*International Journal of Engineering Science Invention*
*ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726*
*www.ijesi.org Volume 2 Issue 10‖ October 2013 ‖ PP.30-38*

# Tense Based English to Bangla Translation Using MT System

[1,]Kanija Muntarina , [2,]Md. Golam Moazzam and [3,]Md. Al-Amin Bhuiyan
*Department of Computer Science and Engineering*
*Jahangirnagar University, Savar, Dhaka-1342, Bangladesh.*

**ABSTRACT :** *This paper presents a tense based Machine Translation (MT) system which operates on English sentences as an input language. The MT system translates the input into its corresponding Bangla sentences using the Natural Language Processing (NLP) conversion technique. It uses context-free grammars for validating the syntactical structure of the input sentences and for parsing it applies the bottom up approach which also helps to verify the accuracy of the generated output sentences. A translation module is added to the MT system to translate out-of-vocabulary words. The effectiveness of this method has been justified over the demonstration of different English sentences with several rules.*

**KEYWORDS :***Machine translation, Natural language processing, Context-free grammar, Parsing, Bottom-up approach.*

## I. INTRODUCTION

Machine translation system design is one of the important fields in modern age for natural language processing. It refers to translation of text or speech from one natural language to another. MT performs simple substitution of words in one natural language for word in another. Using corpus technique more complex translations may be attempted, allowing for better handling of differences in linguistic topology, phrase recognition and translation of Speech, as well as the isolation of anomalies [1], [2]. The demand for machine translation is growing day by day. Analysis and generation are the two major scheme of machine translation. This research deals with the translation of English to Bangla sentences based on tense based rules using MT system and construct parse tree for the both languages. It takes analyzed English sentences as input and generates Bangla sentences as output with its structural representation. The translation and structural representation are done based on some sets of rules and lexicon. The rule sets are generated according to linguistic rule. In rule based approach the source language text is analyzed using various tools like morphological analyzed, parser and transformed to an intermediate representation. A large number of rules is required to natural language processing. These rules transfer the grammatical structure of source language into target language. As the rules increases the system becomes complicated [3].

## II. PREVIOUS DEVELOPMENTS

A fair amount of works have been done on English to Bangla translation using MT in the context of rule based MT. For example, Letter Simplification, English to Bangla MT, Morphological Analysis of Bangla Words, Parsing of Bangla Sentences, and Parsing of English sentences, Bengoli composite letter, English Speech Recognition, English Character Recognition, etc. Most of the architecture follows some common steps: a) Tagging b) Parsing c) transfer of English parse tree to Bangla parse tree d) Generate output translation with morphological analysis [4].An example based machine translation system [5] identifies the phrases in the input through a shallow analysis, retrieves the target phrases using a phrasal example based and finally combines the target language phrases by employing some heuristic based on the phrase reordering rules in Bangla. The NP structure differs in English and Bangla, in English: [specifier/article] [adv] [start noun] [plural marker] [case marker] and in Bangla: [specifier] [adv] [adj] [noun] [plural marker] [case marker]. There are some similarities between English and Bangla pronoun as well, but this differs on gender: Bangla pronouns do not depend on gender information and for second and third persons have more than one forms. So translating pronouns from English to Bangla involves anaphora relation. In Bangla adjective forms (positive, comparative, superlative) are handled in a similar way as in English. This system used a shallow parser to identify the phrases in the source language and tags with the phrase with relevant information, translated these phrases individually and arrange them using some phrase ordering heuristic rules.A phrasal EBMT system for translating English to Bangla [6] works in three steps. In the first step search in direct example base, if not found, then search in generalized tagged example base.

If a match is found in the second step, then extract the English equivalent of the Bangla words from the bilingual dictionary and apply some synthesis rules to generate the word label. If the second step fails, then the tag input headline is analyzed to identify the constituent phrases. The target translation is generated from the bilingual example phrase dictionary and uses the heuristics to reorder Bangla phrases.Another approach has been proposed that treats source and target sentences as strings of letters instead of a collection of words [7]. It treated each word as a sequence of letter, which is translated into a new sequence of letters. This approach reduces the vocabulary size significantly but increased the average sentence length. This system could be useful for closely related languages and languages where very little parallel training data is available.Translation systems are being used now-a-days in machine translation system. A phonetic based translation system for English to Bangla produces intermediate code strings that facilitate matching pronunciations of input and desired output [8]. It used table driven direct mapping between English to Bangla alphabet and a phonetic lexicon enable mapping.

## III. PROPOSED MODEL

The proposed model which can translate grammatical (tense based: present, past or future) English sentence to corresponding Bangla output sentence is shown in Fig.1. To determine the syntactical structure of the input sentence, at first the lexical analyzer component tokenizes the constituents of the input sentence to construct a list of words. If a constituent is not a valid word, it returns the 'invalid'. The generated word list is then used to construct a syntactic tree in the source language from the syntactic structure of the input sentence determined during bottom-up parsing. The lexicon provides the possible features and the corresponding target language word for each input word.

