International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 3 (2011) pp. 746 -754

Structure of Dictionary Entries of Bangla Morphemes for Universal Networking Language (UNL)

Muhammad Firoz Mridha[1], Kamruddin Md. Nur[2], Manoj Banik[3] and Mohammad Nurul Huda[4]

[1]Dept. of Computer Science, Stamford University Bangladesh
Dhaka, Bangladesh
mdfirozm@yahoo.com

[2] Dept. of Computer Science and Engineering, United International University
Dhaka, Bangladesh
kamruddin.nur@gmail.com

[3]Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology, Dhaka, Bangladesh
mbanik99@yahoo.com

[4] Dept. of Computer Science and Engineering, United International University
Dhaka, Bangladesh
mnh@cse.uiu.ac.bd

Abstract: **This paper describes a structure of dictionary entries of Bangla (widely used as Bengali) Morphemes for Universal Networking Language (UNL). The UNL is an artificial language developed for conveying linguistic expressions in order to represent websites information into a standard form. In order to integrate Bangla into this platform it is necessary to develop both a dictionary and a grammar, where Dictionary plays a crucial role in any machine translation (MT) system. Analysis of Grammatical Attributes of Bangla words such as Bangla Roots, Krit Prottoy (primary suffix) and Kria Bivokti (verbal suffix) are also focused in this study. The goal is to make possible Bangla sentence encoversion to UNL and vice-versa. The theoretical analysis of our model proves that the proposed work is successfully able to prepare universal words for Bangla roots, Krit Prottoy and Kria Bivokti along with their grammatical attributes for UNL.**

Keywords**: Universal Networking Language, Universal Words, Knowledge-base, Structure of dictionary entry, Bangla Word dictionary, Morphological Analysis**.

.

I. Introduction

Knowledge is for all, but to be indeed for all, it should be accessible for all those who seek it regardless of their mother tongue. Today, Information Technology (IT) has converted the world into a global village and libraries, as

part of this age, should make use of these technological advancements in achieving the Universality goal and quenching the generation's thirst for knowledge. At present, Machine Translation (MT) techniques have been applied to web environments. The growing amount of available multilingual information on the WWW, as well as the increase of internet users has led to a justifiable interest on this area. The main goal of the Universal Networking Language **(**UNL) system is to provide internet users access to multilingual websites using a common representation [1, 2]. This will allow users to visualize websites in their own language, whether it has been built under a different language or not. This has a growing relevance since the usage of the WWW is generalized across cultural and linguistic barriers. Many languages such as Arabic, French, Russian, Spanish, Italian, English, Chinese or Brazilian Portuguese have already been included in the UNL platform. Our aim is to introduce Bangla into this system. This paper focuses on the structure of dictionary entries for UNL.

Even now in networks, there are a various types of information, which is written in a number of languages but the language barrier is hindering the access of this information. Information written in the UNL can be received by anyone in their mother tongue via a network. Once information written in one language is "enconverted" into UNL it will be able to be shared by anyone in the world [3]. The use of UNL that needs no language analysis will be a great benefit in terms of economy in that it saves time and money. These points alone give the UNL the edge over other ways of translation.
However, the UNL, which is not necessarily accessible to all people, is made up of logical description format that

represents the meaning of sentences and can be understood only by trained specialists. To solve this, an editing system is provided so that documents can be entered in UNL. A number of editing systems could be possible that would meet the needs of particular uses. A UNL writer that does not translate one language into another is different from the typical translator of the past and converts information from any language into UNL. Alternatively, a UNL writer constructs a bridge among all the languages. Currently, the UNL deals with 16 languages including the six official languages of the United Nations (Arabic, Chinese, English, French, Russian and Spanish) in addition to ten other widely spoken languages (German, Hindi, Italian, Indonesian, Japanese, Latvian, Mongol, Portuguese, Swahili and Thai). In its second phase (1999 – 2005) the project will seek to further extend UNL access.

II. UNL Structure

UNL is an artificial language that allows the processing of information across linguistic barriers [10] and that 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, where nodes (that represent concepts) and arcs (that represent relations between concepts) compose the network. UNL contains three main elements:

(i) Universal Words (UW): Nodes that represent word meaning.
(ii) Relation Labels (RL): Tags that represent the relationship between UWs i.e. between two nodes. Tags are the arcs of UNL hypergraph.
(iii) Attribute Labels (AL): Additional information about the UWs.

These elements are combined in order to establish a hierarchical Knowledge-Base (UNLKB) [10] that defines unambiguously the semantics of UWs. The UNL Development Set provides some tools, EnConverter and DeConverters that enable the semi-automatic conversion of natural language into UNL and vice-versa. The main role of EnConverter [11], which implements a language independent parser that provides a framework for morphological, syntactic and semantic analysis synchronously, is to translate natural language sentences into UNL expressions; this allows morphological and syntactical ambiguities resolution. On the other hand, the DeConverter [3, 12] is a language independent generator that converts UNL expressions to natural language sentences.

