Morphological Analysis of Inflecting Compound Words in Bangla

Sajib Dasgupta, Naira Khan, Asif Iqbal Sarkar, Dewan Shahriar Hossain Pavel
and Mumit Khan
BRAC University, Dhaka, Bangladesh.
sajib44new@bracuniversity.net, naira@bracuniversity.net, asif@bracuniversity.net,
pavel@bracuniversity.net, mumit@bracuniversity.net

Abstract

The addition of inflectional suffixes in Bangla com-pound words is fairly complex. A compound is a word that is formed by two or more different words acting as a single entity. One of the key distinguishing features of compounds is the absence of inflectional morphology between the constituents of a compound. In Bangla, however, the constituents may retain inflectional suffixes on either or both the constituents and the resultant compound may then be inflected further as a whole word. Such inflection creates ambiguities as context-free word grammar is unable to recognize whether the inflectional suffix is an inflectional property of the last constituent root-word or of the compound as a whole. We use a feature unification based morphological parser, which can successfully and efficiently parse compound words that retain such inflectional morphology and at the same time resolve such ambiguities.

1. Introduction

Bangla, (ethnonym: Bangla; exonym: Bengali), the 4th most widely spoken language in the world, is extremely productive in terms of its morphology. The Bangla lexicon has a very large number of compound words, i.e. words that have more than one root-word, which can be created from almost any combination of nouns, pronouns and adjectives. While there are existing efforts at building a complete morphological parser for Bangla, all of these can only handle simple words with a single root-word [3, 4]. Our effort here is to develop a morphological system which can parse compound words. The addition of inflectional suffixes to the Bangla compound word introduces ambiguity in the word grammar due to the possible non-deletion of the inflection of the constituent root words. We present a feature-unification based morphotactic structure and word grammar which can successfully parse Bangla compound words, correctly handling any such ambiguity.

2. Morphology and inflection

Morphology is the study of morphemes. For example in Bangla the word অনাধুনিকতার (“anAdUnIktAr”)[1] can be divided into the morphemes “an” (PREFIX), “AdUnIk” (ROOT), “tA” (DERIVATIONAL SUFFIX) and “r” (INFLECTIONAL SUFFIX). Bangla noun and pronoun morphology is predominantly linear, whereas verb morphology exhibits some non-linearity with the root form changing on inflection. Bangla is devoid of infixation which makes the morphotactic analysis a concatenative one [1, 5].

An inflectional suffix is a terminal affix that does not change the word-class (parts of speech) of the root during concatenation; it is added to maintain the syntactic environment of the root in Bangla. For instance, in the above example, “r” (র) is an inflectional suffix as it is grammatically required in certain syntactic environments, whereas “tA” (তা) is a derivational suffix which when added with the root “AdUnIk” (adjective), changes it to a noun.

There are two types of inflectional suffixes in Bangla.

2.1. Nominal and pronominal inflections (Taddhit Suffix)

A nominal or pronominal inflection is an affix that is added to a noun or pronoun. Example: “mAyEr”

[1] Throughout this paper we have used the Roman alphabet to represent Bangla characters. For example “আ” is “a”, “া ” is “A”, “ি ” is “I”, “ক” is “k”, “খ” is “K”, “য়” is “y”, “্ ”(hasanta) is “~” etc. We have also assumed that the storage is in logical order as specified in Unicode. For example খেয়েছি is represented as KEyECI.

(মায়ের = মা + য়ের), “hAtE” (হাতে = হাত + এ) etc. A list of these inflectional suffixes that act as case markers is given below:
“e” (এ), “yE”(য়ে), “y”(য়) ;“tE”(তে), “etE”(এতে);“kE”(কে) ;“rE”(র), “erE”(এরে) ; “r”(র), “er”(এর), “yEr”(য়ের).

2.2. Verbal inflections (Krit suffix)

A verbal inflection is an affix that is added to verbal elements. Example: “krtE” (করতে), “krE” (করে) etc. Here are some verbal inflectional suffixes:
“e” (এ), “yE”(য়ে); “tE”(তে); “lE”(লে).

