

Morphological Analysis of Inflectional Compound Words in Bangla

Abstract.

The addition of inflectional suffixes in Bangla compound words is fairly complex. Normally, when two root words are joined, the corresponding inflectional suffix of each root word is deleted from the final compound word. In Bangla however, the compound word's individual root words may retain their inflectional suffixes even in the final compound word. This non-deletion of inflection creates an ambiguity as the context-free word grammar cannot recognize whether the final inflectional suffix of a compound word is an inflectional property of the last constituent root-word or of the final compound word. We use a feature unification based morphological parser which can successfully and efficiently parse compound words having inflectional suffixes and at the same time resolve such ambiguities.

1. Introduction

Bangla, also known as Bengali, is the 4th most widely spoken language with more than 200 million speakers, most of whom live in Bangladesh and in the Indian state of West Bengal. Modern Bangla morphology is very productive, especially for verbs, with each root verb taking on 168 different forms [2]. Bangla lexicon also has a very large number of compound words, i.e., words that have more than one root, 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 [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 words 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 division of a word into smaller sub-parts, or morphemes. For example, the English word “unforgettable” is divided into 3 morphemes, i.e. “un”, “forget”, and “able”. Similarly, the Bangla word অনাধুনিকতার (“anAdUnIktAr”)¹ is divided into “an” (PREFIX), “AdUnIk” (ROOT), “tA” (SUFFIX) and “r” (INFLECTION). Bangla noun and pronoun morphologies is linear, whereas the verb morphology exhibits some non-linearity, where the root form changes on inflection. It does not have infixation like semitic language, which makes the morphotactic analysis a concatenative one [1,5].

Inflectional suffix is that one which does not change the meaning or parts-of- speech of the root during concatenation; it is added just to maintain structure of a sentence in Bangla. For example “r” (র) is an inflec-

¹ Through out this paper we have used English 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, such as specified in Unicode. For example খেয়েছি is represented as KEyECI.

tional suffix in the above example whereas “tA” (তা) is not inflectional suffix as “tA”, when added with root “AdUnlk” (adjective) makes it noun (change of parts-of-speech).

There are two types of inflectional suffixes in Bangla.

(1) Noun and Pronominal Inflections:

Here inflectional suffix is added with Noun or Pronoun stem. Example: “mAyEr” (মায়ের), “tAdEr” (তাদের) etc. This is the complete list of these inflectional suffixes:

“e” (এ), “yE” (য়), “y” (য়),
“tE” (ত), “etE” (এত),
“kE” (কে),
“rE” (র), “erE” (এর),
“r” (র), “er” (এর), “yEr” (য়ের).

(2) Verbal Inflections:

Here inflectional suffix is added with Verbal root. Example: “krtE”, “krE” etc. Here are some verbal inflectional suffixes:

“e” (এ), “yE” (য়)
“tE” (ত),
“IE” (লি)

There can be one and only one Inflectional suffix in a word having a single root. However, a compound word can have more than one inflectional suffixes which will be described in more detail in the next section. While some linguists consider plural and gender marker suffixes as inflectional variations, we only consider the two types of inflections mentioned above as we limit our discussion to compound words only.

3. Bangla Compound Word

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

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

Compound word’s root words can be joined by a hyphen (-) or nothing. For example:

Hyphenated compound word: “dIn-rAt” (দিন-রাত)
Non-hyphenated compound word: “bIdEsAgt” (বিদ্যাগত)

Bangla has a large number of compound words. Almost all combination of noun, pronoun and adjectives are added with each other. Here are a few examples of compound word:

Noun + Noun = Noun:

মা-বাপ “mA-bAp” (Noun)
= “mA” (Noun) – “bAp” (Noun)
= mother and father (In English)

Noun + Adjective = Adjective:

হাত-গড়া “hAtE-gRA” (Adjective)
= “hAtE” (Noun) – “gRA” (Adjective)

=hand-made (In English)
 Adjective + Noun = Noun:
 লাল-টুপি “lAl-tUpI” (Noun)
 =“lAl” (Adjective) – “tUpI”(Noun)
 = red-cap (In English)

Adjective + Adjective = Adjective:
 দীর্ঘ-স্থায়ী “dIr~G-s~TAyI”
 =“dIr~G” (Adjective) – “s~TAyI” (Adjective)
 = long-lasting (In English)

Pronoun + Noun = Noun:
 আমার-দেশ
 =“amAr-dES”
 =“amAr” (Pronoun) – “dES” (Noun)
 = my-country (In English)

A compound word can have more than two roots:
 “sAt-rAjAr-Dn” (সাত-রাজার-ধন, Seven-king’s-property)

4. Finite State Morphological Parsing

We use a finite state morphological parser based on Kimmo Koskenniemi’s two level morphology [8-10]. There are 3 components of this parsing system:

(1) Lexicon and Morphotactics

It gives the morphological divisions of a certain word given the lexicon and the Finite State (FS) expressing the morphotactics. For example if the following diagram is the FST for Bangla:

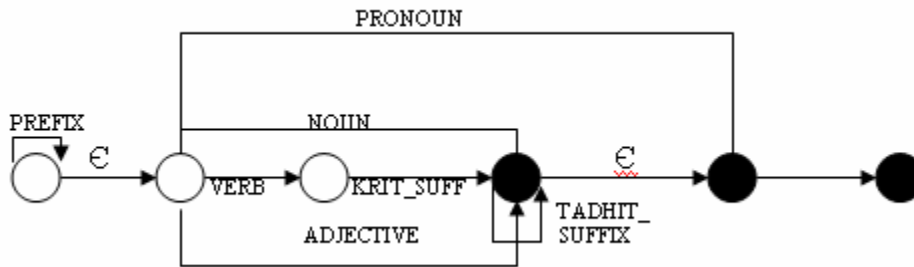


Fig. 1. Finite State for Bangla word. (Here NOUN, ADJECTIVES etc. are lexicon class.)

then we will get the following morphological divisions for the word

“anAdUniktAr” (অন্যুনিকতার)
 = an (PREFIX) + adUnik (ADJECTIVE) + tA (TADHIT-SUFFIX) + r (INFL)
(1)

(2) Morphophonology

Morphophonology handles the combinatory phonic modifications of the morphemes when they are combined. We will not discuss morphophonology in this paper.