A. Universal Words (UW)

UWs are words that constitute the vocabulary of UNL, where 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:

i) Labels are defined to express unit concepts and are called "UWs".
ii) A compound structure of a set of binary relations is grouped together and is called "Compound UWs".

Format: <UW>:= <headword> [<constraint list>]

Example**:** *Curtail (icl>reduce>do, agt>thing, obj>thing);*

Here, curtail is the headword and rest is the constraint list. The keywords *icl*, *agt* and *obj* are taken from *UNL relations.* In the example, *agt* > *thing* implies an agent that is a "thing" or its subclass and similarly, obj stands for object. *Icl* > *reduce* > *do* implies this UW is a subclass of UW "reduce", which is consequently the subclass of "do" that indicates a verb.

B. Relational Labels (RLs)

The relation [1] between UWs is binary that have different labels according to the different roles they play. A RL 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.

C. Attributes Labels (AL)

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.

III. Bangla UNL Dictionary

The Universal Dictionary of concepts strives to include and integrate conceptual lexicons of all natural languages. The UNL dictionary entries require a specific encoding. Each entry of the Word Dictionary is composed of three kinds of elements: the *Headword (HW)*, the *Universal Word (UW)* and the *Grammatical Attributes (GAs)*. A headword is a notation 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. GAs is 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 [7, 8, 19].

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

Here,

HW ← Head Word (Bangla word)
ID ← Identification of Head Word (omitable)
UW←Universal Word (English word from knowledge base)
ATTRIBUTE ← Attribute of the HW
FLG ← Language Flag (we use B for Bangla)
FRE← Frequency of Head Word
PRI← Priority of Head Word

Format of an element of Bangla-UNL Dictionary would be:

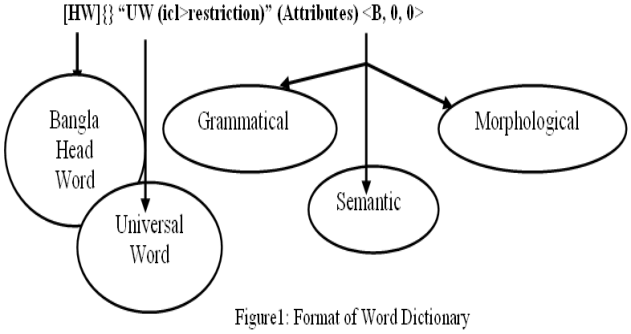

Figure1: Format of Word Dictionary

Some example entries of dictionary for Bangla are given below:

[আমি]{} "I" (1SG, HPRON, PRON, SUBJ) <B,0,0>;
[আমাকে]{} "I" (1SG, HPRON, OBJ, PRON) <B,0,0>;
[আমার]{} "I" (1SG, HPRON, POSS, PRON) <B,0,0>;

The UNL system basically consists of a network and a conversion program between UNL and involved languages including native. All these activities are carried out via the network.

IV. Morphological analysis of Bangla words compatible with UNL structure

Morphological analysis is found to be concentrated on analysis and generation of word forms. It deals with the internal structure of words and how words can be formed [8, 17]. It is applied to identify the actual meaning of the words [6, 7] by identifying the Bangla Prefixes (DcmM©) and Suffixes (cÖZ¨q)

A. Prefixes (DcmM©)

Prefixes are the words that are used before words to express various meanings of the same words. There are around fifty (50) prefixes used in Bangla sentences. In Shangskrit Bangla, we use twenty (20) prefixes[2] say, cÖ(cÖKl©), civ(ˆecixZ¨), Ac(ˆecixZ¨), etc. ; in Bangla we use thirteen (13) prefixes[2] such as ‡e(‰ecixZ¨), Mi(‰ecixZ¨), Ab(Afve) etc.; five(5) foreign prefixes[5] are Mi (bv), `i (wbæm&n), e`& (Lvivc) etc.; four English prefixes[3] are mve(Aaxb A‡_©), ‡nW(cÖavb), dzj(cyiv), nvd(Aa©) etc. and other prefixes[3] say cyit(mg‡þ/mvg‡b), cÖv`yt(`„wó‡MvPi), ewnt(evwn‡i) etc. These prefixes are used before words to make thousands of meaningful Bangla Words.
In our work, we will make separately Word Dictionary entries for all of these prefixes and words, so that they can combinely make meaningful words by applying rules. For example, if we consider prefix "cÖwZ"[10] (means like/similar/every/opposite/against etc.) we can make "cÖwZw`b", "cÖwZkã", "cÖwZcþ" etc. Now, we can make the word "cÖwZ" for dictionary entry. But the word "cÖwZ" has two or more meanings so that we have to represent two or more dictionary entries for the word as follows [16].

