Defining the Structure of Bangla Noun Phrase and Developing Rules for UNL

A. S. M Mahmudul Hasan, Saria Islam, Dr. Jugal Krishna Das, Golum Farook Ahmed

Abstract—The Universal Networking Language (UNL) deals with the communication across nations of different languages and involves with many different related discipline such as linguistics, epistemology, computer science etc. It helps to overcome the language barrier among people of different nations to solve problems emerging from current globalization. The Universal Networking Language (UNL), a project undertaken under the auspices of the United Nations University (UNU) in Tokyo and for a framework for integration of Bangla language to UNL. The mission of the UNU project is to allow people across nations to access information on the Internet in their own language- a step to help bridge the digital divide. Noun phrase is an important analysis for natural language processing. Some of the example noun phrases and there conversion process and rule has been discussed in this thesis. This research is to demonstrate our pioneering efforts in the field of Bengali (Bangla). Here we define Bangla sentence structure and how the noun phrases are used in Bangla. According to the structure of the noun phrase we have developed some rules. By using these rules we can now convert a complex noun phrase into UNL which was not possible before.

Index Terms— Universal Networking Language(UNL), Bangla Noun Phrase, Enconverter, Enconversion rules, UNL relation, UNL attributes, Universal Words.

1 Introduction

The Internet has to face the complexity of multilinguality. People speak different languages and the number of natural languages along with their dialects is estimated to be closed to 4000 [1]. Of the top 100 language in the world, English occupies the top position. That is why English is the main language of the Internet. Understandably not all people know English. Nations are becoming more independent and they need to exchange information [2]. Translation is only means to disseminate information but only with much effort and involving direct and indirect cost [3]. Considering this issue, Universal Networking Language (UNL) is developed by the United Nations University/Institute of Advanced Studies (UNU/IAS) to convert any language to other language [4]. The Universal Networking Language (UNL) is an artificial language for representing, describing, summarizing, refining, storing and disseminating information in a natural-language-independent format [5]. The UNL Program was initially conceived to support multilingual services in Internet being an alternative to classical machine translation systems. The UNL system revolves around a unique artificial language (Universal Networking Language) that pretends to capture the meaning of written documents [6]. Our goal is to include Bangla in this system with less effort.

Internet having other than Bangla likes Spanish, Chinese, Japanese languages. Addressing this issue, it is highly demanding to develop analysis rules to convert Bangla sentence to UNL expression followed by UNL expression to any other native language. In this research, we have developed a structure of noun phrase and some rules to convert the noun phrase into UNL. Theoretically, it proves the successful performance of the rules for converting Bangla to UNL expressions.

2. UNL System - In a Nutshell

Universal Networking Language (UNL) is an Interlingua developed by UNDL foundation. UNL is in the form of semantic network to represent and exchange information. Concepts and relations enable encapsulation of the meaning of sentences. In UNL, a sentence can be considered as a hypergraph where each node is the concept and the links or arcs represent the relations between the concepts. UNL knowledge base provides concepts for the words in the natural language sentences.

- **A. S. M Mahmudul Hasan** is the lecturer of Hamdard University Bangladesh. He received a B.Sc degree from Jahangirnagar University in 2010 and continuing his M.Sc from the same university in Computer Science and Engineering. PH-+8801670140616. E-mail: apon_cse@yahoo.com
- **Saria Islam** obtained her B.Sc (Hons) and continuing M.Sc in Computer Science and Engineering from Jahangirnagar University She is now serving the IBAIS University as a lecturer in computer science and engineering department. PH-+8801670141821. E-mail: saria_islam@yahoo.com.
- **Dr. Jugal Krishna Das** has completed his Ph. D. from Glushkov Institute of Cybernetics, Kiev, Ukraine, in 1993 and M. Sc. from Donetsk Polytechnic Institute,Ukraine, in 1989. Now he is working as a Professor in the department of Computer Science and Engineering,Jahangirnagar University, Savar, Dhaka, Bangladesh. PH-+8801712509082. E-mail:cedas@juniv.edu.
- **Mr. Gulam Farooque Ahmed** is the Chief Executive Director of the Computer Village. PH-+8801819016640. E-mail: villagebd@gmail.com

The UNL consists of Universal Words (UWs), Relations and Attributes and knowledge base.

