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

DESIGN AND DEVELOPMENT OF HUMAN-AIDED RULE-BASED ENGLISH-AMHARIC MACHINE TRANSLATION PROTOTYPE

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

BY

YEHENEW SHIFERAW

JUNE 2004

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

DESIGN AND DEVELOPMENT OF HUMAN-AIDED RULE-BASED ENGLISH-AMHARIC MACHINE TRANSLATION PROTOTYPE

BY

YEHENEW SHIFERAW

Name and Signature of Members of the Examining Board

Ato Getachew Jemaneh, Chairman, Examining Board _____

Professor Narayana Murthy Kavi, Advisor _____

Athanassia Fourla, Advisor _____

Dr. Osei Nana Adjei, External Examiner _____

ACKNOWLEDGEMENT

Time and again, I would like to forward my admiration to my advisors Athanassia Fourla and professor Narayana Murthy Kavi. This thesis would not have come true without their professional comment and recommendation.

I'd like to thank Ato Sebsibe H/Mariam for giving me the inspiration and for helping me sustain in the darkest days of self-doubt. Finally, I'd like to thank my family for having confidence in me at all times.

DEDICATION

THIS THESIS IS DEDICATED TO THE ETERNAL PARTNER:

FOR HELPING ME DISCOVER THE SERENE MEANING (ILUE).

TABLE OF CONTENTS

List of Tables

List of Figures

List of Appendices

ABSTRACT

Today, Machine Translation systems have been developed for different language pairs, which have a relatively wide use nationally and/or internationally. But, the Amharic language, despite its centuries of existence, has not yet been privileged to reap the benefits of this technological arena. This study has tried to develop a prototype translating system that translates texts from English into Amharic. The Source language is chosen to be English due to the fact that together with Amharic, English is serving as a teaching-learning language in the educational system of Ethiopia. The research is an attempt to improve and preserve the role of the Amharic language in the current electronic age of commerce and education.

In the research, designing and structuring the translation process and resources implemented has been customized from the Shakti Machine Translation (MT) system, which translates from English to Hindi. The Shakti kit uses statistical as well as rule-based approaches for processing language. It uses constituent structure at the chunk level and dependency relations at the sentence level of analysis. It has specialized components for word sense disambiguation, parsing, Preposition attachment, phrasal verb identification, transfer grammar, sentence and word generation, and many others.

The thesis, in short, describes the processes of transfer machine translation system using rule-based approach. It discusses the whole process of analyzing the source language, the necessary features of syntactic transfer from the English language into Amharic and synthesizing the target language. The results obtained using sample sentences and phrases taken out from the toy article seem to encourage further large-scale research to be launched, especially with the aim of developing a full-fledged translation system.

List of Abbreviations

AgrO	Agreement in Object
AgrS	Agreement in Subject
ALPAC	Automatic Language Processing Advisory Committee
ALPS	Active Learning Practices for Schools
CFG	Context Free Grammar
DT	Determiner
DTD	Document Type Definition
EBMT	Example-Based Machine Translation
ENGSPAN	English to Spanish Medical Words and Phrases
ET	Event Time
IT	Information Technology
KBMT	Knowledge-Based Machine Translation
LOGOS	MT system for English to French translation
LT	Language Technology
METAL	MT system for German to English translation
MT	Machine Translation
NG	Noun Group
NLP	Natural Language Processing
NN	Singular Noun
NNS	Plural Noun
NP	Noun Phrase
PERL	Practical Extraction and Report Language
POS	Part of Speech
RT	Reference Time
SGML	Standard General Mark-up Language
SOV	Subject-Object-Verb
SSF	Shakti Standard Format
SVO	Subject-Verb-Object
TAM	Tense Aspect Modality
TL	Target Language
VB	Verb, Base Form
VBD	Verb, Past Tense
VBG	Verb, Gerund/Present Participle
VBZ	Verb, 3rd Person Singular Present
VGADV	Adverbial Verb Group
VP	Verb Phrase

CHAPTER ONE

INTRODUCTION

1.1 Background

Numerous researches conducted over the years have proved that Ethiopia is an ancient country with a very rich cultural heritage possessing its own alphabet and literature. One of the distinguishing features of the Ethiopian culture is the existence of different languages. Ethiopia has an old tradition of writing, which existed since the Axum civilization. According to historical facts, the Axumite kingdom was at it's height during 300 A.D., and parallel to this the Amharic language has been experiencing all transitional states that a language requires to fully develop for the longest time ever.

For centuries Amharic has been the official language of the ruling class and still is the official language of the federal government of Ethiopia, and thus has a well-established writing system and well-standardized norms of spoken and written usage, Lesalu, (1973). Amharic is also the widely spoken Semitic language next to Arabic and Hebrew. Together with English, Amharic is serving as a teaching-learning language in the educational system of Ethiopia.

The Amharic language has 33 basic characters with each having seven forms for each consonant-vowel combination. Atelach et al (2003), have reinforced the previous idea by stating that Amharic is a Semitic language of the Afro-Asiatic Language Group that is related to Hebrew, Arabic and Syrian. Amharic, the sellabic language, uses a script, which originated from the Ge'ez alphabet. They further have stated that unlike Arabic, Hebrew and Syrian, the language is written from left to write.

In the Information age that we're dwelling currently the Amharic language spoken by a larger number of people is among the very few languages of the world being incorporated with the computer. Several scholars have analyzed the language for academic purposes. For the past decade a good many people have been using the language for different application and advanced processing tasks. This in turn has resulted in having a large amount of databases and repositories in Amharic stored in electronic form.

Even though there are a number of software packages on the market to be used for the Amharic language, many of them seem to lag behind the standards of current technology being employed by other developed languages. None of the software available or any of the other Amharic software support language specific utilities like spell checking, grammar support, online thesaurus, etc. They further claimed that the absence of such word processing tools for Amharic language has in effect made word processing activities in the language incomplete.

The trend of globalization and the ubiquitous technology of the Internet direct all computational activity around the globe into an intricately interwoven symmetrical cyber village, where all the languages of the world should convey messages compatibly and equivalently in the already available systems. Since an overwhelmingly large number of people cannot make use of the huge information available on the Internet unless translated to local languages, there is a need to translate documents in other languages. This makes the development of machine translation systems more useful.

Unfortunately, only some languages like English have the privilege of utilizing the basic working applications or language technologies and hence dominate the efficiency and productivity of output to be achieved by languages with backward technological backup like Amharic. And this fact has made Amharic to suffer severely from lack of computational linguistic resources. For the language to survive in the information age, it requires the development of basic technological tools for Amharic. Natural language processing tools like morphological analyzers, POS taggers, phrase recognizers, word sense disambiguation programs, parsers, translation aids and so on are required for the development of Amharic language processing software.

Currently, the writer of this proposal has come to learn through personal communication, that Sisay Fisseha is conducting a research on Machine translation for the Amharic language. Apart from that to the best of the writers' knowledge there

is no research conducted directly on Machine Translation (MT) for the Amharic language. Nevertheless, there are related studies in relation to linguistic computation aspects of Amharic language.

There are very few attempts that tried to face the aforementioned problems and were somehow aimed at integrating the existing technological applications and apply Natural Language Processing techniques for the Amharic language. Among these: Mulu (2001) has developed English – Tigrinya Machine-readable bilingual dictionary that is an input for machine translation. Abiyot (2000) had designed and developed word Parser for Amharic language. Atelach (2002) has also designed an automatic sentence parser for Amharic text, and Sebsibe (2001) exercised the development of an English-Amharic electronic dictionary. Along with this Mesfin (2001) and Kibur (2002) have developed an automatic part of speech tagging for Amharic Language and an automatic morphological synthesizer for Amharic perfective verbs, respectively.

The limited number of researches in relation to MT with regard to the Amharic language shows that there is a vital urgency to bridge the gap of the current technological endeavors in the area of MT. This in the end, if properly integrated and implemented would be for the benefit of the majority of the population seeking knowledge, but hindered by the foreign language obstacle.

In order to design a translation system, there are different components to be employed. For this research the majority of designing and structuring the translation process and resources has been adopted from the Shakti Machine Translation system from English to Hindi.

1.2 Statement of the Problem and Justification

Manual translation has been regarded by translators as a repetitive, monotonous and thus boring, but at the same time, difficult job. Since translation requires a profound knowledge both of the source and target languages, people who are qualified for this job are rare and hence there is a shortage of translators. Due to the drawbacks mentioned in the background section, the problem of manual translation is severely reflected in the current situation of our country. The Adult Education Department (1975 E.C) of Ethiopia has identified the following as the major problems encountered from human translation services in Ethiopia:

- Lack of skilled manpower
- Finance
- Absence of authorized body to translate new political, economical, social, cultural and scientific and technological terms [which resulted in] inconsistency is introduced in using or translating such terms.

Some of these problems could be solved using MT. Using MT would minimize cost and maximize speed, both of which are the current demands of the business community. The competitive business community communicates worldwide on daily

bases using the Internet technology. Therefore, they need an urgent inter-language communication tool that can at least provide a draft-quality translation to determine or judge the relevance of the documents. It has repeatedly been proven that MT aids human translators to do their job more efficiently. Locke (1955) puts it statistically that MT increases their productivity by 30% or more.

The situation is fast changing due to the ongoing commercial growth and the influence of new research. MT is becoming a tool for translation by freelancers and small organizations. Cheap translation assistants are also making their way to market to help small companies and individuals write letters, e-mails and business reports in foreign languages.

Various researchers have concluded unanimously on the advantage of having most of the resources on different subjects and fields using the local languages of a country. The following are the privileges to have the literature in Amharic in Ethiopia:

- The mother tongue is the core identity of the people's culture.
- To use both the mother tongue and other languages at the same time is an advantage and is just like becoming a bicultural or polyglot.
- The literature is more convenient to the readers so that they can benefit from it. However, if the literature is in foreign languages, the readers switch off and ignore the publication.

- The literatures should express the specific bicultural, bilingual conditions of the indigenous group, with absolute respect for its cultural uniqueness and human identity.
- It is desirable to provide literacy and literature in everybody's mother tongue.
- Illiteracy may decrease.
- The mother tongue is the principal vehicle for acquiring both linguistic and non-linguistic knowledge.
- Fewer dropouts from schools, fewer failures in examinations, increase of total number of students and greater fluency in the national language.
- Greater number of community projects from bilingual schools, greater reading comprehension and faster learning of the national language; a larger portion of adult literacy.
- In the field of education and social advance, a secure passage toward integration will be opened only through the use of the native language.

Hence, the critical question to answer is how to enable the local people benefit from literature written in the English language. Time and again to communicate with concepts generated by any man on earth, speaking any of the languages of the world, the role that translation plays is quite obvious. Furthermore translation software would overcome the aforementioned problems and could be customized to provide advantages for the majority of the population.

1.3 Objectives of the study

1.3.1 General objectives

The general objective of this thesis is to customize a prototype transfer based translation system, which could translate documents from English into Amharic.

1.3.2 Specific objectives

In line with achieving the general objective stated above, the research would attempt to address the following specific objectives.

1. To study the basic grammatical and structural differences between Amharic and English.
2. To customize an English grammar, parser and tagger or chunker in this specific case.
3. To write a translation module, which would contain equivalence rules to map English tree structures to Amharic.
4. To build a user friendly interface
5. To design an electronic dictionary, to be used in the transfer stage.
6. To test the prototype to be customized.
7. To evaluate the MT output.

1.4 Method

Developing a translation system from English to Amharic requires one to investigate and identify the properties of both languages. For this purpose, a review of related literature was made in the area of English and Amharic sentences, phrases, and word classes. Literature in the area of machine translation, morphological analyzers, and electronic dictionary and lexicon automation was reviewed for this study. Discussion with linguists and experts in the area of Amharic language were made to better understand the phrase structure of the language was commenced frequently.

1.5 Application of Results and Beneficiaries

This prototype system aims to be helpful in many areas of Natural Language Processing (NLP) for the Amharic language. This study aims to have a direct or indirect impact on improving the ability of MT systems to perform real-life MT tasks. It will furthermore improve the quality and efficiency of English–Amharic translation to a significant extent.

Although, the English language is widely used in the educational system of the country, with the exception of educated people and foreigners, the majority of the population does not comprehend this language. There are also many journals, books, newspapers, etc. published abroad in foreign languages that might be very useful to the country. For example, scientific and technological information is mostly published in English. However, the language barrier makes it difficult to use such information by the mass.

Most of the available educational resources for high school and higher education students are written in the English language, and as it has been stated in the background section, this is posing an inconvenience to the Federal government's plan of commencing education in the local languages. Hence the major beneficiaries of this system would be students of different levels looking for resources compiled in English.

1.6 Testing techniques

The prototype to be customized will be tested for effectiveness by using selected English phrases and sentences. Based on the results, the system will be modified and further tests will be conducted time and again to see the effectiveness of the final output.

There are two ways of testing the customized system. The first one is by training it with a limited number of words to be included in the lexicon dictionary, and finally testing it against those data sets. The second method would be to allow the system take its input from users with the knowledge of the available words included in the dictionary and see-test it. The testing and evaluation activity will be explicitly discussed with selected examples in later chapters.

1.7 Limitations of the study

There were different constraints faced during the process of conducting this research. The first and major limitation was the very short time allotted for the research. As all researches this thesis required an intensive study on the basics of the problem. What distinguishes the current thesis is that it demanded a thorough study of Amharic and English grammars. And the succeeding task, which dealt with understanding new tools and programming language, took a considerable amount time. The time given to do these tasks was quite small.

The other obstacle was lack of relevant materials and financial aid that is required by a multidisciplinary research like this. The shortage of software, platform, and appropriate books and journals posed an insurmountable limitation in the course of the thesis. This constraint has been worsened by lack of financial back up to suppress the aforementioned obstacles.

1.8 Scope of the study

The scope is limited to customizing a prototype English-Amharic MT system since it is not possible to build a full-fledged system with the time allotted for the research. The scope of this study will be to translate from English into Amharic technological news published by a private magazine. The narrowed scope of this prototype will be to deal with IT news where the frequency and occurrence of technological English words can be guessed. There are numerous linguistically important reasons for narrowing the scope in this manner to be discussed in chapter three. The domain will

be strictly restricted to news articles dealing with technology, and building a prototype English-Amharic human aided machine translation system that performs the above activity will be scope of this research.

1.9 Organization of the thesis

The thesis is organized in eight chapters. The first chapter discusses the background information, statement of the problem, objective, scope, and significance and methodology of the study. Chapter two introduces the discipline of machine translation. It briefly defines the science, discuss the importance of it, overview the history and outline the different approaches employed since the beginning of the discipline to the current day. This chapter ends with a condensed discussion on the prospects of machine translation.

The third chapter is dedicated to discuss the underlying concepts behind the grammars of the source language (English) and the target language (Amharic). The relevant concepts to the domain of the current thesis are syntactic transfer, and this concept is covered in detail. Finally the differences between these languages, and the effect of these differences are outlined at the end of this chapter.

Chapter four describes the general overview of the translation system to be customized. It introduces each part of the system and gives a brief definition on each of them. The resources used in this thesis and the justification for the preference is explicitly mentioned in this section.

The analysis of the source text that encompasses morphological and lexical analysis that prepares the source language to be transferred to the destined target language is discussed in chapter five. The whole of chapter six is dedicated to the detailed transfer process of the English sentence into Amharic. The explicit discussion on lexical equivalence and transfer rules, along with the final Amharic substitution process is found in this chapter.

Taking sample sentences suitable for the prototype tests the overall performance of the system. Chapter seven contains all the testing mechanisms used, the sample sentences, test results and evaluation of the system. Finally the conclusion drawn from the current research and the recommendations that would make this system more powerful are briefly discussed in chapter eight.

