INFLECTIONAL MORPHOLOGY SYNTHESIS FOR BENGALI NOUN, PRONOUN AND VERB SYSTEMS

Samit Bhattacharya (samit@cse.iitkgp.ernet.in), Monojit Choudhury (monojit@cse.iitkgp.ernet.in), Sudeshna Sarkar (sudeshna@cse.iitkgp.ernet.in) and Anupam Basu (anupam@cse.iitkgp.ernet.in)

Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur, West Bengal, India

Abstract:
Morphological synthesis is an essential part of any natural language generation system. bengali is a highly inflectional language with more than 160 different inflected forms for verbs and 36 different forms for nouns, and 24 different forms for pronouns. Moreover, the choice and ordering of suffixes for nouns and structural changes in verb morphology are dependent on the type (animate/inanimate etc) and syllable structure of the root respectively. We describe here a rule-based approach for Bengali morphological synthesis. For nouns and pronouns, the synthesis engine selects the appropriate suffixes, determines the ordering among them, and concatenates them according to the rules. The verb roots have been classified under 19 categories based on the syllable structure. An FSM is used to recognize the category of the root. Thereafter, a simple algorithm generates the inflected form of the verb based on suffix table and rule tables for the classes.

1.0 INTRODUCTION

Several *natural language processing* (NLP) activities on Indian languages being carried out in our institute require robust *natural language generation* (NLG). *Iconic communication* system which synthesizes valid bengali sentences that can be used for communication, given the input of a sequence of icons [6] and *machine translation* (MT) system [5] where the target language is Bengali are typical examples where we need NLG. Being a highly inflectional language, bengali NLG calls for efficient morphological synthesis. Although many linguistic works on bengali *inflectional morphology* are available [1,7], there is hardly any work from the computational perspective. Moreover, it is not clear how to directly use the linguistic knowledge for developing a computational framework. These lead us to the design of a complete computational model of inflectional morphology for bengali noun, pronoun and verb systems and development of a *morphology synthesizer* for bengali based on it.

Morphology or the composition and organization of the words can be broadly classified into two types – *inflectional* and *derivational* [1]. In *inflectional morphology* the basic meaning and part-of-speech of the root word and the morphed word are the same. The addition of affixes in this case changes certain features of the root such as its number or gender in case of nouns; number, gender, tense and (or) aspect for verbs. *Derivational morphology* on the other hand refers to the derivation of new words with different meaning and possibly different part-of-speech from the root word through various morphological operations. On the basis of the *structural changes* of the root and other morphemes during affixation, morphology can be classified as *linear* or *nonlinear*. In *linear morphology* affixes are added to the root without changing the internal structure of the root, though some *euphonic changes* might take place at the boundary (of the affixes and the root). On the other hand, morphological systems where the internal structure of the morphemes changes during the addition of affixes are classified as *nonlinear morphology* [2]. *Pluralization* and *adjectivization* in english [3] normally pertains to linear morphology whereas Semitic languages [2] and derivational morphology in Sanskrit [4] features nonlinearity. The inflectional morphology of noun and pronoun systems in bengali is linear in nature whereas the system of verbs displays certain nonlinear characteristics.

In this paper, we describe the computational framework of inflectional morphology for bengali noun, pronoun and verb systems and the morphology synthesizer. The paper is organized as follows. Section 2 begins with the motivation for this work by describing the applications, nuances and challenges of bengali morphology. Section 3, 4 and 5 discuss the algorithms and data structures underlying the morphological synthesizers for nouns, pronouns and verbs respectively. Section 6 concludes this paper by identifying possible future works in this direction and raising certain interesting linguistic issues that are worth pondering.

2.0 MOTIVATIONS

Design of algorithms and systems for morphological synthesis is both important and interesting for three reasons. Firstly, it is a pivotal part of any system that communicates in natural language. Secondly, due to inherent complexity of bengali morphology, modeling of it provides interesting computational challenges. Thirdly, it can provide very useful insight into language acquisition, evolution and many other problems of linguistics. In this section, we briefly describe the need of morphological synthesis for NLG. The basic morphological structure of bengali is described thereafter, illustrating the highly inflected nature of the language. The third point will be briefly discussed in the concluding section.