2.1 Universal Words (UWs)

Universal words are UNL words that carry knowledge or concepts. UWs are simply nodes in the UNL graph. There are two type of UWs: permanent and temporary. Permanent UWs represent concepts of common use and are included in the UW dictionary. Temporary UWs may represent new concepts, too specific or not translatable so that they are not included in the dictionary. Examples: bucket(icl>container), water(icl>liquid).

2.2 Relations

Relations are labelled arcs that connect nodes (Uws) in the UNL graph. The relations are binary and usually represent semantic cases and semantic roles. Examples:
agt (break(agt>thing,obj>thing), John(iof>person))

2.3 Attributes

Attributes are annotations used to represent grammatical categories, mood, aspect, etc. Every attribute starts with "@" symbol. Example:
work(agt>human).@past
– means that the tense of the verb *work* is past tense, that is , it is actually *worked* in the sentence.

2.4 Knowledge Base

The UNL Knowledge Base contains entries that define possible binary relations between UWs [7].

3. UNL ENCONVERTER

The EnConverter is a language independent parser, which provides a framework for morphological, syntactic, and semantic analysis synchronously. It would be impossible to solve all the morphological ambiguities if the syntactic or semantic analysis is not performed synchronously. And, it would be impossible to solve every syntactic ambiguity in the absence of semantic analysis.

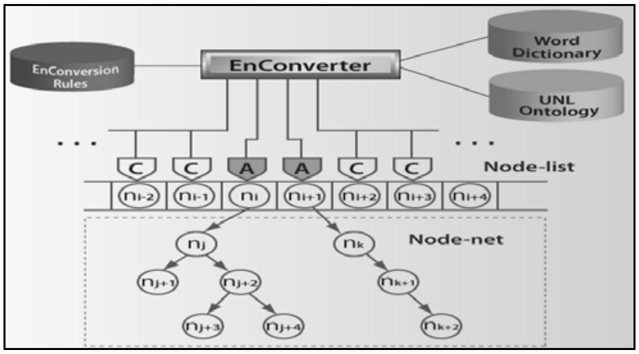


Figure 1: Structure of Enconverter

The EnConverter works in the following way. An input string of text of a sentence is scanned from left to right. When an input string is scanned, all matched morphemes from the beginning (left) of the string are retrieved from the word dictionary and become the candidate morphemes. These candidate morphemes are sorted according to priority. Word selection is done by applying grammar rules of enconversion to these candidate morphemes. Syntactic and semantic analysis is carried out by applying the rules to already selected words to build up a syntactic tree and a semantic network for the input sentence. This process continues until all words of the sentence are inputted, and a complete semantic network of the input sentence is made. The output of this whole process is a semantic network expressed in the UNL format.

4. ANALYSIS OF NOUN PHRASE IN BANGLA

4.1 Structure of Bangla Sentence

A sentence can include words grouped meaningfully to express a statement. A simple sentence can have a verb and some other kinds of meaningful words connected with the verb to express the complete meaning. Those words except verb can be called as noun phrase. In a simple sentence, there are one or more noun phrases centering a verb. The noun phrase can have a case marker to create a relation with the verb.

