Transformational generative grammar (TGG): An efficient way of parsing Bangla sentences

Conference Paper · December 2016

DOI: 10.1109/ICECTE.2016.7879583

3 authors:

Mohammad Kamrul Huq Maroof
cantonment english school & college
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Lamia Alam
Chittagong University of Engineering & Technology
14 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Moshiul Hoque
Chittagong University of Engineering & Technology
72 PUBLICATIONS **210** CITATIONS
SEE PROFILE

International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)
8-10 December 2016, Rajshahi-6204, Bangladesh

Transformational Generative Grammar (TGG): An Efficient Way of Parsing Bangla Sentences

Mohammad Kamrul Huq Maroof, Lamia Alam and Mohammed Moshiul Hoque*
Dept. of Computer Science & Engineering
Chittagong University of Engineering & Technology,
Chittagong, Bangladesh.
Email: kamrulhuqmaroof@gmail.com, lamiacse09@gmail.com and moshiulh@yahoo.com*

***Abstract*— Natural language processing (NLP) refers to the ability of systems to process sentences in a natural language such as Bangla, rather than in a specialized artificial computer language. Computer processing of Bangla language is a challenging task due to its varieties of words formation and way of speaking. The same meaning can be expressed in different ways which is a great challenge to face for translation by an automatic machine translation system. With the advent of internet technology and e-commerce, the demand of automatic machine translation has been increased. Parsing is essential for any type of natural language processing. Parsing of Bangla natural language can be used as a subsystem for Bangla to another language machine aided translation. A parser usually checks the validity of a sentence using grammatical rule. In this paper, we propose a set of transformational generative grammar (TGG) in conjunction with phrase structure grammar to generate parse tree and to recognize assertive, interrogative, imperative, optative and exclamatory sentences of Bangla language. It is applicable for many sentences that cannot be parsed using only phrase structure grammars. The process involves analysis of Bangla sentence morphologically, syntactically where tokens and grammatical information are passed through parsing stage and finally output can be achieved. A dictionary of lexicon is used which contains some syntactic, semantic, and possibly some pragmatic information. We have tested our system for different kinds of Bangla sentences and experimental result reveals that the overall success rate of the proposed system is 84.4%.**

Keywords— natural language processing; machine translation; lexical analysis; syntactic analysis; transformational generative grammar; parse tree; lexicon.

I. INTRODUCTION

Bangla is the native language of millions of Bangladeshi people. Bangla is spoken by about 210 million people of Bangladesh and two states of India [1]. Today, most of the computer based resources and technical journals are in English. Due to the language barrier, the common masses face difficulties to obtain the optimum benefits out of the vast store house of information through internet. Bangla Language Processing (BLP) can help the way to remove this barrier. The success of BLP may have a huge impact particularly to learn and use information and communications technology (ICT) in Bangla for the common people. Without computerization of the language, it can't be achieved. Today internet communication is emerging rapidly. In this revolution, if Bangla language is not properly digitized, this language can't cope with the entire world.

Syntactic analysis is one of the most important phases in NLP. The result of syntactic analysis yields a syntactic representation in a grammar; this form is often displayed in a tree diagram or a particular way of writing it out as text. Sequences of words are transformed into structures that show how the words relate to each other in this analysis. It verifies the validity of a sentence of natural language. Syntactic analysis uses grammar rules such as Context-free grammars (CFG), Context-sensitive grammars (CSG), and Transformational generative grammar (TGG) 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 part of the sentence.

TGG as the name suggests that the grammar it provides is both transformational and generative. Transformation is a method of stating how the structures of many sentences in languages can be generated or explained formally as the result of specific transformations applied to certain basic sentence structures. The other characteristics of TGG are that it is generative. In other words, a grammar must generate all and only the grammatical sentences of a language. To generate is thus to predict what can be sentences of the language or to specify precisely what are the possible sentences of the language

The main contribution of this work is to develop an effective parser that can parse the five types of Bangla sentences such as, assertive, interrogative, imperative, optative and exclamatory sentences by proposing a set of TGG rules.

II. RELATED WORK

A significant number of parser was developed previously which can be mainly categorized into three categories such as CFG Based Parsing, CSG Based Parsing and TGG base Parsing.