2.1 Importance of Morphological Synthesizer in NLG

Generation of syntactically and semantically correct sentences requires appropriate choice among the different forms of nouns, pronouns and verbs. In a sentence, the agreement of the verb with the number, gender and (or) person of the subject, proper case markers for the different nominal forms and expression of the tense, aspect and modality (TAM) of the verb and number, specificity of the nouns are some of the important syntactic constraints governing correct generation. In bengali, many of these features are expressed not by separate lexical items, but by different morphological forms of the verbal, nominal or pronominal roots. Thus, morphological synthesizer is an essential part of bengali NLG.

The problem of morphological synthesis can be formally defined as follows: Given a root word **r** (verb or noun) and a list of relevant attributes, $\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_3}$, …, $\mathbf{a_n}$, the morphological synthesizer generates the word **w**, which has **r** as root and $\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_3}$, …, $\mathbf{a_n}$ as the morphological attributes. For example, consider the english verb root "do" that takes the form "does" in an *indicative* sentence, when the subject is *third person, singular* and the tense is *simple present*. Here, **r** is "do" and the attributes are <*indicative, present, simple, singular, 3*rd *person* >. The corresponding root in bengali is ***kara'*** [1] and the morphed word is ***kare***. In this definition, along with the root, the inputs are the attributes and not the suffixes. This is because the choice of suffix is context dependent and may not be known *a priori*.

It should be mentioned here that in all the above cases morphological synthesis could be avoided by explicit storage of all the inflected forms of words in the lexical database. However, this approach suffers from several problems. There are 56 different forms of a bengali verb root, which can give rise to 112 new forms by the addition of emphasizing suffixes (***i*** and ***o***). Thus a single verb root can have as large as 168 different forms. Similarly a nominal root can have more than 36 different forms and pronouns can have no less than 24 forms. Thus, an all word dictionary is at least 36 times larger than a root dictionary. Large database not only increases the search time,

[1] In this paper, bangali graphemes are written using Roman Script in the ITRANS notation. They are written in bold and italics fonts. Often schwas (***a***) are deleted in the pronunciation of bangali words. Such schwas will be denoted as ***a'***. Important contexts will be underlined.

but also is a real bottleneck if we want to port the applications on small handheld devices with limited memory. Moreover, manually creating this resource is a laborious task, and prone to errors. Secondly, since the morphological rules in bengali are productive, the speaker has the liberty to add suffixes even to the proper nouns. Thus, storage of all possible inflected words is practically impossible. Therefore, morphological synthesis is an important aspect of many bengali NLP applications.

2.2 Morphological Structure of Bengali

Bengali belongs to the Indo-Aryan family of languages that evolved from Sanskrit. It has been highly influenced by the vocabulary and syntactic structure of Persian as well [7]. There are more than 100 different dialects of bengali being spoken in parts of West Bengal, Tripura and Bangladesh. The Standard Colloquial Bengali (SCB) and the other dialects have evolved from an earlier form of bengali known as *Sadhu bhasha* (the tongue of the learned). Unlike Hindi, which is a *gender-number language*, bengali is a *classifier language*. This means that in bengali the verb does not change its form based on the gender and number of the subject/object, as is the case in Hindi; rather it changes with respect to tense, aspect, modality and person only. Including the different spellings, there are 56 inflected forms of a root verb in bengali. Table 1 lists the suffixes for these different forms. Although *compound verbs* are quite common in bengali, in this paper we shall consider synthesis of simple verbs only, because a compound verb, by definition, is a cluster of finite and non-finite forms of several simple verbs all of which are inflected according to the rules to be discussed here.

