaving human translators free to spend time on increasing clarity and improving style, and to translate more important and interesting documents.

According to Arnold (1995), Machine Translation started out with the hope and expectation that most of the work of translation could be handled by a system which contained all the information we find in a standard paper bilingual dictionary. Source language words would be replaced with their target language translational equivalents, as determined by the built-in dictionary, and where necessary the order of the words in the input sentences would be rearranged by special rules into something more characteristic of the target language. In effect, correct translations suitable for immediate use would be manufactured in two simple steps. This corresponds to the view that translation is nothing more than word substitution (determined by the dictionary) and reordering (determined by reordering rules)".

2.6 Multilingual dictionary

Multilingualism is a characteristic of tasks that involve the use of more than one natural language. This fact is most apparent in an increased need for translations and a consequent interest in alternatives to the traditional ways of producing them. The principal alternatives that have been proposed include partially or fully automatic translation, machine aids for translators, and fully or partially automated production of original parallel texts in several languages (Kay, Martin).

Chapter Three:

3 Thesaurus and database of the newly translated Amharic terms

3.1 Data source

The main sources of data used for the construction of the English word thesaurus and the electronic dictionary of the Amharic Science and technology mathematical terms are the Dictionary prepared by Science and technology term translation project for mathematical terms and the sample essays (book one) compiled to illustrate the terms which is prepared by STTTP. The dictionary provides science and technology English terms, its equivalent in Amharic, and definition of the Amharic terms in Amharic language where as the sample essay illustrates terms and its related concepts briefly. Knowledge of experts of the discipline is also a very important component during classification of term hierarchy and association. In addition textbooks of Algebra and analysis are referred to acquire more knowledge of the terms during classification and to make complete the process of acquisition of data.

3.2 Algorithm development

3.2.1 Method of thesaurus construction

According to Lancaster (1986), to construct a thesaurus while the terms are collected, each terms must be organized in to a coherent and cohesive structure. Similarly the mathematical terms are classified into three broad categories as Algebra, Analysis, and Geometry. Algebra, Geometry and Analysis are the three sub disciplines of the subject matter under consideration. In this process some terms belongs in to only one facet, some other in two or all facets. For example, the term "interval" can be grouped in both Algebra and analysis which is not recommended on printed thesaurus since it requires more spaces and gives partial information to the user at a time. For electronic thesaurus it may be also an indication to poorly constructed thesaurus. There are also terms that become difficult to set in any of the facets. For example the term "problem" doesn't belong to any of the defined facets.

Rearranging the hierarchies and association to come up with appropriate relationships among terms and to identify the kind of relationships between terms and facets solves the specified problems.

Once terms are classified in each of the defined facets, the next phase was putting the terms under different sub facets/headings to indicate similarities between terms of the sub disciplines. This process brings together terms that have some sort of relationships. This helps to generate the hierarchies of the terms and to indicate the association of terms.

The hierarchical relationship of terms of a thesaurus shows which are the broader and which are the narrower terms in the hierarchy. In addition to the BT/NT relationship there is also association relationship between terms shown in the thesaurus. The legitimate narrower term should have "a kind of" relationship to its BT.

One of the problem mentioned above, the problem of having more than one BT for a given term, is removed during the process of merging all the associations of terms together under the subject matter "Mathematics" by creating another preferred term entry.

3.3 Application of thesaurus in electronic dictionary

With out the use of the thesaurus, one can find basic information about the given concepts. But the information

for the concept may be either vague due to the complexity of the newly translated Amharic terms or limited (i.e. it may lack definition or illustration). To fully understand such concepts, the availability of the thesaurus is very important. In addition, it provides systematic approach to learn science and technology terms of mathematics. Therefore the narrower and related terms are used as "see also" information.

3.4 Database design

The purpose of the database in this project is to store, retrieve and maintain the subject matter. Any information seeker may use the system to find relevant information in the area of the discipline. The information that should be stored, retrieved, and maintained includes:

- ☐ Amharic equivalent meaning of the given English phrase or word;
- ☐ Amharic definition of the Amharic word or phrase;
- ☐ Amharic illustration of the word or phrase;
- ☐ Pictorial illustration of the term if necessary;
- ☐ A link to associated concepts;

To accomplish the above tasks, a database that organizes the different forms of the available data in a systematic fashion have to be constructed. These include database of the

thesaurus, and the dictionary. Therefore, the proposed table should include the following attributes for the success of the project.

Term code (TCode): A unique code given to each of the terms. A code indicates concepts of the term irrespective of the language; meaning a code, which is given to an English term, can be used to refer Amharic term equivalent of the English term.

English term (ETerm)

Stemmed English term (SETerm): are copies of English term where the phrasal English words are replaced with their equivalent stemmed word lists. For example the phrase "Angle measurement system" will be coded as "angl measur system" in the stemmed term code table. There may be more than one entry in the database that satisfies the user query. In such cases determination of the appropriate record entry is given to the user to determine the best information.

Amharic term equivalent (ATerm)

Amharic term definition(Def) (if any)

Amharic term illustration (Illustration) (if any)

Pictorial representation of the concepts (Picture) (if any)

Broader term of the concept(BTerm)

Narrower term of the concept (NTerm)

Related term of the concept(RTerm)

The following is a table showing all of the above-mentioned attributes together.

Table 3.1

<i>TCode</i>	<i>Eterm</i>	<i>ESTerm</i>	<i>ATerm</i>	<i>Def</i>	<i>Illustration</i>	<i>Picture</i>	<i>BTerm</i>	<i>NTerm</i>	<i>RTerm</i>
									<i>m</i>

To normalize this table, first we have to understand the various formats and natures of the data value to each of the attributes.

In this database, for a given term code that shows unique concepts, we can have different entries as Amharic term (synonymous), English term (synonymous), English stemmed term, narrower terms, and related terms.

In addition, one picture can be used by multiple concepts. It is also assumed in this database that for one concept we have only a maximum of one picture as an illustrative.

The following steps show the normalization process to achieve the normalized tables.

- A. The first normal form eliminates any repeating attributes in the entity set/table. In this case: for a given term code we may have one or more terms having the same concept in different languages. Since these terms should have the same term code, the table should be splitted as a table which hold list of preferred terms together with

their stemmed equivalent in English, list of non preferred terms (synonymous terms) together with their stemmed equivalent in English and the remaining part. Some of the Amharic terms have also synonymous terms that indicate further splitting of the table. As a result, the following table will be resulted.

Table 3.2

<u>TCode</u>	ETerm	SETerm

Table 3.3

<u>TCode</u>	ESTerm	SESTerm

Table 3.4

<u>TCode</u>	ASTerm

Table 3.5

<u><i>TCode</i></u>	<i>ATerm</i>	<i>Def</i>	<i>Illustration</i>	<i>Picture</i>	<i>BTerm</i>	<i>NTerm</i>	<i>RTerm</i>

In addition, in table 3.5, a given concept may have many narrower terms and related terms. This shows the table must be

splitted further as a table containing term code, Amharic term, definition and illustration to the Amharic term and table that holds the rest information. Since, if term X is registered as narrower term of term Y, then technically Y is the broader term of term X. therefore the existence of broader term is simply wastage of memory space. As a result, the following tables are created from the third table.

Table 3.6

<u><i>TCode</i></u>	<i>ATerm</i>	<i>Def</i>	<i>Illustration</i>

Table 3.7

<i>TCode</i>	<i>Picture</i>	<i>NTerm</i>	<i>RTerm</i>

Still table 3.7 is not in the first normal form because narrower terms and related terms are independent attribute. As a result, we have the following splitted tables:

Table 3.8

<i>TCode</i>	<i>Picture</i>	<u><i>NTerm</i></u>

Table 3.9

<i>TCode</i>	<u><i>RTerm</i></u>

- B. The second normal form Shows non-key attribute should depend on the entire key. All the tables shown above satisfy this rule.
- C. The third Normal form shows Non key attributes determined only by the primary key. All the tables shown above satisfy this rule except table 3.2, table 3.3, and table 3.8. In table 3.2 and table 3.3, the stemmed term fields are dependent on the preferred and the non-preferred term respectively and on the primary key. In table 3.8, the non-key attribute field picture, is determined by the non-key attribute TCode. As a result we should split the tables as follow.

Table 3.10

<u>TCode</u>	ETerm

Table 3.11

<u>TCode</u>	SETerm

Table 3.12

<u>TCode</u>	ESTerm

Table 3.13

<u>TCode</u>	SESTerm

Table 3.14

<u>TCode</u>	<u>NTerm</u>

Table 3.15

<u>TCode</u>	<u>Picture</u>

D. The above tables satisfy the Boyce-Codd normal form rule. Finally, by summarizing all the tables from table 3.1 to table 3.15, the normalized lists of tables are shown bellow together with the appropriate table label.

Term Code entity

<u>TCode</u>	ETerm

English synonymous entity

<u>TCode</u>	ESTerm

Amharic synonymous entity

<u>TCode</u>	ASTerm

Stemmed English term entity

<u>TCode</u>	SETerm

Stemmed English Synonymous term entity

<u>TCode</u>	SESTerm

Dictionary entity

<u>TCode</u>	ATerm	Def	Illustration

Picture entity

<u>TCode</u>	Picture

Term Hierarchy entity

TCode	NTerm

Term Relation entity

TCode	<u>RTerm</u>

In general, nine entities are involved in this database for the success of the project.

Since all terms (English/Amharic) have unique term codes and term code can be a numeric data, which is simple to process,

we can replace related terms, and narrower terms by their corresponding term codes. As a result the entity TRelation and Thierarchy described above become:

THierarchy **entity**

<i>TCode</i>	NTCode

Where NTCode is Narrower Term Code

TRelation entity

TCode	RTCode

Where RTCode is Related Term Code

3.4.1 Term Code entity

To increase performance and to utilize memory space, searching can be easily done with the help of numeric values rather than texts. Text processing is very tedious and cumbersome. Therefore in the entity Dictionary, every science and technology mathematical terms that are translated by the STTP will have its own numeric term code.

