Specification of UNL Deconverter for Bangla Language

Aloke Kumar Saha, [1,] Muhammad F. Mridha, [1,] Manoj Banik,[2] and Jugal Krishna Das[3]
Department of CSE, University of Asia Pacific, Dhaka, Bangladesh[1,]
Department of CSE, Ahsanullah University of Science and Technology[2,]
Department of CSE, Jahangirnagar University, Savar, Dhaka, Bangladesh [3,]
aloke71@yahoo.com, mdfirozm@yahoo.com, mbanik99@yahoo.com, drdas64@yahoo.com

Abstract—At present the WWW represents a powerful tool for communication and information interchange. With simple mechanism, it is possible to access innumerable documents about a huge variety of topics, from any place around the world. However, despite the abundance of information, languages very often cause severe problems. When most of the web pages today are written in few most common languages like English, French, Chinese and Spanish etc, it becomes difficult for a person with insufficient knowledge of these languages to access and use this tool of communication and information. This has prompted the need to devise means of automatically converting the information from one natural language to another natural language, called Machine Translation (MT). In this paper we will discusses the interlingua approach to machine translation. Here Universal Networking Language (UNL) has been used as the intermediate representation. Here, we have specified a language independent deconverter for the Bangla language it takes as input a UNL (Universal Networking language) expression. The system takes a set of UNL expression as input and with the help of language independent algorithm and language dependent data generates corresponding Bangla sentence.

Index Terms— Universal Networking Language, morphology, morphological analysis, Bangla language, morphological rules, Deconverter, Syntax analysis.

1 Introduction

Today the regional economies, societies, cultures and educations are integrated through a globe-spanning network of communication and trade. This globalization trend evokes for a homogeneous platform so that each member of the platform can apprehend what other intimates and perpetuates the discussion in a mellifluous way. However the barriers of languages throughout the world are continuously obviating the whole world from congregating into a single domain of sharing knowledge and information.

As a consequence United Nations University/Institute of Advanced Studies (UNU/IAS) were decided to develop an inter-language translation program. The corollary of their continuous research leads a common form of languages known as Universal Networking Language (UNL).The Universal Networking Language (UNL) is a world wide generalizes form of human interactive language in a machine independent digital platform for defining, recapitulating, amending, storing and dissipating knowledge or information among people of different affiliations. The theoretical and applied research associated with this interdisciplinary endeavor facilitates in a number of practical applications in most domains of human activities such as creating globalization trends of markets or geopolitical interdependence among nations
For the purpose of conversion we use Interlingua which follow the UNL specifications proposed by UNU/IAS Tokyo. UNL (Universal Networking language) is a language used to represent a semantic graph equivalent of a concept (contained in text document). The system takes a set of UNL expression as input and with the help of language independent algorithm and language dependent data generates corresponding Bangla sentence.
The Universal Networking Language Programme started in 1996, as an initiative of the Institute of Advanced Studies (IAS) of the United Nations University (UNU) [1, 5] in Tokyo, Japan. The mission of the UNU program is to allow people across nations to access information in Internet in their own languages. The core of the project is UNL, a language independent specification for serving as a common medium for documents in different languages. Researchers involved in this project from different countries have been developing UNL system for their respective native languages. The goal is to eliminate the massive task of translation between two languages and reduce language to language translation to a one time conversion to UNL. For example, Bangla corpora, once converted to UNL, can be translated to any other language given UNL system built for that language. The UNL system does this by representing only the semantics of a native language sentence in a hypergraph. EnConverter [12] (parser) converts each native language sentence to a UNL hypergraph and DeConverter [13] translates from hypergraph to any native language. The main aim of the UNL project is to overcome language barriers. This project currently includes 16 official languages. Bangla is not yet included. We have attempted to demonstrate that we can do similar tasks for *Bangla* as it has been done for other official languages. In this proposal we present a new approach of NLP through UNL for Bangla Language.