There can be one and only one inflectional case marker in a word that has a single root. However, a compound word may have more than one inflection in that suffixes may be attached to each of the constituents (even in this case only one suffix may be added to each of the constituents) and further inflection may be added sequentially to the compound as a whole. This will be described in more detail in the next section.

While plural and gender markers are inflectional suffixes as well, we will only consider the two types of inflections mentioned above and limit our discussion to these in terms of compound words.

3. Bangla compound word

If a word contains more than one root-words then it is called a compound word [2, 6, 7]. For example:

English: “sky-high”
Meaning: as high as the sky
Root words: sky, high
Bangla: “cAd-mUK” (চাদমুখ)
Meaning: moon (cAd)-like face (mUK).
Root words: cAd, mUk

The constituents of a compound may be joined by a hyphen (-), separated by white space or may be written together as a single word. For example:

Hyphenated compound word:
“dIn-rAt” (দিন-রাত)
Non-hyphenated compound word:
“rk~to-lAl” (রক্তলাল)

Bangla has a large number of compound words. A few examples are given below:

Noun + Noun = Noun:
মা-বাপ “mA-bAp” (Noun)
= “mA” (Noun) – “bAp”(Noun)
= mother and father

Noun + Adjective = Adjective:
রক্তলাল “rk~to-lAl”
= “rk~to” (Noun) – “lAl” (Adjective)
=blood-red

Adjective +Adjective = Adjective:
তিক্তমধুর “tik~t-mDur”
= “tikt~t”(Adjective)-“mDur”(Adjective)
= bitter-sweet

A compound word can have more than two root-words:

জল-স্থল-আকাশ-যুদ্ধ “jl-s~Tl-akAS-JudDo” (water-land-sky-war).

4. Finite state morphological parsing

We have used a finite state morphological parser based on Kimmo Koskenniemi’s Two Level Morphology [8-10].

There are 3 components of this parsing system:

4.1. Lexicon and morphotactics

Morphotactics delineates the morphological divisions of a word, given the lexicon and the Finite State (FS), explaining the sequence in which one class of morphemes follows another class inside a word.

For example, the Figure 1 represents a Finite State Machine (FSM) for Bangla:

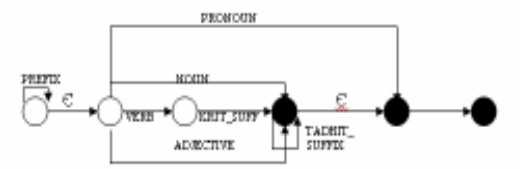


Figure 1: Finite state machine for Bangla words. (NOUN, ADJECTIVES etc. are lexical classes)

Hence, according to the above FSM, we get the following morphological divisions for the word “anAdUniktAr” (অনাধুনিকতার)
= an (PREFIX) + adUnik (ADJECTIVE) + tA (NOMINAL-SUFFIX) + r (INFL)
(Equation 1)

4.2. Morphophonology

Morphophonology studies how phonological factors affect the shape of morphemes and correspondingly how morphological factors affect the shape of phonemes.

Morphophonology will not be discussed in this paper.

4.3. Word-grammar component

This component lists the morphological constraints and tells us which lexical class collocates with which other lexical class. Given a proper word-grammar and feature-unification rules, it uses a chart parser to give us a parse tree [11, 12]. For example the lexical class INFL in Bangla is added with only nouns and pronouns. Therefore, we can try the following word-grammar rule:

Word = Stem INFL
<Stem.pos = n> or <Stem.pos = p>
Word= Stem
Stem= PREFIX Stem
Stem=Stem TADHIT_SUFFIX
Stem=NOUN
Stem=ADJECTIVE
Stem=PRONOUN
Stem=VERB_ROOT KRIT_SUFFIX
//where pos=Feature variable saving parts-of-speech
//and n, p are features denoting noun and pronoun.

