BANGLA TO ENGLISH TRANSLATION ENGINE

M M ASADUZZAMAN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Bangla to English Translation Engine

A thesis submitted by

M M Asaduzzaman

Roll No. 100105009

for the partial fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering (M.Sc.Engg.) examination held on May 31, 2003.

APPROVED BOARD OF EXAMINERS

1. M Masroor

Dr. Muhammad Masroor Ali
Associate Professor and Associate Director
Institute of Information and Communication Technology (IICT)
Bangladesh University of Engineering and Technology (BUET)
Dhaka-1000, Bangladesh

Chairman
(Supervisor)

2.

Dr. Mohammad Kaykobad
Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka-1000, Bangldesh

Member

3.

Dr. Md. Abul Kashem Mia
Associate Professor and Head
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka-1000, Bangladesh

Member
(Ex-officio)

4. Md. Zafar Iqbal 31/5/03

Dr. Mohammad Zafar Iqbal
Professor and Chairman
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet-3114, Bangladesh

Member
(External)

Submitted to

Bangladesh University of Engineering and Technology

in partial fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering (M.Sc.Engg)

CANDIDATE'S DECLARATION

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award if any degree or diploma.

Signature of the Candidate

M M Asaduzzaman

Bangla to English Translation Engine

by

M M Asaduzzaman

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
Dhaka-1000, Bangladesh

2003

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AP	Adjective Phrase
ATN	Augmented Transition Network
CL	Computational Linguistics
FP	First Person
FSA	Finite State Automata
FST	Finite State Transducer
FSTN	Finite State Transition Network
LK	Linguistic Knowledge
MT	Machine Translation
NLP	Natural Language Processing
NP	Noun Phrase
PL	Plural
QNTFR	Quantifier
SG	Singular
SPCFR	Specifier
SP	Second Person
SPH	Second Person Honorific
SPNH	Second Person Non-Honorific
VP	Verb Phrase

Acknowledgements

I am indebted to Dr. Muhammad Masroor Ali, Associate Professor and Associate Director, Institute of Information and Communication Technology (IICT), BUET, the supervisor of the thesis, for his continuous assistance. His inspiration for doing research in machine translation and constant encouragement, infinite patience and understanding, and willingness to let me find my own wings, have all contributed to my humble development as an MT practitioner. I do acknowledge the support from Dr. Md. Abul Kashem Mia, Head, Department of Computer Science and Engineering, BUET. At the center of IICT, I would like to thank Dr. S. M. Lutful Kabir, Director, IICT, for providing me the opportunity to accesss the internet facilities at the Internet browsing center. I found significant support from there.

Finally, I would like to thank my family members for all their support and encouragement throughout my M.Sc. Engg.

Abstract

This thesis work focuses on the issues in designing a translation engine for Bangla to English translation. Machine translation (MT) from one natural language to another is not an easy task. It requires extensive knowledge about the involved languages. Machine Translation is an emerging paradigm for processing natural languages. Machine Translation and Natural Language Processing (NLP) are being used by language industries for translating one language to another. MT invokes a variety of techniques for language translation and machine translation engines are developed using one of the techniques. The MT engine has been developed using linguistic knowledge (LK) architecture. The proposed translation engine can translate only simple Bangla sentences into English. It is upgradeable for translating other types of sentences. The translation process involves analysis the Bangla sentence syntactically, semantically and morphologically. After all sorts of analysis the tokens and grammatical information are transferred through the transfer component. The target system then generates output English sentence from the transferred tokens and grammars. Thus the translation process completes.

Key Words: Machine translation, syntax, semantics, morphology, transfer, synthesis.

Chapter 1

Introduction

1.1 What is Machine Translation?

Machine Translation (MT) [1][2][3] refers to automatic translation by machine. More simply it can be said that machine translation is the application of computers to the task of translating texts from one natural language to another. The translation from one natural language to another is not an easy task by machine. One of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a number of systems are available which produce output of sufficient quality to be useful in a number of specific domains.

1.2 Brief History of Machine Translation

At the end of the 1950s, researchers in the United States, Russia, and Western Europe were confident that high-quality machine translation of scientific and technical documents would be possible within a very few years. After the promise had remained unrealized for a decade, the National Academy of Sciences of the United States published the much cited but little read report of its Automatic Language Processing Advisory Committee (ALPAC). The ALPAC report recommended that the resources that were being expended on MT as a solution to immediate practical problems should be redirected towards more fundamental questions of language processing that would have to be answered before any translation machine could be built. The number of laboratories working in the field was sharply reduced all over the world, and few of them were able to obtain funding for more long-range research programs in what then came to be known as computational linguistics (CL) [3].

There was a resurgence of interest in machine translation in the 1980s and, although the approaches adopted differed little from those of the 1960s, many of the efforts, notably in Japan, were rapidly deemed successful. This seems to have had less to do with advances in linguistics and software technology or with the greater size and speed of computers than with a better appreciation of special situations where ingenuity might make a limited success of rudimentary MT. The most conspicuous example was the METEO system, developed at the University of Montreal, which has long provided the French translations of the weather reports used by airlines, shipping companies, and others. Some manufacturers of machinery have found it possible to translate maintenance manuals used within their organizations largely automatically by having the technical writers use only certain words and only in carefully prescribed ways.

Now machine translation is being practiced as an emerging field of computer science which covers a wide range of knowledge applicable in this field. Besides these practical MT systems are available in the web sites or in form of software products. MT systems such as SYSTRAN, HEISOFT are favorite products on the web for translating sentences between a pair of languages that are supported by the system. Also researchers from all over the world are doing valuable research in MT.

1.2.1 Why Machine Translation is Difficult?

Many factors contribute to the difficulty of machine translation, including words with multiple meanings, sentences with multiple grammatical structures, uncertainty about what a pronoun refers to, and other problems of grammar. But two common misunderstandings make translation seem altogether simpler than it is. First, translation is not primarily a linguistic operation, and second, translation is not an operation that preserves meaning.

1.3 The Structure of Machine Translation Systems

While there have been many variants, most MT systems, and certainly those that have found practical application, have parts that can be named for the chapters in a linguistic text book. They have lexical, morphological, syntactic, and possibly semantic components, one for each of the two languages, for treating basic words, complex words, sentences and meanings. Each feeds into the next until the last one in the chain produces a very abstract representation of the sentence.

There is also a "transfer" component, the only one that is specialized for a particular pair of languages, which converts the most abstract source representation that can be achieved into a corresponding abstract target representation. The target sentence is produced from this essentially by reversing the analysis process. Some systems make use of a so-called "interlingua" or intermediate language, in which case the transfer stage is divided into two steps, one translating a source sentence into the interlingua and the other translating the result of this into an abstract representation in the target language.

1.4 Prospects of Machine Translation

Machine translation can play a vital role in translating instruction manuals for international products such as machinery, medicine, cosmetics, foods, cloths, etc. To introduce Bangladeshi products all over the world it is necessary to supply the instructions for the products in various languages. For human translators this sort of works may not be possible due to a lot of reasons. So an automatic translator is the acceptable candidate in this regard. The translator will be effective mainly for speed of translation. An automatic translator is capable of translating information far more faster than human translators. Only the drawback is that the translation output may not be accurate. So post editing must be done to tune the desired target output.

1.5 Importance of Bangla Machine Translation

Bangla is our mother tongue. Bangla is only the language in the world for which our people sacrificed their lives in 1952. For this reason the 21st February is now observed all over the world as "International Day of Mother Language". This is obviously a pride for us. So the widest possible distribution of Bangla around the world is impossible without machine translation. For human translator this job is not feasible for a lot of reasons. The only way to use machines or computers to do the job of translation.

With the advent of Internet technology and electronic commerce have increased the demand for automatic machine translation of sufficient quality for determining the content of a web page. Demand of machine translation has grown and will continue to grow steadily [2]. A variety of authoring tools and document production techniques have also made linguistic information available in a variety of formats. The information placed on the web is needed to be translated to the language of the user for his/her understanding. If he/she tries to translate the whole contents word-by-word or sentence-by-sentence it will cost a lot of time. An automatic translator can do the job. Translation by machine will save a lot of time in this regard. To produce correct output will require a significant number of problems to be solved. The output produced by the machine is needed to be edited by the expert of that language, if the system is not fully automated. In many cases the gist content of the page or document can be obtained only by machine translation without post editing.

A significant number of MT systems are found on browser over the Internet that can translate a page into one or more languages. Moreover many language industries all over the world are engaged in multilingual translation of a language. Multilingual translation means the translation a language into several languages. The aim of machine translation of a particular language is to introduce the language to others who do not know the language. They will just use a machine for the job of translation. The explosion in electronic communication and the use of computers has explored the way of taking of technology to record, disseminate and maybe even preserve the language and to exploit approaches and techniques that are already tried and tested on more common languages.

1.5.1 Political Importance

All over the world there are thousands of languages. Political issues regarding a nation or a country are found in the language of that country or nation. The problems and solutions of political issues may be applicable to other nations or countries. If the translations of those issues are distributed in several languages then it will bring some good impacts for the nation or country that is eagerly looking for such kind of solutions. As it has mentioned earlier the crisis of human translators will be resolved by introducing MT systems.

The social or political importance of MT arises from the socio-political importance of translation in communities where more than one language is generally spoken. Here the only viable alternative to rather widespread use of translation is the adoption of a single common "lingua franca", which is not a particularly attractive alternative, because it involves the dominance of the chosen language, to the disadvantage of speakers of the other languages, and raises the prospect of the other languages becoming second-class, and ultimately disappearing. Since the loss of a language often involves the disappearance of a distinctive culture, and a way of thinking, this is a loss that should matter to everyone. So translation is necessary for communication - for ordinary human interaction, and for gathering the information one needs to play a full part in society. Being allowed to express yourself in your own language, and to receive information that directly affects you in the same medium, seems to be an important, if often violated, right. And it is one that depends on the availability of translation. The problem is that the demand for translation in the modern world far outstrips any possible supply. Part of the problem is that there are too few human translators, and that there is a limit on how far their productivity can be increased without automation. In short, it seems as though automation of translation is a social and political necessity for modern societies that do not wish to impose a common language on their members [1].

1.5.2 Commercial Importance

The commercial importance of MT is a result of related factors. First, translation itself is commercially important: faced with a choice between a product with an instruction manual in English, and one whose manual is written in Japanese, most English speakers will buy the former - and in the case of a repair manual for a piece of manufacturing machinery or the manual for a safety critical system, this is not just a matter of taste. Secondly, translation is expensive. Translation is a highly skilled job, requiring much more than mere knowledge of a number of languages, and in some countries at least, translators' salaries are comparable to other highly trained professionals. Moreover, delays in translation are costly. Estimates vary, but producing high quality translations of difficult material, a professional translator may average no more than about 4-6 pages of translation (perhaps 2000 words) per day, and it is quite easy for delays in translating product documentation to erode the market lead time of a new product.

1.5.3 Scientific Importance

Scientifically, MT is interesting, because it is an obvious application and testing ground for many ideas in Computer Science, Artificial Intelligence, and Linguistics, and some of the most important developments in these fields have begun in MT. To illustrate this: the origins of Prolog, the first widely available logic programming language, which formed a key part of the Japanese "Fifth Generation" programme of research in the late 1980s, can be found in the "Q-Systems" language, originally developed for MT.

1.5.4 Philosophical Importance

Philosophically, MT is interesting, because it 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. For example, getting the correct translation of uegatively charged electrons and protons into French depends on knowing that protons are positively charged, so the interpretation cannot be something like "negatively charged electrons and negatively charged protons". In this sense, the extent to which one can automate translation is an indication of the extent to which one can automate "thinking".

Despite this, very few people, even those who are involved in producing or commissioning translations, have much idea of what is involved in MT today, either at the practical level of what it means to have and use an MT system, or at the level of what is technically feasible, and what is science fiction. In the whole of the UK there are perhaps five companies who use MT for making commercial translations on a day-to-day basis. In continental Europe, where the need for commercial translation is for historical reasons greater, the number is larger, but it still represents an extremely small proportion of the overall translation effort that is actually undertaken. In Japan, where there is an enormous need for translation of Japanese into English. MT is just beginning to become established on a commercial scale, and some familiarity with MT is becoming a standard part of the training of a professional translator.

1.6 Problems of Machine Translation

Machine translation from one language to another is problematic due to the following reasons:

- Problems of ambiguity
- Lexical and structural mismatches among the languages
- Multiwords units such as idioms and collocations

1.6.1 Problems of Ambiguity

In the best of all possible worlds every word would have one and only one meaning. But, as we all know, this is not the case. When a word has more than one meaning, it is said to be lexically ambiguous. When a phrase or sentence can have more than one structure it is said to be structurally ambiguous. Ambiguity is a well known phenomenon in any natural language. For Bangla it is also an important problem for analyzing or translating. There are several ways for which a sentence may be ambiguous. Word meaning, phrase structure, proverbs, pronouns etc. are the issues for ambiguities in Bangla [4].

1.6.1.1 Lexical Ambiguity

When a word has more than one meaning, it is said to be lexically ambiguous. Lexical ambiguities in Bangla are found due to the use of words in various contexts. The analysis of the words may reveal several meanings which are ambiguous. In Bangla the synonyms (সমার্থক শব্দ) is the classic example for which lexical ambiguity may occur [4].

1.6.1.2 Structural Ambiguity

When a phrase or sentence can have more than one structure it is said to be structurally ambiguous. It occurs when a structure of a sentence provides more than one meaning. When a structure in a sentence is interpreted in different ways, then there will occur a problem regarding the extraction of correct meaning. This types of ambiguities suffer the MT practitioner mostly in finding the appropriate meaning of the sentence [5][6].

1.6.2 Lexical and Structural Mismatches

Lexical and structural mismatches occurs due to lexical holes among the languages. Lexical holes means a phrase of a language can be expressed by a single word in another language and vice versa. This may occur due to grammatical divergences among the languages. This type of mismatches should be treated properly for successful machine translation [1].

1.6.3 Multiwords units

Multiwords units and idioms in MT is problematic due to their translational properties of being as a unit. Idioms are to be translated as a unit meaning. So that individual words in an idiom will not be translated according to their meanings [1].

1.7 Present State of Bangla Machine Translation

Machine translation regarding Bangla language is in rudimentary stage. Very few research activities have been conducted regarding machine translation of Bangla but a significant number of research activities have been conducted on the analysis of Bangla sentences. Among them [7][8] describe the parsing of Bangla sentences. Simple Bangla sentence parsing has been proposed in [7]. The detail analysis of Bangla phrases and different types of sentences are described in [8]. Analysis of complex and compound sentences has not been done yet. The design principle of automatic translation system for Bangla to other natural languages has been described in [9], whereas [10] describes the development of machine translation dictionary for Bangla language. Besides these, [11] focuses on designing a Bangla conversion processor using natural language processing and [12] describes the design and implementation of a bilingual natural language parser for Bangla to English. A significant number of researchers from various institutions of Bangladesh are doing fruitful researches regarding Bangla language.

1.8 Conclusion

Machine translation is important for those who need the gist content of a page or document but not the actual translation. To extract the meaning of a sentence of unknown language is not always possible to do by a human translator because of his/her unavailability. MT systems can solve this problem with a little bit cost though.

