Conversion of Bangla Sentence into Universal Networking Language Expression

Md. Nawab Yousuf Ali[1], Mohammad Zakir Hossain Sarker[2] , Ghulam Farooque Ahmed[3] , Jugal Krishna Das[4]

**[1] Department of Computer Science and Engineering, East West University
Dhaka, Bangladesh**

**[2] MIS/IT Unit, USAID | DELIVER PROJECT
Dhaka, Bangladesh**

**[3] MIS/IT Unit, Computer Village
Dhaka, Bangladesh**

**[4] Department of Computer Science and Engineering, Jahangirnagar University
Dhaka, Bangladesh**

Abstract

Conversion from another language to native language is highly demanding due to increasing the usage of web based application. Firstly, the respective sentence of a native language is converted to Universal Networking Language (UNL) expressions and then UNL expressions can be converted to any native language. UNL system is developed for most of the languages already but a very little effort has been made to convert Bangla language to UNL expressions. In this paper we have described our work to convert Bangla Sentence into UNL Expression. To do this we have described UNL, Bangla grammar, the rules that we have designed for converting Bangla sentence into UNL expression. Finally, we have applied our rules in a Bangla sentence and demonstrated the conversion.

Keywords: *Universal Networking Language, Universal Words, Bangla Roots, Primary Suffix, Verbal Inflexions, Morphological Rules, Semantic Rules.*

1. Introduction

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. Currently, the UNL includes 16 languages [1], 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). 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 [2], where English is the main language of the Internet. But English is not understandable for most of the people. Henceforth, 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 Bangla in this system with less effort.

So far very little effort has been made to convert Bangla language to UNL expressions. We have been working on this topic from the last 3 years. To convert Bangla sentences into UNL expression we needed to go through the Bangla grammar and UNL very rigorously. Simultaneously we have communicated with the UNL center of the UNDL Foundation. They have made two agreements with us: i) Agreement for entering UNLs and ii) UDS agreement, and provided us user name and password to access their resources and utilities. Then we have started converting Bangla sentences into UNL expression successfully. Although we have already worked on Bangla Simple and Compound sentences, but for better understanding of the most of the readers of this paper we have taken a simple affirmative Bangla sentence, “আমি ভাত খাই”, pronounced as ‘Aami vat khai’, meaning of which is ‘I eat rice’ and shown how it can be converted into UNL expression. But it is not limited to this sentence only. This simple example can convert many more sentences of this type.

The organization of this paper is as follow: In Section 2 we describe the Research Methodology, Section 3 has the detail about UNL, Section 4 describes Bangla grammar in detail. In Section 5 and 6 we discuss about the dictionary entries and rules respectively, which we have designed and developed to convert Bangla sentence, Section 7 explains how the rules, developed by us, will be applied to convert, Section 8 shows the result of our work. Finally, Section 9 draws conclusions with some remarks on future works.

2. Research Methodology

For converting Bangla sentence to UNL expressions firstly, we have gone through Universal Networking Language (UNL) [3, 4] where we have learnt about UNL expression, Relations, Attributes, Universal Words, UNL Knowledge Base, Knowledge Representation in UNL, Logical Expression in UNL and UNL systems. All these are key factors for preparing Bangla word dictionary, enconversion and deconversion rules in order to convert a natural language sentence (here Bangla sentence) into UNL expressions. Secondly, we have rigorously gone through the Bangla grammar [5, 6, 7, 8], Verb and roots (Vowel ended and Consonant Ended) [5, 6], Morphological Analysis [7], Primary suffixes [5, 6], Cases and their inflexions [6], construction of Bangla sentence [8] based on semantic structure.

Using above references we extort ideas about Bangla grammar for morphological and semantic analysis in order to prepare Bangla word dictionary (for root, root word suffix etc), morphological rules and enconversion rules in the format of UNL provided by the UNL center of the UNDL Foundation.

3. UNL System – in a nutshell

Although there is an immense proliferation of information through Internet, it is not accessible to vast multitude of people across nations as most of the resources are in English. To overcome this problem, United Nations launched Universal Networking Language project [10] in 1996. The result of the project is universal networking language (UNL), a language neutral specification, and a universal parser specification [11]. 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. The UNL [12] 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 meta-language focuses to express meanings in standardized way. We think that a comprehensive description of UNL specification is necessary though it is available in UNL website. The meaning of 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) [10] 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:

1. Universal Words 2. Attribute Labels 3. Relation Labels 4. UNL Expression 5. Hypergraph 6. Knowledge Base

3.1. Universal Words (UW)

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”.

