A PROFICIENT AUTONOMOUS BANGLA SEMANTIC PARSER FOR NATURAL LANGUAGE PROCESSING

M. F. Mridha[1], Molla Rashied Hussein[1], Md. Musfiqur Rahaman[1] and Jugal Krishna Das[2]
[1]Department of Computer Science and Enginerring, University of Asia Pacific, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh
E-Mail: mdfirozm@yahoo.com

ABSTRACT

In this paper, we present an autonomous semantic parser for analyzing Bangla Language semantics applying Bangla WordNet. A rule based semantic parser engrosses in mapping of Natural Language (NL) Sentence into a complete formal meaning representation language. Semantic parsing augments the stratum of comprehension of natural language processing than the syntactic parsing, which is primarily indulges in untangling the syntactic ambiguities of the words. The rule based approach is used here to develop semantic parser rule as the approach uses the rules and a lexicon for ambiguity resolution of the words. The proposed parser distinguishes grammatical meaning of the words. Tests are carried out in a prearranged schedule on more than 2000 Sentences and Accuracy at different levels is measured.

Keywords: semantic parser, wordnet, lexicon, modification, ontology, syntactic analysis.

INTRODUCTION

According Semantic Parser is a textual analyzer which is constructed by a sequence of tokens to determine its grammatical structure as defined by a given formal grammar. Semantic parsing is the methodology to map a natural language input into a proper meaning rendition. Immerse semantic analysis provides the Sentence rendition in predicate logic or other formal languages which encloses supports for automated reasoning. Conventionally, semantic parsers were constructed manually but this is both fiscally debilitating and laggard in time. Semantic parser analyses the Sentence in terms of its meaning. This task requires access to representations that bond the linguistic elements riveted in the input to the non-linguistic knowledge of the world. Sentence meaning is compiled of entities and interactions between the entities. The fundamental steps for any natural language understanding application is the syntactic analysis. The semantic parser attempts to solve this problem and generates a syntax sovereign representation of Sentence meaning and releasing it from the stiff interpretations constructed by a syntactic analyzer.

Every lexicon uses some language of lexical descriptors to denote lexical information about each word or word sense. Although a language processing system's knowledge about a particular word may be quite far-reaching, often the bulkiness of the knowledge is about the meaning rather than the lexical constraints on the word. Therefore, lexical descriptions tend to be relatively small, and many words, though they may differ significantly in meaning, will have the similar lexical description. Parsing is the de-linearization of linguistic input. That is the use of grammatical rules and other knowledge sources to determine the functions of the words in the Sentence. The need for unambiguous representation has lead to a great endeavor in stochastic parsing. Bangla has a rich system of inflectional ending. The proposed methodology uses a rule based parser. It consists of analyzing process for language Sentence input both syntactically and semantically. Semantic parser rules will be devised with the help of rule based approach as the approach uses the rules and Bangla WordNet to untangle the words' ambiguity.

The organization of this paper is as follows: In Section 2, we describe the Literature Review, Section 3 has the short description about Semantic Parser, Section 4 depicts the design of Semantic Parser, where we discuss about the different approaches, for instance Rule-based, Parts of Speech (POS) Tagging, Predictive, XML and also techniques, such as Tokenizing, Ambiguity Resolution, Syntactic and Semantic Analysis; Section 5 demonstrates our Results. Finally, Section 6 draws the curtains by concluding amid some heeds towards our future work.

LITERATURE REVIEW

For Developing a Semantic parser for Bangla Language, first of all, a comprehensive review has been carried out on the theory of parsing [1,2,3] where Semantic parsing, Syntactic parsing, WordNet, Bangla WordNet, Lexicon, Ontology, Semantic Analysis and Syntactic Analysis were studied as these are the key factors for preparing Bangla Language Parser. Secondly, meticulous overview has been performed on the Bangla Grammar [4,5,6], Morphological Analysis [7,8,9], construction of Bangla Sentence [10] based on semantic structure. Last of all, a throughout inspection was accomplished over the Hybrid Mechanism of merging Rule-based and Statistical Technique into Hybrid Machine Translation (HMT) approach [11] which is suitable for automatically translating morphologically rich and syntactically different languages that abide the Subject Object Verb (SOV) order. Context Free Grammar (CFG) generates language structures and the errors occurred during translation is corrected by applying statistical methodology. Parts Of Speech (POS) tagger was used in that approach to comprehend each word's morphological attachment. Additionally, Predictive Parser and construction of the parse table [12] to recognize Bangla Grammar were summed up. Top down parsing method was understood and avoidance of CFG's left recursion by left factoring was reasoned. Idioms, Phrases and mixed Sentences were left out but increase and supplementary amendment of the production rule would diminish the constraints. Using these references, ideas about Bangla Grammar for morphological and semantic analysis were extracted in order to prepare Bangla WordNet and morphological rules.