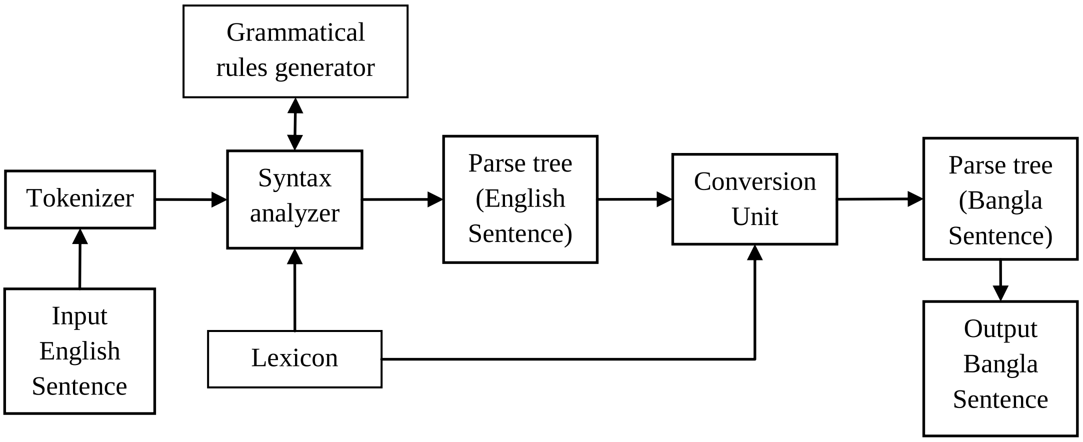


Fig. 1: Proposed model for translating English sentence to corresponding Bangla output sentence

***Tokenizer***

Tokenizer is the program module that accepts the input sentence and parsed into unbroken units, every individual unit called Tokens. Tokens are stored in a list for further access. The Token is then checked into the lexicon for validity. In Lexical analysis phase the stream of characters are sequentially scanned and grouped into tokens according to lexicon. The output of the Tokenizer of the input sentence **"**she loves flowers**"** is as follows [9], [10]:

TOKEN = ("She", "loves", "flowers").

**Syntax Analyzer**

The Syntax Analyzer first takes the tokens and search in the Lexicon. If the tokens are not found in the Lexicon then it shows "Invalid" words. If all the tokens are found in the lexicon it then check the grammatical rules. According to the grammatical rules it parses the English sentence.

**Grammatical Rule Generator**

The most common way to represent grammar is as a set of production rules which says how the parts of speech can put together to make grammatical, or "well- formed" sentences.

**Lexicon**

Lexicon contains the priori tag and suffix for each word. It has two parts: a *rootform lexicon* and a *suffix lexicon*. During the lookup of a word in the lexicon, the *rootform* lexicon is searched first. If the word is found there, the corresponding phrases category is returned. If it fails, the *suffix lexicon* is searched next and for the rest part of that word, root lexicon is searched. This is done until reach to the beginning of the word from ending point. The advantages of only maintaining root forms in the *lexicon* are: generalizations are captured, knowledge is localized and hence more easily maintained, and new word forms are predicted [11], [12]. So, it is clearly desirable to have only headwords in an MT dictionary. This will save time, space, and effort in constructing entries. So, the coverage of our vocabulary of the *lexicon* is very extensive and appropriately selected. The translation equivalence is carefully chosen. The lexicon of our project is given below:

Noun -> book eB | bird cvwL | meal Lvevi | kite Nywo | bus evm | project cÖ‡R± | bycycle evBmvB‡Kj | novel Dcb¨vm | picture Qwe | flowers dzj | flower dzj | baby ev"Pv | car Mvwo | rice fvZ | song Mvb | boy ‡Q‡j | tea Pv | flower dzj | girl ‡g‡q | letter wPwV | football dzUej | cricket wµ‡KU | popy cwc | car Mvwo | mango Avg | hours N›Uv |

Pronoun -> I Avwg | he ‡m | she ‡m | you Zzwg | him Zv‡K| they Zviv | them Zv‡`i | me Avgvq | we Avgiv | his Zvi |

Auxiliary verb -> am nB | will Ki‡m | are nB | does Ki | is nq | was nq |

Main verb -> reading cwo‡ZwQ | stopped _vwgqvwQ | done K‡i‡Q | finished ‡kl K‡i‡Q | gone wM‡q‡Q | ridden Pwoqv‡Q | seen ‡`‡L‡Q | flown ewnqv‡Q | wanted ‡P‡qwQjvg | help mvnvh¨ | crying Kvw`‡Z‡Q | painting AvwK‡Z‡Q | love fvjevmv | loves fvjevmv | sleeping Nygvq‡Z‡Q | sleeps Nygvq | slept Nygvqv‡Q | singing MvB‡ZwQ | reads c‡o | eat Lvq | eats Lvq | drinks cvb K‡i | drink cvb K‡i | drinking cvb Kwi‡Z‡Q | drawn G‡K‡Q | gives ‡`q | give ‡`q | gave w`‡qwQj | writing wjwL‡ZwQ | play ‡L‡j | played ‡L‡j‡Q | driving Pvjvq‡ZwQ | playing ‡Lwj‡ZwQ | plays ‡L‡j | eating LvB‡ZwQ | dancing bvwP‡Z‡Q | eaten ‡L‡qwQ | given w`‡j | written wj‡LwQ | sung ‡M‡qwQ | driven Pvwj‡qwQ | waiting A‡c¶v Kwi‡ZwQ | called ‡W‡KwQ‡j | saw ‡`‡LwQjvg | watching ‡`wL‡ZwQ |

Article -> a GKwU | the wU |

Conjuction ->but wKš‘ | and Ges |

Additive word -> where ‡hLv‡b | there ‡mLv‡bB | that IB |

Interogative -> what Kx |

Adjective -> two `yB | ten `k | mnR easy | fvj good |

Preposition -> for Rb¨ | by `viv |

**Parse Tree**

Parsing refers to a process of finding a parse tree for a given input string. That is, calls to the parsing function PARSE, such as PARSE ("the baby has slept") should return a parse tree with root S whose leaves are "the baby has slept" and whose internal nodes are no terminal symbols [8].

