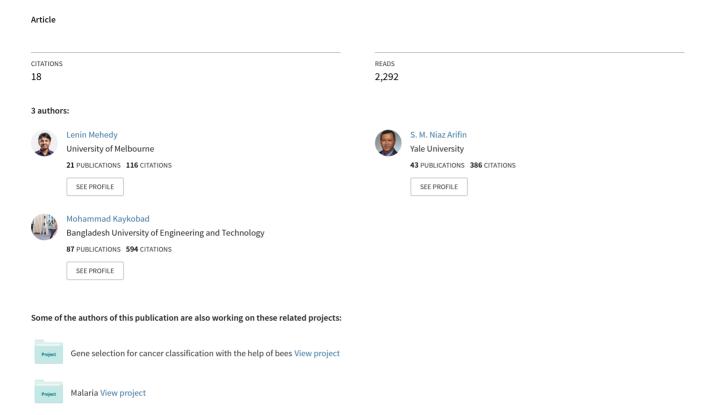
Bangla Syntax Analysis: A Comprehensive Approach



Bangla Syntax Analysis: A Comprehensive Approach

Lenin Mehedy, S. M. Niaz Arifin and M Kaykobad

Department of Computer Science and Engineering,

Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh.

Emails: leninmehedy@yahoo.com, niazarifin@yahoo.com, kaykobad@cse.buet.ac.bd

Abstract: A few researches have been carried out to efficiently VP may be well suited for simple sentences like " সে নিয়মিত recognize natural Bangla sentences. This paper proposes a technique to parse Bangla sentences in a new approach using context-free grammar rules that accepts all types of Bangla sentences including complex, compound, exclamatory and optative sentences. The proposed grammar also allows parsing all five categories of sentences according to Bangla intonation. As the inflection of verb plays a very important role in Bangla parsing, special attention has been paid in decomposing the verb and then extracting the information from the inflection.

1. Introduction

Natural language processing is a very challenging field of modern computer science because of the versatile characteristics of the concerned language, apart from the perspective of the machine translator itself. In particular, MT [3][4][6] is one of the most promising applications of NLP [6][7]. NLP allows people to interact with computers in a natural human language such as Bangla.

Though the computerization of Bangla is an inevitable need, only a few researches [2][5][8] have been made to efficiently recognize natural Bangla sentences. This paper proposes a technique to parse Bangla sentences in a new approach using context-free grammar rules that accept all types of Bangla sentences.

Section 2 describes the previous works in this field and section 3 gives a diagrammatic model of our proposed parser. Section 4 contains a brief review of the basic Bangla grammar [10]. Section 5 lists the first set of grammar rules for the parser, followed by examples of sentence parsing according to the structure. Section 6 lists the second set of rules, followed by examples of parsing according to the intonation of verb. Some particular areas have been pointed out in section 7 for future research.

2. Past Advancements

Murshed [5] proposed a set of grammar rules which starts parsing by breaking a sentence into noun phrase and verb phrase. Sharaf [7] depicted the relation between object-oriented concepts and NLP systems that can be introduced for Bangla. Mortuza and Ali [2] suggested a way to develop an MT dictionary, which can be a useful tool in the lexical analysis phase.

Selim and Iqbal [8] described a way of syntax analysis of different types of Bangla sentences with the help of a transformational generative grammar. They also showed ways to parse different types of Bangla sentences and proposed an algorithm for the parser.

3. The Proposed Model

It has been observed that recent propositions to build a Bangla parser do not cover the wide category of ways to form Bangla sentences. For example, the breaking of a sentence into NP and

লেখাপড়া করে [Se niomito lekhapora kore / He studies regularly]", but sentences that are complex or compound cannot be parsed by this grammar [5][8]. This is why we have proposed the starting rule to parse all types of sentences, including complex, compound, exclamatory and optative sentences, which were not considered before.