Organization of the rest of the thesis is as follows, Chapter 2 discusses the structure of machine translation system. A proposed MT engine and the components of it have been shown. Various steps and components of source and target systems have been dicussed briefly.

Chapter 3 concentrates on the analysis of the source language. For machine translation source language analysis consists of lexical, syntactic, semantic and morphological analyses. The implementations of the analyses have also been presented.

Chapter 4 centers morphological analysis. Morphological features of Bangla words and their analysis have been presented. Three types of morphologies in Bangla are presented with their implementations.

Chapter 5 focuses on transfer of MT system. Syntactic transfer MT has been selected and various issues regarding Bangla-English transfer have been presented with examples.

Chapter 6 deals with the generation of target language output from the target language interface structure. Generation mainly concerns syntactic and morphological generation. Syntactic generation has been presented in form of tree structure which is transferred from the source side by the transfer component. Syntactic and morphological generation have been described with examples.

The concluding chapter describes the goals achieved and emphasizes on the application potentials. It focuses on the shortcomings of the MT engine and the works to be done in future. It also presents some alternative strategies in MT transfer and a comparison among various transfer techniques.

Chapter 2

Machine Translation Engine

2.1 Translation Engine

In machine translation systems the most important non-human component which performs automatic translation is called translation engine [1][2]. MT engines can be classified by their architecture - the overall processing organization, or the abstract arrangement of its various processing modules. Traditionally, MT has been based on direct or transformer architecture engines, and this is still the architecture found in many of the more well-established commercial MT systems. The other architecture is indirect or linguistic knowledge (LK) architecture which having dominated MT research for several years. LK architecture has been adopted in the development of MT engine. So the following section discusses about LK architecture.

2.2 Linguistic Knowledge (LK) Architecture

Linguistic knowledge (LK) architecture is indirect architecture which is a promising one for commercial MT systems. The idea behind this architecture is:

"High quality MT requires linguistic knowledge of both the source and the target languages as well as the differences between them".

The term "linguistic knowledge" refers to extensive formal grammars which permit abstract/relatively deep analyses. With the LK architecture, the translation process

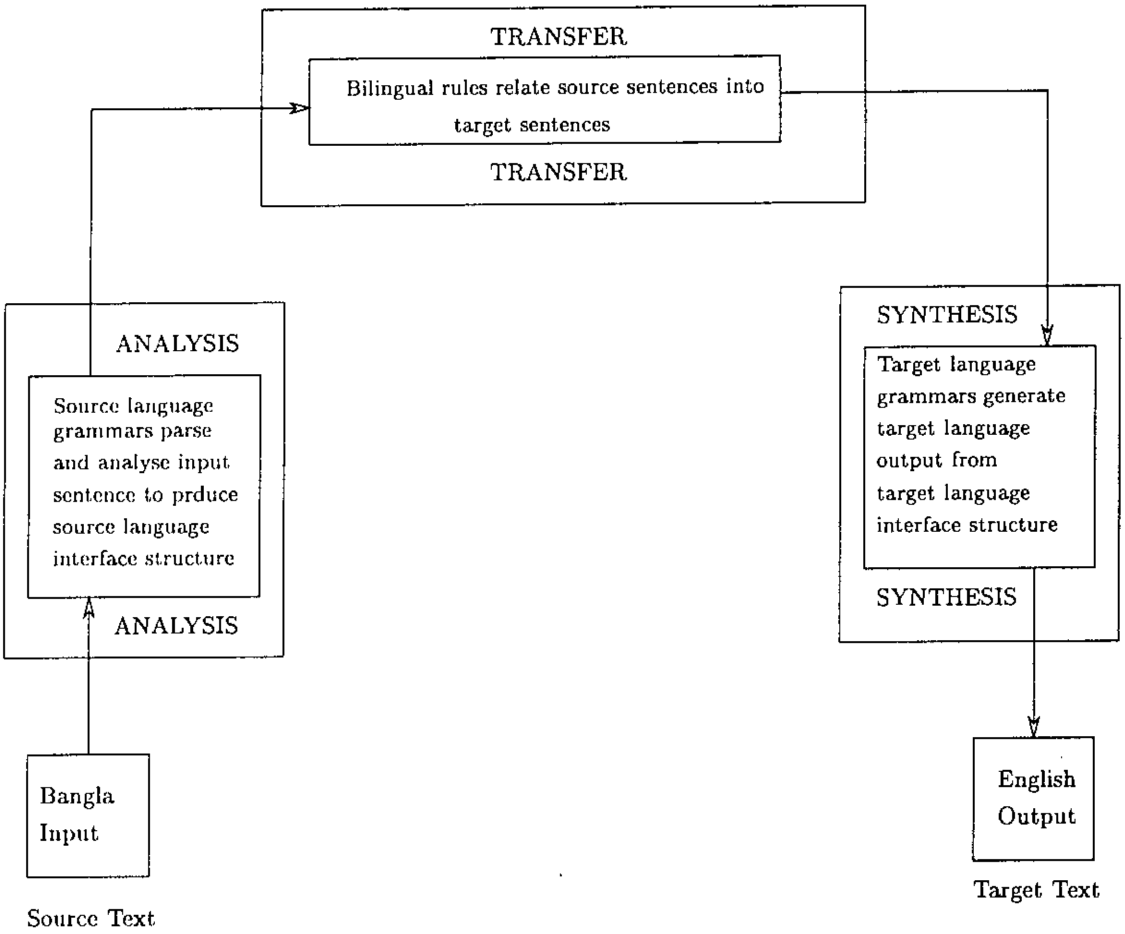

TRANSFER
Bilingual rules relate source sentences into target sentences
TRANSFER
ANALYSIS
Source language grammars parse and analyse input sentence to prduce source language interface structure
ANALYSIS
SYNTHESIS
Target language grammars generate target language output from target language interface structure
SYNTHESIS
Bangla Input
Source Text
English Output
Target Text

Figure 2.1: The components of a transfer system.

relies on extensive knowledge of both the source and the target languages and of the relationships between analyzed sentences in both languages. In short, LK architecture typically accords the target language the same status as the source language.

As can be seen from Figure 2.1, the LK architecture requires two things:

- A substantial grammar of both the source language and the target language. These grammars are used by parsers to analyze sentences in each language into representations which show their underlying structures, and by generators to produce output sentences from such representations.

- An additional comparative grammar which is used to relate every source sentence representation to some corresponding target language representation - a representation which will form the basis for generating a target language translation.

The LK engine for Bangla-English will have grammars for each Bangla and English language. One grammar for Bangla and the other for English. Each of these two grammars is independent entity. That means Bangla grammar has a set of rules for analyzing Bangla sentence and a set of English grammar rules for generating English sentences. Bangla and English grammar rules are written by the specialists of Bangla and English respectively. The important thing here is that for machine translation it is required to have the same deep representation of sentences of the languages. Otherwise structural discrepancies will occur and this would require additional transfer rules for mapping these different structures into each other.

Looking at Figure 2.1 it is clear that if the system is for Bangla to English translation, the first (analysis) step involves using a Bangla parser and the Bangla grammar to analyze the Bangla input. The second (transfer) step involves changing the underlying representation of the Bangla sentence into an underlying representation of an English sentence. The third (synthesis) step and final major step involves changing the underlying English representation into an English sentence, using a generation or synthesis component and the English grammar. The fact that a proper English grammar is being used means that the output of the system are grammatically correct.

This also means that the whole Engine should be reversible, at least in theory. Taking the Bangla-English LK engine in Figure 2.1, we could run the translation from right to left. That is, we could give it English sentences, which would then be analyzed into underlying representations. These representations would be changed into Bangla underlying representations and a Bangla translation would then be synthesized from the result. The same grammars for each language are used regardless of the direction of the translation. In practice few translation engines are reversible, since some rules that are necessary for correct translation in one direction could cause problems if the process was reversed. This is especially true for lexical transfer rules.

2.3 Proposed MT Engine

Based on LK architecture an MT engine has been developed for Bangla-English machine translation. The proposed MT engine is shown in Figure 2.2. The MT engine is

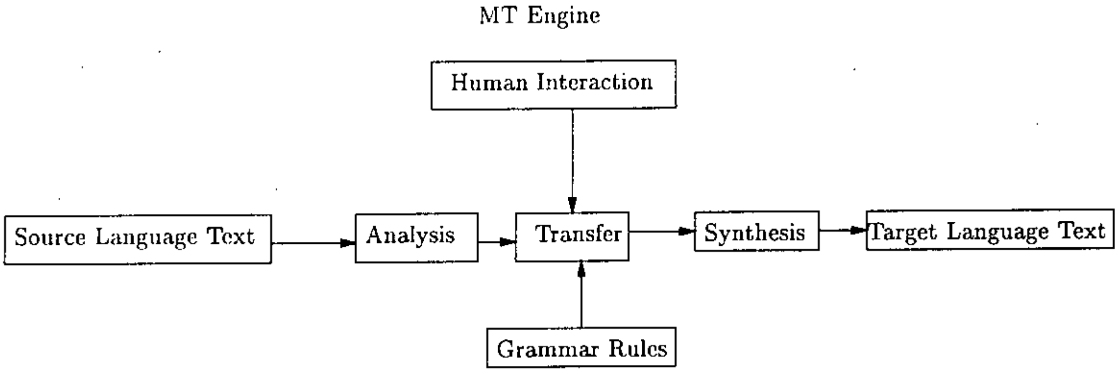


Figure 2.2: The proposed machine translation engine

capable of translating simple Bangla sentence into English and uses linguistic knowledge architecture.

The MT engine consists of,

- Source Language Analyzer,
- Transfer,
- Grammar Rules,
- Human Interaction,
- Synthesis.

2.3.1 Source Language Analysis

The analysis of source language sentence for machine translation centered on lexical analysis, syntactic and semantic analysis, morphological analysis etc. All these analyses are important due a lot of reasons. The details source language analysis will be presented in Chapter 3. The following subsections shows only the introduction of them..

2.3.1.1 Lexical Analysis

Lexical analysis is done through looking up a lexical dictionary for matching the lexicons in an input sentence. Lexical entries contain the root lexicons with their types and attributes. The lexicons along with their types and attributes are retrieved at run time for matching with the input. Based upon the values of the attributes, the morphological analyzer will add the affixes before or after the lexicon to produce the actual form of a word.

2.3.1.2 Syntactic Analysis

Syntax or syntactic analysis uses grammar rules for identifying whether an input sentence structure is correct or not. An input sentence is separated into phrases and the phrases are separated into tokens which are the constituent parts of the sentence. This is called *parsing*. Parsing may be done in top-down or bottom-up strategy. Augmented transition network (ATN) is generally is used to represent the grammars of the parsing. In top-down parsing the sentence is parsed from left to right in depth-first searching technique where each word is in the sentence is first checked whether it is in the lexicon or it is an inflected form.

2.3.1.3 Semantic Analysis

Semantic analyzer determines the semantic meaning of the words in a sentence. For semantic analysis a vocabulary is needed to be defined that will be used on the top of the structure. The way of semantic analysis is discussed in Chapter 3.

2.3.1.4 Morphological Analysis

Morphological analysis is used to analyze the morphological properties of the words. Both source and target language there must be separate morphological analyzers. Morphology is concerned with the analysis and generation of word forms. There are three types of morphologies: Inflectional, Derivational and Compounding. Morphology plays important role in determining the structures of the words in both source and target languages. Morphological properties of words are transferred through the MT engine and the target language synthesis component generates target output by morphological analyzer of that language. Morphological analysis is discussed in Chapter 4.

2.3.2 Transfer

The transfer component of MT engine is responsible for transferring source language structures into target language structures. Figure 2.2 shows transfer is accompanied with human interaction and grammar rules. The grammar rules for transfer system are called comparative grammar because the rules correspond bilingual transformations.

2.3.2.1 Comparative Grammar

The parsers in LK engines typically analyze to relatively abstract, or deep underlying representations. The parser uses the grammars of the language to produce the sort of deep syntactic representation which is called parsing. To translate the deep representation of the source language into target language the transfer component uses a comparative grammar that relates source language representation into corresponding representations for the target language sentences. Just as monolingual grammar of Bangla has a "dictionary" of rules (e.g. N → বই) so also the comparative grammar for Bangla to English has bilingual dictionary rules. In the simplest case, these may just relate source lexical items ("words") to target lexical items. For example the Bangla sentence ami boi pori → আমি বই পড়ি may be translated as word by word basis preserving the sentence meaning.

আমি ↔ I
বই ↔ book
পড়ি ↔ read

The abstract tree representation of the sentence আমি বই পড়ি is shown in Figure 2.3

2.3.3 Human Interaction

As the MT engine is not fully automated, it will need some interaction from the user during translation. To resolve ambiguities of both structural and lexical the engine will request to respond the for some queries. This is indeed necessary for correct translation output.

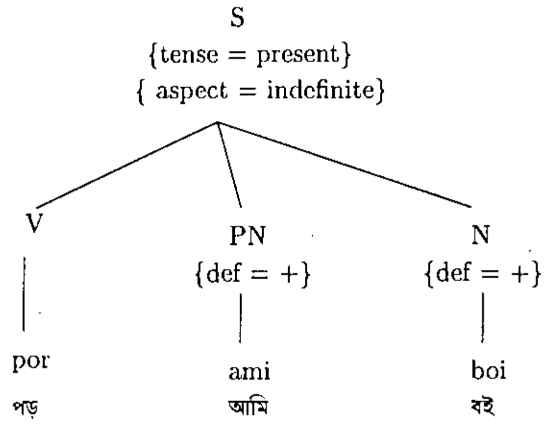


Figure 2.3: Abstract tree representation of আমি বই পড়ি.

2.3.4 Target Language Synthesis

Target language synthesis or generation starts by receiving the structures with the rules from the transfer component. The synthesis component generates target language output in two ways:

- Syntactic Generation,
- Morphological Generation.

2.3.4.1 Syntactic Generation

Syntactic generation is carried out by means of transformations. The syntactic structures transferred from source language are synthesized by target language grammars to find syntactic match. If the syntactic structures are correctly recognized by grammar rules then morphological generation is necessary for producing actual word forms of the target language.

2.3.4.2 Morphological Generation

Morphological rules of the target language are used to generate the word structures of the sentence. It is done through checking the morphological rules of the source language. The rules of the source language transferred and used by the morphological analyzer for producing the word forms of the target language.

Chapter 3

Source Language Analysis

3.1 Syntactic Analysis

Syntactic analysis concentrates on the analysis of syntactic structures of the sentences. This is the first step for MT and NLP systems. The process of determining the syntactic structure of a sentence is known as *parsing* [1][13][14]. The structure of a sentence can be represented as a syntactic tree or as a list. The parsing process is basically the inverse of the sentence generation process since it involves finding a grammatical structure from an input string. When given an input string, the lexical parts or terms must be identified by their types, and then the role they play in a sentence must be determined. These parts are combined successively into larger units until a complete tree structure has been completed. To determine the meaning of a word, a parser must have access to a lexicon [1][13]. When the parser selects a word from the input string it locates the word in the lexicon and obtains the word's possible function and other features, including semantic information. This information is then used to build a tree or other representation structures.

