Framework for Synthesis of Universal Networking Language

Md. Ershadul H. Choudhury
American International University Bangladesh,
Md. Nawab Yousuf Ali
East West University

Abstract

This paper presents the specifications of 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 languages–a step to help bridge the digital divide. 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 systems for their respective native languages. The process basically involves i) building native language to UNL dictionary and ii) deriving language specific syntactic rules called analysis rules for parsing/ translating native language corpora to UNL and vice versa. In this paper we present parallel work for developing a framework for synthesizing Bangla to UNL that involves building a Bangla to UNL dictionary and parsing sentences to UNL. To the best of our knowledge, this is a pioneering work in Bangla.

Keywords

Universal Networking Language, Bangla-UNL dictionary, morphological analysis, universal words, UNL document, parsing

1. Introduction

Although, there is an immense proliferation of information on the Internet, it is still not accessible to the vast multitude of people across nations as most of the resources are in English. To overcome this problem, United Nations has launched the Universal Networking Language project [1] under the auspices of United Nations University, Tokyo. The project team, after reviewing all such previous attempts has developed a universal networking language (UNL), a language neutral specification, and universal parser specification [4], which is considered to be a milestone in overcoming the language barrier for web publication. 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 the UNL system built for that language. The strength of the UNL system lies in the fact that it emphasizes the semantics of a native language sentence, ignoring the complexities of natural languages. An enconverter converts each native language sentence to a UNL document and deconverter translates the UNL document to any native language. The UNL document is itself in English as it is known to linguistics. The development of the native language, specific components-dictionary and analysis rules-is carried out by researchers across the world.

The UNL project currently includes 16 official languages, including Arabic, Chinese, English, French, Russian, Hindi but no work has yet done on Bangla. The infrastructure of UNL is presented in Fig. 1 which shows that UNL documents can be converted to any natural languages through language servers. On the contrary, existing web documents are in HTML and XML and present documents only in English.

In this paper we present a framework to integrate UNL system with Bangla. The main objectives of our research work is i) specification of Bangla-UNL dictionary; ii) development of analysis rules; and iii) translation scheme (parsing). In Sections 2, 3, 4, 5, 6, and 7 we describe the UNL system. In Sections 8 and 9, we present our main work.

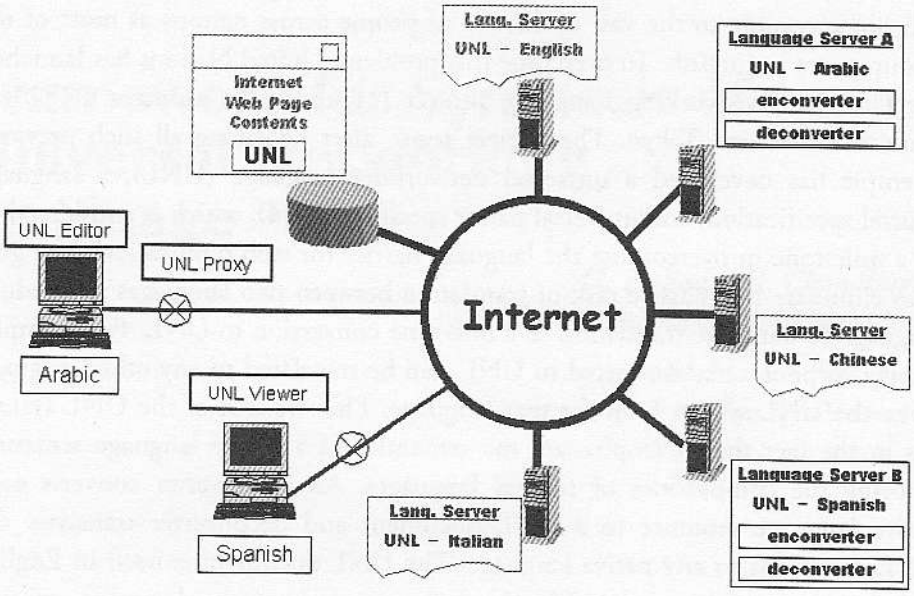

Figure 1. The Infrastructure of the UNL System

2. Universal Networking Language Specification

The UNL [1] has been introduced as a digital meta-language for describing, summarizing, refining, storing and disseminating information in a machine-independent and human language neutral form. This language intends to express meanings in standardized way. We think that a comprehensive description of UNL specification is necessary though it is available in the UNL website. The meaning of a native language sentence is expressed in UNL system as a hypergraph composed of nodes connected by semantic relations. Nodes or Universal Words (UWs) are words loaned from English and disambiguated by their positioning in a knowledge base (KB) of conceptual hierarchies. Function words, such as determiners and auxiliaries, are represented as attributes to UWs or nodes to provide additional information. The core structure of UNL is based on the following elements:

- Universal Words: Nodes that represent word meaning
- Attribute Labels: Additional information about the universal words
- Relation Labels: Tags that represent the relationship between Universal

Words i.e. between two nodes tags, are the arcs of the UNL hypergraph.

2.1 Universal words

Universal words constitute the vocabulary of UNL and 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 this 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".

A UW is a English-language word followed by a list of constraints. The following is the syntax of description of UWs in context-free grammar (CFG):

<UW> ::= <headword> [<constraint list>]
<headword> ::= <character>...
<constraint list> ::= "(" <constraint> ["," <constraint>]... ")"
<constraint> ::= <relation label> { ">" | "<" } <UW> [<constraint list>] |
<relation label> { ">" | "<" } <UW> [<constraint list>]
[{ ">" | "<" } <UW> [<constraint list>]] ...
<relation label> ::= "agt" | euatation and" | "aoj" | "obj" | "icl" | ...

2.2 Headword

The headword is an English word/compound word/phrase/sentence that is interpreted as a label for a set of concepts: the set is made up of all the concepts that may correspond to that in English. A basic UW (with no restrictions or constraint list) denotes this set. There are restricted UW's that are defined by a constraint list. Extra UWs denote new sets of concepts that do not have English-language labels.

2.3 Types of Universal Words

A UW is an English language word with restrictions. UWs do not allow semantic ambiguity as a first principle. The reasons why English words are employed in UW construction are that (i) English is known by all UNL developers; (ii) and there are a lot of good bilingual dictionaries between a local language and English. A UW can express various levels of concepts depending on the restrictions and can be used to express a more specific or particular concept or an instance by giving attributes. The UWs are based on five concepts:

2.3.1 Basic UWs

These are bare headwords with no constraint list.

2.3.2 Restricted UWs

Restricted UW's are headwords with a constraint list. Examples are given below:
state(icl>express(agt>thing, gol>person, obj>thing))
state(icl>country)
state(icl>region)
state(icl>abstract thing)
state(icl>government)

2.3.3 Extra UWs

These are special type of restricted UW; for example:
ikebana (icl>flower arrangement)

2.3.4 Temporary UWs

Such concepts are not necessary to define. For example: http://www.undl.org/

2.3.5 Compound UW's

These are a set of binary relations that are grouped together to express a compound concept. A sentence itself is considered as a compound UW. Compound UWs denote compound concepts that are to be interpreted/understood as a whole so that one can talk about their parts all at the same time. A compound UW is expressed by a scope in UNL expressions. In the example below, ":01" indicates all of the elements that are to be grouped together to define compound UW number 01. An example and translation to UNL is given below:

Women who wear big hats in movie theaters should be asked [to leave].

The UNL translation is as follows:

agt:01(wear(aoj>thing,obj>hat), woman(icl>person).@pl)
obj:01(wear(aoj>thing,obj>hat), hat(icl>wear))
aoj:01(big(aoj>thing), hat(icl>wear))
plc:01(wear(aoj>thing,obj>hat), theater(icl>facilities))
mod:01(theater(icl>facilities), movie(icl>art))
agt:01(leave(agt>thing,obj>place).@entry, woman(icl>person).@pl)

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 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.

For example, the corresponding UW of play is "play (icl>do)". If the word "play" is in the past form in the sentence an attribute @past is tagged with "play (icl>do)". If it is the main word in the sentence then @entry will be tagged such as "play (icl>do), @entry, past".

4. Relational Labels

The relation 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: agt (agent)
Agt defines a thing that initiates an action.
agt (do, thing)
agt (action, thing)
Syntax:
agt[":"<CompoundUW-ID>]"("{<UW1>|":"<Compound UW-ID>} "," {<UW2>|":"<Compound UW-ID>} ")"
An agent is defined as the relation between
UW1 - do, and
UW2 - a thing
Here UW2 initiates UW1, or UW2 is thought of as having a direct role in making UW1 happen.