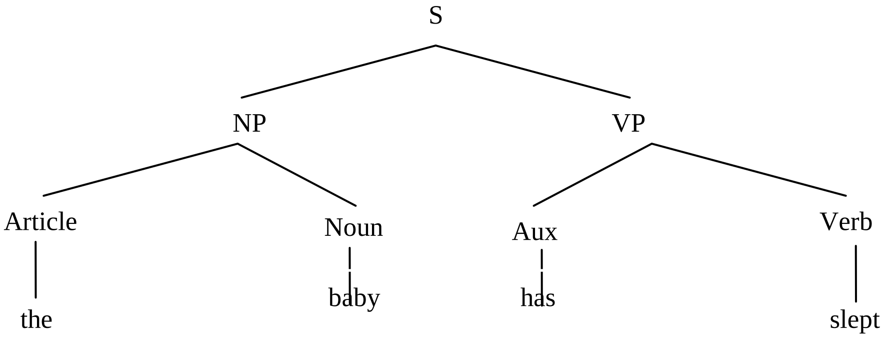


Fig. 2: Parse tree for input English sentence

In linear text, the tree can be written as

[S: [NP: [Article: **the**] [Noun: **baby**]] [VP: [Aux: **has**] [Verb: **slept**]]]

**Conversion Unit**

The converter accepts the input sentence, analyzes and converts into a parse tree/structural representation (SR). Once the SR is created for a particular sentence, it is then converted to corresponding Bangla sentence by conversion unit. The output parse for the bangla sentence will be generated according to the Bangla syntactic structure. The structure of the parse tree will be:

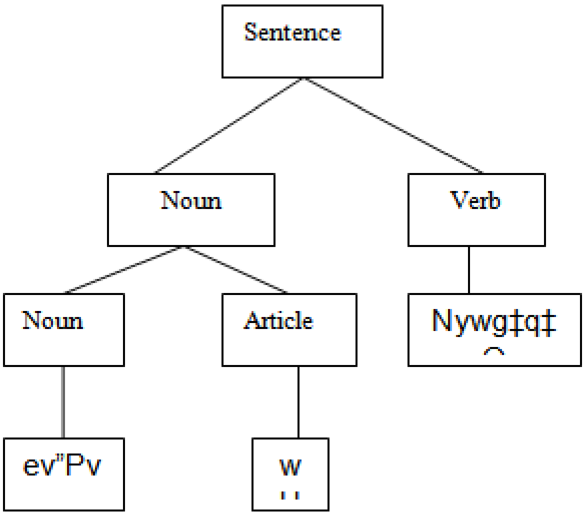


Fig. 3: Parse tree for output Bangla sentence

**Output Bangla Sentence**

Some special enlisted Bangla sentence is the output of the Conversion Unit. In English or Bangla language there are unlimited sentences. Everyone can express his/her feelings within a sentence. This research considered the following tense based grammatical structure of sentences:

- Present Indefinite, Present Continuous, Present Perfect, Present Perfect Continuous
- Past Indefinite, Past Continuous, Past Perfect, Past Perfect Continuous
- Future Indefinite, Future Continuous, Future Perfect, Future Perfect Continuous

**Flowchart**

The flowchart of the proposed model for English to Bangla Translation process is shown below.