Table 1: The suffix table for different forms of the verbs

Person/ TAM	1st	2nd, familiar	2nd, normal	2nd & 3rd formal	3rd
Ind, Pr, Simple	***i***	***isa'***	ϕ	***ena'***	***e***
Ind, Pr, Cont	***chhi***	***chhisa'***	***chha***	***chhena'***	***chhe***
Ind, Pr, Perfect	***echhi***	***echhisa'***	***echha***	***echhena'***	***echhe***
Ind, Pa, Simple	***lAma'/luma'***	***li***	***le***	***lena'***	***la***
Ind, Pa, Cont.	***chhilAma'/ chhiluma'***	***chhili***	***chhile***	***chhilena'***	***chhila***
Ind, Pa, Perfect	***echhilAma'/ echhiluma'***	***echhili***	***echhile***	***echhilena'***	***echhila'***
Ind, Future	***ba***	***bi***	***be***	***bena'***	***be***
Habitual Past	***tAma'/tuma'***	***tisa'***	***te***	***tena'***	***ta***
Imperative	-	***.h/***ϕ	ϕ	***una'***	***uka'***
Neg, Perfect	***ini***	***isa'ni***	***ani***	***ena'ni***	***eni***

Legends: Ind – indicative, Pr – present, Pa – past, Cont – continuous, Neg – negative.

As stated before, bengali verb morphology has some non linear characteristics. Often, the root changes its form when certain suffixes are added to it. For example, the verb ***khA*** (to eat) when followed by the suffix ***chhi*** (present continuous, 1st person) becomes ***khAchchhi***, whereas when followed by the suffix ***echhi*** (present perfect, 1st person) it becomes ***kheYechhi***. Similarly, ***hAra'*** (to lose) followed by these suffixes becomes ***hAra'chhi*** and ***herechhi*** respectively. Thus, the addition of the suffix ***echhi*** changes the root forms of ***khA*** to ***khe*** and ***hAra'*** to ***hera'***, which is an indication of nonlinearity.

Being a classifier language, bengali has several class specifying suffixes that are added to nouns. For example, ***TA*** and ***Ti*** signify singularity and specificity and ***gulo*** or ***guli*** signify plurality and specificity for all except proper nouns. Thus, ***Ama'TA*** means "the mango" whereas ***Ama'gulo*** means "the mangos", where ***Ama'*** stands for "mango". The suffix ***era'*** is *possessive or oblique* marker. Thus, ***Amera'*** means "of mango". These suffixes can be appended one after the other following certain ordering rules. For instance, ***Ama'TAra'*** means "of the mango" and ***Ama'gulora'*** means "of the mangos". The suffixes ***i*** and ***o*** stand for "only" and "also" respectively and can be added to any verbal or nominal form. These suffixes are added at the end of the word after which no other suffix can be added.

The aforementioned facts make the computational modeling of bengali morphology hard and challenging. The large number of inflected forms of verbs and the nonlinearity involved in its morphology make it a complicated problem, which cannot be solved elegantly by the traditional methods of rewrite rules and FSTs [8]. The large number of possible suffixes that can be added one after another provides further computational challenges in bengali noun morphology.

3.0 SYNTHESIS OF NOUN MORPHOLOGY

3.1 Issues in Noun Morphology

The three key issues in noun morphological synthesis are:

- Choice of appropriate suffixes
- Ordering of the suffixes
- Surface-level changes at the boundary while *affixation*

Table 2 gives a list of bengali suffixes, their attributes, constraints and rules for affixation. We have classified the suffixes into three categories – *classifiers*, *case markers* and *emphasizers*. *Classifiers* are suffixes that are added to qualify the class property of the noun such as singularity or plurality, specificity or generality etc. The choice of *classifiers* depends on the type of the noun. For a proper noun, no suffix is added for singularity and specificity, ***rA*** is added for plurality. For other nouns ***TA***, ***Ti***, ***khAnA*** etc. are added for specificity and singularity, whereas ***gulo*** and ***guli*** are added for plurality. *Case markers* are suffixes, which signify the case relation of the noun with the verb or other nouns. The suffixes ***ra'*** (*possessive-marker*) and ***ke*** (accusative-marker) are two examples of case markers in bengali. Both of these suffixes can act as *oblique-markers* as well. *Emphasizers* ***o*** and ***i*** can be added to all of these forms.