Syntactic analysis is essential for any type natural language processing. Because it validates the correctness of the utterance of the writer or the speaker. For further processing the output of the parser is used. For machine translation syntactic analysis performs the primary tasks of translation. The output of parsers of the MT systems are then used by the subsequent components for producing the output translation.

The general parsing process is shown in Figure 3.1.

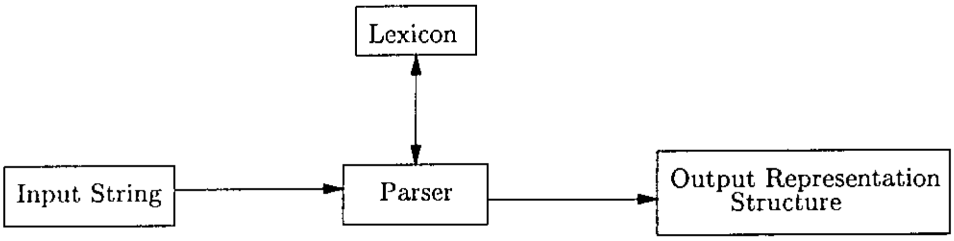


Figure 3.1: Parsing an input to create output structure.

3.1.1 Role of Lexicon

A lexicon is a dictionary of words where each word contains some syntactic, semantic and possibly some pragmatic information. The information in the lexicon are needed to help for determining of the function and meaning of the words appeared in a sentence. Each entry in a lexicon will contain a root form of the word and other information. The inflected forms or derivations are obtained by the morphological analyzer.

The organization and entries of a lexicon will vary from one implementation to another, but they are usually made up of variable length data structures such as list of records arranged in alphabetical order. The word order may also be given in terms of usage frequency so that frequently used words will appear at he beginning of the list facilitating the search.

Typical entries in a Bangla lexicon can be shown as in Table 3.1.

3.1.2 Way of Parsing

There are two ways of parsing.

- Top-Down Parsing : Begin with the start symbol and apply the grammar rules forward until the symbols at the terminals of the tree correspond to the components of the sentence being parsed.
- Bottom-Up Parsing : Begin with sentence to be parsed and apply grammar rules backward until a single tree whose terminals are the words of the sentence and whose top node is the start symbol has been produced.

The choice between these two approaches is similar between forward and backward reasoning [15] in other problem-solving tasks in Artificial Intelligence (AI).

Table 3.1: Typical Entries in Bangla Lexicon

Word	Type	Features
আমি (ami)	Pronoun	Pers1, SG, Human, Personal
ছেলে (chele)	Noun	Pers3, SG, Human, ProperNoun
তিনি (tini)	Pronoun	Pers3, SG, Human, Personal, Honorific
সে (se)	Pronoun	Pers3, SG, Human, Personal, Non-Honorific
তারা (tara)	Pronoun	Pers3, PL, Human, Personal
বই (boi)	Noun	Pers3, SG, Object, Inanimate, Physical
বাবুল (babul)	Noun	Pers3, SG, Human, Masculine, ProperNoun
একটি (ekTi)	Adjective	Specifier, Pers3, SG
এক (ek)	Adjective	Quantifier, Pers3, SG
অনেক (onek)	Adjective	Quantifier, Pers3, PL
ভাল (vhal)	Adjective	Qualitative, Positive
ধীরে (dhiire)	Adverb	Manner
খা (kha)	Verb	Transitive, Intransitive
ভাত (vhat)	Noun	Pers3, Object, Edible, Inanimate, Physical
ঢাকা (Dhaka)	Noun	Pers3, Place, Inanimate, Physical
সকাল (sakal)	Noun	Pers3, Time, Abstract
সমিতি (samiti)	Noun	Pers3, Activity, Collective

Parser uses grammars of the language for determining the syntactic structure. Syntactic analysis is done in top-down strategy from left to right search. The sentence is compared with the grammar to find the phrases and the constituents within the phrases. Syntactic analysis can be done in many ways. The first step of the analysis is the lexical lookup in which a search in the lexical dictionary is made to find the tokens in original forms (root words). After that the syntactic categories (parts of speeches) of the tokens are assigned. Morphological properties of the words are analyzed for identifying the roles they play in the sentence. Then the syntactic categories are combined to form the phrases. The responsibility of an efficient parser is to produce a representation from which the target language structures can easily be generated [14][16][17]. The parsing of simple Bangla sentences proposed in [7][8]. The parser output for two Bangla sentences are shown in Figure 3.6 and Figure 3.8 respectively.

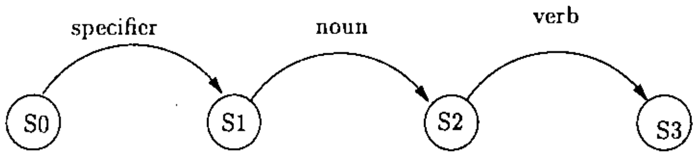


Figure 3.2: Augmented transition network (ATN).

A parsing process generally uses the following two components:

- A declarative representation of syntactic facts about the language, called *grammar*.
- A procedure called *parser*, compares the grammar against input sentence to produce parsed output.

3.1.3 Implementation of Parser

Augmented transition network (ATN) is generally used for parsing. An ATN is a top-down parsing procedure that allows various kinds of knowledge to be incorporated into the parsing system so it can be operated efficiently. The grammars are represented as ATN in which the nodes represent the states and the arcs represent the syntactic categories of the words in a sentence [15]. A path through a transition network corresponds to a permissible sequence of word types for a given grammar. Thus, if a transition network can successfully traversed, it will have recognized a permissible sentence structure. For example, if we try to recognize the Bangla sentence ekTi chele pare (একটি ছেলে পড়ে) then the transition network can be represented as shown in Figure 3.2.

Arcs may be labeled an arbitrary combination of the following:

- Specific words.
- Word categories.
- Pushes to other networks that recognize significant components of a sentence.
- Procedures that perform arbitrary tests on both the current input and on sentence components that have already been identified.

- Procedures that build structures that will form part of the final parse.

Starting at node S0, the transition from node S0 to S1 will be made if a specifier is the first input word found. If successful, state S1 is entered. The transition from S1 to S2 will be made if a noun is found next. The final transition (from S2 to S3) will be made if the last input word is a verb. If the three word category word sequence is not found, the parse fails. Clearly Figure 3.2 will only recognize the Bangla sentences of the form SPCFR N V.

The advantage of this type of representation is that if a match does not occur in a sentence, then the parser backtracks and try to find another match. The ATN is similar to a finite state machine in which the class labels that can be attached to the arcs that define transitions between states has been augmented.

To support a wide range of grammars, the utility of a transition network could be increased if more than a single choice is permitted at some of the nodes. This is possible by introducing several arcs between any pair of nodes. Thus the number of permissible sentence types will be increased substantially. Individual arc may be any parts of speech and jump to the next.

Figure 3.3 shows such a transition network for recognizing Bangla noun phrase. To move from state S0 to S1, it is necessary to find an adjective, a specifier, a noun, a pronoun or none of these by jumping directly to S1. At node S1 there are two choices and the final state should be reached only when there is a post position (অনুসর্গ). In fact, Bangla has a lot of post positions which can be plural markers such as ra (রা), gulo (গুলো), samuho (সমূহ), gon (গণ) etc. and determiners such a Ti (টি), Ta (টা), khana (খানা), khani (খানি) etc. Thus the traversal of a noun phrase will be completed successfully if the paths shown in Figure 3.3 are matched with the input.

The following form of noun phrases will be recognized by the ATN shown in Figure 3.3.

ekti chele (একটি ছেলে)
cheleti (ছেলেটি)
se (সে)
onekgulo chele (অনেকগুলো ছেলে)
ekti vhal chele (একটি ভাল ছেলে)
ami (আমি)
vhal cheleti (ভাল ছেলেটি)
mamun (মামুন)

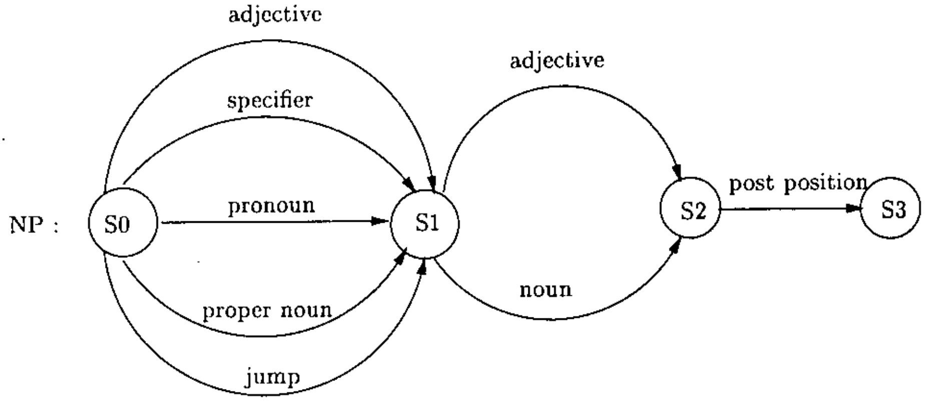


Figure 3.3: Augmented transition network for noun phrase.

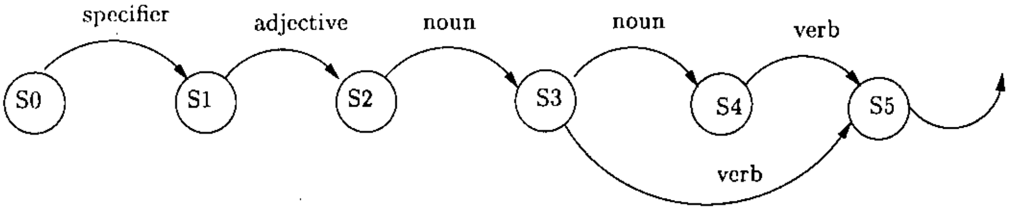


Figure 3.4: Deterministic parsing of Bangla sentence.

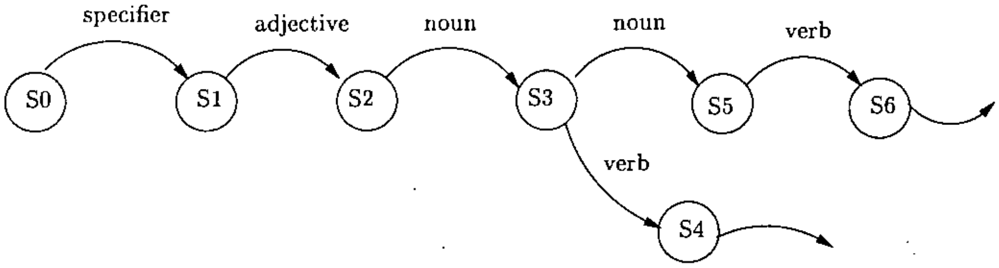


Figure 3.5: Nondeterministic parsing of Bangla sentence.

3.1.4 Deterministic Parser

A deterministic parser permits only one choice for each word category. Thus, each arc will have a different test condition. Consequently, if an incorrect test choice is accepted from some state, the parse will fail since the parser cannot backtrack to an alternative choice [13].

Deterministic parsing of Bangla a sentence is shown in Figure 3.4. The sentence in Bangla which will be analyzed by the parser of Figure 3.4 is NP+N+V or NP+V. The NP here consists of SPCFR+ADJ+N. The drawback of the deterministic parser is that it will not be able to analyze other forms of sentences which are not included in the parser.

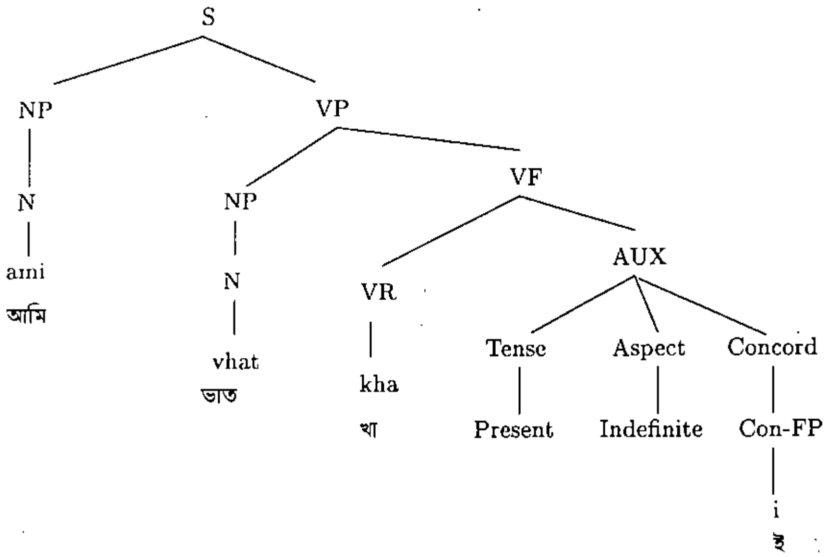


Figure 3.6: Parse tree for the Bangla sentence আমি ভাত খাই.

3.1.5 Nondeterministic Parser

Nondeterministic parser permits different arcs to be labeled with the same test. The analyzer must guess at the proper constituent and then backtrack if the guess is later proven to be wrong. This requires saving more than one potential structure during parts of the network traversal. Nondeterministic parsing of Bangla sentence is shown in Figure 3.5.

This type of parsing covers a wide range of grammars for any language. As it provides backtracking, there is no problem when a mismatch is detected. It checks another structure with the input. If it finds the match, the it then the input is said to be parsed successfully, otherwise an error message will be displayed.

3.2 Parsing Algorithm

Parsing algorithm is top-down depth-first. The input Bangla sentence is parsed from top to bottom and from left to right. The algorithm searches grammar rules for the leaf of the nodes and assigns appropriate category to the nodes. The parsing completes when all the tokens are assigned to their categories. The overall parsing algorithm is shown below:

Scanning

1. The input string is separated into tokens.
2. Tokens are stored in a list for further access.
3. The tokens are checked in the lexical dictionary for matches.

Parsing

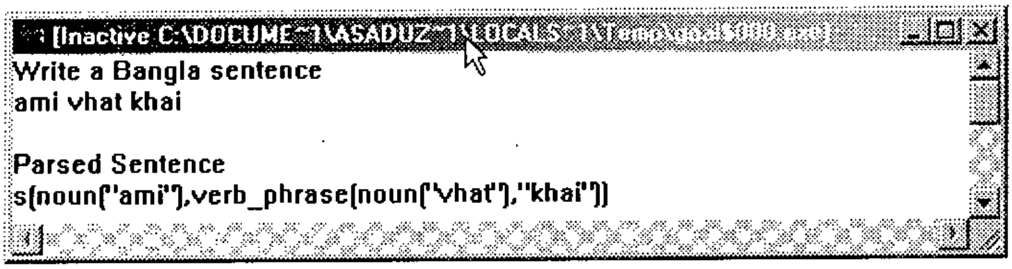
1. The scanned tokens are matched with grammar rules of Bangla. If a rule whose right hand side matches with a token, then the token is assigned with the appropriate category (parts of speech). This step is exactly equivalent to looking up the words in a Bangla dictionary. Given rules of the type PN → ami, N → vhat, and V → kha, this will produce a partial structure.
2. Starting from the left to right hand side of the token list, find every rule whose right-hand side will match one or more of the parts of speech
3. Keep on doing step 2, matching larger and larger bits of phrase structure until no more rules can be applied.