CHAPTER TWO

MACHINE TRANSLATION: *State of the Art*

2.1 Introduction

Automating the translation process is what we call Machine Translation. Machine Translation (MT) may be succinctly defined as the "*mapping of one language into another by electronic means*" Hedden (2000). The same author further defined from a broader perspective that MT can be understood to include such computer applications as compilers and compression programs, etc., which convert a file in one computer language into a file in another computer language. In practical terms, such mapping entails the manipulation of a meaningful string of words of a given natural language, formulated in accordance with the grammar of that particular language.

The information revolution and technological innovations have driven the development of language industries and the expansion of multilingualism. The use of MT has experienced unprecedented growth with many diverse new techniques and demands. From a different horizon, MT is one application of natural language processing that analyzes a text in the Source Language (SL) and generates sentences in target language (TL).

2.2 Importance of MT

There will be enormous commercial, political and social benefits as millions of people are able to transact and communicate cross-language borders online with millions of other people in real-time without the need for a human translation intermediary. The Gartner group has a vision that, the potential for melting language barriers online is staggering and someday this "language conversion utility" must change the world of e-commerce in fundamental ways that will accelerate globalization. Globalization with a human face since humans will use machine translation as a way to broadcast their e-commerce, political, religious and humanitarian message to a global target audience that does not speak the language of the broadcaster.

MT has got social, political, commercial, scientific and intellectual or philosophical importance.

2.2.1 The socio-political importance

Many countries are well known by having communities with multiple languages. The social or political importance of machine translation, according to Arnold (1995), arises from the importance of translation of concepts from one language to another and to keep the social and political stability of the country. From another perspective, the other issue is the vanishing of unique culture associated to the language speakers and the language itself, and the way of thinking will matter to the society. Hence we can conclude that, translation is

the remedy to avoid these problems by facilitating ordinary human transaction and for gathering the information one needs to play a full part in society.

2.2.2 Commercial importance

The commercial importance of MT is a result of factors that are directly related to its social and political importance. To justify some of the commercial importance of MT the following can be considered as fact to the idea. One can select product without language constraint (for example, if we need material to specific topic in the language of Amharic and if the best material is available in English, we can take it and it can be translated to the language of our interest).

Secondly, it will simplify the expensive activity of manual translation, which has been briefly discussed in the first chapter. The process of producing draft translations, along with the often tedious business of looking up unknown words in dictionaries, and ensuring terminological consistency, will become automated, leaving human translators free to spend time on increasing clarity and improving style, and to translate more important and interesting documents.

2.2.3 Scientific importance

MT is an intellectual exercise application and testing ground for many ideas in Computer Science, Artificial Intelligence, and Linguistics, and some of the most important developments in these fields have begun in MT. He illustrates this using first widely available logic programming language called Prolog. Prolog

was formed as key part of the Japanese "Fifth Generation" program of research in the late 1980s, can be found in the 'Q-systems' language, originally developed for MT.

2.2.4 Philosophical importance

Philosophically, MT represents an attempt to automate an activity that can require the full range of human knowledge that is, for any piece of human knowledge, it is possible to think of a context where the knowledge is required. Scientists on machine translation agree that the extent to which one can automate translation is an indication of the extent to which one can automate 'thinking'.

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

2.3 Translation Approaches

There are many different methods for carrying out machine translation. There are three main systems are classified by their strategy for carrying out the translation; these are direct approach, the Interlingua approach and the transfer approach. Literature on the field further categorizes the learning based approach separately.

These approaches and the underlying techniques behind these approaches will be briefly discussed in this chapter.

2.3.1 The direct approach

Direct systems were the first generation of machine translation (MT) and they are usually built with one language pair (translation between two languages) in mind. Direct translation deals with taking a string of words from the source language, removing morphological inflections from the words and then looking up the lemmas in a bilingual dictionary between the source and the target language.

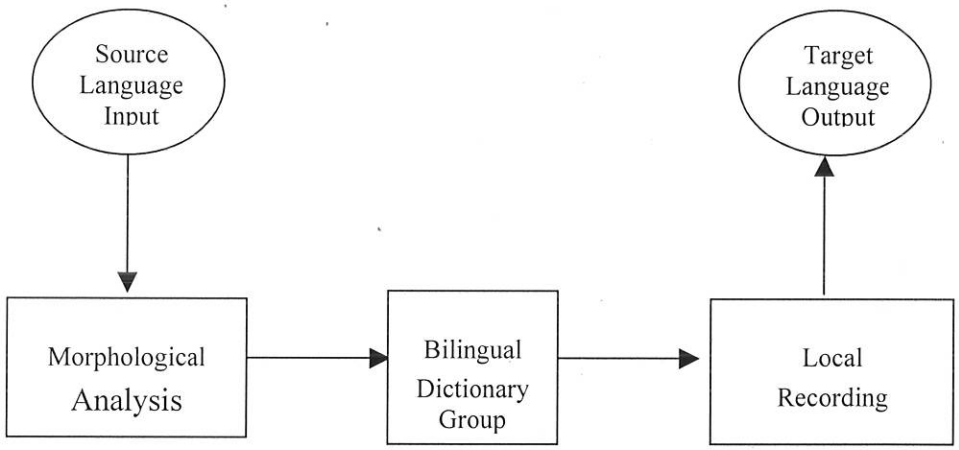


Figure 1: Direct Approach

This system has severe limitations. To start with, an ordinary word in a dictionary has more than one meaning. Meanings of most of the words can be understood from the context in which they appear. This system seems to have additional problem from linguistics point of view, it is highly dependent on the

order of words in different languages. Some of the languages have Subject-Verb-Object form while others have Subject-Object-Verb order. And still some others have Verb-Subject-Object form. Hence an attempt to translate directly in different languages with different word order will end up in severe confusion.

2.3.2 The interlingua approach

The Interlingua approach is the most attractive approach for a multilingual system. In this approach first a sentence in the source language is analyzed, then its semantic content, i.e., meaning, is extracted and represented by a language independent canonical form. Given this abstract representation, a natural language sentence can be generated using a generation module between the representation language and the target language. To include an additional language translator of this type, simply add an analysis module and a generation module for the new language to be represented.

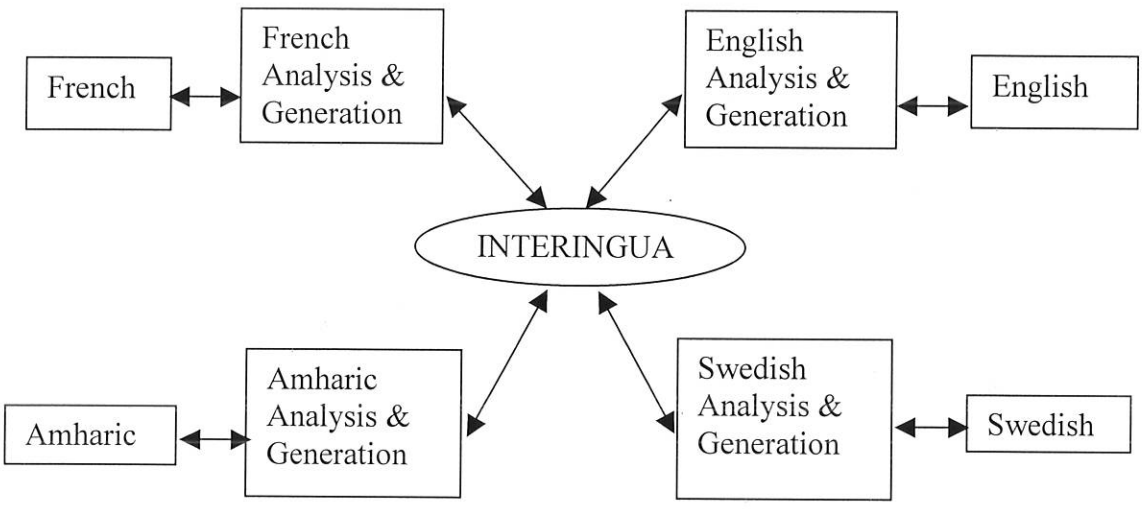


Figure 2: Interlingua Approach

An Interlingua system can translate between all pairs of languages, which were represented. For example, if the system has 6 modules, which are:

- Amharic-Analysis and Generation
- English-Analysis and Generation
- Swedish-Analysis and Generation

Thus this machine will be capable of translating in all of these directions; an Interlingua system can translate from English back to English. This 'back-translation' capability could in fact prove very valuable during system development in order to test analysis and generation modules.

Amharic → English

Amharic → Swedish

Amharic → Amharic

English → Amharic

English → Swedish

English → English

Swedish → Amharic

Swedish → English

Swedish → Swedish

Figure 3: Possible Translation Directions

Nevertheless this approach seems to pose the inefficiency to find a language independent representation, which retains the precise meaning of a sentence in a particular language, which can then be used to generate a sentence in a different language. Noone (2003) emphasizes that considerations, which must be dealt with, are the decision of which representational ontology to use and how to store language specific details in a general representation and a thorough understanding of the repositories.

2.3.3 Transfer approach

In this approach, the source text is analyzed to a certain extant depending on the language pair, the analyzed text can be transferred to a representation of the target language and target text can be generated from this transferred representation. This it the method adopted for implementation in the current thesis.

The transfer approach is used on the basis of the known structural differences between the source and target language. A transfer system is broken into three stages as shown above; Analysis, Transfer and Generation. In the analysis process, different tools are used to get the most possible meaning from the source. Morphological analysis, syntactic analysis and even some semantic analysis may be necessary. In this stage, the transfer method presupposes a parse tree of the input in the source language. Then this parse tree is mapped to a parse tree of the target language and the hand coded transfer rules are applied

to the analyzed text and it is transformed into some representation of the target text.

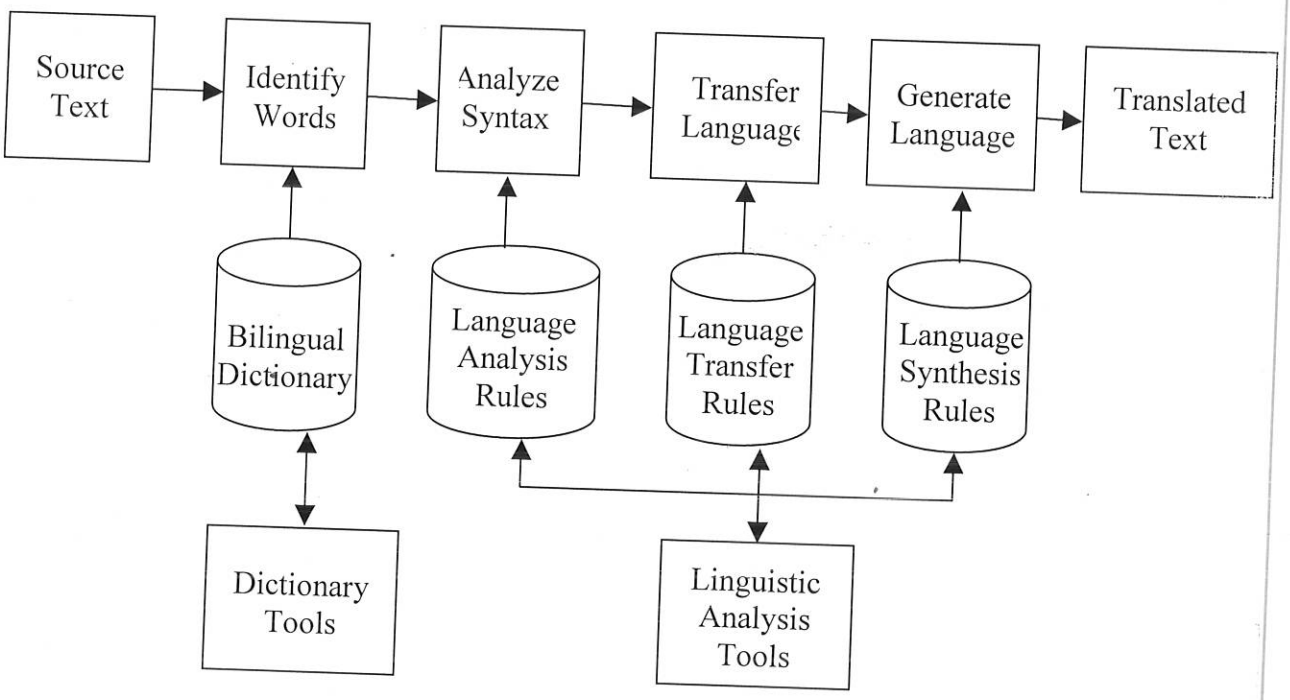


Figure 4: Transfer approach

All of this is done in a transfer module which maps semantically equivalent but syntactically different trees of the source language to the target language. After finding the parse tree of the target language it is put in some grammar module, which will take the tree as input and will output the corresponding natural language sentence.

2.3.4 Learning based translation

This system is based on analyzing similarities and differences between two translation example pairs. There is no linguistic analysis involved in the method and the system totally depends on string matching.

Statistical MT, Knowledge-based MT and Example based MT are also some of the current trends in the discipline. The idea behind Statistical MT approach is to let a computer learn automatically how to translate text from one language to another by examining large amounts of parallel bilingual text. This approach uses statistical data (e.g. which SL lexical unit is translated to which TL word(s) and how often this translation occurs) to perform translation. This statistical data is obtained from an analysis of a vast amount of bilingual texts. Different probabilities are extracted from the bilingual texts automatically by a computer, and these probabilities are vital to the translation process as they are the sole information for calculating how a SL sentence should be translated to the TL form.

In Knowledge based MT (KBMT) the assumption is that high quality translation requires in-depth understanding of the text. A domain model, which supports this in-depth understanding of the meaning and relationship of words in the text, is therefore used to aid the translation process. The motivation behind KBMT is that post-editing is time-consuming and expensive, thus it is worth putting more effort in designing an MT system which can produce high quality output

without human intervention. KBMT tends to be domain specific (especially a domain which is relatively less ambiguous, e.g. technical documents) because it is very complicated and difficult to represent a complete knowledge about the whole world. A specific case for KBMT is Example-based machine translation (EBMT). In this approach is often characterized by its use of a bilingual corpus. In linguistics, corpus (plural corpora) is a large and structured set of texts (now usually electronically stored and processed). A corpus may contain single texts in single language (monolingual corpus) or text data in multiple languages (multilingual corpus). Multilingual corpora that have been specially formatted for side-by-side comparisons are called aligned parallel corpora. This method employs a technology called translation memory, where as the user translates text, the translations are added to translation memory database (which was empty before the user translates the first document with the translation memory). When the same SL sentence occurs again during the translation, the previous translation stored in the database is automatically inserted into the TL document.

CHAPTER THREE

THE AMHARIC-ENGLISH DOMAIN

3.1 Introduction

The first section of this chapter is provided as a quick rundown of grammatical facts and morphological issues of the Amharic language. And major emphasis is given to the 'Verb' word class. This is because lexical transfer of the English verb is based on the morphological tree on its Amharic equivalent verb along with this a brief explanation on the 'noun' word class of Amharic and its morphological variants is presented. And following this a section is dedicated to show the verb and nominal agreement in the Amharic Language. And the second part deals with the differences between the Amharic and English languages, which is written to clarify the fundamental differences between Amharic and English.

3.2 Introduction to the Amharic word Classes

Tesfaye (2002) stated that there are two methods to identify lexical classes of a word. The first is based on the type of suffixes a particular word takes. For example in English nouns cannot take '-ed' as their suffix, but verbs do. The second method is based on the position of a word in a sentence. For example, in Amharic a verb cannot come at sentence initials and a noun cannot come at the end of sentences. Based on the above two criterion the Amharic words are categorized into the following five word classes.

A. **Noun** – comprises words like {-oc} as a plural marker. This group also includes words like /birhan/ 'light', and pronouns like /inne/ 'I', /ante/ 'you', and etc.

B. **Verb** – verbs are those words that come at sentence ends. They also take Clitics like the 2sg.m. {-h}, 1sg.f. {-hu}, and the like.

C. **Adjectives** – are words that come before nouns in sentences. Adjectives serve to modify nouns in sentences.

D. **Adverb** – words in this group are small in number and they do not attach any kind of prefixes and suffixes. The only words belonging to adverbs.

