Addis Ababa University
School of Graduate Studies
Electrical and Computer Engineering Department

MACHINE TRANSLATION SYSTEM FOR AMHARIC TEXT TO ETHIOPIAN SIGN LANGUAGE

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering

By
DAGNACHEW FELEKE WOLDE
October 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

ADDIS ABABA INSTITUTE OF TECHNOLOGY

Machine Translation System for Amharic Text to Ethiopian Sign Language

By

Dagnachew Feleke Wolde

Approval by board of examiners

______________________________	________________
Chairman Dept. graduate committee	Signature
Dr. Eneyew Adugna	________________
Advisor	Signature
______________________________	________________
Internal Examiner	Signature
______________________________	________________
External examiner	Signature

Acknowledgements

Thank you GOD.

I thank you GOD for giving me a wonderful advisor, Dr. Eneyew Adugna, who is willing to hear problems I faced during this research work and saw things in a different angle to tackle them. Especially when things look impossible, YOU put the words of wisdom and encouragement in his heart and mouth. I should also give thanks to YOU about John Glauert of University of East Angelia, who helped on technical issues regarding SiGML Service Player.

Thank you GOD for giving me a family who cares about me more than I do for myself.

Thank you GOD for the friends surrounded me. I wonder how much YOU cared about me when I think of Getnet Bogale, a friend who reminds me my late father (without him I can't think of my postgraduate study); Ezana Adugna, more eager than me to see the fruition of this thesis (and proof-read the draft); Chibo Fikre, matured more than his age (helped me by giving much needed materials); and about others who gave me encouraging words.

Thank you GOD for giving me an opportunity to met Mekane Yesus School for the Deaf staffs. Specially, Assefa Guta and Abiso Wolde (who gave their time to bring change for Deaf students). When I thought of halting this research, YOU speak through them to continue my endeavor.

Thank you GOD about Eyasu Hailu (Ethiopian Sign Language expert), who didn't hesitate to throw words of appreciations and suggestion on future work. Kidane Admasu, Demsachew Workie (both from AAU, linguistics dept.), and Zeleke H/Giorgis (from Menilik preparatory school), who evaluated the system performance are also YOUR gifts to me.

Thank you GOD about Liye, my soul mate. She had been with me at difficult times and her daily words of encouragements showed the bright future YOU planned for me.

Thank you GOD.

Table of Contents

List of Tables

List of figures

List of Appendices

List of Acronyms

AAiT	Addis Ababa institute of Technology
AAU	Addis Ababa University
ALPAC	Automatic Language Processing Advisory Committee
ASL	American Sign Language
BSL	British Sign Language
EBMT	Example Based Machine Translation
ECSA	Ethiopian Central Statistical Authority
eSIGN	Essential Sign Language Information on Government Networks
ESL	Ethiopian Sign Language
ETV	Ethiopian Television
HamNoSys	Hamburg Notification System
HMS	Human Modeling and Simulation
HOLME	Hand-shape, Orientation, Location, Movement, and facial Expression
LFG	Lexical Functional Grammar
MIT	Massachusetts Institute of Technology
MT	Machine Translation
MYSD	Mekane Yesus School for the Deaf-Hossana
NLP	Natural Language Processing
RBMT	Rule Based Machine Translation
SERA	System for Ethiopic Representation in ASCII
SiGML	Signing Gesture Markup Language
SL	Sign Language
SMT	Statistical Machine Translation
ViSiCAST	Virtual Signing: Capture, Animation, Storage and Transmission

Abstract

Machine Translation (MT) system development from one language to another is one of the areas in human computer interaction researches for the last half century. In this thesis the first Amharic to Ethiopian Sign Language (ESL) MT system was designed and implemented for limited vocabulary.

The MT system function is to change or translate a given Amharic text words to their equivalent representation in ESL. To achieve this goal the translation engine should have subsystems to analyze source language (Amharic) and synthesize target language (ESL). Due to close relation of source and target language, in addition to the system is designed for limited vocabulary, Rule Based Machine Translation (RBMT) approach is deployed in this research. Since there was no enough information on grammar of ESL, this thesis scope is limited to morphological level translation; i.e. direct RBMT architecture was followed.

While designing Amharic language morphological analyzer subsystem, it was found that Amharic alphabets (*fidel*) rearrangement based on their phonetic behavior gave a much needed pattern to apply expansion rules. In this thesis, a new arrangement of *fidel* was proposed to be used in computer based researches on Amharic language which involve morphology or phonology.

Amharic language is morphologically rich language, so it's not easy to prepare bilingual dictionary with all possible word derivations. To overcome this problem morphological expansion method was used. In this particular case 40 rules were applied on 20 verbs and 3 rules applied on 10 nouns to get a dictionary with entries more than 800.

Many Sign Language(SL) translation researches, nowadays use universal transcription methods; HamNoSys and SiGML. This thesis also applied these transcription tools for the first time on ESL. For a randomly selected fifty words and all alphabets in ESL, the transcription was done.

The system was designed to map a given Amharic word to ESL equivalent words using morphological analysis and synthesis. If the word is not in the bilingual dictionary, it'll finger spell rather than ignoring it. And performance of the system was found to be more than 80%.

keywords- Sign Language, ESL, Amharic morphology, machine translation, Deaf.

Chapter 1

Introduction

1.1 Background

Ethiopian Sign Language (ESL) is the natural language used primarily by about a million Ethiopian Deaf Community. ESL, as other known Sign Languages (SLs), is accepted as minority language, which coexist with majority languages [22] and is native language for many deaf people. The languages fundamental units are base signs that correspond to traditional notion of morpheme. SLs are composed of three-dimensional (3D) manual and non-manual features. Hand-shape, hand orientation, hand position, and hand movement are the manual components and the non-manual features include position of upper torso, mouth pattern, head and shoulder movement, facial expression and eye-gaze.

Although there are many aged modern and traditional schools for spoken languages in Ethiopia, before 1980's there was no means for the deaf individuals to go to school and, at least, to develop their communication skill. But in 1980s Mekanisa Deaf School, the first deaf/signers school in Ethiopia was opened and ESL with the current structure started then. Later in 1992 the first book for the signers was published by Ethiopian Evangelical Church Mekane Yesus – School for the Deaf School (MYSD) - Hossana, which is intended to help Deaf individuals and their family on how to “speak” using hands and body gesture and “listen” using eyes.

Even though this approach helps the signers to have their own native language, they were marginalized and their communication scope is limited to very narrow geographic area; i.e. where there are signers’ schools. After finishing their elementary and secondary school, many of them get in to ordinary (spoken language user) high schools; and without interpreters, they tried to cope up with non-signers by speech-reading and reading text.

They depended on human interpreters to use the information-communication system and mass-media. But by the year 2001 Ethiopian Television (ETV), which was the only television channel of the country by then, started 20 minutes weekly program for signers with news in sign language and some muted-movies.

Nowadays deaf individuals go to ordinary school, even at the elementary level. There is daily news for ESL users in ETV, and above all the government and public awareness towards sign language and Deaf individuals is promising. One indication of this is the starting of ESL class by Linguistics department of AAU in academic year 2007/2008 (the author of this paper is one of the first students to take this 3 months course). Recently linguistic department of AAU has started a 4 year degree program on ESL, and the first batch will graduate at the end of this academic year.

1.2 Problem Statement

“The purpose of the World Programme of Action concerning Disabled Persons is to promote effective measures for prevention of disability, rehabilitation and the realization of the goals of "full participation" of disabled persons in social life and development, and of "equality". This means opportunities equal to those of the whole population and an equal share in the improvement in living conditions resulting from social and economic development. These concepts shou1d apply with the same scope and with the same urgency to all countries, regardless of their level of development.” [31]. This is the objective of world programme of action concerning disabled persons. This is taken from the UN document regarding disabled persons, printed in 1984. But after more than quarter century in our country Ethiopia the objective is far from reality.

With all above (Introduction part) mentioned efforts, the deaf population is still marginalized. Their contribution to the development of the country is limited regardless of their will and ability. Even though they’ve the right to access information and express their feelings, the environmental setting will not allow them. During the critical language-acquisition years of childhood, they will not start their to-be first language, in this case ESL [2]. Deaf individuals of rural areas, and some urban areas have to go away from their family in order to get education (this problem forced many to go to school at later age) if they are lucky enough (In an informal discussion with MYSD vice-director each year hundreds of deaf students apply to join the school, but they accept only around forty students at preparatory class). They will not participate in community service except in some churches and special programs where interpreters are assigned.

Even though there are many researches done and are ongoing in all higher institutes of Ethiopia, little is towards solving this problem. In half century old AAU, Faculty of Technology (now named Addis Ababa Institute of Technology); the number of research works done so far is not more than three including this one.

There is a video based text to ESL converter developed for limited vocabulary but as will be shown in coming chapter that approach is expensive, both production wise and maintenance wise.

This thesis will try to alleviate some of the problems by developing Amharic text to ESL conversion system. The system will accept Amharic text and outputs its ESL equivalent in Signing Gesture Markup Language (SiGML) which will drive the SiGML Service Player which is found for research purpose at School of Information Systems, University of East Anglia, Norwich, United Kingdom (http://vhg.cmp.uea.ac.uk). The software is aimed to substitute human interpreters to some extent.

One may argue it’s easy to communicate by writing text with the signer. But researches, for English-ASL, showed differently. Studies in United Sates (so far no study done on ESL-Amharic) have shown that deaf high school graduate (approximately 18 years old) have only a fourth grade (approximately 10 years old) English reading level [15].

1.3 Objective

General Objective

The objective of this thesis is to develop Amharic text to Ethiopian Sign Language conversion machine translation system containing Amharic morphological analyzer subsystem, Amharic morpheme to ESL mapping bilingual dictionary, and SiGML transcription of selected ESL words.

Specific Objectives

Specific objectives of this thesis are:

- Transcribing selected ESL words in HamNoSys and SiGML
- Rearranging Amharic alphabets in a way suitable for computer based morphological analysis
- Developing an Amharic word expander upon Amharic word expansion/derivation rules
- Maintaining Amharic words dictionary containing root word and its morphemes
- Designing a Amharic-word to ESL-word(s) mapping system

1.4 Methodology

- ✓ **Literature survey in the area:** There will be enough literature review on previous works done on Amharic and different SL-MT systems and different researches on Amharic language. In addition there will be continuous contact and discussion with individuals and organizations working on sign language projects.
- ✓ **Developing Amharic word expansion program and corpus:** From randomly selected words of different categories, i.e. verbs, nouns, adjectives and adverbs, an Amharic corpus of limited vocabulary will be developed through morphological-expansion by applying different rules.
- ✓ **Developing Amharic to ESL dictionary:** Amharic stem word and morphemes to sign-text dictionary will be developed, which includes sign-texting of selected words in Signing Gesture Markup Language (SiGML).
- ✓ **Designing the MT system :** by applying rules for the translation of closely related languages, the machine translation which analyze given Amharic text and output the corresponding ESL animation system will be developed .
- ✓ **Linking Sign-text to the Avatar:** the sign-text(SiGML) to signing avatar *anna*, which is found from JAsigning project by permission, linking program will be developed.

1.5 Scope of the Study

While working this thesis different researchers were contacted. One of them was Matt Huenerfauth, who has done his PhD dissertation on "Generating American Sign Language

Classifier Predicates For English-To-ASL Machine Translation". With suggestions from him and other colleagues and by taking into consideration the time limitation, the scope of the research is determined.

From the input side, based on phonetic theories of Amharic (see Chapter two for more discussion on subject matter) the twenty-seven simple phonemes are taken. The remaining three complex phonemes can be substituted by combination of simple phonemes as follows:

- { k^w}[**ኵ**] with {kwa}[**ክዋ**]
- {g^w}[**ጐ**] with {gwa}[**ግዋ**]
- {q^w}[**ቍ**] with {qwa}[**ቅዋ**]

Handling all words in one language is a difficult task, and in the case of Amharic it's more difficult due to its morphological complexity. So, for this study randomly selected twenty verbs containing 2 to 4 alphabet used (from now on the term "*fidel"* in italics is used instead of "alphabet" for Amharic characters to avoid confusion with alphabets in ASCII representation). Ten countable nouns with plural forms, five adjectives and adverbs without any morphological analysis are also included, so that it may help for future researchers on MT system for ESL.

