

# Analysis of Bangla Root Word for Universal Networking Language (UNL)

Aloke Kumar Saha  
Department of CSE,  
University of Asia Pacific,  
Dhaka, Bangladesh

M. F Mridha  
Department of CSE,  
University of Asia Pacific,  
Dhaka, Bangladesh

Jugal Krishna Das  
Department of CSE,  
Jahangirnagar University,  
Savar, Dhaka, Bangladesh

## ABSTRACT

Root Word (RW) Analysis Approach for Universal Words (UWs) in Bangla Language is a technique which helps to find out more correct universal words more accurately and quickly than other available techniques. The main problem with universal word creation is that, as Bangla sentences varies semantically depending on the use of the same root in different sentences, it is important to get accurate universal words or universal expressions according to this variation of the sentence structure. The length of the sentence also plays an important role in determining the performance of the output. If length of the sentence is large enough and more complex, it will be problematic for the parser to make specified language which may not imply accurate meaning as required. In this paper a new method, termed as Root word Analysis Approach for converting Bangla to UNL has been investigated Universal Word construction. Root word Analysis Approach begins generation in a constructive fashion by adding some rule/criterion one after another which also shows how to use different meaning of root words using dictionary according to root words in different sentences.

## Keywords

Bangla Root Word, Morphology, Bangla Suffix, Bengali Cases and their inflexions and UNL.

## 1. INTRODUCTION

Bangla is spoken by about 245 million people of Bangladesh and two states of India, but most of the computer based resources and technical journals are in English. Due to the language barrier, the common people face big obstacle to enjoy the optimum benefits of modern information and communication technology (ICT) as well as huge enriched English knowledge database around the globe. The Universal Networking Language (UNL), which is a formal language for symbolizing the sense of natural language sentences, is a specification for the exchange of information. The main aim of the UNL project is to overcome language barrier. Currently, the UNL includes 16 languages, which are the six official languages of the United Nations (Arabic, Chinese, English, French, Russian and Spanish), in addition to the ten other widely spoken languages (German, Hindi, Italian, Indonesian, Japanese, Latvian, Mongol, Portuguese, Swahili and Thai). Bengali is not yet included. In the last few years, machine translation techniques have been applied to web environments. The growing amount of available multilingual information on the Internet and the Internet users has led to a justifiable interest on this area. Hundreds of millions of people of almost all levels of education, attitudes and different jobs all over the world use the Internet for different purposes [1], where English is the main language of the Internet. Because

English is not understandable for most of the people, Interlingua translation programs are needed to develop. The main goal of the UNL system, which allows users to visualize websites in their native languages, is to provide a common representation for accessing Internet of multilingual websites by the majority of the people over the world. For this common representation, lexical knowledge is a critical issue in natural language processing systems, where the development of large-scale lexica with specific formats capable of being used by distinguished applications, in particular to multilingual systems, has been given special focus. Our goal is to include Bengali in this system with less effort.

In this paper we present the Root Word analysis of Bangla Sentences for UNL system. The major components of our research works touches upon i) development of grammatical attributes for Bangla root word to construct *Bangla Word Dictionary* and use of morphological analysis ii) UNL Expression of the Bangla attributes and iii) Bangla sentences analysis. In section 2 we describe the Bangla Sentence structure. In sections 3 and 4, we present our main works that include all the above three components.

## 2. BANGLA SENTENCE STRUCTURE

### 2.1 Basic structure differences

Direct conversion of English sentences into Bengali sentences very often suffers from different kinds of difficulties and ambiguities. For this, conversion is done with the help of the inner significance of the sentence rather than direct word-to-word conversion. In these cases, at first, the purpose of the sentence is understood and then the sentence is interpreted according to that purpose [9][10]. It can be done by proceeding in a systematic way with the help of logical algorithm which is much complex. So direct mapping between the structure of English and Bengali sentences required for computer aided solution. But for many sentences, this mapping is not possible. As for example, we may consider the following sentence:

**Please lend me your ears.** -Its Bengali meaning is “অনুগ্রহ করে আমার কথা শুন।”

In this example, the English word “lend” means “ধার দেয়া” but this is not present in Bengali meaning rather the Bengali word “শুন” is present whose English synonym is “hear” which is not present in the English sentence at all.