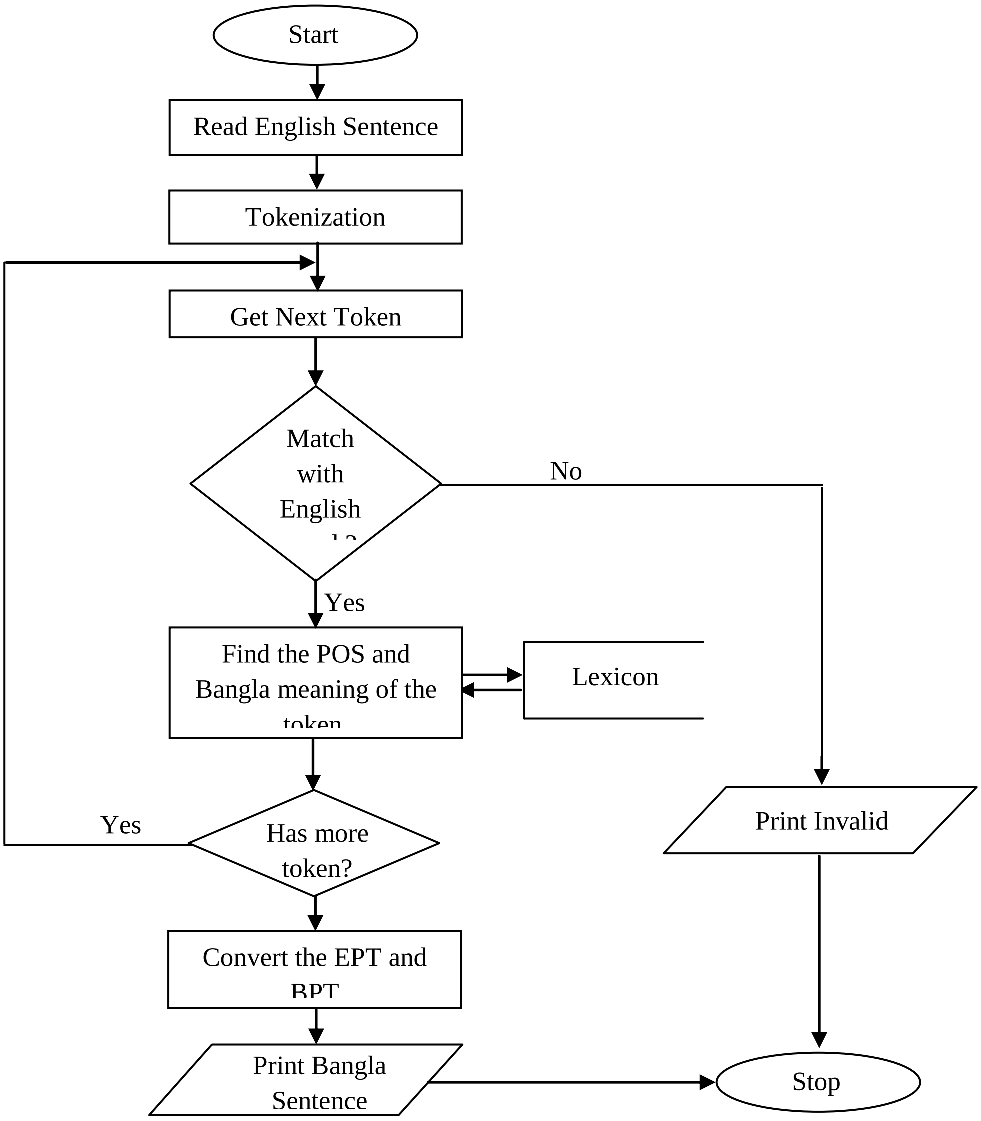
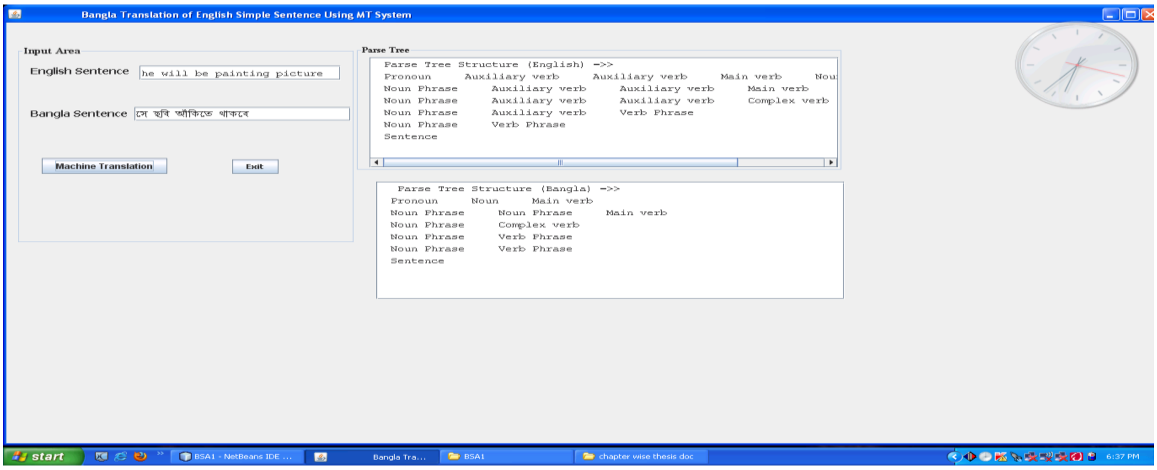


Fig. 4: Flow chart of English to Bangla Translation Process.

## IMPLEMENTATION AND EXPERIMENTAL RESULTS

Natural language processing can be implemented in any programming language. This translation system has been implemented in JAVA 2 platform. The software is actually executable JAR file. The integrated development environment (IDE) that we have used is NetBeans IDE 6.8. For showing Bangla meaning of the input English sentence, a jar file named bswing.jar is used.The sample input, output and parse tree for the input English sentence and parse tree for Bangla output sentence are shown in Fig. 5, Fig. 6, Fig. 7 and Fig. 8 respectively. We have implemented every types of tense based sentence structure. We have also implemented a few types of interrogative sentence structures.



Input Area

English Sentence: he will be painting picture

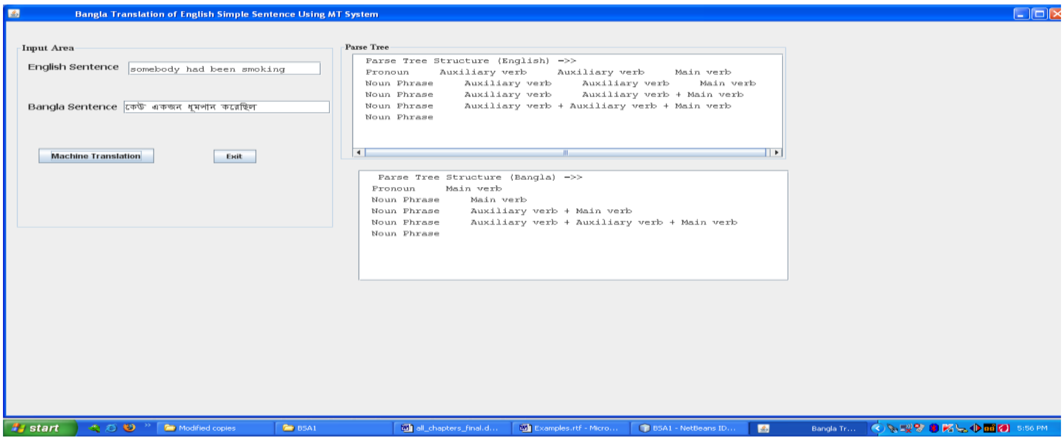
Bangla Sentence: সে ছবি আঁকিতে থাকবে

Parse Tree

```
 Parse Tree Structure (English) ->>
 Pronoun      Auxiliary verb      Auxiliary verb      Main verb      Nou
 Noun Phrase      Auxiliary verb      Auxiliary verb      Main verb
 Noun Phrase      Auxiliary verb      Auxiliary verb      Complex verb
 Noun Phrase      Auxiliary verb      Verb Phrase
 Noun Phrase      Verb Phrase
 Sentence
```

```
  Parse Tree Structure (Bangla) ->>
 Pronoun      Noun      Main verb
 Noun Phrase      Noun Phrase      Main verb
 Noun Phrase      Complex verb
 Noun Phrase      Verb Phrase
 Noun Phrase      Verb Phrase
 Sentence
```