The data dictionary of this entity is

Attribute	Acronym	Data type	size
PTerm	Preferred Term	text	50
Tcode	Term Code	number	long

PTerm is required for each record.

PTerm is a word or phrase that holds science and technology concepts to the discipline of mathematic in one of the foreign language English. TCode is chosen to be the primary key of this entity even though both are candidate keys.

3.4.2 Thierarchy entity

One of the facilities that will give us terms with "associated" concept is, those that are the species or the genius of the given term. The term hierarchy entity also shows the genus/species relationship between the available terms. In this design, one term can have only one genus (BT Broader term) and any number of species (NT narrower terms).

The data dictionary of this entity is

Attribute	Acronym	Data type	size
BTCode	Broader Term Code	Number	long
NTCode	Narrower Term Code	Number	long

Both field values are required.

For a given term, one can easily find its NTCode as well as BTCode, which directly leads to its broader and narrower terms and their synonymous.

The primary key to this entity is NTCode because one broad term may have a number of narrower terms and one narrower term has only one broader term.

3.4.3 TRelation entity

The other facility that we have, to access terms of similar concept, is use of the association relationship of terms. The association relationship shows the existence of similarity between terms not in all aspect but in some context. It is a reciprocal relationship.

The data dictionary of this entity is

Attribute	Acronym	Data type	size
TCode	Term Code	Number	long
RTCode	Related Term Code	Number	long

Both field values are required.

The primary keys of this entity are both TCode and RTCode because one term can be a related term of any number of terms and it may have any number of related terms.

3.4.4 Dictionary entity

The basic information that assists any interested group is contained in the entity called Dictionary. This entity holds

the English term equivalent code, the Amharic term, the definition and illustration of the concept in Amharic language.

The data dictionary of this entity is

Attribute	Acronym	Data type	size
ETCode	English term Code Number		long
ATerm	Amharic Term	text	50
Def	Definition	memo	
Illustration	illustration	memo	

ATerm value is required for all records.

Primary key of this entity is the ETCode attribute.

3.4.5 Picture entity

Some of the terms can be best illustrated with help of pictures/graphs. this entity store the location of the picture file so that one can easily access the information to assist concept understanding.

The data dictionary of this entity is

Attribute	Acronym	Data type	size
TCode	term code	Number	long
Picture	Picture	text	10

Picture field value is required for each record.

TCode is the primary key of the entity.

3.4.6 **ATSynonymous entity**

Some of the translated concepts have two or more Amharic equivalents. Only one of them (the preferred term) will be taken as an attribute in the dictionary entity. The other related synonymous terms will be listed in this entity.

The data dictionary of this entity is

Attribute	Acronym	Data type	size
Tcode	term code	number	long
ASTerm	Amharic	text	50
	synonym term		

TCode field value is required for each record.

ASTerm is the primary key of the entity because one or more synonymous terms may be available to one preferred term.

3.4.7 **ETSynonymous entity**

Some science and technology concepts have two or more equivalent English terms. Only one of them (the preferred term) will be taken as an attribute in the term code entity. The remaining (non preferred term list) terms are registered in this entity.

The data dictionary of this entity is

Attribute	Acronym	Data type	size
Tcode	term code	number	long
ESTerm	English	text	50
	synonym term		

TCode field value is required for each record.

ESTerm is the primary key of the entity because one of more synonymous terms may be available to one preferred term.

3.4.8 SETCode entity

For ease of processing input query, the system use stemmed term rather than the phrases that are associated to the concepts in science and technology. The stemmed term equivalent for the preferred terms will be stored in this entity.

The data dictionary of this entity is

Attribute	Acronym	Data type	size
TCode	Term code	Number	
SETerm	Stemmed English Term	Text	50

SETerm field value is not required for each record. It will be processed automatically at the time of user request.

TCode is chosen as a primary key of this entity by the same reason specified for the entity TermCode.

3.4.9 SESTCode entity

Similarly, the stemmed terms of the non-preferred term are kept in this entity.

The data dictionary of this entity is

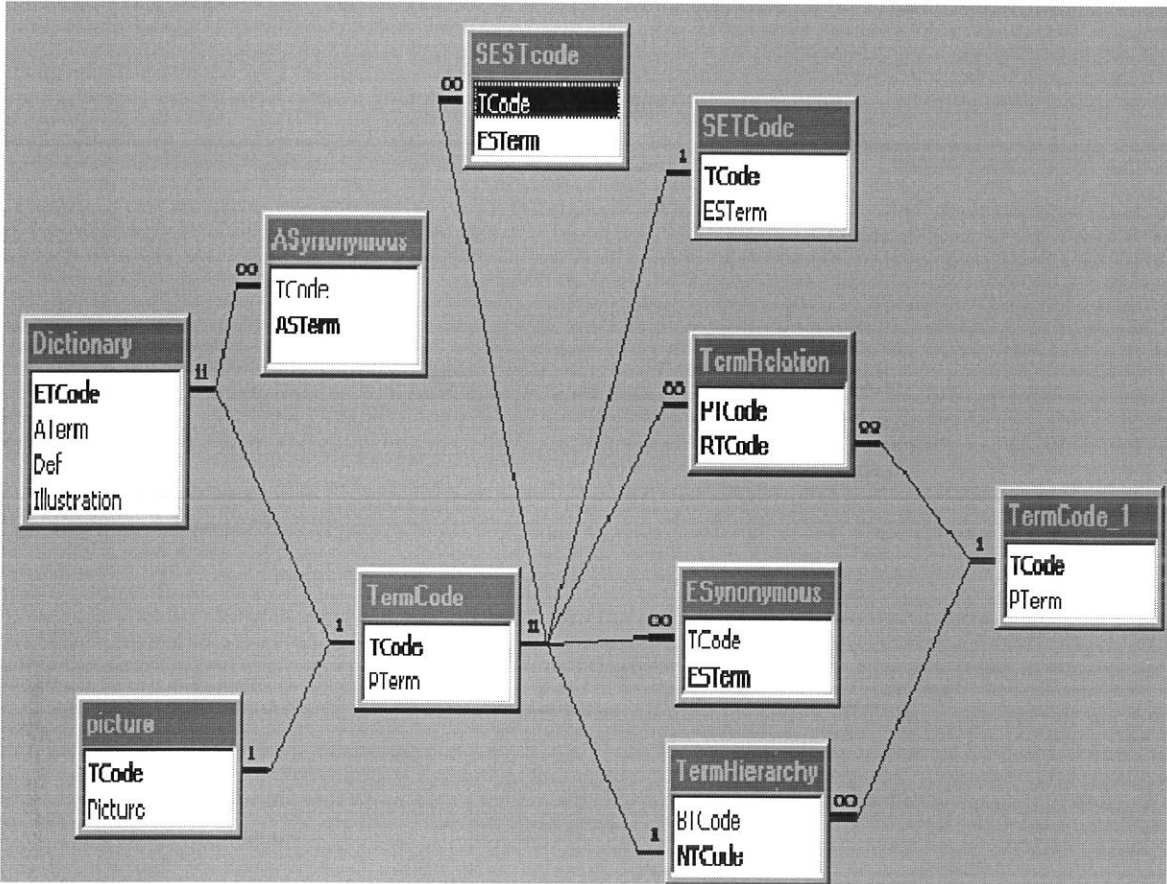
Attribute	Acronym	Data type	size
TCode	Term code	Number	
SESTerm	Stemmed English Synonymous Term	Text	50

SESTerm field value is required for each record. The record will be processed automatically at the time of user request. The primary key of this entity is SESTerm by the same reason specified for the entity ETSynonymous.

3.5 Entity relationship (E-R) diagram

A record in the TCode table can be related to exactly one record of The SETCode table and zero or one record of the TPicture and dictionary table. A record in the TCode table can be related to zero or more records of THierarchy, TRelation, ETSynonymous, and SESTCode tables. A record in the dictionary table is also related to zero or more records of the ATSynonymous table. In general the E-R diagram, constructed using MS-Access, is shown bellow.

fig 3.1



3.6 Referential integrity constraints

Any English term in the term code table has unique numeric value. Any numeric value in any of the other entity has direct relation to the numeric data in the term code table. Therefore, unless the code is registered in the term code table, it has to be controlled registry of the record in any other tables with unknown code. To register new record in any table, first we have to be sure that the English term is registered in the term code table with the appropriate term code. So that we will have the relevant code to register records in each tables. In addition, a record in ATSynonymous table, registry of a record requires preferred term in the dictionary table.

3.7 Phrases matching and processing

A phrase is a language construct with specific syntactic and semantic properties. The phrase component should occur in a common context within the document or query to which the phrase is assigned as content identifier and represent broader concepts (Salton, 1983).

In an information retrieval system supported by a database, records have fixed set of fields and well-organized data formats. The database management systems are concerned about the storage, maintenance, and retrieval of data facts from the database.

Every information retrieval system consists of a set of information items (documents), a set of requests (queries) and mechanism of determining which document meets the requirements of the query (Salton, 1983). In this electronic dictionary, the user provide the request in the form of phrases that specifically identify single concept in the field of mathematics. Each entry of the database field with field name Pterm in the table Term code, lists all possible concepts of the subject mater. In addition to these ideally the various elements of the thesaurus are also unique concepts of mathematics.

The process of automatically searching similarity between the input term and the database information requires phrase level term matching.