3.2 Modelling Noun Morphology

Only one suffix from each of the three groups can be added to a noun root. Thus, at most three suffixes can be added in sequence. The most common ordering of the suffixes is *classifiers* followed by *case markers* followed by *emphasizers*. Thus, ***Ama'gulokei*** (***Ama' + gulo + ke + i***, "these mangos only", accusative) is a well-formed word, whereas, ****Ama'keguloi*** or ****Ama'iguloke*** are not. Rarely, forms such as ***Amera'TA*** (***Ama' + ra' + TA***) are used, but its meaning is different from that of the more common form ***Ama'TAra*** (***Ama' + TA + ra'***, "of the mango"). ***Amera'TA*** is a phrase-level compression for constructs like ***Amera' bAksaTA*** ("the box of mango"), where ***bAksa*** is supposed to be clear from the context.

While affixation, certain surface-level changes take place that are handled by simple set of rewrite rules as shown in table 2. The surface level changes in bengali mainly depend on the last vowel of the root and the suffix added. Thus, when ***ra'*** is added to a word like ***ghara'***, the inflected word becomes ***gharera'***; whereas when it is added to some root ending in ***A***, like ***mAlA*** (garland) it becomes ***mAlAra'*** and not ****mAlera'***.

Table 2: The list of inflectional suffixes that can follow bengali nouns and pronouns

Suffix	Attributes	Rewrite Rules*	Ordering	Constraints/Remarks
CLASSIFIER SUFFIXES:				
TA	1. Singular, specific (***Ama'TA***, ***meYeTi***)	α +***T*x**	Normally, the **first** suffix	1. Only with common noun

Ti ***Te*** ***To***	2. With Numerical constants (***eka'TA***, ***duTo***) 3. With quantifying adjectives & pronouns (***aneka'TA, kataTA***)	→ α*T***x** Where, **x** is ***A***, ***i***, ***e***, or ***o***	after root. Very rarely follows ***era'***	2. ***Te***, ***To*** goes with numerical constants whose last syllable is i and o respectively
khAnA ***khAni***	1. Singular, specific (***Ama'khAnA***, ***raMkhAnA***) 2. With Numerical constants (***tina'khAnA***)	α +***khAn*x** → α***khAn*x** Where, **x** is ***A*** or ***i***	The **first** suffix after the root	1. Normally goes with inanimate nouns
jana'	1. With numerical constants (***chAra'jana'***) 2. With quantifying adjectives (***kaYeka'jana'***)	α + ***jana*** → α***jana***	The **first** suffix after the root.	1. Goes with numerical adjectives qualifying human entities.
tuku	1. With quantifying pronouns (***katatuku***) 2. With concrete nouns (***bhAta'tuku***)	α + ***tuku*** → α***tuku***	The **first** suffix after the root.	It is used in a sense of small amount.
gulo ***guli***	1. Plural, specific (***Ama'gulo***, ***rAstAguli***) 2. With quantifying adjectives (***aneka'gulo***) 3. Pronouns (***kataguli***)	α + ***gul*x** → α***gul*x** Where, **x** is ***o*** or ***i***	Normally, the **first** suffix after root. Very rarely follows ***era'***	1. Goes with common nouns and pronouns standing for and Adjectives qualifying common nouns.
rA	1. Plural, generic (***rAmerA***, ***pAkhirA***)	**αV + *rA*** **→ αV*rA*** **(V ≠*a*)** **α*a'* + *rA*** **→ α*erA***	Normally, the **first** suffix after root.	1. Normally goes with proper nouns or human entities but in a generic sense can also be appended to animate common nouns.
CASE MARKER SUFFIXES:				
ra'	singular (unless some plural marker suffix like ***gulo*** already present) , possessive/oblique. (***Amera'***, ***baiYera'***, ***dAdAra'***)	**α*a'* + *ra'*** **→ α*era'*** **αVV + *ra'*** **→αVV*Yera'*** **αCV + *ra'*** **→ αCV*ra*** **(V ≠*a'*)**	Normally **follows** the Classifier Suffixes (if present)	Can follow any of the classifier suffixes except ***rA***
dera'	Plural, possessive/oblique (***rAmadera'***, ***pAkhidera'***)	α + ***dera'*** → α***dera'***	The **first** suffix after the root.	Normally goes with human entities and proper nouns, but can follow animate common nouns in a generic sense.
ke	Accusative case	α + ***ke*** → α***ke***	**Follows** the classifier suffixes	Goes after all the classifier suffixes except ***rA***.
e	(*Adhikarana karak*) similar to in, on, among, at etc. prepositional phrases (***ghare***, ***baiYe***)	**αCa + *e*** **→ αC*e*** **αCA + *e*** **→ αC*AYa*** **αCVV + *e*** **→αCVV*Ye***	**Follows** the classifier suffixes	Goes after all the classifier suffixes except ***rA***. Does not go after words ending in **C*i***, **C*e*** or **C*u***
te	(*Adhikarana karak*) similar to in, on, among, at etc. prepositional phrases (***mATite***, ***jutote***)	**αV + *te*** **→ αV*te*** **αC*a'* + *te*** **→ αC*ete***	**Follows** the classifier suffixes	Goes after all the classifier suffixes except ***rA***. Follows words ending in **C*i***, **C*e*** or **C*u***
EMPHASIZING SUFFIXES:				
i	only	α + ***i*** → α***i***	The **last** suffix	Can follow any noun, adjective, pronoun or verbs