Fig. 5: Output of Future Continuous tense based sentence



Input Area

English Sentence: somebody had been smoking

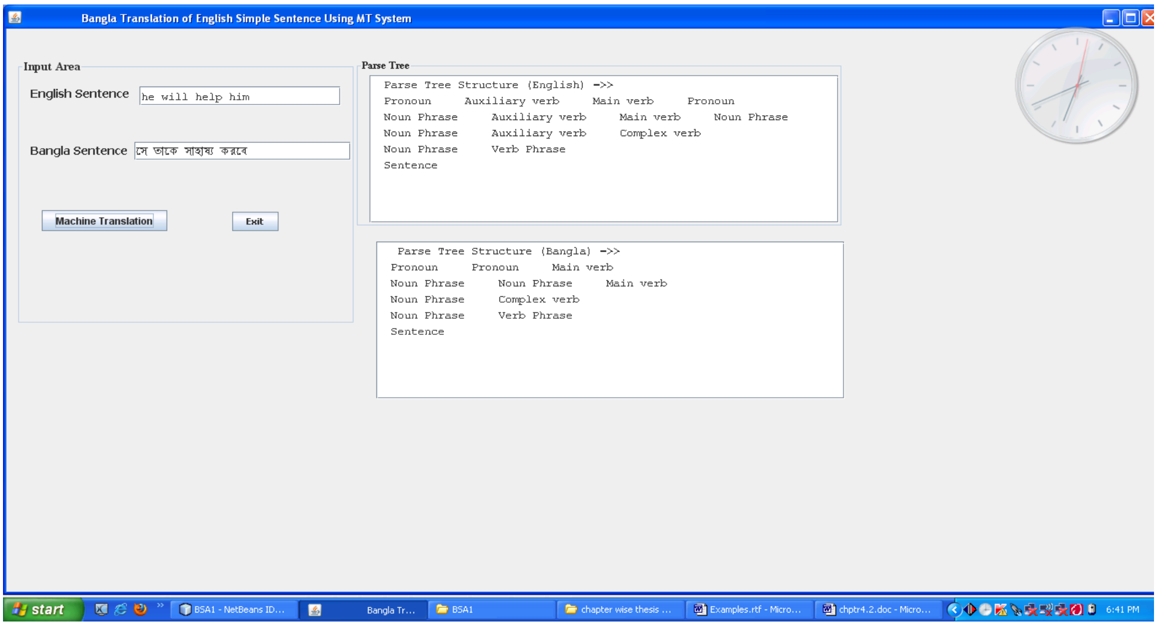
Bangla Sentence: কেউ একজন ধূমপান করেছিল

Parse Tree

```
 Parse Tree Structure (English) ->>
 Pronoun      Auxiliary verb      Auxiliary verb      Main verb
 Noun Phrase      Auxiliary verb      Auxiliary verb      Main verb
 Noun Phrase      Auxiliary verb      Auxiliary verb + Main verb
 Noun Phrase      Auxiliary verb + Auxiliary verb + Main verb
 Noun Phrase
```

```
  Parse Tree Structure (Bangla) ->>
 Pronoun      Main verb
 Noun Phrase      Main verb
 Noun Phrase      Auxiliary verb + Main verb
 Noun Phrase      Auxiliary verb + Auxiliary verb + Main verb
 Noun Phrase
```