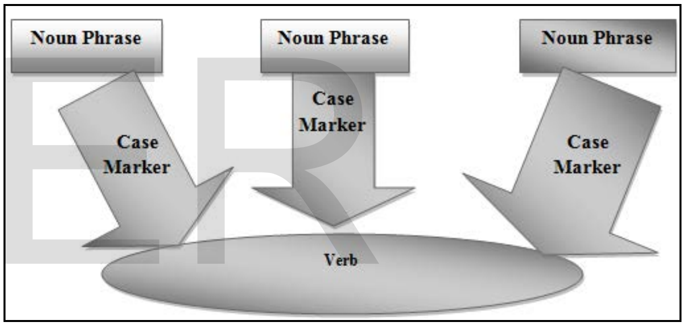


Figure 2: Structure of Bangla Sentence

For example: "আমার বাড়িটি গতকাল ঝড়ে ভেঙ্গে গেছে" pronounce as "Amar barite gotokal jhore bhenge gese" means "My house was broken yesterday by storm".

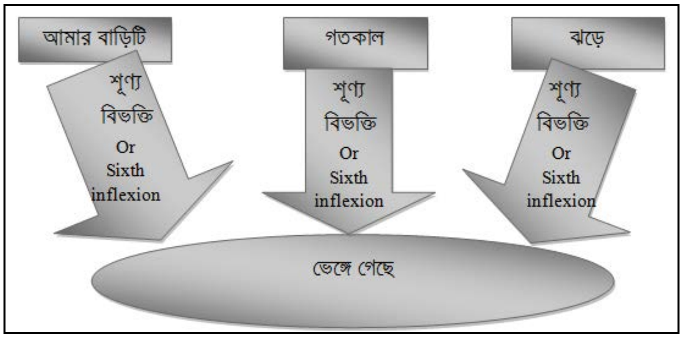

আমার বাড়িটি
গতকাল
ঝড়ে
শূণ্য বিভক্তি Or Sixth inflexion
শূণ্য বিভক্তি Or Sixth inflexion
শূণ্য বিভক্তি Or Sixth inflexion
ভেঙ্গে গেছে

Figure 3: An Example Bangla Sentence

4.2 Structure of a Noun Phrase

A greater noun phrase can have more than one smaller noun phrases. The greater noun phrase can be referred as compound noun phrase and the smaller noun phrase can be referred as simple noun phrase. A compound noun phrase can be formed by recursively joining more than one simple noun phrases.

Noun phrase = [Noun phrase+][Sixth inflexion+][Simple noun phrase]

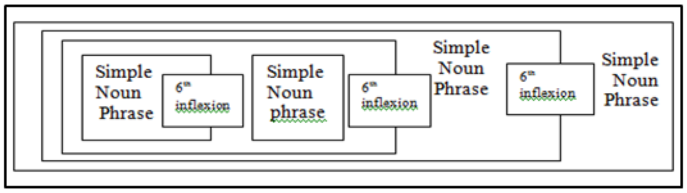


Figure 4: Structure of Noun Phrase

Simple Noun Phrase = [Adjective +][Number Word+] [Conjunction+] [Ordinal Number+] Noun/Pronoun
For example: “আমাদের বাড়ির ছোট মেয়েটির লাল জামাটি” Pronounce as "amader barir choto meyetir lal jamati" is a complete noun phrase.

4.3 Role of Sixth Inflexion in Forming a Noun Phrase

For joining one simple phrase to another to form a compound noun phrase, sixth inflexion is used. Sixth inflexion is used for modifying the next noun phrase in the sentence. For example: “আমাদের বাড়ির ছোট মেয়েটির লাল জামাটি” Pronounce as "amader barir choto meyetir lal jamati".

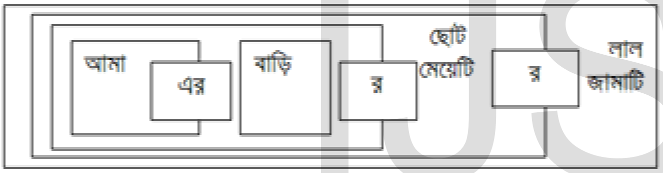


Figure 5: Example of Sixth Inflexion

Here, "আমাদের বাড়ির" is a simple noun phrase. Adding sixth inflexion it can be joined with the next simple noun phrase "ছোট মেয়েটির". Then adding another sixth inflexion it can be joined with the next simple noun phrase "লাল জামাটি". So, formation a compound noun phrase is a recursive procedure of adding simple noun phrase.

