Some Proposed Standard Models for Bangla Dictionary Entries of Bangla Morphemes for Universal Networking Language

Md. Zakir Hossain
Department of Computer Science
Stamford University Bangladesh
Dhaka, Bangladesh

Shahid Al Noor
Department of Computer Science
Stamford University Bangladesh
Dhaka, Bangladesh

Muhammad Firoz Mridha
Department of Computer Science
Stamford University Bangladesh
Dhaka, Bangladesh

ABSTRACT
The Universal Networking Language (UNL) is a world wide generalizes form of human interactive language in a machine independent digital platform for defining, recapitulating, amending, storing and dissipating knowledge or information among people of different affiliations. The theoretical and applied research associated with this interdisciplinary endeavor facilitates in a number of practical applications in most domains of human activities such as creating globalization trends of markets or geopolitical interdependence among nations. In our research work we are presenting a pioneer work that aims to contribute with Development of Models for Bangla Dictionary Entries and Analysis of Grammatical Attributes of Bangla words such as Bangla Roots, Krit Prottoy and Kria Bivokti which will help to create a doorway for converting the Bangla Sentence to UNL and vice versa and subside the barrier between Bangla to other languages.

Keywords
Universal Networking Language (UNL), Morphology, Bangla roots, Krit Prottoy, Kria Bivokti, Model, Dictionary Entry, Morphological Rules.

1. INTRODUCTION
Today the regional economies, societies, cultures and educations are integrated through a globe-spanning network of communication and trade. This globalization trend evokes for a homogeneous platform so that each member of the platform can apprehend what other intimates and perpetuates the discussion in a mellifluous way. However the barriers of languages throughout the world are continuously obviating the whole world from congregating into a single domain of sharing knowledge and information. As a consequence United Nations University/Institute of Advanced Studies (UNU/IAS) were decided to develop an inter-language translation program. The corollary of their continuous research leads a common form of languages known as Universal Networking Language (UNL). UNL acts as an intermediate form computer semantic language whereby any text written in a particular language is converted to text of any other forms of languages [1,2]. UNL, in other words is an artificial language for computers to express information and knowledge that can be expressed in natural language. The rest of the paper is organized as the following. Section 2 outlines the UNL general structure. In Section 3 Bangla roots and primary suffixes (Krit Prottoy, Kria Bivokti) are analyzed morphologically. Section 4 provides a Model for bangla roots and primary suffixes (krit prottoy and kria bivokti) for the Bangla dictionary. Section 5 develops morphological rules for Bangla roots and primary suffixes. This paper is concluded in section 6.

2. STRUCTURE OF UNL
UNL system composed of three parts namely the Universal Words, Attribute labels and Relational Labels. Universal word which is actually nothing but a English like word and is represented by nodes in a hypergraph [3,4]. Nodes associated with a sentence are connected by a relation known as symbolic relation. Each universal word has some attributes that uniquely specifies that word and is placed according to a conceptual hierarchy derives from a knowledge base. However each of the Universal words is comprised of Headword along with some constraints. The headword is considered as the unit form of the English word known as label whereas each of the constraints in a constraint list of the Universal word corresponds to a concept of that word. The attribute lists associated with the individual universal word are used to represent the subjectivity of word based on their grammatical properties [5], [6].

The knowledge base which actually holds every possible combination of semantic relations basically plays two roles. Firstly it defines semantics (concept) of UWs and then provides linguistic knowledge of concepts. The knowledge base however not only provides linguistic knowledge in Computer understandable format but also provides the semantic background of UNL expressions.

In addition to the above parts the UNL system has a Language server which can be fragmented into two distinct parts known as enconverter (EnCo) and deconverter (DeCo). The enconverter builds a framework, independent of the diversity of languages, for morphological, syntactic and semantic analysis and converts the native language text into UNL expressions autonomously. To perform the conversion operation the enconverter uses Word dictionary, Knowledge base and Enconversion rules. In contrast the deconverter acts just the reverse way that the enconverter does. The general format of the dictionary entry is defined by UNL as follows:

[HW] "UW" (ATTRIBUTE1, ATTRIBUTE2, ATTRIBUTE3 ...|…) <FLG, FRE, PRI>

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

3. MORPHOLOGICAL ANALYSIS OF BANGLA ROOTS AND PRIMARY SUFFIXES (KRIT PROTTOY AND KRIA BIVOKTI)

Morphology is the field of linguistics that studies the structure of words. It focuses on patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. In natural language processing (NLP we need to identify words in texts in order to determine their syntactic and semantic properties [7]. In the following section we are analyzing morphologically the different Bangla Roots and Primary suffixes so that we can develop efficient Models for dictionary entries.