To avoid the different formats of a word and non-concept bearing terms, the Porter's word stemming algorithm that improves the recall of the retrieval process is used. Using the stemming algorithm, every word of a given phrase will be converted in to their corresponding stem forms. Porter's stemming algorithm will return blank if the target word is a stop word or a single character word, other wise it returns the stem form of the word by avoiding suffixes and prefixes.

Stemming algorithm is used to reduce the variations of forms of words. It makes various words that share the same stem looks identical. This algorithm improves the recall of a retrieval system by increasing the probability of term matching.

In processing this database, every non-stop word term of the input streams are treated as key terms in the process of term matching with the document in the database and the input phrase. In addition, the English terms of the term code table, excluding the stop word list, can be treated as a control index to the database information (document). One module is constructed to filter non-stop word terms from the given phrase or phrases.

Weighting stemmed words of the input query and the document is very important to generate the rank of the document to the request. The process of weighting the input query term and the database document is shown bellow. In a database system, the input phrases will be automatically stemmed. These stemmed input phrases are free from stop words and each stemmed words have equal importance in the searching process. Therefore each stemmed words in a phrase have identical weight.

If phrase_j is input phrase consists of stem words {W_{1i}, W_{2i}, ...W_{ki}} then

$W_{ij} = 1/K$ Where

W_{ij} is the weight of $stem_i$ in stem phrase $_j$

K is the total number of stemmed words in the stem phrase $_j$

Similarly, the weight of a stem word in the document is given by:

$W_{ij} = N_i/N$ where

W_{ij} is the weight of $stem_i$ in stem phrase $_j$

N is the total number of stemmed word $_i$ in the stem document

N_i the total number of stem words in the document

Since the stored document in the database is a simple phrase, the number of similar stem words in the document (the value of a field in the database) is one.

Therefore $N_i = 1$ and

$$W_{ij} = 1/N$$

To formulate the similarity between the input query phrase and the phrase in the database, the following conditions are taken into consideration.

- Similarity of the phrase is directly proportional to the number of common stem words.

- Similarity of the phrases is inversely proportional to the absolute value of the difference between the number of words in the input query phrase and the database phrase.

Based on the above assumption, the similarity between the stemmed phrases of the query and the phrase of the database can be estimated by the function

$$\text{Sim}(\text{phrase}_i, \text{phrase}_j) = (CM) / (NK) (|N-K|+1)$$

Where

phrase_i is the phrase in the database

phrase_j is the phrase in the query term

C is the number of common stem words in the phrases

N is the number of stem words in phrase_i

K is the number of stem words in phrase_j

$M = \text{Min}(N, K)$

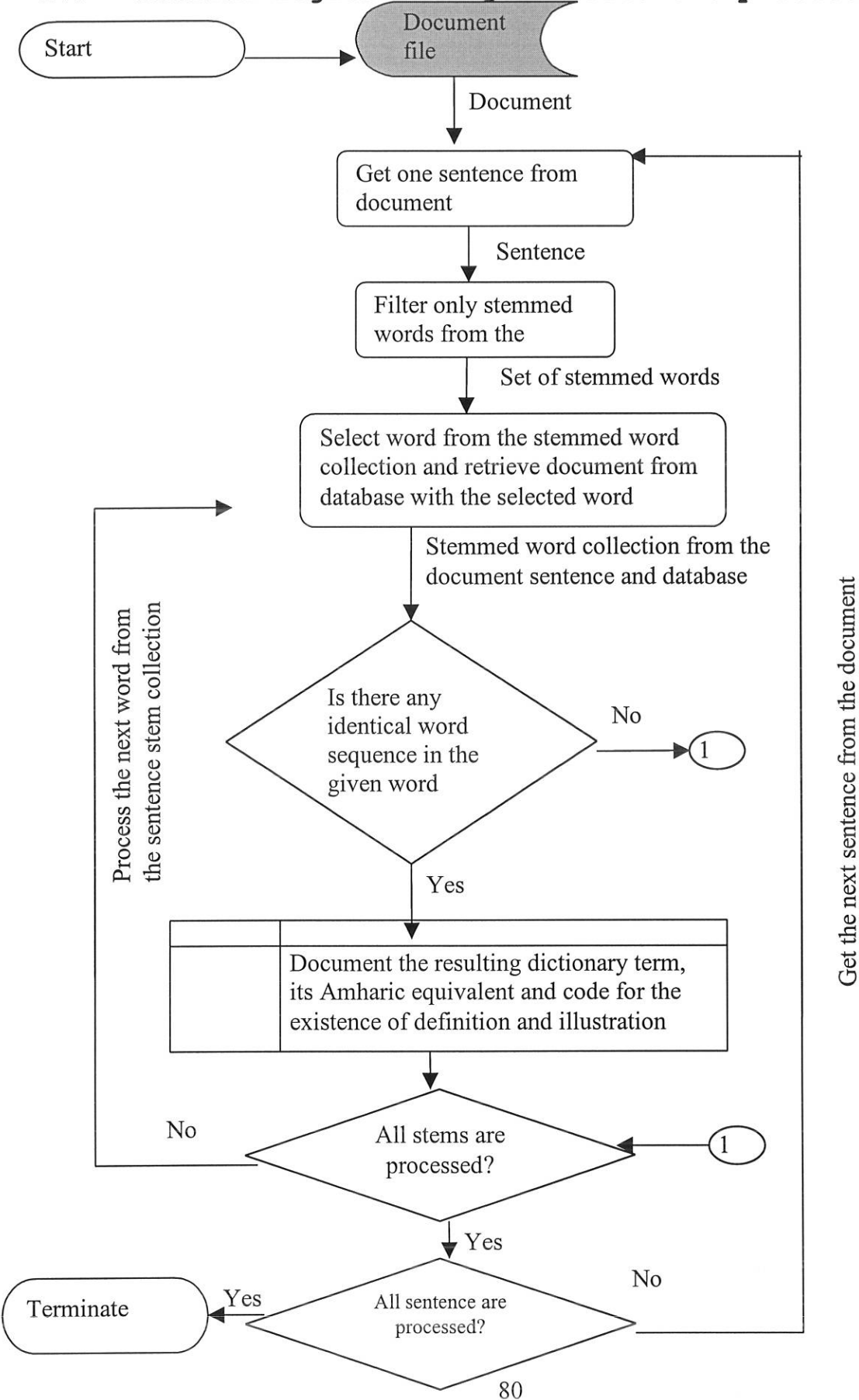
If the value of the similarity function is one, then there is strong relationship between the input phrase and the database term. The related information will be displayed to the user automatically, if there is only one phrase that satisfies the condition. However if there are more than one phrase that has similarity coefficient value 1 to the input phrase or if no record entry has similarity value 1 and some have non zero

similarity, selection will be made by the user to confirm the requirement of the user and the search result.

3.8 Document processing

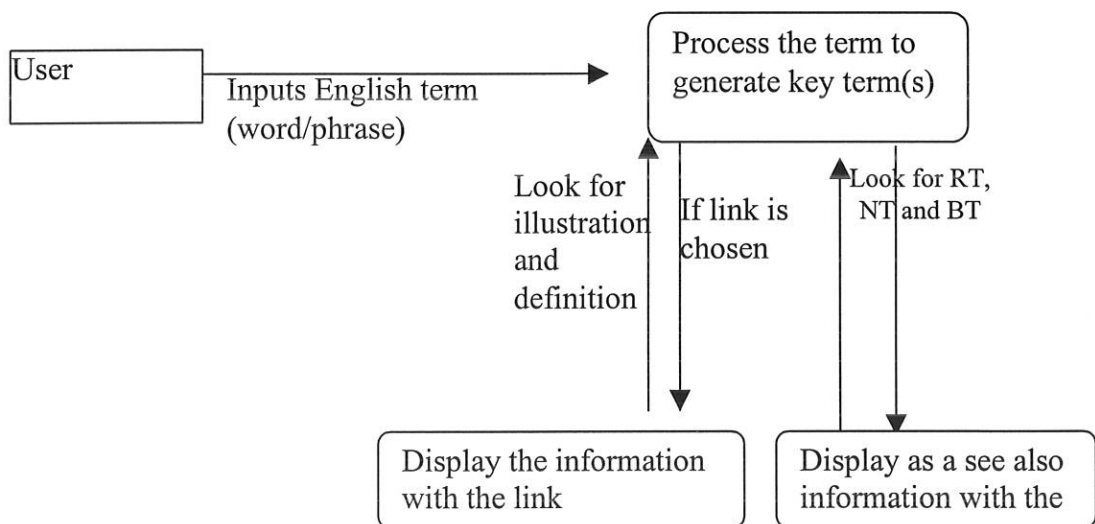
Another method of supporting users in the process of translating science and technology documents from English into Amharic is by providing Amharic equivalent of the key terms of the English document. The process of retrieving concept bearing terms from document is not a simple process. However, to retrieve some of basic key terms with their Amharic equivalent and other basic references to the user, the system takes a sentence and the system will process according to the following logical design.