o	also	α + ***o*** → α***o***	The **last** suffix	Can follow any noun, adjective, pronoun or verbs

In the rewrite rules, α denotes any string, **C** and **V** stand for consonant and vowel respectively. Apart from the constraints mentioned, a word cannot have two suffixes from the same group.

3.3 Implementation

The noun morphology is implemented in a rule-based manner, where the rules are stored in tables indexed by the attributes, as well as the suffixes. For an input $\{\mathbf{r}, \quad < \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots, \mathbf{a}_n>\}$, the system first identifies the attributes as classifiers, case markers or emphasizers. If a classifier attribute (which need classifier suffixes for their expression e.g. are number and specificity), is present the system uses the dictionary to find out the type of the root noun **r**. If the entry is absent, the system assumes it to be a proper noun. Based on the type of **r**, classifier suffixes are chosen from the table. Depending on the classifier suffixes and case marker attributes, appropriate case marker suffixes are chosen next. Emphasizers are independent of type of **r** or any other attributes and thus readily chosen. After obtaining the complete list of suffixes the system concatenates the suffixes in order, taking care of the boundary changes as specified by the rewrite rules.

4.0 SYNTHESIS OF PRONOUN MORPHOLOGY

The system of pronouns in bengali is similar to that of nouns. The only difference lies in the concatenation rules. To illustrate this point, consider the example of the first person pronoun ***Ami*** ("I"), which has the forms ***Ama'rA*** ("we", plural), ***AmAra'*** ("my", singular possessive), ***AmAke*** ("to me", singular, accusative), ***AmAdera'*** ("our", plural possessive) etc. The suffixes used – ***rA***, ***ra'***, ***ke***, and ***dera'*** respectively in the above examples are all used with nouns as well (table 2). The ordering rules are also same. However, the root form should not be considered as ***Ami*** here; rather it should be taken as ***Ama'***. Similarly, for other pronouns we define ***toma'*** (2nd person, normal), ***Apa'nA*** (2nd, 3rd person formal), ***to*** (2nd person, familiar) etc. as roots rather than ***tumi***, ***Apa'ni*** and ***tui*** respectively. The rewrite rules have been designed in accordance with the choices of the roots, which in turn helped to get simpler rules. The uninflected forms ***Ami***, ***Apa'ni***, ***tumi*** etc. are stored. Since the choice of suffixes and implementation are quite similar to that of nouns described above, the details of the pronoun synthesis system will be omitted; instead examples of two concatenation rules are given below to illustrate the formalism. The first rule is for concatenation of the suffix ***rA*** and the second pair of rules is for concatenation of ***ra'***. [α stands for any string of letters.]

α + ***rA*** → α***rA***

α***a'*** + ***ra'*** → α***Ara'***
α(***o/A)*** + ***ra'*** → α***(o/A)ra'***

Thus, ***Ama' + rA*** is ***Ama'rA***, ***Ama' + ra'*** is ***AmAra'*** whereas, ***to + ra'*** is *tora'* and ***Apa'nA + ra'*** is ***Apa'nAra'***. These rules have been implemented using FSTs generated by standard rewrite-rule compilers.