Secondly, according to Bangla intonation(স্বরভঞ্চি)[10], sentences can be divided into five broad categories as given in section 4. The proposed grammar allows parsing all these types in a new approach. Examples of parse trees of all these types have also been provided in sections 5 and 6.

Thirdly, the Bangla grammar has an excellent inherent property in forming the verbs, that is, unlike the English grammar, various necessary information of a sentence such as the tense, the person, the mode of verb (ক্রিয়ার ভাব) etc. can be extracted from a finite verb. Previous works did so by decomposing the verb phrase [5][8]. As we noted that the inflection of verb (ক্রিয়া-বিভক্তি) plays a very important role in this regard, the proposed grammar has paid special attention in decomposing the verb and then extracting the information.

The parser should accept all Bangla sentences that are syntactically correct, parse each of them using the rules of the proposed grammar, or report an error otherwise. Figure 1 shows the model.

The analysis is divided into 3 phases:

- 1. Lexical analysis phase: In this phase the stream of characters are sequentially scanned and grouped into tokens or
- 2. Syntax analysis phase: The parser is the most important tool of this phase. To ensure its validity within the underlying grammar, every sentence must be checked by the parser. The lexicons having a collective meaning are grouped together.

The parser involves grouping of tokens into grammatical phrases that are later used to synthesize the output. Usually, the phrases are represented by a parse tree that depicts the syntactic structure of the input. Some examples are given in section 4 and

3. Semantic analysis phase: This is the last phase of analysis where certain checks are made to ensure that the discrete input components fit together meaningfully. This phase is highly application-dependent and is regulated by the norms and rules of the concerned natural language.

This paper focuses on the formation and use of the grammar rules to be used by the parser in the syntax analysis phase.

4. Bangla Grammar Review

Structurally, there are three types of Bangla sentences:

- a. Simple Sentence (সরল বাক্য),
- b. Complex Sentence (জটিল বাক্য),
- c. Compound Sentence (যৌগিক বাক্য).

Each of these has been defined using clause. Clause Now, a simple sentence can have two parts: (খভবাক্য) [10] is the subpart of a Bangla sentence that has a meaning. There are two types of Bangla clause:

- Principal clause (প্রধান খন্ডবাক্য),
- Subordinate clause (আগ্রিত খন্ডবাক্য)

A simple sentence is formed by an independent clause or **Predicate:** There are two types of predicate: principal clause. Example: শাহেদ স্কুলে যায়

A *complex sentence* is formed by one principal clause and one or more subordinate clause(s).

Example: যদি বৃষ্টি হয় তাহলে যেও না

A compound sentence is formed by two or more principal clauses joined by an indeclinable (অব্যয় পদ). Example: সূর্য ওঠে এবং কাজ শুরু হয়

a. Subject (উন্দেশ্য), b. Predicate (বিধেয়)

Subject: There are two types of subject:

- Simple Subject (সরল উদ্দেশ্য),
- Expanded Subject (সম্প্রসাবিত উদ্দেশ্য)

- Simple Predicate (সরল বিধেয়),
- Expanded Predicate (সম্প্রসারিত বিধেয়)

From this structural point of view we can develop the following rules to parse any type of sentences. Examples are selected so that they follow the most general syntax of Bangla sentences.

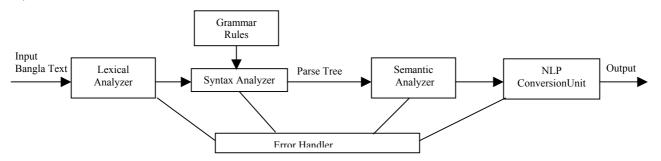


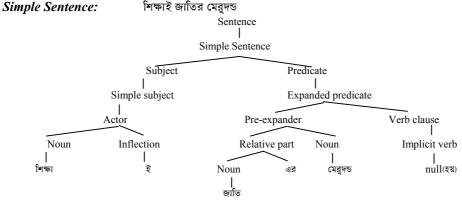
Figure 1: Block diagram of the proposed model