## 2.2 Difference between the Forms of Verbs

In English there are four basic forms of verbs (present, past, continuous and past participle), for example: “go” verb has different forms like: go, went, going and gone. But in Bengali there are a lot of forms of a verb, which differ according to tense, person and number. So it needs to develop a system that is able to find out the appropriate form of verb.

## 2.3 Contextual Problem

Translation process suffers from multiple meaning of the English word. It means which meaning is to choose from several meaning of a single word. This problem can be solved by the determination of the status or class of the subject of a sentence, e.g.

1. Rahim is an engineer. -Its Bengali meaning is “রহিম একজন প্রকৌশলী”

2. Tota is a sweet singer. -Its Bengali meaning is “তোতা একটি মধুর গায়ক”

1. In the first case the word “a” means “একজন”, in the second case, the word “a” means “একটি”. But for a translator, it is very hard to distinguish between these subjects to identify which one is for human being and which one is for bird or animal or other objects;

## 2.4 Multiword Unit (Idioms and Phrases)

Roughly speaking, idioms are expressions whose meaning cannot be completely understood from the meaning of the component parts. For example: 1. He lives from hand to mouth. Its meaning is “সে দিন আনে দিন খায়” and its word to word meaning is “সে হাত থেকে মুখে বাঁচে”. It is hard job to construct a translator that can find out the idioms and collocations of the sentence and translate to the proper meaning.

## 3. MORPHOLOGICAL ANALYSIS OF BENGALI WORDS

Morphology is the field of linguistics that studies the structure of words. It focuses on patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. In natural language processing (NLP) we need to identify words in texts in order to determine their syntactic and semantic properties [7]. In the following section we are analyzing morphologically the different Bengali Roots and Primary suffixes so that we can develop efficient Models for dictionary entries.

### 3.1 Why analysis of Root word need

Verb is the main part of any sentence for any native language. Any Sentence can complete without subject or object. But without verb no sentence is complete. So verb analysis is need to converting from Bengali to UNL. And verb is the combination of root word and suffixes. And root word is tritted as entry node when converting any native language to UNL. And not only verb but also other word is derived from a root word that may have the different transformations. This happens because different morphemes are added with it as suffixes. Therefore, the meaning of the word varies for its different transformations.

## 3.2 Bangla Primary Suffixes (কৃৎ প্রত্যয়)

We know that the core of the verb is called root and if number of suffixes are added to roots then they form verbs. When sound or sounds [8] are added with roots and form nouns or adjectives then the root words are called root verbs and the sound or sounds are added with root verbs are called Primary Suffixes. For example চন্ (Root verb)+অন্ (Primary Suffix)=চলন (Noun) and চন্ (Root verb)+অন্ড (Primary Suffix)=চলন্ড (Adjective). Some others primary suffixes are অন, অনা, অনি, অক, আ etc.

## 3.3 General Prefixes

General Prefixes are the words that are used before words to express various meanings of the same words. There are around fifty (50) prefixes used in Bengali sentences. In Shangkrit Bengali we use twenty (20) prefixes[2] say , প্র(প্রকর্ষ), পরা(বৈপরীত্য), অপ(বৈপরীত্য) etc. , in Bengali we use thirteen prefixes (13) prefixes[2] such as বে(বৈপরীত্য), গর(বৈপরীত্য), অন(অভাব) etc., five(5) foreign prefixes[5] such as গর(না), বদ(খারাপ) etc., four English prefixes[14] such as সাব(অধীন), head (প্রধান), ফুল(পুরা) etc. These prefixes are used before words to make thousands of meaningful Bengali Words.

Here we will make separately Word Dictionary entries for all of these prefixes and words, so that they can combinely make meaningful words by applying rules. For example, if we consider prefix “প্রতি” (means like/similar/every/opposite/against etc.) we can make “প্রতিদিন”, “প্রতিশব্দ” etc. Now we can make the word “প্রতি” for dictionary entry. But the word “প্রতি” has two or more meanings so that we have to represent two or more dictionary entries for the word as follows:

[প্রতি] {} “every (icl>thing)” (ABSTRACT THING) <B,0,0>

[প্রতি] {} “opposite (icl>thing)” (ABSTRACT THING) <B, 0, 0>

Now if we want to represent the concepts of the words say প্রতিদিন, প্রতিশব্দ etc., we need not represent the whole words. We have to represent only the words “দিন”, “শব্দ” in the dictionary entry as per the following format.