5.0 SYNTHESIS OF VERB MORPHOLOGY

5.1 Modelling Verb Morphology

The 2nd person, familiar imperative form of the verb is considered here as the root. Thus, for the verb ***karA*** (to do), ***kara'*** is the root. Based on the syllable structure, the bengali verb roots have been classified into 24 categories as shown in table 3. Only the vowels of the last two syllables are considered during the classification. Empirical observations confirm that roots belonging to the same category behave in the same way under morphological operations. If the root is monosyllabic, then the vowel of the second syllable is considered to be *null*. Thus, the structure of the root can be described by the regular expression:

X*VC*$

where, **C** stands for consonants, **V** stands for vowel, **$** stands for one of the four bengali verb root endings, and **X** stands for any character. The four ways in which a bengali verb root can end are *null*, ***a'***, A or ***oYA***. Therefore, **$** $\in$ {ϕ, ***a'***, ***A***, ***oYA***}.

It should be noted that the graphemic forms of 2nd person familiar imperative for the verbs ***lekhA*** (to write) and ***dekhA*** (to see) are the same – ***lekha'*** and ***dekha'*** respectively, but the pronunciations are different (*lekh* and *dEkh*[2] respectively). These two verbs belong to different categories. The *1st person*, *indicative*, *present continuous* forms of the two are ***likha'chhi*** and ***dekha'chhi*** respectively. Therefore, we shall consider the root form of the former as ***likha'*** and the later as ***dekha'***. This choice has two motivations. Firstly, the bengali graphemic system has no single grapheme representing the sound '*E*'. Secondly, the verb ***lekha'*** has been derived from the Sanskrit verb ***likh.h***, which is what we have chosen as the root.

Like noun morphology, verbs in bengali also take a series of suffixes, each of them marking tense ***(il*** for past tense, ***b*** for future), aspect (***chh*** for continuous**,** ***echh*** for perfect), person (***Ama'*** for 1st person, ***ena'*** for 2nd person formal) and modality (***t*** for habitual past). However, unlike noun morphology, this system is quite complex and irregular. For example, although ***Ama'*** is a suffix marking 1st person as in ***kara'lAma'*** ("I did") or ***kara'tAma'*** ("I used to do"), it is not used in present (***kara'chhi***, "I am doing") or future (***kara'ba***, "I shall do") tenses. In order to avoid such complications, the suffixes for bengali verbs are not considered at the atomic level marking TAM or person; rather, composite suffixes such as ***lAma*** (past, simple, 1st person), ***echhilena*** (past, perfect, 2nd person, formal) are treated as basic units. The list of suffixes is categorized according to the attributes and stored as shown in table 1. We shall call it the *suffix-table*. Under this simplification, verbs can take only one composite suffix, which uniquely determine its TAM and person attributes. The emphasizing suffixes ***i*** and ***o*** are the only ones that can be then added to the inflected verb. The next two subsections describe procedures for identification of the category of a verb root and use of this knowledge for concatenation of the suffixes.

Table 3: Classification of bengali verbs based on syllable structure

$ **V**	ϕ	***a'***	***A***	***oYA***
a	***ha*** [***haoYA***] (to happen)	***kara'*** [***karA***] (to do)	***karA*** [***karAno***] (do, causative)	***saoYA*** [***saoYAno***] (undergo, causative)
A	***khA*** [***khAoYA***] (to eat)	***jAna'*** [***jAnA***] (to know)	***jAnA*** [***jAnAno***] (to inform)	***khAoYA*** [***khAoYAno***] (to feed)
i	***di*** [***deoYA***] (to give)	***likha'*** [***lekhA***] (to write)	***ni~NrA*** [***ni~NrAno***]	--
e	--	***dekha'*** [***dekhA***] (to see)	***dekhA*** [***dekhAno***] (to show)	***deoYA*** [***deoYAno***] (give, causative)
o	***so*** [***so;oYA***] (to lie down)	***tola'*** [***tolA***] (to pick)	***tolA*** [***tolAno***] (pick, causative)	***so;oYA*** [***so;oYAno***] (lie, causative)
u/au	--	--	***ghumA*** [***ghumAno***] (to sleep)	--

There are 24 classes out of which 5 were found to be empty (consisting of no representative verb in use), making total of 19 valid categories. One example for each category is provided along with the verbal noun form in square brackets and the meaning in parentheses.