[cÖwZ]{} "every (icl>thing)" (ABSTRACT THING) <B,0,0>
[cÖwZ]{} "opposite (icl>thing)" (ABSTRACT THING) <B,0,0>
[cÖwZ]{} "against (icl>thing)" (ABSTRACT THING) <B,0,0>

Now, if we want to represent the concepts of the words say cÖwZw`b, cÖwZkã, cÖwZcþ etc., we need not represent the whole words. We have to represent only the words "w`b", "kã" and "cþ" in the dictionary entry as per the following format.

[w`b] {} "day(icl>period>time)"(N,ABSTRACT THING, LIGHT)<B,0,0>
[kã] {} "sound(icl>occure>thing)"(N,ABSTRACT THING)<B,0,0>
[cþ] {} "group(icl>person)"(N,CONCRETE)<B,0,0>

If we have the concepts of the prefix "cÖwZ" and the root words "w`b", "kã", and "cþ" with their grammatical attributes in the Word Dictionary as above format, the conversion rule will make the concepts of the whole words "cÖwZw`b", "cÖwZkã" and "cÖwZcþ", combining the first, second and third concepts of "cÖwZ" respectively. By applying the same rule the EnCo can make all other words used with "cÖwZ", which have the concepts of the words in the word dictionary.
Similarly, if we consider Bangla prefix "ivg" we can make "ivgQvMj", "ivg`v" etc. We can separately represent the concepts of "ivg" , "QvMj" and "`v " in the dictionary entry according to the following format.

[ivg] {}"big(icl>large>thing)"(ADJ, ABSTRACT THING) <B,0,0>
[QvMj] {}"goat(icl>animal>animate thing)"(N, CONCRETE , ANI) <B,0,0>
`v{}"knife(icl>edge_tool>thing)"(N,C) <B,0,0>

Therefore, if we have the concepts of all the words in the dictionary we can make the dictionary entry of all the complete words combined with "ivg". Here, we also can use "QvMj" and "`v" as separate words for other dictionary entries. Finally, we can infer that conversion rules can be applied to prepare thousands of complete Bangla Words combining with prefixes (mentioned above) and words to represent their full concepts in the Bangla-UNL Dictionary.

B.Suffixes (cÖZ¨q)

Morphological analysis describes that every word is derived from a root word. A root word may have different transformations. This happens because of adding different morphemes with it as suffixes. So, the meaning of the word varies for its different transformations. There are four different types of morphologies [9].

C. Noun Morphology: Bangla Nouns have very strong and structural inflectional morphology base on case. Case of noun may be nominative ("‡Q‡j", boy), accusative ("‡Q‡j-‡K", to the boy) and genetive("‡Q‡j-i", of the boy) and so on. Gender and number are also important for identifying proper categories of nouns. Number may be singular ("‡Q‡jÓ, boy or "†Q‡jwU", the boy, "eB", book, "eBwU", the book) plural ("†Q‡jiv", boys "†Q‡j¸wj", the boys "eB¸‡jv", the books etc.). So, from the word "‡Q‡j" we get "‡Q‡ji", "‡Q‡j‡K", "†Q‡jivÓ "†Q‡jwU", "†Q‡j¸wj" etc. and from the word "eB" we get "eBwU", "eB¸‡jv" etc. Some dictionary entries may look like. [‡Q‡j]{} "boy (icl>person)" (N, HN, C, ANI)<B,0,0>
Here, "boy (icl>person)" is the UW for "‡Q‡j" but "i" , "†K" etc. have no UWs. Therefore, they should be represented in the dictionary only with grammatical attributes as follows.

[i] {} "" (3P, SUF, N)<B,0,0>
[‡K] {}""(3P, SUF, N, HUMN, SG)<B,0,0>
[iv] {}"" (3P, PL, SUF, N, HUMN)<B,0,0>
[wU] {}"" (N, SG, SUF,3P) <B,0,0>
[¸wj] {}"" (N, PL, SUF,3P) (<B,0,0>
[¸‡jv] {}""(N, SG, SUF,3P) <B,0,0>

We use 3P, SUF and N as grammatical attributes with "iv", because "iv" is used with third person say "†Q‡jiv", N for noun and SUF as "iv" is a suffix. We have to put meticulous attention while defining the grammatical attributes. Because we use HUMN for human noun as "‡K", "iv" are used with human being only, say ‡Q‡j‡K, Zvnv‡K, ‡Q‡jiv, Zvnviv but not Mi"‡K , Mi"iv etc. But we can not use HUMN with "i", "wU", "¸wj" and "¸‡jv" because they are used with both human and non human, say cvwLi, †Q‡jwU, Mi"¸‡jv, etc.