2 Reason for using UNL

UNL is very powerful platform for language conversion. Once the information is converted to UNL form, it becomes language neutral and it can be converted to other different languages. Thus, it can be used for information exchange between languages. Information in a source language can be converted to UNL using source language Deconverter and then using Enconverter of target language, UNL can be enconverted in to that language. Since, UNL is in logical form, knowledge processing can be done unambiguously to produce useful and desired results.

It enables natural language phenomena to be expressed in formal semantic framework which enables computers to understand natural language. If the UNL is added to the network platforms, the communication status will be changed. UNL will make the communication among people through different Natural Languages possible, which will share information and provide a common educational environment as language is an essential part of the communication process. Communication between different nations will be easier since language barriers will be broken. Breaking language barriers, in turn, will result in, for example,

i) Encouraging mutual understanding among different cultures which is one of the ultimate goals of UNL. Sure, using foreign languages will make nations go through the risk of loosing a big part of their culture; consequently, as time goes, their roots will be lost as well. With the existence of UNL this risk will not exist.

ii) Communication through UNL will make the mission of international organizations, like United Nations and UNESCO, easier as they are concerned about all people with different mother tongues; one of the main problems faced in the exchange of information between the organizations and different nations is the existence of language barriers

3 UNL STRUCTURE

UNL is an artificial language that allows the processing of information across linguistic barriers [10]. This artificial language has been developed to convey linguistic expressions of natural languages for machine translation purposes. Such information is expressed in an unambiguous way through a semantic network with hyper-nodes. Nodes (that represent concepts) and arcs (that represent relations between concepts) compose the network. UNL contains three main elements:

• Universal Words: Nodes that represent word meaning.
• Relation Labels: Tags that represent the relationship between Universal Words i.e. between two nodes. Tags are the arcs of UNL hypergraph.
• Attribute Labels: Additional information about the universal words.

3.1 Universal Words

Universal Words are words that constitute the vocabulary of UNL. A UW is not only a unit of the UNL syntactically and semantically for expressing a concept, but also a basic element for constructing a UNL expression of a sentence or a compound concept. Such a UW is represented as a node in a hypergraph. There are two classes of UWs from the viewpoint in the composition:

• Labels defined to express unit concepts and called “UWs” (Universal Words)
• A compound structure of a set of binary relations grouped together and called “Compound UWs”.

3.2 Relational Labels

The relation [1] between UWs is binary that have different labels according to the different roles they play. A relation label is represented as strings of three characters or less. There are many factors to be considered in choosing an inventory of relations. The following is an example of relation defined according to the above principles.
Relation: There are 46 types of relations in UNL. For example, *agt* (agent), *agt* defines a thing that initiates an action, agt(do, thing), agt(action, thing), obj(thing with attributes) etc.

3.3 Attributes

The attributes represent the grammatical properties of the words. Attributes of UWs are used to describe subjectivity of sentences. They show what is said from the speaker’s point of view: how the speaker views what is said. This includes phenomena technically [4, 5] called speech, acts, propositional attitudes, truth values, etc. Conceptual relations and UWs are used to describe objectivity of sentences. Attributes of UWs enrich this description with more information about how the speaker views these state of affairs and his attitudes toward them.

3.4 UNL Format of Dictionary

The UNDL foundation provides a dictionary format. The Word Dictionary is a collection of the word dictionary entries. Each entry of the Word Dictionary is composed of three kinds of elements: the Headword (HW), the Universal Word (UW) and the Grammatical Attributes. A headword is a notation/surface of a word of a natural language that composing the input sentence and it is to be used as a trigger for obtaining equivalent UWs from the Word Dictionary in enconversion. An UW expresses the meaning of the word and is to be used in creating UNL networks (UNL expressions) of output. Grammatical Attributes are the information on how the word behaves in a sentence and they are to be used in enconversion rules. Each Dictionary entry has the following format of any native language word [5].