Semantic Parser

Parsing is a process by which an input string is analyzed and assigned a suitable structure. The computation of the syntactic structure of a Sentence involves the grammar and the parsing technique. The grammar is a formal specification of the structures allowable in the language and the parsing technique is the method of analyzing an input Sentence to determine its structure as defined by the grammar. The parser takes input as a Bangla Sentence and using both the rule base and the lexical analyzer, analyses each word of the Sentence and returns the base form of each word along with their attributes. The information is analysed to get relations among the words in the Sentences using if then rules.

Design of Bangla Semantic Parser

Bangla Sentence is the input of our Semantic parser which generates a parse tree by using semantic relationship. A parser breaks data into smaller elements, according to a set of rules that describe its structure sequence of tokens to determine its grammatical structure as defined by a native formal grammar. The semantic representation provides a simple description of the grammatical relationships in a Sentence, which can easily be understood and effectively used by people without that specific language proficiency, who want to extract textual relations. The Sentence relationships are represented uniformly as semantic relations between pairs of words [13, 14, 15]. We have design our own lexical parser for getting the tag information and context-free structure grammar representation of source structure. The nouns, pronouns, verb, adverbs, adjectives, singular, plural, persons, tenses etc are stored in a Database.

Rule-based approach

In the rule-based approach, two components can usually be distinguished in an analyzer [10, 16]: a declarative component corresponding to linguistic knowledge and a procedural component which represent the analysis strategy. Linguistic knowledge includes the grammar and the lexicon of the language while analysis strategy is an algorithm which specifies in detail each of the operations involved in the process of analysis.

The rule based semantic parser is implemented in two steps:

1. Undergo syntactic parsing to transform into tokens.
2. The semantic rules assignment and application.

Rule based parsing proceeds by searching through the set of rules in a lexicon to determine which rules may jointly be applied to produce a well formed syntactic structure for a Sentence of the language described in the grammar. If no analysis for a Sentence can be found using the rules in the grammar, then the Sentence is ungrammatical. If more than one analysis is found, then the Sentence is syntactically ambiguous.

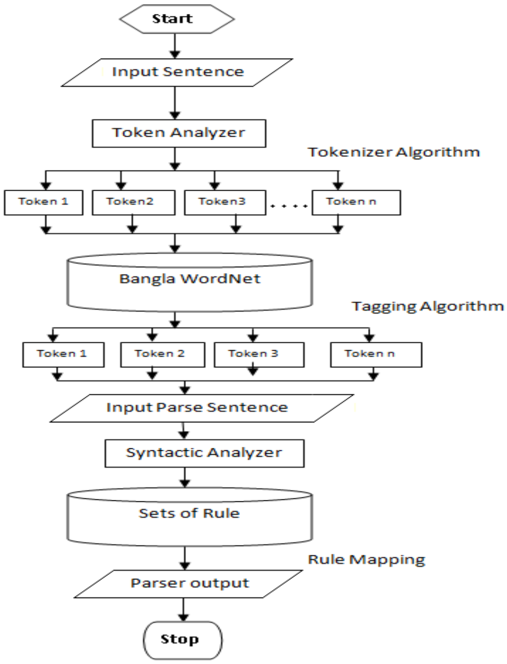
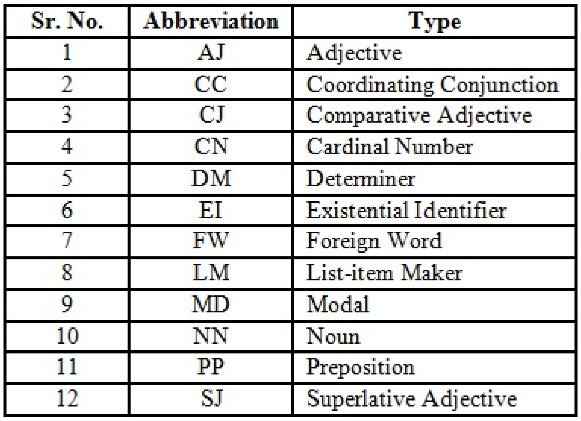