Identifying Grammars

1. To identify the grammar of a sentence, individual attributes of words are retrieved here and checked with grammar rules of Bangla. If they match then the attributes of the words are combined to form the grammatical features of the sentence.
2. The grammatical features are then transformed into logical form to store in the database for transfer so that the features can be used to generate the target sentence.

Let us consider the following example:

```
Input :  ami vhat khai (আমি ভাত খাই)
Scanner Output :TOKL = ("ami", "bhat", "khai")
Parser Output :sen(np("ami"), vp(np("bhat"), "khai")))
("ami", "Pers1", "Sing", "Nom", "Pronoun", "Personal"),
("bhat", "Pers3", "Sing","Objective", "ProperNoun"),
("khai","kha", "Pres", "Ind", "FP")
Grammar :  ("Pres", "Ind", "kha", "ami", "bhat")
```



```
Write a Bangla sentence
ami vhat khai

Parsed Sentence
s(noun("ami"),verb_phrase(noun("vhat"),"khai"))
```

Figure 3.7: Parsing of Bangla sentence আমি ভাত খাই.

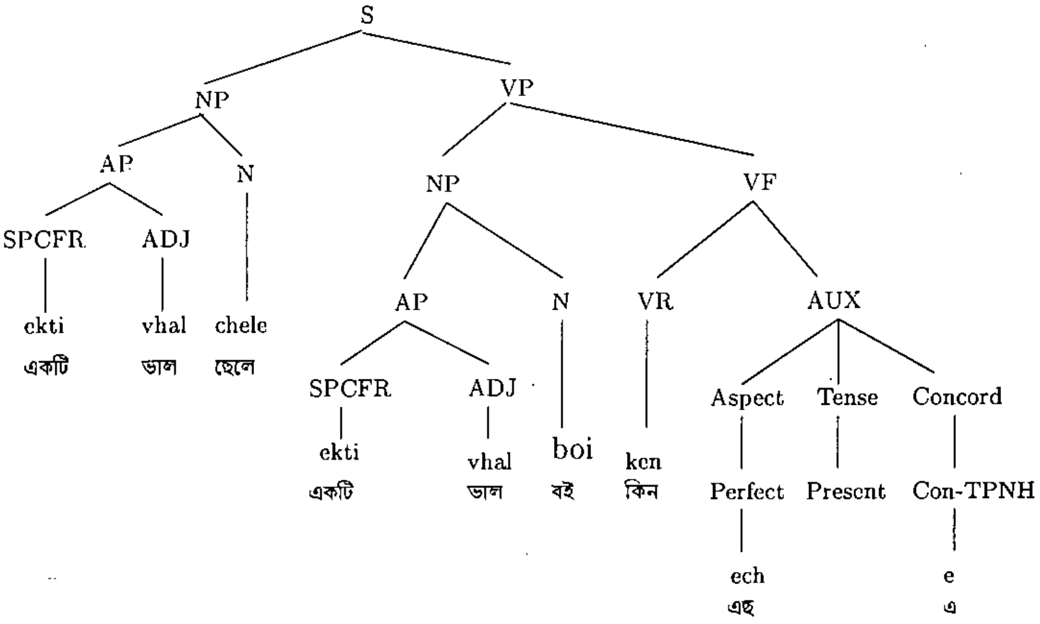


Figure 3.8: Parse Tree for the Bangla sentence একটি ভাল ছেলে একটি ভাল বই কিনেছে.

3.3 Semantic Analysis

Semantic analysis is needed for the resolution of ambiguities. For unambiguous analysis or generation semantic properties of words must be taken into account for actual output generation. Semantic analysis is done by maintaining semantic attributes of words with their categories in the dictionary. When the parser tries to parse a sentence, then the words will be retrieved with their semantic attributes which determine the meaning of the words in a sentence. The semantic attributes are necessary for transfer and generation phases indeed. Semantic attributes must be transferred into target language so that the generation of target output will be semantically correct. The words and their semantic attributes of different categories are maintained in a dictionary and during parse time they are retrieved. The semantic properties of words may include their types, number, gender, person, etc. [18]. Typical semantic properties of words are shown in Table 3.1. To ease semantic analysis, each entry in the lexicon should include the following things:

A. For all catogories:

Word

Category (Noun, Pronoun, Verb, Adjective etc.)

Sense (Activity, Place, Time, Personal etc.)

B. For a specific category:

a. Nouns

Semantic tags (Nominal, Objective, Place, Time, Activity etc.)

Number (Singular, Plural)

Gender (Masculine, Feminine)

b. Verbs

Types (Transitive, Intransitive, Auxiliary, Tense, Aspect, Mode)

c. Adjective

Types (Positive, Comparative, Superlative, Qualitative, Quantitative etc.)

d. Adverbs

Types (Manner, Place, Time, Frequency, Relative)

The information content in the lexical dictionary is taken into account during the determination of semantic meaning i.e. during parsing. When bilingual dictionary is used for transfer, the lexical dictionary is used once again for the lexicons with their semantic properties.

3.4 Morphological Analysis

Morphological analysis concerns the analysis of the morphological properties of words. Morphological analysis produces inflected or derived forms of words based upon the properties of the words in a sentence. Details of morphological analysis have been discussed in Chapter 4.

Chapter 4

Morphological Analysis

4.1 What is Morphology?

Analysis and generation [1][2][15] of word forms is a crucial and basic tool in the processing of natural languages. Morphological analysis is centered on analysis and generation of word form. It deals with the internal structure of words and how words can be formed. Morphology plays important role in applications such as spell checking, electronic dictionary interfacing and information retrieving system where it is important that words that are only inflectional variants of each other are identified and treated similarly. Natural language processing (NLP) and machine translation (MT) system need to identify words in texts in order to determine their syntactic and semantic properties [2]

Morphology is the study of the structure and formation of words. Its most important unit is the morpheme, which is defined as the "minimal unit of meaning". Linguistic textbooks usually define it slight differently as "the minimal unit of grammatical analysis". Let us consider a Bangla word "অসুস্থতা" in Figure 4.1.

There are three morphemes, each carrying a certain amount of meaning. অ means "not" or না, while তা means "being in a state or condition". সুস্থ is a free morpheme because it can appear on its own (as a "word" in its own right). Bound morphemes have to be attached to a free morpheme, and so cannot be words in their own right.

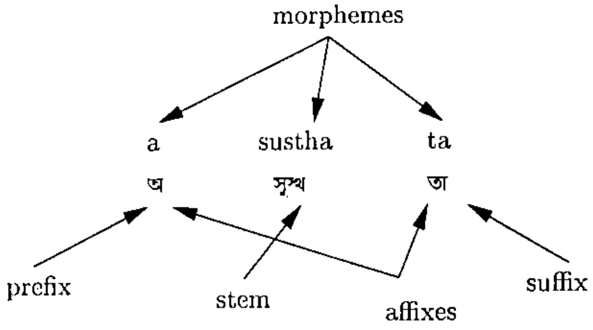


Figure 4.1: Morphological structure of the Bangla word ”অসুস্থতা”.

Armed with these definitions, we can look at ways used to classify languages according to their morphological structures [3].

To build a syntactic representation of the input sentence, a parser must map each word in the text to some canonical representations and recognize its morphological properties. The morphological analysis can be done in two levels. In two-level morphological analysis, the combination of a surface form and its analysis as a canonical form and inflection is called a lemma.

4.2 Morphological Analysis

Morphological analysis of both source and target languages is important due to producing actual underlying representations. Morphological properties of the source language are transferred to target system and the morphological analyzer of that side is responsible for producing the actual word forms in the output sentence. In our MT system we use a morphological analyzer which incorporates the rules by which the words are analyzed. The result of morphological analysis then is a representation that consists of both the information provided by the dictionary and the information contributed by the affixes. Morphological information of words are stored together with syntactic and semantic information of the words and therefore be available to subsequent levels of processing [19]. The overall structure of the morphological analyzer or processor can be shown in Figure 4.2

There are few problems regarding analysis of morphological properties. The main problems are:

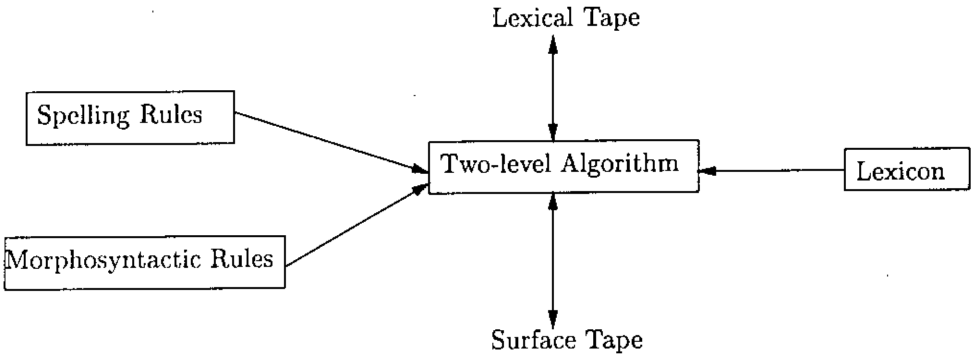


Figure 4.2: Two-level morphological processor.

- Morphological alternations: the same morpheme may be realized in different ways depending on the context.
- Morphtactics: stems, affixes, and parts of compounds do not combine freely, a morphological analyzer needs to know what arrangements are valid.
- Feature-combination: specification of how these morphemes can be grouped and how their morphosyntactic features can be combined.

4.2.1 Finte State Automata for Morphological Analyzer

Finite state machines are used successfully for analyzing rule-based MT systems. So finite state machines can easily implement the rule-based morphological analyzers. Finite state transition network (FSTN) is designed to recognize whether some input is a member of a language or not [15][20]. For practical NLP we usually require more information such as the syntactic category (often know as parts-of-speech POS) and perhaps other information such as tense, number, grammatical person, etc. In finite state automata a transducer is a piece of software that maps one stream of symbols onto another stream of symbols. We can simply say that the program transduces one stream of symbols into another.

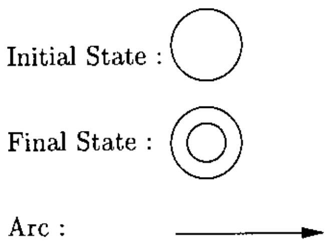
An analysis of an inflected word form is produced by mapping the input form to a sequence of lexical form through the transducers and by composing some output from the annotations on the leaf nodes of the lexical paths that were traversed [13].

Finite state morphological analysis or parsing means if we parse the English word books will produce [book+N+PL] as parsed output.

For finite state morphological analysis we need the following things:

- Lexicon,
- Morphotactics,
- Orthographic rules (Spelling rules).

Finite state transition networks are represented using the following components:



To recognize or analyze a large corpus of words the symbols are put together to form network, which is known as Finite State Transition Network (FSTN).

In the last 10-15 years computational morphology [21] has advanced further towards real-life applications than most other subfields of natural language processing. The quest for an efficient method for the analysis and generation of word-forms is no longer an academic research topic, although morphological analyzers still remain to be written for all but the commercially most important languages.

4.3 Morphology of Bangla Words

Bangla is not morphologically very complex. There are three different types of morphologies are recognized in Bangla. Unlike German, Turkish or Spanish Bangla avoids production of very long words by recursive use of morphological rules [9]. The Bangla morphologies are:

- Inflectional Morphology.
- Derivational Morphology.
- Compounding.

4.3.1 Inflectional Morphology

Inflectional morphology [1][2][15] produces or derives words from another word form acquiring certain grammatical features but maintaining the same part of speech or category. Bangla has a very strong and structural inflectional morphology for its verb forms for different tenses and persons. Using these suffixations, tense and person of a sentence can be detected almost unambiguously. There are a number of inflectional suffixes denoting number of the nouns and pronouns of a sentence, but fortunately; only few of them are in common uses. There are well-formed suffixes to represent cases (কারক) of the nouns. These suffixes perform the job of English prepositions. Notations for identifying genders are also available but these are rather intractable for morphological analysis because of their irregularities in occurrences [9].

4.3.2 Morphology of Verb Phrases

Every sentence in Bangla must have a verb phrase. The verb phrase consists of a noun phrase and a verb. The verb consists of a compulsory verb part, which is called verb root (কৃৎ প্রত্যয়). The noun phrase may contain an NP or an NP and an AP. The detailed analysis of the forms of verb auxiliary can be found in [7][8]. The morphological properties of the verb forms are discussed here with implementation. Let us consider the Bangla sentence `se vhal kaj karche` (সে ভাল কাজ করছে). Here the verb in the Bangla sentence is `karche` (করছে). The morphological analysis of the verb involves the following issues:

Firstly, the morphological component will recognize `karche` (করছে) is an inflected form of the verb root `kar` (কর).

Secondly, we would like to retain the information carried out by the affix so that it will be easy to generate output sentence. The verb here `karche` means that the verb is finite, or tensed (present continuous). This is important since it allows the verb to occur as the only verb of a main clause.

Thirdly, the information that we gather from the inflection is the fact that the verb is third person (as opposed to first person, occurring with ami (আমি) or amra (আমরা), and as opposed with second person, occurring with `tumi` (তুমি) or `tui` (তুই) or `tora` (তোরা), and that it is singular (rather than third person plural, which occurs with tara, or with

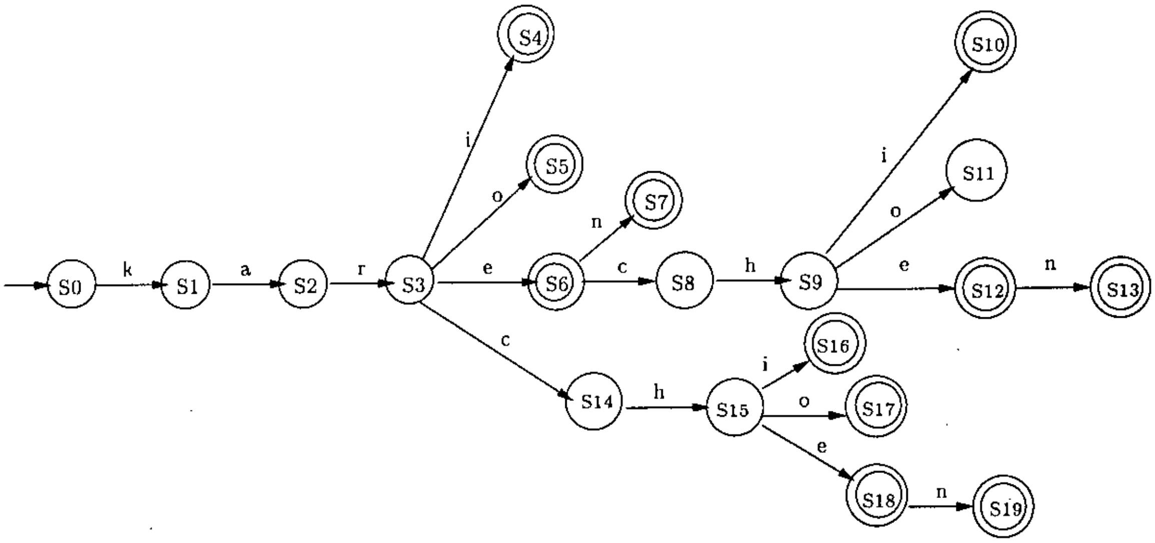


Figure 4.3: Finite state transition network (FSTN) for analyzing Bangla Verbs consisting the root কর in present tense only.

a plural noun).