978-1-5090-5785-6/16/$31.00 ©2016 IEEE

Some works have been done in syntax analysis Bangla simple sentences using CFG in [1-6]. In [1,2], a comprehensive approach for Bangla syntax analysis was developed where a formal language is defined as a set of strings. A parsing methodology for Bangla natural language sentences is proposed in [3] and shows how phrase structure rules can be implemented by top-down and bottom-up parsing approach to parse simple sentences of Bangla. In [4,5] a predictive parser was introduced which is effective for detecting a paragraph contains of almost all types of traditional and non -traditional Bangla sentences. The accuracy rate was in between 70-80%. In [6], English to Bangla machine translation (MT) system was built using context free grammar (CFG) which was able to translate only assertive sentence.

Quite a few works have been done using CSG. Among them, the most notable is a method to analyze syntactically Bangla sentence using context-sensitive grammar rules in [7,8]. The parser developed in [7] accepted almost all types of Bangla sentences, including simple, complex and compound sentences and then interpret the input Bangla sentence into English using the NLP conversion unit. But the limitation of that work was it couldn't handle the ambiguity of a sentence. In the work [8], 28 decomposition rules were used and the success rate was around 90%. It proposed a technique to parse Bangla sentences in a new approach using context-sensitive grammar rules and then convert them to English by NLP conversion unit. The principal goal was to design a parser that is capable of accepting all types of Bangla sentences, from the structural viewpoint. Arefin et. al proposed a set of CSG's to parse the Bangla sentences, including assertive, interrogative and imperative [8]. The proposed framework can parse Bangla of sentences with over 80% accuracy, but when sentence length increases the success rate decreases.

TGG based parsing is still a new concept to the natural language processing. Only a single work [9] relates to this type of parsing. In that work, they dealt with assertive sentence, interrogative sentence and imperative sentence. But in their work, they focused solely on the transformational aspects of TGG rather than handling variation of a sentence, besides, they didn't implement the parser for those TGG rules which confined their work to the theoretical concept only.

In [10] authors proposed a structure for detecting Bangla grammar which recognizes correct Bangla sentences and rejects incorrect ones based on Head-Driven Phrase Structure Grammar (HPSG). They used the Linguistic Knowledge Building (LKB) system for implementing their designed Bangla Grammar and it works only for simple assertive sentences.

In contrast to these, we proposed a set transformational generative grammar which is applicable for many usable sentences that cannot be parsed using only phrase structure grammars. In this paper, we developed a framework for the parser to parse five types of Bangla Sentence based on intonation i.e. assertive, interrogative, imperative, optative and exclamatory sentence. We considered simple sentences only in this paper.

III. PROPOSED SYSTEM

The proposed Bangla natural language parser is shown in fig. 1. The proposed parser module is capable of parsing of five types of Bangla sentence based on purpose, i.e. assertive, interrogative, imperative, optative and exclamatory sentences in syntactic pattern respectively. Details processing of this sentence by using this module is given in the following subsections.

A. Input sentence

Bangla sentences are taken as input for the proposed framework. In this system, only assertive, interrogative, imperative, optative and exclamatory sentences are considered as input for implementation.

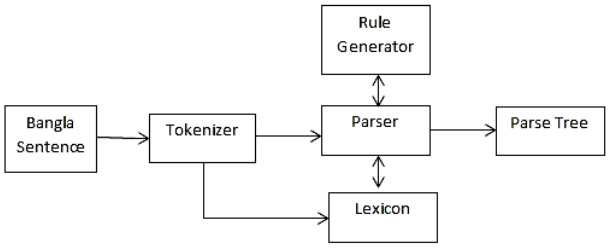

Fig. 1. Proposed Parser

B. Tokenizer

Tokenizer is the program module that accepts a sentence to be parsed as an unbroken string and breaks it into individual words called tokens. Tokens are stored in the list for further access. The token is then checked into the lexicon for the validity. From input sentence, tokenizer will generate tokens. For example:

Input Sentence: একটি মেয়ে বই পড়ছে

Output of tokenizer: ("এক", "টি", "মেয়ে", "বই", "পড়ছে")

The outputs of tokenizer are used as input of parsing to get the source language structure.

C. 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 is needed to help determine the function and meanings of the words in a sentence. The entries in a lexicon could be grouped and given by word category (by determiner, nouns, verbs, and so on), and all words contained within the lexicon listed within the categories to which they belong. Entries in a dictionary will be equivalent to collections of attributes and values (i.e. features).