Figure-1. System architecture of Semantic parser.

POS Tagging Approach

POS tagger can be used to tag the input Sentences to supply particulars about the word's pronunciation and its morphological attachment. As the tag sets are usually language autonomous, an entire tag set of having 20-30 tags or more could be prepared for all languages. A paradigm [11] of POS tag set is illustrated in Table-1.

Table-1. POS Tag Set.



Sr. No.	Abbreviation	Type
1	AJ	Adjective
2	CC	Coordinating Conjunction
3	CJ	Comparative Adjective
4	CN	Cardinal Number
5	DM	Determiner
6	EI	Existential Identifier
7	FW	Foreign Word
8	LM	List-item Maker
9	MD	Modal
10	NN	Noun
11	PP	Preposition
12	SJ	Superlative Adjective

Predictive Approach

In this phase, Predictive approach is a novel way of executing recursive decent parsing with the help of managing the stack of activation record. Predictive parser initially has an input, then a stack, after that a parse table and finally receives the output. Tag set description for Bangla Grammar is demonstrated in Table-2.

Table 2. Tag Set Description for Bangla Grammar.

Tag Name (Symbol)	Example
Noun (NN)	বই, কলম
Pronoun (PN)	সে, তুমি, আমি
Adjective (AD)	অল্প, বেশি, ভালো, মন্দ
Verb (VB)	পড়া, হাঁটা
Conjunction (CN)	এবং, থেকে
Negative Description (ND)	নয়, নি, না
Modifier (MR)	একদিন, এ, একটি