E. **Prepositions** – comprises words that are used to form adverbial phrases appearing before nouns they serve. Prepositions could not serve as base for the generation of other words nor do they conjugate for any kind of grammatical formation as for number, gender, etc. {lä}, {kä}, {bä}, {yä}, {slä} are some prepositions of the language.

For the convenience of this thesis, only the verb and Noun would be briefly discussed as follows related with morphology.

3.3 Issues in the morphology of Amharic

Morphology is defined as the study of morpheme or the meaningful word part of a word that contains no smaller meaningful parts of a language, and the way in which they are joined together to make words. For instance "gun" is one morpheme; "gun-s" contains 2 morphemes; "gun-fight-er" contains 3 morphemes. Each component has

also semantic or grammatical information to add to the overall meaning of the word, such as tense, number and word class.

3.3.1 Types of morphological processes

Linguists dealing with morphology agree that there are two broad (and partially overlapping) classes of ways to form words from morphemes: inflection and derivation. Inflection is the combination of a word stem with a grammatical morpheme, usually resulting in a word of the same class as the original stem, and usually filling some syntactic function like agreement. And derivation is the combination of a word stem with a grammatical morpheme, usually resulting in a word of a different class, often with a meaning hard to predict exactly.

3.3.2 Constituents of a morph

It has been mentioned that words are formed from a combination of morphemes. Word forms are produced from morphs in a number of ways. Trost (2000) presents affixation, reduplication, and compounding as the major ones. Since the testing words in the dictionary to be designed for this thesis are chosen to be intentionally simple (with less morphological complexity), it's the first method that the system deals with, hence will be briefly discussed.

Affixation attaches affixes to free morphs. Suffixes and prefixes are the two most common affixes found in many languages. But some languages have also circumfixes and infixes. A crucifix is the combination of a prefix and a suffix, which together express some feature.

3.4 The Amharic Verb

The Amharic verb has a root that consists in a number of "root letters", or "radicals" (most commonly three). To indicate person, tense, mood, etc. the forms of these radicals can change; prefixes and suffixes also can be attached; but the radicals themselves remain, and so identify the verb for us and as all Semitic languages Amharic exhibits root-pattern morphology. Accordingly identification of the most frequently occurring consonant and vowel patterns has been a logical starting point in most attempts to organize Amharic verbs into classes.

3.5 Agreement phenomena in Amharic

3.5.1 Verbal agreement

Amharic has a complex system of agreement in both verbals and nominals. It has distinct person, number and gender markings, with a possibility of fusion of the gender or number affix with the person marker in the perfective, and fission of the same affixes in the imperfective.

3.5.1.1 The gerund

There are two types of gerunds in Amharic, vis, verbal and normal. The system of the verbal gerund has the form of systems in the imperfective aspect, but unlike other such systems, which are prefixing, the verbal gerund is suffixing. The suffixes are inflection of person, number and gender.

3.5.1.2 Representation

In this section the formal representation of Amharic verbal agreement will be discussed. .

The surface ordering of functional (morphological) categories is reflective of the underlying configurations of constituents of clause structures, the representation of the Amharic clause should be along the line in (figure) next page.

Baye (1996) suggested that, a language like Amharic with an S-O-V word order is expected to show a verbal complex of the order V-AgrO-AgrS, reflecting a hierarchical relation in which AgrS dominates AgrO and AgrO dominates VP. What is shown in (Figure 3.1) next page is a clear violation to this principle. So the facts of Ethio-Semitic languages like Amharic seem to pose a problem for the two universal claims made on the hierarchical organization of constituents and on the linear ordering of affixes.

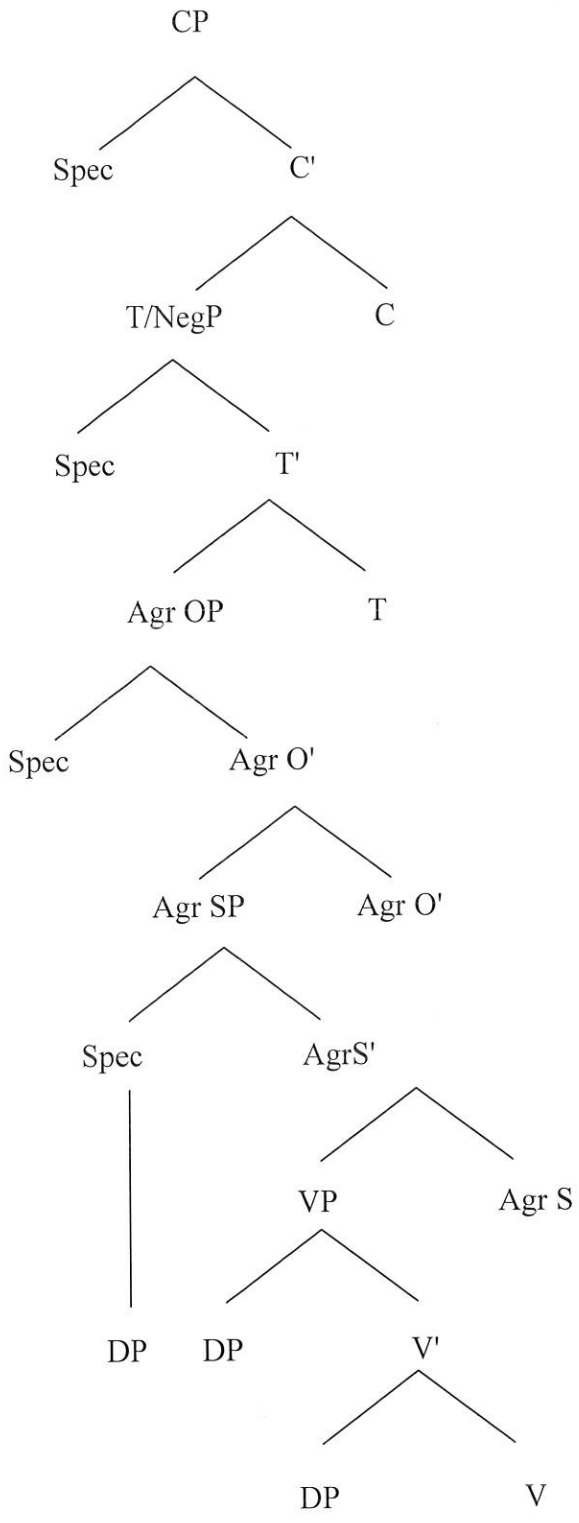


Figure 5: Representation of Amharic Clause

3.5.2 Nominal Agreement

Amharic nouns have different inflectional morphemes that create agreement of gender, number, definiteness and case.

3.5.2.1 Number

Number is a grammatical character that describes quantity. As many Semitic languages, Amharic nouns are divided into two: finite and infinite. Quantities that can be counted and represented by numbers are further divided into singular, where the element is only one or plural, where the elements are two or more.

Singular nouns do not have a special feature to show that they are singular. But, we can represent Amharic plural nouns in two ways: The first one is by adding the suffixes '***-oc***' and '***-woc***' on the main word, and the second one is by repeating the noun twice with a linking vowel '***-a***' in between. Let us see the two cases with examples:

a) By adding suffixes: ***-oc*** and ***–woc***,

Singular	**Plural**
ayat (a grandparent)	ayat**oc** (grandparents)
bet (a house)	bet**oc** (houses)
Temari (a student)	Temari**woc** (students)
Wenber (a chair)	Wenber**oc** (chairs)
Makina (a car)	Makina**woc** (cars)

Table 1: Regular Nouns

b) By repeating the word itself by a linking vowel '*a*'

Singular	Plural
Get (a jewelry)	get**a**get (jewelries)
Til (a worm)	til**a**til (worms)
Cherk (a garment)`	cherk**a**cherk (garments)
Biret (a metal)	biret**a**biret (metals)

Table 2: Irregular Plural Nouns

3.5.2.2 Gender

Gender in Amharic can be classified into lexical and grammatical. The grammatical gender has some inflection on the root word to show either 'masculine' or 'feminine'. In Amharic the 'masculine' nouns are usually the main words and when the suffix '***-itu***' or when the prefix '***yichi-***' is added on it, the word will show a feminine noun.

Example for this is:

Masculine	Feminine
beg (sheep)	beg**itu** (female sheep)
demet (cat)	demet**itu** (female cat)
zaf (tree)	zaf**itu** (female tree)
makina (car)	makina**itu** (female car)

Table 3: Gender implying Suffixes

Pronouns also show gender difference like this:

Masculine	Feminine
ante (you)	anchi (you for a girl)
issu (he)	Issua (she)

Table 4: Pronouns Showing Gender

There are words in Amharic, which imply gender lexically, without a prefix or suffix.

Masculine	Feminine
abat (father)	enat (mother)
agot (uncle)	akist (aunt)
wondim (brother)	ehit (sister)
bere (ox)	lam (Cow)

Table 5: Lexical Gender Implication

3.5.2.3 Definiteness

The grammatical role of nouns in a sentence is what we consider as definiteness. Their definiteness might be either as an object, or showing possession. And it also shows gender along with word class in Amharic. Along with the possessive feature, it also shows number.

Noun	Morphology		Possessive pronoun
	Masculine	Feminine	
Bet (a house)	bet**u** (the house)	bet –**wa** **-itu** **-itwa** (the house)	bet**e** (my house) bet**u** (his house) bet**wa** (her house) bet**achn** (our house) bet**achew** (their house)

Table 6: Definiteness in a Noun (gender)

And for the number case:

Noun	Singular	Plural
Bet	bete (my house) betu (his house) betwa (her house)	betoche (my houses) betochu (his houses) betochua (her houses) betochchen (our houses) betochachew (their houses)

Table 7: Definiteness in a Noun (number)

3.5.2.4 Case

The last feature of Amharic nouns is that many of the nouns do not show specific gender and number as a root word directly. But, when suffixes like '*-u-n*', '*-u-wa-n*', '*-itwa-n*', '*-itu-n*' are added we could clearly see the difference in meaning.

Noun	Accusative (DO)		Dative (IDO)	
	Masculine	Feminine	Masculine	Feminine
Bet (a house)	Bet-**u-n** (the house)	Bet-**u-wan** (the house)	Le-bet-**u** (for the house)	Le-bet-**wa** (for/to the house)
Lij (a child)	Lij-**u-n** (the child)	Lij-**i-twan** (the child) Lij-**i-tu-n** (the child)	Le-lij-**u** (for the child)	Le-lij-**itwa** or Le-lij-**itu** (for the child)

Table 8: Case Specification in Amharic

3.6 Basic Differences between Amharic and English

A quick comparison between Amharic and English shows that in the former the abstract form is much greater, in morphological and semantic forms. While in English most of the abstract nouns are derived either from adjectives, for example: justice, wisdom, bravery, truth, etc., or from nouns, but only such nouns which mark human beings or abstract notions, for example: manhood, heroism, friendship, etc. But otherwise no nouns serve as the base of abstract nouns, except if they first became adjectives. In Amharic on the other hand, the derivational possibilities are much more diversified and, consequently, the semantic field covered by this form is much larger.

As it has been tried to discuss briefly in the previous sections, the most typical Semitic characteristic of Amharic is its root-and-pattern morphology. As in all Semitic languages, verbal roots consisting of several consonants (most commonly, three) are the bearers of essential semantic values. Amharic also has prepositions, which may assign case and o-role to the right. On the other hand, there is also a set of postpositions (which are typical of Cushitic languages) that are often used in combination with a preposition but sometimes occur without one. Since the verb is situated on the right of both subject and object, Amharic appears to qualify as a "head-final" language, assigning, both case and o-role to the left (indirectly as far as the subject is concerned). '

One other important characteristic of Amharic is that it is "pro-drop", or 'null-subject", language. That is, it is not necessary, in Amharic that the subject position be

filled with a lexical NP. For instance, a speaker might express the essential content of the following:

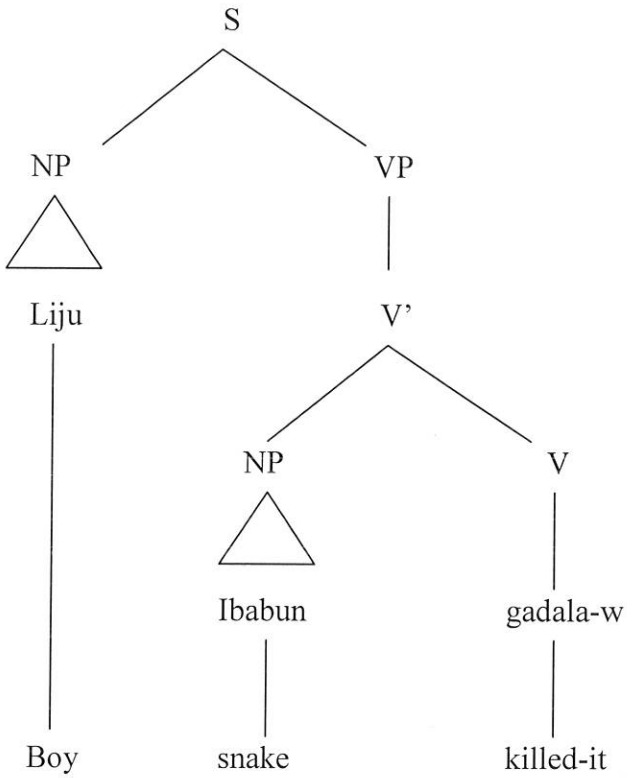


Figure 6: The Undetailed Structure of a Simple Declarative Sentence in Amharic

By saying the sentence without the lexical subject liju 'the boy' as follows:

ibabun gadala-w
snake killed-it
(def:acc) (3.s.m.)
'he killed the snake'

Taking morphological information in the previous section, one can see that the person suffixes in Amharic are 'genitive pronominals' referring to a possessor NP. The number marker is '***-oc***', 'plural' and the gender marker is '***-it***', 'feminine'. The definiteness marker '***-u***', is phonetically identical with the third person genitive

possessive suffix '***-u***'. In fact, one can say that these two are also similar semantically in the sense that they specify a noun in terms of its possessive or deictic reference. Hence, a noun like 'lig-u' is ambiguous between the definite readings 'the child' or the possessive readings, 'his child'. However, syntactically the two homophonous affixes differ since the definite, and not the possessive, '***-u***'/ occur with a modifier when there is one. The possessive, '***-u***' is always found attached to the head noun. The following are examples showing this difference (Baye, 1996).

(a) lig-u
Child-def
(b) tilli'k-u lig(-u)*
big-def child
(c) tilli'k-u lig-u
big-def child-3smGen
'The his big child' / 'the big child of his'

When plural reference is made, the number marker comes following the stem and preceding the genitive suffix. And like in definiteness, the adjectival modifier agrees with the noun head in number or gender by showing the plural '***-oc***' or the feminine marker '***-it***' in the manner shown in (a) and (b) below:

(a) habtam-oc-u lig-oc-e
rich-pl-def child-pl-1sGen
'The rich children of me'
(b) habtam-it-u lig-e
rich-fem-def child- 1sGen

Finally, abstract nouns in Amharic can be directly derived from almost all parts of speech, with the exception of verbs, prepositions and conjunctions. Now, all the parts of speech which can be used as bases of the abstract nouns, namely adjectives, participles, nouns, pronouns, adverbs and numbers, have this in common that they can

serve, contrary to verbs, prepositions or conjunctions, as part of the predicate in copula sentences. Consequently, the abstract nouns can be considered as the transformation of a copula sentence. As O.Jespersen (1924) wrote about the abstract noun, in his philosophy of grammar, "the idea of 'being' is smuggled into the word". Amharic makes a full use of the capacity of a copula sentence to be transformed into a noun.

3.7 Effect of the language difference on the translation system

Lexical and structural ambiguities arise in the course of translating from a positional language like English to a language with very rich morphology like Amharic. As it would be discussed in the testing and evaluation section of the thesis, these ambiguities would make the translation system less efficient.