It's known that Amharic is morphologically rich language, especially the verbs. From many morphological expansion rules, forty are implemented to transitive verbs. Therefore, the number of Amharic words in the dictionary is more than eight hundred.

The system is designed to finger-spell the Amharic text if it fails to find the exact match. In the finger-spelling sub system the consonant {e}[**እ**] is not handled. This is done to avoid the complexity on the coding and shortage of time to handle its vowel form and consonant form.

1.6 Organization of the Thesis

This project is organized in 4 chapters.

The following chapter, chapter 2, is dedicated for studying the previous works done on the area. Since there is no Amharic-to-ESL MT system; other SL conversion systems are discussed in addition to the study made on previous researches on Amharic language Natural Language Processing (NLP).

The analysis and design of Amharic text analyzer subsystem is discussed on the third chapter, while the overall MT system design and implementation goes to chapter four. In these chapters every necessary detail is touched while avoiding repetitions.

In chapter 5, results of the system evaluation by end-users are presented. Every finding is explained based on the theories and the system architecture. The last chapter is dedicated to the conclusion and recommendations for future work.

Chapter 2

Literature Review on Amharic to ESL MT Systems

2.1 Introduction

There are many researches in the development of Machine Translation (MT) system for spoken language to sign language. There are four big projects for English to American Sign Language(ASL) MT system development ; European Union's ViSiCAST project, ZARDOZ,ASL Workbench, TEAM by University of Pennsylvania. Also other countries developed their own speech/text-to-sign language conversion software , such as: Polish Sign Language, Japanese Sign Language , Chinese Sign Language, South African Sign Language, and Sign Language of the Netherlands [14][15]. But there is no previous work for Amharic to ESL.

To translate from one language to another, the following knowledge sets are necessary

- Source language knowledge (for this research it's Amharic)
- Target language knowledge (which is ESL)
- Knowledge of various correspondence between source language and target language
- Knowledge of the subject matter, including ordinary general knowledge and 'common sense'.
- Knowledge of the culture, social conventions, custom and expectation of the source and target language users (speakers and/or signers).

Knowledge of language includes the following parts

- Phonological(orthography) knowledge: sound and writing system of a language,
- Morphological knowledge: how words are constructed in the language
- Syntactic knowledge: how sentence and phrases can be made up out of words
- Semantic knowledge: what the words and phrases mean

In this chapter previous works on other SL-MT systems is discussed and the language properties of Amharic and ESL are covered then. The background information on MT systems is mainly summarized from Hutchins [17][18][19].

2.2 Machine Translation Systems

The main idea behind MT systems is that a given source language (input) is transformed into target language (output) by carrying out the simplest possible parse, replacing source words with their target language equivalent as specified in a bilingual dictionary, and then roughly rearranging their order to suit the rules of the target language.

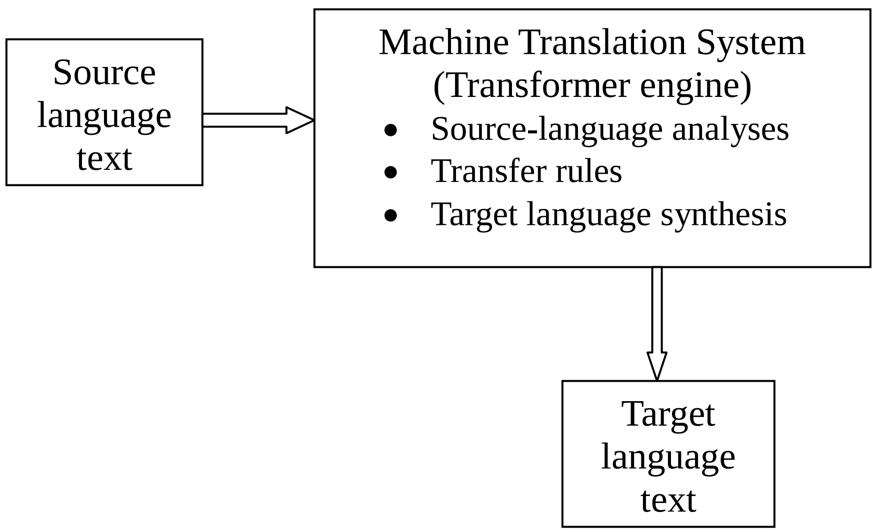


Figure 2.1. MT System architecture

As shown in figure 2.1, the MT system first accepts the word(s) from the source language and analyzes to get language information necessary for bilingual dictionary. Then the bilingual dictionary maps to target language by using transfer rules. Finally target language is synthesized according to its own linguistic structure.

Before going into the details of MT systems, it's important to have a look at its history.

2.2.1 History of Machine Translation

In 1933 the first practical suggestions were made with two patents issued in France and Russia to Georges Artsrouni and Petr Trojanskij respectively. Artsrouni's patent was for a general-purpose machine which could also function as a mechanical multilingual dictionary. Trojanskij's patent, also basically for a mechanical dictionary, went further with proposals for coding and interpreting grammatical functions using 'universal' symbols in a multilingual translation device [16].

Andrew Booth and Warren Weaver met in 1946 and 1947 and put forward the first tentative ideas for using the newly invented computers for translating natural languages. In 1948 Booth worked with Richard H. Richens on morphological analysis for a mechanical dictionary. By this time, the idea of mechanical translation had occurred independently to a number of people, and in July 1949 Warren Weaver (a director at the Rockefeller Foundation) put forward specific proposals for tackling the obvious problems of ambiguity, based on his knowledge of cryptography, statistics, information theory, logic and language universals. This memorandum was the stimulus for MT research in the United States.

In 1951 Yehoshua Bar-Hillel was appointed to do research at the Massachusetts Institute of Technology (MIT). After visiting all those interested in the subject he wrote a state-of-the-art report, in which he outlined some of the basic approaches to MT questions; and in June 1952, he convened the first MT conference at MIT.

In 1954 the first demonstration of MT system for carefully selected Russian sentence to English was presented by IBM and Bar-Hillel. The demonstration stimulates the large scale funding of MT research in the USA and research on MT begun in USSR.

In 1950s and 1960s two types of MT research methods are followed. The first one is an empirical trial-and-error approach (also called brute-force) which aims to develop a system producing useful if crude quality translations in the near future. The second method is based on theoretical approaches (also called perfectionist) which involve research in computational linguistics.

From year 1956 to 1966 three basic models of MT approaches were evolved. The first was the 'direct translation' model, where programming rules were developed for the translation specifically from one source language into one particular target language with a minimal amount of analysis and syntactic reorganization.

The second approach was the ‘interlingua’ model, based on abstract language-neutral representations, where translation would then be in two stages, from source language to interlingua and from interlingua to target language.

The third approach was less ambitious: the ‘transfer approach’, where conversion was through a transfer stage from abstract (i.e. disambiguated) representations of source language texts to equivalent target language representations; in this case, translation comprised three stages: analysis, transfer, and generation (or synthesis).

The year 1966 Automatic Language Processing Advisory Committee (ALPAC), set up by the National Science Foundation at the instigation of the U.S. sponsoring bodies, which concluded that MT was slower, less accurate and twice as expensive as human translation, that “there is no immediate or predictable prospect of useful machine translation” and furthermore that there was no shortage of technical and scientific translators in the United States [3]. It is unfortunate that the public image of MT has been formed by the disastrous and grossly expensive mistakes of the early work on MT. There is perhaps no other scientific enterprise in which so much money has been spent for so little return. By 1965 it has been estimated that U.S. government agencies had supported MT research at 17 institutions to the tune of almost 20 million dollars. A sudden and abrupt end came with the report and henceforth MT was to be regarded as an expensive failure. This resulted in a decade of quite time [17][18][19].

In the decade after ALPAC, more systems were coming into operational use and attracting public attention. For many years, the systems were based on direct translation by using bilingual dictionaries. By the 1980s, however, advances in computational linguistics allowed much more sophisticated approaches, and a number of systems adopted an indirect approach to the task of translation. In these systems, texts of the source language are analyzed into abstract representations of ‘meaning’, involving successive programs for identifying word structure (morphology) and sentence structure (syntax) and for resolving problems of ambiguity (semantics). Included in the latter are component programs to distinguish between homonyms (e.g. Amharic words such as {felege}[**ፈለገ**], which can mean either want/need or search) and to recognize the correct semantic relationships. The abstract representations are intended to be unambiguous and to provide the basis for the generation of texts into one or more target languages.

There have in fact been two basic ‘indirect’ approaches. In one the abstract representation is designed to be a kind of language-independent ‘interlingua’, which can potentially serve as an intermediary between large numbers of natural languages. Translation is therefore in two basic stages: from the source language into the interlingua, and from the interlingua into the target language. In the other indirect approach (in fact, more common approach) the representation is converted first into an equivalent representation for the target language. Thus there are three basic stages: analysis of the input text into an abstract source representation, transfer to an abstract target representation, and generation into the output language. By late 1980s, systems of all these kinds were developed, and it is true to say that all current commercially available systems are also classifiable into these three basic system types: direct, interlingual and transfer.

The best known of the MT systems for mainframe computers are in fact essentially of the ‘direct translation’ type. They are however improved versions of the type; unlike their predecessors, they are highly modular in construction and easily modifiable and extendable. In particular, the Systran system, originally designed for translation only from Russian into English, is now available for a very large number of language pairs: English from and into most European languages (French, German, Italian, Spanish, Portuguese), Japanese, Korean, etc. Logos, originally marketed for German to English, is also now available for other languages: English into French, German, Italian and Spanish, and German into French and Italian. The Fujitsu Atlas system, on the other hand, is still confined to translation between English and Japanese (in both directions).

In the late 1980s, Japanese governmental agencies began to sponsor an interlingua system for Asian languages, involving co-operation with researchers in China, Thailand, Malaysia and Indonesia.

Recently the MT systems are mainly statistical based and its derivation corpus based. Also speech translation is one of the research areas and there are systems that use hybrid approach. Since mid 1990, it has become common to provide online MT systems from the vendors.

2.2.2 Machine Translation Approaches

MT systems can be classified into three depending on the way they map from source language to target language: rule based, corpus based, and Hybrid. Each approach is briefly described below.

2.2.2.1Rule based Approach

Rule based approach can be classified in to three depending on the depth of the linguistic rules applied on the system. Fig. 2.2 shows the three ruled-based MT and their linguistic analysis/synthesis stages.