Fig. 6: Output of Past Perfect Continuous tense based sentence



Input Area

English Sentence: he will help him

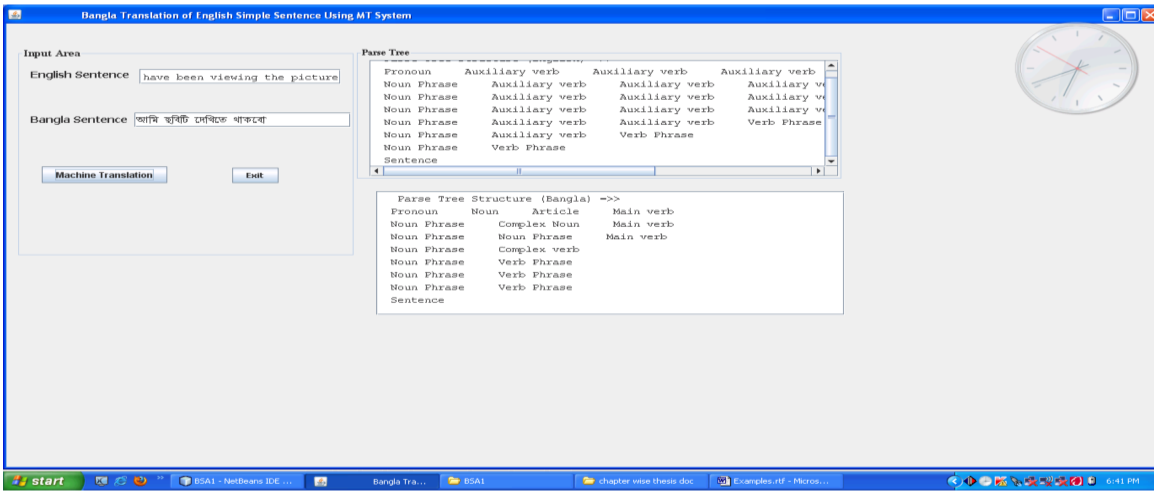
Bangla Sentence: সে তাকে সাহায্য করবে

Parse Tree

```
 Parse Tree Structure (English) ->>
 Pronoun      Auxiliary verb      Main verb      Pronoun
 Noun Phrase      Auxiliary verb      Main verb      Noun Phrase
 Noun Phrase      Auxiliary verb      Complex verb
 Noun Phrase      Verb Phrase
 Sentence
```

```
  Parse Tree Structure (Bangla) ->>
 Pronoun      Pronoun      Main verb
 Noun Phrase      Noun Phrase      Main verb
 Noun Phrase      Complex verb
 Noun Phrase      Verb Phrase
 Sentence
```

Fig. 7: Output of Future Indefinite tense based sentence


Bangla Translation of English Simple Sentence Using MT System
Input Area
Parse Tree
Machine Translation
Exit
start

English Sentence: have been viewing the picture

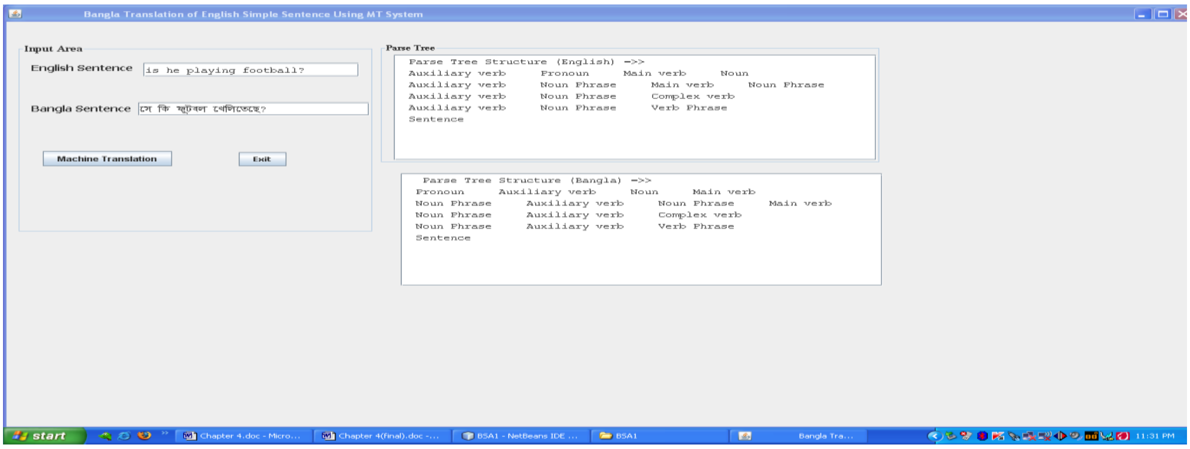
Bangla Sentence: আমি ছবিটি দেখিতে থাকবো

```
Pronoun      Auxiliary verb     Auxiliary verb     Auxiliary verb
Noun Phrase      Auxiliary verb     Auxiliary verb     Auxiliary ve
Noun Phrase      Auxiliary verb     Auxiliary verb     Auxiliary ve
Noun Phrase      Auxiliary verb     Auxiliary verb     Auxiliary ve
Noun Phrase      Auxiliary verb     Auxiliary verb     Verb Phrase
Noun Phrase      Auxiliary verb     Verb Phrase
Noun Phrase      Verb Phrase
Sentence
```

```
 Parse Tree Structure (Bangla) ->>
Pronoun     Noun      Article     Main verb
Noun Phrase      Complex Noun      Main verb
Noun Phrase      Noun Phrase      Main verb
Noun Phrase      Complex verb
Noun Phrase      Verb Phrase
Noun Phrase      Verb Phrase
Noun Phrase      Verb Phrase
Sentence
```

Fig. 8: Output of Present Perfect Continuous tense based sentence

In this research work we have considered some interrogative sentence type structures. Some snapshots of the output for interrogative sentence type structures are shown in Fig. 9 and Fig.10.


Bangla Translation of English Simple Sentence Using MT System
Input Area
Parse Tree
Machine Translation
Exit
start

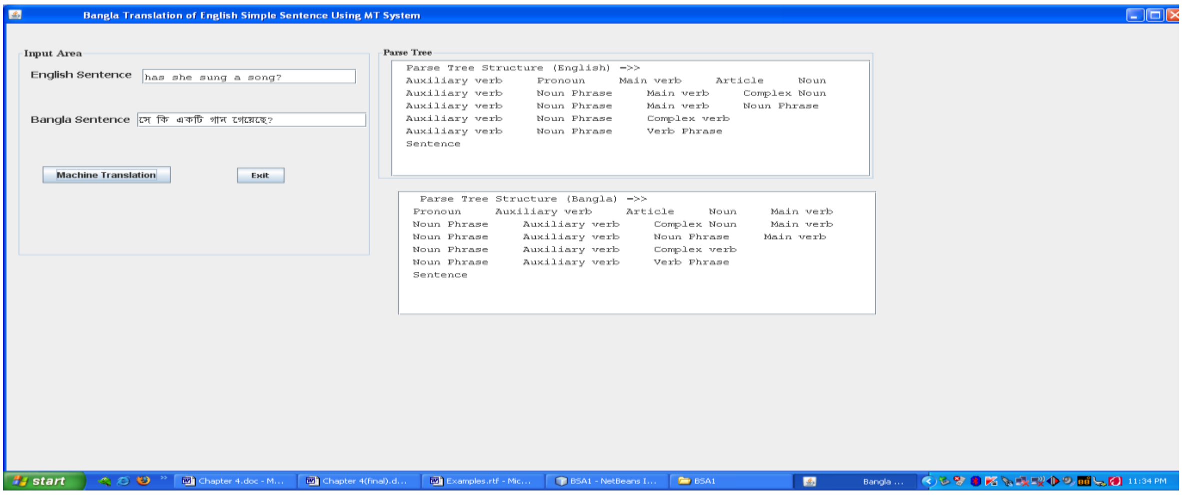
English Sentence: is he playing football?