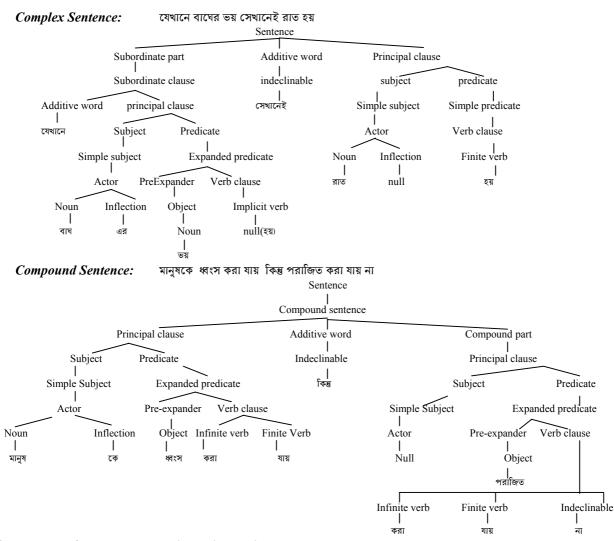
5. Set 1: Basic rules to parse a sentence*

- 1. Sentence -> Simple sentence | Complex sentence | Compound sentence:
- **2.** Simple sentence -> Principle clause;
- 3. Complex sentence -> Subordinate part + Additive word + Principal clause | Principal clause + Additive word + Subordinate part, 4. Subordinate part -> Subordinate clause | Subordinate clause + Additive word + Subordinate part,
- **5.** Subordinate clause -> Additive word + Principal clause;
- **6.** Additive word -> Indeclinable | Null; **7.** Compound sentence -> Principal clause + Additive word + Compound part;
- **8.** Compound part -> Principal clause | Compound sentence; **9.** Principal clause -> Subject + Predicate;
- **10.** Subject -> Simple subject | Expanded subject;
- 11. Predicate -> Simple predicate | Expanded predicate; 12. Simple Subject -> Actor (কর্তৃপদ);
- 13. Actor -> Noun + Inflection | Pronoun + Inflection | Implicit (উহা) Actor; 14. Pronoun -> Person ; 15. Person -> FP | SP | TP; 16. FP -> aami | aamraa; 17. SP -> SPH | SPNH | SPP; 18. SPH-> aapni | aapnaara; **19.** SPNH -> tumi | tomraa;
- **20.** SPP -> tui | toraa; **21.** TP -> TPH | TPNH; **22.** TPH -> tini | taaraa; 23. TPNH -> shey | taaraa; 24. Implicit Actor -> Null;

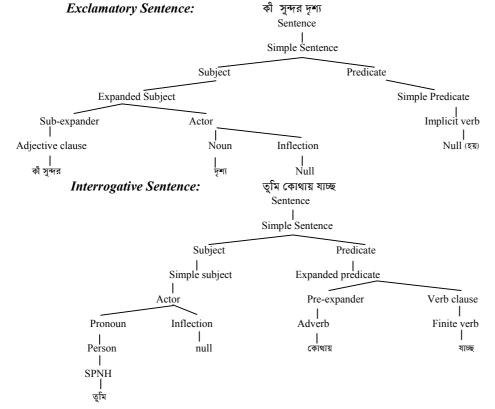
- **25.** Expanded Subject -> Sub-expander + Subject;
- **26.** Sub-expander -> Adjective | Adjective + Infinite verb Adjective clause | Relative part (সম্বন্ধ পদ/পদসমষ্টি) | Relative part + Adjective | Adverbial clause;
- 27. Relative part -> Noun + এর (er) | Pronoun + এর | Adjective + এর , 28. Simple predicate -> Verb clause | Implicit verb;
- 29. Implicit verb -> Null; 30. Expanded predicate -> Preexpander + Verb clause;
- 31. Pre-expander -> Adverb | Adverb + Adverb | Adverb + Object (কর্মপদ) | Adjective + Object | Adjective expander (বিশেষণের বিশেষণ) + Adjective + Object | Object | Adverbial clause;
- 32. Object -> Noun | Pronoun | Relative part + Noun | Relative part + Pronoun | Null:
- 33. Verb clause -> Infinite verb + Finite verb | Finite verb | Implicit verb | Infinite verb + Finite verb + Indeclinable | Finite verb + Indeclinable(অব্যয় পদ);
- 34. Indeclinable -> ना (na) | Other;
- * The '->' sign means the phrase "can have the form of", the '|' sign indicates an alternative rule for the left-side term and the '+' sign means *join* of two terms of a sentence.