Data Format:
[HW] {ID} “UW” (Attribute1, Attribute2,…) <FLG, FRE, PRI>
Here, HW ← Head Word (Bangla word),
ID ← Identification of Head Word (omitable),

UW ← Universal Word,
ATTRIBUTE ← Attribute of the HW,
FLG ← Language Flag(we use B for Bangla),
FRE ← Frequency of Head Word, PRI ← Priority of Head Word

Each entry in Word Dictionary includes native language Head Word, corresponding UW, and the attributes. Attributes include grammatical and semantic attributes. An example of an entry in Bangla Language Word Dictionary Attributes can be:

[বই]{}“book(icl>thing)” (N,C,INANI,PHY)<N,0,0>

Here, [বই] is the Bangla Head Word, book(icl>thing) is UW and (N,C,INANI,PHY) is the attribute list.

Some example,
[শহ্র]{}“city(icl>region)” (N, PLACE) <B,0,0>
[আমি] {} “i(icl>person)” (1P,1SG,PRON, HPRON,SUBJ) <B,0,0>
[খাই]{}“eat(icl>consume>do)”(V,ROOT,SORNT) <B,0,0>
[ভাত]{}“rice(icl>cereal>thing)”(N,CONCRETE,HF) <B,0,0>
[প্রচুর] {} “huge(icl>many)” (ADJ) <B,0,0>
[পাখি] {} “bird(icl>animal>animate thing)” (N, ANI, SG, CONCRETE) <B,0,0>

3.5 UNL Deconverter

A "deconverter" is software that automatically deconvert UNL into native languages. It is important to achieve a high quality and correct results. It is also important that the basic architecture of the "deconverter" is widely shared throughout the world, in order to treat all languages with the same quality and precision standards. Technology developed for a language can be applied to other languages as long as the architecture is shared. A "Deconverter", which generates natural language from UNL, plays a core role in the UNL system. It is very significant that "deconverter" is capable of expressing UNL information with very high accuracy.

A tool called *DeCo* has been designed by UNU/IAS as a language independent generator that provides synchronously a framework for morphological, syntactic and semantic generation and word selection for UNL to natural language. Its structure has been shown in figure 1.

It can deconvert both context-sensitive and context-free languages. It uses target language specific Word Dictionary, Co-occurence Dictionary and Deconversion rules to generate the target language. So, developing a Deconverter for a language means developing dictionaries and writing deconversion rules, which are understood by the DeCo and these are language dependent.

The UNL expressions are converted in to semantic net called Node-net. The UWs are replaced with corresponding native language Head Words. If it is not possible to unambiguously decide the correct Head Word for a given UW, Co-occurence dictionary is used. Co-occurence dictionary contains more semantic information for proper word selection without the ambiguity. But the use of Co-occurrence dictionary is optional.

Node-net represents the hyper graph (a representation of UNL expressions) that has not yet been visited. Each node contains certain attributes initially loaded from the Language Dictionary and sometime generated by *DeCo* during runtime. These attributes can be read or deleted or new attributes can be added. This is governed by deconversion rules. Each node in the Node-net is traversed and inserted in to the Node-list.

Node-list shows the current list of nodes that the Deconverter can look at through its windows. Node-list includes two generation windows circumscribed by condition windows. At the initial stage before any deconversion rule application there are three nodes in the Node-list, *Sentence Head* node, *Entry* node and *Sentence Tail* node[13].