D. Rule Generator

The rule generator generates a set of phrase structure (PS) rules that relate to the sentence structure. List of used rule of Bangla to parse different type of sentences are given in table I.

E. Parser

Parsing is a process of taking tokens of input sentence and producing a parse tree or structural representation (SR) according to TGG rules.

1) Parsing Algorithm

The input Bangla sentence is parsed by bottom-up parsing. The steps are given below:

Step-1: The input string is separated into tokens and tokens are stored in a list for further access. For example, Bangla sentence "একটি মেয়ে বই পড়ছে" will be tokenized in to "এক", "টি", "মেয়ে", "বই", and "পড়ছে".

TABLE I. LIST OF PS RULES

Rule No.	TGG Rules	Example
1.	S→NP VP	রহিম পড়ছে
2.	NP→N_{animal}/N_{object}/$N_{creator}$	ছেলে/বই/আল্লাহ
3.	NP→(Det) NP	একটি বল
4.	NP→NP (Det)(Biv)	বলটি,স্কুলে,লোকটি কে
5.	NP→NP Biv NP	আমার বই
6.	Det→(Qtfr) Art	দুইটি, টি
7.	NP→ AP NP (EM)	সুন্দর পাখি
8.	NP→Person	সে, আমি
9.	Person→FP/SP/TP	আমি/তুমি/সে
10.	SP→SPHF/SPNHF/SPP	তুমি/ আপনি/ তুই
11.	TP→TPHF/TPNHF	সে, তারা / তিনি, তাঁরা
12.	VP→(IW)(NP) VF	ফুটবল খেলে
13.	VF→VR AUX (IM)	খেয়েছ, খেয়েছ?"
14.	Aux→ Aspect Tense Con	খেয়েছিলাম
15.	Aspect→ Imperative/Indefinite/Continuous/Perfect	
16.	Tense→ Past/ Present/Future	করলাম, করেছিলাম,করব
17.	Con→Con-FP/Con-SPHF/Con-SPNHF/Con-SPP/Con-TPHF/Con-TPNHF	ই(করি),এন,উন(করেন/করুন),অ, ও (কর/করো)
18.	NP→Null	
19.	NP→ (Adj) NP	গভীর সমুদ্র, বিশাল আকাশ
20.	VP→VF VF	খেলতে □াও, খেতে আসো
21.	VF→ VR Biv	খেতে, খেয়ে, □তে, খেলে, খেলতে
22.	AP→Ad Ads	কী চমৎকার, আশ্চর্য সুন্দর

[Abbreviations: S: Sentence, NP: Noun phrase, N: Noun, PN: Pronoun, VP: Verb phrase, VF: Verb form, V: Verb, Qntfr: Quantifier, PP: Preposition, Biv: Bivokti (inflection), AP: Adjective phrase, Adj: Adjective, Con: Concord, Aux: Auxiliary, Det: Determiner, Gr: Gerund, IW: Interrogative word, IM:Interrogative marker, FP: First Person SP: Second Person, TP: Third Person, SPNHF: Second person non-honorific , SPHF: Second Person Honorific, SPP: Second person pejorative, TPHF: Third person non-honorific , TFNHF: Third person honorific]

Step-2: The tokens are checked in the lexicon for the validity. Tokens ("এক", "টি", "মেয়ে", "বই", "পড়ছে") will be checked for validity

Step-3: The tokens are matched with the grammar. If a rule whose right hand side matches with a token, then the token is assigned with the appropriate category. This is equivalent to looking up the words in Bangla dictionary. For example, Qtfr->এক, Art->টি, TPNHF ->মেয়ে, N->বই, VR->পড়, Biv->ছে will produce a partial structure.

Step-4: Starting from left to right hand side of token list, check every rule whose right hand side will match one or more of the parts of speech. If a right hand side of a rule matches with appropriate parts of speech, then we have to select that rule.

Step-5: Finally, generate the corresponding parse tree.

The parser output for the example will be: NP[Det[Qtfr[এক]Art[টি]]N[Person[TP[TPNHF[মেয়ে]]]]VP[NP[N[বই]Biv[]]VF[V[পড়]Aux[Aspect[Cotinuous[ছ] Tense[Present[] Con[Con-TPNHF[এ]]]]]. Figure 2 illustrates the corresponding parse tree.