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” | and” | “aoj” | “obj” | “icl” | ...
```

3.2. Attributes

The attributes [4] 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. Attributed of UWs enrich this description with more information about how the speaker views these states-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”.

3.3. Relational Labels

The relation [4] 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 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)

3.4. 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, 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:[4]
<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.

3.5. 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 of in a scope is 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 hyper-node of UNL is the following:[4]
<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>”.

3.6. Knowledge Base

The UNL Knowledge Base (KB) [10] 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 credibility is.

4. Bangla Grammar

After studying and understanding the UNL concept we have realized that we need to develop the following in order to convert Bangla sentence into UNL expression.

- A Bangla word dictionary
- Morphological and Semantic rules

To develop the above we have rigorously gone through the following part of Bangla grammar so far.

- Root, Vowel and Consonant ended
- Primary Suffixes (Krit Prottoy and Verbal Inflexions)
- Cases and their inflexions
- Verbs

In the next chapter we have discussed about Bangla verb which covers the other topics also.

4.1 Verb

Diversity of verbs in Bangla is very significant [5, 6]. Morphological analysis is applied to verbs to get roots and suffixes. Many different words (nouns, adjectives or verbs) can be derived from a single root. For example, the verb ‘করিতেছি’ (koritechi), is analyzed into root ‘কর’ (kor) and suffix ‘ইতেছি’ (itechi). In Bangla there is a significant number of roots. A good number of suffixes are combined with these roots to form verbs or nouns or adjectives [5, 6, 7]. The suffixes that are combined with roots are divided two groups: [6]

4.1.1 Verbal Inflexions (ক্রিয়া বিভক্তি, pronounce as ‘kria bivokti’): The suffixes that are combined with roots to form verbs are known as Verbal Inflexions (VIs). For instance, the VIs ‘ই’(e), ‘বেন’ (ben), ‘চ্ছে’ (chhe) and ‘চ্ছিলেন’ (chhilen) make verbs যাই(jai), যাবেন (jaben), ‘যাচ্ছে’ (jachhe) and ‘যাচ্ছিলেন’ (jachhilen) respectively combining with verb root ‘যা’(ja) means ‘go’ in English

4.1.2 Primary Suffix (কৃৎ প্রত্যয়, pronounce as ‘krit prottoy’): These types of suffixes are combined with roots to form nouns or adjectives. For example

$$\underbrace{\text{চল্(chol)}}_{\text{Root}} + \underbrace{\text{অন্ত (onto)}}_{\text{Primary Suffix}} = \underbrace{\text{চলন্ত(cholonto)}}_{\text{Adjective}}$$

$$\underbrace{\text{ধর(dhor)}}_{\text{Root}} + \underbrace{\text{আ (aa)}}_{\text{Primary Suffix}} = \underbrace{\text{ধরা(dhora)}}_{\text{Noun}}$$

In this paper, we have focused on first type of suffixes named Verbal Inflexions that are combined with roots to form verbs.

Verb = Root + Verbal Inflexion

4.1.3 Roles of Root and Verbal Inflexion in the formation and meaning of a verb: A root contains the core meaning, which relates with the action or state of the verb, whereas verbal inflexion (VI) defines the formation of the verb and reflects person, tense (in case of finite verb) and other properties. For instance, the root ‘খা’ (kha) means ‘eat’ indicates the action of the verb ‘খাইতেছি’ (khaitechi), to get food through mouth where as the VI ‘ইতেছি’ (itechi) indicates the person (1st person) and tense (present continuous) of that verb. In UNL, person of an inflexion plays role in morphological and syntactic analyses of the verb but has no importance in semantic analysis of the verbs. That means it does not add or change any semantic relation and attribute. On the other hand, tense of an inflexion plays a significant role in semantic field. It adds or changes semantic attributes in the UNL expressions but does not affect on the relation.