D. Adjective Morphology: As Adjective we can consider Bangla words "mvnm", "my›`iÓ and "fv‡jv" meaning "bravery", "beautiful" and "good" in English, respectively. From the first word, we get mvnmx (mvnm+B), mvn‡mi (mvnm+Gi). Again, from the second and third words we get my›`ix, fv‡jvi, fv‡jvUv etc. We have to have the dictionary entries for mvnm, my›`i, fv‡jv, B, Gi, i, Uv to make the meaningful words mvnmx, mvn‡mi, my›`ix, fv‡jvUv etc. by combining the morphemes with the root words using analysis rules.
For example, if we consider a Bangla sentence, "mvnmxiv mvn‡mi mv‡_ Ab¨v‡qi c^{a}wZev` K‡i|Ó We can represent the sentence as "mvnm-B-iv mvnm-Gi mv‡_ Ab¨vq-Gi c$^{a}$wZ-ev` Ki-G|". Here, B, iv, Gi, ev`, G etc. are morphemes. So, we can see that a number of morphemes are added with the root words to make the full meaning of the new words as well as sentence.

E. Pronoun Morphology: Here, we can consider the word root "Zvnv"(he/she). From this we get Zvnv-iv, Zvnv-‡K, Zvnv-‡`i, Zvnv-‡`i-‡K, Zvnv-w`M‡K etc. So, we have to consider these morphemes iv, ‡K, ‡`i, w`M‡K for dictionary entries to form words with "Zvnv" as above.
***F. Verb Morphology*:** Diversity of verb morphology in Bangla is very significant. If we consider 'hv' (means go) as a root, we can represent this root in the dictionary as [hv]{} "go (icl>move>do)" (V, @present) <B,0,0>.

Some transformations based on the persons and tenses are.

- For first person:

[B]{} "" (SUF, PRESENT, 1P)<B,0,0> for hvB (hv+B))
[B‡ZwQ] {}"" (SUF, PRESENT, 1P) <B,0,0> (for hvB†ZwQ (hv+B‡ZwQ)) etc.

- For second person:

[B‡ZwQ‡jb] {}""(SUF, PAST)<B,0,0> (for hvB‡ZwQ‡jb (hv+B‡ZwQ‡jb))

- For third person:

[†e] {}"" (SUF, FUTURE)<B,0,0> (for hv‡e (hv+‡e))

For resolving the ambiguities of the words wM‡qwQ, wM‡qwQjvg, wM‡q‡Qb, wM‡qwQ‡jb, hvB‡Z _vK‡eb, wMqv‡Q etc. we have to define them as full words for dictionary entries. For instance, [wM‡qwQjvg]{} "go(icl>move>do)"(V, PAST, INDEF, 1P). Using the same procedure we can make dictionary entries for different transformations of other roots such as Ki& (do), wjL& (write), ‡` (give) etc.

Moreover, there are a huge number of Primary (K…r cÖZ¨q) and Secondary (ZwØZ cÖ¨Zq) suffixes used with roots and words. Each of them has own meaning [4]. For example, ‡jŠwKK (†jvK+BK), gvwmK(gvm+BK), ^`wbK(w`b+BK) etc. Here, BK is a suffix added with "†jvK", "gvm" and "w`b" words to form new words.

G. Primary suffixes (K…r cÖZ¨q): The suffixes that are used after roots to form new meaningful words called primary suffixes [4]. In the examples above B, B‡ZwQ, B‡ZwQ‡jb, †e are all primary suffixes. In addition to these there are many more primary suffixes like Ab (euva&+Ab=euvab, bvP&+Ab=bvPb), Av(co&+Av=cov) etc.

H.Secondary suffixes (ZwØZ cÖZ¨q): The suffixes that are used after words to form meaningful new words called secondary suffixes. Examples are given above with noun, adjective and pronoun morphologies. In addition, there are many other secondary suffixes like AB (cuvP+AB=cuvPB, mvZ+AB=mvZB) AvB (wgVv+AvB=wgVvB, XvKv+AvB=XvKvB, cvUbv+AvB=cvUbvB) etc.

We will outline Dictionary entries for all these primary and secondary suffixes along with their grammatical attributes, so that we can prepare thousands of Bangla Words combining with roots and words for building Bangla Word Dictionary for UNL

V. Structure of dictionary entry for Bangla morphemes

The structure of dictionary raises significant questions in UNL. The information held about lexical items must be stored efficiently and should be easily accessible by the system and by the user. Moreover, the dictionary should be readily extendable. The Universal Dictionary of Concepts must include three principal components:

i) The repository of concepts commonly referred to as the dictionary of UNL.

ii) The network of relations between concepts, which can be referred to as the UNL Knowledge Base (UNLKB).

iii) The local dictionaries, which link concepts with words of various natural languages.

A. Structure of Bangla Roots:

[HW]{}"UW"(ROOT,BANJONANT/SORANT, URoot, KPG1, KPG2…) <FLG, FRE, PRI>

HW←Head Word (Bangla Word). Here, we considered Bangla root.

UW←Universal Word (English word from knowledge base).

ROOT←Root which is an attribute for Bangla roots. This attribute is fixed for all Bangla roots.

BANJONANT/SORANT←This is another important attribute for Bangla roots. Every root is ended with vowel or consonant

URoots←This type of attribute is optional for Bangla roots. There are few Roots which are included in URoots.

KPG1←KPG1 means attribute for the name of the group 1 of Kritprottoy.