3.9 General logical design to document processing



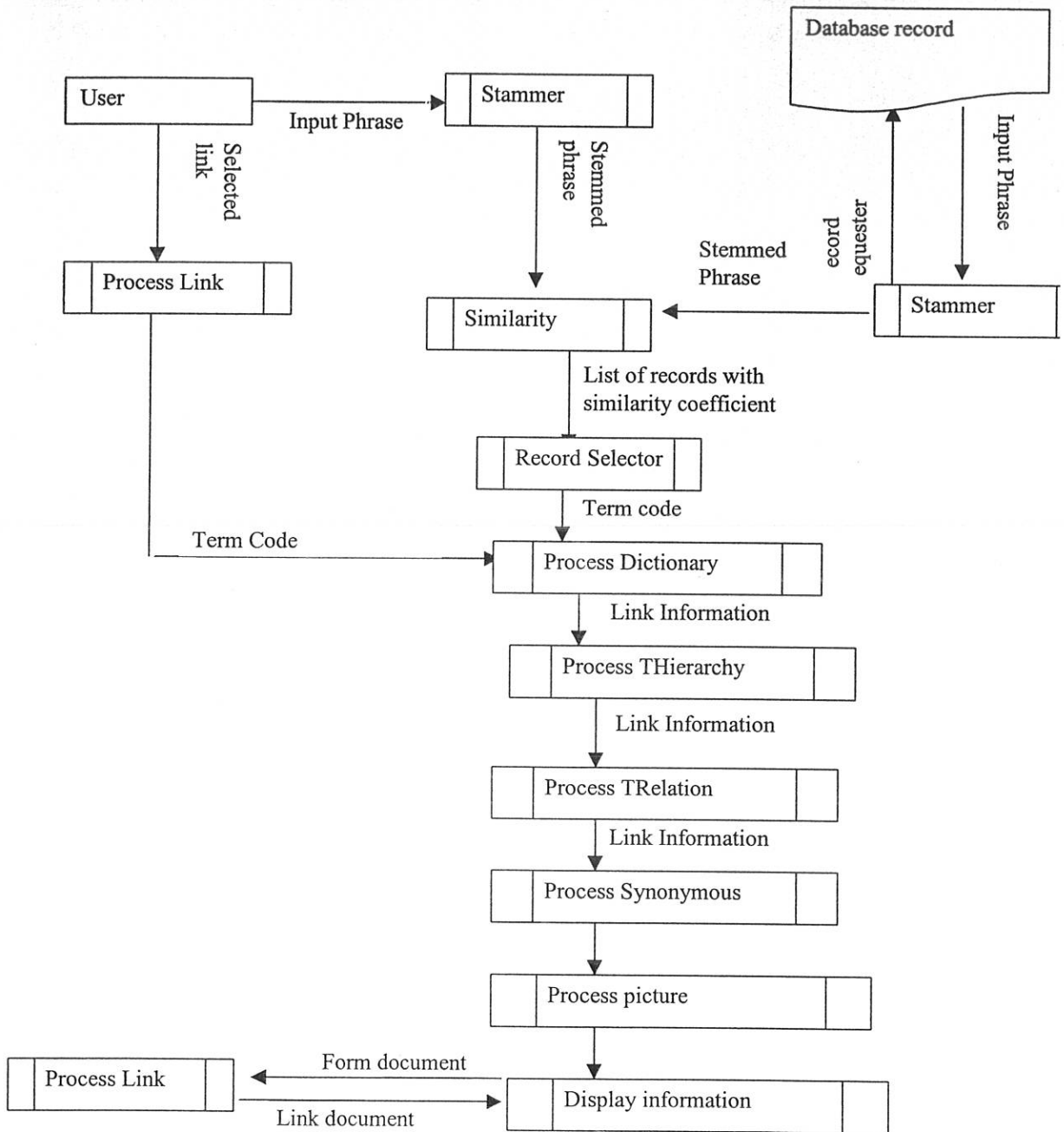
Sentence in a document, in this system, is defined as a group of words, symbols and other character sets that terminate with a period. Even though this definition seems incomplete, it partially helps us to identify whether or not two words belongs to the same sentence or not.

3.10 General logical design of phrase processing



3.11

Detailed logical design of phrase processing



3.12 Amharic software

Beside the various problems especially in line with standardization, electronic processing of the Amharic document takes one part of the prototype development.

The Amharic software that I used to process the Amharic document is the Power Ge'ez 2000™, which has various weak and strong side. Power Ge'ez 2000™ is a Windows-based software that allows users to do word-processing in Ethiopic fonts. The software is designed to work with all types of Windows applications. Once installed onto your system, it will reside in memory making itself readily available to all active Windows applications.

One of the strong side of power Ge'ez is its ability to read and convert documents written in different Ethiopic fonts.

Chapter Four:

4 Program Development

4.1 System input

The system input of the sample prototype can be classified in to two as document input and phrase input. In the case of Document input, the system automatically detect any list of key terms from the document and retrieve their Amharic equivalent for readers to understand the concepts with in the document. However, the phrasal input, by which the user provide to the system at the time he/she request some information about single science and technology concepts.

4.1.1 Phrasal input

The input English term is a key term (can be a word or phrase) that the professional and/or information seekers want to have the Amharic conceptual equivalence, its definition and illustrations in Amharic. This term can be as simple as a morpheme or as complex as a phrases or some times as sentence. But to map the input to the concepts in the database there are various procedures and functions developed by scholars and researchers. Porter's Stemming algorithm is used in the process of mapping the inputs to the document that is available on the database. Porter's stemming algorithm code is available on the Internet using Pascal and Java programming

language. I have used this programming logic to code the stemming algorithm in visual basic. Array implementation is used to list out possible stop word lists, prefixes, suffixes and the method of morpheme modification.

The general algorithm of the main module is shown bellow:

If ***there is input English phrase*** then

- Avoid the stop words from the input English phrase and find out the corresponding stem words to each phrase
- Open the database to look for related information
- For each stem word, execute the query that select records from term code table that includes words like the stem word followed by any character in the field ETerm.
- To the retrieved list of records, find out the similarity between the input phrase and the phrase retrieved from the record, which helps us to return the appropriate term code.
- Sort the similarity coefficient in descending order.
- Display if there is unique database document with similarity coefficient one otherwise display the retrieved list with similarity different from zero.

The program code is shown bellow.

```
Private Sub CMDProcess_Click()  
Dim DBEDictionary As Database  
Dim temp As String, word As String  
Dim keyword As String  
Dim qry As QueryDef  
Dim qryTerm As Recordset  
Dim count As Integer  
    SimWord.last = 0  
If ((frmMainForm.txtEnTerm.text <> AETerm(cindex)) And (frmMainForm.txtEnTerm.text  
<> "")) Then
```

```

Call Library.InitializeVariable
EnTerm = frmMainForm.txtEnTerm.text
keyword = EnTerm
temp = keyword
Call Stemming.stripAffixesFromPhrase(temp)

While (temp <> "")
    Call Library.getWord(temp, word)
    word = Stemming.stripAffixes(word) & "*"
    Set DBEDictionary = OpenDatabase("c:\dictionary97.mdb")
    Set qry = DBEDictionary.CreateQueryDef("", "SELECT SETCode.* FROM SETCode
WHERE SETCode.ESTerm like '" & word & "'")
    Set qryTerm = qry.OpenRecordset
    Call processCode(qryTerm, keyword)
Wend

If SimWord.last >= 1 Then
    Call ModuleSorting.quickSort(SimWord, 1, SimWord.last)
    count = ModuleSorting.NbStemWithCoefOne(SimWord)
    If count = 1 Then
        Call ProcessDictionary(SimWord.lists(1).code, True)
    ElseIf count > 1 Then
        Call frmAlternative.ShowAlternative(count)
    Else
        Call frmAlternative.ShowAlternative(SimWord.last)
    End If
Else
    ' look for in the synonymous term table
    temp = keyword
    Call Stemming.stripAffixesFromPhrase(temp)
    While (temp <> "")
        Call Library.getWord(temp, word)
        word = Stemming.stripAffixes(word) & "*"
        Set DBEDictionary = OpenDatabase("c:\dictionary97.mdb")
        Set qry = DBEDictionary.CreateQueryDef("", "SELECT SESTCode.* FROM
SESTCode WHERE SESTCode.ESTerm like '" & word & "'")
        Set qryTerm = qry.OpenRecordset
        Call processCode(qryTerm, keyword)
    Wend
    If SimWord.last >= 1 Then
        Call ModuleSorting.quickSort(SimWord, 1, SimWord.last)
        count = ModuleSorting.NbStemWithCoefOne(SimWord)
        If count = 1 Then
            Call ProcessDictionary(SimWord.lists(1).code, True)
        ElseIf count > 1 Then
            Call frmAlternative.ShowAlternative(count)
        Else
            Call frmAlternative.ShowAlternative(SimWord.last)
        End If
    End If

```

```

Else
    MsgBox "not found in the dictionary"
End If
End If
End If
End Sub

```

4.2 Document file input

Occasionally, users may need to provide the document file to the system, so that the system will retrieve relevant information to the user by identifying the list of key words that are available within the document. These can help the users to identify the key terms and their Amharic equivalent and a remark on the presence and absence of additional information like definition and illustration.

The process is made by identifying one sentence of the document file and repeats the process for all sentences of the document. The code that manipulate a sentence is shown bellow:

```

Sub processSentence(sentence As String)
Dim word As String
Dim i As Integer
Dim pos As String
Dim temp As String
Dim foundphrase As String
    Call stripAffixesFromPhrase(sentence)
    temp = Trim(sentence)
    SimWord.last = 0
While (temp <> "")
    Call Library.getWord(temp, word)
    word = word & "*"
    Set DBEDictionary = OpenDatabase("c:\dictionary97.mdb")
    Set qry = DBEDictionary.CreateQueryDef("", "SELECT SETCode.*

```

```

&_
    FROM SETCode WHERE SETCode.ESTerm like ' ' & word & ' ')
    Set qryTerm = qry.OpenRecordset
    Call processCode(qryTerm, sentence)
Wend

If (SimWord.last <> 0) Then
    For i = 1 To SimWord.last
        SimWord.lists(i).Similarity = &_
        Similarity.countWord(SimWord.lists(i).name)
    Next i
    Call ModuleSorting.quickSort(SimWord, 1, SimWord.last)
    For i = 1 To SimWord.last
        foundphrase = Trim(subSentence(sentence,
SimWord.lists(i).name))
        If foundphrase <> "" Then
            pos = InStr(sentence, foundphrase)
            If Not foundbefore(keyWordFound, SimWord.lists(i).name)
Then
                keyWordFound.last = keyWordFound.last + 1
                keyWordFound.list(keyWordFound.last).name = &_
                SimWord.lists(i).name
                keyWordFound.list(keyWordFound.last).code = &_
                SimWord.lists(i).code
            End If
        End If
    Next i
End If
End Sub

```

4.3 Record selection from term code table

Three different cases are encountered in the process of identifying record that has some sort of similarity with input queries. The three cases are indications to the relevance of the document to the query. These are:

1. The query stem words could be mapped to exactly one record in the database table with similarity coefficient one. In such cases this record and its associate will be displayed to the information seeker.