4.1.4 Variations of roots: Some roots change their forms when they combine with some specific VIs to make verbs. For example, the root ‘যা’ (ja) means ‘go’ remains unchanged when it combines with VI ‘ইতেছি’ (itechi) to make verb, ‘যাইতেছি’ (jaitechi) but the same root changes it’s form to ‘গি’ (gi), ‘গে’ (ge) and ‘যে’ (je) when it combines with VIs ‘য়াছিলাম’(achilam) to make verb ‘গিয়াছিলাম’ (giachilam), ‘লাম’ (lam) to make verb ‘গেলাম’ (gelam) and ‘তাম’ (tam) to make verb ‘যেতাম’ (jetam) respectively. All the variations of a root appear in the lexicon at different entities, though they all contain same UW but in case of grammatical attributes we use ALT, ALT1 and ALT2 with verb roots ‘গি’ (gi), ‘গে’ (ge) and ‘যে’ (je) as they are the first, second and third alternatives of verb root ‘যা’ (ja) etc. and rest of the attributes will be the same for all variations shown in table 1, table 2 and table 3.

5. Proposed Dictionary Entries

After having a good understanding on UNL and Bangla grammar discussed in section 3 and 4 respectively we have started developing Templates for dictionary entries of Bangla Root and Verbal Inflexion.

Table1: Variations of Vowel Ended Roots and their Verbal Inflexions of VEG1, VEG1.1, VEG2 and VEG3 for First Person

			Vowel Ended Roots							
Persons		**Tenses**	পা (pa)	খা (kha)	গা (ga)	চা (cha)	ছা (ccha)	নি (ni)	দি (ni)	যা (ja)
First Person	**Present**	Present Indefinite	ই	ই	ই	ই	ই	ই	ই	ই
		Present Continuous	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি
		Present Perfect	পা>পে য়েছি	খা>খে য়েছি	গা>গে য়েছি	চা>চে য়েছি	ছা>ছে য়েছি	য়েছি	য়েছি	যা>গি য়েছি
	Past	Past Indefinite	পা>পে লাম	খা>খে লাম	গা>গাই লাম	চা>চাই লাম	ছা>ছাই লাম	লাম	লাম	যা>গে লাম
		Past Habitual	পা>পে তাম	খা>খে তাম	গা>গাই তাম	চা>চাই তাম	ছা>ছাই তাম	তাম	তাম	যা>যে তাম
		Past Continuous	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম
		Past Perfect	পা>পে য়েছিলাম	খা>খে য়েছিলাম	গা>গে য়েছিলাম	চা>চে য়েছিলাম	ছা>ছে য়েছিলাম	য়েছিলাম	য়েছিলাম	যা>গি য়েছিলাম
	Future	Future Indefinite	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব	বো, ব
			VEG 1			VEG1.1		VEG2		VEG3

Table 2: Variations of Vowel Ended Roots and their Verbal Inflexions of VEG4, VEG5 and VEG6 for First Person

			Vowel Ended Roots							
Persons		**Tenses**	ছুঁ (cchu)	থু (thu)	শু (shu)	ধু (dhu)	ন (no)	দু (du)	নু (nu)	রু (ru)
First Person	**Present**	Present Indefinite	ই	ই	ই	ই	ই	ই	ই	ই
		Present Continuous	চ্ছি	চ্ছি	চ্ছি	চ্ছি		চ্ছি	চ্ছি	চ্ছি
		Present Perfect	য়েছি	য়েছি	য়েছি	য়েছি		য়েছি	য়েছি	য়েছি
	Past	Past Indefinite	লাম	লাম	লাম	লাম		লাম	লাম	লাম
		Past Habutual	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম		চ্ছিলাম	চ্ছিলাম	চ্ছিলাম
		Past Continuous	য়েছিলাম	য়েছিলাম	য়েছিলাম	য়েছিলাম		য়েছিলাম	য়েছিলাম	য়েছিলাম
	Future	Past Perfect	ব	ব	ব	বো, ব		বো, ব	বো, ব	বো, ব
			VEG4				VEG5	VEG6		

Table 3: Variations of Vowel Ended Roots and their Verbal Inflexions of VEG7, VEG8 and VEG9 for First Person

			Vowel Ended Roots							
Persons		**Tenses**	হ (ha)	ধা (dha)	না (na)	বা (ba)	ক (ko)	ব (bo)	র (ro)	ল (lo)
First Person	**Present**	Present Indefinite	ই	ই	ই	ই	ই	ই	ই	ই
		Present Continuous	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি	চ্ছি
		Present Perfect	য়েছি	ধা>ধে য়েছি	না>নে য়েছি	বা>বে য়েছি	য়েছি	য়েছি	য়েছি	য়েছি
	Past	Past Indefinite	লাম	ধা>ধাই লাম	না>নাই লাম	বা>বাই লাম	ক>কই লাম	ব>বই লাম	র>রই লাম	ল>লই লাম
		Past Habutual	তাম	ধা>ধাই তাম	না>নাই তাম	বা>বাই তাম	ক>কই তাম	ব>বই তাম	র>রই তাম	ল>লই তাম
		Past Continuous	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম	চ্ছিলাম