KPG2←KPG2 means attribute for the name of the group 2 of Bangla Kritprottoy.

FLG← Flag (We use B for Bangla Language)

FRE← Frequency of Head Word

PRI← Priority of Head Word

Here, some attributes are written with all capital letters and some are written with both capital and small letters. ROOT, BENJONANT/SORANT contains all capital letters because these attributes are fixed for all Bangla roots but URoot is such type attribute which will not present for all Bangla roots. Again, in case of Akpg1, Akpg2…. they will be replaced with such as AA (আ), OWA (ওয়া), YEA (ইয়ে) etc .This is for our convention.

Now, if we consider "Bangla roots" for dictionary entry:

[পড়]{}"read (icl>do)"(ROOT, BANJONANT,AA, EI,ANOW,ANTO,OAN,TI,YEA,OA,UWA)<B,0,0>

[চল]{}"go(icl>do)"(ROOT,BANJONANT,AA,EI,ANOW,ANTO,OAN,TI, YEA) <B, 0, 0>

[নাচ]{}"dance(icl>do)"(ROOT,BANJONANT,AA,EI,ANOW,ANTO,OAN,YEA) <B, 0, 0>

[পচ]{}"decay(icl>occur)"(ROOT,BANJONANT,AA,OAN,YEA,UWA) <B, 0, 0>

[মার] {} "beat (icl>do)" (ROOT, BANJONANT, AA, EI, OAN, OA,) <B, 0, 0>

[হার]{}"defeat(icl>occur)"(ROOT,BANJONANT,AA,EI,ANOW,YEA) <B, 0, 0>)

[জিত]{}"defeat(icl>occur)"(ROOT,BANJONANT,AA,EI,ANOW, YEA, OA,) <B, 0, 0>

[ধর]{}"catch(icl>do)"(ROOT,BANJONANT,AA,EI,ANOW, OAN,TI,OA) <B, 0, 0>

[কাঁদ]{}"cry(icl>do)"(ROOT,BANJONANT,AA,EI,ANOW, ANTO, OAN, YEA) <B, 0, 0>

[মর]{}"die(icl>Occur)"(ROOT,BANJONANT,AA,EI,OAN,TI) <B, 0, 0>

[ডুব]{}"sink(icl>do)"(ROOT,BANJONANT,UROOT,AA, EI, ANOW, ANTO, OAN) <B, 0, 0>

B. Structure for Bangla Kria Bivokti:

[HW] {} "" (BIV, V, PERSON, TENSE …) <FLG, FRE, PRI>

HW←Head Word (Bangla Word-KriaBivokti (ক্রিয়াবিভক্তি))

UW←Universal Word (In case of KriaBivokti, UW will be null)

BIV←Bivokti which is an attribute of KriaBivokti (ক্রিয়াবিভক্তি)

V←Verb which is an attribute of KriaBivokti (ক্রিয়াবিভক্তি) because it makes verb to add with Bangla root (**ক্রিয়ামূল**) as Suffixes.

PERSON←Attribute person which is an attribute of KriaBivokti (ক্রিয়াবিভক্তি) because the form of verb is varied according to Bangla Person (**পুরুষ**).

TENSE ←Attribute Tense which is an attribute of KriaBivokti (ক্রিয়াবিভক্তি) because the form of verb is also varied according to Bangla Tense (**কাল**).

Here, some attributes are written with all capital letters and some are written with both capital and small letters. In BIV, V contains all capital letters because these attributes are fixed for all KriaBivokti but Aperson will contain as 1P, 2P…..etc and Atense will contain as Present Indefinite, Present continuous, and Past Indefinite etc. This is our convention.

Here, I list some structures for dictionary entry of KriaBivokti (ক্রিয়াবিভক্তি)

[এ] {} "" (BIV, V, 3PG, PRI) <B, 0, 0>
[ইতেছে] {} "" (BIV, V, 3PG, PRC) <B, 0, 0>
[ইয়াছে] {} "" (BIV, V, 3PG, PRP) <B, 0, 0>
[উক] {} "" (BIV, V, 3PG, IMPS) <B, 0, 0>

[এন]{}“”(BIV,V,3PR,2PR,PRI) <B, 0, 0>
[ইতেছেন]{}“”(BIV,V, 3PR,2PR, PRC) <B, 0, 0>
[ইয়াছেন]{}“”(BIV,V, 3PR,2PR, PRP) <B, 0, 0>
[উন] {} “” (BIV, V, 3PR,2PR,IMPS) <B, 0, 0>
[অ] {} “” (BIV, V,2PG , PRI) <B, 0, 0>
[ইতেছ] {} “” (BIV, V, 2PG , PRC) <B, 0, 0>
[ইয়াছ] {} “” (BIV, V, 2PG, PRP) <B, 0, 0>
[ও] {} “” (BIV, V, 2PG,IMPS) <B, 0, 0>

Here 3PG← 3rd Person General, PRI← Present Indefinite, PRC← Present Continuous and PRP← Present Perfect

C. Structure for Bangla Krit Prottoy:

[HW] {} “UW” (KPROT, BENJONANT/SORANT, NOUN/ADJECTIVE/PROJOJOK KRIA, Gname…...) <FLG, FRE, PRI>

HW← Head Word (Bangla Word-(কৃপ্রত্যয়))

UW← Universal Word (English word from knowledge base). KPROT ← KritProttoy which is an attribute of Bangla KritProttoy.