[দিন] {} “day (icl>period>time)” (N, ABSTRACT THING, LIGHT) <B, 0, 0>

[শব্দ] {} “sound (icl>ocure>thing)” (N, ABSTRACT THING) <B, 0, 0>

## 3.4 UNL Structure of Bengali Case

As in the traditional understanding case (in Bengali Kaarok) in Bengali denotes the relationship of the nominals with the main verb of the clause. The cases are broadly classified into 6 categories, each having a finer categorization into sub-types. The traditional case with their inflexions is shown in table.

**Table 1 Bengali Cases and their inflexions**

Classical Case vs. Bengali Kaarok	Inflexions (Case maker)	
	Singular	Plural
Nominative case(Kartri Kaarok)	0 (null), ঐ(e),ঐ(oy), তে(te)	রা(ra), ঐরা(era)
Accusative case (Karma Kaarok)	0(null), কে(ke), রে(re), ঐরে(ere)	দিগকে(digoke)
Instrumental case (Karan Kaarok)	দারা(dara),দিয়া(dia), কর্তিক(kortik)	দের দারা(der dara), দিগের দারা(diger dara)
Dative case (Sampradaan Kaarok)	0(null), কে(ke), রে(re), ঐরে(ere), ঐ(e),ঐ(oy), ইত(et)	দিগকে(digoke)
Ablative case (Apaadaan Kaarok)	হইতে(hoite),থেকে(theke)	দের হইতে(der hoite), দিগের হইতে(diger hoite)
Genitive case (Sambandha pad)	র(r), ঐর(er)	দের (der), দিগের(diger)
Case of time-place (Adhikaran Kaarak)	য়া(e), ঐ(oy), তে(te)	সকলে(sokole),দের মধ্যে(der modde)

Table summarizes the correspondence between Bengali cases and the UNL relations.

**Table 2 Bengali cases versus UNL Relations**

Classical Case	Corresponding UNL relations
Kartri Kaarok	agt, cag, ptn, aoj, cao, and
Karma Kaarok	obj, ben, cob
Karan Kaarok	ins, met
Sampradaan Kaarok	ben,gol,pur, rsn
Apaadaan Kaarok	frm, src, plf, tmf
Sambandha pad	mod, pos, pof
Adhikaran Kaarak	plc, plt, tim, tmt, to, gol, scr, opl

### 3.5 Grammar based Analysis

We can Analysis root word kor(কর), ধর(dhor), পড়(por), গড়(gor), চল(chol), খেল(khel), লেখ(lekh) using these following rules that discus above.

**Table 3 Bengali root word কর (kor)**

Tense	Language	First person	Second person	Third person
Simple Present	Shadhu (সাদু) Suffix-0	করি কর+0+ই	কর কর+0	করে কর+0+ঐ
		কর+ই Kor+i	কর Kor	কর+ঐ Kor+e
	Cholito (চলিত) Suffix-0	করি কর+0+ই	কর কর+0	করে কর+0+ঐ
		কর+ই Kor+i	কর Kor	কর+ঐ Kor+e
Continuo	Shadhu	করিতেছি	করিতেছ	করিতেছে