Finally, here the person is non-honorific third person se (সে) rather that honorific third person tini (তিনি). However in Bangla there is no difference between se, or tara as well as tini or tara. In both cases the affixes for present continuous tense will be che and chen respectively.

There are various ways of describing this. One of the simplest forms is to use rules of the following form: [lex = V, cat = v, +finite, person = 3rd, number = sing, tense = pres, aspect = cont] ↔ [V + che]

Here we have introduced a rule which says that finite verbs which are third person singular and have present tense [cat=V, +finite, person=3rd, number=sing, tense = pres, aspect=cont] can be formed by adding che to the base form (the base form is represented as the value of the attribute lex). The rule can also be read in the opposite direction: if a word can be divided into a string of characters that matches with the verb root and che, then it may be a finite verb with third person singnlar for present continuous tense. Whether something is indeed the base form of the verb can be verified in the monolingual dictionary. So, if the morphological analyzer encounters a word like karche(করছে), it will check whether the monolingual dictionary contains an entry with the features cat = v, lex = kar. Since it does, karche can be represented by

Table 4.1: Words recognized by the FSTN show in Figure 4.3

Bangla Verbs	Transducer Path
kari(করি)	$kar + V + Pres + Indf + 1P$
karo (কর)	$kar + V + Pres + Indf + 2P$
kare (করে)	$kar + V + Pres + Indf + 3P + Non - Honrific$
karen (করেন)	$kar + V + Pres + Indf + 3P + Honrific$
karchi (করছি	$kar + V + Pres + Cont + 1P$
karcho (করছ)	$kar + V + Pres + Cont + 2P$
karche (করছে)	$kar + V + Pres + Cont + 3P$
karchen (করছেন)	$kar + V + Pres + Cont + 3P$
karechi (করেছি)	$kar + V + Pres + Perf + 1P$
karecho (করেছ)	$kar + V + Pres + Perf + 2P$
kareche (করেছে)	$kar + V + Pres + Perf + 3P + Non - honorific$
karechen (করেছেন)	$kar + V + Pres + Perf + 3P + honorific$

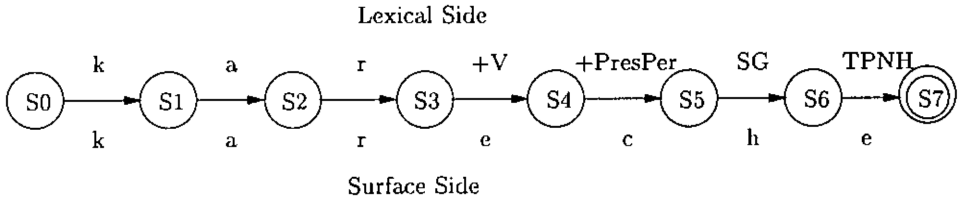


Figure 4.4: Lexical and surface structures of the Bangla verb করেছে.

means of the lexical entry, with some of the information supplied by the rule. The result of morphological analysis then is a representation that consists of both the information provided by the dictionary and the information contributed by the affix.

Each and every path of Figure 4.3 will constitute a lexical path by which the verb form will be recognized. The lexical path will show the real analysis of the verb form. Morphological analyzer is necessary for two reasons :

- To produce the actual form with given root and other morphological properties.
- To analyze an inflected form of verb in order to find its morphological properties.

The first one is obviously opposite to second one. But both of them are equally important for natural language analysis and synthesis. Lexical transducer shown in Figure 4.4 is the best way to show the morphological analysis of a verb form. Let us

Table 4.2: Morphological Features of Some Bangla Verbs

Bangla Root	Morphological Features	Inflected Words
kha	$+V + 1SG + Pres + Ind$	*khai*
kha	$+V + 3SG + Pres + Perf + Non - Honorific$	*khaehce*
kha	$+V + 3SG + Pres + Perf + Honorific$	*khaechen*
kha	$+V + 3SG + Pres + Cont + Honorific$	*khachen*
por	$+V + 2SG + Past + Ind + Pejorative$	*porli*
por	$+V + 2SG + Past + Ind + non - honorific$	*porle*
por	$+V + 2SG + Past + Ind + honorific$	*porlen*
ja	$+V + 2PL + Future + Ind + non - honorific$	*jabe*
ja	$+V + 3PL + Future + Ind + non - honorific$	*jabe*

consider another Bangla verb kareche (করেছে). The lexical side of the transducer shows the root form of the verb and the morphological properties of the verb in a sentence. If this is feed into lexical side, the surface side will produce the output verb form based on the attributes. This way of producing surface structures from lexical forms is important for generation phase. And the opposite direction i.e. to find morphological properties of verbs is essential for source language analysis.

The way presented here for analyzing morphology can be done for all verb forms in Bangla. Unlike English, Bangla has no irregular form of verbs. All the verb forms in Bangla are easy to analyze to find their morphological properties. That means the verb forms have a uniform relation with tense, aspect, mode, person, number, gender etc. The FSTN shown in Figure 4.3 is shown only for present tense. In the same way other tenses can also be analyzed. Table 4.2 shows some verbs of Bangla with their morphological features.

4.3.3 Morphology of Noun Phrases

Bangla noun phrases have different categories of nouns with suffixes, quantifiers, preposition, adjectives etc. Bangla nouns may be concrete or abstract. Concrete nouns can be classified as proper noun (পলাশ, ঘর), common noun (শহর, পাখি), material noun (তামা, পানি), and collective noun (দল, সমিতি). There are also two classes of abstract nouns:

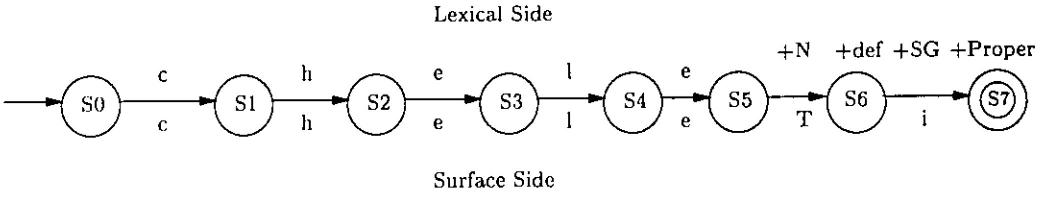


Figure 4.5: Lexical and surface structures of the Bangla noun ছেলেটি.

proper abstract noun (ক্ষমা, উদারতা), and verbal noun (হাটা, পড়া). Gender and case are also important for identifying proper categories of nouns. Number may be singular (বই) or plural (বইগুলো) Gender of nouns can be masculine (ভাই), feminine (বোন), common (শিশু), and neuter (কলম) [22][6]. Case of a noun may be nominative (ছেলে), accusative (ছেলেকে), genitive (ছেলের), and locative (ঘরে) [5][10].

For analyzing nouns in Bangla a lot issues come in front. These are quantifiers such as ek (এক), dui (দুই), bohu (বহু), onek (অনেক), determiners (নির্দেশক) such as ti (টি), ta (টা), khana (খানা), khani (খানি), bivakti (বিভক্তি) such as jon (জন), ke (কে), der (দের), digake (দিগকে), and plural markers ra (রা), era (এরা), gulo (গুলো), gon (গণ), borga (বর্গ), samuho (সমূহ) etc. A Bangla inflected noun form can be obtained by adding the quantifiers with or without preposition before the noun and the plural marker after the noun. For example ekjon lok (একজন লোক), bohu lok (বহু লোক), cheleti (ছেলেটি), chelera (ছেলেরা), boigulo (বইগুলো) etc. Besides these, Bangla has a significant number of prepositions (অনুসর্গ) and post position (উপসর্গ) which can be added before and after of a word to generate new words [23]. All these forms of nouns are needed to be analyzed appropriately for correct translation.

For example the Bangla noun cheleti (ছেলেটি) consists of root word chele (ছেলে) and the Bangla determiner (নির্দেশক) ti (টি) that can be analyzed as $chele + N + Def + Proper + SG + Nom$. This rule will be helpful in transfer of the noun so that actual sense should be unaltered in the target language. In the lexical side the entry $chele + N + Def + Proper + SG + Nom$ means the surface side of the transducer will produce cheleti as output. In the opposite way if we would like to analyze cheleti the morphological analyzer will produce $chele + N + Def + Proper + SG + Nom$w as output.

Let us consider another Bangla noun boigulo (বইগুলো), which can be analyzed as: $boi + N + Def + PL$. Bangla nouns are of different forms. Each and every category of

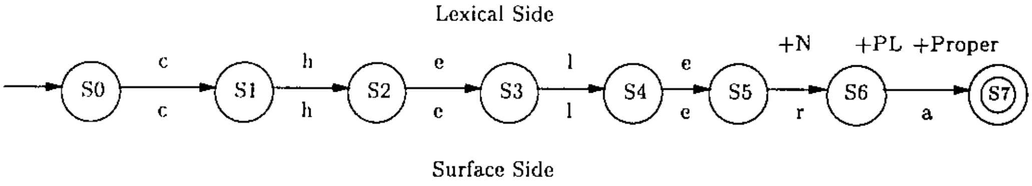


Figure 4.6: Lexical and surface structures of the Bangla noun ছেলেরা.

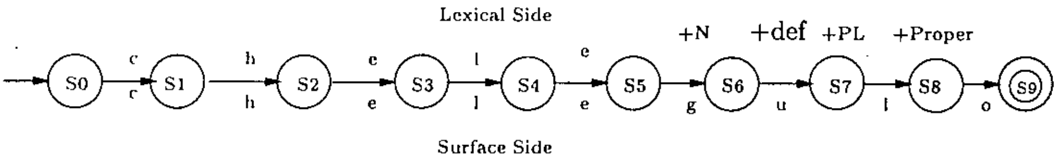


Figure 4.7: Lexical and surface Structures of the Bangla noun ছেলেগুলো.

noun is needed to be analyzed to produce the exact target language meaning.

Table 4.3 shows some Bangla nouns with their morphological features.

4.3.4 Derivational Morphology

Derivational morphology is simple and a word rarely uses the derivational rule in more than two or three steps. The first step forms nouns or adjectives from verb roots (কৃৎ প্রত্যয়). The next steps form new nouns and adjectives. Unlike English, where derived long verbs are very common, long verbs are rarely derived in Bangla. Instead, a derived noun describing the act of the verb is appended with a "do" verb. For example, English forms the "Industrialize" from the nominal "Industry", and this verb can receive any of the inflectional suffixes. But, in Bangla, a noun "শিল্পায়ন" is appended with the do-verb (কর) to represent the same sense. As a result the derived Bangla verb will be "শিল্পায়ন কর". Our linguists categorize them as compound verbs [9]. The morphological analysis of such kind of words may be represented as N + V.

4.3.5 Compounding

There are various types of compound nouns which form compounding in Bangla. Semantics of the compound words may correspond to any or both of the constituents, or none of them. Compound words are difficult for morphological analysis but the rules are needed to generate fluent Bangla from Interlingua. Some forms of compounding are discussed below:

Table 4.3: Morphological Features of Some Bangla Nouns

Bangla Root	Morphological Features	Inflected Words
chele	$+N+3SG+Proper+Nom$	*chele*
chele	$+N+3PL+Proper+Nom$	*chelera*
chele	$+N+3SG+Def+Proper+Nom$	*cheleti*
chele	$+N+3PL+Def+Proper+Nom$	*chelegulo*
chele	$+N+3SG+Def+Proper+Objective$	*cheletike*
chele	$+N+3PL+Def+Proper+Objective$	*cheleguloke*
boi	$+N+3SG+Def+Proper+Objective$	*boiti*
boi	$+N+3PL+Def+Proper+Objective$	*boigulo*

- Noun + noun : This type of compound noun form by concatenating two simple nouns and the meaning is extracted as a whole. For example শিক্ষক সমিতি, বাড়ী ঘর, নদী তীর. This form of compounding can be analyzed as N + N + morphological features. There may be other variations. One form of compound nouns where the second noun is the part of first noun and the first one appears with Bangla bivakti (বিভক্তি). For example কলেজের গেট, বইয়ের দোকান. The morphological rule for this may be N+Biv+N+morphological features. In other form the first noun indicates the place and time of the second one. For example রাস্তার দোকান, বগুড়ার দই, গরমের ছুটি. This can also be analyzed as N+Biv+N+morphological features.

- Noun + verb : A noun and a verb constitutes this types of words in Bangla and the whole words provides a single meaning. For example ফুল তোলা, গাড়ী চালানো, পাখি শিকার. The morphological rule in this case may be N+V+morphological features.

4.4 Conclusion

Within the context of MT, it is clearly desirable to have a similar approach, where monolingual and transfer dictionaries only contain the head words and no inflected words [1]. In order to achieve this a system must be capable of capturing the regular patterns of inflections. This can be done by adding a morphological component to the system, which describes all the regular inflections in general rules, with additional explicit rules for irregular inflection, thus allowing dictionary writers to abstract away from inflected forms as much as possible. The morphological component will be able to map inflected words onto the appropriate head words and will retain the information provided by the inflectional affix by adding the relevant features.

Morphological analysis of Bangla words is an interesting research topic for the dictionary builders for MT systems. It requires extensive efforts to build a complete morphological analyzer for a language. In the thesis work, we use a morphological analyzer for very limited number of words that are required for our translation. But interested researchers can do research for complete morphological analyzer for Bangla words.

Chapter 5

Transfer of MT

5.1 Transfer Machine Translation

In transfer machine translation the source language (SL) is analyzed into an SL-dependent representation which is then transferred into a target language (TL)-dependent representation from which a TL sentence is generated via some generation procedure. At a minimum, transfer systems require monolingual modules to analyze and generate sentences, and transfer modules to relate translationally equivalent representations of those sentences [2]. Figure 5.1 shows the minimum components of a multilingual transfer system for three languages.

Transfer systems are generally regarded as a practical compromise between the efficient use of resources of interlingua systems, and the ease of implementation of direct systems. However, it is clear that for a general multilingual system the number of transfer modules grows polynomially in the number of languages. That is, for n languages, we need at least [n(n-1)/2] transfer modules. This is because for each n languages, there are (n-1) possible TL in a fully multilingual system. If these modules are reversible, then only half of these number of transfer modules are required. This is an obvious disadvantage of transfer systems, since they become more expensive the more languages they have.