BENJONANT/SORANT ← KritProttoy will be added with Benjonanto or Soranto. So it is another attribute.

NOUN/ADJECTIVE/PROJOJOKKRIA←to add with root Kritprottoy will make Noun or Adjective or Projojok Kria or any combination of these.

Gname ← Group Name of Krit Prottoy.

FLG← Flag (We use B for Bangla Language)

FRE← Frequency of Head Word.

PRI← Priority of Head Word.

So, structure will be like as follows:

[কৃপ্রত্যয়] {} “”(KPROT, BENJONANT/SORANT, NOUN/ADJECTIVE/PROJOJOK KRIA, Gname…...) <FLG, FRE, PRI>.

The dictionary Entries of all Kritprottoys (কৃপ্রত্যয়) are as follows:

[আ]{}“”(KPROT,BANJONANT,NOUN, AA)<B,0, 0>
[ওয়া]{}“ (KPROT, SORANT, NOUN, OWA) <B, 0, 0>
[ই]{}“”(KPROT, BANJONANT, NOUN, EE)<B,0, 0>
[আও]{}“”(KPROT,BANJONANT,NOUN,AO)<B,0,0>
[আনো] {} “” (KPROT, BANJONANT, NOUN, AANO) <B, 0, 0>
[অন্ত]{}“”(KPROT,BANJONANT,ADJECTIVE,AANTO) <B, 0, 0>
[অন]{}“”(KPROT,BANJONANT,NOUN,ON)<B,0, 0>
[তি]{}“”(KPROT,BANJONANT,NOUN,ADJECTIVE,TI) <B, 0, 0>
[ইয়ে]{}“”(KPROT,BANJONANT,ADJECTIVE,EIA)<B,0, 0>
[ও]{}“”(KPROT,BANJONANT, NOUN, ADJECTIVE, OO) <B, 0, 0>
[উয়া]{}“”(KPROT,BANJONANT,NOUN,ADJECTIVE,UWA) <B, 0, 0>

VI. Morphological rule generation for Bangla Root and Bivokti

Bangla is a Semantic language, and its basic characteristic is the rich morphology in which most of its words are derived from roots. Inflections and derivations are generated by changing vowels and insertion of consonants. Bangla sentences are characterized by a strong tendency for agreement between its constituents, between verb and noun, noun and objective, in matters of numbers, gender, definitiveness, case, person etc. These properties are expressed by a comprehensive system of affixation [5,6]. To satisfy these grammatical properties, generation rules are expected to be complex, to handle the processing of generating grammatically correct Bangla sentences from UNL expression and structure. The linguistic attributes of roots that have been used in the dictionary are basically: SORANT, BANJANT and CASE MARKER. Finally, the variations in the written forms of Bangla are also handled by making entry for each of these forms in the dictionary.

A database system has been developed for the classification and features adding for each entry in the dictionary. The system gets the UW and tries to get the equivalent Bangla word from a Bangla - UNL dictionary. The selected Bangla word is then classified to Noun or Verb or Particle. For example, if the word is denoted as having a broken plural, the system will ask the user to add this entry and both forms are linked to the same UW. It is a genitive construction in which two words are linked up in such a way that the second (second particle of the construction) qualifies or specifies the application of the first (first particle of the construction). The relation mapping is implemented in the enconversion rules.

In our system, we managed to handle this rich and complicated morphology by implementing a modular approach to coding the rules. Our implemented process of morphological generation starts by choosing the right stem which is set to accept prefixes or suffixes depending on its position and role in the sentence. Now, if we consider a sentence such as: করিম বইটি পড়িতেছি (Pronunciation is: Karim boi poriteche). To form an UNL expression it is needed Morphological, Syntactic and Semantic Analysis. But here I am concerned only Morphological Analysis. So, from the sentence “করিম বইটি পড়িতেছি” Here I just consider the word “পড়িতেছি” (Pronunciation is poriteche) for our morphological analysis. From the analysis of Bangla Root and KriaBivokti one can readily find that they agree right composition rule.

Right Composition Rule: (For Bangla root and KriaBivokti only)

-: C {ROOT :::} {BIV, V:-BIV ::}

In right composition rule, “-:C” defines the function of concatenating the string of the UW of the left-node to the string of the UW of the right-node. 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 [7].
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.