Bangla Sentence: সে কি ফুটবল খেলিতেছে?

```
 Parse Tree Structure (English) ->>
Auxiliary verb      Pronoun      Main verb      Noun
Auxiliary verb      Noun Phrase      Main verb      Noun Phrase
Auxiliary verb      Noun Phrase      Complex verb
Auxiliary verb      Noun Phrase      Verb Phrase
Sentence
```

```
 Parse Tree Structure (Bangla) ->>
Pronoun      Auxiliary verb      Noun      Main verb
Noun Phrase      Auxiliary verb      Noun Phrase      Main verb
Noun Phrase      Auxiliary verb      Complex verb
Noun Phrase      Auxiliary verb      Verb Phrase
Sentence
```

Fig. 9: Output of Interrogative (Present Continuous tense) based sentence


Bangla Translation of English Simple Sentence Using MT System
Input Area
Parse Tree
Machine Translation
Exit
start

English Sentence: has she sung a song?

Bangla Sentence: সে কি একটি গান গেয়েছে?

```
 Parse Tree Structure (English) ->>
Auxiliary verb      Pronoun      Main verb      Article      Noun
Auxiliary verb      Noun Phrase      Main verb      Complex Noun
Auxiliary verb      Noun Phrase      Main verb      Noun Phrase
Auxiliary verb      Noun Phrase      Complex verb
Auxiliary verb      Noun Phrase      Verb Phrase
Sentence
```

```
 Parse Tree Structure (Bangla) ->>
Pronoun      Auxiliary verb      Article      Noun      Main verb
Noun Phrase      Auxiliary verb      Complex Noun      Main verb
Noun Phrase      Auxiliary verb      Noun Phrase      Main verb
Noun Phrase      Auxiliary verb      Complex verb
Noun Phrase      Auxiliary verb      Verb Phrase
Sentence
```

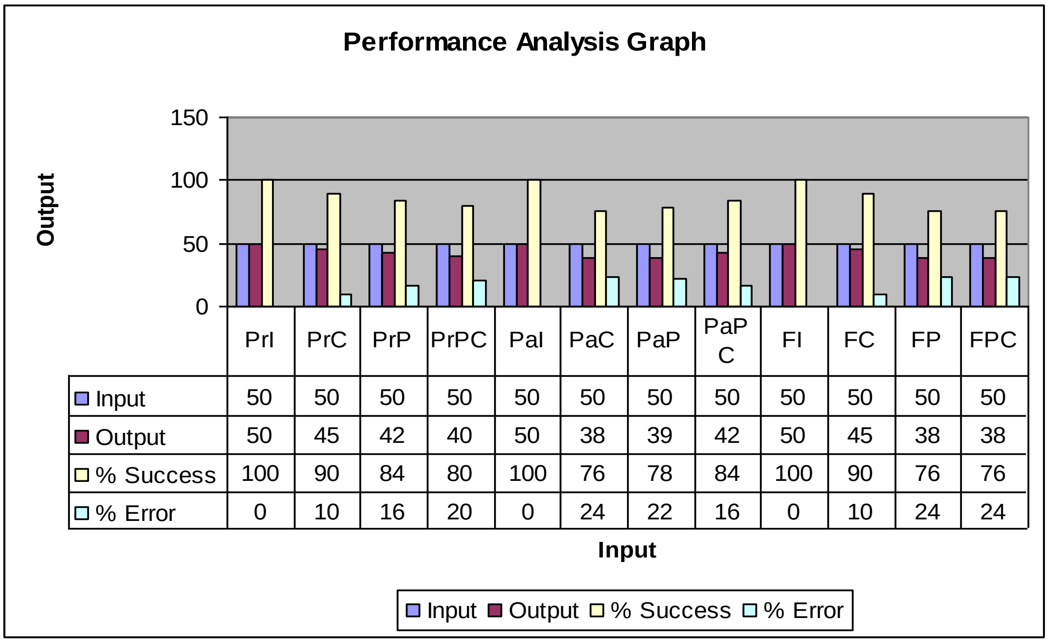
Fig. 10: Output of Interrogative (Present Perfect tense) based sentence

## IV. PERFORMANCE ANALYSIS

In order to justify the robustness of the system, we have used a total of (50*12) English sentences and the results are summarized in the table I. The performance of the system has also been analyzed graphically in the bar diagram of Fig. 11.