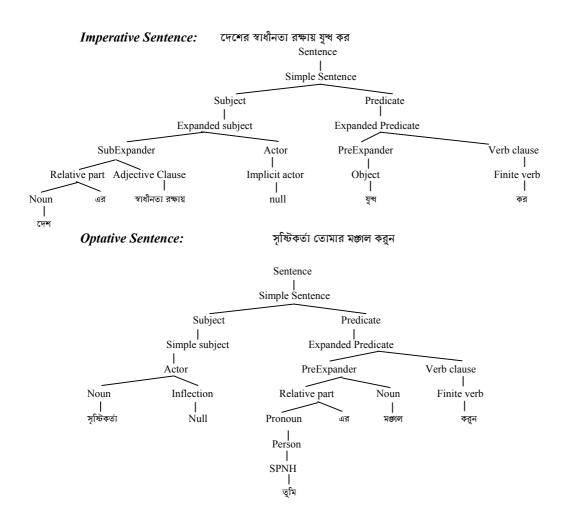
5.1 Examples of sentences according to structure:





5.2 Examples of sentences according to intonation:





As mentioned earlier, depending on intonation (স্বৰভঞ্জি), Bangla sentences can be divided into five categories:

- a. Assertive Sentence (বিবৃতিমূলক বাক্য),
- b. Interrogative Sentenc(প্রশ্নুসূচক বাক্য),
- Exclamatory Sentence (বিস্ময়সূচক বাক্য), c.
- d. Optative Sentence (ইচ্ছাসূচক বাক্য),
- e. Imperative Sentence (আদেশ বাচক বাক্য).

By the rules given in set 1, we can parse all of them. Some examples are shown below.

The first example of a simple sentence also serves as an example of an assertive sentence.

6. Parsing of Bangla Finite Verbs

Murshed [4] proposed parsing method for different forms of Bangla present tense. We extend the set by proposing methods for all other types. The types of Bangla tense are shown in figure2.

Murshed [5] used a name 'AUX' for the inflection of Bangla verb (ক্রিয়া-বিভক্তি). In fact, the inflection of Bangla verb can have different forms depending on the tense, the person and the class of subject of the verb. The forms are described in table 1.

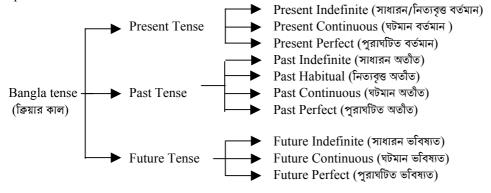


Figure 2: Types of Bangla tense

After a thorough observation of table 1, we propose the second Set 2: Rules to decompose finite verbs set of grammar rules to further decompose and extract information from the finite verb of a sentence. Three examples of parse trees involving less frequently used tenses are also follow to show the decomposition process.