5.2 Identification of the Category of the Verb Roots

The first step to verb morphological synthesis is identification of the category of the root. This task is accomplished by an FSM shown in Figure 1. The FSM scans the verb root from right to left and

[2] '*E*' stands for the phoneme often represented graphemically by '***yA***' as in ***byAkula'*** or ***e*** as in ***eka'***.

makes transitions for each character. When it reaches a final state, the label of that state indicates the category of the root. During the execution of the FSM, it automatically identifies the substrings represented by **$, C*, V** and **X*** in the regular expression **X*VC*$** for the root.

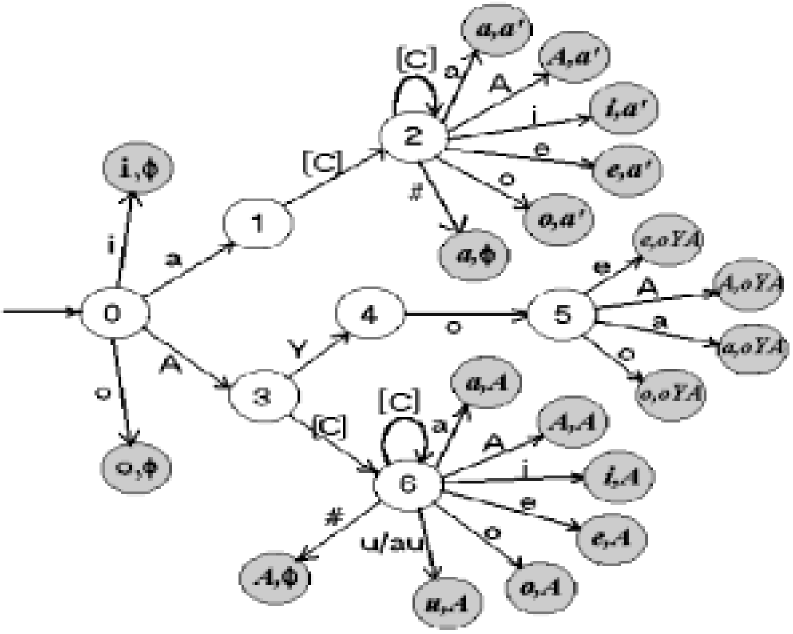

Figure 1: FSM for recognition of the class of a bengali verb root

The final states are drawn in gray and the labels indicate the category as per table 2. [C] and # denotes any consonant and the end of input respectively. The other alphabetical symbols are bengali graphemes in ITRANS notation.

5.3 Suffixes Concatenation

The process of appending the suffixes to the root verb is not trivial in bengali; that is where the information about the category of the verb is required. The suffixes added to the verbs are independent of the category. However, depending on the category, there are changes that take place at the boundary and at the internal structure of the root while concatenation. The rules for handling the changes during the concatenation process are stored in the *rule-tables*. There are rule-tables for each of the verb categories. A typical rule for concatenation is of the form

X*AC*a’ (root**) + echhi (**suffix**) → X*eC*echhi** (morphed form)

For example, ***hAra’*** + ***echhi*** **→** ***herechhi*** (“I have lost”). Here, the substrings **X*** and **C*** are ***h*** and ***r*** respectively. After identification of the category and the different substrings (**X*, C*** etc.) of the root by the method described in the last section, the relevant *rule-table* is selected. Based on the input attributes, the suffix is selected from the *suffix-table*. The corresponding entry in the rule-table for that suffix stores the rule for concatenation. As the suffix and root forms are already known, the rule-table needs to store information about the type of the change only, which for the above example would be the expression **X*eC***. This information is enough along with the suffix-table to generate the inflected form of the verb. The complete algorithm for synthesis of verb inflectional morphology is presented in figure 2.

```
procedure Inflect_verb
Input: root, tense, aspect, modality, person
Output: inflected_word
begin
1. Run the FSM of figure 2 on root (from right to left) as input.
2. Output of 1:  the substrings X*, V, C*, $ and category_of_root
3. suffix =  suffix_table[tense][aspect][modality][person]
4. rule_table = Tables[category_of_root]
5. rule = rule_table[tense][aspect][modality][person]
   /* Note that rule is a string containing X*, C*, V and/or $ */
6.1 replace the expressions X* and C* in rule by
      the substrings X* and C* identified in step 2
6.2 If rule has V and/or $ replace them by the same substrings identified in 2.
      Call the new substring stem
7. inflected_word = stem+suffix  /* '+' stands for string concatenation*/
end
```