us Present	(সাদু) Suffix-ইতে	কর+ইতে+ছি Kor+ite+chi	কর+ইতে+ছ Kor+ite+ch	কর+ইতে+ছে Kor+ite+che
	Cholito (চলিত) Suffix-0	করছি কর+0+ছি কর+ছি Kor+chi	করছ কর+0+ছ কর+ছ Kor+ch	করছে কর+0+ছে কর+ছে Kor+che
Perfect Present	Shadhu (সাদু) Suffix-ইয়া	করিয়াছি কর+ইয়া+ছি Kor+iya+chi	করিয়াছ কর+ইয়া+ছ Kor+iya+ch	করিয়াছে কর+ইয়া+ছে Kor+iya+che
	Cholito (চলিত) Suffix-ঐ	করেছি কর+ঐ+ছি Kor+e+chi	করেছ কর+ঐ+ছ Kor+e+ch	করেছে কর+ঐ+ছে Kor+e+che
Simple Past	Shadhu (সাদু) Suffix-ই	করিলাম কর+ই+লাম Kor+i+lam	করিলে কর+ই+লে Kor+i+le	করিল কর+ই+ল Kor+i+l
	Cholito (চলিত) Suffix-0	করলাম কর+0+লাম Kor+lam	করলে কর+0+লে Kor+le	করল কর+0+ল Kor+l
Continuo us Past	Shadhu (সাদু) Suffix-ইতে	করিতেছিলাম কর+ইতে+ছি+লাম Kor+ite+chi+lam	করিতেছিলে কর+ইতে+ছি+লে Kor+ite+chi+le	করিতেছিল কর+ইতে+ছি+ল Kor+ite+chi+l
	Cholito (চলিত) Suffix-0	করছিলাম কর+0+ছি+লাম Kor+chi+lam	করছিলে কর+0+ছি+লে Kor+chi+le	করছিল কর+0+ছি+ল Kor+chi+l
Perfect Past	Shadhu (সাদু) Suffix-ইয়া	করিয়াছিলাম কর+ইয়া+ছি+লাম Kor+iya+chi+lam	করিয়াছিলে কর+ইয়া+ছি+লে Kor+iya+chi+le	করিয়াছিল কর+ইয়া+ছি+ল Kor+iya+chi+l
	Cholito (চলিত) Suffix-ঐ	করেছিলাম কর+ঐ+ছি+লাম Kor+e+chi+lam	করেছিলে কর+ঐ+ছি+লে Kor+e+chi+le	করেছিল কর+ঐ+ছি+ল Kor+e+chi+l
Simple Future	Shadhu (সাদু) Suffix-ই	করিব কর+ই+ব Kor+i+bo	করিবে কর+ই+বে Kor+i+be	করিবে কর+ই+বে Kor+i+be
	Cholito (চলিত) Suffix-0	করব কর+0+ব Kor+bo	করবে কর+0+বে Kor+be	করবে কর+0+বে Kor+be
Continuo us Future	Shadhu (সাদু) Suffix-ই	করিতে থাকিব কর+ই+তে/থাক+ই+ব Kor+i+te/thak+i+bo	করিতে থাকিবে কর+ই+তে/থাক+ই+বে Kor+i+te/thak+i+be	করিতে থাকিবে কর+ই+তে/থাক+ই+বে Kor+i+te/thak+i+be

	Cholito (চলিত) Suffix-0	করতে থাকব কর+0+তে/ থাক+0+ব কর+তে/থাক +ব Kor+te/thak +bo	করতে থাকবে কর+0+তে/থা ক+0+বে কর+তে/থাক+ বে Kor+te/thak+ be	করতে থাকবে কর+0+তে/থাক+0+ বে কর+তে/থাক+বে Kor+te/thak+be
Perfect Future	Shadhu (সাদু) Suffix- ইয়া	করিয়া থাকিব কর+ইয়া/থাক +ই+ব Kor+iya/thak +i+bo	করিয়া থাকিবে কর+ইয়া/থাক +ই+বে Kor+iya/thak +i+be	করিয়া থাকিবে কর+ইয়া/থাক+ই+বে Kor+iya/thak+i+be
	Cholito (চলিত) Suffix-এ	করে থাকব কর+এ/থাক+ 0+ব কর+এ/থাক+ ব Kor+e/thak+ bo	করে থাকবে কর+এ/থাক+0 +বে কর+এ/থাক+ বে Kor+e/thak+b e	করে থাকবে কর+এ/থাক+0+বে কর+এ/থাক+বে Kor+e/thak+be

#### 4. BANGLE TO UNL ENCONVERSION

An En-Converter is a software that automatically or interactively en-converts natural language text into UNL. As it generates UNL from natural languages, enables peoples to make UNL documents without any knowledge about UNL. It means that users of the UNL system do not need to learn UNL. En-Converter is a language independent parser that provides synchronously a framework for morphological, syntactic and semantic analysis. It would be impossible to solve an ambiguity in morphological analysis without the use of syntactic or semantic information. Also, it would be impossible to solve an ambiguity in syntactic analysis without the use of semantic information. En-Converter generates UNL expressions from sentences (or lists of words of sentences) of a Bengali language by applying en-conversion rules. In addition to the fundamental function of en-conversion, it checks the formats of rules, and outputs the messages for any errors. It also outputs the information required for each stage of en-conversion in different levels. With these facilities, a rule developer can easily develop and improve rules by using En-Converter. En-Converter loads the en-conversion rules and the rule checker works while converting rules. Once the rules are made, they are stored automatically and can be used directly the next time without rule conversion.

1. Convert or load the rules.
2. Secondly, it inputs a string or a list of morphemes/words of a sentence of Bengali language.
3. Input a Bengali sentence.