The current translation system works on a transfer approach, which first parses the source text, then transforms the source language syntax tree into the target language and finally use the target language syntax tree to generate a translated sentence. Lack of wide coverage computational grammar for Amharic language is one obstacle faced during this thesis. Besides, the performance of parser-based translation system is limited by the performance of the current syntactic parsing systems. Murthy (1999) stated that even the best available parsers are not good enough, and the same problem posed inefficiency on this system. Finally, the non-availability of suitable substitutable equivalent Amharic words for the technological English words is one of the major drawbacks of the Amharic language.

CHAPTER FOUR

DESIGN OF THE SYSTEM

4.1 Introduction

In this chapter, the methods and resources of this system that are used in the process of translating from English to Amharic will be discussed. Following that the fundamental modules and operational flow of the system will be described.

As it has been mentioned in chapter one, for this research majority of designing and structuring the translation process and resources has been adopted from the Shakti Machine Translation (MT) system, which translates from English to Hindi. Shakti uses statistical as well as rule-based approaches for processing language. It uses constituent structure at the chunk level and dependency relations at the sentence level of analysis. It has specialized components for word sense disambiguation, parsing, Preposition attachment, phrasal verb identification, transfer grammar, sentence and word generation, and many others.

The architecture of Shakti is highly modular. The complex problem of MT has been broken into smaller sub-problems. Every sub-problem is a task, which is handled by an independent module. The modules are put together using a common extensive representation using trees and feature structures, called the Shakti Standard Format (SSF).

4.2 System Outline

This prototype is unidirectional, i.e. it can translate from English to Amharic. The system can be used through an interface. *As* can be seen in Fig 5.1, the system will take the path shown, by taking the input English sentence and finally producing an Amharic output. During the first half of this chapter, the translation process is outlined.

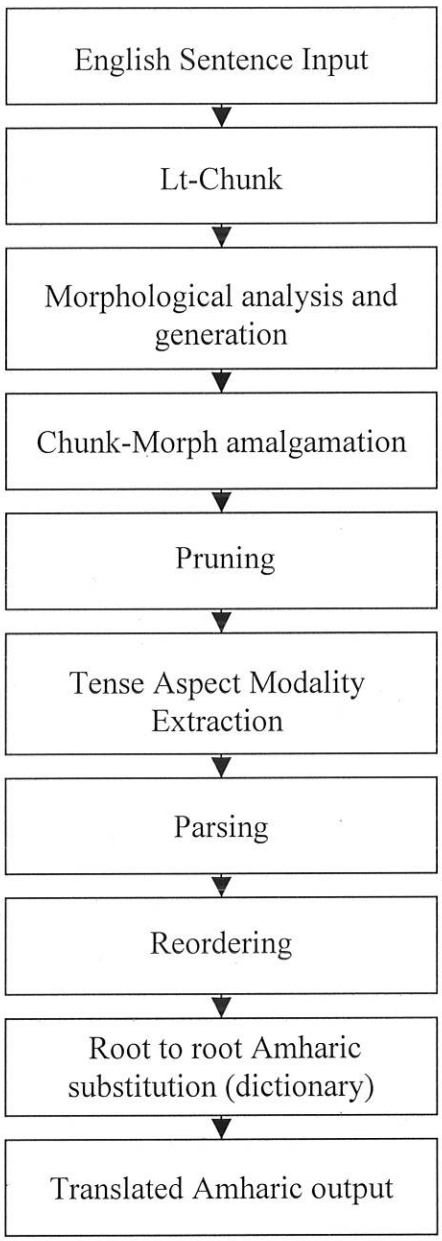


Figure 7: Outline of the Translation System

4.3 Translation from English to Amharic

The translation process from English to Amharic will be outlined here. In the diagram 5.1 each stage is given a name. The translation process is done in a modular manner and each module in the system is responsible to achieve a certain specific task clearly put in the coming sections. The aforementioned modules perform the remaining task of translating. Brief functional description of each module is dealt with in the following sections.

4.3.1 The Input File

Initially, the user provides English input text. There is an input file designed for this purpose. This file holds the input sentence temporarily, and transfers to the succeeding modules that perform the translating task.

4.3.2 LT-CHUNK

This specific LT_CHUNK software is used for the disambiguation of words part-of-speech and the identification of simple noun and verb groups in English texts. It consists of :

ltchunk - text chunking program;

lttok - English tokenizer build on Finite State Automata

ltpos - probabilistic part-of-speech tagger based on Hidden Markov Models plugged with Maximum Entropy probability estimators;

sgtransduce - a grammar interpreter which is supplemented with grammars to recognize verb and noun groups; and other supporting utilities;

4.3.3 Morphological analysis and generation

These are two morphological processes performed by the system. In both cases, morphological analyzer and morphological generator built in the toolkit accesses the lexical database containing the repository of available stems and flexes. The relevant morphosyntactic features (verb, lemma, person, number, etc.) of the chunked sentence will automatically calculated by this module of the system.

4.3.4 Amalgamation of Morphological Information with Chunker

In this module, the lt-chunk and the morphological analyzer's output are being combined together. First this module reads a line from chunker output and stores the morphological information in an array for each lexical item of that sentence. Afterwards the words and all its probable tags are stored in an array. And this morphological information obtained at this stage is added to the initially chunked output from the chunker.

4.3.5 Pruning

The main objective of this module is to prune the output given be the preceding module of the system. When the morphological information is merged with the Chunker output, the system displays all the possible occurrences of the lexical items in the sentence. It is impossible to translate unambiguously when there are two or more options of morph result. Therefore the system should identify the relevant form that can be used under the given context, and pruning at this stage

means selecting only the correct option and leaving all the other alternatives as irrelevant.

The temporarily pruned morphological output of the system does not take the contextual and inherent features described into consideration. The system uses morphological category information with the chunk group information to prune. Hence this module finalizes the pruning process by integrating the outputs of the categorical information and the temporarily pruned morphological entry. There are cases where the morphological output might be two, under such condition this module of the system takes the first one and gives an appropriate output.

4.3.6 Extracting Tense Aspect Modality (TAM)

By the terms Tense, Aspect and Modality (TAM) reference is made to intrinsically interwoven semantic domains that mainly appear across languages as formal grammatical categories, but are also expressed at the lexical, syntactic and even pragmatic level.

he three primary grammatical categories for which a verb may be inflected are Tense, Aspect, and Mood. These three relate respectively to location in time, temporal constituency, and modality in much the same way as grammatical Gender relates to sex and grammatical Number relates to enumeration. This module extracts the TAM of the pruned verb groups that might be ambiguous.

4.3.7 Parsing

Parsing is conventionally defined as a combination of recognizing an input string and assigning some structure to it. Syntactic parsing is the task of recognizing a sentence and assigning a syntactic structure to it. Basically the main duty of this module of the system is to take care of the syntactic structure of the English input sentence.

Parsing the English input is the process of discovering words in a sentence that are related in structure and in meaning. In this module the parser tries to determine functionality of words and their roles in sentence structure according to the specified grammar.

4.3.8 Reordering

After the Parsing is done, the next task to be performed by the system would be reordering the SVO order of the English sentence by the appropriate SOV order of Amharic so that the substitution of Amharic words would be facilitated. The system takes the parsed English output and reorders it according to the conventional Amharic word order.

4.3.9 Root-to-root Amharic substitution

This is the module that performs the lexical transfer of the English words by the Amharic words. By lexical transfer it means the selection of the word, or words, in the Amharic language that best render(s) the concept represented by the English language word or expression. It is at this step that the designed

dictionary is referred by the system to replace the parsed and reordered English output by its equivalent Amharic root word.

4.3.10 Translated Amharic output

The final output of the system after the necessary root-to-root substitution and the appropriate morphological synthesis is finished the system gives out the processed Amharic sentence output. This output is temporarily stored into the file, which is integrated with the previous module.

4.4 Resources used

1. LT-POS (Tagger)
2. LT-CHUNK (Chunker)
3. PERL (programming language employed)
4. QT-Designer (programming language used to design the user interface)

As it has been mentioned earlier developing this translation prototype made use of customizing functional units of the Shakti translation tool-kit. The reason for choosing the Shakti tool-kit has been mentioned at the beginning of chapter four. The components used and their operational description is discussed as follows:

4.4.1 Part of Speech tagging (POS-tagging)

The task of POS-tagging is to assign part of speech tags to words reflecting their syntactic category. But often, words can belong to different syntactic categories in different contexts. For instance, the string "***books***" can have two readings: in

the sentence *he books tickets* the word "***books***" is a third person singular verb, but in the sentence *he reads books* it is a plural noun. A POS-tagger should segment a word, determine its possible readings, and assign the right reading given the context.

There are two major paradigms for organizing the lexicon: the word-list lexicon where each word is declaratively stored together with its morpho-syntactic features and the morphological lexicon where the base forms of words (stems) are provided with the rules for the formation of their inflectional and derivational variants. Regardless of how the lexicon is implemented, the morphological classifier retrieves for a word its possible parts of speech and other related morpho-syntactic features such as number, case, gender, etc. In the POS-tagging framework, each combination of morpho-syntactic features is unambiguously mapped into a POS-tag. For example, a singular noun can be tagged as NN and a plural noun as NNS, the base form of a verb can be tagged as VB and its past form as VBD. The examples are tags from a set known as the "Penn Treebank tag-set". In this tag-set more than one tag can be assigned (as in the case of "***books***" above") the classifier returns a set of possible POS-tags, also called a POS-class. However, no lexicon contains all possible words. When the classifier comes across a work that is not in the lexicon, a POS-guesser tries to guess the POS-class for the unknown word.

4.4.1.1 LT POS

A group of researchers at the University of Edinburgh (2000), discussed that LT POS incorporates a part of speech guesser, which employs a number of different guessing strategies. LT POS achieves 88-92% accuracy on unknown words. In the output LT POS will then give information saying the word was found in the lexicon, or a suffix-guessing strategy was used, or all guessing failed and a default strategy was used, etc.

When tagging SGML marked-up text, LT POS will read in the document's Document Type Definition (DTD), or any other DTD provided. It is possible to ask LT POS to output the tagged text as SGML. Depending on the details in the document's DTD, LT POS will then produce something like the following:

```
<SENTENCE>
<W TAG="PPS">He</W><W TAG="VBZ">books</W><W
TAG="NNS">tickets</W>
</SENTENCE>
```

LT POS incorporates a tokeniser that will determine sentence and word boundaries. But it is possible to switch off the internal tokeniser and to use LT POS with a different tokeniser.

4.4.2 LT CHUNK

4.4.2.1 Features and limitations

LT CHUNK is a syntactic chunker or partial parser. LT_CHUNK is software for the disambiguation of words part-of-speech and the identification of simple noun and verb groups in English texts. It uses the part-of-speech information provided by the tagger described earlier and employs mildly context-sensitive grammars to detect boundaries of syntactic groups. The chunker leaves all previously added information in the text and creates a structural element that includes the words of the chunk. According to Mikheev (2000), currently it is capable of recognizing boundaries of simple noun and verb groups. Noun groups do not include prepositional or clausal post-modifiers:

[The most important man] of [our little group].

[The red book] which I bought yesterday.

The chunker does not break two conjunctive noun groups if the second one doesn't have a clear start (such as a determiner):

As [*a* tokenizer LT POS] *uses* ....

Some errors occur because of an ambiguity in the text. Consider:

This sentence <has capitalized>[words] .

The chunker gets it wrong, by not recognizing that "capitalized words" belong together as a single chunk. But note that its analysis is correct for

This program <has capitalized>[words].

Although the chunker tries to correct some obvious mis-taggings of the tagger it usually produces a wrong grouping if a severe mis-tagging occur.

4.4.3 Why use PERL?

PERL, which is an acronym for **P**ractical **E**xtraction and **R**eport **L**anguage, is chosen for this system mainly because it has very distinguishing and strong features that enable it to manipulate strings. And it highly convenient for many translation procedures employed due to it's built in functions for string analysis and synthesis, and I did not see any reason to change the programming language that the Shakti system used.

For this special purpose of translation PERL happen to be convenient because it fully parses and pre-"compiles" any given script before execution. The PERL script first takes an inventory of the kinds of information present in the semantic representation, and generates a formulaic phrase for each one. And effectively eliminates the potential for runtime SYNTAX errors.

Besides, PERL is good for prototyping in that it handles many data handling details such as regular expressions (pattern matching), string and list processing, data formatting, file input and output and database access.

Here are some of the distinguishing features that convinced me to use PERL:

- It is used to automate repetitive and/or complex tasks, and it is favorable for report generation.
- PERL is portable, in part, due to its widespread availability and due to its inherent cross-platform compatibility.
- PERL has sufficient number of built in features that you don't often have to call non-portable external executables.
- PERL has numerous built-in optimization features which makes it run faster than other scripting languages
- PERL represents the things to be done in fewer lines of code as compared with other programming languages and besides takes better advantage of the underlying Operating system.

4.5 Sample Translation

Following is a short example that briefly shows what exactly happens at each step of the prototype. This is planned to give a brief introduction of the underlying operational details given in chapters five and six.

------- **Input Sentence** ----------

he is reading the book

------- **Lt-Chunk output**---------

<S>[[he_PRP]] ((is_VBZ reading_VBG)) [[the_DT book_NN]]</S>

------- **The Morphological analysis and generation** ---------

SENTENCE
He //he,P,m,s,3|
is //be,v,m,s,3|
reading //reading,n,m,s,3,0,,|//read,v,m,s,3|
the //the,det,m,s,3|
book //book,n,m,s,3|//book,v,m,s,3|

End SENTENCE

----- Amalgamation of the generated morphological information with chunker output----

```
SENTENCE
1   ((          NP
1.1 he          PRP     //he,P,m,s,3|
    ))
2   ((          VG
2.1 is          VBZ     //be,v,m,s,3,0|
2.2 reading     VBG     ////reading,n,m,s,3,0,,|read,v,m,s,3|
    ))
3   ((          NP
3.1 the         DT      //the,det,m,s,3|
3.2 book        NN      //book,n,m,s,3,0,,|//book,v,m,s,3|
    ))

End SENTENCE
```

---------- pruning ----------

```
SENTENCE
1   ((          NP
1.1 he          PRP     //he,P,m,s,3,0,,|
    ))
2   ((          VG
2.1 is          VBZ     //be,v,m,s,3,0,,|tense='PRES'|
2.2 reading     VBG     //read,v,m,s,3,0,,|aspect='PROG'|
    ))
3   ((          NP
3.1 the         DT      //the,det,m,s,3,0,,|
3.2 book        NN      //book,n,m,s,3,0,,|

End SENTENCE
```

---------After extracting TAM ----------

```
SENTENCE
1     ((        NP
1.1   he        PRP     //he,P,m,s,3|
      ))
2     ((        VGADV
2.1   ((        VG      //read,v,m,s,3,is_ing|aspect='PROG'
2.1.1 is        VBZ     //be,v,m,s,3|tense='PRES'|
2.1.2 reading   VBG     //read,v,m,s,3,0,,|aspect='PROG'|
      ))
      ))
3     ((        NP
3.1   the       DT      //the,det,m,s,3|
3.2   book      NN      //book,n,m,s,3|
      ))

End SENTENCE
```