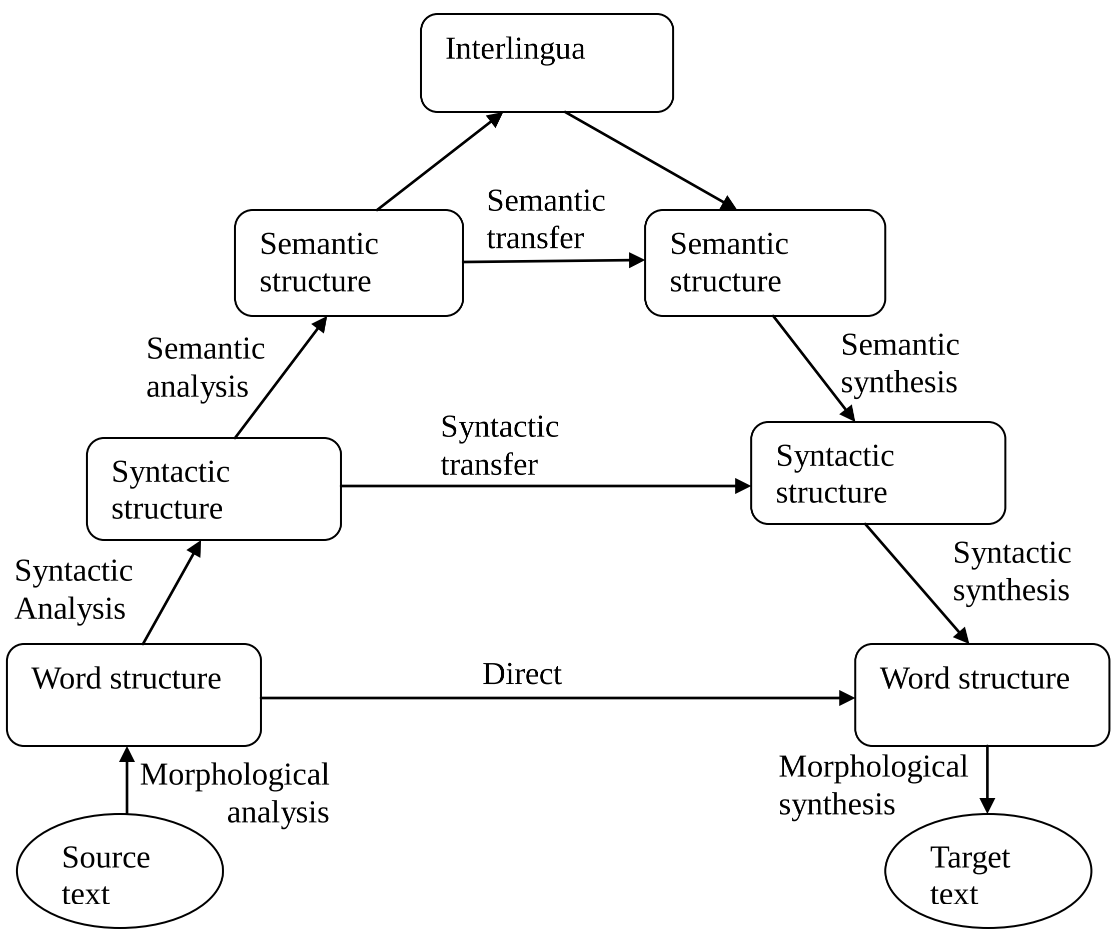


Figure 2.2. rule-based translation systems [10]

Direct translation

Direct translation is the oldest approach to MT. The direct translation strategy passes each sentence text to be translated through a series of standard stages. If the MT system uses direct translation, this usually means that there is no syntactic or semantic analysis after the morphological analysis for the source language. The translation is based on large dictionaries and word-by-word translation with some simple grammatical adjustments e.g. on word order and morphology. Due to lack of grammatical information about ESL (it's in research phase) and ESL corpus, this thesis can be done only by direct translation and some form of transfer approach based on the authors knowledge of the language [10].

Interlingua-based translation

The basic idea of the interlingual (pivotal) architecture for MT is that the analysis of source language text should result in representation of text that is independent of the source language. In effect, in Interlingua there is no transfer map, and the MT model thus has phases: analysis and generation. In a standard multilingual system with X source languages and Y target languages, the transfer approach will involve XY transfer maps; moreover, we need X analysers and Y generators. In the Interlingua approach, only X parsers and Y generators are needed per language. Interlingua based MT is done via an intermediary (semantic) representation of the source language text. Translation needs two phases: analysis from the source language to the Interlingua (universal language) and generation from the universal language to the target language [10][26].

Transfer based translation

Transfer systems are a middle course between direct and Interlingua MT strategies. They divide translation into steps which clearly differentiate source language and target language parts. In the transfer approach there is therefore no language independent representation: the source language intermediate representation is specific to a particular language, as is the target language intermediate representation. There is no necessary equivalence between the source and target intermediate representations for the same language. In the transfer strategy a source language sentence is first parsed into an internal representation. Thereafter a transfer is made at both lexical and structural levels into equivalent structures of the target language. In the third stage a translation is generated. Whereas the Interlingua approach requires complete resolution of all ambiguities in the source language text so that translation into any other language is possible, in the transfer approach only those ambiguities inherent in the language in question are tackled. This approach is a development over direct translation and this was lexically driven [26].

2.2.2.2 Corpus based Approach

While the classical rule-based approach can achieve reasonably high translation quality, it also takes many person years to create a system that can handle a wide variety of sentence constructions and will perform well on unseen text. Other approaches, which currently receive considerable treatment in the research community, are corpus-based approaches, namely Statistical MT (SMT) and Example-Based MT (EBMT). Much attention has recently been given to the rapid deployment of MT systems: given a language pair, how quickly can we build a system that can produce reasonable quality output? [25].

Statistics-based MT

SMT is the application of statistical techniques to the task of translating texts from one natural language to another. Key notions involved are those of the language model and the translation model. With this information (which is extracted automatically from the corpus), the translation model can, for a given S, calculate P(T|S) (that is, the probability of T, given S). This is the essence of the approach to statistically-based MT, although the procedure is itself slightly more complicated in involving search through possible source language sentences for the one which maximizes P(S)*P(T|S).

A variant of Baye's rule is used to show that the probability that a string of words (T) is a translation of a given string of words (S) is proportional to the product of the probability that a string of target words is a legal utterance in the target language and the probability that a string of words in source language is a translation of string of words in the target language.

$$P(T|S) = \frac{P(T)*P(S|T)}{P(S)} \text{ --(2.1)}$$

Because P(S) is fixed for a given source sentence, it can be ignored while trying to maximize P(T|S). Therefore

$$P(T|S) \sim P(T)*P(S|T) \text{ --(2.2)}$$

One needs to find T that maximizes the right hand side of equation (2.2).

It should be clear that in an approach such as this there is no role whatsoever for the explicit encoding of linguistic information, and thus the knowledge acquisition problem is solved.

On the other hand, the general applicability of the method might be doubted, since it is heavily dependent on the availability of good quality bilingual or multilingual data in very large proportions, something which is currently lacking for most languages [10][32].

Example-Based MT

EBMT builds a database of word- and phrase-level associations from a bilingual corpus. The approach is based on the observation that the bilingual corpus is essentially a database of correct translations. If the system can determine which parts of the source language sentences translate into which parts of the target language sentences, then it can build a repository of known correct translations. At run-time, a source language sentence is matched as a whole or in part against the translation examples, and is translated using those examples. The partial translations are then merged to form a full translation of the source language sentence [25][4].

2.2.2.3 Hybrid Systems

Hybrid Systems combine the Rule-based approach and statistics-based approach to get a better result. The system combines linguistic information and statistical information. By doing this, the problems of ambiguity on the rule-based system are solved by the statistical part and complex sentences, which are difficult for SMT are solved by the rule-based systems. This approach is the most common one now a days by MT researchers [25][10].

2.2.3 Machine Translation Systems for Sign Languages

As discussed in chapter one ESL come to public attention only recently and there are very few publications on ESL and, to best of my knowledge, there is no previous work on Amharic to ESL MT system.

But for the American counterpart of ESL, American Sign Language (ASL), there are four standard speech-to-ASL MT systems. Other countries have developed their own speech/text-to-sign language conversion software, such as: Polish Sign Language, Japanese Sign Language, Chinese Sign Language, South African Sign Language, Greek Sign Language, and Sign Language of the Netherlands.

Here is the brief description about the four English-to-ASL MT systems as discussed by Huenerfauth [15].

ZARDOZ

The ZARDOZ system was proposed (and somewhat implemented) English-to-Sign-Languages translation system using a set of hand-coded schemata as an interlingua for a translation component. While the implemented portion of the system focused on American Sign Language, the authors were developing their framework with British, Irish, and Japanese Sign Language also in mind. Some of the research foci of this system were the use of Artificial Intelligence (AI) knowledge representation, metaphorical reasoning, and a blackboard system architecture; so, the translation design is very knowledge-based and AI-reasoning heavy. During the analysis stage, English text would undergo sophisticated idiomatic concept decomposition before syntactic parsing in order to fill slots of particular concept/event/situation schemata. The advantage of the logical propositions and labeled slots provided by a schemata-architecture was that commonsense and other reasoning components in the system could later easily operate on the semantic information. Because of the amount of hand coding needed to produce new schemata, the system would only be feasible for limited domains. To compensate for these limitations, if a schema did not exist for a particular input text, the authors suggest that the system could instead perform word-for-sign transliteration (producing Signed English output) [15].

ASL Workbench

The ASL Workbench was a proposed and partially implemented English-to-ASL machine-assisted translation system which incorporated modern research on ASL phonological models. The sophisticated phonological model used by this system is particularly robust and is based on the modern Movement-Hold model of ASL phonology. The design uses a lexical-functional grammar (LFG) for analysis of the English text into a functional structure, handcrafted transfer rules for converting an English f-structure into an ASL one, and LFG rules to produce ASL output. The system uses a transfer-specific lexicon to map English words/phrases to analogous ASL signs/phrases.

When the system encounters difficulties in lexical choice or other translation tasks, it asks the user of the system for advice. The system also records a very simplistic discourse model from the English input (consisting of a flat list of discourse elements and their spatial locations – all of which are specified by the human operator). The ASL Workbench system was designed to be a tool to assist professional translators in converting from English to American Sign Language. In fact, the system cannot operate correctly without human intervention. It makes no attempt to resolve discourse referents in an English text; it requires the user to note when two phrases refer to the same entity [15].

TEAM

TEAM was an English-to-ASL translation system built at the University of Pennsylvania that employed Synchronous Tree Adjoining Grammar rules to build an ASL syntactic structure simultaneous to the parsing of an English input text. The output of the linguistic portion of the system was a string-like ASL “gloss notation with embedded parameters” that encoded limited information about morphological variations, facial expressions, and sentence mood as features associated with the individual words in the string. This project took advantage of graphics research at the University of Pennsylvania by using one of the Center for Human Modeling and Simulation’s (HMS) animated virtual humans as the signing avatar. In particular, this project explored how the HMS technology’s ability to modify the “manner” of an animated character’s movements (by modifying a small number of parameters in their animation software) could be used to modify the performance of some ASL signs to indicate adverbial information or perform morphological inflection [15].

ViSiCAST (Virtual Signing: Capture, Animation, Storage and Transmission)

As part of the European Union’s ViSiCAST project, researchers at the University of East Anglia implemented a system for translating from English text into British Sign Language. Their approach uses the CMU Link Parser to analyze an input English text, and then uses Prolog declarative clause grammar rules to convert this linkage output into a Discourse Representation Structure. During the generation half of the translation process, Head Driven Phrase Structure rules are used to produce a symbolic sign language representation script. This script is in the system's proprietary "Signing Gesture Markup Language," a symbolic coding scheme for the movements required to perform a natural sign language [15].

The ViSiCAST project was the first of series of projects on SL by European Union and now it's in its fourth cycle Dicta-Sign project (Dicta-Sign is a three-year EU-funded research project that aims to make online communications more accessible to sign language users). It passes through **eSIGN** (Essential Sign Language Information on Government Networks) [11] project and recently finished the JASigning Project.

In this research, as mentioned in previous chapter, virtual animator (avatar) *anna* is used as signer. It's available in the official web site of JASiging project (http://vhg.cmp.uea.ac.uk) and can be downloaded and used for research purpose. In addition to the avatar player software (SiGML Service player), there are also other tools available which help to transcribe in SL. As the name indicates, SiGML service player accepts SiGML as input and plays/signs it.

Regarding the signer there are basically two possible approaches: either using a recorded video or designing a virtual signer. When transmission over internet and storage cost are considered, different researches have shown that the later one is advisable. For example in eSIGN project it was shown that the data used to drive the video data (with the best available compression, MPEG-4) for one second data in 320×240 resolution is 64 times of what required for avatar. When the resolution is 640×480 it increases further to 230 times of the avatar. There are also other issues like flexibility, production cost, and customization [30].

Video usage will preserve the naturalness of signing, but due to its "static" property of representation of linguistic utterance, it cannot be easily modified. It's known that human language is dynamic by nature, and there should be means to easily update the material. In the case of video, one should re-record again, and remove the previous one. With avatar there is no need to change it, but simply change the command (in this case the SiGML representation of that particular word).

Another area of research focus in SLs is how to present or write signs. From many proposed notification systems SignWriting by Valerie Sutton, Stokoe Notation Joe Martin, and HamNoSys by Hamburg University are the dominant ones. From the three notifications systems HamNoSys is the most suitable for SL-MT systems [23].