Then, it starts to apply rules to the Node-list from the initial state. En-Converter applies en-conversion rules to the Node list. The process of rule application is to find a suitable rule and to take actions or operate on the Node-list in order to create a syntactic functionalities and UNL network using the nodes in the Analysis Windows. If a string appears in a window, the system will retrieve the Word Dictionary and apply the rule to the candidates of word entries. In this case, if a word satisfies the conditions required for the window of a rule, this word is selected and the rule application succeeds. This process will be continued until the syntactic functions

and UNL network are completed and only the entry node remains in the Node-list [s7].

- a. Apply the rules and retrieve the Word Dictionary Finally, it outputs the UNL network (Node-net) to the output file in the binary relation format of UNL expression.
- b. Output the UNL

Output the UNL expressions with the exception of the first process of rule conversion and loading, once En-Converter starts to work, it will repeat the other processes for all input sentences.

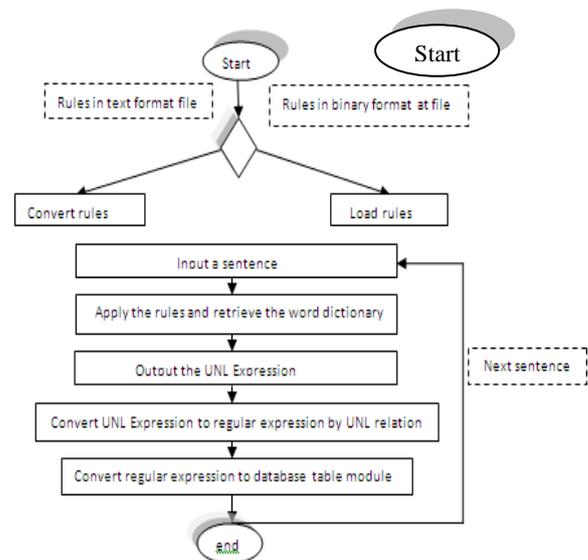


Figure: 1 Flowchart of the Enconversion Process

En-Converter analyses a sentence using the Word Dictionary, Knowledge Base, and En-conversion Rules. It retrieves relevant dictionary entries from the Word Dictionary, operates on nodes in the Node-list by applying En-conversion Rules, and generates semantic networks of UNL by consulting the Knowledge Base. The word entries of Bengali language are stored in the Word Dictionary. Each entry of the Word Dictionary is composed of three kinds of elements: the Headword, the Universal Word (UW) and the Grammatical Attributes. A headword is a notation/surface of a word of a Bengali language that composes the input sentence, and it is to be used as a Bengali En-conversion Rules. Apply the rules and retrieve the Word Dictionary. Figure 2 shows the flowchart of enconversion process. Then, the UNL network (Node-net) is outputted to the output file in the binary relation format of UNL expression then we convert the UNL Expression to regular expression by UNL relation. Finally, we need to convert the regular expression to database table module. Flow chart of Bengali to UNL En-Conversion:

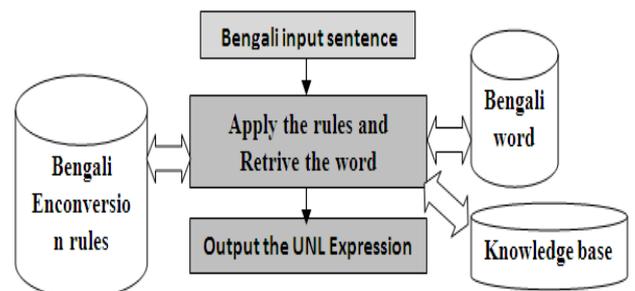


Figure: 2 References between each process and the dictionary and the rule files

**Example:** Consider the following simple English sentence “It is very hot in Dhaka today”.

The corresponding Bengali text and its word by word meaning in English is given as,

ঢাকায় আজ খুব গরম - Dhaka-in today very hot

The node list is enclosed within “<<” and “>>”. The analysis window is enclosed within “[” and “]”. The nodes delimited by “/” are those explored and fixed by the system. Morphological analysis is carried out in a series of steps.

1. /<</[ঢাকা]/[য়]/“আজ খুব গরম”/>>/
2. /<</[ঢাকা য়]/“আজ খুব গরম”/>>/
3. /<</ঢাকা য়/আজ/[খুব]/[গরম]/>>/
4. /<</ঢাকা য়/[আজ]/[গরম]/>>/
5. /<</[ঢাকা য়]/[গরম]/>>/
6. /<</[গরম]/>>/

**Step 1:** The noun ঢাকা (*Dhaka*) of the type *PLACE* and case maker য় (*in*) are combined and an attribute *PLC* is added to the noun to indicate that *plc* relation can be made between the main predicate of the sentence and this noun.