When the morphological divisions in Equation 1 are given to the above word grammar we get the following parse tree:

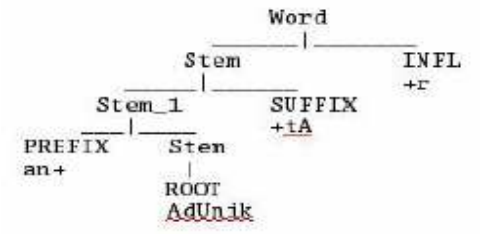


Figure 2: Parse tree for "anAdUniktAr" (অনাধুনিকতার)

5. Morphological parsing of Bangla compound word

A compound word is formed by joining two or more root-words by hyphens (-) or Null (""). Normally when two root-words join together the inflectional suffix of the first root-word may be deleted in the resultant com-pound word. For example, the compound word "mAmA-bArI" (মামা-বাড়ি) is actually the noun phrase "mAmAr bAri" (মামার-বাড়ি) where "r" is the inflectional suffix (genitive marker) for the root-word "mAmA". This "r" is deleted when the compound word is formed. This is called inflection deletion in compound words. So, when an inflectional suffix is found at the end of a compound word, it is presumed to be the inflectional suffix of the whole compound, and not of the final root-word. Hence, accordingly the correct parse tree for the word "mAmA-bArItE" (মামা-বাড়িতে) should be Figure 3(a) whereas Figure 3(b) is deemed incorrect. [2, 6]

Figure 3: (a) correct parse tree for "mAmA-bArItE" (b) incorrect parse tree for "mAmA-bArItE"

If all compound words followed the above form of inflection deletion then we could conclude that whole compounds only get inflected terminally. Based on that, we modify the FSM and word-grammar for Bangla compound as shown in Figure 4 [9]:

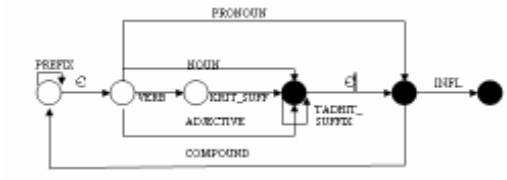


Figure 4: FSM for a compound word (version 1)

Word Grammar:

Word=Word INFL
Word=Word COMPOUND Word //here COMPOUND={'-', 0}
Word= Stem
Word= PREFIX Stem
Stem=Stem TADHIT_SUFFIX
Stem=NOUN
Stem=ADJECTIVE
Stem=PRONOUN
Stem=VERB_ROOT KRIT_SUFFIX

(Grammar 1)

5.1. Non-deletion of inflectional suffix

The above hypothesis of just one inflectional suffix per compound word is not always true. There are many compound words whose individual stems retain their own inflectional suffixes. In other words, inflection deletion as specified above does not hold true for many compound words [2, 6]. For example:

"GrE-bAhIrE" (ঘরে-বাহিরে)
= "Gr" + "e" – "bAhIr" + "e"
= (NOUN + INFL) – (NOUN + INFL)

In the above example the inflectional suffix "e" remains "undeleted" in the compound word. The same is true for many other compound words such as:

"mAmAr-bArI" (মামার-বাড়ি)
= "mAmA" + "r" – "bArI"

"hAtE-pAyE"(হাতে-পায়ে)
= "hAt"+ "e" – "pA" + "yE"

To account for these inflectional suffixes in terms of compound words we change the FSM and grammar as shown in Figure 5.