4.4 Compound noun phrase

A compound noun phrase is formed by joining more than one simple noun phrase. By adding sixth inflexion at the end of a simple noun phrase we can form a compound noun phrase. At the end of a compound noun phrase, a case marker has to be added.

4.5 Complete noun phrase

A simple or compound noun phrase that has a case marker to create a case relation is called complete noun phrase.

4.6 Simple noun phrase

A simple noun phrase can have the following elements:

i. Noun
ii. Pronoun
iii. Adjective
iv. Number
v. Case marker
vi. Number inflexion
vii. Ordinal number
viii. Conjunction
ix. Sixth inflexion

5. Binary Relations and Attributes Used in Formation of a Noun Phrase

According to our research the following binary relations can be used for enconverting Bangla Noun phrase to UNL:

i. and (conjunction)
ii. or (disjunction)
iii. per (proportion, rate or distribution)
iv. pos (possessor)
v. qua (quantity)
vi. mod (modification)

List of attributes used in formation of noun phrase:

i. @def
ii. @indef
iii. @ordinal

6. Rules for Noun Phrase

Here some rules are developed to enconvert the noun phrase into UNL. The rules for noun phrase are given bellow:

6.1 Rule 1

If there is a subjective pronoun, a possessive pronoun, an objective pronoun or a possessive vowel ended or consonant ended noun a in a sentence then it will be considered as a noun phrase. The rules are -

Rule 1.1: R{PRON,SUBJ,^np:-PRON,np::}{BLK:::}
Rule 1.2: R{PRON,#POS,^np:-PRON,np::}{BLK:::}
Rule 1.3: R{PRON,#OBJ,^np:-PRON,np::}{BLK:::}
Rule 1.4: +{N,VEND,^#POS:#POS,-N,np::}
{BIV,6TH,VEND:::}
Rule 1.5: +{N,CEND,^#POS:#POS,-N,np::} {BIV,6TH,CEND:::}
Rule 1.6: +{np,VEND,^#POS:#POS::}
{BIV,6TH,VEND:::}
Rule 1.7: +{np,CEND,^#POS:#POS::}
{BIV,6TH,VEND:::}

In the word dictionary we entered an attribute to differentiate subjective, possessive and objective pronoun. So, when Rule 1.1, 1.2 and 1.3 will be executed it will change the attribute to noun phrase (np) and shift right. For Rule 1.4 and 1.5 it will create a composite node and change the attribute to noun phrase (np).
For Example:
[আমার]{}"i(icl>person)"(PRON,HPRON,1P,SG,#POS)<B,1,1>
Here, "আমার" pronounced as "amar" means "my" is a possessive pronoun. If it appears in the left analysis window it will be

transformed into a noun phrase.