		Past Perfect	য়েছিলাম	ধা>ধে য়েছিলাম	না>নে য়েছিলাম	বা>বে য়েছিলাম	য়েছিলাম	য়েছিলাম	য়েছিলাম	য়েছিলাম
	Future	Future Indefinite	ব	ব	ব	ব	বো, ব	বো, ব	বো, ব	বো, ব
			VEG7	VEG8			VEG9			

5.1. Dictionary Entries

We know that Dictionary Entries are made using HW (Head Word), UW (Universal Word) and GA (Grammatical Attributes) [10]. HW indicates the native language word; here, Bangla word in this case, UW is corresponding to the concept from Knowledge Base and GAs are grammatical behaviors of that particular word in that particular language. For example, if we consider Bangla word "পাখি" (pakhi) means 'bird' then its dictionary entry is: [পাখি]{} "bird(icl>vertebrate>thing)" (N, ANI, FLY) <B, 0, 0>

where, “পাখি” is Bangla HW, “bird(icl>vertebrate>thing)” is UW from Knowledge Base and N denotes noun, ANI for animate thing and FLY for flying thing respectively are the grammatical attributes of the above word.

5.1.1 Template for Bangla Verb Root: The template that we are designing here for Bangla roots is depicted bellow:
[HW] {} “UW (icl/iof…>concept1>concept2, REL1>.., REL2>…,” (ROOT, VEND/CEND [,ALT/ALT1/ALT2..] VEG_n/CEG_n, #REL1, #REL2,.. <FLG, FRE, PRI>
where,
HW← Head Word (Bangla Word; in this case it is Bangla root);
UW← Universal Word (English word from knowledge base);
icl/iof/… means *inclusion/instance of* …to represent the concept of universal word
REL1/REL2.., indicates the related relations regarding the corresponding word.
ROOT ← It is an attribute for Bangla roots. This attribute is immutable for all Bangla roots.
CEND and VEND are the attributes for consonant ended and vowel ended roots respectively as every root is ended either with consonant or vowel;
VEG_n ← means attribute for the group number of vowel ended roots
CEG_n ← means attribute for the group number of consonant ended roots
ALT, ALT1, ALT2 etc. are the attributes for the first, second and third alternatives of the vowel or consonant ended roots respectively. If the root is default, then no alternative is used.
#REF1, #REF2 etc. are the possible corresponding relations regarding the root word
Here, attributes say, ROOT, CEND/VEND are fixed for all Bangla roots whereas ALT or ALT1 etc. does not necessary for all roots, they are used only for alternative roots.
In the following examples we are constructing the dictionary entries for some sample *Bangla roots* using our designed template:
[যা]{}“go(icl>move>do, plf>place, plt>place, agt>thing)” (ROOT, VEND, VEG3, #PLF, #PLT, #AGT)<B, 0, 0>
[গি]{}“go(icl>move>do, plf>place, plt>place, agt>thing)” (ROOT, VEND, ALT, VEG3, #PLF, #PLT, #AGT) <B,0,0>
[খা]{}“eat(icl>consume>do,agt>living_thing, ins>thing, obj>concrete_thing, plf>thing, tim>abstract_thing)” (ROOT, VEND, VEG1, #PLF, #PLT, #AGT)<B, 0, 0>
In the examples above, for first two entries the relation *plf* (place from) indicates from where agent go/goes, plt (place to) means to where go/goes, agt(agent) for who go/goes and attribute ALT indicates that root “গি “(gi) is the first alternative of root “যা “(ja) discussed in table 1. Attributes #PLF, #PLT and #AGT indicate that relations *plf, plt* and *agt* can be made with roots “গি “(gi) and “যা “(ja). Similarly, other entries can be developed according to the format above.

5.1.2 Template for Verbal Inflexion: In the previous section we designed a template for *Bangla verb roots*. However the template for Verbal Inflexion is very similar to that of Bangla verb roots. One thing is verbal inflexions do not have any universal word and they have only grammatical attributes and differ each other with attributes they use.
[HW]{} “” (VI, V, Aperson [,ALT/ALT1,ALT2...], GEN/RES/NEG, Atense, SD/CH, VEG_n/CEG_n/^VEG_n/ ^CEG_n.) <FLG, FRE, PRI>
HW← Head Word (Verbal Inflexion of Bangla Verb Root); UW← Universal Word (In case of Verbal Inflexion, UW is null); VI← is an attribute of Verbal Inflexion, V← Verb, since Verbal Inflexions form verb when it is added with Bangla verb root as Suffixes so we keep the ‘V’ as an attribute.