XML Storage for Words

In this phase, currently, the eXtensible Markup Language (XML) is used ubiquitously for its minimalism. Any kind of data can be stored in this sophisticated format. If Bangla word is the tag_name and its corresponding POS is the value, then general format of XML tag having <tag_name>value</tag_name> would look like the following:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<WORD>
        <আমি>PN</আমি>
        <একটি>MR</একটি>
        <রুটি>NN</রুটি>
        <খাই>VB</খাই>
        <নি>ND</নি>
```

Figure-2. XML File's Data Format.

Tokenizing Algorithm for Bangla Sentence

1. Take input Sentence
2. Tokenize input Sentence.
3. Store all words into a Database.
4. Select each word singly from the Database.
5. Search in and compare selected word with Bangla WordNet.
6. If word is found one or more times, then store associated tag(s) of word into Database.
7. If one tag is stored in Database, then go to 9.
8. Select one or more linguistic rules and search most appropriate tag for a word by applying rules.
9. Display the word with associated tag as an output.

Ambiguity Resolution

If an unknown the rules considering the tags for surrounding words are used for resolving ambiguities at different levels. Before the step of ambiguity resolution, each word is attached with number of tags. Since a particular word may have number of tags, there is need to check which tag is applicable to a particular word in a Sentence, for example a word present in a noun list in Bangla WordNet, can be tagged with a noun as well as an adjective tag. For this purpose, there is need to apply certain rules depending upon the grammatical category of preceding or succeeding words. These rules are prioritized [17, 18]. First level of ambiguity exists when a particular word can have number of tags of different grammatical category. The rules should check the grammatical category for the surrounding words so that it can conclude the tag of that particular word, e.g. considering the two Sentences **"সে বই পড়ে।"** means "He reads a book." and **"সে হাঁটতে গিয়ে পড়ে গেল।"** means "He went for a walk and fell." Here, the same word **"পড়ে"** has two different meanings.

Syntactic and Semantic Analysis

Semantic parser works on the dependencies between the words identified after syntactic analyzer. The token is generated where target words and syntactic arguments are identified and which are then matched against the rules. The token also identifies the word level semantics and relations that have direct syntactic correspondence. It also identifies Sentence types like assertion, query, and command. Another important feature of analyzer is that it is based on feature of augmented grammar that has the ability of detecting the grammatically incorrect Sentences and subsequently grammatically incorrect Sentences will be rejected. **"আমি কলম খাই।"** in English is "I eat pen.", which is semantically wrong but syntactically correct.

Semantic Parsing Algorithm:

1. The output from *Tokenizing Algorithm* acts as input to the Semantic/Syntactic Analyzer.
2. The tokens generated from the *Tokenizing Algorithm* are stored in Database. These tokens have grammatical relations.