3.1 Bangla Roots

Every Language consists of several verbs. The center part of those verbs is called roots .In another way if we split the verbs we get two parts Root and Suffix. From verbs if we remove suffix we get root. For example 'নাচে' (Nache) means Dance is a verb. The two segments of নাচে (Nache) are নাচ (Nach) and এ (a) denotes the root and suffix (kria Bivokti) respectively Some other Bangla root verbs are চল (Chol), পড় (Por), ধর (Dhor) etc [9]

3.1.1 Soranto Roots

Verb roots that are ended with sharo-chinnoh (vowel) are called Soranto root. As for example, চা (Cha), হ (Ho), পা (Pa) etc [9].

3.1.2 Banjonant Roots

Verb roots that are ended with banjon-borno (consonant) are called Banjonant root. For example, কর্ (Kor),চল্ (Chol) ,পড় (Por) etc [9].

3.1.2 URoots

All the Roots are added with আ (AA) KRIT PROTTOY and form a meaningful word. For example:
কর্ (Kor) + আ (AA)=করা (Kora)
পড় (Por) + আ (AA)=পড়া (Pora)

However in additioin to the above form we also find some exceptional roots which don't make any meaningful word after adding with AA krit prottoy. As for example:
দুল্ (Dul) + আ (AA)=দুলা (Dula)
খুল্ (Khul)+ আ (AA)=খুলা (Khula)

Do not represent any meaningful word. Instead we use Dol (দোল) and Khol (খোল) in place of Dul and Khul respectively when we add the AA krit prottoy with them. Hence those words forms like the following procedure:
দোল (Dol) + আ (AA)=দোলা (Dola)
খোল (Khol) + আ (AA)=খোলা (Khola)

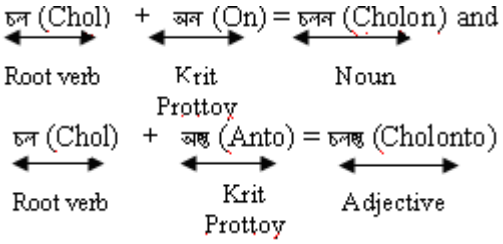
To solve this problem we divide BANJONANT Roots into two categories. One is general BANJONANT that is attributed as BANJONANT and another is attributed with URoots.

3.2 Primary Suffixes

The suffixes that are used after roots to form new meaningful words are called Primary suffixes. There are two types of Primary Suffixes: i) Krit Prottoy ii) Kria Bivokti.

3.2.1 Krit Prottoy

When sounds are added with root and form noun or adjective then the root word is called root verb and those sounds are called Krit Prottoy. For example



Some others Krit Prottoy are অন (On), অনা (Ona),অনি (Oni) etc [9].

Here we are discussing about the morphological Analysis of Bangla Roots, Krit Prottoy and Kria Bivokti so that we can analyze Bangla words. To develop Models for word Dictionary and rules we divided Bangla roots and Krit Prottoy into 9 groups: *আ (*AA GROUP); *ই গ্রুপ (*EI GROUP); *অন্ গ্রুপ (*ON GROUP); *আনো গ্রুপ (*AANO GROUP); *অন্ত গ্রুপ (*ANTO GROUP); *তি গ্রুপ (*TI GROUP); *ওয়া গ্রুপ (*OWA GROUP); *ও গ্রুপ (*O GROUP); *উয়া গ্রুপ (*UWA GROUP);

3.2.2 Kria Bivokti

The Bangla letter or group of letters which are added with Bangla roots as suffixes and form complete verbs (সমাপিকা ক্রিয়া) are called Bangla Kria Bivokti.[10]. As for example "ই" (E), "বেন" (Ben) etc.