Aperson← Attribute person, this is an important attribute because verb varies according to Bangla Person.
ALT/ALT1/ALT2 ← Attributes for alternative roots. These attributes are used as attributes of verbal inflexions when they are combined with respective verb roots.
GEN/RES/NEG← Attributes for verbal inflexions when they are combined with verb roots to form general (GEN), respective (RES) and neglect (NEG) verbs in respect to person. They are used as attributes with the VIs that are combined with verb roots to form verb for second and third persons
Atense ← Attribute Tense, This is also an important attribute because verb varies according to Bangla Tense.
SD/CH← Attribute for types of languages. SD for ‘shadhu’, which is literature language and CH for ‘cholti’, which is conversation language. They are used as attributes with the VIs as they form SD or CH types of verbs.
VEG_n/CEG_n/^VEG_n/^CEG_n← Attributes indicate for vowel ended group number or consonant ended group number of not for vowel or consonant ended group. They are used as attributes of VIs as they are combined with respective groups or not. Like *verb roots* some attributes like VI, V are fixed for all *Verbal Inflexioni* but Aperson can be either attributes ‘P1’(for first person), ‘P2’ (for second person) or ‘P3’ (for third person) and Atense can be any tense such as attributes ‘PRS’ (for Present Indefinite), ‘PRG’ (progress for Present continuous) CMPL (complete for perfect tense) etc. If the tense is past continuous then two attributes are used consecutively such as attribute ‘PST’ (for past) and ‘PRG’ (for continuous) and for future tense FUT is used.

Some examples of dictionary entries of *Verbal Inflexions* according to the proposed template are given below:

[য়েছিলাম] “ ”{}(VI,P1,PST, PER,ALT,CH,VEG1,VEG9)
[চ্ছিলাম] “ ”{}(VI,P1,PST,PRG,CH)
[বি] “ ”{}(VI,P2,NEG,FUT,CH)
[চ্ছেন] “ ”{}(VI,P2,RES,PRT, PRG,CH)

Here, VI, ‘য়েছিলাম’ can be combined with first alternative roots(as ALT is used to define attribute) with verb roots of *vowel ended group 1* or *vowel ended group 1.1* for past perfect tense (attribute PST for past and CMPL for perfect) to create the verbs of conversation language (CH attribute for conversation language) for first person (attribute is P1). Similarly, attributes for other dictionary entries are defined.

6. Proposed Morphological and Semantic Rules

If we consider a sentence say, “আমি ভাত খাই” (pronounce as *aami vaat khai*.) meaning, “I eat rice.” for conversion process. Assuming that all the words and morphemes of the given sentence are in the dictionary as follows:

[আমি]{}"i(icl>person)"(PRON,HPRON,1P,SG,SUBJ)<B,1,1>
[ভাত]{}"rice(icl>food)"(N)<B,0,0>
[খা]{}"eat(icl>consume>do,agt>living_thing,obj>concrete_thing)"(ROOT,VEND,#AGT,#OBJ,VEG1)<B,0,2>
[ই]{}""(VI,VEND,1P,PRS)<B,0,0>

where, attributes PRON denotes pronoun, HPRON for human pronoun, 1P for first person, SG for singular number, SUB for subject, N for noun, ROOT for verb root, VEND for vowel ended root, #AGT for agent (means agent relation can be made with root ‘খা’ (kha) , #OBJ for object (like as agent relation), VEG1 means the root is fall in the vowel ended group 1, VI is the attribute for the verbal inflexion that can combine with root to make verb while PRS means present tense. 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]. 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 figure 1.

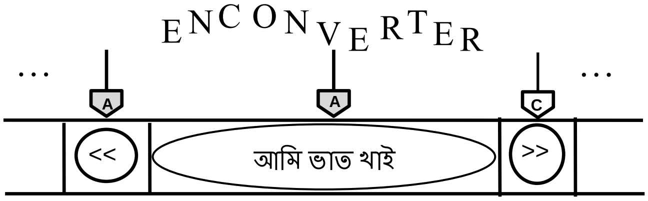


Figure 1: Initial state of the Analysis Windows and the node list

After insertion of the input file with our given sentence the following rules, which we have developed, will be applied step by step to complete the conversion process of the sentence to UNL expressions.