---------- **Parsed Outputs**---------

```
Parse Number 1 =
SENTENCE(1,3) -> SDECL
    SDECL(1,3) -> RBSTAR NP VGADV NPSTAR PPSTAR NPSTAR PPSTAR
RBSTAR
        RBSTAR(1,-1) -> (null)
        NP(1,1) -> he
        VGADV(2,2) -> reading
        NPSTAR(3,3) -> NPSTAR NP
            NPSTAR(3,-1) -> (null)
            NP(3,3) -> the book
        PPSTAR(4,-1) -> (null)
        NPSTAR(4,-1) -> (null)
        PPSTAR(4,-1) -> (null)
        RBSTAR(4,-1) -> (null)

Parse Number 2 =
SENTENCE(1,3) -> SDECL
    SDECL(1,3) -> RBSTAR NP VGADV NPSTAR PPSTAR NPSTAR PPSTAR
RBSTAR
        RBSTAR(1,-1) -> (null)
        NP(1,1) -> he
        VGADV(2,2) -> reading
        NPSTAR(3,-1) -> (null)
        PPSTAR(3,-1) -> (null)
        NPSTAR(3,3) -> NPSTAR NP
            NPSTAR(3,-1) -> (null)
            NP(3,3) -> the book
        PPSTAR(4,-1) -> (null)
        RBSTAR(4,-1) -> (null)
Parse Number 3 =
SENTENCE(1,3) -> NP VGADV PHR
    NP(1,1) -> he
    VGADV(2,2) -> reading
    PHR(3,3) -> NP
        NP(3,3) -> the book


Parse Number 4 =
SENTENCE(1,3) -> NP VGADV NP
    NP(1,1) -> he
    VGADV(2,2) -> reading
    NP(3,3) -> the book
```

---------- **After reordering** ----------

```
1     ((          NP      //he,P,m,s,3|role='subj:2'|
1.1   he          PRP     //he,P,m,s,3|
      ))
2     ((          NP      //book,n,m,s,3|role='obj:2'|
2.1   the         DT      //the,det,m,s,3|
2.2   book        NN      //book,n,m,s,3|
      ))
3     ((          VGADV   //read,v,m,s,3,0,,is_ing|aspect='PROG'| name=2|
3.1   ((          VG      //read,v,m,s,3,0,,is_ing|aspect='PROG'|
3.1.1 reading     VBG     //read,v,m,s,3,0,,|aspect='PROG'|
      ))
      ))
```

----------**After root-to-root Amharic substitution**---------

```
THE AMHARIC ::
SENTENCE
1     ((          NP      //issu,P,m,s,3,0,,|role='subj:2'|
1.1   issu        PRP     //issu,P,m,s,3,0,,|
      ))
2     ((          NP      //matsihaf,n,m,s,3,0,,|role='obj:2'|
2.1   the         DT      //the,det,m,s,3,0,,|
2.2   matsihaf    NN      //matsihaf,n,m,s,3,0,,|
      ))
3     ((          VGADV   //manbaab,v,m,s,3,0,,is_ing|aspect='PROG'| name=2|
3.1   ((          VG      //manbaab,v,m,s,3,0,,is_ing|aspect='PROG'|
3.1.1 eyannebaabaa VBG    //manbaab,v,m,s,3,0,,|aspect='PROG'|
      ))
      ))
```

----------**Translated Amharic output** ----------

Issu matsihaf eyannebaabaa naw.

CHAPTER FIVE

THE ANALYSIS STAGE

5.1 Introduction

In this chapter the techniques used to analyze the input sentences will be described. And please note that in all the analytical steps to be discussed an input English sentence "*He is reading the book*" is taken to demonstrate the outputs of each of the modules of the system.

5.2 Chunker output

As it has been discussed in the features and limitations of the LT CHUNK is a syntactic Chunker or partial parser. It creates a structural element that includes the words of the chunk and is capable of recognizing boundaries of simple noun and verb groups. Noun groups do not include prepositional or clausal post-modifiers. This module handles the input toy sentence as follows:

<S>[[he_PRP]] ((is_VBZ reading_VBG)) [[the_DT book_NN]]</S>

5.3 The Morphological analysis and generation

Morphological analysis and morphological generation are two tasks shared by many applications in the field of NLP. Closely connected to these tasks, one of the main issues at hand when designing such applications is how to organize and store in the lexicon the morphological information needed to analyze and generate words.

Machine Translation is typically one of the applications that need to handle analysis and generation processes of the source language at the same time. Linguistic databases for MT systems need to be designed so that the knowledge they store is as much process independent as possible. Thus designing declarative lexicon databases in MT applications is a mandatory.

The morphological analyzer and morphological generator accesses the lexical database containing the repository of available stems and flexes. Both stems and flexes have certain lexical information [Linfo] associated. This lexical information has to meet at least two conditions:

1. The relevant morphosyntactic features (verb lemma, person, tense, mood, number, etc.) are either directly conveyed by [Linfo] or can be automatically calculated from it by some computational machinery.

2. The [Linfo] stored for each stem and flex must also be enough to ensure that only legal stem+flex sequences are retrieved from the lexical database. We will assume that both grammatical information and lexical information are represented and stored in the form of feature-value pairs.

Therefore the generated output of the module after morphological analysis describes that the morpheme '*he*', is a pronoun, the gender is male; the number is third person singular. '*Is*' is analyzed shares the above features with the pronoun and it are considered as a verb to be. The analyzer generates two possible options for the word

'*reading*' and '*book*', in the first instance it can be considered as a verb and in the later as a noun. Choosing the appropriate form from the two options is handled in the later stages. The word '*the*' is analyzed as a determiner. The output of this module after analyzing the morphology states the morphological information for each lexical item of the input sentence looks like:

```
He      //he,P,m,s,3|
is      //be,v,m,s,3|
reading //reading,n,m,s,3|//read,v,m,s,3|
the     //the,det,m,s,3|
book    //book,n,m,s,3|//book,v,m,s,3|
```

5.4 Merging the Morphological Information with Chunker output

The morphological component in this Natural Language Tool employed provides a formalism for writing a 'grammar' of words which comprises a base lexicon for words and affixes, rules for combining them, as well as spelling rules to account for spelling changes.

In this module, the lt-chunker output and the morph analyzer output is being used. First this module reads a line from chunker output and stores the morphological information in an array for each lexical item of that sentence. Afterwards the words and most probable tag are stored in an array. And this morphological information obtained at this stage is added to the initially chunked output from the chunker. And the output of this stage for our specific example would look like:

```
1     ((        NP
1.1   he        PRP     //he,P,m,s,3|
      ))
2     ((        VG
2.1   is        VBZ     //be,v,m,s,3|
2.2   reading   VBG     //reading,n,m,s,3|read,v,m,s,3|
      ))
3     ((        NP
3.1   the       DT      //the,det,m,s,3|
3.2   book      NN      //book,n,m,s,3|//book,v,m,s,3|
      ))
```

In the previous section it was shown that for the word *reading* there are two alternative morphological information in the dictionary, i.e. the first one considers it as a noun while there is an equally likely probability of the word to act as a verb in the sentence. The Chunker information states that the word is found in the verb group and when this is merged with the analysis output. Likewise the system also presents two probabilities for the word '*book*', here again *book* can act as a verb in one circumstance and as a verb in another, but when this module informs that there is a determiner '*the*' preceding it. Hence when the morphological information is added to the chunker output, it gives us all the possible occurrences of the lexical items.

5.5 Pruning

Linguistic-based pruning, which is applied on the extracted morphological analysis translation alternatives in order to filter and detect terms and their translation TAM to the Source language that are morphologically close enough. Morphological knowledge such as Part-of-Speech (POS), context of terms extracted from thesauri, conflict analysis of the chunker and the morphological analyzer could be valuable to filter and prune the test morphological result. POS tags are assigned to each source term via morphological analysis.

As well, a Source language morphological analysis will assign POS tags to the translation candidates. For English-Amharic pair of languages, English nouns and verbs are compared to Amharic nouns and verbs, respectively. English adverbs and adjectives are compared to Amharic adverbs and adjectives, because of the close relationship between adverbs and adjectives in English.

The main objective of this module is to prune the output given by the preceding step in the system. As it has been shown, when the morphological information is added to the chunker, the system displays all the possible occurrences of the lexical items in the sentence. It is impossible to translate unambiguously when there are two or more options of translating a given word. Therefore the system should identify the relevant form that can be used under the given context, and pruning at this stage means selecting only the correct option and leaving all the other alternatives as irrelevant.

The temporarily pruned morphological output of the system does not take the contextual and inherent features described into consideration. And taking the aforementioned issues into consideration the pruned output of the toy sentence would be:

```
1     ((        NP
1.1   he        PRP     //he,P,m,s,3|
      ))
2     ((        VG
2.1   is        VBZ     //be,v,m,s,3|
2.2   reading   VBG     //read,v,m,s,3|
      ))
3     ((        NP
3.1   the       DT      //the,det,m,s,3|
3.2   book      NN      //book,n,m,s,3|
      ))
```

5.6 Extracting TAM

When using the terms Tense, Aspect and Modality (TAM) reference is made to inherently related semantic domains that mainly appear across languages as formal grammatical categories, but are also expressed at the lexical, syntactic and even pragmatic level. Before stating the process of identifying the TAM of the pruned categorical sentence, let's try to see what TAM basically is:

The three primary grammatical categories for which a verb may be inflected are Tense, Aspect, and Mood. These three relate respectively to location in time, temporal constituency, and modality in much the same way as grammatical Gender relates to sex, and grammatical Number relates to enumeration. However, the relationship between grammatical form and its referent is not always exact.

If the object of the proposed investigation is to determine the primary referent of the grammatical form (inflection) (i.e., whether the verb is inflected for Tense, Aspect, or Mood), we need to determine whether the form primarily indicates (1) location in time, (2) temporal constituency, or (3) modality.

To summarize some grammatical pairings,

Grammatical category	Referent
Gender	Sex
Tense	Location in time
Voice	Agency
Mood	Modality
Aspect	Temporal constituency
Number	Enumeration

Table 9: Summary of Some Grammatical Pairings

5.6.1 Tense

Tense, according to Comrie (1994), is "*grammaticalized expression of location in time*". This location in time is the relation of Event time to Speech time. In the case of written texts, if the Event occurred before the author wrote, it is Past; if the Event was ongoing at the time of writing, it is Present; if the Event had not yet happened at time of writing, but was expected to happen, it is Future.

5.6.2 Aspect

Like Tense, Aspect is concerned with time, but where Tense is concerned with relating the time of the situation to another time-point, Aspect is concerned with the temporal constituency of the situation.

5.6.3 Mood

At any point in time, it is not known what will happen next; there are multiple alternative futures. For every undetermined possibility, there is a hypothetical world in which that possibility is true. Only one of these possible worlds turns out to be the "real" world. Even for events that happen in the past, at the time of the event, it is not known which of the possible worlds will be the actual world.

Hence this module in the system has a code that takes the aforementioned conditions into consideration and enables it to clearly identify the TAM of the input sentence. The verb group that this module deals with are VBP, which designates a verb used for Past tense. VBG, which designates a verb, used for Present tense, and finally VBZ, which designates that the verb used, is an

auxiliary verb. And based in the tense results the aspect is taken as progressive.

Therefore, this module results in the following output after extracting the TAM:

```
1       ((              NP
1.1     he              PRP         //he,P,m,s,3|
        ))
2       ((              VGADV
2.1     ((              VG          //read,v,m,s,3,is_ing|aspect='PROG'
2.1.1   is              VBZ         //be,v,m,s,3|tense='PRES'|
2.1.2   reading         VBG         //read,v,m,s,3|aspect='PROG'|
        ))
        ))
3       ((              NP
3.1     the             DT          //the,det,m,s,3|
3.2     book            NN          //book,n,m,s,3|
```

CHAPTER SIX

THE TRANSFER STAGE

6.1 Introduction

Here what goes on during the transfer stage of the translation process will be explained. Structural divergences, which were encountered when mapping complements of a sentence in English to their equivalent Amharic complements, are mentioned in this module. The equivalence rules and lexical transfer rules will be examined during the second part of this chapter.

6.2 Translation of Verb Complements between the Source and Target languages

Just to recap a structural divergence is when a complement of the verb in one language is a different type of complement in the other language. Therefore, if a verb phrase in English consists of a verb and a noun phrase while the equivalent Amharic verb phrase consists of a verb and a prepositional phrase. Many structural divergences occur when transferring verb complements between English and Amharic in this translation machine.

6.3 Equivalence Rules

Syntactic transfer systems rely on mappings between the surface structures of sentences. A collection of tree-to-tree transformations is applied to the analysis tree of the source language in order to construct a target language analysis tree.

6.3.1 Finding the English tree using equivalence rules

This equivalence rule is given an Amharic tree and has to find the corresponding English tree. There are a set of equivalence rules, each of which deal with a certain type of tree as input, some of these trees represent sentences including finite verbs, verbs with separable prefixes, verb groups including more than one verb etc. All of these rules ultimately do the same thing; they all call the same series of predicates with some minor alterations. When the Amharic tree is passed into the equivalence method it is unified with the basic tree structure representing its sentence type, in this case the basic tree is, the one representing a sentence in the present tense (i.e. with one verb).

6.4 Parsing

One of the most important components of the machine translation system is to transfer the grammatical structure of the source language to its equivalence in the target language.

Parsing the English sentence and reordering the SVO format of the English grammar to the Amharic SOV will help the transformation of grammar from source to target language.

Parsing input is the process of discovering words in a sentence that are related in structure and in meaning. A good parser will have to be able to determine functionality of words and their roles in sentence structure according to the specified grammar.

There are often many ways to parse the input texts. Parsing input strings one sentence at a time appears to be easier for the parser to handle. This is another area where grammatical structure comes in to play.

A tagger will tag all words in a sentence to find their purposes of being in the sentence. At the same time, all words will be stored in a tree of words. For this specific case the structural information that is taken includes subject, object (direct, indirect), transitivity and etc. When it determines the functionality of all words, those words will be mapped to the words that have the same meaning and same functionality in the other language.

Setting rules for mapping words from one language to another need extra cautious because often times a word have many meanings. Machine translation needs to establish a connection between the original and the target language.

There are many algorithms available to be used, but the Shakti kit has a rule based parser which takes the English sentence input and apply the following rules, context-free grammar and regular expression, which are aimed to handle the different word classes accordingly.

6.4.1 Context-free grammar and regular expressions

A regular expression is a formula and a special language that is used for specifying simple classes of strings. Thus they can be used to specify search strings as well as to define a language in a formal way. Regular expression search requires a pattern that we want to search for, and a corpus of texts to search through. A regular expression search function will search through the corpus returning all texts that contain the pattern. And another important use of regular expressions is in substitution: a string characterized by one regular expression to be replaced by a different regular expression.

A context-free grammar consists of a set of rules or productions, each of which expresses the ways that symbols of the language can be grouped and ordered together, and a lexicon of words and symbol.

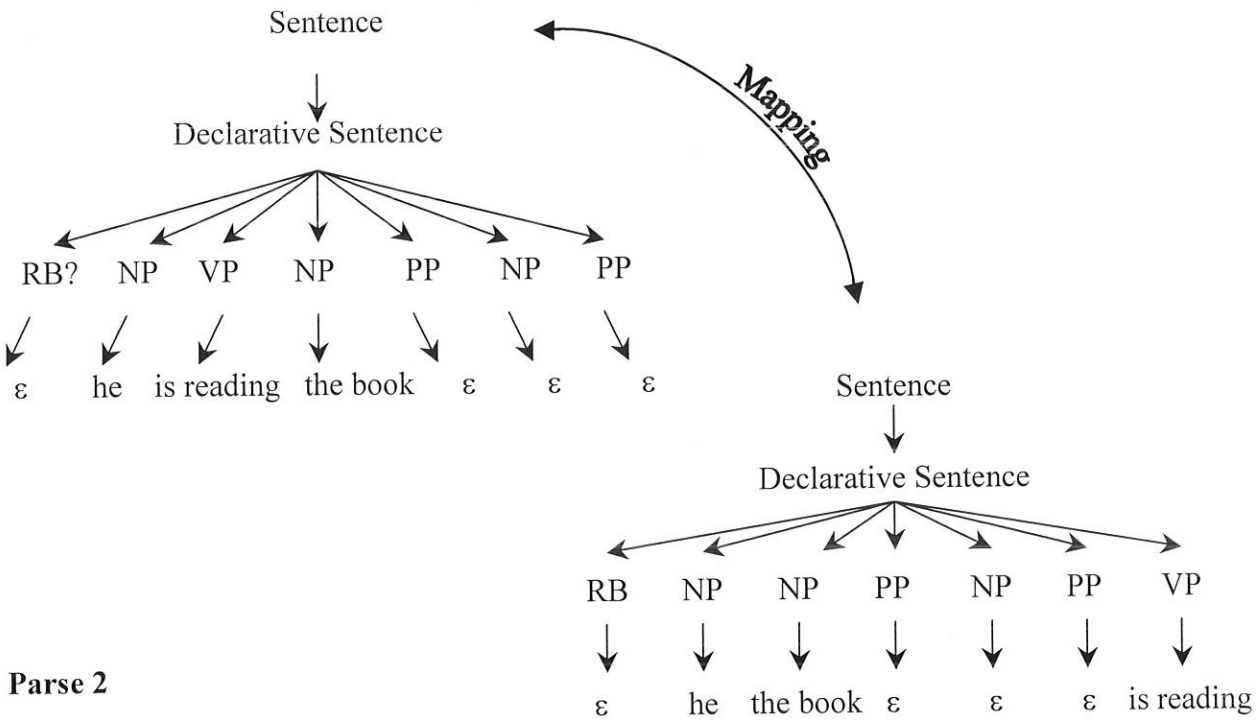
The symbols that are used in a CFG are divided into two classes. The symbols that correspond to words in the language are called terminal symbols: the lexicon is the set of rules that introduce these terminal symbols. The symbols that express clusters or generalizations of these are called non-terminals. In each context-free rule, the item to the right of the arrow (→) is an ordered list of one or more terminals and non-terminals, while to the left of the arrow is a single non-terminal symbol expressing some cluster or generalization.