3. Look up in Rule base we are mapping the tokens with Rules in the Rule base. This matching is semantic (meaningful) and check the structure of the Sentence should be correct.
4. After matching the selected Sentence kept as another database.
5. Identify which rule can satisfy the condition.
6. If condition is satisfied then display the message "The Sentence is matched with defined rule".

RESULT ANALYSIS

In this paper, in this research work, we measure the efficiency of our Semantic parser system for Bangla Sentences. In this regard, Bangla Sentences are used for testing. The training corpus is gathered from Bangla WordNet.

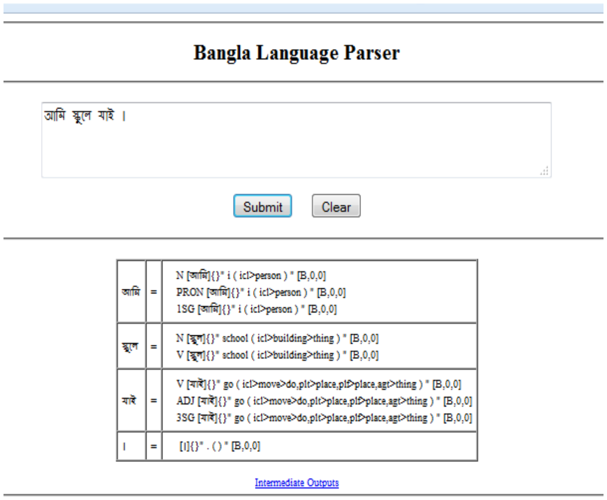
The corpora were manually tagged with our own tagset. More than 2000 simple Sentences were selected for testing which contains approximate 9250 tagged words, among them the frequency of noun is 4250, pronoun is 720, verb is 2350, adjective is 1630 and adverb is 300. During training and testing data, we have classified the words according to the part of speech and calculated the accuracy of each type of Sentences. The evaluated results are shown in Table-3.

Table-3. Numbers of categorized words.

Part of Speech	Total Words
Noun	4250
Pronoun	720
Verb	2350
Adverb	300
Adjective	1630

When more Sentences are tested and rules will be added, then accuracy will be increased. The GUI of Semantic is classified into three windows:
Submit Textbox
Clear
Intermediate output.

The outputs of tested Sentences are given below:

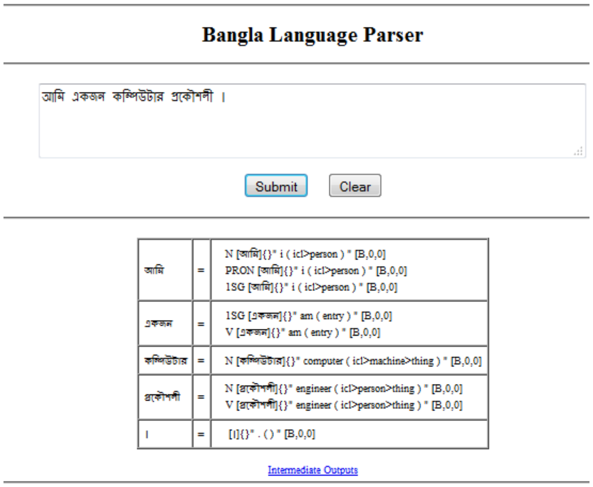


Bangla Language Parser

আমি স্কুলে যাই ।

আমি	=	N [আমি]{}" i (icl>person) " [B,0,0] PRON [আমি]{}" i (icl>person) " [B,0,0] 1SG [আমি]{}" i (icl>person) " [B,0,0]
স্কুলে	=	N [স্কুল]{}" school (icl>building>thing) " [B,0,0] V [স্কুল]{}" school (icl>building>thing) " [B,0,0]
যাই	=	V [যাই]{}" go (icl>move>do,plt>place,plf>place,agt>thing) " [B,0,0] ADJ [যাই]{}" go (icl>move>do,plt>place,plf>place,agt>thing) " [B,0,0] 3SG [যাই]{}" go (icl>move>do,plt>place,plf>place,agt>thing) " [B,0,0]
।	=	[।]{}" . ( ) " [B,0,0]

Figure-3. Parser Output (1).



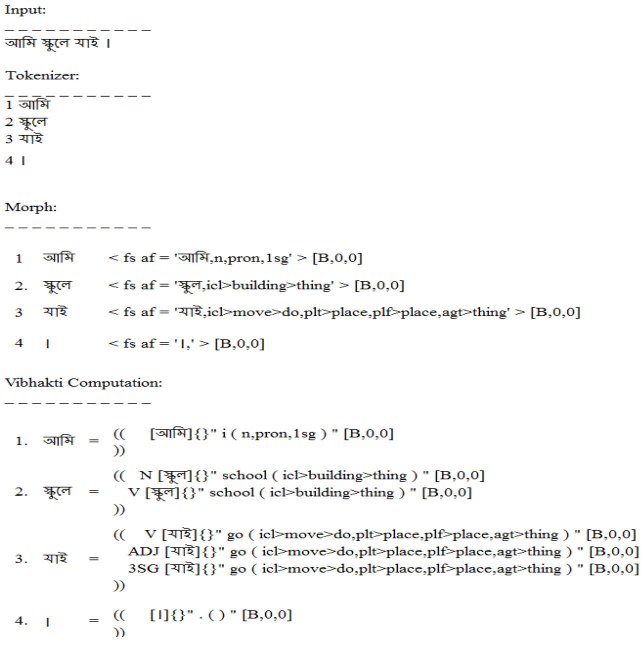
Bangla Language Parser

আমি একজন কম্পিউটার প্রকৌশলী ।

আমি	=	N [আমি]{}" i (icl>person) " [B,0,0] PRON [আমি]{}" i (icl>person) " [B,0,0] 1SG [আমি]{}" i (icl>person) " [B,0,0]
একজন	=	1SG [একজন]{}" am (entry) " [B,0,0] V [একজন]{}" am (entry) " [B,0,0]
কম্পিউটার	=	N [কম্পিউটার]{}" computer (icl>machine>thing) " [B,0,0]
প্রকৌশলী	=	N [প্রকৌশলী]{}" engineer (icl>person>thing) " [B,0,0] V [প্রকৌশলী]{}" engineer (icl>person>thing) " [B,0,0]
।	=	[।]{}" . ( ) " [B,0,0]

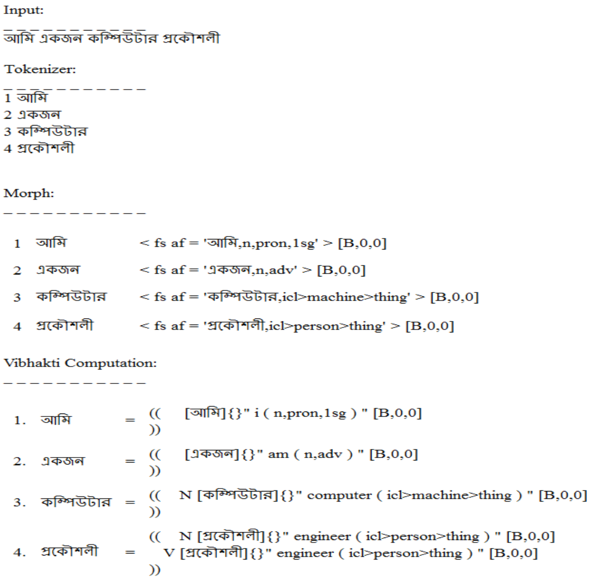
Figure-4. Parser Output (2).

The system read Bangla text; when terminator symbol is found then system compare each word of Sentence with lexicon entries. If Sentence is matched with the defined rule then information will display parsing output of the given Sentence. If the word found having multiple tags then the system search most appropriate rule for the word. By applying rule, system selects most appropriate tagged word according to the context of Sentence. If word is not found in the lexicon, then system will not display the information.