```
Rule 1:
R{SHEAD:::}{PRON,SUBJ:::}P10;
Rule 2:
DR{SUBJ,^blk:blk::}{BLK:::}P10;
Rule 3:
R{PRON,SUBJ:::}{N:::}P10;
Rule 4:
DR{N,^blk:blk::}{BLK:::}P10;
Rule 5:
R{N:::}{ROOT,^VERB:::}P10;
Rule 6:
+{ROOT,VEND,^ALT,^VERB:+VERB,-
ROOT,+@::}{KBIV,VEND:::}P10;
Rule 7:
:{:::}{VERB,KBIV:-KBIV,-VEND,-
CEND::}P10;
Rule 8:
>{N::obj:}{VERB,#OBJ:::}P10;
Rule 9:
>{HPRON,SUBJ::agt:}{VERB,#AGT:::}P10;
Rule 10:
R{SHEAD:::}{VERB,^&@entry:+&@entry::}P1
0;
Rule 11:
R{VERB:::}{STAIL:::}P10;
```

We have developed the above rules by following the UNL specifications described in the next two sub-sections.

6.1. Morphological rules

Morphological analysis is found to be centered on analysis and generation of word forms. It deals with the internal structure of words and how words can be formed [13]. It is applied to identify the actual meaning of the words [14] identifying the Prefixes and Suffixes. Morphological study comes here to help with rules for analyzing the structure and formation of the Bangla verbs. Rules for morphological analysis of verbs are used to combine the roots with their corresponding inflexions to complete the meaning of the verbs. An Enconversion Rule (morphological/semantic) is composed of Conditions for the nodes placed on the Analysis Windows and Condition Windows, and Actions and/or Operations for the nodes placed on the Analysis Windows. Such enconversion rules describe the kind of actions and/or operations that should be carried out for all phenomena of a language, and under what conditions. EnConverter will find the most suitable rule every time, and create a partial syntactic tree and/or UNL expression. A set of UNL expressions of a sentence will finally be completed after having applied a set of all the necessary rules. Out of 15 different types of rules [main] 2 rules are used for morphological analysis. One is left composition rule (<) and another is right composition rule (>)

6.1.1. Left Composition (+): The basic type of this group is “+”, this type of rule is used basically to create a syntactic tree with the two nodes on the Analysis Windows [4.] By applying this type of rule, the two headwords of the left and right nodes are combined into a **composite node**, the original left and right nodes are replaced with the composite node in the Node-list, and the sub-syntactic tree and attributes of the **left node** are inherited. If the operator "@" appears in the <ACTION> field of the rule for the left node, the attributes of the right node are also inherited.

Application of this rule implies the deletion of the original two nodes from the Node-list and the insertion of the new **composite node** into the Node-list. The position of the new **composite node** is on the right Analysis Window.

6.1.2. Right Composition (-): The basic type of this group is “-“, this type of rule is used basically to create a syntactic tree with the two nodes on the Analysis Windows. By applying this type of rule, the two headwords of the left and right nodes are combined into a **composite node**, the original left and right nodes are replaced with the composite node in the Node-list, and the sub-syntactic tree and attributes of the **right node** are inherited. If the operator "@" appears in the <ACTION> field of the rule for the right node, the attributes of the left node are also inherited [4].

Application of this rule implies the deletion of the original two nodes from the Node-list and the insertion of the new **composite node** into the Node-list. The position of the new **composite node** is on the right Analysis Window.

6.2. Semantic Rules

Semantic rules are used for creating semantic relations. Semantic relation describes the relations between the words in the sentence. They are used to form semantic network of the UNL [4]. Two types of rules are used for this relation.

6.2.1. Left Modification Rule (<): This type of rule creates a Syntactic Tree and a Semantic Relation for the two nodes on the Analysis **Windows.** The right node becomes the modifier of the left node. This rule deletes the right node from the Node-list, while the left node becomes the head of this partial syntactic tree and remains in the Node-list. It creates a semantic relation, according to the designation of the relation in the <RELATION> field, with the node where the relation is described in the <RELATION> field as the **to-node** and the partner node as the **from-node** of the semantic relation. It adds the semantic relation to the list of semantic relations of the left node, and outputs it in the result of UNL expressions when the enconversion is completed.

If the operator "@" appears in the <ACTION> field of the rule for the left node, the attributes of the right node are also inherited. After applying this type of rule, the left node moves to the right Analysis Window.