Bivokties are changed according to Bangla tense (ক্রিয়ার কাল) and Bangla person (পুরুষ) [10]. The following table shows the Kria Bivokties for present indefinite Tense with different Person.

Table 1. Form of KriaBivokties in Present Tense

কাল (Tense)	উত্তম পুরুষ (1st P) (All)	মধ্যম পুরুষ (সাধারণ) (2nd P) (Gen)	মধ্যম পুরুষ (তুচ্ছ) (2nd P) (close)	মধ্যম পুরুষ (সমভ্রমাত্মক) (2nd P) (Res)	নাম পুরুষ (সাধারণ) (3rd P) (Gen)	নাম পুরুষ (সমভ্রমাত্মক) (3rd P) (Gen)
	আমি (Ami)	তুমি (Tumi)	তুই (Tui)	আপনি (Apni)	সে (se)	তিনি (Tini)
সাধারণ বর্তমান (Present Indefinite Tense)	-ই (E)	-অ (o)	-ইস (Es)	-এন (N)	-এ (A)	-এন (N)

Here, P means person; Gen means General; Res means Respected. Similarly the Bivokties are changed in other tenses and persons.

4. MODEL FOR DICTIONARY ENTRY OF BANGLA ROOTS AND PRIMARY SUFFIXES (KRIT PROTTOY AND KRIA BIVOKTI)

We know that Dictionary Entries are made using HW (Head Word), UW (Universal Word) and GA (Grammatical Attributes) [8]. HW is Bangla word in this case, UW is corresponding to the concept from Knowledge Base and Grammatical Attributes are Grammatical behaviors of that particular word in that particular Language. For example if we consider Bangla word "গ্রাম" (Gram) means village then its dictionary entry is:
[গ্রাম] (Gram) {} "village (icl>region)" (N, PLACE) <B, 0, 0>
where "গ্রাম" is Bangla HW, "village" is UW from Knowledge Base and N, and PLACE are its grammatical Attributes. Grammatical attributes are used frequently in morphological, syntactic and semantic analysis to develop rules for enconversion and vice-versa which is shown in details in section 5.1. If the Grammatical Attributes remain indiscriminate in dictionary then it is very difficult for us to accomplish the dictionary entry and consequently requires huge time to complete the dictionary. Moreover it might be experienced lot of complexities for those who want to further develop this Dictionary. Our total Bangla Grammar is fragmented into several sections such as Parts of speech, Prottoy, Bivokti etc. All the Bangla words in a particular section have some grammatical attributes which resemblance each other. So we are developing a standard format referring here as Model for every category of Bangla Grammar so that we can efficiently find out grammatical attributes of the Bangla words of that particular category for Dictionary Entry. Furthermore it requires less time and minimizes complication.

4.1 Model for Bangla Root

The Model that we are designing here for Bangla roots is depicted bellow:
[HW] {} "UW" (ROOT, BANJONANT/SORANT, URoot, Akpg1, Akpg2…) <FLG, FRE, PRI>
HW← Head Word (Bangla Word; in this case it is Bangla root);
UW← Universal Word (English word from knowledge base);
ROOT ← It is an attribute for Bangla roots. This attribute is immutable for all Bangla roots.
BANJONANT/SORANT ← This is another important attribute for Bangla roots. Every root is ended either Banjonant or Sorant;
URoots← This type of attribute is optional for Bangla roots.. As we discussed earlier there are some exceptional Roots which are included in URoots.
Akpg1← Akpg1 means attribute for the name of the group 1 of Kritprottoy.
Akpg2← Akpg2 means attribute for the name of the group 2 of Kritprottoy.

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 whereas URoot doesn't necessarily present for all Bangla roots. The attributes Akpg1, Akpg2 etc can be any group of Prottoy such as AA (আ), OWA (ওয়া), YEA (ইয়ে).

In the following examples we are constructing the dictionary entries for some sample Bangla roots using our designed Model:
[পড়] {} "read (icl>do)" (ROOT, BANJONANT, AA, EI, ANOW, ANTO, OAN, TI, UWA) <B, 0, 0>
[খা] {} "eat (icl>do)" (ROOT, SORANT, OWA, EI, ANOW, OAN) <B, 0, 0>.