```
Input:
_ _ _ _ _ _ _ _ _ _ _
আমি স্কুলে যাই ।

Tokenizer:
_ _ _ _ _ _ _ _ _ _ _
1 আমি
2 স্কুলে
3 যাই
4 ।


Morph:
_ _ _ _ _ _ _ _ _ _ _

 1    আমি      < fs af = 'আমি,n,pron,1sg' > [B,0,0]

 2.   স্কুলে     < fs af = 'স্কুল,icl>building>thing' > [B,0,0]

 3    যাই      < fs af = 'যাই,icl>move>do,plt>place,plf>place,agt>thing' > [B,0,0]

 4    ।        < fs af = '।,' > [B,0,0]

Vibhakti Computation:
_ _ _ _ _ _ _ _ _ _ _

 1.  আমি   =  ((    [আমি]{}" i ( n,pron,1sg ) " [B,0,0]
              ))
              ((   N [স্কুল]{}" school ( icl>building>thing ) " [B,0,0]
 2.  স্কুলে  =     V [স্কুল]{}" school ( icl>building>thing ) " [B,0,0]
              ))
              ((    V [যাই]{}" go ( icl>move>do,plt>place,plf>place,agt>thing ) " [B,0,0]
 3.  যাই   =     ADJ [যাই]{}" go ( icl>move>do,plt>place,plf>place,agt>thing ) " [B,0,0]
                 3SG [যাই]{}" go ( icl>move>do,plt>place,plf>place,agt>thing ) " [B,0,0]
              ))

 4.  ।     =  ((    [।]{}" . ( ) " [B,0,0]
              ))
```

Figure-5. Intermediate Parser Output (1).



```
Input:
_ _ _ _ _ _ _ _ _ _ _ _
আমি একজন কম্পিউটার প্রকৌশলী

Tokenizer:
_ _ _ _ _ _ _ _ _ _ _ _
1 আমি
2 একজন
3 কম্পিউটার
4 প্রকৌশলী

Morph:
_ _ _ _ _ _ _ _ _ _ _ _

 1  আমি          < fs af = 'আমি,n,pron,1sg' > [B,0,0]

 2  একজন         < fs af = 'একজন,n,adv' > [B,0,0]

 3  কম্পিউটার     < fs af = 'কম্পিউটার,icl>machine>thing' > [B,0,0]

 4  প্রকৌশলী      < fs af = 'প্রকৌশলী,icl>person>thing' > [B,0,0]

Vibhakti Computation:
_ _ _ _ _ _ _ _ _ _ _ _

 1.  আমি        =  ((    [আমি]{}" i ( n,pron,1sg ) " [B,0,0]
                   ))

 2.  একজন       =  ((    [একজন]{}" am ( n,adv ) " [B,0,0]
                   ))