The assumption is that, since there is only one record in the appropriate table that match the input query, this should answer the user needs. Otherwise the user will get his/her information need from the list of possible links.

2. The query stem words could be mapped to more than one record in the database table with a similarity coefficient of one. This shows the existence of many related concepts so that the user should select the best key phrase/word from the list of options displayed in the alternative form. The selected item with its associated information will be displayed.
3. The query stem words couldn't be mapped to any number of records in the database table with similarity coefficient one. However, there may be list of terms from the database with similarity coefficient greater than zero and less than one. This indicates the absence of exact match term and presence of at least a term that have some sort of similarity. This may also be encountered due to spelling error, use of synonymous term, which doesn't exist in the database. In such condition, all terms which have non zero similarity coefficient will be displayed to the user in descending order of their similarity so that the user can select his need if exist or it gives a chance to reformulate the query.

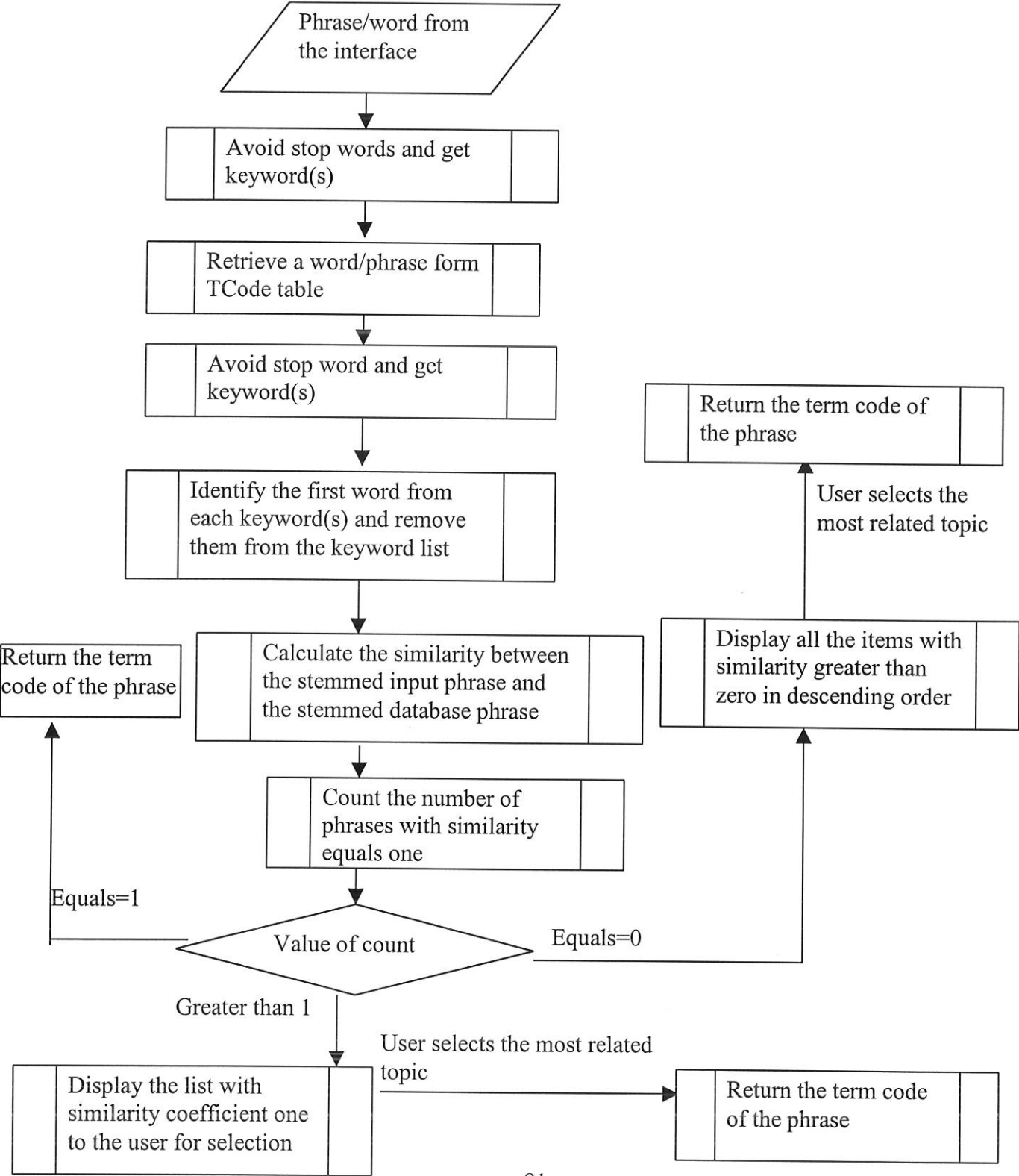
In the process of searching appropriate record to a query, different forms of query can be used to refer to various records. For example, in the complex number representation system, $a + bi$, there are two parts: the real and the imaginary part. To retrieve information about the imaginary part, the input key word may have the following forms:

- ☐ imaginary part
- ☐ imaginary number
- ☐ imaginary part of complex number
- ☐ the imaginary number of a complex number
- ☐ imaginary side of complex number
- ☐ complex number
- ☐ number etc

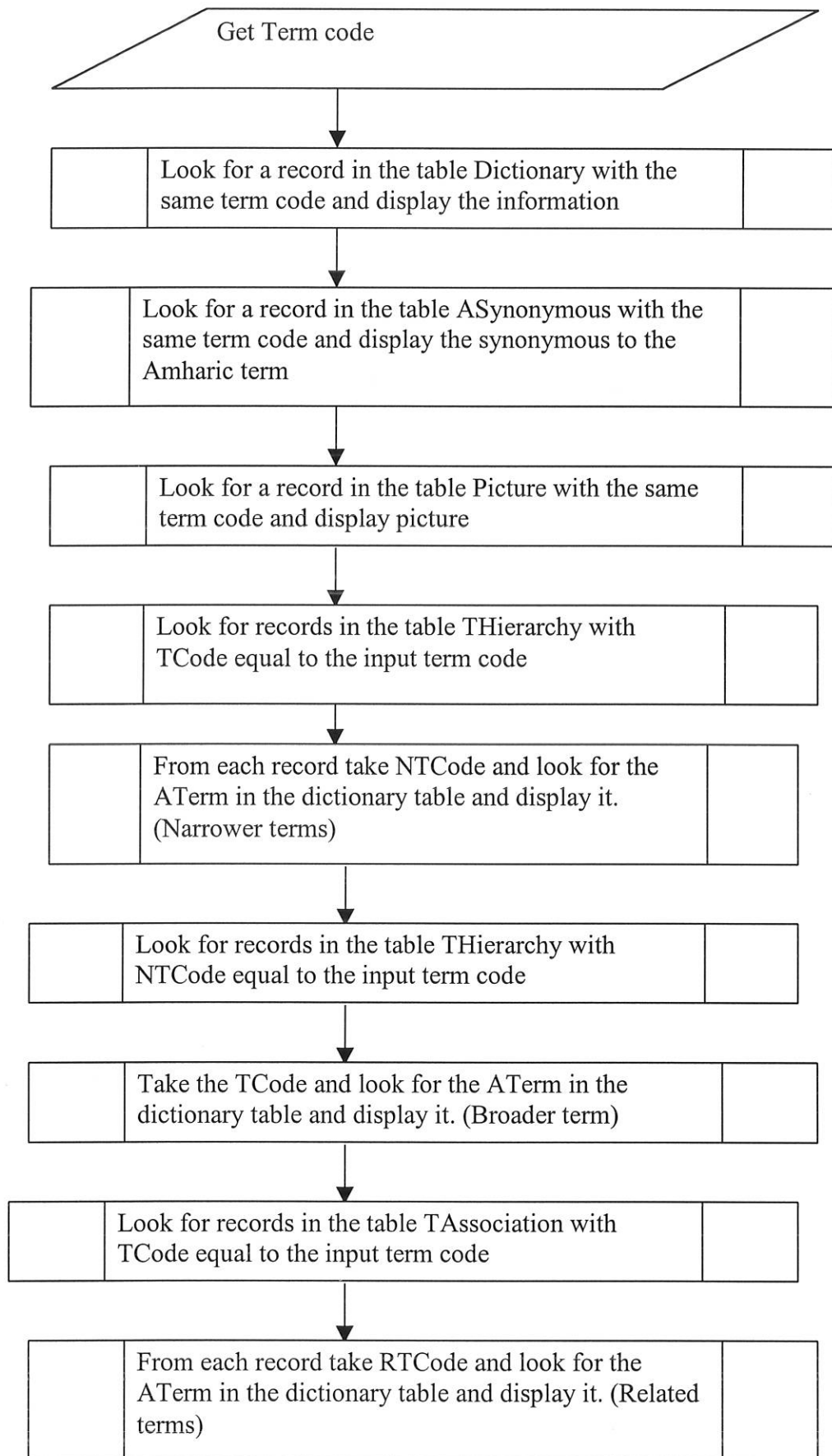
The first five are almost conceptually identical that are represented by only one preferred term (imaginary number) in the **term code** table. These phrases have at least one word in common 'Imaginary' which can be an indication to their similarity and to the concept the user need to retrieve. Using only one word may not be fully informative since imaginary can be used to show either axis or numbers or others. Furthermore, some time there may be various occurrences that do not have any common word, in such cases it may require further word processing, concept matching and other process. The last two

are wider in context, which can be easily retrieved through term relation.