**Table I: Performance measurement**

| Tense | Short Form | Input | Output | Percentage of Success | Percentage of Error |
|---|---|---|---|---|---|
| Present Indefinite | PrI | 50 | 50 | 100 | 0 |
| Present Continuous | PrC | 50 | 45 | 90 | 10 |
| Present Perfect | PrP | 50 | 42 | 84 | 16 |
| Present Perfect Continuous | PrPC | 50 | 40 | 80 | 20 |
| Past Indefinite | PaI | 50 | 50 | 100 | 0 |
| Past Continuous | PaC | 50 | 38 | 76 | 24 |
| Past Perfect | PaP | 50 | 39 | 78 | 22 |
| Past Perfect Continuous | PaPC | 50 | 42 | 84 | 16 |
| Future Indefinite | FI | 50 | 50 | 100 | 0 |
| Future Continuous | FC | 50 | 45 | 90 | 10 |
| Future Perfect | FP | 50 | 38 | 76 | 24 |
| Future Perfect Continuous | FPC | 50 | 38 | 76 | 24 |



|  | PrI | PrC | PrP | PrPC | PaI | PaC | PaP | PaP C | FI | FC | FP | FPC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| □ Input | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| ■ Output | 50 | 45 | 42 | 40 | 50 | 38 | 39 | 42 | 50 | 45 | 38 | 38 |
| □ % Success | 100 | 90 | 84 | 80 | 100 | 76 | 78 | 84 | 100 | 90 | 76 | 76 |
| □ % Error | 0 | 10 | 16 | 20 | 0 | 24 | 22 | 16 | 0 | 10 | 24 | 24 |

Fig. 11: Performance analysis graph

In this research work several samples have been taken for the purpose of performance measurement. From the above result it is being found that for present indefinite, past indefinite and future indefinite tense based sentence structure the success rate is 100%. There are some fluctuations in present continuous, past continuous and future continuous tense based sentence structure. Because according to the structure we have to use some auxiliary verbs in English sentences. But in Bangla sentence the auxiliary verb is totally neglected. In addition, in the perfect and perfect continuous based structure is very complex. Such types of structure follows prepositions and some indefinite complex structure rules which have been ignored here. Due to these reasons some inconsistencies have been found.

## REFERENCES

[1] M. M. Asaduzzaman and Muhammad Masroor Ali, Transfer Machine Translation – An Experience with Bangla English Machine Translation System, In the Proceedings of the International Conference on Computer and Information Technology 2003.

[2] Shah Anqur Rahman, Kazi Shahed Mahmud, Banani Roy. and K. M. Azharul Hasan -- English to Bengali Translation Using A New Natural Language Processing Algorithm, published in Proceedings of International Conference on Computer and Information Technology 2003.

[3] Tamanna Haque Nipa, Muhammad Harun-Owr-Roshid, Salahuddin Mohammad Masum, Mortuza Ali -- Bangla and English Morphological Rules for Machine Translation Dictionary, published in Proceedings of International Conference on Computer and Information Technology 2003.

[4] M. Sipser, Introduction to the Theory of Computation, (2nd Edition, McGraw-Hill, 2005.

[5] SK Naskar and S Bandyopadhyay. Handleing of prepositions in english to bengali machine translation. In *Proceedings of The EOCL 2006 Workshop*, 2006

[6] SK Naskar and S Bandyopadhyay. A phrasal ebmt system for translating english to bengali. In *Proceedings of The Workshop on language, Artificial Intelligence, and Computer Science for Natural language Processing Applications (LAICS-NLP)*, 2006.

[7] De Vilar, JT Peter, and H NEY. Can we translate letters? *In Proceedings of ACL Workshop* on SMT, 2007.

[8] Naushad UzZaman, Arnab Zaheen, and Mumit Khan. A comprehensive Roman (English to Bangla) transliteration scheme. In *Proceedings of International Conference on Computer Processing on Bangla*, 2006.

[9] Mohammed Moshiul Hoque and Muhammad Masroor Ali -- A Parsing Methodology for Bangla Natural Language Sentences, published in Proceedings of International Conference on Computer and Information Technology 2003.

[10] Lenin Mehedy, S. M. Niaz Arifin and M Kaykobad -- Bangla Syntax Analysis: A Comprehensive Approach, published in Proceedings of International Conference on Computer and Information Technology 2003.

[11] Stuart J. Russel and Peter Norvig, Artificial Intelligence – A Modern Approach, Prentice-Hall of India, New Delhi, 2003.

[12] D. W. Patterson, Introduction to Artificial Intelligence and Expert Systems, PHI Private Limited, New Delhi, 2003

**Kanija Muntarina** completed her B.Sc (Hons) in Computer Science and Engineering from Jahangirnagar University in 2003 and M.S. in Computer Science and Engineering from the same University in 2005 respectively. Currently, she is a lecturer in the Dept. of Computer Science and Engineering, Dhaka City College, Dhaka, Bangladesh. Her research interests include Artificial Intelligence, Neural Networks, Computer Vision, Image Processing and so on.

**Md. Golam Moazzam** completed his B.Sc (Hons) in Electronics and Computer Science from Jahangirnagar University in 1997 and M.S. in Computer Science and Engineering from the same University in 2001 respectively. He is now an Associate Professor in the Dept. of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh. His research interests include Digital Image Processing, Artificial Intelligence, Computer Graphics, Neural Networks, Computer Vision and so on.

**Md. Al-Amin Bhuiyan** received his B.Sc (Hons) and M.Sc. in Applied Physics and Electronics from University of Dhaka, Dhaka, Bangladesh in 1987 and 1988, respectively. He got the Dr. Eng. degree in Electrical Engineering from Osaka City University, Japan, in 2001. He has completed his Postdoctoral in the Intelligent Systems from National Informatics Institute, Japan. He is now a Professor in the Dept. of CSE, Jahangirnagar University, Savar, Dhaka, Bangladesh. His main research interests include Image Face Recognition, Cognitive Science, Image Processing, Computer Graphics, Pattern Recognition, Neural Networks, Human-machine Interface, Artificial Intelligence, Robotics and so on.