- **35.** Finite verb -> VR + AUX;
- 36. VR -> কর্ (kor) | যা (ja) | Other verb roots(ক্রিয়ামূল)
- **37.** AUX -> Aspect + Tense + Concord;
- 38. Tense -> Present | Past | Future

- **39.** Aspect -> Indefinite | Continuous | Perfect | Imperative | Habitual
- **40.** Concord -> NonImperativeCon | ImperativeCon;
- 41. Present -> null; 42. Past -> ল (1) | ইল (il)
- 43. Future -> ব (b) | বো (bo); 44. Indefinite -> Null;
- 45. Continuous -> ছ (chh) | চছ (chchh)
- 46. Perfect -> এছ (echh); 47. Imperative -> Null;
- 48. Habitual -> ত (t)
- **49.** NonImperativeCon -> NIConFP | NIConSPH | NIConSPNH | NIConSPP | NIConTPH | NIConTPNH
- **50.** ImperativeCon -> IConSPH | IConSPNH | IConSPP | IConTPH | IConTPNH
- 51. NIConFP -> ই (i) | আম (am) | Null | ও (o);
- 52. NIConSPH -> এন (en); 53. NIConSPNH -> Null | এ (e)
- 54. NIConSPP -> ইম্ (ish) | ই (i); 55. NIConTPH -> এন (en); 56. NIConTPNH-> এ (e) | Null
- 57. IConSPH -> উন (un) | এন (en); 58. IConSPNH -> ও (o);
- **59.** IConSPP -> ইস্ (ish) | Null
- 60. IConTPH -> এন (en) | উন (un); 61. IConTPNH-> এ (e) | উক (uk); 62. Infinite Verb -> Verb root + Inflection.