Examples of "agt" relation:

John breaks ...	agt(break(agt>thing,obj>thing), John(icl>person)
Mary broke the window	agt(break(icl>do).@entry.@past, Mary)

Some other relations in UNL are as follows:

aoj (thing with attribute)
bas (standard (basis) of comparison)
cag (co-agent)
con (condition)
dur (duration)
equ (equivalent)
gol (goal: final state)
iof (an instance of)
mod (modification)
plc (place)
pur (purpose or objective)
rsn (reason)
src (source: initial state)
tim (time)

5. Unl Expression

The UNL expresses information or knowledge in the form of semantic network. UNL semantic network is made up of a set of binary relations where each binary relation is composed of a relation and two UWs that hold the relation. A binary relation of UNL is expressed in the following format:

<relation> (<uw1>, <uw2>)

In <relation>, one of the relations defined in the UNL specifications is described. In <uw1> and <uw2> the two UWs that hold the relation given at <relation> are described.

6. Hypergraph

The UNL expression is a hyper-semantic network. That is, each node of the graph, <uw1> and <uw2> of a binary relation can be replaced with a semantic network. Such a node consists of a semantic network of a UNL expression and is called a "scope". A scope can be connected with other UWs or scopes. The UNL expressions in a scope can be distinguished from others by assigning an ID to the <relations> of the set of binary relations that belong to the scope. The general description format of binary relations for a hypernode of UNL is the following:

<relation> :< scope-id> (<node1>, <node2>)

Where,

- <scope-id> is the ID for distinguishing a scope. <scope-id> is not necessary to specify when a binary relation does not belong to any scope.
- <node1> and <node2> can be a UW or a <scope node>.
- A <scope node> is given in the format of ":< scope-id>".

7. Knowledge Base

The UNL Knowledge Base (KB) gives possible binary relations between UWs. The knowledge base is a set of knowledge-base entries. The format of knowledge-base entries is as follows.

<Knowledge Base entry>::= <Binary relations> "=" <degree of certainty><Binary Relation> ::= <Relation Label> "("<UW1>","<UW2>")"<degree of certainty> ::= "0" | "1" | ... | "255"

When the degree of certainty is "0", it means the relation between two UWs is false. When the degree of certainty is more than "1", it means the relation between two UWs is true, and the bigger the number is, the more the credible it is.

The UW system has been introduced to:

- generate a word when a concept is not included in a language and
- reduce the number of knowledge-base entries which can be deductively inferred.

For this purpose the "icl" relation was introduced so that it could inherit properties from upper UW's. Each UW is categorized according to the role of concept to other concepts. For example, a proposition such as "A dog can eat food." is expressed in the following manner:

icl(icl>animal), animal(icl>living thing))=1

agt (eat(icl>do(obj>thing), animal(icl>living thing))=1

obj(eat(icl>do(obj>thing), food(icl>functional thing))=1

verb roots. The next steps form new nouns and adjectives. We have examined derivational morphology for UNL-Bangla dictionary too. Examples of both forms are given for the root word "kar".

The UNL form:

kar]{} "do(icl>do)"(List of Semantic and Syntactic Attribute)<B,0,0>;

Inflectional Morphology:
[-ebe] "ebe" (VMORP, FUTURE),<B,0,0>

Derivation Morphology:
[a] "a" (NMORP),<B,0,0>

In the following we have given morphological analysis of a Bangla verb word. We can select the head words as the Longest Common Lexical Unit (LCLU) of all the possible transformations of the word. We can give the example of the Bangla word "ja" which means "to go". The corresponding UW in basic form is "go". The dictionary entry is [ja] { } "go (icl>do)" where 'ja' is the head word and (icl>do) is from the knowledge base. Some possible transformations of 'ja' in the Bangla to UNL dictionary are given as follows [3, 9]:

A snapshot of the Bangla to UNL dictionary can be seen in what follows:
[ja] { } "go (icl>do)" (V, 3P, @present) <B, 0, 0>
// root word is "ja". Other derivations follow:
[-i] { } "go (icl>do)" (V, 3P, present) <B, 0, 0>
[gelo] { } "go (icl>do)" (V, 3P, @past) <B, 0, 0>
[-be] { } "go (icl>do)" (V, 3P, @future) <B, 0, 0>
...

Such dictionary order with root word followed by derivations will help in any quick search to find UW and the attributes of a Bangla word.