4.4 Logical design to English input word processing to get the term code



Once the term code is achieved for the English key word, then it become simple to retrieve the basic information from the dictionary table that includes Amharic equivalence, definition to the Amharic term, and illustration. The supplementary information as picture and the see also information should be retrieved from their corresponding table. The logical design on how to process the database to retrieve the information is shown bellow.



The function "process dictionary" is the visual basic implementation of the logical design. The following code shows how this function process the data in the database.

Sub ProcessDictionary(AmTermCode As Long, WithEnTerm As Boolean)

If AmTermCode <> -1 Then *' indicate absence of related information in the database*

If Not WithEnTerm Then

Call InitializeVariable *' initialize public variable*

Enterm = ProcessTermCodeTable(AmTermCode) *' return the English term*

End If

Call ProcessDictionaryTable(AmTermCode) *' set AmTerm, Definition and illustration*

Call ProcessPictureTable(AmTermCode) *' set file name for the picture*

Call ProcessRealtionTable(AmTermCode) *' set see also information*

Call DevelopLinkInformation(definition, txtDefinition) *' create the document link information*

Call DevelopLinkInformation(illustration, txtIllustration) *' shows the link to the user*

Call refreshMyForm

Call DisplaySeeAlso(cindex + 1) *' show related term, narrower term and association*

Call SetDefaultColor

Call UpdateStack

Call ShowLink(txtDefinition)

Call ShowLink(txtIllustration)

Else

MsgBox "Sorry We don't have such information"

End If

End Sub

After the required information is displayed on the output screen we have to process the output information to make the output more transparent. A link should be developed to

phrases/word that have further elaboration which includes the see also information, selected words/phrases in the definition/ illustration. These processes also require Amharic document processing like stemming, indexing, morpheme identification and others to automatically identify which phrase/word has entry point in the database.

Even though there are researches already conducted in the area of automatic Amharic word stemming, indexing, and processing, it becomes very difficult to find it and incorporate into this project. However to the development of the prototype, manual concept-tagging method is used to create links for the Amharic document, which can serve to elaborate the linked item.

Now a day, automatic text processing of Amharic document is at its infant stage. As a result, automatic identification of Amharic texts from a given document that can be linked to another document in the database is very difficult. Therefore manual method of concept identification and tagging is used in the process of development of the prototype.

To elaborate the difficulty of automatic concept identification, assume we have a phrase that consists of three words. Assume also all the words together, each word independently, and all possible combination of two words have associated concepts in the database. In this case to

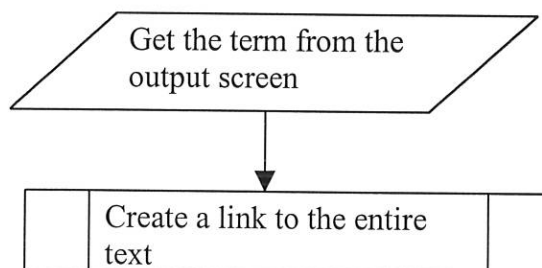
automatically select which word(s) should have a link, we should have a mechanism that analyze the semantic of the Amharic document and other morphological analysis are required. Therefore the manual tagging method should be used with out hesitation even though it is tiresome and bulky.

The symbol ጽ (š) is used as a starting tag and ጽ (ÍD) is used as a closing tag for the Amharic key terms having associated information to be linked. The link code associated to the document to be retrieved when the user clicks on the link is attached immediately the closing tag separated by blank space.

If there are two or more similar items in a document, which has an entry in the database, only the first item will be tagged for link development.

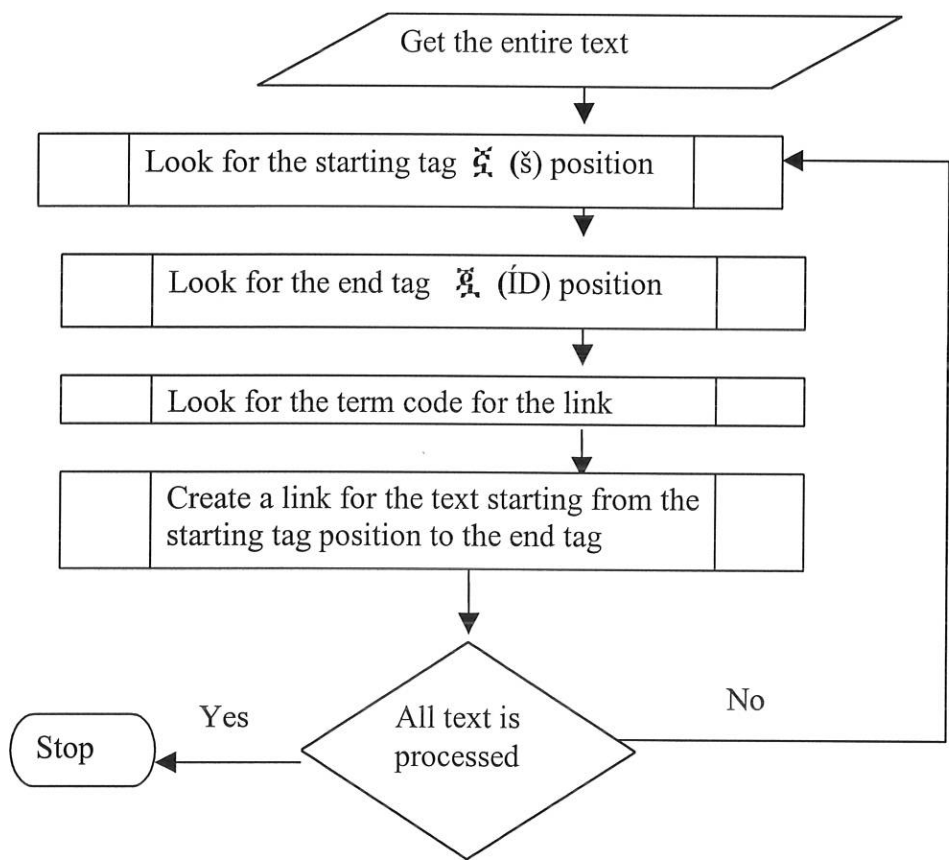
4.5 Logical design of creating a link to Amharic document

A) For see also information



The see also information are the broader, narrower, and related concepts of the target item. To avoid confusion, each concept has different link color.

B) For definition/illustration



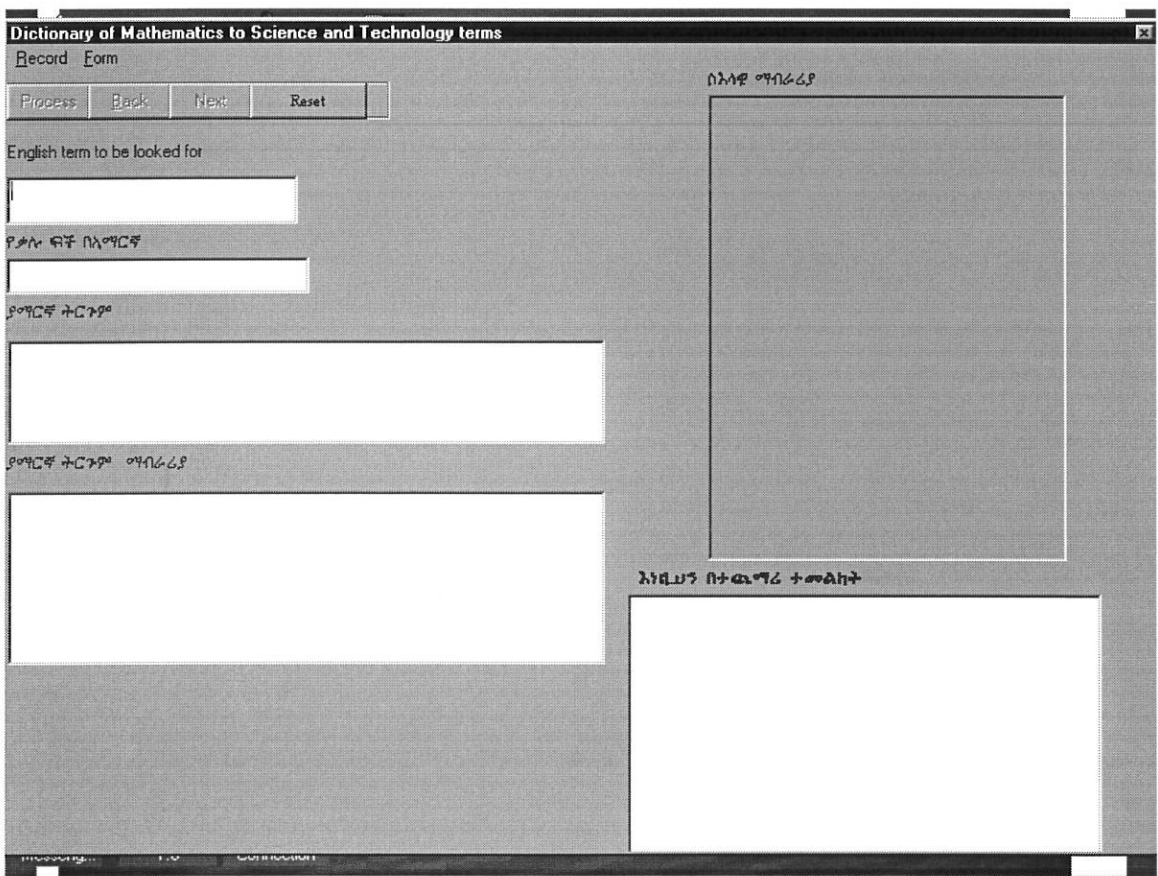
4.6 The user interface

The user interface is the interface that the user request information from the stored database and the output information will be displayed to the user. Before displaying the output information to the input query, there may be

different dialog boxes displayed to confirm the relevance of the out put information to the user inquiry.