4.2 Model for KriaBivokti

In the previous section we designed a Model for Bangla roots. However the Model for Kria Bivokti is very similar to that of Bangla roots. They only differs each other with attributes they use.
[HW] {} "" (BIV, V, Aperson, Atense,..) <FLG, FRE, PRI>
HW← Head Word (Bangla Word-KriaBivokti); UW← Universal Word (In case of KriaBivokti, UW is null);
BIV← Bivokti, which is an attribute of KriaBivokti; V← Verb, Since KriaBivokti forms verb when it is added with Bangla root as Suffiex so we keep the verb as an attribute.
Aperson← Attribute person, This is an important attribute because verb varies according to Bangla Person.
Atense ← Attribute Tense, This is also an important attribute because verb varies according to Bangla Tense.
Like Krit Prottoy some attributes are written with all capital letters and some are written with both capital and small letters. BIV, V contains all capital letters because these attributes are fixed for all KriaBivokti but Aperson can be either 1st, 2nd or 3rd person and Atense can be any tense such as Present Indefinite, Present continuous, Past Indefinite etc.

In the following examples we are introducing the dictionary entries for some sample Kria Bivokties following our proposed Model:
[এ] {} "" (BIV, V, 3PG, PRI) <B, 0, 0>; [ইতেছে] {} "" (BIV, V, 3PG, PRC) <B, 0, 0>
Here, 3PG← 3rd Person General, PRI← Present Indefinite, PRC← Present Continuous

4.3 Model for Krit Prottoy

Like Bangla Root and Kria Bivokti, the Krit Prottoy also has some attributes. The Model for Kritprottoy is defined as follows:
[HW] {} "UW" (KPROT, BENJONANT/SORANT, N/ADJ, Gname…...) <FLG, FRE, PRI>
HW← Head Word (Bangla Word-Krit Prottoy); UW← Universal Word (In case of KriaBivokti, UW is null);
KPROT ← Krit Prottoy; BENJONANT/SORANT ← This is an important attribute since we need to know whether the Krit Prottoy will be added with Banjonanto or Shoranto or both of them;
N/ADJ← Noun/Adjective, The Krit Prottoy adding with Root as suffix can form either Noun or Adjective. So this is another vital attribute; Gname ← Group Name of Krit Prottoy.

Using the above Model we are building here some sample dictionary entries of Krit Prottoy
[আ] {} "" (KPROT, BANJONANT, NOUN, AA); [ওয়া] {} "" (KPROT, SORANT, NOUN, OWA)

5. MORPHOLOGICAL RULE GENERATION FOR BANGLA ROOT AND PRIMARY SUFFIXES (KRIA BIVOKTI AND KRIT PROTTOY)

The enconverter starts applying the rules from the initial state as soon as a Bangla Word is inserted into the Node-list. EnConverter applies enconversion rules to the Node-list. The process of rule application finds a suitable rule and takes actions or operates on the Node-list in order to create a syntactic functionalities and UNL network using the nodes in the Analysis Windows. If a string appears in a window, the system will retrieve the Word Dictionary and apply the rule to the candidates of word entries. If a word satisfies the conditions required for that window, this word is selected and the rule application succeeds. This process is continued until the syntactic functions and UNL network are completed and only the entry node remains in the Node-list. Finally, it outputs the UNL network (Node-net) to the output file in the binary relation format of UNL expression.

5.1 Morphological Rule Generation for Bangla Root and Kria Bivokti

Morphological rules for Bangla root and Kria Bivokti are required frequently when morphological analysis of Bangla verb is appeared in Enconverter. Thus our designed Models of Bangla Root and Kria Bivokti can have tremendous impact for developing Bangla dictionary as well as building enconversion rules.

Let us consider the following example, আমি ফুটবল খেলিতেছি। (AAmi Football Khelitechi) means I am playing football.

To form an UNL expression it is needed Morphological, Syntactic and Semantic Analysis. But here we are concerned only Morphological Analysis of Verbs. So from the above sentence 'AAmi Football Khelitechi' we are considering the word "খেলিতেছি" (Khelitechi) which means playing for our morphological Analysis.