Many factors make transfer system of MT an attractive issue [2]. These are

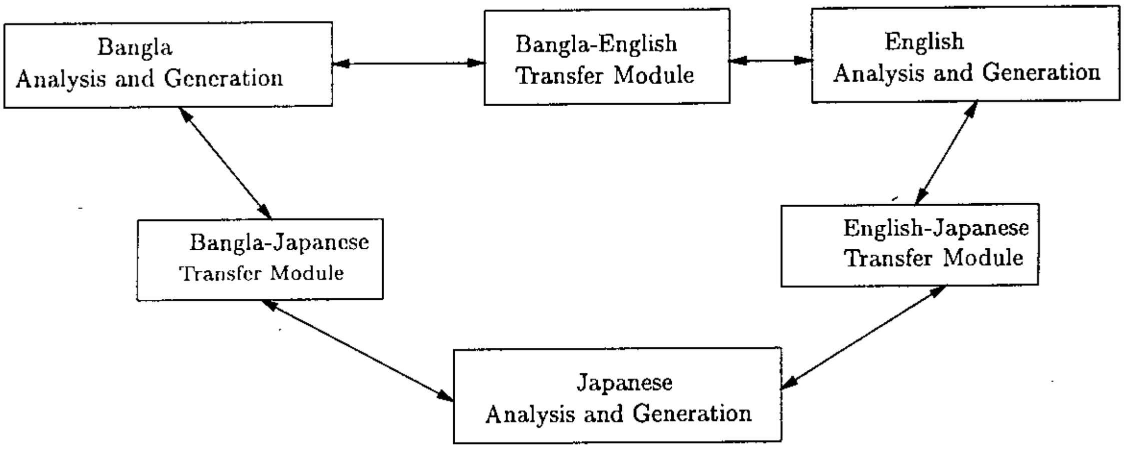


Figure 5.1: Minimal transfer architecture for three languages.

- Many systems are bilingual, or their principal use for translation in one direction between a limited number of languages.

- Where full multilinguality is required it is possible to have a hub language into and out of which translation is done.

- Portions of transfer modules can be shared when closely related languages are involved.

5.2 Role of Bilingual Dictionary

For source to target language transfer bilingual dictionary of the two language plays the key role in source to target translation. Dictionaries are crucial parts for MT systems. Dictionaries are the largest components of an MT system in terms of the amount of information they hold. The size and quality of dictionary limits the scope and coverage of a system, and the quality of translation that can be expected. That is why the dictionaries where the end user can and expected to be able to contribute most to a system [1]. The contents of the dictionaries must be maintained efficiently and be readily extendable as and when required by the end user [24]. Several issues for creating machine translation dictionary have been mentioned in [10]. For our MT system we use a bilingual dictionary of limited vocabulary for our purpose. The dictionary here we use contains Bangla to English translation of words and other grammatical information that are necessary for correct translation.

MT systems may of three kinds. These are:

a) Syntactic transfer MT,

b) Semantic transfer MT,

c) Lexicalist MT.

Syntactic transfer MT has been adopted into our MT system. The following sections describe syntactic transfer MT with respect to Bangla-English translation.

5.3 Syntactic Transfer MT

The syntactic structures are transferred in this technique. The phrase to phrase transfer is done with their grammatical features [25].

Syntactic transfer systems rely on mappings between the surface structures of sentences: a collection of tree-to-tree transformations is applied recursively to the analysis tree of the SL sentence in order to construct a TL analysis tree. Figure 5.2 illustrates a simple tree to tree transfer module to translate the Bangla noun phrase একটি ভাল ছেলে (ekTi vhal chele) into its English translation a good boy. The transformations include translation variables, indicated by tv, that relate translationally equivalent portions of the source and target structures.

Given that transformations result in complete parse trees in the target language, the notion of generation as understood here is not applicable in this approach. In fact, it is possible to build syntactic transfer systems where the only generation undertaken is morphological generation. However, some syntactic transfer systems include tree-to-tree transformations during generation. These transformations deal with syntactic and semantic discrepancies not accounted for during transfer. Their purpose is to simplify transfer by allowing it to produce incorrect TL. This aids multilinguality, since it shifts the workload from the bilingual component to the monolingual one.

The tree-to-tree transformation algorithm is a recursive, non-deterministic, top-down process in which one side of the tree-to-tree transfer rules are matched against the input structure, resulting in the structure on the right-hand side. The transformation algorithm is then called recursively on the value of the transfer variables to yield the corresponding TL structure.

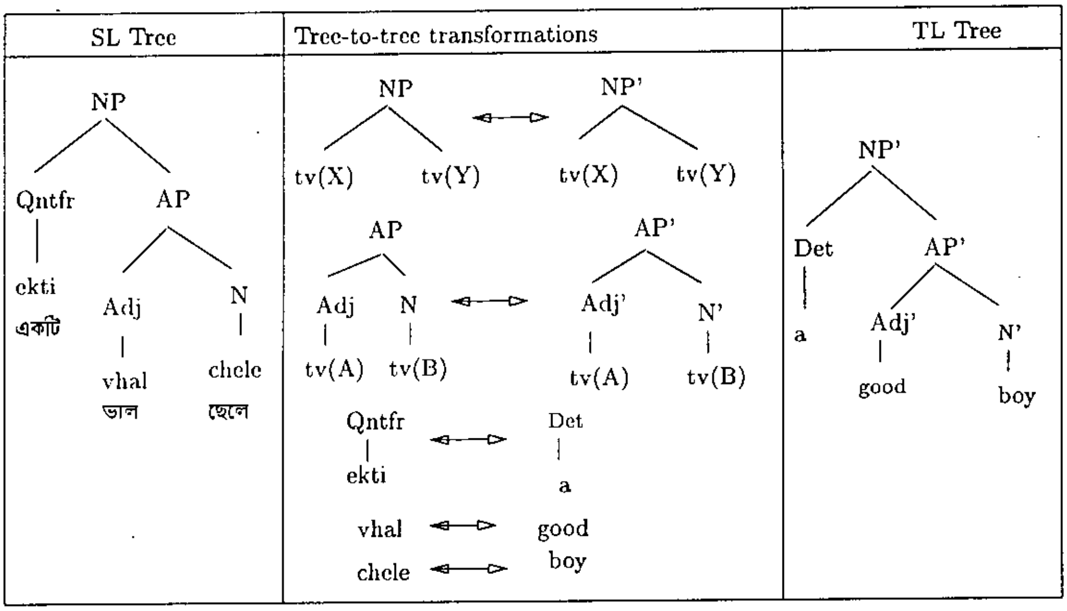


Figure 5.2: Syntactic transfer of adjective phrase from Bangla to English.

5.4 Transfer Test Cases for Bangla-English

Most problems of MT concern disambiguation, anaphora resolution, robustness etc. A number of such problems have been identified during the transfer. These problems are peculiar to translation because the divergences and structural mismatches between source and target sentences. A translation divergence normally implies that the meaning is conveyed by the translation, although the syntactic structure and semantic distribution of meaning components is different in two languages and translation mismatch implies a difference in information content between the source and target sentences [2]. The divergences experienced during the transfer so far are,

- Thematic,
- Head Switching,
- Structural,
- Lexical Gap,
- Collocational,
- Multi-lexeme and Idiomatic.

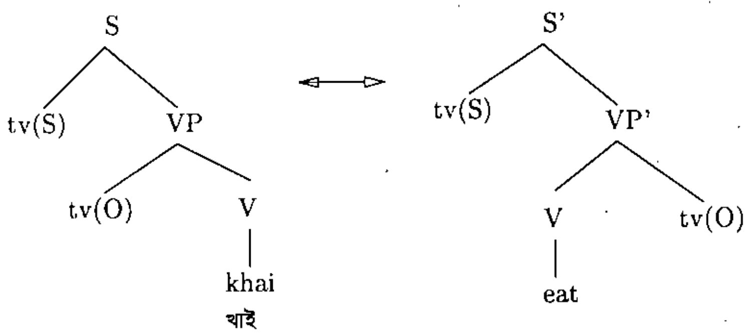


Figure 5.3: Thematic divergence in Bangla-English transfer.

5.4.1 Thematic

Thematic divergences relate to changes in the grammatical role played by arguments in a predicate. The example shown Figure 5.3 is the thematic divergence between Bangla and English.

Bangla : ami vhat khai (আমি ভাত খাই)
English : I eat rice

The grammatical object in Bangla appears before the verb but in English the object takes place after the verb. A mapping between the syntactic structures of these types of sentences needs access to both the subject and object within the same transformation and place them in different positions within the target structure.

Such transformations can potentially involve a clash in the agreement properties of the verb, as well as discrepancies in the case marking and clitic structure of the target sentence.

5.4.2 Head Switching

In head switching the syntactic head of the source language is not translated as a syntactic head, but as a modifier, a complement, an auxiliary or some other constituents.

Bangla : cheleTi vhat khaeche (ছেলেটি ভাত খেয়েছে)
English: The boy has eaten rice

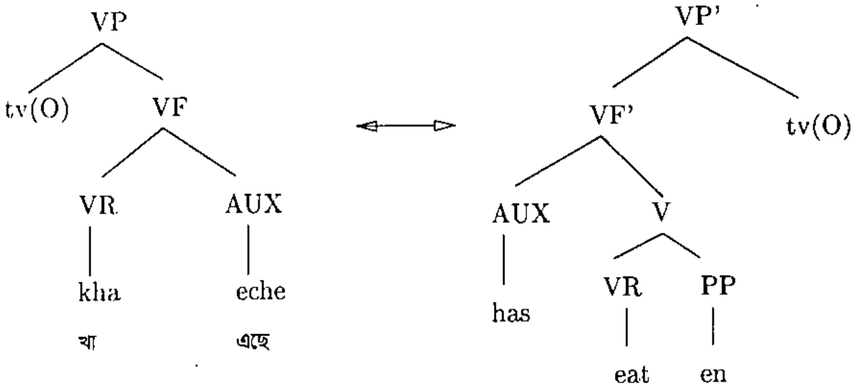


Figure 5.4: Head switching in Bangla-English transfer.

Here khaeche (খেয়েছে) is the syntactic head in Bangla. Its translation in English is not as syntactic head but with an auxiliary before the verb and it is in participle (eaten) form instead of its base form (eat). In Bangla the affix (এছে) after the verb root kha (খা) indicates present perfect tense with with third person non-honorific, singular number as subject. This grammatical information transferred and the target English translation shows the correct contrast with source by auxiliary has and a participle before and after the verb root eat.

5.4.3 Structural

Structural divergence occurs due to structural discrepancies between source and target language. In Bangla VP consists of NP with complements and the verb whereas in English the VP consists of the Verb and the NP. This structural discrepancies can be illustrated by the following example:

Bangla : lokTi chairTite baslo (লোকটি চেয়ারটিতে বসল)

English : The man sat on the chair

Figure 5.5 shows the structural divergence between Bangla and English. The Bangla word "chairTi" (চেয়ারটি) will be translated as "the chair" in English. The preposition "on" will be added before "the chair" because of Bangla bivakti "te" (তে) after the word "chairTi" (চেয়ারটি).

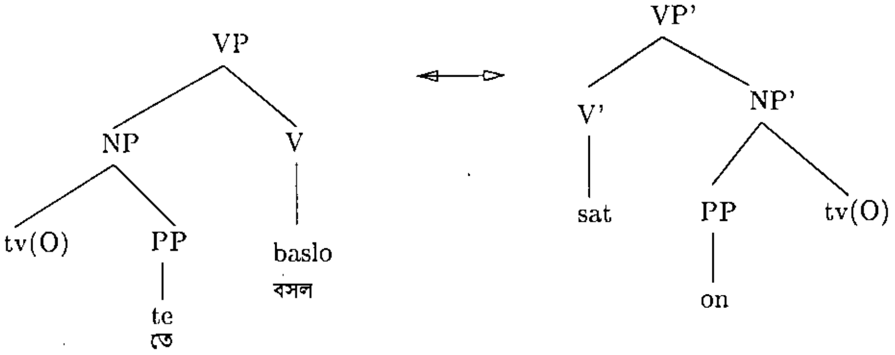


Figure 5.5: Structural divergence in Bangla-English transfer.

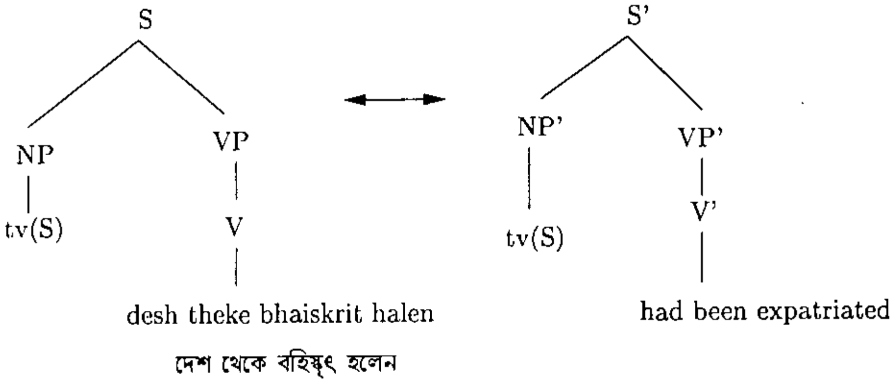


Figure 5.6: Lexical gap in Bangla-English transfer.

5.4.4 . Lexical Gap

Lexical gaps are single-word concepts in one language which can be rendered by two or more words in other language.

Bangla :তিনি দেশ থেকে বহিষ্কৃৎ হলেন

English :He had been expatriated

Figure 5.6 shows the lexical gap between Bangla and English for the Bangla sentence তিনি দেশ থেকে বহিষ্কৃৎ হলেন.

5.4.5 Collocational

Collocational divergences arise when the modifier, complement or head of a word is different from the default translation. Collocational divergence occurs in transitive verbs between two languages where the verb should be translated based on the object. Verbs play the key role in the whole semantics of a sentence. Therefore, a good translation of verbs is essential for improving the accuracy of the translated sentence.

Table 5.1: Examples of collocations for a verb তৈরি করা in the dictionary

Bangla	Collocational divergence in English
সে খাবার তৈরি করছে	She is preparing foods
সে একটি বাড়ি তৈরি করেছে	He has built a house
তিনি একটি গাড়ি তৈরি করেছেন	He has manufactured a car
ছেলেটি প্রোগ্রামটি তৈরি করেছে	The boy has developed the program
তিনি একটি মার্কেট তৈরি করেছেন	He has established a market
তিনি একটি সম্পর্ক তৈরি করেছেন	He has created a relation

Hence, for accurate determination of target verbs, collocational information is necessary to maintain [26].

Table 5.1 shows the collocations Bangla verb তৈরি করা and its translation into English based on the objects. To preserve the sense of meaning in such collocations a collocation dictionary is to be maintained for selecting the target words.

5.4.6 Multi-lexeme and Idiomatic

Mutli-lexeme and idiomatic divergences include those in which a phrase in one language corresponds to a phrase or a word in another language without there being clear translation relation between their individual words. In Bangla there are some proverbs that can not be translated into English or other languages according to the meaning of individual words. The proverb should be translated as a complete meaning in the target language. Let us consider the Bangla sentence tini ekTi durghaTanae paTal tullen (তিনি একটি দুঘটনায় পটল তুললেন) which has a proverb পটল তুললেন means died. This idiomatic phrase must be translated as a unit in English and the appropriate verb will be die.