8.3 Bangla-English Dictionary

Bangla to English dictionary is the source of building a Bangla to UNL dictionary as universal words are English words mandated by UNL. Such dictionaries also provide all attributes along with the meaning of a word. Any entry in the dictionary is put in the following format:

[HW] {ID} "UW" (ATTRIBUTE1, ATTRIBUTE2 . . .) <FLG, FRE, PRI>

8. Bangla-UNL Dictionary

We present below the framework of Bangla-UNL dictionary on the basis of morphological analysis, standard bi-lingual dictionaries, and the UNL specified knowledge base.

8.1 Morphological Analysis

The first task is to analyze Bangla morphological structures. Morphological analysis results in lexical structure or hierarchy of the local vocabularies with the entries of the root word and morphemes [7,8]. Morphology is concerned with the ways in which words are formed from basic sequences of morphemes. It acts as the crossroads between phonology, lexicon, syntax, semantics, and context. Two types are distinguished:

- Inflectional Morphology
- Derivational Morphology

Inflectional Morphology defines possible variations on a root form. Inflectional morphology produces or derives words from another word form acquiring certain grammatical features but maintaining the same part of speech or category. Bangla has a very strong structural inflectional morphology for its verb forms. For example, there are nearly (10x5) forms for a certain verb-word in Bangla with 10 tenses and five persons and a root verb changes in form according to tense and person. For example roots such as "ja" or "kha" has such forms [7]

8.2 Morphological Analysis of Verbs

Morphological analysis is applied to identify the actual meaning of the word by identifying the suffix or morpheme of that word. Every word is derived from a root word. A root word may have different transformations. This happens because of different morphemes which are added with it as suffixes. So the meaning of the word varies for its different transformations. For example, if we consider 'kor' as a root word then after adding 'echilo' we get the word 'kor-echilo' which means a work done in the past. Similarly after adding 'be', we get the word 'kor-be'. Here, one word represents the past indefinite tense of the root word 'kor' and another represents the future indefinite tense. Therefore, by morphological analysis we get the grammatical attributes of the main word. For this reason we have applied morphological analysis to find out the actual meaning of the word.

Derivational morphology is simple and a word rarely uses the derivational rule in more than two or three steps. The first step forms nouns or adjectives from

Here,

HW ← Head Word (Bangla word)
ID ← Identification of Head Word (omitable)
UW ← Universal Word
ATTRIBUTE ← Attribute of the HW
FLG ← Language Flag
FRE ← Frequency of Head Word
PRI ← Priority of Head Word

Some example entries of dictionary for Bangla are given below:
shohor {} "city(icl>region)" (N, PLACE) <B,0,0>
prochur {} "huge(icl>big)" (ADJ) <B,0,0>
Here the attributes,

N stands for Noun
PLACE stands for place
ADJ stands for Adjective
FLG field entry is B which stands for Bangla

A universal knowledge base is defined in UNL specification. This knowledge base is language independent and each native language word should be referenced to this knowledge base. The knowledge base of universal words is a hierarch of concepts.

9. Enconversion System

To translate Bangla sentences into UNL form, we have to use EnCo, a universal converter system [4] built by the UNL organization. The process model known as enconverter system is given in Fig. 2. It is a language independent parser which provides morphological and syntactic analysis synchronously.

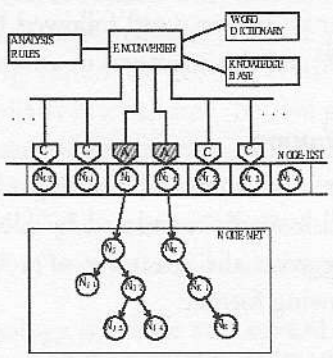

Figure 2: Structure of Enconverter of UNL Project

From Figure 2, it can be seen that analysis rules for parsing, Bangla to UNL dictionary and knowledge base of UNL are the components for converting Bangla to UNL. EnCo operates on the nodes of the node-list through its windows which are of two types, namely, analysis window (AW) and condition window (CW). The analysis windows are used to check two adjacent nodes in order to apply one of the analysis rules. If there is an applicable rule, EnCo adds lexical attributes to or deletes lexical attributes from these nodes, and/or creates a partial syntactic tree and/or UNL network according to the type of the rule.