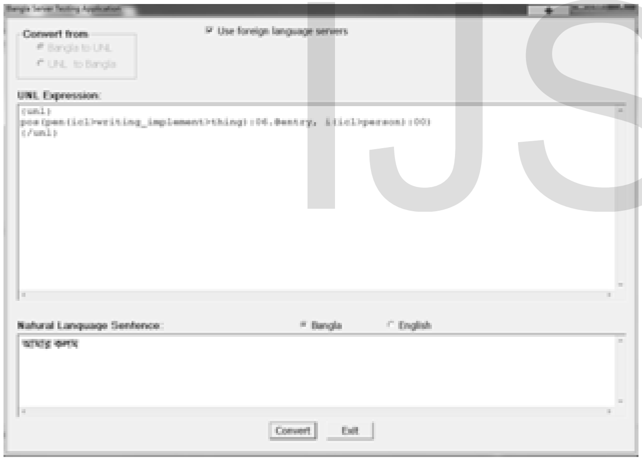
6.2 Rule 2

If we get a noun phrase with 6th inflexion in the left analysis window and another noun phrase in the right analysis window then this rule will work.
For example: "আমার কলম" pronounced as "amar kolom" means "my pen"
The following dictionary entries are needed for converting the above:
[আমার]{}"i(icl>person)"(PRON,HPRON,1P,SG,#POS) <B,1,1>
[কলম]{}"pen(icl>writing_implement>thing)"(N,NCOM,#OBJ,CEND) <B,0,0>
The rule is:
Rule 2.1: >{np,#POS::pos:}{np:::}
Rule 2.2: >{np,#POS::mod:}{np:::}

The UNL expression for the above example is:

```
{unl}
pos(pen(icl>writing_implement>thing):06.@entry,
        i(icl>person):00)
{/unl}
```

Enconverter output:


Bangla Server Testing Application
Convert from
Bangla to UNL
UNL to Bangla
☑ Use foreign language servers
UNL Expression:
Natural Language Sentence:
● Bangla
○ English
Convert
Exit

```
{unl}
pos(pen(icl>writing_implement>thing):06.@entry, i(icl>person):00)
{/unl}
```

আমার কলম

Figure 6: Enconverter output 1

6.3 Rule 3

If the left analysis window have a noun and the right analysis window have an article then the following rule will execute.
For example:
"পাখিটি" pronounce as "pakhiti" means "the bird"
The following dictionary entries are needed for converting the above:
[পাখি]{}"bird(icl>vertebrate>thing)"(N,VEND)
[টি]{} "" (ART)
The rule is:
+ {N,VEND,^ART:np,-N,-VEND,@def: : }{ART: : :}
Here, N for noun, VEND for vowel ended, ART for article.

6.4 Rule 4

If we get an adjective in the left analysis window and anything except noun in the right analysis window then this rule will work.
The rule is:
Rule 4.1: R{:::}{ADJ:::}
Rule 4.2: R{ADJ:::}{^N:::}P0
Here, N for noun and ADJ for adjective.

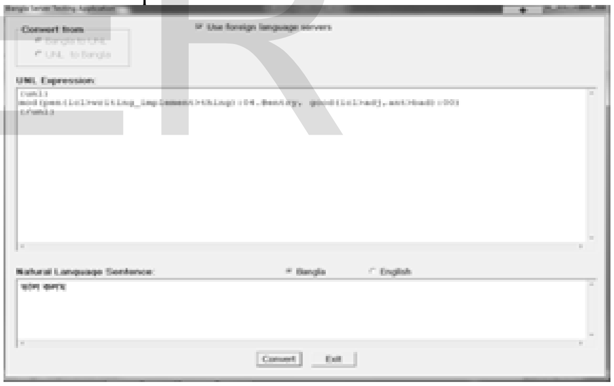
6.5 Rule 5

If we get an adjective in the left analysis window and a noun or noun phrase in the right analysis window then this rule will work.
For example: "ভাল কলম" pronounce as "bhalo kolom" means "good pen"
The following dictionary entries are needed for converting the above:
[ভাল]{}"good(icl>adj,ant>bad)"(ADJ,VEND)<B,0,0>
[কলম]{}"pen(icl>writing_implement>thing)"(N,NCOM, #OBJ,CEND)<B,0,0>
The rule is:
Rule 5.1: >{ADJ::mod:}{N,^np:np,-N::}P10
Rule 5.2: >{ADJ::mod:}{np:::}P10

The UNL expression for the above example is:

```
{unl}
mod(pen(icl>writing_implement>thing):04.@entry,good(icl>adj,ant>bad):00)
{/unl}
```

Enconverter Output:


Bangla Server Testing Application
Convert from
Bangla to UNL
UNL to Bangla
☑ Use foreign language servers
UNL Expression:
Natural Language Sentence:
● Bangla
○ English
Convert
Exit

```
{unl}
mod(pen(icl>writing_implement>thing):04.@entry, good(icl>adj,ant>bad):00)
{/unl}
```