Here is the brief description of HamNoSys put by Susanne Bentele, from Humberg University, on SignWriting web:

> *HamNoSys was developed by a group of hearing and deaf people as a scientific/research tool and first made publicly available in 1989. The purpose of HamNoSys, unlike SignWriting, has never been an everyday use to communicate (e.g. in letters) in sign language. It was designed to fit a research setting and should be applicable to every sign language in the world. It consists of about 200 symbols covering the parameters of handshape, hand configuration, location and movement. The symbols are as iconic as possible and are easily recognizable. The order of the symbols within a string is fixed, but still it is possible to write down one and the same sign in lots of different ways. The notation is somewhat phonemic, we're working on a 'phonologization' of the system at the moment. Hence, the transcriptions are very precise, but on the other hand also very long and cumbersome to decipher. It is possible to note down facial expressions, but their development isn't quite finished yet. ... HamNoSys is still being improved and extended all the time as the need arises. The system is used, for example, in research institutions in Finland, Australia, New Zealand, Switzerland and Germany [23].*

HamNoSys has been designed following the basic principles:

- Language independence: HamNoSys is not specific to a particular SL, it allows to describe any signs of any language. This feature stems from the original motivation that led to the implementation of HamNoSys: written to provide a means for researchers to describe signs.
- Description of posture and movement, not of meaning: The meaning of a sign is described in HamNoSys not only describes the position and movement of hands.

A sign may have different meanings in different contexts, but if it is in the same way, HamNoSys transcription is the same.

- The omission of irrelevant information: Only describe the parts of posture and movement that are important in creating a sign. Most signs are represented through the hands and face, so that the position of, for example, shoulders and elbows is not important, just have to adapt naturally to the hand position. Thus, shoulders and elbows have no notation in most of the signs.

ViSiCAST is the first project to use HamNoSys for storing the phonetic form of individual signs in the lexicon and for combining signs into sign language utterances. ESign Editor Tool, designed in the project eSign provides an editing environment that transcribes the signs of a quick and easy, also incorporating the tools needed to describe the non-manual signs. Since computers cannot process the syntax of these descriptions of HamNoSys, University East Anglia designed SiGML [13].

SiGML contains exactly the same information as the description HamNoSys, but in a language that computers can process. SiGML is an XML application language defined in the ViSiCAST project for the purpose of specifying signing sequences. The major component of SiGML allows sign language gestures to be defined at the phonetic level, based on the model of signing phonetics embodied in HamNoSys [24].

Since there is no previously done transcription of ESL alphabets and words in HamNoSys notification system, it is one of the tasks of this research work (for detail see chapter 4).

2.3 Amharic Language

Amharic, which belongs to the Semitic language family, is the official language of Federal Government of Ethiopia. According to the latest census results, Amharic is a mother tongue of more than 17 million people. The language is also used as a second language for over 5 million people [12]. In Semitic language family, Amharic stands second in its number of speakers next to Arabic, and has five dialectical variations (Addis Ababa, Gojjam, Gonder, Wollo, and Menz) spoken in different regions of the country. The speech of Addis Ababa has emerged as the standard dialect and has wide currency across all Amharic-speaking communities [27].

2.3.1 Amharic Phonetics

As with all of the other languages, Amharic has its own characterizing phonetic, phonological and morphological properties. For example, it has a set of speech sounds that is not found in other languages. For example the following sounds are not found in English {P}[ጰ],{S}[ጸ],{x}[ሽ], and {q}[ቀ].

Amharic also has its own inventory of speech sounds. It has thirty consonants (27 simple and 3 complex) [7] and seven vowels. The consonants are generally classified as stops, fricatives, nasals, liquids, and glides. Tables 2.1 and 2.2 show the classification of Amharic consonants and vowels respectively.

Table 2.1. Categories of Amharic Consonants

Manner of Articulation	Voicing	Place of Articulation											
		Labials		Alveolar		Palatals		Velars		Labio-velars		Glottals	
Stops	Voiceless	P	ፕ	T	ት	ʧ	ች	k	ክ	kʷ	ኵ	ʔ	ዕ
	Voiced	B	ብ	D	ድ	ʤ	ጅ	g	ግ	gʷ	ጕ		
	Glottalized	p'	ጵ	t'	ጥ	ʧ'	ጭ	q	ቅ	qʷ	ቍ		
Fricatives	Voiceless	F	ፍ	S	ስ	ʃ	ሽ					H	ህ
	Voiced	V	ቭ	Z	ዝ	ʒ	ዥ						
	Glottalized			s'	ሽ								
Nasal	Voiced	M	ም	N	ን	ɲ	ኝ						
Liquids	Voiced			L	ል								
Glides	Voiced	W	ው	R	ር	y							

Table 2.2. Categories of Amharic Vowels

	Front		Central		Back	
High	I	ኺ	ɨ	ኽ	U	ኹ
Middle	E	ኼ	ə	ኸ	O	ኾ
Low			A	ኻ		

2.3.2 Amharic Writing System

Amharic uses the Ge'ez writing system, called *Fidel*. In this system the symbols are consonant based consonant-vowel combination and are arranged in a row and column. According to Baye [7] , each row is dedicated to the thirty consonants and the columns or orders represent the seven different phonemes resulted on the application of the vowels in a regular fashion. From this it’s apparent that Amharic writing system is partially phonetic: i.e. there is more or less a one-to-one correspondence between the phones and the symbols [28].

The thirty three core characters by seven orders gives 231 distinct symbols. In addition to this there are others that contain special features usually representing labialization like {kwa}[ኳ], {gwa}[ጓ], {qwa}[ቋ]. Regarding the punctuation marks, there are about seventeen, of which only few are commonly used [1]. Thus approximating the writing system can be achieved by taking out the redundant symbols without losing essential understanding [21].

2.3.3 Amharic Morphology

Amharic, as other Semitic Languages, is morphologically rich language. Its morphology is extremely complex, but is relatively well understood [7][21][28].

The smaller units which are used to form words are called morphemes. Morpheme is a meaningful combination of phonemes, and can be divided in to two; free morpheme and bound morpheme. A free morpheme can stand by itself to give a meaningful word, while the bound morpheme give meaning only when combined with another morpheme [7][29]. For example, {lam}[ላም]-(cow) is a free morpheme and has meaning. But, the morpheme {-E}[-ኤ] in word {lame}[ላሜ] – (my cow) cannot stand by itself or does not have any meaning in Amharic.

Words are formed by combining one or more morphemes. In word formation, Amharic shows root-pattern morphological phenomenon [8]. Different words can be formed by affixation, reduplication or compounding [29]. Affixation attaches affixes to free morphs, and they can be either prefixes, suffixes, or infixes. The following example shows affixation in Amharic.

Example: Root word: {bela}[በላ]

Prefix: {as-bela}[አስበላ]

Suffix:{bela-n}[በላን]

Infix:{byi}[ብዪ]

The second type of word formation is combining two free morphemes to form another word form. And the third type is reduplication, in which portion of a freely standing morpheme is duplicated to form another word.

Example:

Combination: {tmhrt-bEt}[ትምህርት ቤት]

Reduplication:{bela-bela}[በላ በላ]

Different categories of words exhibit different derivation from morphemes. Different researchers categorize Amharic words in eight groups (Noun, Verb, Adjective, Adverb, Preposition, Pronoun, Conjunction, and Interjection) or five groups (Noun, Verb, Adjective, Adverb, and Preposition) depending on their chronology [8][6]. Since, categorizing in five is the recent one; it's used in this thesis. Each category is discussed below.

2.3.3.1 Nouns

Amharic nouns are free morphemes [7] and are used to name or identify any of a class of things, people, place or ideas [6][29]. Nouns can further be classified as countable and uncountable, pronoun, ideas, indicative pronoun, possessive pronouns and so on.

Depending on their classes, countable nouns may take the suffix {-oc}[ኦች] or prefix {Ine-}[እነ-]to make plural form.

2.3.3.2 Verbs

In Amharic verbs are words that usually come at the end of complete declarative sentence, and shows action or event. According to Baye [7], they are mostly composed of first column/order of Amharic *fidel*(*Ge'ez* form) and can take the prefix {Iye-}[**እየ-**]. In addition they also can change the last *fidel* from first order (*Ge'ez*) to seventh order (*sabI*).

2.3.3.3 Adjectives

Adjectives are words that come before nouns in sentence, and serve as modifiers or describers of nouns they precede.

2.3.3.4 Adverbs

According to Tesfaye [29] words in adverb class are small in number and they do not attach any kind of prefixes and suffixes. But Atelach [6] and Daniel [8] add other adverb classes which are found in bound-morpheme format (for more information refer to [6], pp.44-45 and [8], pp.44-45). In general Adverbs refer to places, time, and circumstances of the action mentioned by the verb.

2.3.3.5 Prepositions

Amharic prepositions, like adverbs, are few in number and are bound-morphemes:i.e. they don't have meaning by themselves. Their purpose is to coordinate or subordinate nouns [7].

2.4 Ethiopian Sign Language

Sign Language is a fully-fledged natural language. That is, it is not simply a signed version of some other natural language such as English, Amharic, Afan Oromo or Haddiyisa , but has its own syntax and grammar rules. Hence, the problem of machine translation from a given source language to Sign Language is as difficult (if not more) as the machine translation between English and, say, Amharic. There is need for a machine translation system from Amharic to Sign Language. The Deaf community in Ethiopia, particularly in the previously disadvantaged communities, has a low literacy rate. As such, the usual ways of information communication (TV, newspapers) is often ambiguous, misleading, or non-existent for this group of the population.

Communication in commonplace situations (for example, clinics, hospitals, post offices and police stations) is difficult if not impossible without Sign Language interpreters. The only feasible way of communication information to the Deaf community at large, is currently by using Sign Language interpreters (on ETV, for example). This is expensive, and it is impractical to translate all information in this manner. A computer translation system would enable mass translation for more situations in an affordable manner.

As any other sign languages ESL is language which uses manual communication, body language and lip patterns instead of sound to convey meaning, simultaneously combining hand shapes, orientation and movement of the hands, arms or body and facial expressions to express fluidly a speaker's thought. ESL commonly developed in deaf communities, which can include interpreters and friends and families of deaf people as well as people who are deaf or hard of hearing themselves.

Linguistics of ESL

In linguistic terms, sign languages are as rich and complex as any oral language, despite the common misconception that they are not "real languages". Professional linguists have studied many sign languages and found them to have every linguistic component required to be classed as natural languages.

Sign languages are not pantomime- in other words; signs are largely arbitrary and have no necessary visual relationship to their referent, much as most spoken language is not onomatopoeic. Nor are they a visual rendition of an oral language. They have complex grammars of their own, and can be used to discuss any topic, from the simple and concrete to the lofty and abstract.

Sign languages, like oral languages, organize elementary, meaningless units into meaningful semantic units. The elements of a sign are **H**and-shape (or Hand-form), **O**rientation (or Palm Orientation), **L**ocation (or Place of Articulation), **M**ovement, and Non-manual markers (or Facial **E**xpression), summarized in the acronym **HOLME**.

Common linguistic features of deaf sign languages are extensive use of classifiers, a high degree of inflection, and a topic-comment syntax. Many unique linguistic features emerge from sign languages' ability to produce meaning in different parts of the visual field simultaneously. For example, the recipient of a signed message can read meanings carried by the hands, the facial expression and the body posture at the same moment. This is in contrast to oral languages, where the sounds that comprise words are mostly sequential (tone being an exception).

Sign languages' relationships with oral languages

A common misconception is that sign languages are somehow dependent on oral languages, that is, that they are oral language spelled out in gesture, or that they were invented by hearing people. ESL is not a signed Amharic or other Ethiopic Language. And ASL or British sign language (BSL) is not signed English. Also ESL is not the same as ASL though, two signers of different sign language can communicate easily than two different spoken language users. For example an ESL-only signer (“speaker”) and ASL-only signer (“speaker”) will communicate better than an Amharic-only speaker and English-only speaker will do each other.

The manual alphabets of Amharic are used in ESL. But proper names, common names and technical words have their own ESL vocabularies.

On the whole, deaf sign languages are independent of oral languages and follow their own paths of development. For example, BSL and ASL are quite different and mutually unintelligible, even though the hearing people of Britain and America share the same oral language.

Similarly, countries which use a single oral language throughout may have two or more sign languages; whereas an area that contains more than one oral language might use only one sign language. South Africa , which has 11 official oral languages and a similar number of other widely used oral languages is a good example of this. It has only one sign language with two variants due to two major educational institutions for the deaf which serve different geographic areas of the country.

Spatial grammar and simultaneity

Sign languages exploit the unique features of the visual medium. Oral language is linear.