Present Imperative Tense: আক্রান্ত হলে রুখে দাঁড়াও

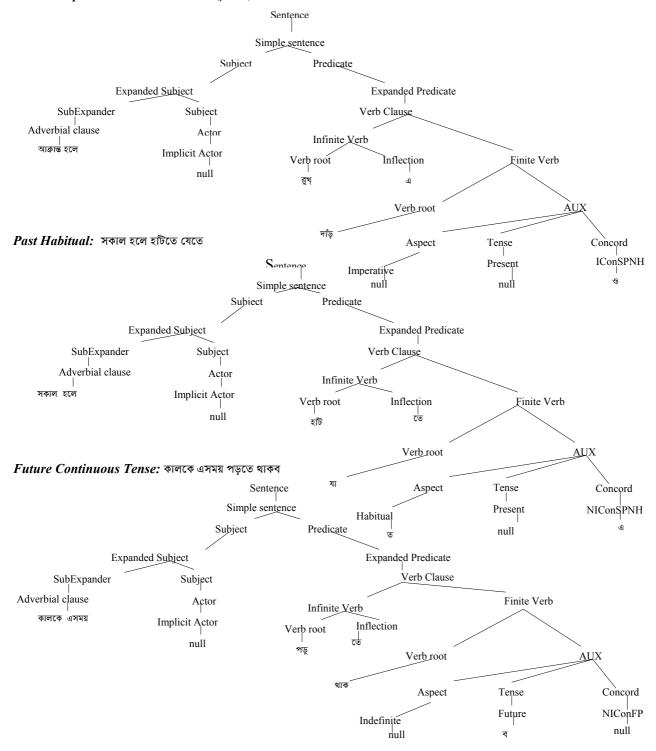


Table 1: Forms of Inflection of Verb

Tense	FP	SPH	SPNH	SPP	ТРН	TPNH
	(আমি , আমরা)	(আপনি , আপনারা)	(তুমি , তোমরা)	(তুই, তোরা)	(তিনি, তাঁরা)	(সে, তারা)
PrInd PrCon	ই (i)	এন (en)	অ (o)	ইস্ (ish)	এন (en)	ه (e)
	কর্ + ই = করি	কর্ + এন = করেন Kor + en =koren	কর্ + অ = কর Kor + o = koro	কর্ + ইস্ = করিস্	কর্ + এন = করেন	কর্ + এ = করে Kor + e =kore
	Kor + i = kori ছ/চছ + ই = ছি/চিছ	Kor + en = koren ছ/চছ + এন = ছেন/চেছন	ছ/চছ + অ = ছ/চছ	Kor + ish= Korish ছ/চছ + ইস = ছিস/চিছস	Kor + en =koren ছ/চ্ছ + এন =	ह/ह्य + व =
FFCon	(chhi/chchhi)	(chhen/chchhen)	(chho/chchho)	(chhish/chchhish)	ছেন/চেছন	ছেন/চেছ
	কর্ + ছি = করছি	কর্ + ছেন = কর্ছেন	কর্ + ছ = করছ	কর্ + ছিস্ = করছিস্	(chhen/chchhen)	(chhe/chchhe)
	Kor + chhi =	Kor + chhen= Korchhen	Kor + chho = Korchho	Kor + chhish = Korchhish	কর্ + ছেন = কর্ছেন Kor + chhen=	কর্ + ছে = কর্ছে
	Korchhi			Korennisn	Korchhen	Kor + chhe= Korchhe
PrPer	এছ/এচছ + ই =	এছ/এচছ + এন =এছেন/এচেছন	এছ/এচ্ছ +অ = এছ/এচ্ছ	এছ/এচ্ছ + ইস্ =	এছ/এচছ + এন	এছ/এচছ + এ
	এছি/এচিছ	(echhen/echchhen)	(echho/echchho)	এছিস্/এচিছস্	=এছেন/এচেছন	=লক্ষে/লচ্ছে
	(echhi/echchhi) কর্ + এছি = করেছি	কর্ +এছেন = করেছেন Kor + echhen=	কর্ + এছ = করেছ Kor + echho = Korechho	(echhish/echchhish) কর্ + এছিস্ = করেছিস্	(echhen/echchhen) কর্ +এছেন = করেছেন	(echhe/echchhe) কর্ +এছে = করেছে
	Kor + echhi =	Korechhen	Koi + ecinio - Korecinio	Kor + echhish =	Kor + echhen=	Kor + echhe=
	Korechhi			Korechhish	Korechhen	Korechhe
PrImp	[Not applicable]	উন (un)	অ (o)	Null	উন (un)	উন (uk)
		কর্ +উন = করুন	কর্ + অ = কর	কর্ + null = কর্	কর্ +উন = করুন	কর্ +উক = করুক
		Kor + un =korun	Kor + o = koro	Kor + null = kor	Kor + un =korun	Kor + uk =koruk
PtInd	ল+আম = লাম (lam)	ল+ এন= লেন(len) কর + লেন = করলেন	ল+ এ= লে(le) কর + লে = করলে	ল+ ই= লি(li) কর + লি = করলি	ল+ এন= লেন(len) কর + লেন = করলেন	ল+ null= ল(lo) কর + ল = করল
	কর + লাম= করলাম	Kor + len = Korlen	Kor + le = Korle	Kor + li = Korli	Kor + len =	Kor + lo =
	Kor + lam=				Korlen	Korlo
	Korlam					
PtHab	ত+আম = তাম	ত+ এন= তেন(ten)	ত+ এ= তে(te)	ত+ ই= তি(ti)	ত+ এন= তেন(ten)	ত+ null= ত(to)
	(tam)	কর্ + তেন = করতেন	ক্র্ + তে = ক্রতে	কর্ + তি = করতি	ক্র্ + তেন = ক্রতেন	কর্ + ত = করত