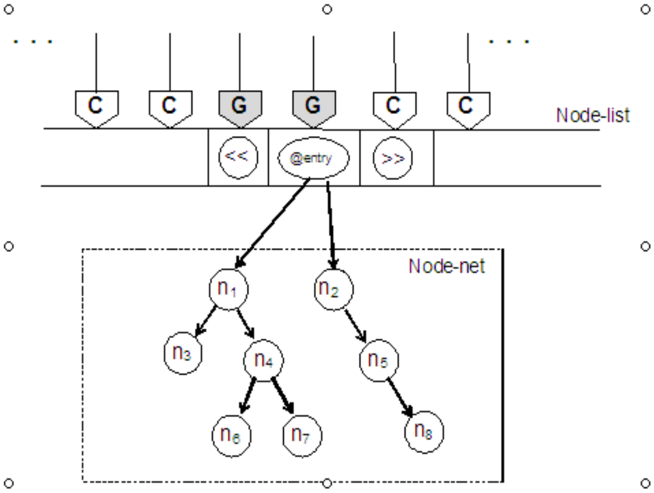

Figure 1 shows the initial state of the Generation Windows and the Node-list

The generation occurs at the generation windows, when the conditions in the condition windows are satisfied. The result of rule application is operation on the nodes in Node-list like changing attributes, copy, shift, delete, exchange etc. and/or insertion of nodes from Node-net to Node-list. The rule application halts when either Left Generation Window reaches the *Sentence Tail* node or Right Generation Window reached the *Sentence Head* node.

If post-editing is required the Deconverter will start applying post editing rules. Post editing rule has not been used for UNL Bangla Deconverter. At the end, the nodes in the Node-list represent the generated sentence [13].

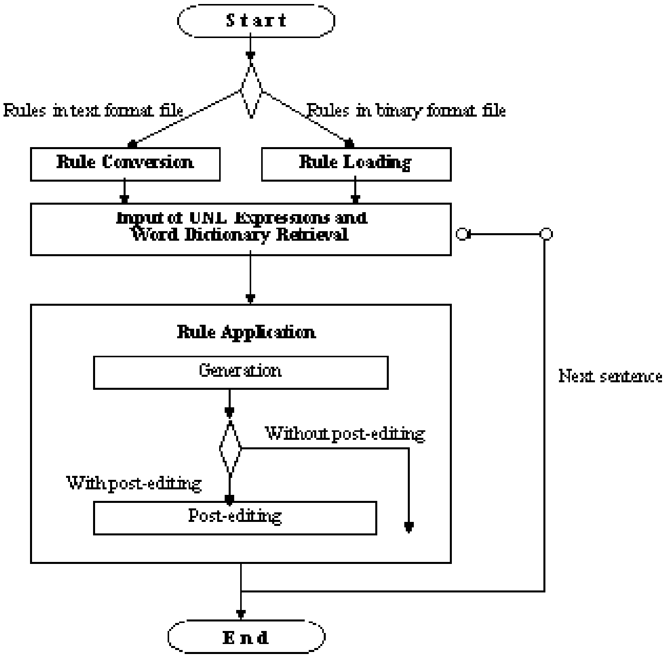

Figure 2: Flowchart of deconversion system

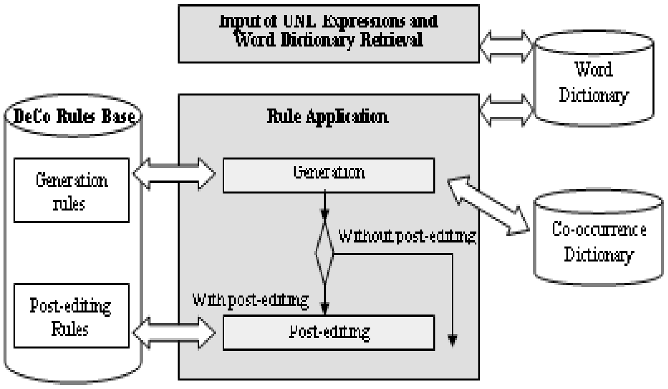

Figure 3: process and the dictionary and rule files.

4 Proposed Model

In our proposed model a "Deconverter", which generates Bangla language from UNL, plays a core role in the UNL system. It is very significant that "deconverter" will be capable of expressing UNL information with very high accuracy. It will consist of word dictionary and conversion rules for a language. This will be language independent software that is applicable for any languages. This engine takes UNL expression as input and generates target language (Bangla) sentence with the help of various database files like lexicon files, morphological rule files [1,2].

In our proposed model we have used two phases to convert from UNL to Bangla as:

1) Syntax analysis phase
2) Morphological phase

Syntax analysis module is responsible for Bangla sentence formation by syntax analysis phase. The syntax analysis phase is aimed at generation of proper sequence of words for the Bangla language. In order to get the correct Bangla sentence as the output of the DeConverter system, all the rules should be applied in proper order. It is also responsible for proper word formation though morphology generation. This module handles noun, verb and adjective morphology generation. This module not only inflects the root words, but also introduces conjunctions, case markers and any other new words if necessary.