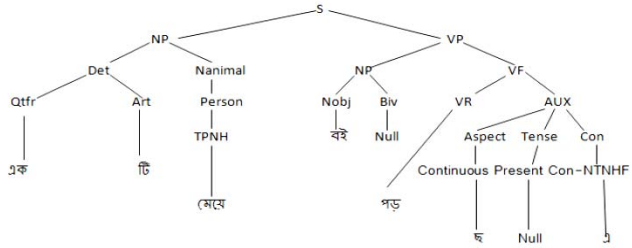

Fig. 2. Parse tree of "একটি মেয়ে বই পড়ছে"

F. Parsing Sentences with TGG

In this paper, we have used PS rules to represent deep structure of a sentence. Some transformational rules (TR) rule is applied over the deep structure(DS) to yield surface structure(SS) of other sentence. List of some used TR to parse different type of Bangla sentences are given in table II.

For example,let us consider a assertive sentence of ambigus SS:

i) একটি মেয়ে বই পড়ছে
ii) মেয়ে একটি বই পড়ছে
iii) মেয়েটি বই পড়ছে

TABLE II. LIST OF TRANSFORMATIONAL RULES (TR)

Type of Sentence	Transformational Rule
Assertive Sentence	i. #Qtfr+Art+N_{animal}#→ #N_{animal}+Qtfr+Art# ii. #N_{animal}+Qtfr+Art#□#N_{animal}+Art# Condition: if Qtfr=এক
Interrogative Sentence	iii.#Person+IW+NP+VF+IM#→#Person+NP+VF+IM# iv. #Person+NP+VF+IM#→#NP+VF+IM# Condition: Person=SP v. #Person+IW+NP+IM# → #NP+IW+Person+IM#
Imperative Sentence	vi. #Person+N+Biv+VF#→#N+Biv+VF# Condition: Person=SP
Optative Sentence	PS rules are adequate, No TR required
Exclamatory Sentence	vii. #N+Art+AP+EM→AP+N+EM#

[a] # (hash) symbol represents the end of string.

All these three sentences have the same deep structure. Among these three, the elementary sentence is "একটি মেয়ে বই পড়ছে",which is parsed using the PS rules described in table I and fig. 2 illustrates the parse tree.

Our system gnerates the parese tree for for sentence (ii) and (iii) using TR(i) and(ii) described in table II.

Applying TR (i) structure of "একটি মেয়ে" is transformed to structure of "মেয়ে একটি" which is shown in fig. 3.

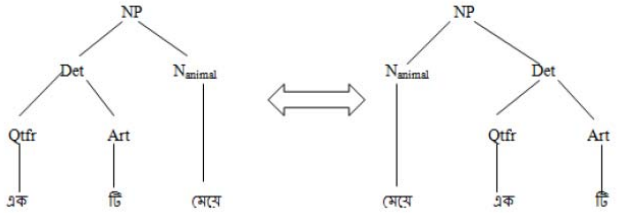

Fig. 3. Position Interchange Transformation applied in sample sentence

Applying TR (ii) structure of is transformed to structure of which is shown in fig. 4.

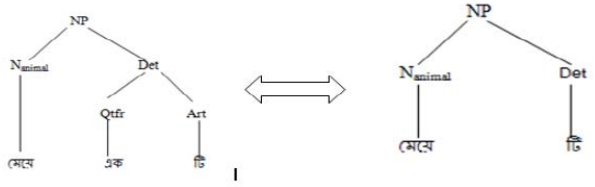

Fig. 4. Subractive Transformation applied in sample sentence

The transformation yields parse tree for sentence (ii) and (iii) which are shown in figure 5 and 6 respectively.

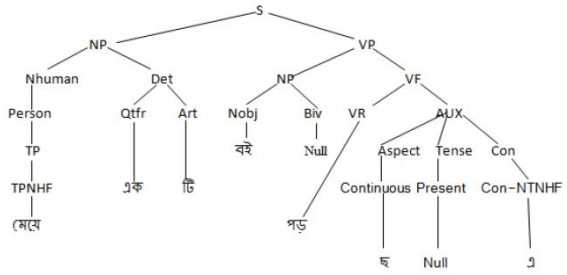

Fig. 5. Parse tree for "মেয়ে একটি বই পড়ছে"

IV. IMPLEMENTATION

We developed this system to parse for assertive, interrogative, imperative, optative and exclamatory sentences. Fig. 7 represent the implementation snapshots of our proposed parser for assertive sentence. The parser detects the sentence type and it shows the syntactic structure of input sentence as list of tokens. The PS rules used to generate the list of token is shown in PS rules field. Applied TR is shown in TR field.