	কর্ + তাম= করতাম Kor + tam=	Kor + ten = Korten	Kor + te = Korte	Kor + ti = Korti	Kor + ten = Korten	Kor + to = Korto
	Kortam				Korten	Korto
PtCon	ছ + ই ল+আম =	ছ + ই ল+ এন=	ছ + ই ল+ এ=	ছ + ই ল+ ই =	ছ + ই ল+ এন=	ছ + ই ল+ null=
	ছিলাম	ছিলেন(chhilen)	ছিলে(chhile)	ছিলি(chhili)	ছিলেন(chhilen)	ছিল(chhilo)
PtPer	(chhilam) কর +ছিলাম=	কর্ + ছিলেন = করছিলেন Kor + chhilen =	কর্ + ছিলে = করছিলে Kor + chhile = Korchhile	কর্ + ছিলি = করছিলি Kor + chhili =	কর্ + ছিলেন = করছিলেন	কর্ + ছিল = করছিল Kor + chhilo =
	করছিলাম	Korchhilen	Koi + cililic – Koicililic	Korchhili	Kor + chhilen =	Korchhilo
	Kor +chhilam=				Korchhilen	
	Corchhilam এছ + ই ল+আম =	এছ + ই ল+ এন=	এছ + ই ল + এ=	এছ + ই ল+ ই =	এছ + ই ল+ এন=	এছ + ই ল+ null=
PtPer	এছ 🛨 ২ গ+আম =	এছ দ ২ গদ অন্ত্র এছিলেন(chhilen)	এছিলে(chhile)	এছ দ ২ গ দ ২ = এছিল(chhili)	এছ 🛨 ২ গ্ৰাম এম = এছিলেন(chhilen)	এছ + ২ ল+ null= এছিল(chhilo)
	(echhilam)	কর্ + এছিলেন = করেছিলেন	কর্ + এছিলে = করেছিলে	কর্ +এছিলি = করেছিলি	কর্ + এছিলেন =	কর্ + এছিল =
	কর্ +এছিলাম=	Kor + echhilen =	Kor + echhile =	Kor + echhili =	করেছিলেন	করেছিল
	করেছিলাম Kor	Korechhilen	Korechhile	Korechhili	Kor + echhilen = Korechhilen	Kor + echhilo = Korechhilo
	+echhilam=				Korcennien	Koreciiiio
	Korechhilam					
FuInd	ব + null/ ও =	ব + এন =বেন (ben) কর্ + বেন = করবেন	ব + এ =বে (be) কর্ + বে = করবে	ব + ই =বি (bi) কর্ + বি = করবি	ব + এন =বেন (ben) কর + বেন = করবেন	ব + এ =বে (be) কর্ + বে = করবে
	ব/বো (bo) কর্ + ব = করব	क्र्म + ५५० = क्र.५० Kor + ben =	φη + τη = φητη Kor + be =	কর্ + ।ব = করাব Kor + bi =	কর্ + বেন = করবেন Kor + ben =	কর্ + বে = করবে Kor + be =
	Kor + bo =	Korben	Korbe	Korbi	Korben	Korbe
	Korbo					
FuCon	তে থাকব (Te thakbo)	তে থাকবেন (Te thakben)	তে থাকবে (Te thakbe)	তে থাকবি (Te thakbi)	তে থাকবেন (Te thakben)	তে থাকবে (Te thakbe)
FuPer	এ থাকব	এ থাকবেন	(1e thakbe) এ शाकरव	এ থাকবি	এ থাকবেন	এ থাকবে
	(e thakbo)	(e thakben)	(e thakbe)	(e thakbi)	(e thakben)	(e thakbe)
FuImp	Not applicable	ব + এন =বেন (ben)	3 (0)	ইস(ish)	ব + এন =বেন (ben)	ব + এ =বে (be)
		কর্ + বেন = করবেন Kor + ben =	কর্ + ও = করো Kor + o = koro	কর্ + ইস = করিস Kor + ish= Korish	কর্ + বেন = করবেন Kor + ben =	কর্ + বে = করবে Kor + be =
		Kor + ben = Korben	KOI + O = KOIO	KOI T ISII— KOIISII	Kor + ben = Korben	Kor + be = Korbe
	i	· · · · · · · · · · · · · · · · · · ·				