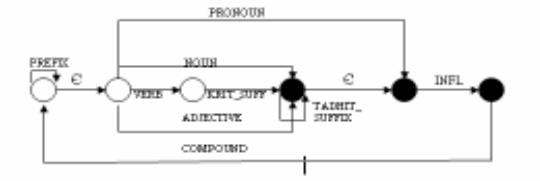


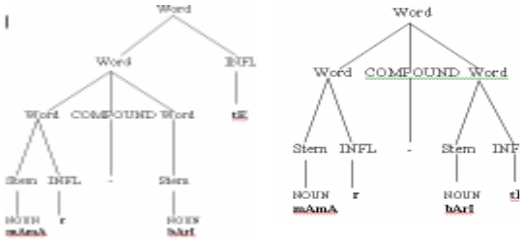
Figure 5: FSM for compound word (version 2)

New Word Grammar:
Word=Word INFL
Word=Word COMPOUND Word
Word= Stem
Word=Stem INFL // new addition to the previous grammar
Word=PREFIX Stem
Stem=Stem TADHIT_SUFFIX
Stem=NOUN
Stem=ADJECTIVE
Stem=PRONOUN
Stem=VERB_ROOT KRIT_SUFFIX

(Grammar 2)

5.2. Ambiguous word-grammar

The grammar shown above (grammar 2) turns out to be an ambiguous one as it gives two different parse trees for the same compound word. As a result, we cannot recognize whether the final inflectional suffix of a compound word is the inflectional property of the final constituent (last root-word) or of the compound as a whole. For example, the parse tree given by the above grammar for the word "mAmAr-bArItE" (মামার-বাড়িতে) is shown in Figure 6.



(a) correct (b) incorrect

Figure 6: Two parse trees for the word "mAmAr-bArItE" (মামার-বাড়িতে)

Here we cannot determine whether the final inflectional suffix "tE" is the inflectional property of the compound word ("mAmAr-bArI") as shown in figure 6(a) or of the last root-word ("bArI") as shown in figure 6(b). But, according to Bangla grammar, the parse tree in Figure 6(a) is the correct one, not the one in Figure 6(b).

Similarly, for the word "GrE-bAhIrE" (ঘরে-বাহিরে), the parse tree shown in Figure 7(b) is the correct one, not the one in Figure 7(a).

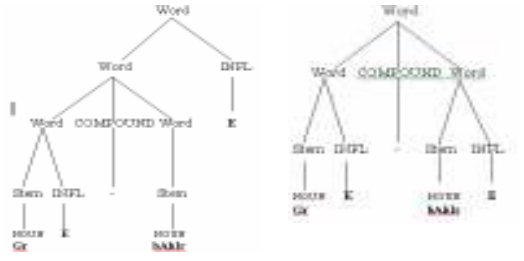


Figure 7: Two parse trees for the word "GrE-bAhIrE" (ঘরে-বাহিরে)

5.3 Ambiguity resolution

To resolve the ambiguities stated above, we define two new features and carry out feature unification which ensures that there is just one parse tree for every com-pound word. The two new features are derived in the following way:

We classify nominal and pronominal inflections into 5 categories and define the feature variable inflType to denote the following inflectional categories:

inflType= Ie ["e" (এ), "yE"(য়ে), "y"(য়)]
inflType= It ["tE"(তে), "etE"(এতে)]
inflType= Ik ["kE"(¬কে)]
inflType= Ire ["rE"(¬র), "erE"(এরে)]
inflType= Ir ["r"(র), "er"(এর), "yEr"(¬য়ের)]

There are maximally 3 types of inflectional suffixes in each category. These 3 types are actually added as an inflectional suffix with 3 different types of nouns/pronouns. For example, the "e" inflection is added with nouns whose last character is a consonant (e.g. "hAt", হাত); the "yE" inflection is added with nouns whose last character is a vowel and has 2 characters (e.g. "pA", পা); the "y" inflection is added with nouns whose last character is a vowel and has more than 2 characters (e.g. "kAdA", কাদা) [2, 6].