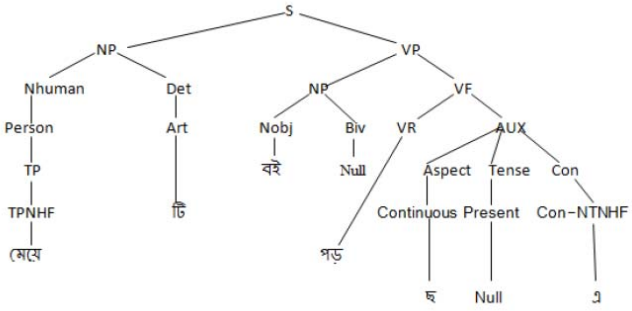

Fig. 6. Parse tree for" মেয়েটি বই পড়ছে"

V. EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of our proposed system, we have tested the system for about 770 different kinds. These sentences were collected from Bangla grammar book [10] and the book in reference [11]. We measure the overall success rate of the proposed framework and expressed in terms of percentage.

A. Success Rate

From our analysis, we have found that total 650 of generated outputs are correct among them about 210 sentences are assertive, 190 sentences are interrogative, 80 sentences are imperative, 79 sentences are optative and 100 sentences are exclamatory. Success rate denotes the ratio between total no. of correctly parsed sentences and total no of input sentences. We have calculated success rate for

assertive, interrogative, imperative, optative and exclamatory sentences separately and finally we have calculated the overall success rate of our system. The success rate for assertive, interrogative, imperative, optative sentences and exclamatory sentences are 84%, 90.5%, 80%, 87.5% and 77% respectively. Overall success rate of proposed system is 84.4%. Table III illustrates the success rate of different types of sentence. From the table, we can see that our proposed system parses Bangla interrogative sentence most successfully, whereas parsing exclamatory sentence has the least success rate. This is due to the fact that we used a limited number of PS rules and TRs' listed in table II, which are not able to reflect the linguistic competence and knowledge of a native speaker. In order to increase the success rate, we need to introduce more PS rules and TRs'. We are optimistic about future research in this regard.

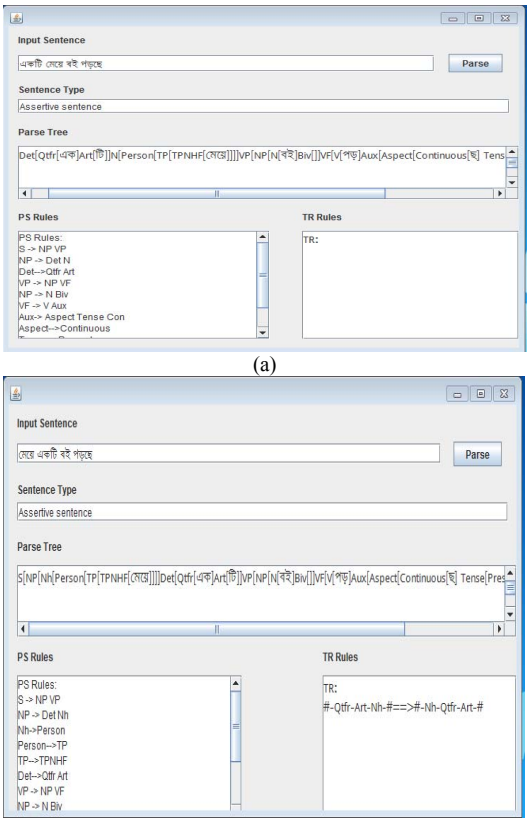

Fig. 7. Parsing of of the assertive sentence (a)"একটি মেয়ে বই পড়ছে" and (b) "মেয়ে একটি বই পড়ছে"

TABLE III. SUCCESS RATE OF DIFFERENT SENTENCE TYPES

Type of Sentence	Total no. of Sentence	Correctly performed syntax analysis	Success Rate (%)	Overall Success Rate (%)
Assertive	250	210	84	84.4
Interrogative	210	190	90.5	
Imperative	100	80	80	
Optative	80	70	87.5	
Exclamatory	130	100	77	

VI. Discussion

All sentences of a language cannot be generated using only PS rules, that's why we proposed a set transformational generative grammar. Let's take a Bangla interrogative sentence "তুমি কি ভাত খেয়েছ?": a parser can detect the sentence using PS rules as an interrogative sentence, but sentences such as "তুমি ভাত খেয়েছ?" and "ভাত খেয়েছ?" are not possible to be detected let alone using PS rules.