9.1 Enconvertion of Bangla Expression to UNL

We can use a language independent parser like EnCo for Bangla to UNL conversion. For that purpose, we have to provide the grammatical rules called analysis rules for Bangla. The universal parser [4] of the EnCo scans input Bangla sentences from left to right. In this parsing it will identify each word and find out its corresponding UW, attribute and morphemes. It will then form the relations between the UWs with the help of the given grammatical rules. An example of Bangla to UNL conversion is given below.

A simple Bangla expression is:
"oddhapok Ali gotokal shokal-e gram-er gorib-der ortho diechen"

According to [3], the corresponding UW and head word of each Bangla word is given below:

Gorib {} "poor (icl>group)"
Oddhapok {} "professor (icl>title)"
Die { } "give (icl>do)"
Gotokal {} "yesterday (icl>day")
Ortho {} "money (icl>thing)"
Shokal {} "morning (icl>heavenly phenomenon)"
Gram {} "village (icl>region)"
Ali {"Ali (icl>male)"

The UNL relations among the UWs i.e. the final UNL expression, is as follows:

aoj(professor(icl>title), Ali(icl>person))
mod(yesterday(icl>day),morning(icl>heavenly phenomenon))
tim(give(icl>do).@entry.@past, yesterday(icl>day))
to(money(icl>functional thing), poor(icl>group).@pl)
obj(give(icl>do).@entry.@past,money(functional thing))

plc(poor(icl>group).@pl, village(icl>region))
agt(give(icl>do).@entry.@past, Ali(icl>person))

The hypergraph of the sentence is shown in Figure 3

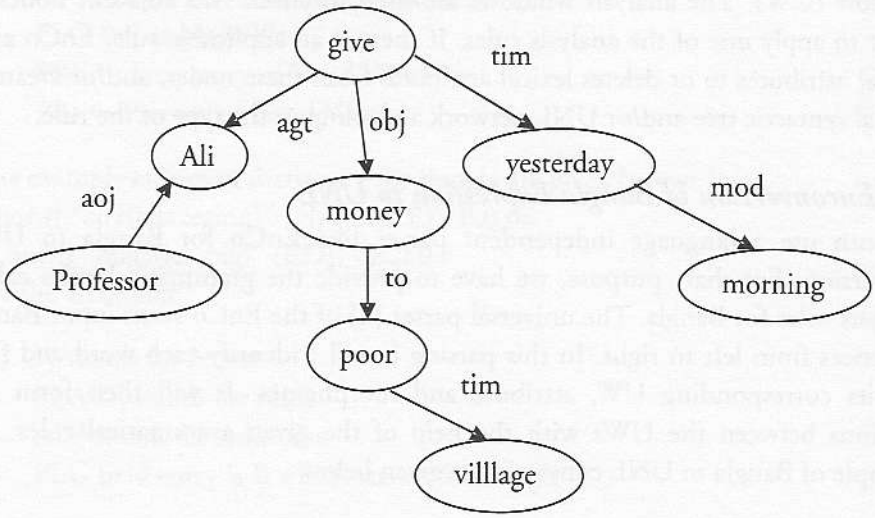

Figure 3. The UNL Hypergraph

9.2 The Encoding of a Bangla Conjunctive Sentence

The encoding process will be performed by shift/reduce parsing [5]. We have observed that the Hindi language has syntactic similarities with Bangla and the Hindi to UNL [6] system developed at the Indian Institute of Technology, Bombay, can serve as a reference for us. Here is an example of encoding a simple Bangla sentence. We assume that analysis rules and the dictionary of Bangla to UNL are given to the analyzer system EnCo.

"Karim ebong Rahim school-e ja-be."

The node list is shown here within "<<"and ">>". The analysis window is within "["and "]". The nodes delimited by "/" are those explored and fixed by the system.
/<</[Rahim]/[ebong]/"Karim school-e ja-be"/>>/

The noun "Rahim" and the conjunction "ebong" will be combined and a "and" relation can be made between "Rahim" and other noun or pronoun.
/<</[Rahim ebong]/[Karim]/"school-e ja-be"/>>/

A "and" relation will be made between two nouns "Rahim" and "Karim" and then "ebong Karim" will be deleted/reduced. Current sentence in the analysis windows are as follows:
/<</[Rahim]/[school-e]/" ja-be"/>>/

"Rahim" cannot be resolved with "school" because there is no Bangla syntactic rule like noun followed by another noun except with a "," or conjunction like "ebong" in between. Therefore, the analyzer looks ahead further right (make right shift) to get a verb "ja". Since the noun "school" can be resolved with the verb "ja", a reduction in the action will take place because of the relation between the noun and verb.