ভাল কলম

Figure 7: Enconverter Output 2

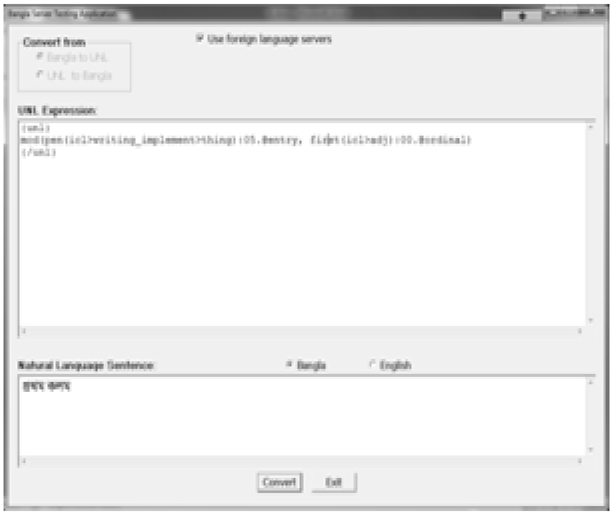
6.6 Rule 6

If we get ordinal number in left analysis window and noun phrase in the right analysis window this rule will work.
For example: "প্রথম কলম" pronounced as "prothom kolom" means "first pen"
The following dictionary entries are needed for converting the above:
[প্রথম]{}"first"(NUM,ORD)<B,1,1>
[কলম]{}"pen(icl>writing_implement>thing)"(N,NCOM, #OBJ,CEND)<B,0,0>
The rules are-
Rule 6.1: >{NUM,ORD:&@ordinal:mod:}{N:np,-N::}
Rule 6.2: >{NUM,ORD:&@ordinal:mod:}{np:::}

The UNL expression for the above example is-

```
{unl}
mod(pen(icl>writing_implement>thing):05.@entry,first(icl>adj):00.@ordinal)
{/unl}
```

Enconverter output:



Convert from
○ Bangla to UNL
○ UNL to Bangla

☑ Use foreign language servers

UNL Expression:

```
{unl}
mod(pen(icl>writing_implement>thing):05.@entry, first(icl>adj):00.@ordinal)
{/unl}
```

Natural Language Sentence: ● Bangla ○ English

প্রথম কলম

Figure 8: Enconverter Output 3

7. Conversion of a Bangla Noun Phrase into UNL

We have already discussed about all the rules in detail. Now we will show the enconvertion of a greater noun phrase as an example. The example noun phrase is "আমার প্রথম ভাল লাল কলম" pronounced as "amar prothom valo lal kolom" means "my first good red pen". Assuming that all the words and morphemes of the given sentence are in the dictionary as follows:

[আমার]{}"i(icl>person)"(PRON,HPRON,1P,SG,#POS)<B,1,1>
[প্রথম]{}"first"(NUM,ORD)<B,1,1>
[লাল]{}"red(icl>adj,equ>crimson)"(ADJ,CEND)<B,0,0>
[ভাল]{}"good(icl>adj,ant>bad)"(ADJ,VEND)<B,0,0>
[কলম]{}"pen(icl>writing_implement>thing)"(N,NCOM,
#OBJ,CEND)<B,0,0>

Enconverter (EnCo) can input either a string or a list of words for a sentence of a native language. A list of morphemes or words of a sentence must be enclosed by [<<] and [>>] [4, 5]. When the sentence is taken into EnCo, it places the sentence head (<<) in the LAW, sentence texts or morphemes or words in the RAW and the sentence tail (>>) in the RCW shown in the following figure.