So, we classify every noun/pronoun in the lexicon into 3 categories and define feature variable rootInflType to store the noun/pronoun categories.

rootInflType=Nc [noun whose last char is a consonant, e.g., "hAt" (হাত)]
rootInflType=Nv [noun whose last char is a vowel and has two characters, e.g., "pA" (পা)]
rootInflType=Nv2 [noun whose last char is a vowel and has more than two characters, e.g., "kAdA"(কাদা)]

We modify the lexicon to add the two features in the following way:

Lexicon: NOUN
(1){ hAt (হাত) \feature Nc}
(2){ "pA" (পা) \feature Nv}

Lexicon: INFL
(3){ "e" \feature Ie, Nc}
(4){ "yE" \feature Ie, Nv }
(5){ "y" \feature Ie, Nv2 }
(6){ "kE" \feature Ik, {Nc | Nv | Nv2}
//as "kE" can be added with any Noun categories.}

We then classify the compound words with inflectional suffixes into 4 different categories: [2,6]

5.3.1. Category 1. In this category, if there are two root-words (both nouns) in the compound word, then there are two inflectional suffixes, and the category of inflectional suffix of the first root-word is the same as the category of inflectional suffix of the second root-word, and that category is inflType=Ie as described above. For example:

"GrE-bAhIrE" (ঘরে-বাহিরে)= ("Gr" + "e") – ("bAhIr" + "e")
"jAlE-kAdAy" (জলে-কাদায়) = ("jAl" + "e") – ("kAdA" + "y")
[Here "e" and "y" are inflection of category Ie]

The grammar incorporating the above feature constraint is as follows:

Word=Word_1 COMPOUND Word_2
<Word_1 inflType> = < Word_2 inflType> = Ie
<Word_1 pos> = < Word_2 pos > = N
//here pos=parts-of-speech of a word and N means category Noun

5.3.2. Category 2. In this category, if there are two root-words (both pro-noun), then there are two inflectional suffixes, and the category of inflectional suffix of the first root-word is the same as the category of inflectional suffix of the second root-word. For example:

"tOmAr-amAr" (তোমার-আমার) = ("tOmA" + "r") – ("amA" + "r")
"tOmAtE-amAtE"(¬ তোমার-আমাতে) = ("tOmA" + "tE") – ("amA" + "tE")

The grammar incorporating the above feature constraint is as follows:

Word=Word_1 COMPOUND Word_2
<Word_1 inflType> = < Word_2 inflType>
<Word_1 pos> = < Word_2 pos > = Pr
//here pos=parts-of-speech of a word.
//Pr means category Pronoun

5.3.3. Category 3. In this category, the inflectional suffix of the first root word is retained but the second root-word has no inflectional suffix, and the second root-word is an adjective. For example:

"hAtE-kATA" (হাতে-কাটা) = ("hAt" + "E") – ("kATA")

The grammar incorporating the above feature constraint is as follows:

Word=Word_1 COMPOUND Word_2
<Word_1 inflType> != 0
<Word_2 inflType> = 0
<Word_2 pos> =Adj
//here pos=parts-of-speech of a word.
//Adj means category Adjective

5.3.4. Category 4. In this category, the inflectional suffix of the first root-word is of category Ir (as defined above) and the inflectional suffix of the second root-word is not present. For example:

“mAmAr-bArI” (মামার-বাড়ি) =(“mAmA”+“r”)–(“bArI”)
“mAmAr-kArA” (মামার-করা) =(“mAmA”+“r”)–(“kArA”)

The grammar incorporating the above feature constraint is as follows:

Word=Word_1 COMPOUND Word_2
<Word_1 inflType> = Ir
<Word_2 inflType> = 0

Now we consider the words “mAmAr-bArItE” (মামার-বাড়িতে) and “GrE-bAhIrE” (ঘরে-বাহিরে) which resulted in two parse trees with the previous ambiguous grammar (Figures 6 and 7).

The parsing of “mAmAr-bArItE”, shown in Figure 6(a), holds because of compound word rule category 4.

“mAmAr-bArItE”= “mAmA”+“r”–“bArI”+ “tE”
= ((“mAmA”+“r”)–“bArI”) + “tE”

The parsing of “mAmAr-bArItE”, shown in Figure 6(b), does not hold because of compound word rule category 1.