7. Future Research Areas

The following points should be considered by the future researchers to enhance and improve the proposed model. Due to the constraint in space, we could not include grammar rules to

handle various Bangla punctuation symbols. As this paper mainly focuses on the syntax analysis phase, it sets aside the job of extracting additional information about the other parts of speech such as the noun, the pronoun, the adjective, the adverb and the indeclinable which will be of great use in the semantic

analysis phase. The concepts of the change of voice, narration and other special concepts of Bangla grammar such as the composition of words (সমাস) and inflection of the noun or pronoun (নাম-বিভক্তি) should be further analyzed As the major emphasis was given to parse the finite verb of the sentence, other types of clauses like the adverbial or adjective clauses were not parsed further.

8. Conclusion

This paper proposes a technique to parse Bangla sentences in a new approach using context-free grammar rules. The principal goal was to design a parser that is capable of accepting all types of Bangla sentences, both from the structural viewpoint and the viewpoint of intonation of verb. Innovative efforts in the areas of future expandability can treat this paper as a starting milestone.

Abbreviations used MT Machine Translation, NLP Natural Language Processing, NP Noun Phrase, VP Verb Phrase, FP First Person SP Second Person, SPH Second Person Honorific, SPNH Second Person Non Honorific, TP Third Person SPP Second Person Pejorative, TPH Third Person Honorific, **TPNH** Third Person Non Honorific NonImperativeCon Non Imperative Concord, **NIConFP** Non Imperative Concord for FP NIConSPH Non Imperative Concord for SPH, **NIConSPNH** Non Imperative Concord for SPNH NIConSPP Non Imperative Concord for SPP, **NIConTPH** Non Imperative Concord for TPH **NIConTPNH** Non Imperative Concord for TPNH, ImperativeCon Imperative Concord **IConFP** Imperative Concord for FP, **IConSPH** Imperative Concord for SPH, IConSPNH Imperative Concord for SPNH, **IConSPP** Imperative Concord for SPP **IConTPH** Imperative Concord for TPH, **IConTPNH** Imperative Concord for TPNH PrInd Present Indefinite, PrCon Present Continuous, PrPer Present Perfect, PrImp Present Imperative PtInd Past Indefinite, PtHab Past Habitual, PtCon Past Continuous, PtPer Past Perfect, SubExpander Subject Expander

FuInd Future Indefinite, **FuCon** Future Continuous, **FuPer** Future Perfect, **FuImp** Future Imperative, **PreExpander** Predicate Expander

Acknowledgements

We would like to thank Mrs. Shamim Jahan Ahsan of Viqarunnisa Noon School, Dhaka, for her invaluable suggestions and comments.

References

- 1. Alfred V. Aho, Ravi Sethi and Jeffrey D. Ullman, *Compilers: Principles, Techniques and Tools*, Addison-Wesley, 2000.
- 2. Mortuza Ali and Muhammad Masroor Ali, "Developments of Machine Translation Dictionaries for Bangla Language", in Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2002.
- 3. D. Arnold, L. Balkan, S. Meijer, L. Humphreys and L. Sadler, *Machine Translation An Introductory Guide*, Blackwells-NCC, London, 1994.
- 4. M. A. Islam, M.L. Rahman and M. A. Mottalib, "Developments of Machine Translation: A Review", in Proceedings of National Conference on Computer and Information Systems (NCCIS), Dhaka, Bangladesh, 1997.
- 5. Md. Manzoor Murshed, "Parsing of Bengali Natural Language Sentences", in Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 1998.
- 6. S. J. Russell and P. Norvig, *Artificial Intelligence A Modern Approach*, ch. 22-23, Prentice-Hall International, Inc., 1995.
- 7. Abdul Baqi M. Sharaf, "An Object Oriented Model for Natural Language Processing", in Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2000.
- 8. Mohammad Reza Selim and Mohammad Zafar Iqbal, "Syntax Analysis of Phrases and Different Types of Sentences in Bangla", in Proceedings of International Conference on Computer and Information Technology (ICCIT), SUST, Sylhet, Bangladesh, 1999.
- 9. P.C. Wren & H. Martin, *English Grammar & Composition*, S. Chand & Company LTD, 2000.
- 10. বাংলা ভাষার ব্যাকরণ, জাতীয় শিক্ষাক্রম ও পাঠ্যপুস্তক বোর্ড, ঢাকা ।