5.4.7 Transfer Rules

Since a good measure of compositionality is a prerequisite for transfer to be possible at all, the transfer relation (trf) has to be recursively defined in terms of relation between progressively smaller subparts of the structural descriptions of the input text, up to

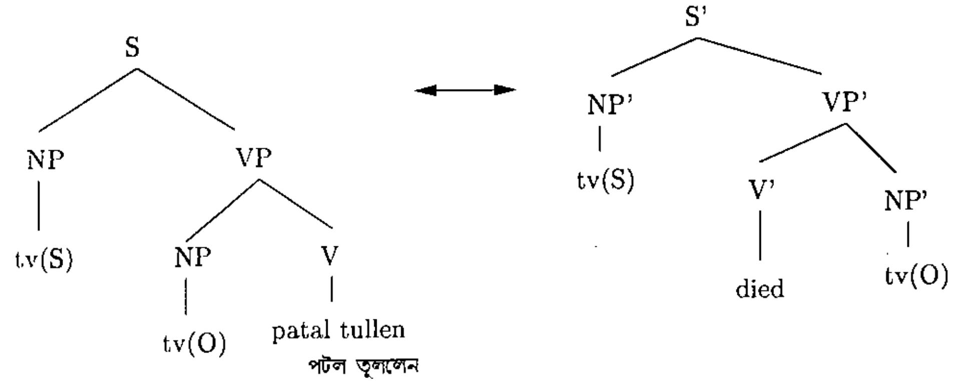


Figure 5.7: Idiomatic divergence in Bangla-English transfer.

the level of basic lexical units. Depending on the degree of compositionality involved, three types of rules can be distinguished [27].

The rules of Bangla and English grammars are presented in Appendix A and Appendix B respectively. The following pseudocode in Visual Prolog may be indicative of the transformations shown in Figure 5.2.

```
trf(nounp(Spcfr, Adj, Noun),  np(Detr, EAdj, ENoun)):-
                              trf(Spcfr, Detr),
                              trf(Adj, EAdj),
                              trf(Noun, Enoun).
          trf(Spcfr, Detr):-  bedict(Spcfr, Detr).
            trf(Adj, EAdj):-  bedict(Adj, EAdj).
        trf(Noun, ENoun):-    bedict(Noun, ENoun).
```

5.5 Bangla-English Transfer Algorithm

The Bangla-English transfer component subsequently converts a transformed Bangla tree to a English tree in a bottom-up and parallel manner along with Bangla tree. The transfer algorithm for Bangla-English can be presented as:

1. At first, the Bangla-English transfer dictionary is searched for all Bangla words that are terminals of the Bangla parse tree. This is done at the lowest level sub-tree of the Bangla parse tree.

2. An upper level Bangla sub-tree is converted to a corresponding English sub-tree by using the Bangla-English transfer rules and by using the Bangla-English transfer results of the current level Bangla sub-trees. The category of the top node of the upper sub-tree determines which set of Bangla-English transfer rules is to be applied.

3. During the transfer of sub-trees, semantic processing is performed according to the data in the Bangla-English transfer dictionary.

5.6 Problems of Syntactic Transfer

- Syntactic transfer modules are heavily dependent on the grammatical formalism and geometric structure assigned to sentences.
- It suffers from unbounded dependency constraints.
- It depends on extensive encoding of transformation details in the lexicons and in its pure form, the diminished capabilities for complex semantic reasoning.

5.7 Conclusion

Syntactic transfer strategy has been presented in this chapter with several test cases between Bangla-English transfer machine translation. The problems of syntactic transfer have also been mentioned. Actually MT depends to a large extent on the acquisition and development of large lexical, grammatical and terminological resources. Human expertise in an approach and experience of its use will affect the success of the approach. Different MT applications adopt different transfer systems. The approach which to be selected must be started with initial point to make it effective for the application.

Chapter 6

Generation

6.1 Generation

The generation or synthesis component of the target language produces one or more English language sentences from a English tree which conveys English syntax, English equivalent and other information. The roles of generation component are to generate English auxiliary verbs, to determine appropriate English equivalent of adverbs, negation, determiners and conjunctions including subordinate clauses etc. English auxiliary verbs are generated based on English verb information, such as the original form of the verb, the conjugation type of the verb, tense, aspect, voice and modality.

6.1.1 Syntactic Generation

Syntactic generation is carried out by means of transformations. In accordance with the theory of generative and transformational grammar, transformations occur in an orderly manner in an ascending cycle, that is to say from inside, outwards, starting with the most subordinate clauses. Actually syntactic generation starts by receiving the transferred tree structure from the transfer component. Then the grammar rules from the source side are also considered for the generation of target sentence. Figure 6.1 shows abstract tree structure of the Bangla sentence আমি ভাত খাই and its English `I eat rice`. From the figure we see that the root of each tree contains the grammars and the leaf nodes contain the lexicons with their semantic features. The abstract tree

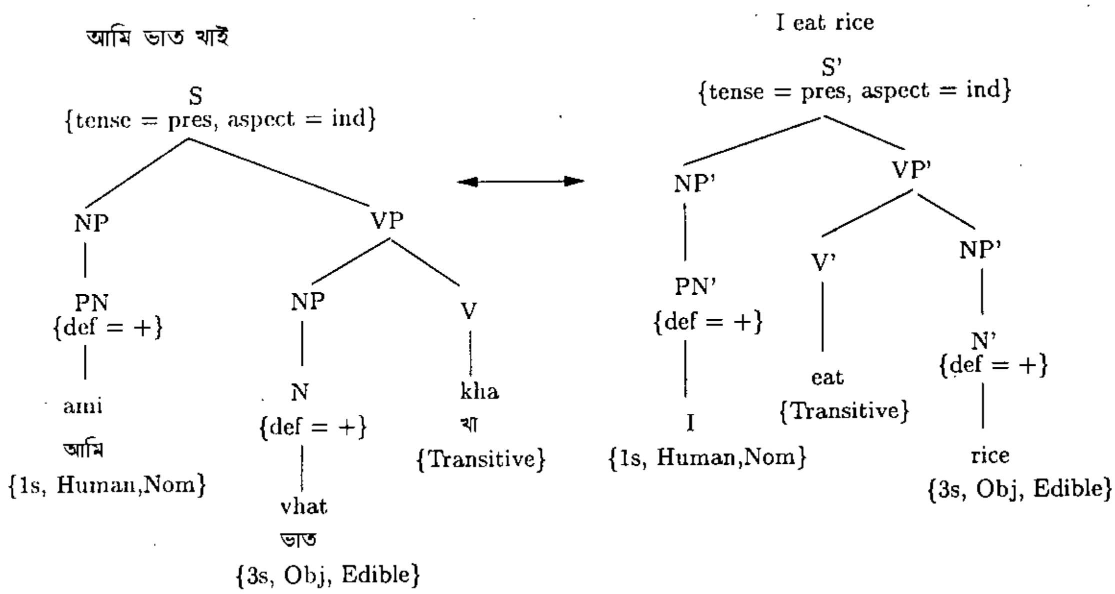


Figure 6.1: Target sentence generation of Bangla input আমি ভাত খাই

represented at the right side of Figure 6.1 shows the way of syntactic generation. In syntactic generation the grammatical features are received from the transfer component and the nodes of the tree are translated according to bilingual dictionary. The semantic properties are used by the morphological analyzer of the target side for generating actual word forms.

6.1.2 Morphological Generation

The morphological generation is made up of subject-verb agreement, conjunction, noun quantifier, insertion of determiner, noun agreement, adjective agreement, placement of adjective, elision and contraction. For example, the Bangla word cheleti (ছেলেটি) has the lexicon "chele" with the determiner (নির্দেশক "টি") "ti" and the semantic features as (`Noun, 3s, Definite, Human, Nominative`). The lexicon "chele" will be generated as "boy" in English. But the actual translation will be "the boy". The morphological analyzer of the target side using the semantic information will generate the translation of the Bangla word ছেলেটি as "the boy".

The generation of complete sentence is carried out syntactically and morphologically in parallel. Each leaf node is first generated syntactically and then morphological generation adds appropriate affixes to the leaf nodes for actual representation in

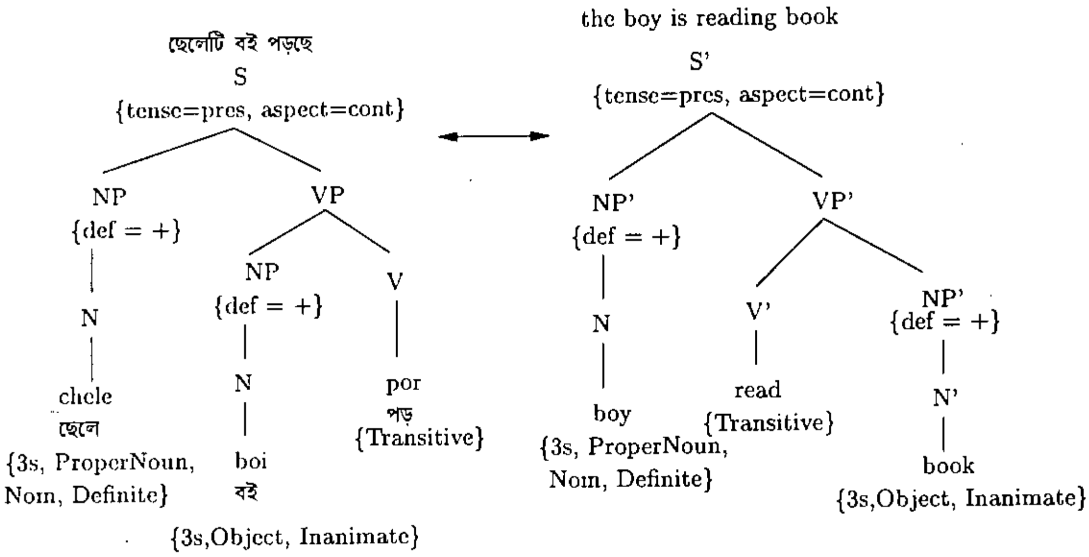


Figure 6.2: Target language sentence generation of Bangla input ছেলেটি বই পড়ছে.

the sentence. Looking at Figure 6.1, the root contains the grammar of the Bangla sentence আমি ভাত খাই as `tense = pres, aspect = ind` that means the sentence is in present indefinite tense. The leaf nodes contains the subject, object and verb of the sentence with their semantic properties. Semantic properties and the grammar will be transferred to the target language i.e. English so that actual target forms of words are generated. The subject and object here are in the lexical forms. But the verb খাই is separated by খা and ই. The morphological analyzer of Bangla will analyze the verb খাই as `lex = kha, cat=v, tense= pres, aspect = ind, mod = FP` and the transfer component will translate "kha" as "eat" and the morphological analyzer at the target side will use the transferred rule of grammar for the generation of the verb form "eat" which is the desired translation of খাই.

Figure 6.4 shows the translation of the Bangla sentence একটি ভাল ছেলে একটি ভাল বই কিনেছে. The parser output and the constituent parts are shown also. Finally, the English output "a good boy has bought a good book" which is produced by the MT engine has also been presented. Different analyses stages are not shown here. But they can be presented as:

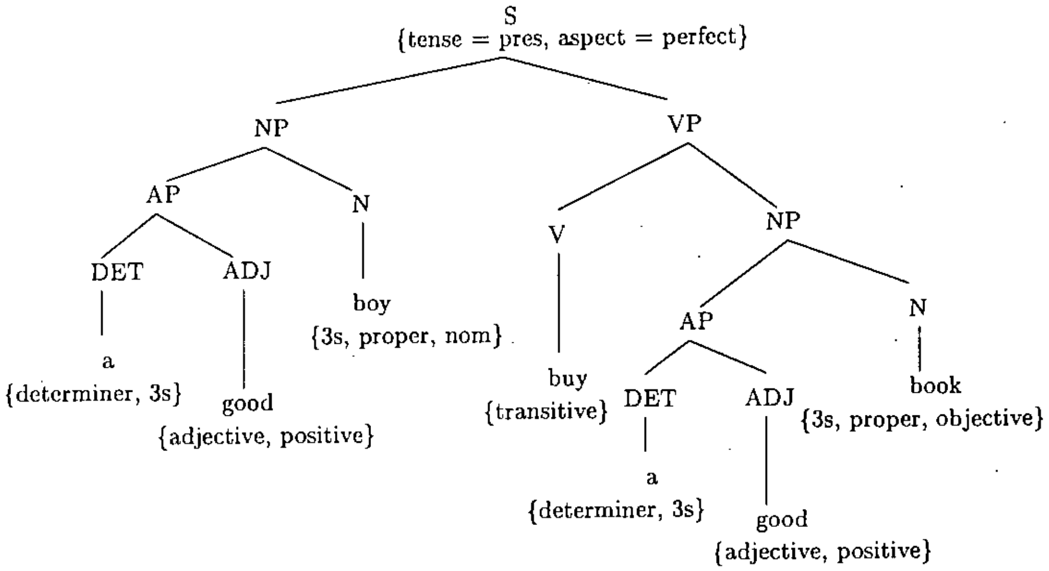
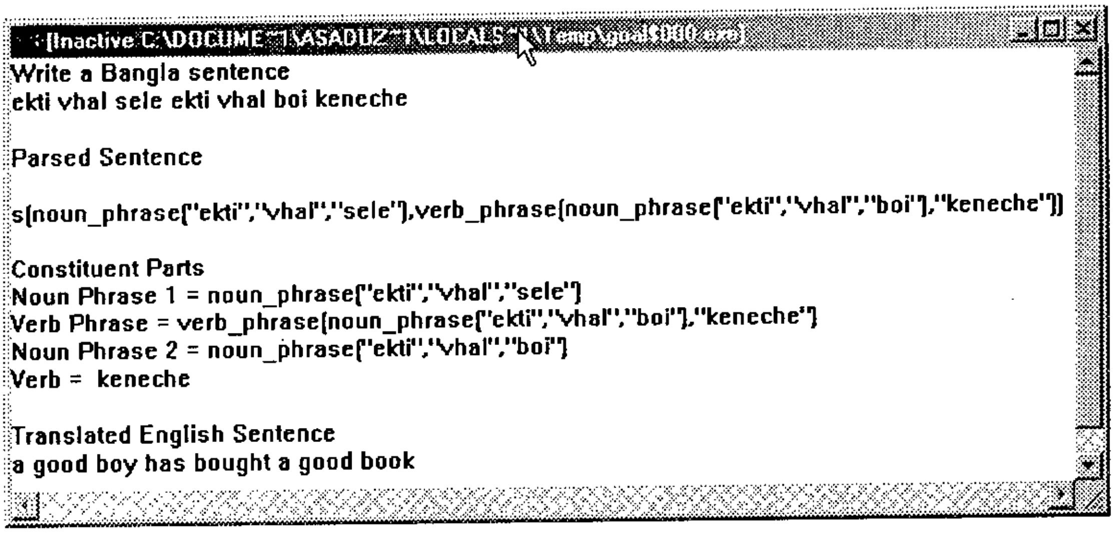


Figure 6.3: Target interface structure of the sentence একটি ভাল ছেলে একটি ভাল বই কিনেছে.

```
Input :            ekTi vhal chele ekTi vhal boi keneche (একটি ভাল ছেলে একটি ভাল বই
                   কিনেছে)
Scanner Output :   TOKL = ("ekTi", "vhal", "chele", "ekTi", "vhal", "boi",
                   "keneche")
Parser Output :    sen(np("ekTi", "vhal", "chele"), vp(np("ekTi", "vhal",
                   "boie"), verb("kenecehe"))))
                   ("ekTi", "Pers3", "Sing", "Qunatifier"),
                   ("vhal", "Adjective", "Positive", "Qualitative"),
                   ("chele", "Pers3", "Sing", "Human", "Nominative"),
                   ("boi", "Pers3", "Sing", "Objective", "ProperNoun"),
                   ("keneche", "ken", "Pres", "Perf", "TPNH")
Grammar :          ("Pres", "Perf")
```

The synthesis component will receive the transferred tree structure and the grammar for generating English output. The transferred tree structure is shown in Figure 6.3. The syntactic and morphological rules of English will be applied to each and every leaf of the tree and forms the phrase structure in English as well as the target result. After these operation the target output will be generated as: `a good boy has bought a good book`.



```
Write a Bangla sentence
ekti vhal sele ekti vhal boi keneche

Parsed Sentence

s(noun_phrase(["ekti","vhal","sele"]),verb_phrase(noun_phrase(["ekti","vhal","boi"]),"keneche"])

Constituent Parts
Noun Phrase 1 = noun_phrase(["ekti","vhal","sele"])
Verb Phrase = verb_phrase(noun_phrase(["ekti","vhal","boi"]),"keneche"]
Noun Phrase 2 = noun_phrase(["ekti","vhal","boi"])
Verb =  keneche

Translated English Sentence
a good boy has bought a good book
```

Figure 6.4: Translation of Bangla sentence একটি ভাল ছেলে একটি ভাল বই কিনেছে.