Only one sound can be made or received at a time. Sign language, on the other hand, is visual; hence a whole scene can be taken in at once. Information can be loaded into several channels and expressed simultaneously. As an illustration, in Amharic one could utter the phrase, "{mgb belahu} [**ምግብ በላሁ**]" (I ate food). To add information about the amount, one need to add a phrase "{beTam bzu}[**በጣም ብዙ**]" (too much). He/she would have to make even longer if he/she want to show the feeling, like satisfaction or dissatisfaction. But in ESL, all information can be conveyed at the same time by changing the way the hand moves (fast, slow, strong, repeated) and at the same time by applying non-manual features such as body posture, facial expression and so on. Even though the Amharic phrase "{beTam bzu mgb belahu}[**በጣም ብዙ ምግብ በላሁ**]"(I ate to much food) is longer than "{mgb belahu} [**ምግብ በላሁ**] "(I ate food), in ESL the two signs may have equal length.

Chapter 3

Preparing Amharic for the MT System

3.1 Introduction

In the previous chapter general linguistic theories behind Amharic and ESL are discussed. Also the history and requirement in MT are briefly explained. The morphological richness of Amharic language made it difficult to manually list all available words. But many researchers showed that there is a certain pattern to generate words from stem (root) words. In this chapter, morphological expansion of Amharic words is discussed in detail, which is implemented in this thesis work as a basic part on the MT system.

3.2 Amharic *Fidel* Rearrangement

Mesfin [21] suggested a column and row movement as a morphological expansion method for Amharic words. He found that the approach failed to work at some point (see [21] pp.52-53). Baye [7] has shown that this phenomenon happen to alveolar phonemes tend to shift towards palatals when followed by the vowel {i}[**ኢ**](pp 28-29).

Based on the language property, the *fidel* is rearranged based on their phonological characteristics. It's found that this approach gives some uniformity to the morphological expansion program. Table 3.1 shows the rearranged *fidel*.

Table 3.1. phonetically arranged Amharic *fidel*

	ሳድስ	ግዕዝ	ካዕብ	ሳልስ	ራብዕ	ሃምስ	ሳብዕ
		E	U	I	A	E	O
ፐ	P	Pe	Pu	Pi	Pa	pE	Po
ተ	T	Te	Tu	Ti	Ta	tE	To
ቸ	C	Ce	Cu	Ci	Ca	cE	Co
ከ	K	Ke	Ku	Ki	Ka	kE	Ko
በ	B	Be	Bu	Bi	Ba	bE	Bo
ደ	D	De	Du	Di	Da	dE	Do
ጀ	J	Je	Ju	Ji	Ja	jE	Jo
ገ	G	Ge	Gu	Gi	Ga	gE	Go
ጰ	P	Pe	Pu	Pi	Pa	PE	Po
ጠ	T	Te	Tu	Ti	Ta	TE	To
ጨ	C	Ce	Cu	Ci	Ca	CE	Co
ቀ	Q	Qe	Qu	Qi	Qa	qE	Qo
ፈ	F	Fe	Fu	Fi	Fa	fE	Fo
ሰ	S	Se	Su	Si	Sa	sE	So
ሸ	X	Xe	Xu	Xi	Xa	xE	Xo
ሀ	H	He	Hu	Hi	Ha	hE	Ho
ቨ	V	Ve	Vu	Vi	Va	vE	Vo
ዘ	Z	Ze	Zu	Zi	Xa	zE	Zo
ዠ	Z	Ze	Zu	Zi	Za	ZE	Zo
ጸ	S	Se	Su	Si	Sa	SE	So
መ	M	Me	Mu	Mi	Ma	mE	Mo
ነ	N	Ne	Nu	Ni	Na	nE	No
ኘ	N	Ne	Nu	Ni	Na	NE	No
ረ	R	Re	Ru	Ri	Ra	rE	Ro
ለ	L	Le	Lu	Li	La	lE	Lo
የ	Y	Ye	Yu	Yi	Ya	yE	Yo
ወ	W	We	Wu	Wi	Wa	wE	Wo
ዐ	I	E	U	I	A	E	O

3.3 Representation of Amharic *Fidel*

From the above table one can observe that *sadis* fidel is represented by a single ASCII character and the other forms (from *ge'ez* to *sabI*) are *sadis* followed by corresponding vowel.

This approach is suggested by Daneil [9], and is called SERA. The SERA representation found wide acceptance among researchers and is also adopted for this thesis work.

It's apparent that for real time applications, it's better to work with numbers than strings or characters. Bringing this in mind, for the morphological expansion/retrieval, Amharic characters/words are changed to array of integer numbers.

Table 3.2. number assignment of Amharic characters

Amharic	*ASCII*	*Number*
ፕ	P	0
ት	T	1
ች	C	2
ክ	K	3
ብ	B	4
ድ	D	5
ጅ	J	6
ግ	G	7
ጵ	P	8
ጥ	T	9
ጭ	C	10
ቅ	Q	11
ፍ	F	12
ስ	S	13
ሽ	X	14
ህ	H	15
ቭ	V	16
ዝ	Z	17
ዥ	Z	18
ጽ	S	19
ም	M	20
ን	N	21
ኝ	N	22
ር	R	23
ል	L	24
ይ	Y	25
ው	W	26
ዕ	I	27
እ	-	27
አ	E	28
ኡ	U	29
ኢ	I	30
ኣ	A	31
ኤ	E	32
ኦ	O	33

Example

{bela}[በላ] → 4,28,24,31

{TeTa}[ጠጣ] →9,28,9,31

{qome}[ቆመ] → 11,33,20,28

In the morphological expansion this numbering method will be used. This numbering will be done by the program and there is no need to hand-code each word. Also, the program handles converting from number array to ASCII format.

3.4 Morphological Expansion

Amharic words expansion or derivation follow predefined pattern depending on the type of the word. In chapter two it was shown that Amharic words can be classified in 5 categories; noun , verb, adjective, adverb, and preposition. Each category has its own derivation rule. For this research work, nouns and verbs are considered because they are the ones that are morphologically complex and appear more frequently than others in sentences.

3.4.1 Nouns

Nouns can be classified as countable and uncountable. Countable nouns can add the plural marker suffix {-oc}[ኦች] (,33,2) as shown in Table 3.3.

Table 3.3. noun pluralizing

Countable name	Plural form
{sew}[ሰው] (13,28,26)	{sewoc}[ሰዎች] (13,28,26,33,2)
{wenber}[ወንበር] (26,28,21,4,28,23)	{wenberoc}[ወንበሮች] (26,28,21,4,28,23,33,2)
{beg}[በግ] (4,28,7)	{begoc}[በጎች] (4,28,7,33,2)
{temari}[ተማሪ] (1,28,20,31,23,30)	{temariwoc}[ተማሪዎች] (1,28,20,31,23,30,26,33,2)
{elga}[አልጋ](28,24,7,31)	{elgawoc}[አልጋ](28,24,7,31,26,33,2)
{ejabi}[አጃቢ](28,6,31,4,30)	{ejabiwoc}[አጃቢ](28,6,31,4,30,26,33,2)

Algorithm for pluralizing can be driven from the pattern shown in Table 3.3 for regular noun expansion and is shown in figure 3.1 below.

```
Begin:
 For a given countable noun in number array
     If the last integer is less than 27 then
          Append 33 and 2 at the end of the array
     If not append 26, 33 and 2 at the end of the array
End:
```

Figure 3.1 algorithm for noun pluralizing

Another property of nouns is they allow expansion by applying the definite article indicator morpheme {-u}[-**ኡ**] (,29) which is shown below in Table 3.4.

Table 3.4. definite article suffixation

Name	Definite article suffixation
{beg}[**በግ**] (4,28,7)	{begu}[**በጉ**] (4,28,7,29)
{dem}[**ደም**] (5,28,20)	{demu}[**ደሙ**] (5,28,20,29)
{elga}[**አልጋ**](28,24,7,31)	{elgaw}[**አልጋው**](28,24,7,31,26)
{wxa}[**ውሻ**](26,14,28)	{wxaw}[**ውሻው**](26,14,28,26)

The corresponding algorithm for definite article is shown in figure 3.2.

```
Begin:
 For a given noun in number array
     If the last integer is less than 27 then
          Append 29 at the end of the array
     If not append 26 at the end of the array
End:
```

Figure 3.2. definite article suffixation algorithm

Nouns can be also expanded by adding the suffix {-E}[**-ኤ**](,32), as shown in Table 3.5 and corresponding algorithm, Figure 3.3 , to show something/someone is possessed by first person singular

Table 3.5. Possession suffix

name	Possession by first person singular
{dem}[ደም] (5,28,20)	{demu}[ደሜ] (5,28,20,32)
{doro}[ዶሮ](5,33,23,33)	{doroyE}[ዶሮዬ](5,33,23,33,25,32)
{wxa}[ውሻ](26,14,28)	{wxayE}[ውሻዬ](26,14,28,25,32)

The corresponding algorithm for definite article is:

```
Begin:
 For a given noun in number array
     If the last integer is less than 27 then
          Append 32 at the end of the array
     If not append 25 and 32 at the end of the array
End:
```

Figure 3.3. possession suffixing algorithm

There are also other rules which are not indicated in the above algorithms, but one can follow the method shown to derive rules. But, it should be noted that there are few irregular nouns and they should be added to the corpus by hand. For example, pronouns will not follow the above rules; so, they are inserted into the system by hand and there is no implemented rule of expansion upon them.

3.4.2 Verbs

Different researchers propose different ways to expand a given verb. Mesfin [21] argue that the best way from computational point of view is to consider masculine past tense as the stem (root) verb instead of approaches followed by some linguists(for further reading see [21] pp.45-47). Though a different approach is used regarding the expansion program, his approach is adopted for stem verb assignment.

Amharic verbs can contain information about the subject, object, gender, time and others.

They can be classified as transitive and intransitive verbs.

Some of the rules common to all verbs (except few irregular verbs which are handled separately) are shown below with examples and corresponding algorithm.

In Table 3.6., One of the rules that can be applied on verbs to get another word is adding the prefix {Iye-}[እየ](27,25,28,), is shown. The Algorithm for this rule is also shown in figure 3.4.

Table 3.6. prefix {Iye-}[እየ]

Words	Prefix {Iye-}[እየ]
{wesede}[ወሰደ](26,28,13,28,5,28)	{Iyewesede}[እየወሰደ](27,25,28,26,28,13,28,5,28)
{hEde}[ሄደ](15,32,5,28)	{IyehEde}[እየሄደ](27,25,28,15,32,5,28)
{meTa}[መጣ](20,28,9,31)	{IyemeTa}[እየመጣ](27,25,28,20,28,9,31)
{bela}[በላ](4,28,24,31)	{Iyebela}[እየበላ](27,25,28,4,28,24,31)

The algorithm of this affixation is:

```
Begin:
 For a given verb in number array
     Append integers 27, 25 and 28 at the start of the array
End:
```

Figure 3.4. Algorithm for prefixation of {Iye-}[እየ]

The second property of verbs is they can change the last *fidel* to seventh order (*sabI*).

Table 3.7. last *fidel*'s Change to *sabI*

Words	Last *fidel* changed to sabI
{wesede}[ወሰደ](26,28,13,28,5,28)	{wesedo}[ወሰዶ](26,28,13,28,5,33)
{hEde}[ሄደ](15,32,5,28)	{hEdo}[ሄዶ](15,32,5,33)
{meTa}[መጣ](20,28,9,31)	{meTto}[መጥቶ](20,28,9,1,33)
{bela}[በላ](4,28,24,31)	{belto}[በልቶ](4,28,24,1,33)

The following Algorithm, figure 3.5, changes the last *fidal* to sabI as shown in Table 3.7

```
Begin:
 For a given verb in number array
     If the last integer is equal to 31 then
          Remove the last entry and add 1 and 33
     If not replace the last entry with 33
End:
```

Figure 3.5.algorithm for last *fidel*'s Changing to *sabI*

Information about the subject can be attached on a verb and two examples are shown in Tables 3.8 and Table 3.9, and the derived algorithms are shown in figure 3.6 and 3.7 respectively.