The morphological rules are governed by UNL relations and attributes. Morphological rules due to UNL relations are called relation label morphology. Some relation label morphology rules have been shown in the table.

For example 1: ‘ Abul lives in Bangladesh’ ; relation ‘agt’ appears in relation between ‘Abul; and ‘live’ and ‘plc’ appears in shows relation between ‘live’ and ‘in Bangladesh’. Figure 4 Graph representation of above UNL expressions

UNL:
{unl}
aoj(live(icl>be,com>style,aoj>person,man>uw).@entry.@present,anis)
plc(live(icl>be,com>style,aoj>person,man>uw).@entry.@present,dhaka)
{/unl}
Outputঃ আবুল বাংলাদেশে বাস করে।

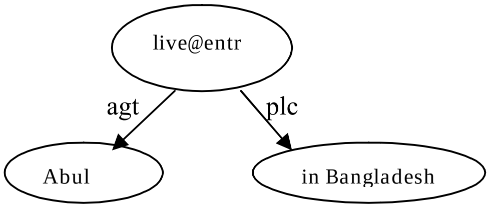

Figure 4: Graph representation of above UNL expressions

For example 2: ‘ Rahim is reading a book’ ; relation ‘agt’ appears in relation between ‘Ragim’ and ‘Reading’ and ‘obj’ appears in shows relation between ‘reading’ and ‘ a book’. Figure 5 Graph representation of above UNL expressions

UNL :
{unl}
agt(read(icl>see>do,agt>person,obj>information).@entry.@present.@progress,kerim(icl>name>abstract_thing,com>male,nam<person))

obj(read(icl>see>do,agt>person,obj>information).@entry.@present.@progress,book(icl>publication>thing).@indef)
{/unl}
Output: রহি্ম একটি বই পড়তেছে।

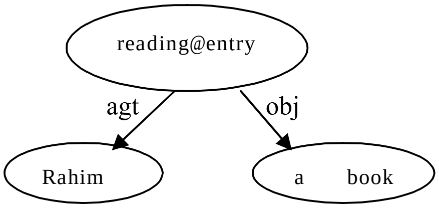

Figure 5: Graph representation of above UNL expressions

For example 3: ' I eat mango' ; relation 'agt' appears in relation between 'Ragim' and 'Reading' and 'obj' appears in shows relation between 'reading' and ' a book'. Figure 6 Graph representation of above UNL expressions

UNL:
{ unl}
agt(eat(icl>consume>do,agt>living_thing,obj>concrete_thing).
@entry.@present,i(icl>person))
obj(eat(icl>consume>do,agt>living_thing,obj>concrete_thing).
@entry.@present,mango(icl>grain>thing))
{/unl}
Output: আমি আম খাই।

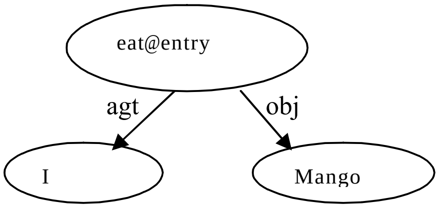

Figure 6: Graph representation of above UNL expressions

For example 4: ' Rahim and Karim are brother' ; relation 'aoj' appears in relation between 'Rahim' and 'brother' and 'aoj' appears in shows relation between 'brother' and ' Karim'. Figure 7 Graph representation of above UNL expressions

UNL :
{unl}
and(kerim(icl>name>abstract_thing,com>male,nam<person),rahim)aoj(brother(icl>person,pos>person).@entry.@pl.@present,kerim(icl>name>abstract_thing,com>male,nam<person))
{/unl}
Outputঃ রহি্ম এবং করিম ভাই।