So, from our example Bangla word "পড়িতেছি" morphologically found "পড়" and "ইতেছি" where "পড়" is Bangla ROOT and "ইতেছি" is KriaBivokti

[পড়]{}"read(icl>see>do)"(ROOT,BENJONANT,AA,EE,AA NO,AANTO,TI,EIA,OO,UWA)
[ইতেছি] {} "" (BIV, V, 1P, PRC) <B, 0, 0>

EnCo can input either a string or a list of words for a sentence of a native language. A list of morphemes of a sentence must be enclosed by [<<] and [>>] [6]. When we input our word into EnCo, the Sentence Head (<<) will be on LAW, sentence texts/morphemes/words will be on RAW and the Sentence Tail (>>) will be on Right Condition Window (RCW) shown in figure 2. EnCo uses CWs for checking the neighboring nodes on both sides of the AWs in order to judge whether the neighboring nodes satisfy the conditions for applying an analysis rule or not [13].

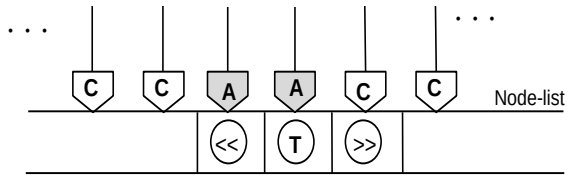

Figure 2 Initial state of the analysis window

When the sentence is input, the input word is scanned from left to right. Here, left most word of our sentence is "পড়". When an input string "পড়" is scanned, all matched morphemes with the same string characters e.g. প, পা, পড়, পড়া, পড়ি etc. are retrieved from the Word Dictionary and become the candidate morphemes according to a rule priority in order to build the syntactic tree and the semantic network (UNL expressions) of UNL for the sentence.

This rule is applied to insert the subject "পড়" of the sentence into the node-list and word "পড়" is shifted left to the next window which is LAW.

Now, EnCo analyzes the next word of the sentence "ইতেছি"like as "পড়".
When it is inserted in Enconverter it will be looked like as follows:

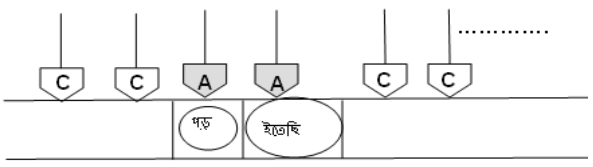

Figure 3. State of two morphemes in the analysis window

Now EnCo starts morphological analyses with the word "পড়িতেছি", to find the actual meaning of the word. It first breaks the word into "পড়" and "ইতেছি", which are available in the dictionary. The analyzer then adds the root (পড়) and suffix"ইতেছি" and find out the actual meaning of the word "পড়িতেছি" from the dictionary.
To form an UNL expression it is needed Morphological, Syntactic and Semantic Analysis. But, here I am concerned about only morphological analysis. So, from the sentence "বইটি পড়া হয়েছে". Here, we just consider the word "পড়" for our morphological analysis.
When it is inserted in Enconverter it will be looked like as follows:

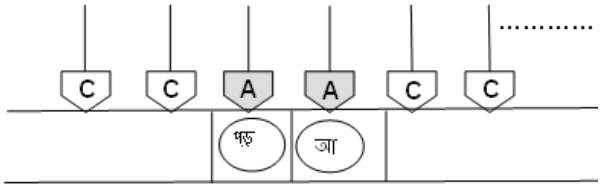

Figure 4. State of root with suffix in the analysis window

So, we see that "পড়" will be in left analysis window of the Enconverter and "আ" will be in right analysis window. EnConverter will search in our proposed Bangla dictionary and it will find the dictionary entry for "পড়" and "আ" like as follows:

[পড়]{}"read (icl>see>do)" (ROOT, BANJONANT, AA,EE,AANO,AANTO, TI, EIA, OO, UWA)
[আ] {} "" (KPROT, BANJONANT, NOUN, AA)

Again, it will search in Bangla Dictionary to find a rule for joining these two morphemes. If there is any rule exists in the dictionary then, these two morphemes will be joined and form a meaningful word. But if no rule exists then these two morphemes will not be joined. In this case it will find a rule which can join these two morphemes. According to the types of Enconversion rules [13] we know that there is a rule which is known as Right Composition [14].

Using this rule we can develop the following rules for joining the above two words.
Right Composition Rule: (For Bangla root and KritProttoy only)

-:C{ROOT,AA:::}{KPROT,AA:-KPROT,-BANJONANT,-AA ::}

Using this rule, the root "পড়"(when it is in LAW) is added with the suffix "আ" (when it is in the RAW) to form a meaningful word "পড়া". It describes that if there is a consonant ended root in group AA [আ] is in LAW and suffix AA [আ] is in RAW then two headwords will be added to make "পড়া". This rule also describes that all the attributes of the node of RAW (attributes for আ) are added with the attributes of the new word and the following attributes KPROT, BANJONANT and AA are deleted [18].