The main dialog box accepts user request in the form of English phrase and display the relevant information to the request. It has various kinds of objects like buttons, menus and edit boxes as shown bellow.

Fig 4.1

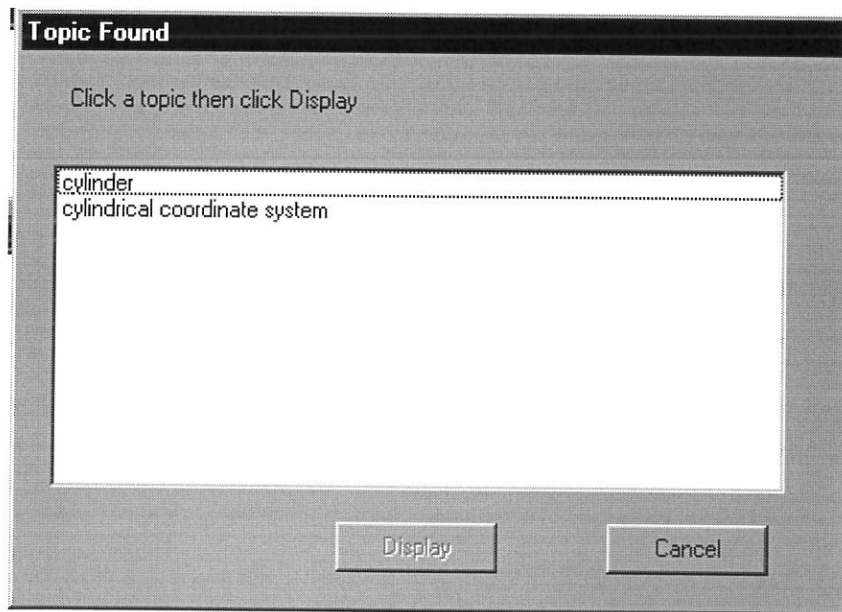


If necessary, confirmation of message window will be displayed with list of possible options available in the database. Some of the confirmation dialog boxes are:

- If the searching mechanism found out two or more equally important records in the database all

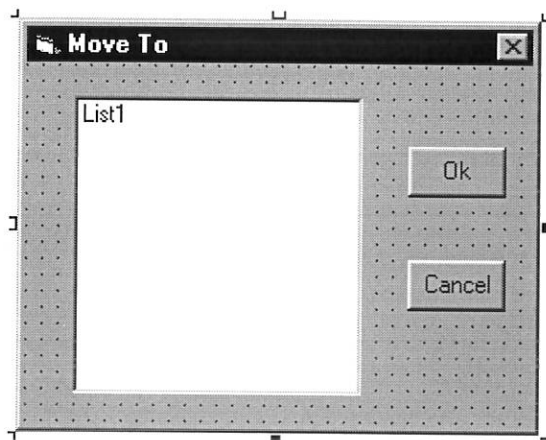
possible records that satisfy the condition will be displayed on the form frmTopFound. This form also be used in the condition where no records are found with similarity value one but if there is any record with similarity value greater than zero.

Fig 4.2



- If the user wants to move to specific page which is already browsed by the user (frmMoveTo form)

Fig 4.2



4.7 Sample test result

4.7.1 Document file processing result

The database consists of 846 concept-bearing terms, out of which 18 terms are synonymous terms. That means there are 826 unique English concept bearing terms. Out of these, 793 (96%) have Amharic equivalent, 303(36.7%) have Definitions and 93 (11.25%) have illustrations. The high percentage of terms without definitions and illustrations indicates that a lot should be done in the process of elaborating the terms already translated into Amharic language. Each term is placed as a narrower term and /or related term in the term association hierarchy. The prototype program is designed to provide relevant information to the user of the system. It provides Amharic equivalents to terms, definitions and illustrations to the Amharic term. the system is tested in two ways: by providing text document and simple phrases to the system.

The sample input document is the second unit of the document from grade 9 textbook. This document contains various terms related to the main topic "set". From my observation the systems identifies almost all terms that are registered in the term database and provides the Amharic equivalents to the terms and inform the presence and absence of definition and illustration.

From the input file containing 2409 words, which contains various kinds of mathematical terms, 65 different science and

technology terms (phrases/words) together with their Amharic equivalent are retrieved. The retrieved terms consist of 100 words, which is 4.15% of the total words in the document. This shows by average from every 24 words in the document there is one science and technology term. Even though there is no module that checks the semantics of the term, from the list of the retrieved document only one term semantically provides invalid meaning, namely the term unit. Unit is defined in mathematics as a standard to define measurements, however, the term is used to mean beginning of new chapter. Including this term, the following statistics are observed in line with the presence and absence of definition and illustration to the Amharic term.

Table 4.1

Term definition		Term illustration	
Absent	Present	Absent	Present
33 (50.8%)	32 (49.2)	40 (61.5%)	25 (38.5%)

4.7.2 Performance evaluation of the system for user query

To evaluate the performance of the system for the input query, 30 different queries are taken from 3 different subject specialists.

The list of queries can represent the entire population because 7 of them have 3 concept bearing words, 12 of them have 2 concept bearing word and 11 of them have one concept bearing words. Therefore there are about 54 concept-bearing words in the sample queries. This is 6.5% of the total unique concept bearing terms registered in the database system.

The list of queries and summary of the system response is given bellow:

Table 4.2

	Query	System response			
		Proper Amharic meaning	Definition	Illustration	# of Alternatives
1	What is limit?	Yes	No	No	1
2	What is Neighborhood?	No	No	No	1
3	What is continuity?	Yes	No	No	1
4	What is partion?	Yes	No	No	1
5	What is sequence?	Yes	Yes	No	1
6	What is convergence?	Yes	No	No	2
7	What is integral?	Yes	No	No	1
8	Give me information a bout test of convergence?				6
	For convergence test	Yes	Yes	No	
9	What is divergence?	Yes	No	No	1
10	What is Regular partition?	No	No	No	5
11	What is the relation between group and field				6
	For field	Yes	No	No	
	For Group	Yes	Yes	No	
12	What is a division ring?	Yes	No	No	1
13	What do mean by similar triangle?				4
	For similarity	Yes	No	Yes	
	For Triangle	Yes	No	Yes	

14	What is a limit of a function?				2
	For Function	Yes	Yes	Yes	
	For limit	Yes	No	No	
15	Give some explanation about derivative				3
	For derivative	Yes	No	No	
16	What do mean countable set?				9
	For Set	Yes	Yes	Yes	
17	What is a parallelogram?	Yes	Yes	No	1
18	Is Trapezium a parallelogram?				2
	For Parallelogram	Yes	Yes	No	
	For Trapezium	Yes	Yes	No	
19	Compare circle and polygon				5
	For circle	Yes	No	No	
	For polygon	Yes	Yes	yes	
20	Is natural numbers integer?				8
	For natural number	Yes	No	No	
	For Integer	Yes	Yes	yes	
21	What is a field?	Yes	No	Yes	1
22	What is a sub group?	Yes	No	No	1
23	What is a cosset?	No	No	No	1
24	What is an algebraic Structure?				11
	For structure	Not sufficient	No	No	
25	What is a group homomorphism?				4
	For group	Yes	Yes	No	
	For homomorphism	Yes	No	No	
26	What is an algebraic numbers?	No	No	No	13
27	What is a uniformly convergent function?				7
	For Uniformly continuous function	Yes	No	No	
	For convergent	Yes	No	No	
28	What is a simple ordered set?				4
	For Ordered field	Yes	No	No	
	For Ordered group	Yes	No	No	
	For Ordered pair	Yes	Yes	Yes	

29	What is a double angle formula?	Yes	No	No	1
30	What is a left unitary module?	No	No	No	3

To each of the queries, the system response is classified in to five based on the accuracy of the output to the query as:

High with accuracy value 3, when the system provides appropriate meaning with or without related information to the entire request.

Satisfactory with accuracy value 2, when the system provides appropriate meaning to each of the concept bearing words/phrases with or without related information to each concept and the user can easily grasp some idea of his need.

Fair with accuracy value 1, when the system provides appropriate meaning to some of the concept bearing words/phrases with or without related information to each concept.

Poor with accuracy value 0, when the system doesn't provides any information at all or the information is irrelevant.

Table 4.3

Performance	Number of concept bearing words in a query				
	1	2	3	Total	Average
3	9 (82%)	3 (25%)	2 (28.6%)	14 (46.6%)	4.7
2	0 (0%)	5 (42%)	2 (28.6%)	7 (23.3%)	2.3
1	0 (0%)	2 (17%)	2 (28.6%)	4 (13.3%)	1.3
0	2 (18%)	2 (17%)	1 (14.3%)	5 (16.6%)	1.7
Total	11 (100%)	12 (101%)	7 (100.1%)	30 (99.8%)	

From this data, we can summarize the following conclusions:

1. The system provides highly accurate information to the query if the query is a single word query (82%)
2. The system provides satisfactory information to the query if it consists of two concept-bearing words (42%)
3. To all type of queries, the system provides basic information (highly accurate (46.6%) or satisfactory (23.3%))
4. High probability of poorly accurate output is observed in the case of a single concept query. This is the condition where the term is not registered at all in the system.

4.7.3 Verification of the formula

$$\text{Sim}(\text{phrase}_i, \text{phrase}_j) = (\text{CM}) / ((\text{NK}) (|\text{N}-\text{K}|+1))$$

This process is used for verification of the above formula. In the process of retrieving relevant information, if there is no intermediary alternative list, it is an indication of strong similarity between the input query and the database information.