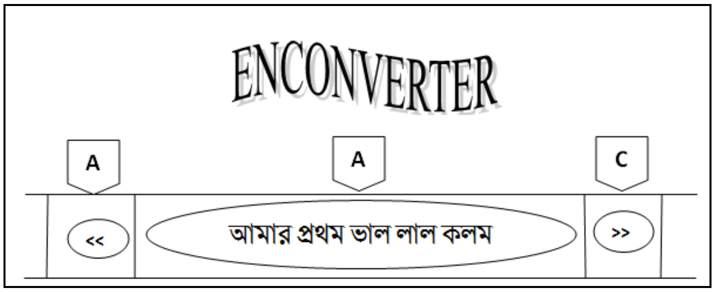


Figure 9: Example of Enconvertion process

After insertion of the input file (InputFile.txt) with our given sentence the following rules will be applied step by step to complete the conversion process of the sentence to UNLexpressions. One rule can be used more than once.

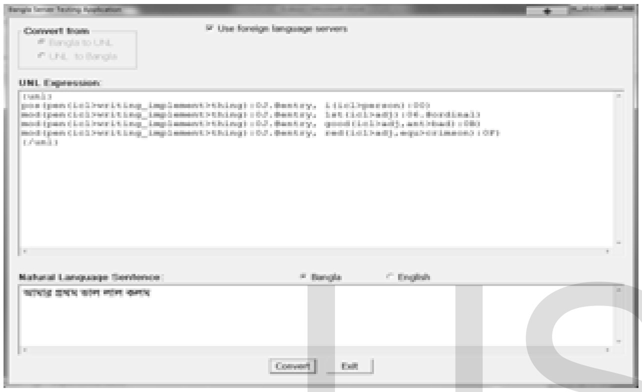
Step 1: R{SHEAD:::}{PRON:::}P2;
Step 2: R{PRON,#POS,^np:-PRON,+np::}{BLK:::}P0;
Step 3: R{:::}{^PRON,^N,^VERB,^ROOT,^ADJ,^ADV,
s^ABY,^KBIV,^BIV:+N,+PROP::}(BLK)P10;
Step 4: R{N:-N,+np::}{^biv,^BIV,^6TH:::}P0;
Step 5: DL{BLK:::}{:::}P10;
Step 6: R{:::}{ADJ:::}P0;
Step 7: R{ADJ:::}{^N:::}P0;
Step 8: DL{BLK:::}{:::}P10;
Step 9: R{:::}{ADJ:::}P0;
Step 10: R{ADJ:::}{^N:::}P0;
Step 11: DL{BLK:::}{:::}P10;
Step 12: >{ADJ::mod:}{N,^np:+np,-N::}P10;
Step 13: >{ADJ::mod:}{np:::}P10;
Step 14: >{NUM,ORD:+&@ordinal:mod:}{np:::}P0;
Step 15: DL{BLK:::}{:::}P10;
Step 16: >{np,#POS::pos:}{np:::}P0;
Step 17: R{:::}{np:::}(STAIL)P0;
Step 18: R{np:+&@entry::}{STAIL:::}P0;

Step 1 is Right Shift rule that describes that when sentence head is in the Left Analysis Window (LAW) and word 'আমার' (amar) is in the Right Analysis Window (RAW) then AWs will be shifted to right after rule application. In step 2 Right Shift rule will be applied again as the word 'আমার' (amar) is in the LAW and a blank space is in the RAW. Step 3 Shift the analysis window to right and the blank space comes to the LAW and the word 'প্রথম' (prothom) comes to the RAW. Step 4 shifts right again. Step 5 deletes blank space. Step 6 shifts the AWs to the right and LAW have the word 'প্রথম' (prothom) and the RAW have the word 'ভাল' (valo). Step 7 is right shift rule and LAW have the word 'ভাল' (valo) and RAW have blank space. Step 8 deletes the blank space. Step 12 creates a 'mod' relation between the word 'লাল' (lal) and 'কলম' (kolom). Step 13 creates a 'mod' relation between the word 'ভাল' (valo) and 'কলম' (kolom). Step 14 creates a 'mod' relation between the word 'প্রথম' (prothom) and 'কলম' (kolom) and includes an attribute 'ordinal'. Step 16 creates a 'pos' relation between the word 'আমার' (amar) and 'কলম' (kolom). Finally, in step 18 is applied to place the sentence tail (STAIL) on the LAW to complete the conversion process.