VII. Conclusion

Use of TGG in BLP is still at its rudimentary stage. Our implemented system can parse those sentences which was not possible to parse by PS rules only. It can analyze assertive, interrogative, imperative, optative & exclamatory sentences and generate parse tree for those input sentences efficiently. It can also detect the sentence category to which it belongs. It can parse those sentences having identical meaning, but of different surface structure. We have tested a good amount of sentence in our system and the testing result indicates, 84.4% of success rate.Our system mainly deals with parsing structurally simple sentences. Parsing of complex and compound sentence are left as future issues.

Reference

[1] Bengali language [Online]. Available: https://global.britannica.com/topic/Bengali-language

[2] L. Mehedy, S. M. Arefin and M. Kaykobad, "Bangla Syntax Analysis: A Comprehensive Approach", in Proc. *International Conference on Computer and Information Technology* (*ICCIT'03)*, Dhaka, Bangladesh, vol. 5, pp. 287-293, 2003.

[3] M. S. Islam, "*Research on Bangla Language Processing in Bangladesh: Progress and Challenges*", in Proc. 8th International Language & Development Conference, Dhaka, Bangladesh,pp. 527-533, 2009.

[4] M. M. Hoque and M. M. Ali, "A Parsing Methodology for Bangla Natural Language Sentences", in Proc. *International Conference on Computer and Information Technology* (*ICCIT'03)*, Dhaka, Bangladesh, vol. 2, pp. 277-282, 2003.

[5] K. M. A. Hasan, Al-Mahmud, A. Mondal and A.Saha, "*Recognizing Bangla Grammar using predictive parser*", International Journal of Computer Science and Network Security (IJCSNS), Vol. 3, No. 6, ,pp.61-73, Dec. 2011.

[6] K. M. A. Hasan, A. Mondal, A.Saha, "A Context Free Grammar and its Predictive Parser for Bangla Grammar Recognition", in Proc. *13th International Conference on Computer and Information Technology (ICCIT '10)*, Dhaka, Bangladesh,pp.87-95,2010.

[7] S. S. Ashrafi, M. H. Kabir, M. M. Anwar and A. K. M. Noman, *"English to Bangla Machine Translation System Using Context-Free Grammars"*, International Journal of Computer Science and Network Security(IJCSNS), Vol.10, Issue 3, No. 2, pp. 144-153 May 2013.

[8] M. M. Hoque and M. M. Ali, "Context-sensitive Phrase Structure rules for Structural Representation of Bangla Natural Language Sentences ", in Proc. Int. Conf. on Computer and Information Technology (ICCIT'04), pp. 615-620, 2004.

[9] M. S. Arefin, L. Alam, S. Sharmin and M. M. Hoque, "An empirical framework for parsing Bangla assertive, interrogative and imperative sentences," *in Proc. International Conference on Computer and Information Engineering (ICCIE'15)*, Rajshahi, 2015, pp. 122-125.

[10] M. M. Anwar, M. Z. Anwar and Md. A. Bhuiyan, "*Syntax Analysis and Machine Translation of Bangla Sentences*", International Journal

of Computer Science and Network Security (IJCSNS), Vol.9, Issue 8, pp. 317-326, Aug. 2009..

[11] M. A. Islam, K. M. A. Hasan and M. M. Rahman, "Basic HPSG structure for Bangla grammar," in Proc. *15th International Conference on Computer and Information Technology (ICCIT'12),* Chittagong, pp. 185-189, 2012.

[12] M. R. Selim and M. Z. Iqbal, "*Transformational Generative Grammar for Various Types of Bengali Sentences*", SUST Studies, Vol. 12, No. 1, pp. 99-105, 2010.

[13] বাংলা ভাষার ব্যাকরণ, জাতীয় শিক্ষাক্রম ও পাঠ্যপুস্তক বোর্ড, ঢাকা।

[14] H. Azad, "*Bakkyatotto* (বাক্যতত্ত্ব)", Dhaka University, Dhaka, Second Edition, 1994