 3.  কম্পিউটার   =  ((   N [কম্পিউটার]{}" computer ( icl>machine>thing ) " [B,0,0]
                   ))
                   ((   N [প্রকৌশলী]{}" engineer ( icl>person>thing ) " [B,0,0]
 4.  প্রকৌশলী    =    V [প্রকৌশলী]{}" engineer ( icl>person>thing ) " [B,0,0]
                   ))
```

Figure-6. Intermediate Parser Output (2).

Table-4. Calculated Accuracy of Sentences.

Tested Sentences	Correct matched Sentences	Accuracy (%)
2000	1725	86.25%

Testing is performed with 2000 Sentences and accuracy at different levels is calculated. The first phase which resolves the ambiguity for different grammatical category and assigns tag to each word in a Sentence was found to have approximately 86.25% accuracy.

CONCLUSIONS

We have portrayed a rule based approach of semantic parsing for Bangla language. This approach is closer to the ways human surmise the language. This approach analyses the constituent of a Sentence and analyze the meaning of entire Sentence on the basis of smaller semantic units. This semantic parser relies on the rules salvaged from the Bangla WordNet. In this paper our developed Bangla semantic Parser shows how a Bangla Sentence is parsed into tokens and then the relationship between tokens is unearthed by using grammar and semantic representation, generating a parse tree. Rule-based approach is used to resolve the ambiguity of words. Tagging and tokenization algorithms were developed and implemented for Bangla Sentences. The accuracy of 86.25% was achieved from the Semantic Parser. During the experiment, it has been scrutinized that the accuracy was getting diminutive when we tested ambiguous words and Sentences furthermore.

REFERENCES

[1] Ji Luning, Lu Qin, Li Wenjie and Chen YiRong. 2007. "Automatic Construction of core Lexicon For Specific Domain", Sixth International Conference on Advanced Language Processing and Web Information Technology, pp.183-188.

[2] Akshar Bharati and Rajeev Sangal. 1993. "Parsing Free Word Order Languages in the Paninian Framework", ACL93: Proc. Of Annual Meeting of Association for Computational Linguistics, New Jersey, pp. 105-111.

[3] Gildea and D. Jurafsky. 2000. "Automatic labeling of semantic roles". In Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), pages 512520, Hong Kong, October.

[4] D. S. Rameswar. 1996. "Shadharan Vasha Biggan and Bangla Vasha", Pustok Biponi Prokashoni, November, pp.358-377

[5] Asad H. 1994. "Bakkotottyo", Second edition, Dhaka.

[6] D. C. Shuniti Kumar. 1999. "Bhasha-Prakash Bangala Vyakaran",Rupa and Company Prokashoni, Calcutta, July, pp.170-175.

[7] P. Abney. Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, "Principle-Based Parsing: Computation and Psycholinguistics", pages 257–278. Kluwer, Dordrecht.

[8] Bharati R. Sangal D.M. Sharma and L. Bai 2006. Anncorra. "Annotating corpora guidelines for pos and chunk annotation for Indian languages". LTRC-TR31.

[9] D.M. Shahidullah. "Bangla Baykaron", Ahmed Mahmudul Haque of Mowla Brothers prokashani, Dhaka-2003.

[10] Vijay Kumar, Pankaj K. Sengar. 2010. "Segmentation of Printed Text in Devanagari Script and Gurmukhi Script", International Journal of Computer Applications Vol. 3, No.8, pp. 0975 –8887.

[11] Sangeetha J., S. Jothilakshmi and Devendra Kumar. 2014. "An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism." International Journal of Engineering & Technology (0975-4024) 6.4.

[12] Hasan K. M., Al-Mahmud Amit Mondal and Amit Saha. 2011. "Recognizing Bangla Grammar using Predictive Parser." International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No. 6, Dec.

[13] Pawan Goyal, Vipu Arora and Laxmidhar Behera 2009. “Analysis of Sanskrit text: Parsing and Semantic Relation”, Springer-Verlag Berlin Heidaelberg, pp. 200-218.

[14] Ms Vaishali M. Barkade. *et al*. 2010. “English to Sanskrit Machine Translation Semantic Mapper”, International Journal of Engineering Science and Technology Vol.2, No. 10.

[15] Ms Vaishali M. Barkade. Prof. Prakash R. Devale, Dr.Suhas H. Patil. 2010. “English to Sanskrit Machine Translator Lexical Parser and Semantic Mapper”, National Conference On” Information and Communication Technology”(NCICT-10).

[16] Javed Ahmad Mahar, Ghulam Qadir MEMON. 2010. “Rule Based Part of Speech Tagging of Sindhi Language", International conference on Signal Acquisition and processing, pp.101-106.

[17] Khaled Shaalan 2010. “Rule-based Approach in Arabic Natural Language Processing”, International Journal on Information and Communication Technologies, Vol. 3, No. 3.

[18] Ji Luning, Lu Qin, Li Wenjie and Chen YiRong. 2007. “Automatic Construction of core Lexicon For Specific Domain”, Sixth International Conference on Advanced Language Processing and Web Information Technology, pp.183-188.