6.2 Successful Translation Cases

The MT egnine has been designed for translating simple Bangla sentence into English. Simple and short sentences are translated by the MT engine successfully. Appendix A shows the Bangla grammars for the sentences which are covered by the MT engine. The analogous English grammars are shown in Appendix B. All grammar rules have been tested with various examples.

6.3 Partially Successful Cases

Words in a Bangla sentence may be of any order, i.e. Bangla is a free word order language. Verb and object may interchange their positions in a sentence without changing the meaning of the sentence. The MT engine can parse any sentence in Bangla of the order subject-object-object (SOV) or subject-verb-object (SVO) for most of the cases. But it is not able to translate long sentence having noun phrases of large number of contituents.

6.4 Unsuccessful Cases

The MT engine can not translate interrogative, negative, negative-interrogative, optative, complex and compound sentences. The analysis, transfer and generation for those types are not included in the MT engine, if they are included the MT engine will be capable of translating them.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

An MT system has been developed and tested for Bangla to English translation. Very simple and short sentences are selected for translation. Since one of the prime goals is to improve the quality materials for human use, the MT engine presented here can only be indicative of possible arena for improvement. Translation from Bangla to English and vice versa is an interesting and pride tasks for us as we are Bangladeshi. The analysis, transfer of Bangla sentences and the generation of English sentences have presented throughout the chapters. We have shown the syntactic, semantic and morphological analyses of Bangla sentences and in the target side the generation of English sentences has also been discussed. The aim of the thesis work was to develop an MT engine for Bangla to English translation. But machine translation from Bangla to other languages has not been done yet. So we select our goal to translate only simple sentences. To include other types of sentences in Bangla, their analyses, transfers and syntheses should be included to upgrade the MT engine.

Our MT engine uses a Bangla parser for parsing input sentence that have syntax, semantic and morphological analyzers. After all the analyses the syntactic transfer strategy has been used to transfer the sentence into target system. In the target side the synthesis component of English uses English analyzer for generating English output.

However successful translation from one language to another requires controlled language for the creation of source text, which will be translated. So for Bangla language we must have to define a controlled language that will restrict the creation of Bangla sentences which will not be possible to translate by the machine. Another point is that machine translation is always applicable to a specific domain, such as weather reports, technical documents, email messages and instruction manuals etc. So it is wise to select the application domain first and design MT engine for that domain.

7.2 Future Work

There are a lot of research scopes regarding Bangla to other languages machine translation. As mentioned earlier machine translation of Bangla is in rudimentary stage, the following areas may be the research interests for the researchers.

- Syntax analyzer for complex and compound sentences.
- Powerful semantic analyzer for Bangla.
- Efficient morphological analyzer for Bangla.
- Design of controlled language for Bangla.
- Interlingua representations of Bangla.
- Transfer modules for Bilingual translation.

7.3 Alternative Strategies

The transfer technique adopted in the system is syntactic. The syntactic transfer MT system has been discussed in Chapter 5 with its demerits. The following sections are two alternative strategies in machine translation transfer. They have some features which can eliminate the problems of syntactic transfer. Comparisons of them are given in Table 7.1.

7.3.1 Semantic Transfer MT

One of the problems with syntactic transfer MT system is that in order to account for grammatical differences between two languages, many transformations are only

Table 7.1: Comparison of transfer strategies

Strategy	Advantages	Disadvantages
Syntactic	Simpler analysis, Faster development of grammar and Simpler automatic grammar induction	Complex transfer rule and interactions, Expensive to maintain
Semantic	Simpler transfer rules, Application other than MT and Theoretically motivated semantics	Expertise may be scarce and Need for changes as semantic theory develops
Lexicalist	Transparent transfer rules and Simpler acquisition of transfer modules	Difficult of include non-lexical information during transfer and Danger of inefficiency

variations of each other. For example auxiliary verbs in Bangla are attached with the verb roots based on the tense, aspect and mode of the sentence, but the translation into English shows that the auxiliaries in English appear before the verbs and the verbs in their inflected forms.

Semantic transfer MT sees the translation as a relation between language-dependent representations, which nevertheless neutralizes many language-specific idiosyncrasies. Although the representations are semantically oriented, they maintain some of the structure of their language of origin and therefore easier to use for analysis and generation [2].

Many representations of semantic transfer MT have been invented and some of them have been used for semantic transfer. Quasi-Logical Form (QLF) [28] is the prominent one in that case. In QLF transfer at the semantic level removes many language specific structures and replaces them with more standardized forms such as QLF. Still, it is possible to structure QLFs such that similarity between QLFs varies between source and target languages. For example, one could treat prepositions are relations between event variables and qterms (quantified terms), or they could have wider scope over the QLF for the VP they modify. Such decisions should ultimately be motivated by the semantic theory adopted, but in practice, it is difficult to ensure that equivalent QLF representations are built for translationally equivalent sentences. Semantic transfer rules overcome discrepancies that arise in the meaning representations produced by different grammars.

7.3.2 Lexicalist Transfer MT

A major source of complication with the syntactic and semantic transfer methodologies presented above is the recursive character of their representations. In the former, analysis trees themselves consist of analysis trees. In the latter, the argument to a QLF predicate can itself be a QLF. Problems arise when the transfer structure in SL and TL are markedly different elements of the SL structures that are geometrically distant may need to be in close proximity in the TL. Without additional mechanisms to cope with such divergences, transfer modules fail to capture useful and interesting cross-linguistic generalizations. These mechanisms will have non-local effects on the structures being transferred, which can decrease the perspicuity of the system and add to the difficulty in transfer rule development. This and other issues relating to scope and ambiguity have given rise to non-recursive approaches to transfer. Lexicalist MT is such a non-recursive approach in transfer.

In LexMT, cross-linguistic relationships are established at the level of lexemes or sets of lexemes, rather than more abstract representations. The principal advantage being that such relationships can be verified empirically by inspection of bilingual corpora, bilingual dictionaries or through validation by a bilingual. In addition, the automatic acquisition of contrastive data can be facilitated since transfer relationships are established in a format close to that found in bilingual corpora. Finally, the relations are reusable in the sense that they are to some extent independent of the syntactic and semantic theory of any particular system, and can therefore be adapted to different transfer approaches. This is unlike the previous two strategies in which one must have a significant amount of knowledge about the transfer representation and its behavior to be able to determine whether two structures stand in a transfer relation or not, and for which painstakingly acquired contrastive knowledge cannot be easily ported to other systems [2].

Appendix A

Bangla Grammar Rules

Rule Bng1a : S → NP V

আমি খাই (Pron + V)

সে পড়ে (Pron + V)

বাবুল খেলে (N + V)

আমি পড়ব (N + V)

Rule Bng1b : S → NP V

ছেলেটি খায় (N + Det + V)

ছেলেরা খায় (N+ PM + N)

ছেলেগুলো খায়(N+ PM + N)

Rule Bng1c : S → NP V

একটি ছেলে খায় (Spcfr + N + V)

অনেক ছেলে খেলছে

বহু লোক এসেছে

Rule Bng1d : S → NP V

ভাল ছেলেটি খায় (Adj + N + Det + V)

ভাল ছেলেরা খায় (Adj + N + PM + V)

ভাল ছেলেগুলো খায় (Adj + N + PM + V)

Rule Bng1e : S → NP V

একটি ভাল ছেলে খায় (Spcfr + Adj + N + V)

Rule Bng2a : S → NP NP V

আমি ভাত খাই (Pron + N + V)

বাবুল ভাত খায় (N + N + V)

Rule Bng2b : S → NP NP V

আমি একটি বই পড়ি (Pron + Spcfr + N + V)

Rule Bng2c : S → NP NP V

আমি একটি ভাল বই পড়ি (Pron + Spcfr + Adj + N + V)

ভাল ছেটি ভাল বই পড়ে (Adj + N + Det + Adj + N + V)

একটি ভাল ছেলে বইটি কিনেছে (Spcfr+ Adj + N + N + Det + V)

একটি ভাল ছেলে ভাল বইটি কিনেছে (Spcfr+ Adj + N + Adj + N + Det + V)

Rule Bng3a : S → NP NP V

আমি লোকটিকে চিনি (Pron + N + Biv + V)

Rule Bng4a : S → NP NP V

আমার বাবা ঢাকায় থাকেন (NP + Biv + N + N + Biv + V)

Appendix B

English Grammar Rules

Rule Eng1a : S → NP V

I eat (Pron + V)

He eats (Pron + V)

Babul plays (N + V)

I will read (Pron +V)

Rule Eng1b : S → NP V

The boy eats (Det + N + V)

Boys eat (N + V)

The boys eat (Det + N + V)

Rule Eng1c : S → NP V

A boy eats (Det + N + V)

Many boys eats (Det + N + V)

Rule Eng1d : S → NP V

The good boy eats (Det + Adj+ N + V)

The good boys eat

Good boys eat (Adj + N + V)

Rule Eng2a : S → NP V NP

I eat rice (Pron + V + N)

Babul eats rice (N +V +N)

Rule Eng2b : S → NP V NP

I read a book (Pron + V + Det + N)

Rule Eng2c : S → NP V NP

I read a good book (Pron + V + Det + Adj + N)

The good boy reads good book (Det + Adj + N + v + Adj +N)

A good boy has bought the good book (Det + Adj + N + V + Det + Adj + N)

Rule Eng3a : S → NP V NP

I know the man (Pron +V +Det + Noun)

Rule Eng4a : S → NP V NP

My father lives in Dhaka

Bibliography

[1] D. Arnold, L. Balkan, S. Meijer, R. L. Humphreys, and L. Sadler, *Machine Translation - An Introductory Guide.* NCC Blackwell Ltd., London, 1994.

[2] A. Trujillo, *Translation Engines: Techniques for Machine Translation.* Springer-Verlag, London, 1992.

[3] M. Kay, "Machine translation." Xerox-PARC, Palto Alto, CA and Stanford University.

[4] M. Chowdhury and M. H. Chowdhury, *Bangla Vashar Bakoron.* National Curriculum and Text Book Board, 2001.

[5] H. R. Thompson, *Essentials Everyday Bengali.* Bangla Academy, 1994.

[6] A. Humayun, *Bakyatattya.* The University of Dhaka, 1994.

[7] M. M. Murshed, "Parsing of Bengali natural language sentences," in *Proc. International Conference on Computer and Information Technology, ICCIT'98*, (Dhaka, Bangladesh), pp. 185–189, 1998.

[8] M. R. Selim and M. Z. Iqbal, "Syntax analysis of phrases and different types of sentences in Bangla," in *Proc. International Conference on Computer and Information Technology, ICCIT'99*, (Sylhet, Bangladesh), pp. 175–186, 1999.

[9] S. Asaduzzaman and M. M. Ali, "A system for automatic translation between Bangla and other natural languages: The design principle," in *Proc. International Conference on Computer and Information Technology, ICCIT'98*, (Dhaka, Bangladesh), pp. 194–198, 1998.

[10] M. M. Ali and M. M. Ali, "Development of machine translation dictionaries for Bangla language," in *Proc. International Conference on Computer and Information Technology, ICCIT'2002*, (Dhaka, Bangladesh), pp. 267–271, 2002.

[11] M. I. A. Khan, A. K. M. A. Hossain, and R. C. Debnath, "A Bangla conversation processor using natural language processing," in *Proc. International Conference on Computer and Information Technology, ICCIT 2002*, (Dhaka, Bangladesh), pp. 262–266, 2002.

[12] M. M. Murshed and M. A. Mottalib, "Design and implementation of a bilingual natural language parser (for Bangla to English)." M. Sc. thesis, Department of Computer Science, University of Dhaka, Dhaka, Bangladesh.

[13] D. W. Patterson, *Introduction to Artificial Intelligence and Expert Systems.* Printice Hall of India Pvt. Ltd., 2002.

[14] E. Charniak and D. McDermott, *Introduction to Artificial Intelligence.* Addison-Wesley Longman Inc., 1999.

[15] E. Rich and K. Knight, *Artificial Intelligence.* Tata-McGraw-Hill Ltd, 1991.

[16] A. V. Aho and J. Ullman, *The Theory of Parsing Translation and Compiling*, vol. 1-1: Parsing. Printice-Hall Inc., 1972.

[17] A. V. Aho, R. Shethi, and J. Ullman, *Compilers, Principles, Techniques and Tools.* Addison-Wesley, 2002.

[18] A. J. Thomson and A. V. Martinet, *A Practical English Grammar.* Green View Publishers, Dhaka, Bangladesh, 2001.

[19] R. Evans and G. Gazder, "Datr: A language for lexical knowledge representation," *Computational Linguistics*, vol. 22, no. 2, pp. 167–216, 1996.

[20] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation.* Pearson Education, Inc., 2001.

[21] L. Karttunen, R. M. Kaplan, and A. Zaenen, "Two-level morphology with composition," in *Proc. International Conference on Computational Linguistics, COLING '92*, (Nates, France), pp. 141–148, 1992.

[22] A. M. F. U. Bhuiyan, *Nuton Poddotite Bangla Bakoron.* Ideal Library, 1997.

[23] R. 1slam, *Bangla Backaron Samiikkha.* Globe Library Pvt, Bangladeshd, 2001.

[24] M. M. Wood, E. Pollard, H. Horsfall, N. holden, B. Chandler, and J. Carrol, "Dictionary organization for machine translation: The experience and implication of the umist japanese project." Center for Computational Linguistics, UMIST, Manchester, UK.

[25] D. Rao, P. Bhattacharya, and R. Mamidi, "Natural language generation for English to Hindi human-aided machine translation," in *Proc. International Conference on Knowledge Based Computer Systems,* (Mumbai, India), pp. 171–189, 1998.

[26] Y. Kim, B. T. Zhang, and Y. T. Kim, "Collocational dictionary optimization using wordnet and k-nearest neighbour learning," *Machine Translation,* vol. 16, no. 2, pp. 89–108, 2001.

[27] P. Isabelle and E. Mackiovitch, "Transfer and modularity." Canadian Workplace Automation research Centre, Quebec, Canada.

[28] H. Alshawi, D. Carter, B. Gamback, and M. Rayner, *Swedish-Elglish QLF Translation,.* Chapter 14, pp. 277-309, 1992.