/<</Rahim /[school-e]/[ja-be]/>>/

An "obj" relation is created between the noun "school-e" and the verb "ja-be" and "school-e" is deleted.

/<</ [Rahim]/[ja-be]/>>/

Here, "agt" relation is created between "Rahim" and the verb "ja-be".

/<</ jabe/ [>>]/

A right shift is performed and the right shift rule attaches an attribute @entry to the last word "ja-be" left in the node list. Bangla to UNL dictionary has been searched to replace "ja-be" by the UW "(go (icl>do).@future)". A verb is the main word of a sentence and most of the relations are created by the main word. The UNL output of the corresponding sentence is:

and (Karim(icl>male),Rahim(icl>male))
obj(go(icl>do).@entry.@future,school(icl>place))
agt(go(icl>do).@entry.@future, Rahim(icl>male))

We have also verified that some other type of sentences such as assertive and interrogative sentences can be translated similarly.

10. UNL Project For Bangla

The authors of this paper of the Computer Science and Engineering Department of East West University have undertaken the work of synthesizing Bangla to UNL [10] and are working on constructing a Bangla-UNL dictionary and analysis rules (Bangla to UNL translation rules). The first author has become a member of the

UNL Society of Universal Networking Digital Language (UNDL) Foundation which permits members to use the official resources. However, a UNDL Foundation recognized UNL Centre for Bangla is yet to be established.

11. Conclusion

Our preliminary work gives indication that Bangla-UNL system can be built like similar systems built for other official languages. The following conclusions can be arrived at:

- The UNL system is a step towards multilingual translation and the system is shown to be workable.
- Emphasis on semantics of natural languages is the most important attribute of UNL
- It is an instrument towards overcoming the digital divide.
- UNL has the potential to become a language of the web like HTML and XML. UNL document can be converted to any natural language for which the system is built but HTML and XML presents documents and data only in English.
- Our introductory work – UNL concept testing for Bangla – can serve as a basis for further research and development work.
- We have explained in brief how to build Bangla-UNL dictionary and all previous works on Bangla morphological analysis can be of help.
- We have given an example of how to discover rules for Bangla-UNL conversion which includes morpho-syntactic and semantic phenomena. The process can be automated.
- A UNL center can be established for Bangla like the sixteen other official of the UNL project languages to perform effective research and development and coordinate activities among researchers, developers, and the Universal Networking Digital Language (UNDL) Foundation.

References

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

[2] Shibprashanna Lahiri, (1999) *Beboharik Vasha Bichitra,* Shahitya Prkashoni, Dhaka, 1999

[3] Bangla Academy (2004), Bengali-English Dictionary, Dhaka

[4] UNU/UNL Centre, (2000) Enconverter Specification Version 2.1, Tokyo 150-8304, Japan

[5] Earley, J., (1970) An Efficient Context Free Parsing Algorithm, *Communications of the ACM,* 13(2), 1970

[6] Bhattacharyya, (2001) Multilingual Information Processing Using Universal Networking Language, in *Indo UK Workshop on Language Engineering for South Asian Languages* LESAL, Mumbai, India

[7] Dashgupta Sajib, Khan Mumit (2005) Morphological Parsing of Bangla Words using PCKIMMO, *International Conference on Computer and Information Technology (ICCIT).,* Dhaka,

[8] Asaduzzaman, M.M., Ali, Mohammed Mashroor (2003) Morphological Analysis of Bangla Words for Automatic Machine translation, *International Conference on Computer, and Information Technology (ICCIT), Dhaka*

[9] Serrasset Gilles, Boitel Christian, (1999) UNL-French Deconversion as Transfer & Genera-tion from an Interlingua with Possible Quality Enhancement through Offline Human Interac-tion. *Machine Translation Summit-VII,* Singapore

[10] Choudhury Md. Ershadul, Ali Md. Nawab Yousuf, Sarkar, Zakir Hossain, Razib, Ahsan (2005), Bridging Bangla to Universal Networking Language- A Human Language Neutral Meta-Language, *International Conference on Computer, and Information Technology (ICCIT), Dhaka,*