Table 3.8. first person singular as subject in verbs

Words	First person singular
ወሰደ{wesede}(26,28,13,28,5,28)	ወሰድሁ[2]{wesedhu}(26,28,13,28,5,15,29)
ሄደ{hEde}(15,32,5,28)	ሄድሁ{hEdhu}(15,32,5,15,29)
መጣ{meTa}(20,28,9,31)	መጣሁ{meTahu}(20,28,9,31,15,29)
በላ{bela}(4,28,24,31)	በላሁ{belahu}(4,28,24,31,15,29)

Algorithm for first person singular subject in verb :

```
Begin:
 For a given verb in number array
     If the last integer is equal to 31 then
          Append 15 and 29 at the end of the array
     If not remove the last entry and append 15 and 29 at
     the end
End:
```

Figure 3.6. Algorithm for first person singular as subject in verbs

Table 3.9. third person singular feminine

Words	Third person singular feminine
ወሰደ{wesede}(26,28,13,28,5,28)	ወሰደች{wesedec}(26,28,13,28,5,28,2)
ሄደ{hEde}(15,32,5,28)	ሄደች{hEdec}(15,32,5,28,2)
መጣ{meTa}(20,28,9,31)	መጣች{meTac}(20,28,9,31,2)
በላ{bela}(4,28,24,31)	በላች{belac}(4,28,24,31,2)

And the algorithm for Table 3.9 is:

```
Begin :
 For a given verb in number array
     Append 2 at the end of the array
 End:
```

Figure 3.7. algorithm for third person singular feminine

In the same way rules for other pronouns (1st person plural, 2nd person singular and plural, 3rd person singular and plural) are developed.

One algorithm worth mentioning is the command form of a verb for second person singular feminine, which exhibit valuable phonetic property(there are also other cases which follow this characteristic while expanding, but one example is enough to show how to handle such a case). This property, as mentioned previously, is the main reason behind rearranging Amharic *fidel* (see Table 3.2) to get some pattern. It's evident that by relating Table 3.2 and Table 2.1 the rearrangement is sound from linguistic point of view. From computational angel, as shown in the following algorithm, it simplifies derivation/expansion of words by generating some pattern.

Table 3.10. third person singular feminine, command form: the case of alveolar phones

Words	Third person singular, command
{meta}[መታ](20,28,1,31)	{mci}[ምቺ](20,2,30)
{hEde}[ሄደ](15,32,5,28)	{hji}[ህጂ](15,6,30)
{geza}[ገዛ](7,28,17,31)	{gZi}[ግዢ](7,18,30)
{bela}[በላ](4,28,24,31)	{byi}[ብዪ](4,25,30)
{lefefe}[ለፈፈ](24,28,12,28,12,28)	{leffi}[ለፍፊ](24,28,12,12,30)
{lemene}[ለመነ](24,28,20,28,21,28)	{lemNi}[ለምኚ](24,28,20,22,30)
{seTe}[ሰጠ](13,28,9,28)	{sCi}[ስጪ](13,10,30)
{esebe}[አሰበ](28,13,28,4,28)	{esbi}[አስቢ](28,13,4,30)
{dabese}[ዳበሰ](5,31,4,28,13,28)	{dabSi}[ዳብሺ](5,31,4,14,30)

From Table 3.10, it can be observed that alveolar phonemes to shift to palatals on the application of vowel {-i}[-ኢ](,30). Figure 3.8 shows the algorithm for this case (there are cases still to be handled about third person feminine, see recommendation part for future works suggestion).

```
Begin:
Define variable alveolar {1,5,9,13,17,21,24}
 For a given verb in number array
     Staring from end of array remove the last two integers
     greater than 26
     If the last item in array is in alveolar then add one
     to that number
     Append 30 at the end of the array
End:
```

Figure 3.8. third person singular feminine, command form: the case of alveolar phones

3.5 Amharic Dictionary Development

Developing an Amharic dictionary containing all available words is difficult if not impossible. But as described earlier, it's possible to generate most forms by applying morphological expansion rules on stem verbs.

Table 3.11. implemented rules and their morphological information for transitive verbs

Rule number	Meaning
0	3rd person singular, masculine, past tense, affirmative
1	1st person, past tense, affirmative
2	1st person, past tense, affirmative
3	2nd person singular, masculine, past tense, affirmative
4	2nd person singular, feminine, past tense, affirmative
5	2nd person plural, past tense, affirmative
6	3rd person singular, feminine , past tense, affirmative
7	3rd person plural , past tense, affirmative
8-15	{rules 1-8} change subject to object
16-31	{Rules 0-18 } change past tense to continues tense
32-38	{Rules 0-8} negative instead of affirmative Rules for 2nd person masculine and 3rd person feminine are similar
39	2nd person, masculine, command
40	2nd person, feminine, command

Chapter 4

The MT system Design and Implementation

4.1 Introduction

As indicated in chapter one of this thesis, the objective of this research is to propose, design and implement the MT system for Amharic to ESL. As mentioned time and again, this is the first of its kind for ESL. After extensive study on the subject, the system is designed and the implementation for small vocabulary is done.

This chapter goes through each part of the system with explanatory examples whenever necessary.

4.2 Overall System

From the end user point of view the system can be described as a system that accepts Amharic words in text format and outputs the equivalent in ESL by using signing avatar. To achieve this function three processing main sub and one supporting systems are required.

The figure below shows the block diagram of the overall system. Consecutive sub-topics explain each sub-system in detail; except the morphological expander unit, which is covered in previous chapter.

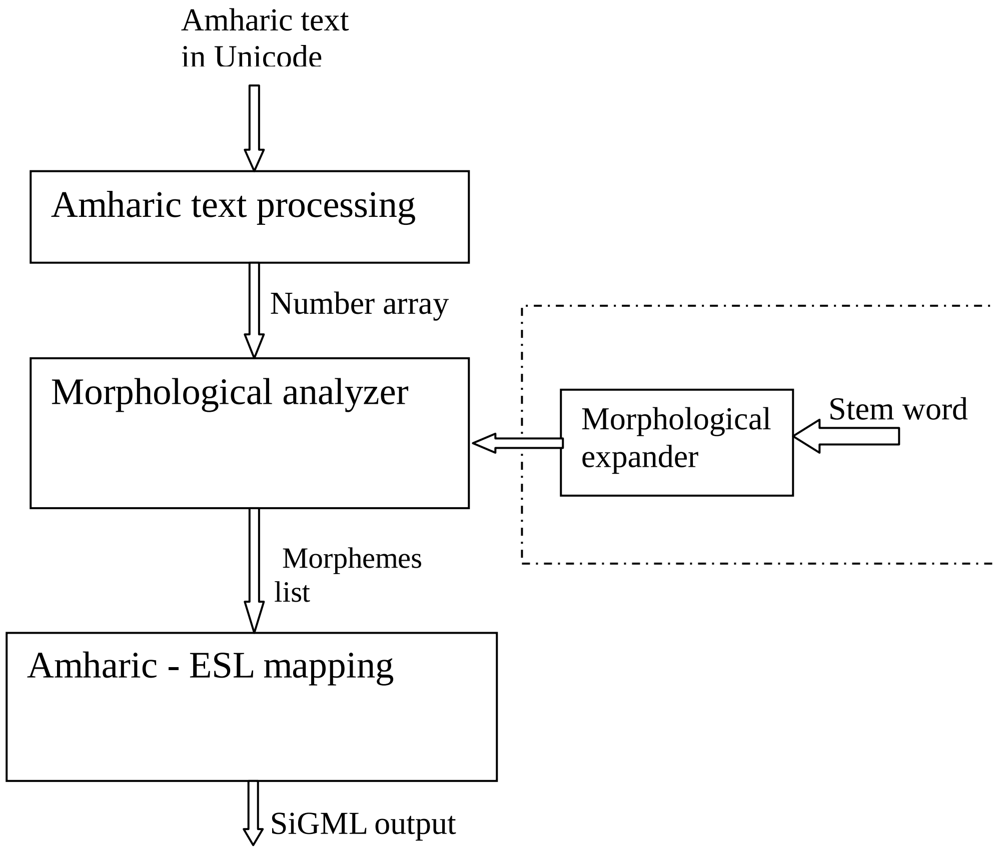


Figure 4.1. Amharic-to-ESL MT system architecture

4.3 Amharic Text Processing

Text processing part is responsible to change a given Unicode written Amharic input to ASCII representation, by using the SERA representation. This part also implements the conversion from *Ge'ez* or SERA to number array and vice versa.

There are at least two good reasons to represent the words in integer format. The first reason is to avoid the overwriting of some verbs in the operating system environment as shown below.

For computational simplicity, each word is saved with its expanded word forms in a single file, with the root word as the file name. If the ASCII is used as it is some words will be missed, due to overwriting, from the dictionary. For example, while working on this research it is found that from the two words {meta}[መታ] and {meTa}[መጣ] only one can be stored on the same folder (checked on windows® operating systems). But by using numbered representation approach the former will be (20,28,1,31) and the later takes (20,28,9,31).

The second advantage of integer representation is computational efficiency. It is needless to say that integer operations are more efficient (speed and memory usage) than string operations.

4.4 Morphological Analyzer

This sub-system is the backbone of the MT system. It contains the morphological expander, maintains the Amharic dictionary and when invoked it'll load the dictionary on the memory with all the needed information.

The morphological expander was dealt in detail in the previous chapter. For a given word depending on its type the morphological expander applies available linguistic rules to form other possible words. Each word found by morphological expansion will be added to the Amharic dictionary.

The dictionary of each word contains information about the stem word and the rule applied upon the stem to generate it. The generated word is used as a key in the dictionary and the value is morphological information attached to it. In other words, the morphological analysis is done in a forward fashion (the other method is backward analysis in which the stem and other morphemes extracted from a given verb by applying rules on the complex or expanded word).

On this sample dictionary which contains more than eight hundred words, if the keys are put on a random, the processing time will be in order of O(n). This will make the system slower as the size of the dictionary grows larger and larger. One can increase the size of the dictionary easily by adding rule(s) on the existing words or by adding a word(s) and applying the existing rules. But if the dictionary keys are arranged, the binary search method can be applied and the computation time will be in O(log n).

The second issue related with the performance is where to put the dictionary. One of the possibilities is to put it on the storage (hard disk) which will save memory space at the expense of speed. The other alternative is loading the whole dictionary on the memory which speeds up the operation.

Considering the high probability of using this MT in real time environment, the arranged-dictionary is loaded to the memory. To minimize the memory usage, which will be the bottleneck for large dictionary, the keys and values are put in an array of integer (primitive data type).

There are two possible scenarios for a given Amharic word input

1. The word is found in Amharic dictionary
2. No entry in the Amharic dictionary

The sub system sends the corresponding information to the mapping sub unit. If the word is found in the dictionary, in that case there is a key equal to the input word, the sub-system sends the root word and a rule applied on it to the mapping sub system.

If there is no entry in the Amharic dictionary, the word will be sent to the mapping system with the tag showing it's not a recognizable word by the system.

The system is designed to accept multiple words (separated by space) and can process a combination of recognizable and unrecognizable words.

4.5 Amharic –ESL Mapping Sub System

The ultimate goal of this system is to change a given text word(s) in Amharic to ESL. The morphological analyzer sub-system prepares the word(s) and this sub-system holds the conversion key.

As discussed in the previous sub topic, the output from the morphological analyzer can be a root word with corresponding expansion-rule, or an unrecognized word with the tag indicating it's not found in the dictionary.

If the word is found then the mapping system looks for the corresponding meaning in the ESL dictionary. One Amharic entry may result in more than one ESL entry. For example the Amharic word {elbelam}[**አልበላም**] in ESL is a combination of signing equivalent of {bela}[**በላ**] plus {aydelem}[**አይደለም**]. In this fashion the mapping can be done for words found in both dictionaries.

Some Amharic words may not have ESL equivalent (e.g. {gXbet}[**ግሽበት**], {qelaweTe}[**ቀላወጠ**], {wudqt}[**ውድቅት**]). This is not the case only for this system, even professional translators use finger-spelling when the word does not have ESL equivalent. So, when the sub-system found a word of this kind, it'll map to the finger-spelling (manual signing) ESL.

If the word which comes from the morphological analyzer is unrecognized word in Amharic dictionary, the mapping will be sent directly to the manual signing without looking in the ESL word dictionary.