Jurafsky (2000) stated that, the formal language defined by a CFG is the set of strings that are derivable from the designated start symbol. Each grammar must have one designated start symbol, which is often called S.Since context-free

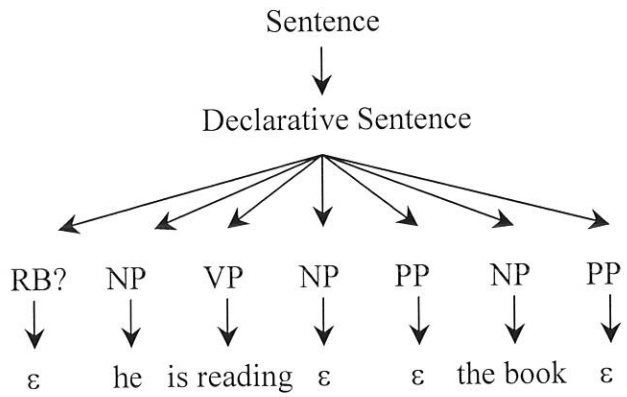
grammars are often used to define sentences, S is usually interpreted as the "sentence" node, and the set of strings that are derivable from S is the set of sentences in some simplified version of English.

The Parsed Outputs (Longest Parse trees of SENTENCE) of the given sentence in our example would be:

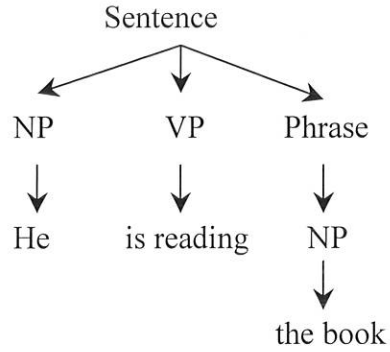
Parse 1



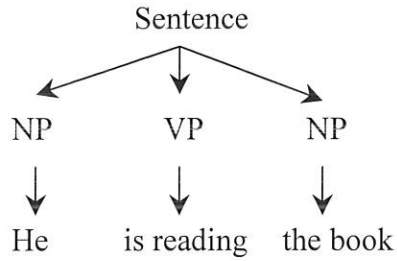
Parse 2



Parse 3



Parse 4

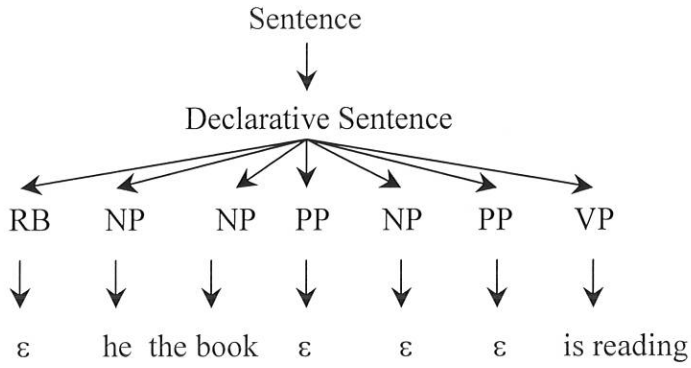


The above four parse results indicate the structural ambiguity that should be disambiguated. Such structural disambiguation is not the concern of the research. In this work the first parse result will be taken. The choice of the parse tree will have a direct effect on the structure of the target language. If the right parse is not chosen, the output may perform wrongly.

6.5 Reordering

After the Parsing is done, the next thing to be done by the system would be reordering the SVO order of the English sentence by the appropriate SOV order of Amharic.

The system takes the selected parser result and reorders the place of the words according to the predefined Amharic equivalent order



6.6 Root-to-root Amharic substitution

Lexical transfer is defined on a base lexeme defined in the lexicon. The fact that one can not list all word forms in the lexicon means one needs some recursive means of expressing meaning, i.e. compositional translation of the words along the line of compositional treatment of sentences. This raises an important question about the structure of the lexicon in general and the form of the lexical units which serve as bases for lexical transfer in particular.

This is the module that performs the lexical transfer of the English words by the Amharic words. By lexical transfer we mean the selection of the word, or words, in the Amharic language that best render(s) the concept represented by the English language word or expression. The consequence of a particular lexical transfer choice will then be treated in the structural transfer phrase. Lexical transfer has been characterized as the bottleneck of MT systems. Basically, lexical transfer is a problem because there is no isomorphism among lexical fields across language. The same word may have several different (and often seemingly unrelated) translations.

This problem is very annoying for translation in general, and for MT in particular, because the correct choices can be conditioned by circumstances ranging from general syntactic restriction to idiosyncratic properties at the lexical level. Moreover, the well-known problem of lexical gaps means that different languages choose different basic concepts to be realized as words and as phrase. To making things worse, even the way words interact in a single language does not seem to be a function of their meaning alone. And finally, there are idioms, frozen phrase that have lost their literal meaning and that simple to be memorized to be understood. Idioms are very culture-dependent, and more often than not cannot be literally translated. All of this raises doubt about the usefulness of word-to-word translation.

The foundation of lexical transfer is dictionary lookup in a cross language dictionary. The translation equivalent may be a single word or it may be a phrase. Furthermore, sometimes a generation process must subsequently inflect words in such phrases, as in this case.

It is a well-known fact that a word may have several possible translations. *Bank* is such a word. It can mean 'a place where money is kept' and also 'the land near river, etc.' fortunately there is at least two ways to tackle this problem: in the parsing or in the generation stage. This system employs the second method. This methodology treat such words as having only one meaning, and to handle the selection among multiple translations by using constraints imposed by the target language during generation. In practice, these cases are more often dealt with in the parsing stage, as the algorithms for lexical choice during generation are high overhead, especially for content words.

After the lexical transfer aided by the dictionary, the Amharic substitution of the module would be:

```
1     ((                NP     //issu,P,m,s,3|role='subj:2'|
1.1   issu              PRP    //issu,P,m,s,3|
      ))
2     ((                NP     //matsihaf,n,m,s,3|role='obj:2'|
2.1   the               DT     //the,det,m,s,3|
2.2   matsihaf          NN     //matsihaf,n,m,s,3|
      ))
3     ((                VG     //manbaab,v,m,s,3,is_ing|aspect='PROG'| name=2|
3.1   ((                VG     //manbaab,v,m,s,3,is_ing|aspect='PROG'|
3.1.1 eyannebaabaa      VBG    //manbaab,v,m,s,3|aspect='PROG'|
      ))
      ))
```

6.6.1 The dictionary

The dictionary incorporated with the system is used as a cross-reference for the root-to-root Amharic substitution of the reordered English text. Since the prototype is designed to translate only from English to Amharic, this dictionary is useful in synthesizing and generating the target language. For this specific purpose, the dictionary contains root English words, their part of speech and the equivalent root word for each in Amharic. The words included in the dictionary are selected based on the frequency of each and their relevance to the sample article to be translated by the prototype. There dictionary contains three hundred seventy one (371) words and it's included in the appendix part.

6.7 Post-processing: Handling Articles

In this section the articles 'the', 'an', and 'a' are handled. The articles 'a' and 'an' do not have any significance on the semantics of a word in Amharic. For instance 'a book' and 'book' have the same meaning. 'a' doesn't show number or possession,

and the same goes to 'an'. But, the article 'the' shows possession and the suffix '-u' is added to the word following 'the' in the English input. The final duty to be performed by the system is to produce the one line Amharic sentence output. The output of this file in our specific example would be:

Issu matsihaf eyannebaabaa naw.

CHAPTER SEVEN

EXPERIMENTATION

7.1 Testing

In order to evaluate/test the prototype the following procedure will be followed. There are one hundred pre-edited sentences, since the system is human aided the input sentences to be given to it are edited based on the constraints to be discussed in section 7.4 and one hundred phrases from an IT magazine (the article is attached on the appendix part).

The sentences and phrases are chosen on the basis of linguistic simplicity, i.e. all the sentences have only one clause. In this thesis the result is evaluated by the final output. It is also possible to evaluate the output at each phase of the output as it has been shown in the analysis and transfer stage in chapters five and six.

In evaluating the input English sentence at phase level, the step-by-step transformation of the input sentence is detected; hence the component responsible for the possible distortion or failure of the final output can easily be traced. But, since it has been demonstrated in the previous chapters in detail to show the inner working of the system, only one sentence and one phrase are taken to evaluate the output at each stage. The rest is neglected in this section to avoid repetition

7.2 Results of the prototype

a. Sample sentences translation output

Translation number	English input sentence	Amharic output sentence by the prototype	Amharic output sentence by a human translator
1	The improvement was good.	Imrovementu Xiru neber	mashashalu xiru neber
2	The science is acceptable.	saayinsu taqebay new	Saayinsu taqebaynet alaw
3	They need explanation.	Innasu falag maabrariya	innasu maabrariya yifalagalu
4	The numbers change	quxirwochu lewuxi	Quxirwochu telewuti
5	The office has many branches.	biirou aleshi many qirincafwoch	biirou bizzu qirincafwoch alew
6	Building information society is important.	building meraja hibirateseb asifalagi new	meraja hibireteseb maganbat asifalagi new
7	It is important to build information society.	meraja hibirateseb ganab ganab new it	meraja hibireteseb maganbat asifalagi new
8	What is quality?	min agolamash new	agolamash mindin new
9	Society benefits from knowledge.	hibirateseb xiqimwoch timihirit ka	hibirateseb katimihirit yixtqemal
10	Give me an answer	Six meles	Meles six
11	He is my brother	Issu wnndim new	Issu wondime new
12	Technology is good	Technology xirru	Technology xittu new
13	Breakfast makes clever	Kuris gobaz sera	Kuris xiru new
14	He is building business	Issu nigid eyeganaba new	Issu nigid eyesera new
15	Come to me	Mexa wedeinie	Wedeinie na
16	What is your name?	Man new simih?	Simih man new?
17	She is sleeping	Issua tegnitatech	Issua tegnitalech
18	Rise and shine	Teneshina abreqreqi	Tenesh
19	You are the one	Anchi andua nesh	Ante neh
20	He took it	Issu wesedew	Wesedew
21	When did he come?	Meche meta issu?	Ke yet new yemexaw?
22	It is decided	Tewesinual	Tewesinual
23	He reported yesterday	Issu tinantina report	Tinantina report argual
24	All of them	Hulum inessu	Hulum
25	When will you go?	Meche ante tihedaleh	Meche tihedaleh
26	It is not fair	Fithawi aydelem	Fithawi aydelem
27	He is my uncle	Issu yenie agot new	Agote new

28	I love him	Inie fiqir issu	Ewedewalehu
29	I care	Inie exeneqeqalehu	Exeneqeqalehu
30	I need explanation	Inie mabrariya ifeligalehu	Mabrariya efeligalehu
31	She needs a divorce	Issua mefatat tifeligalech	Mefatat tifeligalech
32	Evaluate the document	Dosiewn gemgimew	Dosiewun gemgimew
33	Drink the water	Wuhaun xexa	Wuhawun xexa
34	He eats banana	Issu muz bella	Muz yibelal
35	History is good	Tarik xirru new	Tarik xirru new
36	It it above the law	Higgu belay new	Ke hig belay new
37	Mathematics is important	Hissab asfalagi new	Hisab asfelagi new
38	My father is a manager	Abate astedadari new	Abate astedadari new
39	He increased the money	Issu birrun chemere	Birrun chemere
40	She moved his car	Issua mekina anqasaqasech	Mekinawun ankesakesechiw
41	He improved	Issu teshashilual	Teshashale
42	We prevented the problem	Igna chigir tekelakel	Chigirun asqerenew
43	We sleep at his home	Igna bettu tegnan	Issu bet tegnan
44	Winter is coming	Kiremt iyemeta new	Kiremt iyemexa new
45	This is a university	Yihe university new	Yihe university new
46	He is watching the dog	Issu wushawun iyetemelekete new	Wushawun iyexebeke new
47	He is writing	Issu iyetsafe new	Iyetsafe new
48	Sugar is not good	Sikuar xirru aydelem	Sikuar xirru aydelem
49	She is a soldier	Issua wetader nech	Wetader nech
50	Society needs peace	Hibreteseb selam felege	Hibretesebu selam yasfeligewal
51	I have a small girl	Inie tinish lig	Tinish set lij alechign
52	She is preparing food	Issua migib mazegajet	Migib iyazegajech new
53	He is short of words	Issu achir kalat	Kalat axerut
54	She is writing a poem	Issua qinie tsafech	Qine iyetsafech new
55	He is satisfied	Issu rektual	Rektual
56	He is selling	Issu iyeshexe new	Iyeshexe new
57	The police is here	Polisu izieh new	Polisochu metewal
58	He received orders	Issu tizazmaqebel	Tiizaz teqebele
59	Religion is good	Haimanot xirru new	Haymanot xirru new
60	She is eating rice	Issua ruz eyebelach new	Ruz iyebelach new
61	She comes here	Issua izieh metach	Izzih timexalech
62	This is my region	Yehe yene kilil new	Yihe yene kilil new
63	It is not me	Inie aydelehum	Inne aydelehum
64	Knowledge is important	Iwket asfalag new	Iwqet asfelagi new
65	She informed me	Issua negragnalech	Negerechign
66	We are invited	Igna tegabzenal	Tegabzenal
67	She moved her leg	Issua igir manqesaqes	Igruan anqesaqesech