“mAmAr-bArItE”=“mAmA”+“r”–“bArI” + “tE”
= (“mAmA” + “r”) – (“bArI” + “tE”)
[“r” and “tE” are of different inflType]

The parsing of “GrE-bAhIrE”, shown in Figure 7(a), does not hold because of compound word rule category 3.

“GrE-bAhIrE”= “Gr” + “e” – “bAhIr” + “e”
= ((“Gr” + “e”) – “bAhIr”) + “e”
[“bAhIr” is not adjective]
The parsing of “GrE-bAhIrE”, shown in Figure 7(b), holds because of compound word rule category 1.

“GrE-bAhIrE”= “Gr” + “e” – “bAhIr” + “e”
= (“Gr” + “e”) – (“bAhIr” + “e”)

5.4. Final grammar: (in PC-KIMMO format)

```
RULE
Word_1 -> Word_2 INFL
          <Word_2 cmpCheck> = +
<Word_2 rootInflType>= <INFL rootInflType>
<Word_1 inflType> = <INFL inflType>
          <Word_1 cmpCheck> = +
          <Word_1 pos> = <Word_2 pos>
          {
            <Word_2 pos> = NN
           /
            <Word_2 pos> = PRO
}
RULE
Word_1 -> Word_2 COMPOUND Word_3
<Word_1 pos> = <Word_3 pos>
<Word_1 cmpCheck> = +
<Word_1 rootInflType>= <Word_3 rootInflType>
{
          ;category1
 <Word_2 inflType> = <Word_3 inflType>
 <Word_2 inflType> = IE
 <Word_2 pos> = <Word_3 pos>
 <Word_2 pos> = NN
 <Word_1 inflType> = IE
/
;category2
<Word_2 inflType> = <Word_3 inflType>
<Word_2 pos> = <Word_3 pos>
<Word_2 pos> = PRO
<Word_1 inflType> = <Word_2 inflType>
/
;category3
<Word_2 inflType> = !ZR
<Word_3 inflType> = ZR
<Word_3 pos> = ADJ
<Word_1 inflType> = ZR
/
;category4
<Word_2 inflType> = IR
<Word_3 inflType> = ZR
<Word_1 inflType> = ZR
/
;category5 (no inflections)
<Word_2 inflType> = <Word_3 inflType>
<Word_2 inflType> = ZR
<Word_1 inflType> = ZR
          }
RULE
Word -> Stem
<Word pos> = <Stem pos>
<Word inflType> = ZR
<Word cmpCheck> = -
<Word rootInflType>= <Stem rootInflType>
RULE
```

```
Word -> Stem INFL
<Stem rootInflType>= <INFL rootInflType>
<Word pos> = <Stem pos>
<Word inflType> = <INFL inflType>
<Word cmpCheck> = -
{
   <Stem pos> = NN
    /
   <Stem pos> = PRO
   }
RULE
Stem_1 -> NPREFIX Stem_2
  <Stem_1 pos> = <Stem_2 pos>
 <Stem_1 rootInflType> = <Stem_2 rootInflType>
RULE
Stem_1 -> Stem_2 TADHITSUFFIX
  <Stem_1 pos> = <Stem_2 pos>
  <Stem_1 rootInflType> = <TADHITSUFFIX rootIn-
flType>
RULE
Stem -> NOUNROOT
  <Stem pos> = <NOUNROOT pos>
  <Stem rootInflType> = <NOUNROOT rootInflType>
RULE
Stem -> ADJROOT
  <Stem pos> = <ADJROOT pos>
RULE
Stem -> PRONOUNROOT
  <Stem pos> = <PRONOUNROOT pos>
  <Stem rootInflType> = <PRONOUNROOT rootIn-flType>

END
```

Note: Here we have shown only those feature unifications that are associated with ambiguity resolution of compound words.