**Step 2:** In this step, the system right shifts here because there is no combination or modification rule between noun and adverb.

**Step 3:** The system recognizes গরম (*hot*) as a predicate of the sentence. So, it generates *man* relation between

the adverb খুব (*very*) and the predicate গরম (*hot*). The adverb খুব (*very*) is deleted afterwards.

**Step 4:** Then the analyzer looks ahead further right beyond the noun phrase আজ (*today*) to get the predicate গরম (*hot*). A *tim* relation is created between আজ (*today*), and গরম (*hot*) and finally, আজ (*today*) is deleted.

**Step 5:** By using *PLACE* attribute of the noun ঢাকা (*Dhaka*), the system generates *plc* relation between the noun ঢাকা and the predicate গরম.

**Step 6:** A right shift at this point brings the Sentence Tail (STAIL) under the LAW and thus signals the end of analysis. This right shift rule also attaches the attribute *@entry* to the last word left in the Node-list and thus the predicate গরম (*hot*) is preserved till the end.

The UNL output of the corresponding sentence is as follows:

```
[unl]
man(hot(icl>state).@entry.@present,very(intensifier))
tim(hot(icl>state).@entry.@present,today(icl>period))
plc(hot(icl>state).@entry.@present,Dhaka(icl>place))
[/unl]
```

## 5. CONCLUSION

Root word Analysis Approach targets to improve the universal word creation ability of Bangla language. A new approach proposed in this thesis. The new technique uses a constructive approach to the universal language of Bangla language to determine. The novelty of this method is that it is simple and easy approach is used to modify the properties and diversified use of words on proposals for a given word that is to be determined. This analysis functions initially tried to solve this problem for a number of proposals for example, the approach required for semantically correct rate for the true meaning of giving.

## 6. REFERENCES

- [1] H. Uchida, M. Zhu. The Universal Networking Language (UNL) Specification Version 3.0 Edition 3 ,Technical Report, UNU, (2005/6-UNDL Foundation, International Environment House, Tokyo, 2004)
- [2] H. Uchida, M. Zhu, “The Universal Networking Language (UNL) Specification Version 3.0”, Technical Report, United Nations University, Tokyo, 1998
- [3] S. Abdul-Rahim, A.A. Libdeh, F. Sawalha, M. K. Odeh, “Universal Networking Language(UNL) a Means to Bridge the Digital Divide”, Computer Technology Training and Industrial Studies Center, Royal Scientific Society, March 2002.
- [4] M. M. Asaduzzaman, M. M. Ali, “Morphological Analysis of Bengali Words for Automatic Machine Translation”, International Conference on Computer, and Information Technology (ICIT), Dhaka, 2003, pp.271-276
- [5] Bengali Academy (2004), Bengali-English Dictionary, Dhaka.
- [6] Enconverter Specifications, version 3.3, UNL Center/UNDL Foundation, Tokyo, Japan 2002.
- [7] Enconverter Specification Version 3.3, (UNU Centre, Tokyo 150-8304, Japan 2002)
- [8] DeConverter Specification, Version 2.7, (UNL Center, UNDL Foundation, Tokyo 150-8304, Japan 2002)
- [9] D.M. Shahidullah. Bengali Baykaron, (Ahmed Mahmudul Haque of Mowla Brothers prokashani, Dhaka 2003)
- [10] Zakir Hossain, Shahid Al Noor, Muhammad Firoz Mridha Some Proposed Standard Models for Bengali Dictionary Entries of Bengali Morphemes for Universal Networking Language. IJCSNS International Journal of Computer Science and Network Security, V OL.12 No.11, November 2012 .
- [11] Bouguslavsky, I., Frid, N. and Iomdin, L. (2000). Creating a Universal Networking Module within an Advanced NLP system. Proceedings of the 18th International Conference on Computational Linguistics, pp. 83-89.
- [12] Aloke Kumar Saha, Muhammad F. Mridha, Manoj Banik, and Jugal Krishna Das. Specification of UNL Deconverter for Bengali Language. International Journal of Scientific & Engineering Research, Volume 3, Issue 9, September-2012 ISSN 2229-5518.