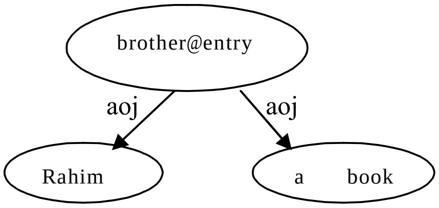

Figure 7: Graph representation of above UNL expressions

For example 4: ' He is going home' ; relation 'agt' appears in relation between 'He' and 'going' and 'plt' appears in shows relation between 'going' and ' home'. Figure 8 Graph representation of above UNL expressions

UNL :
{unl}
agt(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@present.@progress,he(icl>person))
plt(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@present.@progress,home(icl>building>thing))
{/unl}
Output: সে বাড়িতে যাইচেছে।

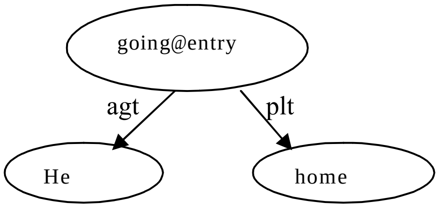

Figure 8: Graph representation of above UNL expressions

Syntax analysis is the process of linearizing the Semantic hyper-graph, i.e., it decides the word-order in the generated sentence. To make this process rule driven, we make several important assumptions:
The syntax analysis phase is aimed at generation of proper sequence of words for the target sentence. These phases first reads the input UNL file and convert it into semantic net like structure known as nodenet. We use lexicon files to map the UWs to target language worlds.
The process of deconversion involves syntax analysis phase and morphology phase. The syntax planning analysis is aimed at generation of proper sequence of words for the target sentence. These phases first reads the input UNL file and convert it into semantic-net like structure known as nodenet. Nodenet is a directed acyclic graph structure, which defines the sentence in the form of Directed Acyclic Graph. We use lexicon files to map the UWs to target language worlds. After generating a nodenet, the problem of the syntax plan generation get reduce to the problem of Directed Acyclic Graph traversal. Proper traversal of the node net generates the syntax plan of the target sentence. This syntax plan needs to be processed by the case-marking file, which apply proper case marker for each and every relations. This case-marking phase is next processed by the morphology phase. The morphology phase gives a final form of the target sentence.

Deconversion (or "generation" in general) rules describe the conditions for rule application: the way of rewriting the attributes of nodes that satisfy those conditions, as well as the way of composing a native language sentence. DeConverter looks at, and operates on, the nodes in the Node-list through its windows, and the conditions and actions of a deconversion rule are matched to the windows [1]. Each part of the rule expresses the conditions of, or actions on, the adjacent nodes in the Node-list in the order of the Left Condition Windows (LCW or PRE), the Left Generation Window (LGW), the Middle Condition Windows (MCW or MID), the Right Generation Window (RGW), and the Right Condition Windows (RCW or SUF).

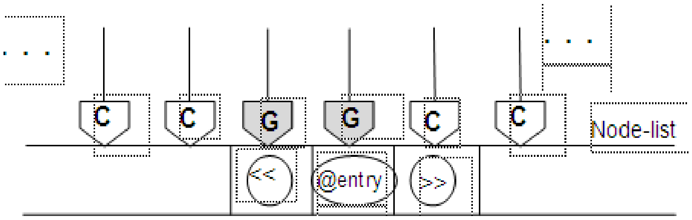

Figure 9: Initial state of the GW and the Node-list

DeCo can input either a string or a list of words form UNL to convert its native language. A list of entry node from UNL must be enclosed by [<<] and [>>] [6]. When we input the word into DeCo, the Sentence Head (<<) will be on LGW, sentence texts/morphemes/words will be on RGW and the Sentence Tail (>>) will be on Right Condition Window (RCW) shown in figure 9. DeCo uses CWs for checking the neighbouring nodes on both sides of the GWs in order to judge whether the neighbouring nodes satisfy the conditions for applying a generation rule or not.