To simplify the mapping, in the ESL dictionary folder ESL-words are saved by using Amharic equivalent numbering method. The contents in each file are the SiGML representation of that ESL-word.

4.6 ESL Representation

There is a growing trend to use signing avatar as an SL output instead of recorded videos. The pros and cons of each system are shown in detail in chapter two. Different research groups developed different avatar depending on the requirement of their projects. Some of them are commercial and others are open for research purpose.

For this thesis the avatar '*anna'* developed by School of Information Systems, University of East Anglia, Norwich, United Kingdom at JASigning project is used. As many SL animation projects, JASigning uses SiGML text as an input to the avatar.

SiGML is XML-based markup language and is suitable for the computer system and contains text tags written using ASCII character set only. It is developed with TCP/IP application in mind. But from the transcriber point of view it's a difficult task to represent a sign language in text tag.

Different sign language writing methods are suggested to overcome this problem. One of them, the HamNoSys has found a wide range of acceptance. It's developed for computer based researches on sign language by Hamburg University research group.

4.7 ESL Alphabets

Since ESL is created or born in Ethiopia, it shares some characteristics of Amharic language. One of the similarities between Amharic and ESL is their alphabetic representation and ordering. Like Amharic, ESL has thirty-four (including {ve}[ቨ] basic alphabetic shapes with seven type of movement each. Also as the case with Amharic, there is special movement for Amharic *diqala fidelat* in ESL.

As it's done to Amharic, it is better to represent the thirty-four basic shapes and attach the seven uniform movements than transcribing the 238 (34×7) HLOMEs. The author transcribed these thirty-four *fidel* shapes(see Table 4.1) and the seven movements(see Table 4.2).

Table 4.1. ESL alphabets

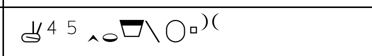
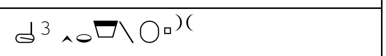
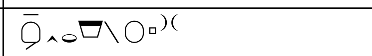
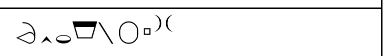
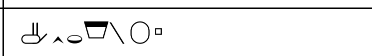
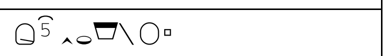
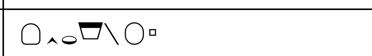
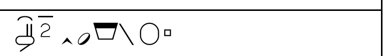
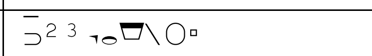
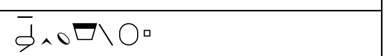
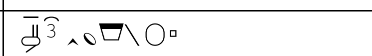
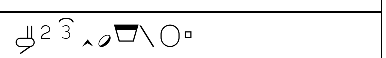
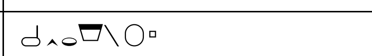
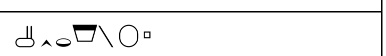
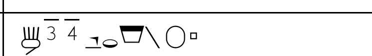
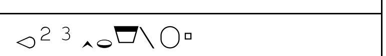
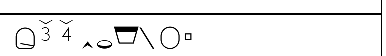
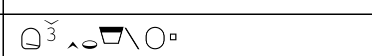
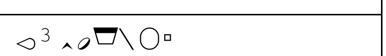
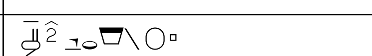
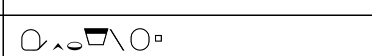
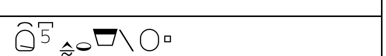
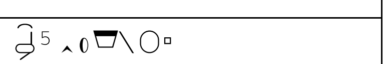
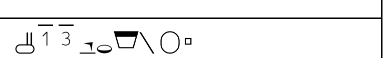
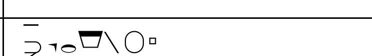
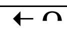
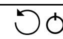
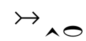
Amharic *fidel*	ESL in HamNoSys	Amharic *fidel*	ESL in HamNoSys
ሀ{he}		ለ{le}	
ሐ{he}		መ{me}	
ሠ{se}		ረ{re}	
ሰ{se}		ሸ{Se}	
ቀ{qe}		በ{be}	
ተ{te}		ቸ{ce}	
ኀ{he}		ነ{ne}	
ኘ{Ne}		አ{e}	
ከ{ke}		ኸ{he}	
ወ{we}		ዐ{e}	
ዘ{ze}		ዠ{Ze}	
የ{ye}		ደ{de}	
ጀ{je}		ገ{ge}	
ጠ{Te}		ጨ{Ce}	
ጰ{Pe}		ጸ{Se}	
ፀ{Se}		ፈ{ፈe}	
ፐ{pe}		ቨ{ቨe}	

Table 4.2. the seven movements of ESL alphabets

Ge’ez	kaIb	Sals	rabI	Hams	sads	sabI
(no movement)						

In chapter two, it is indicated that the avatar of SiGML Service Player accepts SiGML notification and there is a one to one correspondence between HamNoSys and SiGML. By using the eSIGN-editor, which is available on JASigning project site (http://vhg.cmp.uea.ac.uk), each transcription in HamNoSys is converted to SiGML. For example the SiGML equivalent of {le}[ለ] is

```
Line-1   <hamfinger23spread/>
Line-2   <hamthumbacrossmod/>
Line-3   <hamringfinger/>
Line-4   <hampinky/>
Line-5   <hamextfingeru/>
Line-6   <hampalmd/>
Line-7   <hamshoulders/>
Line-8   <hambetween/>
Line-9   <hamhead/>
Line-10  <hamlrbeside/>
Line-11  <hamclose/>
```

Interpretation of the SiGML:

The first four lines represent the hand shape, from which the first line declares hand shape to be as shown in figure 4.2 (number notation in SiGML for fingers is 1- thumb, 2-inex finger, 3- middle finger , 4-ring finger, and 5- pinky). Therefore, `finger23spread` means the hand shape of fist but with index finger and middle finger are in straight mode and are spread.

Figure 4.2. Hand-shape: finger23-spread

But one can see that this is not the desired hand-shape. First, the thumb should be across the fingers and secondly stretched fingers should be ring-finger and pinky. This is done by line number 2 (thumb mode), 3 (ring finger), and 4 (pinky) of the above SiGML code.

Line numbers 5 and 6 made hand-orientation in a way palm facing the observer. The last four lines indicate that the location is between head and shoulder and is close to body. Since it's in first order there is no movement. The resulting output is shown below as signed by *anna*.

Figure 4.3. *fidel* {le}[ለ] signed by *anna*

In this way all alphabets and selected words are transcribed in SiGML.

Full table of HamNoSys with its SiGML description and Unicode table can be found from Hamburg official web site (www.sign-lang.uni-hamburg.de.) it's also included on this thesis accompanying CD which can be found at the graduate study program of AAiT, AAU.

4.8 Words in ESL

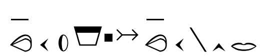
The books prepared for sign language beginners by MYSD-Hossana [5][20] contain about 400 ESL words. For this particular study more than ten percent of the words are transcribed in to HamNoSys. Since there is no one to one correspondence between Amharic and ESL, these more than 50 words can represent much more words in Amharic. For example the ESL word for {bela}[**በላ**] can represent many derivations of that word depending on the context, such as {belehu, belan, lbla, enbla} and so on.

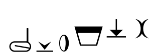
ESL words that are transcribed into HamNoSys are given in appendix B. The words that need some rectification are discussed in chapter six.

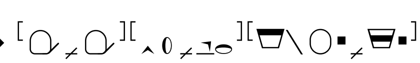
4.9 Mapping rules

While doing mapping, there are at least three different cases that the MT faces. The first and the simplest one is there is an exact much to the Amharic word in the ESL dictionary.

For example:

{belahu}[**በላሁ**] ➔ 

{InE}[**እኔ**] ➔ 

{egziabhEr}[**እግዚአብሔር**] ➔ 

In this case the MT just peaks the appropriate word. Many Amharic words to a single ESL word can be included in this category. For example {InE}[**እኔ**] and {lenE}[**ለኔ**] can be represented by the same ESL word.

But this is not always the case, there is the second type of word; a word in Amharic may map to two or more ESL word.

For example:

{msganayE}[**ምስጋናዬ**] ➔ [illegible] + [illegible]

*(msgana[**ምስጋና**] + yene[**የኔ**])*

{etbelam}[**አትበላም**] ➔ [illegible]

*(eswa[**እሷ**] + bela[**በላ**] + eydelem[**አይደለም**])*

(In the above example {eswa}[**እሷ**] may be omitted depending on the context).

The third type, and that is a rich area of research for future researchers, is ambiguous words. In this type of mapping the word may take different ESL meaning depending on the context of Amharic sentence. One good example is {felege}[**ፈለገ**]. It can be mapped in ESL either to ‘he wanted’ or ‘he searched’. Another example is {eleN}[**አለኝ**] which can mean ‘he said to me’ , ‘I have’ , or ‘I felt’ and all three are different in ESL.

The MT system developed in this work implements the first two cases. To handle the last case the system should go beyond morphological analysis and it’s not feasible to handle all the problems in one MSc thesis.

Chapter 5

Evaluation

5.1 Introduction

To best of my knowledge, this Amharic text to ESL conversion system is the first of its kind, the design and implementation was all done as a starting research regarding the ESL part. Normally, MT systems for SLs are group projects and take years for completion. The motivation behind this thesis is also to open door for such type of big project on ESL.

Regardless of the above mentioned achievement and faced problems (like shortage of reference materials), there should be some way to evaluate the system for further research and improvement. From different possible methods to evaluate MT system, evaluation by native signers was chosen. This chapter discusses finding of the evaluation.

5.2 Evaluation

The method followed to evaluate this MT system is as follows. The evaluators (three deaf individuals, which are native ESL users) were asked to write down the ESL word signed by the avatar from a decoded input. They were also asked to write any comment they had on that specific word, such as, needs little refinement, satisfactory and so on. For the finger spelling part, each alphabet is presented and they were asked to give values 1(very poor), 2(poor), 3(satisfactory), 4(good), and 5 (perfect). The following table, Table 5.1 and 5.2, summarize evaluation data.

Table 5.1. Assessment of finger spelling

Signer	No of alphabets	Assessment value				
		1	2	3	4	5
1	34	5	1	3	0	25
2	34*	2	8	6	5	12
3	34	0	4	8	5	17

* signer-2 jumped one alphabet with-out giving value

Table 5.2 words evaluation in ESL dictionary

Signer	No of words	Correct	Incorrect	
			replaced	Jumped
1	28	21	1	6
2	28	23	1	4
3	28	25	1	2

5.3 Discussions

From evaluation table for finger spelling more than 50% of the alphabets are perfectly transcribed. If a word which is not in ESL dictionary is given to the system, the output sign can be perceived, score greater than or equal to 3, is about 80% . i.e. 81 out of 101.

If the word is in the ESL dictionary, they can perceive more than 82% (67 out of 84). The overall system performance is, therefore, more than 80%.

With additional effort on ESL transcription part, the percentage of retrieval can be increased and as a first work the result is very promising for further research.

In addition to above mentioned evaluation, there was a discussion on the signs with them and their suggestion is included in recommendation part for future work.

Chapter 6

Conclusion and Recommendations

6.1 Conclusion

The purpose of this thesis was to design and implement the first Machine Translation system for Amharic text to Ethiopian Sign Language conversion. The end result expected from this research was a system that is capable of analyzing Amharic text and synthesizes equivalent representation in ESL.

To achieve the result, this research needed at least three areas of study: source language, target language, and MT systems.

Amharic: source language

In this thesis much emphasis was given for the phonetic and morphologic characteristics of Amharic language. On phonetic study, there were two problems identified. The first problem is the standard way of Amharic alphabets (*fidel*) arrangement doesn't have appreciable relationship to their phonetic behavior. This disordered arrangement hinders some researchers from applying phonetic rules easily using computers (computer programs). One of the contributions of this research is suggesting a new arrangement of Amharic *fidel* for computerized linguistic analysis.

The second problem faced was related to computational efficiency. There are more than 231 phoneme-symbols in Amharic. But, it's evident that their number can be reduced, again by applying phonetic theory of the language. By doing so the number of necessary symbols for computer systems was reduced to 34. In addition to this, each symbol was represented by an integer rather than using characters. Since the symbols are in number format it'll reduce the computational time for mathematical operations, which were used in the morphological expander sub-unit.