6.2.2. Right Modification Rule (>): This type of rule creates a Syntactic Tree and a Semantic Relation for the two nodes on the Analysis Windows. The left node becomes the modifier of the right node. This rule deletes the left node from the Node-list, while the right node becomes the head of this partial syntactic tree and remains in the Node-list. It creates a semantic relation, according to the designation of the relation in the <RELATION> field; with the node where the relation is described in the <RELATION> field as the **to-node** and the partner node as the **from-node** of the semantic relation. It adds the semantic relation to the list of semantic relations of the right node, and outputs it in the result of UNL expressions when the enconversion is completed.

If the operator "@" appears in the <ACTION> field of the rule for the right node, the attributes of the left node are also inherited.

After applying this type of rule, the right node remains in the right Analysis Window.

7. Conversion of Bangla sentence into UNL applying the proposed rules

Let us consider the sentence, “আমি ভাত খাই” pronounce as “*aami vat kha”i* meaning, “I eat rice.” for conversion process and apply the rules given in section 6.

Rule 1 is *Right Shift rule* that describes that when sentence head is in the Left Analysis Window (LAW) and word ‘আমি’ (aami) is in the Right Analysis Window (RAW) then AWs will be shifted to right after rule application. In this situation, the Enco will retrieve the word, ‘আমি’ (aami)
from the Word Dictionary file and remains in the LAW and ‘ভাত খাই’ (vat khai) will be in the RAW.
Rule 2 is *Right Node Deletion rule*, it deletes the right node which is blank space between ‘আমি’ (aami) and ‘খাই’ (vat) and only the noun ‘ভাত’ (vat) will be placed in the RAW while the verb ‘খাই’ (khai) will be placed in the Right Condition Window (RCW).
Again *right shift rule* (rule 3) is applied to shift the windows to right and *Right Node Deletion rule* (rule 4) is applied to delete the space between ‘ভাত’ (vat) and ‘খাই’ (kaai) so that the word ‘ভাত’ (vat) is retrieved from the Word Dictionary and remains in the LAW and the verb ‘খাই’ (kaai) is divided into root ‘খা’ (kha) which remains

in the RAW and verbal inflexion 'ই' (i) remains in the RCW.

Now, again *right shift rule* (rule 5) is applied to place 'খা' (kha) in the LAW and 'ই' (i) in the RAW to perform the morphological analysis.

In this situation, Enco retrieves the dictionary entries of 'খা' (kha) and 'ই' (i) from the word dictionary (input file) and will apply the analysis rule (rule 6) that is if there is a vowel ended root (in our example, 'খা') is in the LAW and verbal inflexion (in our example, 'ই') is in the RAW then after applying the rule the two headwords of the left and right nodes are combined into a composite node to complete the morphological analysis of the verb 'খাই' (khai) [section 6.1.1]

Rule 7 is *attribute changing rule* used to rewrites the attributes by deleting attributes VI, VEND, and CEND the for verb 'খাই' (khai) that remains in the RAW.

Now Enco starts semantic analysis between the words of our sentence by applying semantic rules 8 and 9.

Rule 8 is *Right Modification Rule* (>), indicates that if noun 'ভাত' is in the LAW and verb 'খাই' is in the RAW then an object relation (obj) is made between them where 'খাই' remains in the RAW and 'ভাত' is deleted. [section 6.2.2]

Now the word 'আমি' (aami) comes to the LAW and an agent relation (agt) is made between 'আমি' (aami) and 'খাই' (khai) by applying rule 9, so that 'আমি' (aami) is deleted from the node-list and the verb 'খাই' (khai) remains in the RAW which is the main predicate of the sentence.

After that right shift rule 10 is applied to shift the windows to right and *&@entry* attribute is added to the verb as verb 'খাই' (khai) is the main word of the sentence.

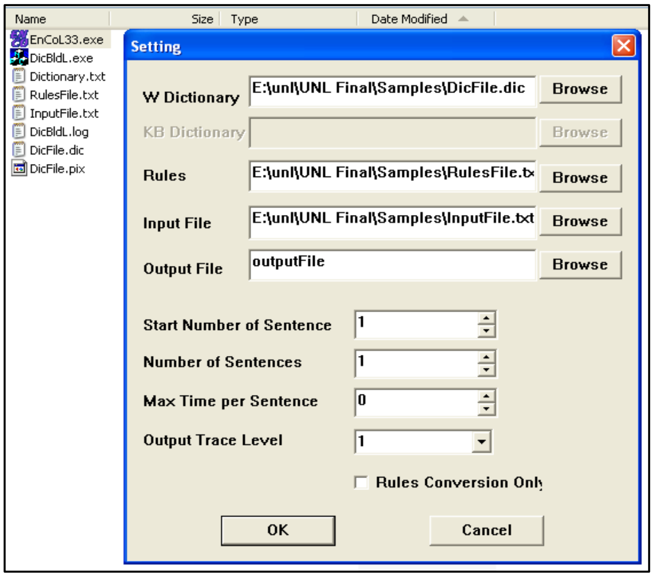
Finally, rule 11 is applied to place the sentence tail (STAIL) on the LAW to complete the conversion process [4]

8. Result

To convert the Bangla sentence "আমি ভাত খাই" we have used the following files.

1. Input file
2. Output file
3. Rules File
4. Dictionary

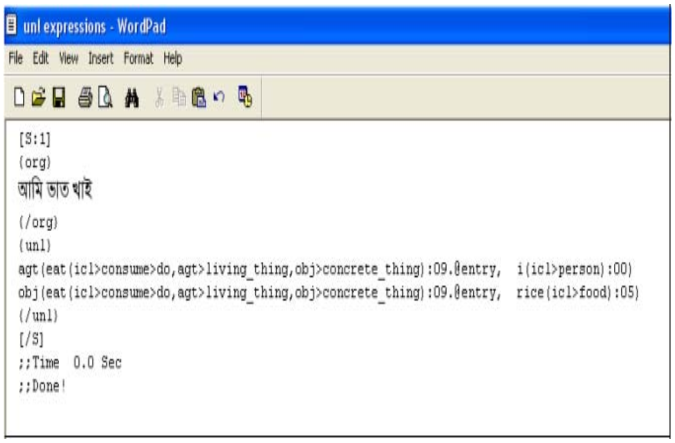
We have used an Encoder (EnCoL33.exe) and a dictionary builder file (DicBldL.exe) provided by the UNDL Foundation of UNL center, which we have downloaded from [9]. Screen 1 shows the way of selecting files.



W Dictionary: E:\unl\UNL Final\Samples\DicFile.dic
KB Dictionary:
Rules: E:\unl\UNL Final\Samples\RulesFile.tx
Input File: E:\unl\UNL Final\Samples\InputFile.txt
Output File: outputFile
Start Number of Sentence: 1
Number of Sentences: 1
Max Time per Sentence: 0
Output Trace Level: 1
□ Rules Conversion Onl

Screen 1: Selecting Files in enco.exe

Screen 2 shows the output file, generated by the encoder, which contains the UNL expression of the Bangla sentence "আমি ভাত খাই"