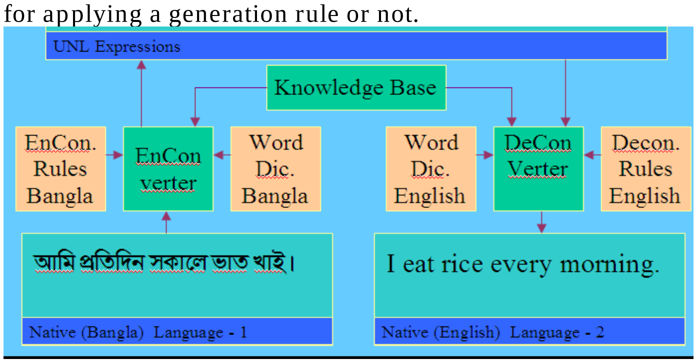

Figure 10: UNL process

6 Conclusion

This paper has described the development of UNL Bangla Deconverter, a Bangla language generator. Using the UNL system with its language components it has been proved to be a powerful environment for man machine communication. On the other hand, other machine translation systems will not be able to provide such environment for education and exchange of information as they are away from universality. They will never be inter-lingual. This will make their value limited to only the one or two languages involved in the translation. Consequently, communication and the distribution of information will be negatively affected. However one drawback can be that UNL has not yet conceived as a fully automatic machine translation system.

The Bangla language could successfully be generated from UNL hyper semantic networks with a high degree of accuracy. The main skeleton of Bangla sentence structure has been handled however many problems remain unsolved such as generating passive structures, correct ordering of modifiers of the same type, selecting the correct word representing universal words which represents the main challenges of the future work.

References

[1] Muhammad Firoz Mridha, Mohammad Nurul Huda, Chowdhury Mofizur Rahman, Jugal Krishna Das, “Development of Morphological Rules for Bangla Root and Verbal Suffix for Universal Networking Language”. ICECE 2010, 18-20 December 2010, Dhaka, Bangladesh.

[2] Muhammad Firoz Mridha, Manoj Banik, Md. Nawab Yousuf Ali, Mohammad Nurul Huda,Chowdhury Mofizur Rahman, Jugal Krishna Das, “Formation of Bangla Word Dictionary Compatible with UNL Structure,” SKIMA’10, Paro, Bhutan, August, 2010.

[3] H.Uchida, M. Zhu, and T. C. D. Senta, Universal Networking Language, NDL Foundation, International environment house, 2005/6, Geneva, Switzerland.

[4] D. M. Shahidullah, “Bangala Vyakaran”, Maola Brothers Prokashoni, Dhaka, August 2003, pp.110-130

[5] Muhammad Firoz Mridha, Kamruddin Md. Nur, Manoj Banik and Mohammad Nurul Huda, “Structure of Dictionary Entries of Bangla Morphemes for Morphological Rule Generation for Universal Networking Language”. International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) 2011

[6] Muhammad Firoz Mridha, Kamruddin Md. Nur, Manoj Banik and Mohammad Nurul Huda, “Generation of Attributes for Bangla Words for Universal Networking Language(UNL)”. International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) 2011.

[7] H. Uchida, M. Zhu, “The Universal Networking Language (UNL) Specification Version 3.0”, Technical Report, United Nations University, Tokyo, 1998

[8] Muhammad Firoz Mridha, Manoj Banik, Md. Nawab Yousuf Ali, Mohammad Nurul Huda,Chowdhury Mofizur Rahman, Jugal Krishna Das, “Formation of Bangla Word Dictionary Compatible with UNL Structure,” SKIMA’10, Paro, Bhutan, August, 2010.

[9] S. Dashgupta, N. Khan, D.S.H. Pavel, A.I. Sarkar, M. Khan, “Morphological Analysis of Inflecting Compound words in Bangla”, International Conference on Computer, and Communication Engineering (ICCIT), Dhaka, 2005, pp. 110-117

[10] Bangla Academy, “Bengali-English Dictionary” Bangla Academy Dhaka 2007

[11] http://www.unl.ru/deco

[12] EnConverter Specification, Version 3.0, UNL Center, UNDL Foundation, Tokyo 150-8304, Japan 2002

[13] DeConverter Specification, Version 2.7, UNL Center, UNDL Foundation, Tokyo 150-8304, Japan 2002.