Regarding morphology of Amharic, different approaches of researches were studied. Many linguists suggested different methods to analyze Amharic morphology. One of which was morphological expansion method. After thorough study, the expansion method was selected as the best fit for MT systems of limited vocabulary.

Morphological expansion program was implemented based on the language property. From different categories of Amharic words, due to their morphological expandability, verbs were made to expand up to 40 forms by applying different rules. By doing so on 20 verbs the Amharic dictionary of more than 800 entries was maintained.

ESL: target language

A common problem for all researchers on MT system with Amharic language was the lack of enough resources to carry out their research; and the problem worsens for ESL. Unfortunately, MT system design requires not only knowledge of source language but also knowledge of target language. Books published so far on ESL are not more than 5. So, one cannot depend on them to get enough knowledge. There were two options to tackle the problem; either hire (or find volunteer) professional ESL user, or learn the language. Learning ESL was feasible solution and the author studied ESL class as part of this research work.

There are many research projects going on different sign languages. These projects, many of them are large scale, were studied carefully and contact with different researchers were made. As a result, it was decided to use the HamNoSys representation to transcribe/store ESL words and alphabets.

MT system design

Before designing, different types of MT systems were studied. The study showed that rule based approach was better fit for this research. From three available RBMT systems (direct, transfer, and interlingua) direct approach was realized. Because of unavailability of enough information on grammatical structure of ESL, realization of the other two RBMT was unadvisable.

In this direct RBMT system, source language, Amharic, passed through morphological analysis stage only. Result of analysis was passed to transfer/mapping system. Final stage of the system synthesizes ESL morphology. For words which are not recorded on ESL dictionary, manual signing was used instead of just ignoring unrecognized ones.

The output of the system was given to the SiGML Service Player of JASigning project. This player animates a given ESL word by using signing avatar.

Evaluation of the system was done by native signers and overall transfer ratio was calculated to be more than 80%.

6.2 Recommendations

While working on this thesis, there were many problems found that need future research. Below some of them are listed.

- Morphological irregularities

  Some phonemes like; {ge}[ገ],{we}[ወ], and {he}[ሀ] tend to change the morphological expansion rule pattern. There should be some linguistic theory behind the problem and there should be a way to predict and handle such cases using an algorithm.

- Ambiguity

  Ambiguity is and has been the ever existing problem of MT systems. Different researchers tried to handle ambiguity of Amharic words. But most of them are not based on context. It was shown in this thesis that one Amharic word can have three different meanings in ESL depending on the context.

- Better transcription

  From discussion with evaluators of this MT system and other individuals, there are some words/alphabets need further refinement. Some of the problems arise from ESL hand shape complexity. Are the problems related to HamNoSys or the avatar? And how to solve it (The HamNoSys guys promise to add more symbols if the need arises).

- Grammar of ESL

  One good research area, most probably to linguistics students, is grammatical structure of ESL. At the time of this research, there is no publication on the subject area. Solving this problem opens a door for other researches like transfer based MT system for ESL.

- MT systems for other Ethiopian languages

  There is a great need to translators from one Ethiopian language to another, especially in official cases involving both regional and federal government offices. It'll have academic and socio-economic advantage if the is MT system, say, from Afan Oromo to Amharic and vice versa.

References

[1] Abiyot Bayou. (2000). Design and Development of Word Parser for Amharic Language. Unpublished Masters' Thesis. Addis Ababa University, Addis Ababa.

[2] Admasu A. & Derso T.(1991). *Deaf people communicating with professionals*. In Partnership between Deaf People and Professionals. Proceedings of a Conference, Rabat, Malta.

[3] ALPAC. (1966). *Language and machines: computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee. Washington, DC: National Academy of Sciences.

[4] Arnold, D.J., Balkan, L., Meijer, S., Humphreys R.L., & Sadler L. (1996). *Machine Translation: an Introductory Guide*. London: NCC Blackwell Ltd.

[5] Asefa Guta & Behiru Tamiru. (2006). ***የምልክት ቋንቋ ለጀማሪዎች - 2ኛ እትም*** *[Sign Language for beginners. 2nd ed.]*. Hossana: Mekane Yesus School for the Deaf.

[6] Atelach Alemu Argaw. (2002). Automatic Sentence Parsing for Amharic Text: An Experment using Probabilistic Context Free Grammars. Unpublished Master's Thesis, School of Graduate Studies, Addis Ababa University.

[7] Baye Yimam. (2002). ***አጭርና ቀላል የአማርኛ ሰዋስው*** *[short and simple Amharic Grammar]*. Addis Ababa: Alpha publishers.

[8] Daniel Gochel Agonafer. (2003). An Integrated Approach to Automatic Complex Sentence Parsing for Amharic Text. Unpublished Master's Thesis, School of Graduate Studies, Addis Ababa University.

[9] Daniel Yacob. (1997). The System for Ethiopic Representation in ASCII (SERA) – 1997

standard. Retrieved May 24, 2011: http://www.abyssiniacybergateway.net/fidel/sera-97.html

[10] Dorr, B.J., Jordan, P.W., & Benoit, J.W. (1998). *A survey of Current Paradigms in Machine Translations*. (Technical Report: LAMP-TR-027/UMIACS-TR-98-72/CS-TR-3961). University of Maryland, College Park.

[11] eSIGN Project website (*Essential Sign Language Information on Government Networks*). Retrieved April 30, 2009, http://www.sign-lang.uni-hamburg.de/eSIGN.

[12] Ethiopian Central Statistical Authority (ECSA). (1998) . *The 1994 Population and Housing Census of Ethiopia: Results at Country Level.* Vol. 1 Statistical Report48. Addis Ababa.

[13] Hank, T. (Ed.).(2001). *ViSiCAST Deliverable D5-1: Interface Definition*. (Technical Report).

[14] Huenerfauth, M. (2003). A Survey and Critique of American Sign Language Natural Language Generation and Machine Translation Systems. (Technical Report MS-CIS-03-32). Computer and Information Science, University of Pennsylvania.

[15] Huenerfauth, M. (2006). Generating American Sign Language Classifier Predicates For English-To-ASL Machine Translation. PhD dissertation. University of Pennsylvania.

[16] Hutchins, W.J. & Lovtsky, E. (2000). Petr Petrovich Tryanskii (1894-1950): a forgotten pioneer of machine translation. *Machine Translation*, 15(3), 187-221.

[17] Hutchins, W.J. (1979). *Linguistics models in machine translation*. *UEA Papers in Linguistics*, 9, 29-52.

[18] Hutchins, W.J. (1999). The development and use of machine translation systems and computer-based translation tools. *International Conference on Machine Translation & Computer Language Information Processing*, Proceeding of conference, Beijing, China.

[19] Hutchins, W.J. (2005). The history of machine translation in a nutshell. Retrieved April 10, 2011,from http://ourworld.compuserve.com/homepages/wjhutchins/Nutshell.htm

[20] Mekonen Mulat, Asefa Guta, & Atsedemariam Milkyas. (2010). ***የምልክት ቋንቋ ለጀማሪዎች፡ 3ኛ እትም*** *[Sign Language for beginners. 3^{rd} ed.]*. Hossana: Mekane Yesus School for the Deaf.

[21] Mesfin Birile. (2008). Synthetic Speech Trained – Large Vocabulary Amharic Speech Recognition System (SST-LVASR). Unpublished Master's Thesis, school of Graduate Studies, Addis Ababa University.

[22] Neidle, C., Kegle, D., MacLaoughline, D., Bahan, B., & Lee, R.G.(2000). *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. Cambridge, MA. MIT Press.

[23] Official SignWriting site. Retrived June 12,2011, http://www.signwriting.org

[24] Prillwitz, S., Leven, R., Zienert, H., Hanke, T. , & Henning, J.(1989). *Hamburg Notation System for Sign Languages – An introductory Guide*. International Studies on Sign Language and the Communication of the Deaf, Volume 5, Institute of German Sign Language and Communication of the Deaf, University of Hamburg.

[25] Probst, K. (2005). Learning Transfer Rules for Machine Translation with Limited Data. Unpublished Doctoral Dissertation. Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.

[26] Selam, Y.(2009). A generic framework for Arabic to English machine translation of simplex sentences using the Role and Reference Grammar linguistic model. Unpublished Master's Thesis, Institute of Technology Blanchardstown, Dublin, Ireland.

[27] Solamon Teferra Abate. (2005). Automatic Speech Recognition for Amharic. Unpublished Doctoral dissertation. University of Hamburg.

[28] Tadesse Anberbir, Gasser, M., Takara, T., & Yoon, K.D. (2011). Grapheme-to-Phoneme Conversion for Amharic Text-to-Speech System. Conference on Human Language Technology for Development, Alexandria, Egypt.

[29] Tesfaye Bayu Bati. (2002). Automatic Morphological Analyzer: An Experiment Using Unsupervised and Autosegmental Approach. Unpublished Master's Thesis. Addis Ababa University.

[30] UEA eSIGN project. Retrieved June 3, 2011, http://www.visicast.cmp.uea.ac.uk/esign.

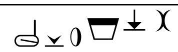
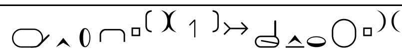
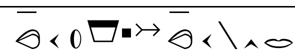
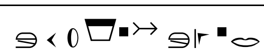
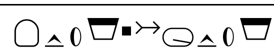
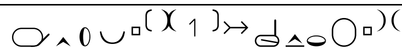
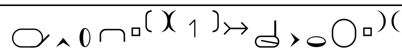
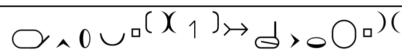
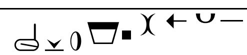
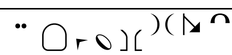
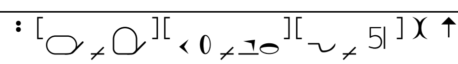
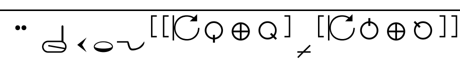
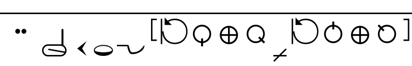
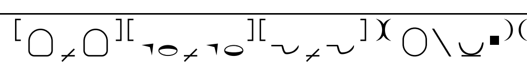
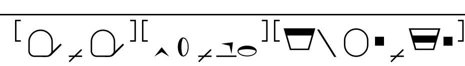
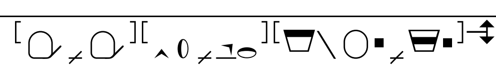
[31] United Nations. (1983). *World Programme of Action Concerning Disabled Persons*. New York.

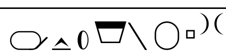
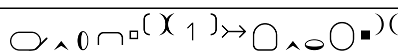
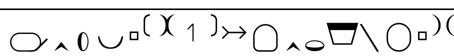
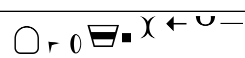
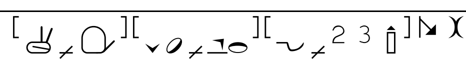
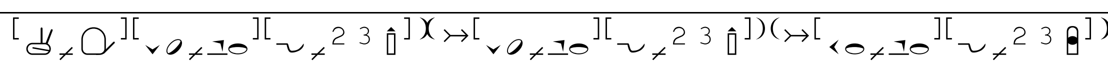
[32] Wang, Y. (1998). Grammar Inference and Statistical Machine Translation. Doctoral Dissertation. School of Computer Science, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA.

Appendixes

Appendix A. ESL manual alphabets [5][20]

Appendix B. transcription of selected ESL words in HamNoSys

Word	ESL in HamNoSys
እኔ	
አንተ	
በላ	
ጠጣ	
ወሰደ	
አንቺ	
እሱ	
እሷ	
እኛ	
ምስጋና	
ረዳ	
ሄደ	
መጣ	
ተኛ	
እግዚአብሔር	
ዛፍ	

0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
የኔ	
የአንተ	
የአንቺ	
የኛ	
ቆመ	
ወደቀ	

Declaration

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

Name: Dagnachew Feleke Wolde Signature:______________
Place: Addis Ababa
Date of submission: October 2011

This thesis has been submitted for examination with my approval as advisor.
Dr. Eneyew Adugna Signature: __________________
Advisor