Three input queries are avoided from the above table (table 4.4), since there is no record in the database that has at least one common stem word. Precision is zero if no match occurred.

From the given list of queries, which have similarity coefficient value one, there is one query with poor precision measure. This is because the term is registered in the database without any Amharic equivalence.

To prove or disprove the existence of significant association between the similarity coefficient and the precision measure, the following hypotheses are taken.

H_0 = there is no significant association between the variables.

H_1 = there is significant association between the variables.

The following frequency is observed for all possible combination of similarity coefficient and precision measures.

Table 4.4

Similarity	High	Satisfactory	Fair	Poor	Total
1	12	0	0	1	13
0.66667	1	1	0	0	2
0.5	1	10	2	0	13
0.33333	0	3	6	0	9
Total	14	14	8	1	37

From this table we can see that highest percentage of highly accurate document is retrieved with similarity value of one

(85.7%). Using the Chi Square test, the existence of association between the variables similarity coefficient and presison measure can be shown as follow.

$$\mathbf{X}_{cal}^2 = \sum \sum (\mathbf{O}_{ij} - \mathbf{E}_{ij})^2 / \mathbf{E}_{ij}$$

Where

$\mathbf{X}_{cal}^2$ is calculated chi square value

$\mathbf{O}_{ij}$ is the observed frequency

$\mathbf{E}_{ij}$ is the expected frequency

i is row value

j is coulumn value

$$\mathbf{E}_{ij} = (\mathbf{f}_i * \mathbf{f}_j) / \mathbf{f}_{tot}$$

Where

$\mathbf{f}_i$ is the total frequency of row i

$\mathbf{f}_j$ is the total frequency of column j

From the above data, $\mathbf{X}_{cal}^2$ equals = 40.86264

Total degree of freedom = 9.

The tabulated chi square value at $\alpha = 0.05$ for the above degree of freedom is:

$$\mathbf{X}_{tab}^2 = 16.919$$

Therefore $\mathbf{X}_{cal}^2 > \mathbf{X}_{tab}^2$

This shows:

The null hypothesis is not valid so that we have to accept the alternative, which says there is significant relation between the similarity coefficient and the precision measure.

Chapter five

5 Conclusions and recommendations

5.1 Conclusion

The development of local language is of great importance to many aspects of the country. This can be achieved by doing a lot of related researches and experiments that direct the way to make local language developed in line with science and technology. The existence of science and technology materials written in the language of the society will provoke the interest of the society to read and write simple/complex documents using their language which is an indication of modernization of the language.

This experiment shows the possibility of using science and technology facilities, like computers, that can help professionals of mathematics and interested groups in the process of modernizing the local language (Amharic).

It is also indicated that the similarity coefficient between two phrases is a vital indication of the importance of the output information to the input query.

The importance of the output information decreases as the number of concept bearing words in the input query increase.

5.2 Recommendation

The designed system doesn't consider the concepts of standard mathematical symbols during the process of identifying keywords from input document file. In addition, the research is limited to only one science and technology discipline "mathematics". The first recommendation is to invoke researchers to test the possibility of automating the above listed problems.

The next recommendation concentrates on how to make this system functional. The system provides basic concepts to information seekers. However, some of the database fields have no value especially for those fields that provide further illustration like definition and illustration of the Amharic equivalent.

Therefore, it is strongly recommended to design a project that makes this system functional.

6 References

1. Abraham Demoz. (1968) "Amharic for modern usage" Ethiopian Journal of Education. Addis Ababa.
2. Amsalu Aklilu. (1983). "Principle for the creation of New Science and Technology terms in Amharic." Wiesbaden: Otto Harrassowitz.
3. Amsalu Aklilu. (1986). Academy of Ethiopian Language Science and Technology Term Translation Project "The linguistic theory and methodology of the science and technology term translation project" International seminar on terminology: Addis Ababa.
4. Anber, Taghrid. (1987). The Academy of Ethiopian Languages Science and Technology Term Translation Project. "Egypt's Experience in language planning." Proceedings of the international seminar on terminology.
5. Arnold D J (1995). Why MT Matters <http://clwww.essex.ac.uk/~doug/book/node5.html>
6. Arnold D J (1995). Interlinguas <http://cslu.cse.ogi.edu/HLTsurvey/node36.html>
7. Assefa G/Mariam. (1982). Ethiopian Studies. "Technical terms in Amharic: Problems, solutions." Proceedings of 7th international Conference.
8. Bisrat Dilnessahu (1987). The Academy of Ethiopian Languages Science and Technology Term Translation Project, Proceedings of the international seminar on terminology
9. Brook, Terrence A. (1998). Orthography as a fundamental impediment to on line information retrieval. Journal of the American Society for Information Science . 49 (8).
10. የኢትዮጵያ ቋንቋዎች አካዳሚና በኢ.አ.ዩ ቋንቋዎች ኢንስቲትዩት. (1972 ዓ.ም). "Which translation in which language would each citizen of this country understand best." የትርጉም ሴሚናር ከግንቦት 14-16 ቀን አድስ አበባ
11. የኢትዮጵያ ቋንቋዎች አካዳሚና በኢ.አ.ዩ ቋንቋዎች ኢንስቲትዩት. (1972 ዓ.ም) "የማስተማሪያ መጽሀፍትን በማዘጋጀት በኩል የተደረጉ ጥረቶችና ችግሮች፡፡" የትርጉም ሴሚናር ከግንቦት 14-16 ቀን አድስ አበባ

12. Christine, McNab. From traditional practice to current policy: The changing pattern of language use in Ethiopia.
13. Diego, San (1996).
http://homes.line.ru/vrl/mt_typs.htm
14. Ermias Dagne and Demissu Gemed. (1974E.C.). Science and Technology Term Translation Project: 2nd translation workshop Report. Addis Ababa
15. Ermias Dagne and Demissu Gemed. (1987). Amharic Science and Technology Terminology Project: An Overview; A paper presented to the international seminar on terminology, June 16-18.
16. Girma Yilma. (1987). The Academy of Ethiopian Languages Science and Technology Term Translation Project. "Opening Address." Proceedings of the international seminar on terminology
17. He, Shaoyi. (1998). "Concept similarity and conceptual information alteration via English-Chinese translation of Medical article titles." Journal of the American Society for Information Science Vol49 (2)
18. Hedden, Thomas D. (2000). Machine translation: a brief introduction [On line]
19. http://www.iiijnet.or.jp/edr/E_Struct.htm
20. <http://www.sil.org/sociolx/LangDev.htm>
21. http://www.iiijnet.or.jp/edr/E_Struct-WD.html
22. Jordan, Sova R. (1977). "Computerized Russian Translation at ORNL." Journal of the American Society for Information Science . 28 (1).
23. Kay, Martin. Overview.
<http://cslu.cse.ogi.edu/HLTsurvey/ch8node5.html>. Xerox Palo Alto Research Center, Palo Alto, California, USA
24. Lancaster, F.W. (1986). Vocabulary control for information retrieval. Second edition. Arlington: Information resource press.
25. Leslalu, Wolf (1972). While working on the English-Amharic Context Dictionary

26. Livonen, Mirja and Sonnenwald, Dane H. (1998). From Translation to Navigation of different discourse: A model of search term selection during the pre-online stage of search process. . Journal of the American Society for Information Science. 49(4).
27. Losee, Robert M.; Haas Stephanie W. (1995). "Sub language terms: dictionary, usage, and Automatic Classification." Journal of the American Society for Information Science. 46 (7).
28. Martin Kay <http://cslu.cse.ogi.edu/HLTsurvey/ch8node3.html>
29. Massamba, D.P.B. (1987). The Academy of Ethiopian Languages Science and Technology Term Translation Project. "Terminology Development: the Tanzanian experience." Proceedings of the international seminar on terminology.
30. Muddamalle, Manikya Rao (1998). Natural Language versus Controlled vocabulary in information Retrieval: A case study in soil Mechanics. Journal of the American Society for Information Science . 49(10).
31. Ngra, E.A (1976). Language styudy and The African University: The necessity for change. Bulletin of The association of African University. Vol 3(1).
32. Nguyen, Cecile et al. (1998). Education in Ethiopia: Are children getting what they deserve? [online].
33. Pocock, R.J.(2000). <http://www.comp.leeds.ac.uk/atibcbs/ai5/mrd/mrd.htm>. "Machine readable Dictionaries and Lexical Databases."
34. Pocock R.J. [online]. Machine-Readable Dictionaries And Lexical Databases.
35. Rahaingoson, Heneri. (1987). The Academy of Ethiopian Languages Science and Technology Term Translation Project. "The Malagasy experience in terminology development." Proceedings of the international seminar on terminology.
36. Salton, gerald and McGill, Michael J. (1983). Introduction to modern information retrieval. New York, McGraw-Hill.
37. The Structure of the EDR Electronic Dictionary [On

line]

38. The University of Western Ontario. (1998). Thesaurus Construction [online]
39. Van Horn, Christopher and Tilman, Dave. (1997, July 8) Education in Ethiopia [Online].
40. Yanni, Yigal. (1987). The Academy of Ethiopian Languages Science and Technology Term Translation Project. "The experience of the of the Academy of Hebrew language in terminological Dictionary making." Proceedings of the international seminar on terminology.

DECLARATRIION

THIS THESIS IS MY ORIGINAL WORK AND HAS NOT BEEN PRESENTED FOR
A DEGREE IN ANY OTHER UNIVERSITY.



SEBSIBE HAILEMARIAM

THE THESIS HAS BEEN SUBMITTED FOR EXAMINATION WITH MY APPROVAL
AS A UNIVERSITY ADVISOR

DR. DEMISSU GEMEDA

ATO DEREJE TEFERI