VII. Conclusion

This paper has presented structure for Bangla Roots, Krit Prottoy and Kria Bivokti which is able to generate dictionary entries for them. The following conclusions are drawn from the study.

i) In this proposed work, we can assign grammatical attributes for the roots and their suffixes.
ii) It can also develop rules for morphological analysis for Bangla words (Especially for roots and suffixes) which will be useful for conversion of Bangla sentences to UNL expressions and vice-versa.
iii) Theoretically, it proves that this model works correctly for Bangla words even though the limited number of words and rules are considered in this thesis.

In near future, the author would like to consider Bangla words in boarder perspective.

References

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

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

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

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

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

[6] P. Bhattacharyya. "Multilingual information processing using UNL". in Indo UK workshop on Language Engineering for South Asian Languages LESAI, 2001.

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

[8] D.C. Shanti, "Vahsa-prokash Bangla Byakaran". Rupa and company prokashani, Calcutta, July 199, PP.170-175

[9] UNLspecifications, http://www.undl.org/unlsys/unl/unl2005/

[10] M. E. H. Choudhury, M. N.Y. Ali, M.Z.H. Sarkar, R. Ahsan, "Bridging Bangla to Universal Networking Language- A Human Language Neutral Meta-Language", International Conference on Computer and Information Technology (ICCIT), Dhaka, 2005,pp.104-109

[11] M.E.H. Choudhury, M.N.Y. Ali, "Framework for synthesis of Universal Networking Language", East West University Journal, Vol. 1, No. 2, 2008, pp. 28-43

[12] 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

[13] Enconverter Specifications, version 3.3, UNL Center/ UNDL Foundation, Tokyo, Japan 2002.

[14] 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". 6th International Conference on Electrical and Computer Engineering,ICECE 2010, Dhaka, Bangladesh, 18-20 December 2010,pp.570-573.

[15] Md. Sadequr Rahman, Sangita Rani Poddar, Muhammad Firoz Mridha, Mohammad Nurul Huda, "Open Morphological Machine Translation: Bangla to English". NWESP'2010, November, India, page, 460-465, ISBN: 978-1-4244-7817-0.

[16] Muhammad Firoz Mridha, Md. Zakir Hossain, Manoj Banik, Mohammad Nurul Huda, Chowdhury Mofizur Rahman, Jugal Krishna Das, " Development of Grammatical Attributes for Bangla Root and Primary Suffix for Universal Networking Language," SKIMA'10, Paro, Bhutan, August, 2010,pp.61-65.

[17] Md. Zakir Hossain,Shahid Al Noor,Muhammad Firoz Mridha,"Some Proposed Standard Models for Bangla Dictionary Entries of Bangla Morphemes for Universal Networking Language", International Journal of Computer Applications(IJCA) (0975 – 8887),Volume 12– No.6, December 2010.

[18] Muhammad Firoz Mridha, Md. Zakir Hossain, Shahid Al Noor, "Development of Morphological Rules for Bangla Words for Universal Networking Language" IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.10, October 2010.

[19] 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,pp.49-54.

Author Biographies

First authors Profiles:

Muhammad Firoz Mridha is now a full time faculty in Stamford University Bangladesh. He obtained M.Sc. in Computer Science and Engineering from United International University (UIU) in 2010 and B.Sc. in Computer Science Engineering 2004 from Khulna University of Engineering and Technology(KUET). His research interest includes Natural Language Processing, Artificial Intelligence and Software.

Second authors Profiles:

Kamruddin Md. Nur attended Victoria University of Wellington, New Zealand (VUW) for his bachelor in Computer Science (2004) and United International University (UIU), Bangladesh for his masters in Computer Science (2010). He is a permanent faculty member of Computer Science department at Stamford University Bangladesh and his research interests include – Software Engineering Approaches, Software Visualization, Web Engineering, Ubiquitous and Mobile Computing, Usability Engineering, Artificial Intelligence etc.

Third authors Profiles:

Manoj Banik was born in Habiganj, Bangladesh in 1972. He completed his B.Sc. Engineering in Computer Science and Engineering from BUET, Dhaka, Bangladesh and M.Sc. Engineering in Computer Science and Engineering from UIU, Dhaka, Bangladesh. Now, he is working as an Assistant Professor in the Department of Computer Science and Engineering of Ahsanullah University of Science and Technology, Dhaka, Bangladesh. His research interests include Speech Recognition, Phonetics, Universal Networking Language and Algorithms. He is a member of Bangladesh Engineers Institute (IEB).

Fourth authors Profiles:

Mohammad Nurul Huda was born in Lakshmipur, Bangladesh in 1973. He received his B. Sc. and M. Sc. in Computer Science and Engineering degrees from Bangladesh University of Engineering & Technology (BUET), Dhaka in 1997 and 2004, respectively. He also completed his Ph. D from the Department

of Electronics and Information Engineering, Toyohashi University of Technology, Aichi, Japan. Now he is working as an Associate Professor in United International University, Dhaka, Bangladesh. His research fields include Phonetics, Automatic Speech Recognition, Neural Networks, Artificial Intelligence and Algorithms. He is a member of International Speech Communication Association (ISCA).