After executing all the rules the encoverter will convert the noun phrase into UNL as follows:

```
{unl}
pos(pen(icl>writing_implement>thing):0J.@entry,
i(icl>person):00)
mod(pen(icl>writing_implement>thing):0J.@entry,
first(icl>adj):06.@ordinal)
mod(pen(icl>writing_implement>thing):0J.@entry,
good(icl>adj,ant>bad):0B)
mod(pen(icl>writing_implement>thing):0J.@entry,
red(icl>adj,equ>crimson):0F)
{/unl}
```

The following screen will show the enconverter output:


Bangla Server Testing Application
Convert from
Bangla to UNL
UNL to Bangla
☑ Use foreign language servers
UNL Expression:
Natural Language Sentence:
● Bangla
○ English
Convert
Exit

```
{unl}
pos(pen(icl>writing_implement>thing):0J.@entry, i(icl>person):00)
mod(pen(icl>writing_implement>thing):0J.@entry, 1st(icl>adj):06.@ordinal)
mod(pen(icl>writing_implement>thing):0J.@entry, good(icl>adj,ant>bad):0B)
mod(pen(icl>writing_implement>thing):0J.@entry, red(icl>adj,equ>crimson):0F)
{/unl}
```

[illegible]

Figure 10: Enconversion

8. Conclusion

This research is about Bangla noun phrase and how to convert the noun phrases to UNL expression. Here a demonstration of converting a Bangla noun phrase has been shown by taking a noun phrase as an example. Although all the cases of noun phrase formation has been discussed but for better understanding a demonstration has been shown separately. By using these rules in UNL enconverter we can convert a noun phrase into an UNL expression.

Though there are so many exceptions in Bangla noun phrase structure, we tried to build a standard form. Our future work is handling those exceptions and makes it more convenient. Here we have developed rules for enconvertion. In future, we will develop the rules for deconvertion also.

I hope, this research will demonstrate our pioneering efforts in the field of Bengali (Bangla) language.

References

[1] S. Abdel-Rahim, A.A. Libdeh, F. Sawalha, M.K. Odeh, "Universal Networking Language(UNL) a Means to Bridge the Digital Divide", Computer Technology Training and Indistrial Studies Center, Royal Scientific Sciety, March 2002.

[2] Md. N. Ali, J.K. Das, S.M. A. Al-Mamun, Md. E. H. Choudhury,,"Specific Features of a Converter of Web Documents from Bengali to Universal Networking Language", International Conference on Computer and Communication Engineering 2008 (ICCCE'08), Kuala Lumpur, Malaysia.pp. 726-731.

[3] Md. N.Y. Ali, J.K. Das, S.M. A. Al-Mamu, A.M. Nurannabi, "Morphological Analysis of Bangla Words for Universal Networking Language", Third International Conference on Digital Information Management (ICDIM 2008), London, England.pp. 532-537.

[4] Md. E. H. Choudhury, Md. N. Y. Ali, M. Z.H. Sarkar, A. Razib, "Bridging Bangla to Universal Networking Language- A Human Language Neutral Meta-Language", International Conference on Computer, and Information Technology (ICCIT), Dhaka, 2005

[5] Wikipedia, (2012, August 07th). Introduction to UNL. Retrieved from http://www.unlweb.net/wiki/ index.php/Introduction_to_UNL.

[6] Jesús Cardeñosa, Alexander Gelbukh, Edmundo Tovar. "Universal Networking Language: Advances in Theory and Applications". February, 2012.

[7] Robin, (2012, August 29). Universal Networking Language. Natural Language Processing. Retrieved from http://language.worldofcomputing.net/unl/universal-networking-language-unl.html.