68	Many people will come	Bizzu sewoch metu	Bizzu sewoch yimexalu
69	It is lack of interest	Yefilagot manes new	Yefilagot manes new
70	I love Africa	Inie afrka iwedalehu	Afrikan iwedalehu
71	The manager comes here	Astedadariw mata	Astedadariw izzih yimetal
72	He did not kill	Issu algadalam	Algedelem
73	She is coming	Issua mexach	Iyemexach new
74	He was not drinking	Issu iyexexa alneberem	Iyexexa alneberem
75	They are eating	Inessu eyebelu	Iyebelu new
76	He is defending her.	Issu eyetekelakele issua	Iyetekerakerelat new
77	I was dreaming of you	Inye anchin salim neber	Iyalemkush neber
78	We had lunch	Igna missa neberen	Missa belan
79	She is drying it	Issua iyaderekech new	Iyadereqechiw new
80	God is here	Igziabher izizh new	Igziabher izzih new
81	She is going	Issua iyehedech new	Iyehedech new
82	I am giving him food	Inye migib iyesexehut negn	Migib iyesexehut new
83	Who is evaluating them?	Man new innesun gamagama?	Yemigemegimachew man new?
84	She has decided to go	Issua hedech wesenech	Lemehed wesenech
85	He is calculating his money	Issu birrun demere	Genzebun iyasela new
86	They dream a lot.	Inessu bizzu hilem	Bizzu yalimalu
87	Forgive me	Yiqirta innie	Yiqirta argilign
88	She is explaining to him	Issua lessu mabrariya	Iyasredachiw new
89	He had the chance	Issu idilu neberew	Idilu neberew
90	Love will come to you	Fiqir ante yimexal wede	Fiqir wedeante yimexal
91	She gives him love	Issua fiqir sexech issu	Fiqir sexechiw
92	I will come by car	Innie makina mexa be	Bemakina imexalehu
93	God will give	Igziabher sexe	Igziabher yisexal
94	He is closing the business	Issu nigidun zega	Nigdun iyaqome new
95	She is coming here	Issua izzih eyemexach new	Izzih iyemexach new
96	He was born in Ethiopia	Issu ityopia meweled neber	Yeteweledew ityopia new
97	They will not have been coming	Inessu eyemexu alneberem	Laymexu yichilu neber
98	They had not been going	Inessu eyehedu alneberem	Layhedu yichilu neber
99	This is mine	Yihe yene new	Yih yene new
100	He is the applicant	Issu kesash new	Amelkachu issu new

Table 10: Sample Sentence input and Output by the machine & a human translator

b. sample phrases translated output

Translation number	English input sentence	Amharic output sentence by the prototype	Amharic output sentence by a human translator
101	Above accomplishment	kelay Kinwane	kakinwane belay
102	Lack of information	ixoot meraja ka-	yameraja ixoot
103	Watching the trouble	tamalakat girigiru	girgirun mamalakat
104	the present politics	presentu polatika	yahunu polatika
105	Information and society	meraja and hibirateseb	merajana hibirateseb
106	an attempt to baptize	an mukara xamaq	yamaxamaq mukara
107	the applicant appealed	applicaniu kasas	amalikachu kasas
108	Change his citizenship	lawax yeisuu zeginet	zeginetun lawax
109	Face the trouble	tagafax girigiru	girgirun tagafax
110	Challenge of information	fatagn-huneta meraja ka-	yamaraja fetagn-huneta
111	The ability to do	Chilotawu yeesrat	Yemesrat chilota
112	About the attempt	Mukerawu sele	Sile mukeraw
113	According to him	Kehone indessu	Indessu kehone
114	Accepting God	Meqebel Igziabherin	Igziabherin meqebel
115	Assessing the accomplishment	Eyegemegeme kinwaniewun	kiniwanewun megemgem
116	According to Sintayehu	Kehone Sintayehu wede	Sintayehu indalew
117	Bad attempt	Metfo mukera	Mexfo mukera
118	After baptism	Behuala timiqet	Keximqet behuala
119	Basic benefit	Meseretawi xeqim	Meseretawi xeqemeta
120	Before appealing	Befit igbagn	Keyigbagn befit
121	Any answer	Manignawum mels	Manignawum mels
122	Between challenges	fetenawoch mekakel be	Be fetenawoch mekakel
123	Big change	Tiliq lewx	Tiliq lewx
124	Bless him	Mebarek issu	Barkew

125	Anywhere between	Yetim kekakel	Beyetignawum bemekakel
126	Both citizens	Huletum zegoch	Huletum zegoch
127	Bringing the change	Eyamexa lewux	Lewxun mamxat
128	Business collection	nigd tiriqim	Yenigd tiriqim
129	Brave colleague	Defar yesira baldereba	Defar yesira baldereba
130	Cost benefit calculation	Waga silet xiqim	Yegudatina tiqim silet
131	Damaging Ethiopia	Eyegoda ityopia	Ityopian megudat
132	Educating the people	Eyetemare hizbu	Hizbochun mastemar
133	Defending Ethiopia	Eyetekelakele ityopia	Ityopian mekelakel
134	Culture of the country	Bahil hageru ye	Yeageritu bahil
135	Evaluating the decision	Eyegemegeme wusaniewun	Wusanewun megemgem
136	Explaining the decision	Eyabrara wusaniewun	Wusanewun mabrarat
137	Far from divorce	Ruq mefatat ke	Kemefatat yeraqe
138	Educating the government	Eyetemare mengistu	Mengistun mastemar
139	History should change	Tariq lewux neber	Tarik melewex alebet
140	Having many branches	Eyenore quirinchafich bizzu	Yebizzu qirinchafoch menor
141	Giving love	Eyesexe fiqir	Fiqirin mesxet
142	Measuring the improvement	Eyemezene meshashalu	Meshashalun melekat
143	Loving the country	Eyewedede hageru	Agerituan mewded
144	Educating mathematics	Eyexemare hisab	Hisab mastemar
145	Lack of education	Ixox timhirt ye	Yetimihirt ixox
146	Increasing the money	Eyechemere genzebu	Genzebun mechemer
147	Listening to music	Eyadamexe muzika wede	Muziqa madamex
148	Information measurement	Mereja melekat	Merejan melekat
149	Infront of me	Kefitlefit innie ke	Kefitlefite
150	Important improvements	Asfalg meshashaloch	Asfelagi meshashaloch

151	Defending the law	Eyetekelakele higu	Higin maskeber
152	Interesting movement	Eyemareke inqisqase	Maraki inqisiqase
153	Nation on movement	Hager inqisqase be	Beinqisiqase lay yale hager
154	Insurance for money	Wastina genzeb ye	Yegenzeb wastina
155	Natural measurement	Tefetro melekia	Yetefetro melekia
156	Inviting him	Eyegabeze issu	Iyegabeznew
157	The need to prevent	Asfalagi mekelakel	Yemekelakel asfelaginet
158	Measuring the quantity	Eyemelekat mextenu	Mexenun melekat
159	Questioning the religion	Eyeteyeke haimanotun	Haymanotun mexeraxer
160	Receiving the invitation	Yetefetro melekia	Gibzhawun mekebel
161	Ordering the rich	Habtamochun mazez	Habtamochun mazez
162	Preventing problem	Chigirin megtat	Chigirin megtat
163	Building the road	Mengedun megenbat	Mengedun megenbat
164	Parental satisfaction	Yewelaj irkata	Yewelaj irkata
165	Satisfying the people	Hizbochun market	Hizbochun market
166	Evaluating the students	Temariwochun megemgem	Temariwochun megemgem
167	Personal security	Yegil dehninet	Yegil dehninet
168	When somebody comes	Meche Yehone sew simexa	Yehone sew simexa
169	Coming by car	Memxat makina be	Bemakina memxat
170	Having the chance	Magnet idilun	Idilun maggnet
171	Thanking the society	Mamesgen Hibretesebun	Hibretesebun mamesgen
172	Preparing the shelf	Mazegajet Mederderiyawun	Mederderiyawun mazegajet
173	Security over satisfaction	Dehininet Beirkata lay	Beirkata lay dehninet
174	The region in need	Kililu chigger be	Bechigir lay yalew kilil
175	Near the police	Qirib polisu	Kepolisochu axegeb
176	Far from intelligence	Yerqe kebilxet	Kebilxet yeraqe

177	Movement for politics	Inqisiqase poletika ye	Yepoletika inqisiqase
178	Choice of the citizens	Mircha zegochu ye	Yezegochu mircha
179	Brave soldiers	Defar wetaderoch	Defar wetaderoch
180	Contributing to the people	Astewatsio Lehizbochu	Lehizbochu astewatsio madreg
181	Coming to here	Memxat Izzih	Izzih memxat
182	When knowledge lacks	Meche siyaxir Iwqet	Iwqet siyaxir
183	Evaluating intelligence	Megemgem Gubzinan	Gubzinan megemgem
184	All of them	Hulum inessu	Hulum
185	Coming from America	Memxat amerika ke	Keamerika memxat
186	Preserving the politics	Maqoyet Poletikawun	Poletikawun maqoyet
187	Moving the citizens	manqesaqes Zegochun	Zegochun manqesaqes
188	Building business	Megenbat Nigdun	Nigdun megenbat
189	Completing the collection	Maxenaqeq Sibisibun	Sibisibun maxenaqeq
190	Causing the damage	Mensie tifat	Yetifatu mensie mehon
191	Explaining for a friend	Mabrarat guadegna le	Leguadegna mabrarat
192	Directing the government	Memrat Mengistun	Mengistun memrat
193	Having an idea	Alegn hasab	Yehasab menor
194	Have been coming	Alegn memxat	Iyemeta neber
195	Being an applicant	Mehon Amelkach	Amelkach indemehonu
196	Loving you	Iyewededkuh anten	Iyewededkuh
197	Having said this	Norogn yihe tenagere	Yihenin kalku behuala
198	Knowledge for us	Iwqet igna le	Iwqet leigna
199	Computer for development	Computer idget le	Computer leidget
200	Importance of information	Xeqemeta Yemereja	Yemereja xeqemeta

Table 11: Sample Phrase Input and Output by the Machine& a Human Translator

7.3 Evaluation of the prototype

This prototype translates only one sentence at a time. If the entire article were entered into the system without pre-editing a sound translation would have not been feasible. When pre-editing the source sentences from the article, the number of prepositions in a sentence has been reduced to one and all the sentences joined by conjunctions and disjunction have been treated separately as independent sentences. The pre-editing is done based on the capabilities of the system. As it has been already discussed the system handles a single clause input English sentences, and punctuation is removed.

All the words used in the sample test sentences exist already in the dictionary; hence the system does not have to treat any new word out of the dictionary. However, in, under such circumstances the prototype returns the words back unaltered. This is done to improve the system's robustness.

The other drawback (to be improved in the future) of the system is that if the input sentence is given in any different word order that the usual SVO, like translation number 7 in table 7.1, it would not give the correct translation. And as it can be seen in translation number 1 and 6, the system gives out the adjectives without translating them as they are in the target language.

Another problem with this system is with the parser used. The parser of the system takes the first parse without analyzing the other alternatives, this is because the system does not have any semantic information at its disposal, and it cannot chose the

best choice if a number of possible translations arise. This definitely resulted in mismatches of meanings and misplacement of the prepositions in translation number10, 12 and 20.

The system gives good results for a sentence consisting of five words with a single preposition. And if the root word is in the dictionary it properly understands the TAM of the word. As it can be seen in the result section the output of the system has better efficiency for phrases than sentences. Even under a condition where a word has more that one form, the system gives out a logical translation due to the chunker which primarily identify the word class as in has been demonstrated in chapter five. When the system cannot translate an entire phrase or sentence, a direct translation of every word is a possible solution. This may lead to inaccurate translations, but at least the system gives its best attempt.

To sum up with, the results shown above for this prototype can be considered as encouraging. Even though the whole process of pre-editing is time-taking, if it is done properly following the rules set for the system, it will in the end lead to better results. It has been confidently shown that the system works satisfactorily for sentences broken down to a single clause level. An example that show how the system work was given in chapter one, and some more examples that take one input sentence and one phrase from the sample test data are taken and demonstrated below.

CHAPTER EIGHT

CONCLUSIONS AND RECOMMENDATIONS

8.1 Conclusions

The primary objective of this thesis is to customize a prototype transfer based translation system, which could translate documents from English into Amharic. It is based on the architecture of the English- Hindi translation system, which used powerful tools LT-POS and LT- CHUNK, which highly facilitated the translation process. It uses constituent structure at the chunk level and dependency relations at the sentence level of analysis. It has specialized components for word sense disambiguation, parsing, Preposition attachment, phrasal verb identification, transfer grammar, sentence and word generation, and many others.

The thesis began by justifying the need of machine translation in the Ethiopian context. At the end of the first chapter a sample sentence with all the steps of translation is included to briefly show how the entire system operates. Following this, a condensed overview on the discipline of machine translation, its history, approaches, importance and its prospect in the years to come is discussed. In this section the scientific approach that is being employed these days with the help of statistical and knowledge based approaches is appropriately discussed.

Then the core of the translation system that is the linguistic analysis is presented in detail. In this section the basic grammatical rules of Amharic are reviewed in comparison with the source language to be translated, the English language. Since the machine tries to approach the translation process from a syntactic transfer perspective, the parse trees of Amharic along with its generation and analysis are discussed. The differences between the two languages and the way they will affect the translation process are outlined at the end of this section.

The succeeding parts of the thesis give a deeper explanation on the analysis and transfer steps of the system. It has been tried to present all the resources used during the translation process with a detailed discussion, which relates the linguistic part with the technical aspect.

Finally all the customization and designing is tested against a pre-edited set of sample sentences. These sentences have been selected from a linguistics point of view. Had there been more time the dictionary would have been further developed and many more structurally various types of sentences would have been chosen to test the system. The output of the system is considered as satisfactory as compared with the constraints face during the thesis work. The reasons for not being able to extend the translation system are:

1. Lack of time allotted for the research
2. Financial constraint to perform a research of this scale
3. Lack of electronic bilingual dictionaries vital to a translation system

4. Lack of parallel corpora in Amharic for detailed experimentation
5. Lack of detailed resource on the lexical analysis of the Amharic language

Apart from such constraints, in the process of achieving the primary objective there were many parallel contributions of this thesis in the field of NLP for Amharic language. The thesis dealt with grammatical issues of Amharic starting from morpheme up to semantic analysis of complicated sentences. It has been shown that Amharic is a language with complex morphology and that it requires still a more detailed research in the field to easily accommodate the language to computer application software.

This thesis presented a condensed and very explicit comparison between the English and Amharic languages. This comparison can be extended and used as an accessory in the relevant NLP areas such as morphology related software and domains that require syntactic transfer.

The output of the system, with the necessary modification and update would to a certain extent serve as a communicating media for the majority of Ethiopians that cannot speak the English language. And if other researches continue developing and customizing this system from English or other foreign languages to any of the local languages used in Ethiopia, this thesis would have contributed significantly.

8.2 Recommendations - Future improvements

Many components of the system require further and in-depth research to come up with a better functional translation system. The following are some of the areas of research that can be conducted based on this thesis:

1. This system demands better parsing algorithms that can offer all the possible parses and choose the best one among the alternative outputs.
2. The numbers of words that the system uses are numbered and it cannot be a full-fledged system with these limited words.
3. The lexical and structural ambiguity during transfer from the English language to Amharic has a lot yet to be done.
4. This system handles only single sentences and much has to be done to come up with a system that handles more number of sentences.
5. The system is poor in identifying semantic content of words; hence this has to be improved.
6. Preposition handling is a key task that awaits to be worked on.
7. The system is highly dependent on the order of the input source language and this could be improved.
8. The system can be made to translate from English to Amharic and back from Amharic to English.
9. The system would operate better if there were adequate and functional morphological analyzers, generators and synthesizers for the Amharic language.

10. The role that an efficient Amharic parts of speech tagger plays in such a system is remarkable. Any structured and domain oriented attempt in that direction would make the system fair and sound.
11. The backbone of this system is the bilingual dictionary. Therefore coming up with a dictionary that can be used in support of a variety of natural language tasks in view of cross-reference, morphological analysis of inflections, synonyms, linguistic concepts and additional grammatical information would improve the performance of the system considerably.
12. Making the system web-based and preparing an interface that can easily integrate it with other existing translation engines would facilitate the improvement of the machine and to a certain extent would enrich the Amharic language.
13. Working on a system that can handle the pre-editing and post-editing task of the human translator would make the translation system full-fledged.
14. The design of a semantic analyzer as a module of the Translation System would yield priceless contribution to the efficient and productive full-fledged system.

REFERENCES

Abiyot Bayou. 2000. Design and Development of Word Parser for Amharic Language, Master's Thesis at the School of Information Studies for Africa, Addis Ababa University, Addis Ababa.

Arnold d. J. 1995. Why MT matters. http://clwww.essex.ac.uk/doug/book/ node5 .html

Atelach Argaw. 2002. Automatic Sentence Parsing for Amharic Text: An Experiment Using Probabilistic Context Free Grammars, Master's Thesis at the School of Information Studies for Africa, Addis Ababa University, Addis Ababa.