Figure 2: Algorithm for synthesis of verb inflectional morphology

6.0 CONCLUSION

This paper has described some novel approaches for bengali morphological synthesis of the noun, pronoun and verb systems. The systems described have been implemented in java and are platform independent. They have been provided with user-friendly GUI not only for testing, but also for using these products as bengali learning tools. The performance of the systems has been tested on a large number of randomly selected words and was found to be mostly accurate. There are some exceptions to the verb morphology, such as the verb ***yA*** ("go") in *perfect* aspect becomes ***ge*** as in ***gelAma'*** (past simple, 1st person) or ***gechhe*** (present perfect, 3rd person). The verb ***Achha'*** ("be") in past tense becomes ***chhi*** as in ***chhilAma'*** (1st person). Such exceptions have been handled by explicitly storing all the inflected forms of the *defective verbs*. There are very few defective verbs in bengali and hence the list is quite small.

One interesting question to ask in the context of bengali verb morphology is how to explain the nonlinear behavior of the verbs, when the language does not show any such nonlinearity in the contexts of nouns, pronouns or adjectives. The answer to this question lies in the evolutionary history of today's bengali (SCB). SCB originated from *Sadhu Bhasha*, which has quite structured verb morphology. To illustrate this point, consider the morphed forms ***hAriYAchhe*** (***hAra'*** + ***iYAchhe***, 3rd person, present perfect) or ***kariYAchhe*** (***kara'*** + ***iYAchhe***, 3rd person, present perfect), which are quite symmetric compared to their counterparts ***herechhe*** and ***karechhe*** of SCB. The conversion of ***A*** to ***e*** in SCB for the root ***hAra'*** during the morphological transformation seems spooky. However, it is not so if we consider the forms ***hAriYAchhe*** and ***kariYAchhe*** and apply a series of identical phonological transformations to both to obtain the present forms. We believe that an optimization model can be constructed based on simple phonological phenomena like *harmony*, which when applied, can convert the inflected forms of verbs in *Sadhu bhasha* to their present forms. Since the inflected forms in the *Sadhu bhasha* are quite structured, these would provide an alternative and elegant solution to bengali verb morphological synthesis. There is a lot of research scope in this direction of evolutionary modeling of bengali and other languages, which can solve not only many NLP application problems, but can also answer many fundamental questions of linguistics.

7.0 REFERENCES

[1] S. Sen, *Bhashara Itibritta*, Ananda Publishers Pvt. Ltd., pp. 122 – 263, 2001

[2] G. A. Kiraz, *Computational Nonlinear Morphology with Emphasis on Semitic Languages*, Cambridge University Press, UK, 2001

[3] R. Sproat, *Morphology and Computation*. MIT Pres, Cambridge, MA. 1992

[4] A. A. Macdonell, *Vedic Grammar*. Munshiram Manoharlal Publishers Pvt. Ltd., New Delhi, 2000

[5] Nirenburg S. (Ed.), *Machine Translation: theoretical and Methodological Issues*. Cambridge University Press, Cambridge, 1987

[6] A. Basu *et al*, “Vernacular Education and Communication Tools for People with Multiple Disabilities” in *Proceedings of DYD02*, Bangalore, Dec. 2002

[7] S. K Chatterji, *The Origin and Development of the Bengali Language*, Rupa Co., New Delhi, 2002

[8] Kaplan R. M., Kay M., “Regular models of phonological rule systems” in *Computational Linguistics*, Vol. 20, no. 3, pp. 331-378, Sept. 1994