```
[S:1]
(org)
আমি ভাত খাই
(/org)
(unl)
agt(eat(icl>consume>do,agt>living_thing,obj>concrete_thing):09.@entry,  i(icl>person):00)
obj(eat(icl>consume>do,agt>living_thing,obj>concrete_thing):09.@entry,  rice(icl>food):05)
(/unl)
[/S]
;;Time  0.0 Sec
;;Done!
```

Screen 2: Output file with UNL expressions

9. Conclusion

In this paper, we have discussed and demonstrated how a simple assertive sentence can be converted into UNL expression. But we have already worked on Simple and Compound sentences but for better understanding of the most of the readers of this paper we have only demonstrated a simple assertive sentence. We have mentioned this in Section 1 (Introduction). In section 2 we have outlined the methodology by which we have carried out the research. We have discussed about UNL and Bangla Grammar in section 3 and 4 respectively. In section 5, 6 and 7 we have described our work elaborately. Section 8 shows the practical implementation of our research and it is found that our rules, dictionary, etc. worked perfectly. We understand that it just a start of a long journey and hope that we would be able to reach the destination.

References

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

[2] 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 Industrial Studies Center, Royal Scientific Society, March 2002

[3] http://www.undl.org last accessed on January 30, 2011

[4] H. Uchida , M. Zhu , T.G. D. Senta "Universal Networking Language", 2005/6-UNDL Foundation, International Environment House.

[5] D.M. Shahidullah, "Bangla Baykaron", Ahmed Mahmudul Haque of Mowla Brothers prokashani, Dhaka-2003.

[6] D. C. Shuniti Kumar, "Bhasha-Prakash Bangala Vyakaran", Rupa and Company Prokashoni, Calcutta, July 1999, pp.170-175

[7] D. S. Rameswar, "Shadharan Vasha Biggan and Bangla Vasha", Pustok Biponi Prokashoni, November 1996, pp.358-377

[8] H. Azad , "Bakkotottyo", Second edition, 1994, Dhaka

[9] http://www.undl.org/index.php?option=com_content&view =article&id=53&Itemid=99&lang=en# - To download Enconverter

[10] H. Uchida , M. Zhu, " The Universal Networking Language (UNL) Specification Version 3.0 1998, Technical Report, UNU, Tokyo, 1998", 2005/6-UNDL Foundation, International Environment House

[11] M.N.Y. Ali, J.K. Das, S.M. Abdullah Al Mamun, M. 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.

[12] J. Parikh, J. Khot, S. Dave, P. Bhattacharyya, "Predicate Preserving Parsing", Department of Computer Science and Engineering , Indian Institute of Technology, Bombay

[13] M.N.Y. Ali, J.K. Das, S.M. Mamun, A. M. Nurannabi, "Morpholoical Analysis of Bangla worfs for Universal Networking Language", International Conference on Digital Information Management, icdim, 2008, London, England, pp. 532-537

[14] M.N.Y. Ali, S.A.Noor, M.H.Z. Sarker, J.K. Das, "Development of Analysis Rules for Bangla Root and Primary Suffix for Universal Networking Language", International Conference on Asian Language Processing, IALP 2010, Harbin China.

[15] Enconverter Specification Version 3.3, UNU Foundation, Tokyo 150-8304, Japan 2000

First Author: Md. Nawab Yousuf Ali is a full time faculty member of East West University Bangladesh in the Department of Computer Science and Engineering. He obtained his M.Sc. in Computer Engineering from Lvov Polytechnic Institute, Lvov, Ukraine, USSR in 1992. He is the author of one journal and eleven international conference papers in home and abroad. His research interest includes Natural Language Processing especially Universal Networking Language.

Second Author: Mohammad Zakir Hossain Sarker is the MIS Specialist of a USAID funded project run by an American Company named JSI and also was a former faculty member of East West University Bangladesh in the Department of Computer Science and Engineering. He obtained his M.Sc. in Computer Engineering from Dhaka University, Dhaka, Bangladesh in 1998. He is the author of 7 Journal and more than 15 International Conference papers in home and abroad. His research interest includes Natural Language Processing especially Universal Networking Language.

Third Author: Mr. Ghulam Farooque Ahmed is the Chief Executive Director of the Computer Village. He has invented Bangla Font software Lekoni which is very popular in Bangladesh. He has been working on Bangla Language Processing for more than 20 years.

Fourth Author: 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. He is the author of 13 Journal and more than 16 International Conference papers in home and abroad. His research interests include Network Protocols, Universal Networking Language, Distributed Systems and so on.