Atelach Alemu, Lars Asker, and Mesfin Getachew. 2003. Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward, http://www.sciences.univ.fr/irin/taln2003/articles.alemu.pdf

Comrie, Bernard 1994. 'Tense,Aspect,Mood' in: The Encyclopedia of Language and Lingutistics. Oxford, Pergamon Press, 4558-63.

Hedden, Thomas D. 2000. Machine Translation: A Brief Introduction, http:/www.he.net/hidden-intro_mt.htm

Hutchins, W.J. and Somers, H.L. 1992. An Introduction to Machine Translation. Academic Press, London.

Jespersen O. 1924. The Philosophy of Grammar, London,136.

Jurafsky D. and James H. Martin 2000. Speech and Language Processing, Prentice Hall, Upper Saddle River, New Jersey.

Kibur Lisanu. 2002. Design and Development of Automatic Morphological Synthesizer for Amharic Perfective Verbs, Master's Thesis at the School of Information Studies for Africa, Addis Ababa University, Addis Ababa.

Lesalu, Wolf. 1995. A Reference Grammar of Amharic. Wiesbaden, Harrassowitz, Belgium.

Locke, W.N and Booth A.D, (Editors). 1955. Machine Translation of Languages, New York, Wiley.

Mesfin Getachew. 2001. Automatic Part of Speech Tagging for Amharic Language: an Experiment Using Stochastic Hidden Markov Model HMM Approach, Master's Thesis at the School of Information Studies for Africa, Addis Ababa University, Addis Ababa.

Mikheev Andrei. 2000. A Workbench for Finding Structure in Texts, www.ltg.ed.ac.uk/software/pos/steve-anlp.ps

Mulu G/ Egziabher. 2001. Developing English-Tigrinya Machine Readable Bilingual Dictionary: An Input for Machine Translation, Master's Thesis at the School of Information Studies for Africa, Addis Ababa University, Addis Ababa

Murthy K. N. 1999. 'MAT: A Machine Assisted Translation System,' in: Proceedings of the NLPRS-99 Fifth Natural Language Processing Pacific Rim Symposium, Beijing, China, Nov 5-7.

Noone G. 2003. Machine Translation: A Transfer Approach, www.cs.tcd.ie/courses/csll/nooneg2003.pdf

Sebsibe H/Mariam. 2001. Construction of English-Amharic Electronic Subject Dictionary for Science and Technology Terms: An Experiment with Mathematical Terms, Master's Thesis at the School of Information Studies for Africa, Addis Ababa University, Addis Ababa.

Tesfaye Bayu. 2002. Automatic Morphological Analyser: An Experiment Using Unsupervised and Autosegmental Approach, Master's Thesis at the School of Information Studies for Africa, Addis Ababa University, Addis Ababa.

Trost. 2000. Morphology. http://www.iai.uni-sb.de/docs/morphology.pdf.

APPENDICES

APPENDIX 1. LISTOF WORDS FOUND IN THE DICTIONARY

"ability","N","ciloota"

"able","Adj","yemicil"

"about","Adv","sila-"

"about","Prep","sila-"

"above","Prep","belay"

"above","Adj","kelay"

"accept","VT","taqabal"

"acceptable","Adj","taqebay"

"acceptance","N","teqebayinat"

"accomplish","VT","kawen"

"accomplishment","N","Kinwane"

"according","Prep","bezih"

"activity","N","kinwun"

"afraid","Adj","fari"

"after","Prep","behuala"

"Africa","N","africa"

"again","Adv","indegana"

"always","Adv","hulgizie?"

"am","AuxV","negn"

"anti","N","tsare"

"and","Conj","ina"

"animal","N","ensisa"

"answer","malis",""

"any","Adj","minim"

"anybody","Pron","manim"

"anything","Pron","manignawum"

"anywhere","Adv","yetim"

"appeal","N","mamalkat"

"appeal","V","kasas"

"applicant","N","amalkac"

"application","N","mamalkaca"

"at","Prep","ba-"

"attack","N","xiqaat"

"attack","VT","axaq"

"ask","VT","xayaq"

"assess","VT","gamagam"

"attempt","N","mukara"

"attempt","VTI","mokar"

"attend","VTI","masataf"

"baby","N","hitsen"

"bad","Adj","maxifo"

"bank","N","bank-bet"

"bank","N","malika"

"baptize","VT","xamaq"

"baptism","N","ximiqat"

"basic","Adj","masarata"

"basket","N","qiricat"

"is","V","new"

"are","V","nachew"

"was","V","neber"

"were","V","neberu"

"before","Adv","qadim"

"before","Prep","bafit"

"begin","VT","jamar"

"behind","Prep","houala"

"benefit","N","xiqim"

"benefit","VT","xaqam"

"besides","Prep","kagon"

"between","Prep","mekakel"

"between","Prep","1.ke_bIca"

"bible","N","matsihaf-qidus"

"big","Adj","gizuf"

"bless","VT","barak"

"blessed","Adj","yatabaraka"

"body","N","akalat"

"book","N","matsihaf"

"born","v","tawalad"

"both","Conj","hulatum"

"brain","N","aimiro"

"branch","N","qirincaf"

"brave","Adj","jagina"

"brave","VT","jagan"

"break","VT","sabar"

"breakfast","N","quris"

"bring","VT","amax"

"brother","N","wondim"

"brown","Adj","buna"

"build","N","ginibata"

"build","VT","ganab"

"business","N","nigid"

"but","Conj","negar-gin"

"by","Prep","ba-"

"calculate","V","asal"

"can","N","xasa"

"can","V","cal"

"cancel","VT","saraz"

"car","N","makina"

"care","N","Inkibkabie"

"carry","VT","tashekam"

"category","N","madab"

"chair","N","wenber"

"challenge","N","fatagn-huneta"

"chance","N","idil"

"change","N","zirzir"

"change","N","lewuxi"

"change","V","lawax"

"choice","N","miricaa"

"choose","V","marax"

"citizen","N","zeginet"

"city","N","katama"

"clean","Adj","nitsuh"

"clean","V","atsad"

"clever","Adj","gobaz"

"close","Adj","zig"

"close","V","zag"

"coffee","N","bunaa"

"cold","Adj","qizqaaza"

"colleague","N","baldaraba"

"collect","V","sasab"

"collection","N","sibsib"

"come","V","max"

"community","N","mehiberaseb"

"complete","VT","fatsam"

"condition","N","hunatie"

"construct","VT","ganab"

"continent","N","ahigur"

"contribute","VT","awawax"

"contribution","N","mawaaco"

"cost","N","wega"

"country","N","hager"

"culture","N","baahil"

"cut","VTI","qorax"

"damage","N","gudat"

"damage","VI","god"

"daughter","N","set-lij"

"day","N","qan"

"dead","N","mot"

"decide","V","wasan"

"decision","N","wusanie"

"defend","V","takalakal"

"direct","Adj","qaxitaa"

"direct","VT","amalakat"

"directly","Adv","baqaxitaa"

"direction","N","aqixaaca"

"divorce","N","ficii"

"divorce","VT","fat"

"do","V","sar"

"do","VT","sar"

"document","N","sanad"

"dog","N","wusha"

"down","Prep","wada-taach"

"dream","N","hilim"

"dream","V","alam"

"drink","N","maxaxi"

"drink","V","xax"

"dry","Adj","direq"

"dry","VI","daraq"

"dry","VT","daraq"

"dust","N","abuara"

"each","Adv","iyandandu"

"earth","N","midir"

"eat","VT","bal"

"educate","VT","astamar"

"education","N","timihirit"

"Ethiopia","NP","ethiopia"

"evaluate","V","gamgam"

"even","Adj","mulu-quxir"

"evening","N","meta"

"every","Adj","hul"

"exam","N","fatana"

"explain","VT","abirar"

"explanation","N","maabrariya"

"far","Adj","ruq"

"for","Prep","le-"

"form","Prep","ka-"

"friend","N","guadanga"

"from","Prep","ka"

"fuel","N","nedaj"

"full","Adj","mulu"

"future","N","Yawadafit"

"give","VT","sax"

"go","V","heda"

"God","N","igiziabher"

"god","N","xaot"

"good","Adj","Xiru"

"government","N","mengist"

"have","V","al"

"had", "V","neber"

"has","V","al"

"he","Pron","isuu"

"head","N","iras"

"health","N","xeena"

"her","Pron","yesua"

"here","Adv","izihi"

"him","Pron","lesu"

"his","Pron","yeisuu"

"history","N","taarik"

"home","N","menoria_bet"

"hour","N","saat"

"house","N","bet"

"how","INT","indeit"

"human","Adj","yesew"

"human","N","sew"

"hundred","N","meto"

"I","Pron","inie"

"idea","N","hasab"

"if","Conj","bihon"

"important","Adj","asifalagi"

"improve","V","ashashal"

"improvement","N","mashashal"

"in","Prep","wusix"

"increase","V","camaraa"

"infornt","Prep","fitlefit"

"inform","VTI","asawak"

"information","N","meraja"

"inside","Prep","wusix"

"insurance","N","medin"

"intelligence","N","ginizabie"

"interest","N","filagot"

"invitation","N","gibizja"

"invite","VT","gabaz"

"is","V","new"

"knife","N","bilawa"

"knowledge","N","iwuket"

"labour","N","gulibat"

"lack","V","aax"

"lack","N","ixoot"

"language","N","quanqua"

"large","Adj","tiiliq"

"law","N","hig"

"lazy","Adj","sanaf"

"leaf","N","qixal"

"love","N","fiqir"

"love","VTI","afakar"

"male","N","wonid"

"manager","N","halaafi"

"many","Adj","bizu"

"mathematics","N","hisab"

"me","Pron","lanie"

"measure","V","lakk"

"measurement","N","malakiya"

"mine","pron","yene"

"money","N","genzeb"

"move","V","anqasaqas"

"movement","N","inqisqaasie"

"music","N","muziiqaa"

"my","Pron","yenie"

"name","N","sim"

"name","VT","sayam"

"nation","N","zegoch"

"national","Adj","biherawi"

"nationality","N","zeginet"

"natural","N","tafaxiro"

"nature","N","tafaxiro"

"near","Prep","qirb"

"need","V","falag"

"new","Adj","adis"

"no","Det","ayidelem"

"not","Adv","ayidalem"

"now","Adv","ahun"

"number","N","quxir"

"of","Prep","ka-"

"office","N","biiro"

"on","Prep","layi"

"on","Adv","layi"

"or","Conj","wayim"

"order","N","qidamtakatal"

"other","Det","lelooch"

"our","Pron","yegna"

"over","Adv","balay"

"over","Prep","balay"

"parent","N","walaji"

"pass","V","alaf"

"past","N","halafi"

"people","N","hizib"

"phrase","N","harag"

"plant","N","itsiwaat"

"plant","N","takil"

"plant","V","tak"

"poem","N","qinee"

"police","N","tsaxita-askabari"

"politics","N","polatika"

"prepare","V","azagaj"

"present","N","ahun"

"present","V","aqarab"

"preposition","N","mastawadid"

"prevent","V","takalakal"

"price","N","wega"

"problem","N","chigir"

"quality","N","agolamash"

"quantity","N","maxan"

"question","VTI","xayaq"

"receive","VI","taqabal"

"red","Adj","qey"

"refrigerator","N","maqazqezsa"

"region","N","killil"

"religion","N","hayimanot"

"rice","N","ruz"

"rich","Adj","habtam"

"river","N","waniz"

"road","N","mengad"

"satisfaction","N","irikataa"

"satisfy","V","rak"

"science","N","saayins"

"sea","N","bahir"

"season","N","weqit"

"security","N","dehininet"

"sell","V","shax"

"sentence","N","arafita-negar"

"she","Pron","isua"

"sheep","N","beg"

"shelve","N","mederderia"

"shirt","N","shemiz"

"short","Adj","acir"

"sleep","N","magnita"

"sleep","V","tagn"

"small","Adj","tinish"

"society","N","hibirateseb"

"soldier","N","watadar"

"some","Det","yetawasana"

"somebody","Pron","yehona-saw"

"soon","Adv","tolo"

"sugar","N","sikuar"

"sun","N","tsehay"

"thank","VT","misigana"

"that","Det","ya"

"their","Pron","yeinnasu"

"these","Det","inazihi"

"they","Pron","innasu"

"thin","Adj","qacin"

"this","Pron","yih"

"this","Det","yih"

"those","Pron","inaziya"

"thousand","N","shii"

"three","N","sost"

"Thursday","N","hamus"

"time","N","gizie"

"to","Prep","wada"

"today","Adv","zarie"

"today","N","zarie"

"trip","N","shirishir"

"trouble","N","girigir"

"understand","VTI","tarade"

"university","N","yuniversity"

"up","Prep","wedalayi"

"watch","N","yaiji-saat"

"watch","VTI","tamalakat"

"water","N","wuha"

"way","N","mangad"

"we","Pron","ignaa"

"weak","Adj","dakamaa"

"what","Det","min"

"what","Interro","min"

"when","Det","meche"

"when","Interro","meche"

"where","Det","yet"

"where","Interro","yet"

"which","Det","yetu"

"which","Interro","yetu"

"white","Adj","naci"

"who","Det","man"

"who","Interro","man"

"whom","Interro","leman"

"whom","Det","leman"

"whose","Interro","yeman"

"whose","Det","yeman"

"why","Det","lemin"

"why","Interro","lemin"

"will","0","yi-"

"window","N","maskot"

"winter","N","kiremt"

"with","Prep","ke-"

"write","VT","tsaf"

"wrong","Adj","yetasasata"

"yes","N","aawo"

"you","Pron","aante"

"you","Pron","innaante"

"yours","pron","yenante"

APPENDIX 2. GENERAL DESCRIPTION: THE LT_CHUNK USED IN THE SYSTEM

This specific LT_CHUNK software is used for the disambiguation of words part-of-speech and the identification of simple noun and verb groups in English texts. It can handle both plain ASCII and (n)SGML/XML marked-up files. It includes:

ltchunk - text chunking program;

lttok - English tokenizer build on Finite State Automata

ltpos - probabilistic part-of-speech tagger based on Hidden Markov Models plugged with Maximum Entropy probability estimators;

sgtransduce - a grammar interpreter which is supplemented with grammars to recognize verb and noun groups; and other supporting utilities;

A. Running the chunker over ASCII texts

An example of chunking an ASCII text is

>> cat EXAMPLES/text | bin/ltchunk

The output of the program will contain noun groups enclosed into [[ ]] brackets and verb groups enclosed into (( )) brackets:

[[previous government investigations]]that

[[both Rear Admiral Husband E. Kimmel]]and [[his Army counterpart]],

[[Maj. Gen. Walter C. Short]], ``((committed))[[errors]]of [[judgment]".

We can also output part-of-speech information with every word using the flag show_tags:

```
>> cat EXAMPLES/text    | bin/ltchunk -show_tags
```

In this mode the chunker will output words as word_TAG structures:

```
[[ A_DT Pentagon_NNP study_NN ]](( re-affirmed_VBD ))
[[ the_DT conclusion_NN ]]of_IN [[ previous_JJ government_NN investigations_ ]]
that_IN [[ both_DT Rear_NNP Admiral_NNP Husband_NNP E._NNP Kimmel_NNP
]]
and_CC [[ his_PRP$ Army_NNP counterpart_NN ]],_, [[ Maj._NNP Gen._NNP
Walter_
C._NNP Short_NNP ]],_,
``(( committed_VBD ))[[ errors_NNS ]]of_IN [[ judgment_NN ]]"._.
```

APPENDIX 3. LT POS (tags)

LT POS can be used with the Penn Treebank tagset. It is the tag set used in the LT POS demo. Here are the most important tags.

The Penn Treebank. In *Computational Linguistics*, vol. 19, no. 2, pp313-330.

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, suPERLative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*'s*
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, suPERLative	best
RP	particle	give *up*

TO	to	*to* go, *to* him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Table 12: Some Important Penn Treebank Tagset