6. Implementation

We have implemented the above morphological analyzer for compound words in PC-KIMMO version 2, which is based on two-level morphology [4,13,14]. We have primarily used compound-words found in Bangla grammar books [2,6,7] to produce our test cases and obtained 100% correct results. The word-grammar we proposed here is a generalized one and incorporates almost all possible compound word combinations. Therefore, it should work for any given inflecting com-pound word whether it has been tested or not. Here is the PC-KIMMO output for the word “GrE-bAhIrE” for which we previously got 2 parse trees (Fig 7).

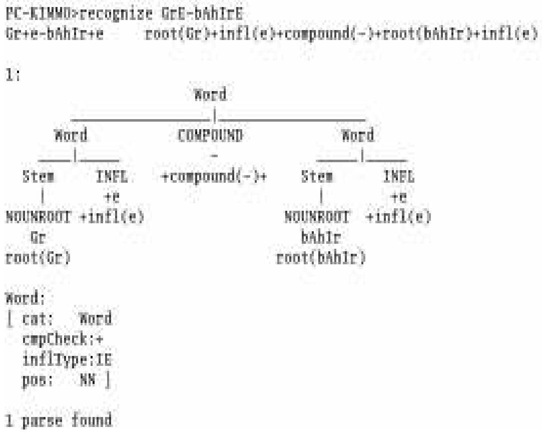


Figure 8: Disambiguated output of the word “GrE-bAhIrE” (ঘরে-বাহিরে).

7. Conclusion

We have presented a morphological parser for Bangla compound words, which handles the ambiguities resulting from inflection deletion, or the lack thereof. Combined with the morphological rules for simple words found in the literature [2,6,7], we have presented a word-grammar which can successfully parse all inflected variations of compound words. We have implemented the word grammar in PC-KIMMO, and tested it on a large number of commonly found compound words with very good results. Hopefully our effort here will help in implementing a complete morphological parser for Bangla in future.

8. References

[1] B. Comrie, ed., “The World’s Major Languages”, Oxford University, New York, 1987.

[2] S.K. Chottapaday, “Vasha Prokash Bangla Bakaran”, May 1989.

[3] S. Bhattacharya, M. Choudhury, S. Sarkar and A. Basu, “Inflectional Morphology Synthesis for Bengali Noun, Pronoun and Verb Systems”, *Proc. of the National Conference on Computer Processing of Bangla (NCCPB 05)*, pp. 34 - 43, Dhaka, Bangladesh, Mar 2005.

[4] S. Dasgupta and M. Khan, “Morphological Parsing of Bangla Words using PC-KIMMO”, *Proc. ICCIT 2004*, Dhaka, Bangladesh, December, 2004.

[5] P. Sengupta and B.B. Chaudhuri, “Morphological processing of Indian languages for lexical interaction with application to spelling error correction”, *Sadhana, Vol. 21, Part. 3*, pp. 363-380, 1996.

[6] P. Sarkar, “Bangla Rupthatter Bhumica”, 1997.

[7] S. Rameshar, “Sadaran Vhasabiggan and Bangla Vhasa”, Ananda Press, 1996.

[8] K. Koskenniemi, “Two-level morphology: a general computational model for word-form recognition and production.”, Publication No. 11. Helsinki: University of Helsinki Department of General Linguistics, 1983.

[9] A. Spencer and A.M. Zwicky, “The Handbook of Morphology”, Blackwell Publishers, 2001.

[10] D. Jurafsky and J.H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, Prentice Hall, 2000.

[11] S.M. Shieber, “An introduction to unification-based approaches to grammar.”, *CSLI Lecture Notes No. 4*. Stanford, CA: Center for the Study of Language and Information, 1986.

[12] E.L. Antworth, “Morphological Parsing with Unification-based Word Grammar.”, A paper presented at North Texas Natural Language Processing Workshop, May 23, 1994.

[13] E.L. Antworth, “PC-KIMMO: A two-level processor for morphological analysis”, *Occasional Publications in Academic Computing No. 16*. Dallas, Texas: Summer Institute of Linguistics, 1990.

[14] PC-KIMMO, available at www.sil.org/pckimmo.