From the analysis of Bangla Root and KriaBivokti it is found that they agree the right composition rule [11]. 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. This rule forms a conglomerate node by coalescing two headwords of the left and right nodes. The conglomerate node supersedes the original left and right nodes and the sub-syntactic tree and attributes of the right node are inherited. If the <ACTION> field of the rule for the right node contains the operator "@", the attributes of the left node are also inherited.

So from our example Bangla word "খেলিতেছি" (Khelitechi) which means playing we morphologically deduce that "খেল" (khel) is a Bangla root corresponds to play and "ইতেছি" (Etechi) is a Bangla Kria Bivokti.
Thus we can arrange the following dictionary entries for our given example:
[খেল](Khel) {} "play (icl>do)" (ROOT, BANJONANT, AA, EI, ANOW, YEA, OA) <B, 0, 0>
[ইতেছি](Itechi) {} "" (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 [>>] [11]. When we input our word into EnCo, the Sentence Head (<<) is contained in Left Analysis Window (LAW), that of texts/morphemes/words in Right Analysis Window (RAW) and the Sentence Tail (>>) is in Right Condition Window (RCW) as shown in figure 2. [11]. 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.

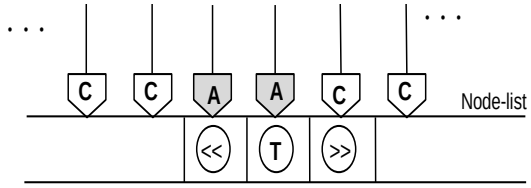


Figure 1. Initial state of the analysis window

Here, the input string "খেল" (Khel) is analysed; all matched morphemes with the same string characters associated with খ (Kha) and ল(La) e.g. খ (Kha), খেল (Khel), খেলা (Khela) , খেলি (Kheli) etc. are retrieved from the Word Dictionary and become the candidate morphemes according to a rule priority. This rule is applied to insert the subject খেল"(Khel) meaning play of the sentence into the node-list and word "খেল"(Khel) is shifted left to the next window which is LAW.
Now EnCo analyzes the next word of the sentence "Itechi" in the same way as "khel".
When it is inserted in Enconverter it is looked like as follows:

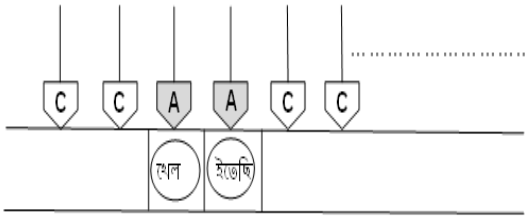


Figure 2. State of two morphemes in the analysis window

Now EnCo starts morphological analysis with the word "Khelitechi" to find the actual meaning of the word. It first breaks the word into "Khel" and "Itechi" which are available in the dictionary. The analyzer then adds the root (Khel) and suffix "Itechi" and find out the actual meaning of the word "Khelitechi" from the dictionary.

6. CONCLUSION AND FUTURE WORK

This paper presents Models for Bangla Root, Krit Prottoy and Kria Bivokti which avail dictionary entries of root, Krit Prottoy and Kria Bivokti. In this proposed work, we can assign grammatical attributes for the roots and their suffixes as well as can 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. Even though the limited number of words and rules are considered in this paper, it theoretically shows that the designed model works perfectly for Bangla words. All the Bangla words and rules will be considered in future.

7. REFERENCES

[1] Ronaldo Teixeira Martins, Lúcia Helena Machado Rino, Maria das Graças Volpe Nunes, Osvaldo Novais Oliveira Jr. The UNL distinctive features: inferences from a NL-UNL enconverting task.

[2] Uchida, H., Zhu, M. and Della Senta, T. (2000). UNL: A Gift for a Millennium. The United Nations University.

[3] Gilles St~Rasset, Christian Boitet. On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter.

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

[5] Bouguslavsky, I., Frid, N. and Iomdin, L. (2000). Creating a Universal Networking Module within an Advanced NLP system. Proceedings of the 18th International Conference on Computational Linguistics, pp. 83-89.

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

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

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

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

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

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

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

[13] M. M. Asaduzzaman, M. M. Ali, "Morphological Analysis of Bangla Words for Automatic Machine Translation", International Conference on Computer and Information Technology (ICCIT), Dhaka, 2003, pp.271-276.