(3) Word-grammar component

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

```
Word = Stem INFL
      <Stem.pos = n> or <Stem.pos = p>
Word= Stem
Stem= PREFIX Stem
Stem=Stem TADHIT_SUFFIX
Stem=NOUN
Stem=ADJECTIVE
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:

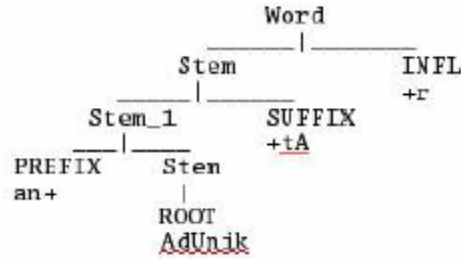


Fig. 2. Parse tree for “anAdUniktAr” (অনাধুনিকতার)

5. Morphological Parsing of Bangla Compound Word

Compound word is formed by joining two or more stems by hyphens (-) or *Null* (“”). Normally when two stems join together the inflectional suffix of the first stem remains unspecified in the resulting compound word. For example, the compound word “mAmA-bArI” (মামা-বাড়ি) is actually the word “mAmAr bArI” (মামার বাড়ি) where “r” is the inflectional suffix for stem “mAmA”. That “r” is deleted when compound word is formed. This is called *inflection deletion* in compound words [2,6]. So, when a inflectional suffix is found at the end of a compound word, it is presumed to be the inflectional suffix of the compound word, and not the inflectional suffix of the last stem. For example, the parse tree for the word “mAmA-bArItE”(মামা-বাড়িত) should be like Figure 3(a) not Figure 3(b).

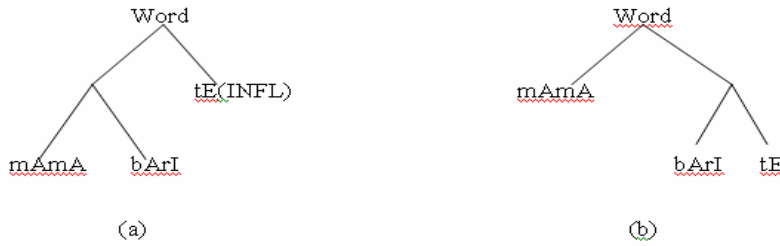


Fig. 3. (a) correct parse tree for “mAmA-bArItE” (b) incorrect parse tree for “mAmA-bArItE”

If all compound words followed the above inflection deletion then we should conclude that *there is just one inflectional suffix for every compound word*. Based on that, we modify the FS and word grammar for Bangla compound as follows[9]:

Finte State:

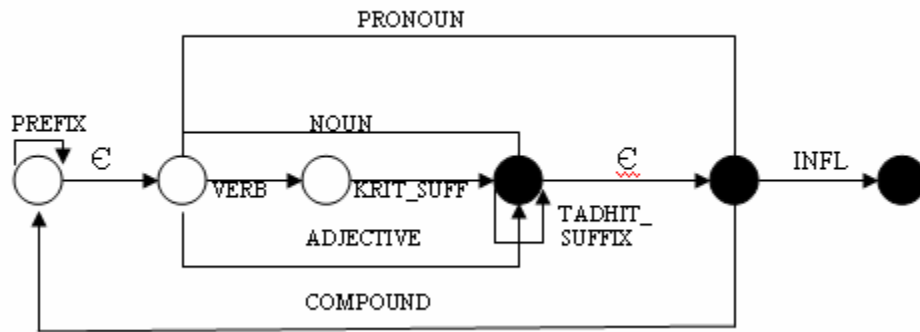


Fig. 4. Finite State for Compound word (version 1)

Word Grammar:

- Word=Word INFL
 - Word=Word COMPOUND Word // here COMPOUND={'-', 0}
 - Word= Stem
 - Word=Stem PREFIX
 - Stem=Stem TADHIT_SUFFIX
 - Stem=NOUN
 - Stem=ADJECTIVE
 - 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 for many compound words [2,6]. For example:

$$\begin{aligned}
 &\text{“GrE-bAhIrE” (ঘর-বাহির)} \\
 &= \text{“Gr”} + \text{“e”} - \text{“bAhIr”} + \text{“e”} \\
 &= (\text{NOUN} + \text{INFL}) - (\text{NOUN} + \text{INFL})
 \end{aligned}$$

In the above example “e” inflectional suffix remain “un-deleted” in the compound word. Same is true for many other compound words like the following:

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

“hAtE-pAyE”(হাত-পায়)
 = “hAt”+ “e” – “pA” + “yE”

To accommodate the inflectional suffixes in the compound word we change the FST and grammar:

New Finite State:

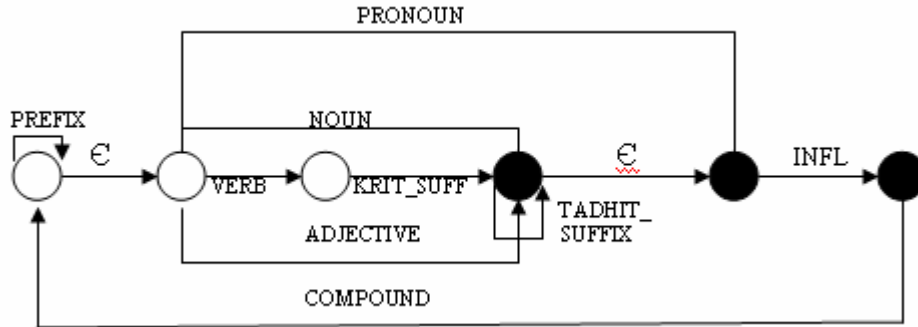


Fig. 5. Finite State for Compound word (version 2)

New Grammar:

- Word=Word INFL
 - Word=Word COMPOUND Word
 - Word= Stem
 - Word= Stem INFL //This is the new addition to the previous grammar
 - Word=Stem PREFIX
 - Stem=Stem TADHIT_SUFFIX
 - Stem=NOUN
 - Stem=ADJECTIVE
 - Stem=VERB_ROOT KRIT_SUFFIX
-grammar (2)

5.2. Ambiguous Grammar

But 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. The result is that we cannot recognize whether the final inflectional suffix of a compound word is inflectional property of last root-word or the final compound word. For example, the parse tree given by the above grammar for the word “mAmAr-bArItE” is as follows:

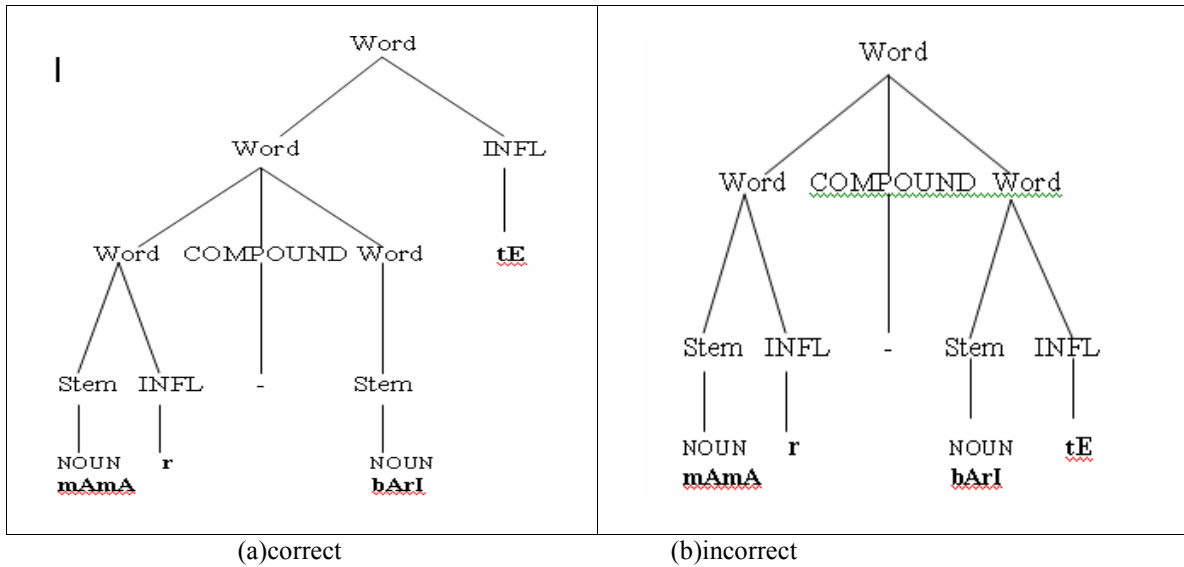


Fig. 6. Two parse trees for the word “mAmAr-bArItE” (মামার-বাড়িত)

Here we cannot determine whether final inflectional suffix “tE” is the inflectional property of compound word (“mAmAr-bArI”) as shown in figure 6(a) or 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” (ঘর-বাহির), the parse tree shown in Figure 7(b) is the correct one, not the one in Figure 7(a).

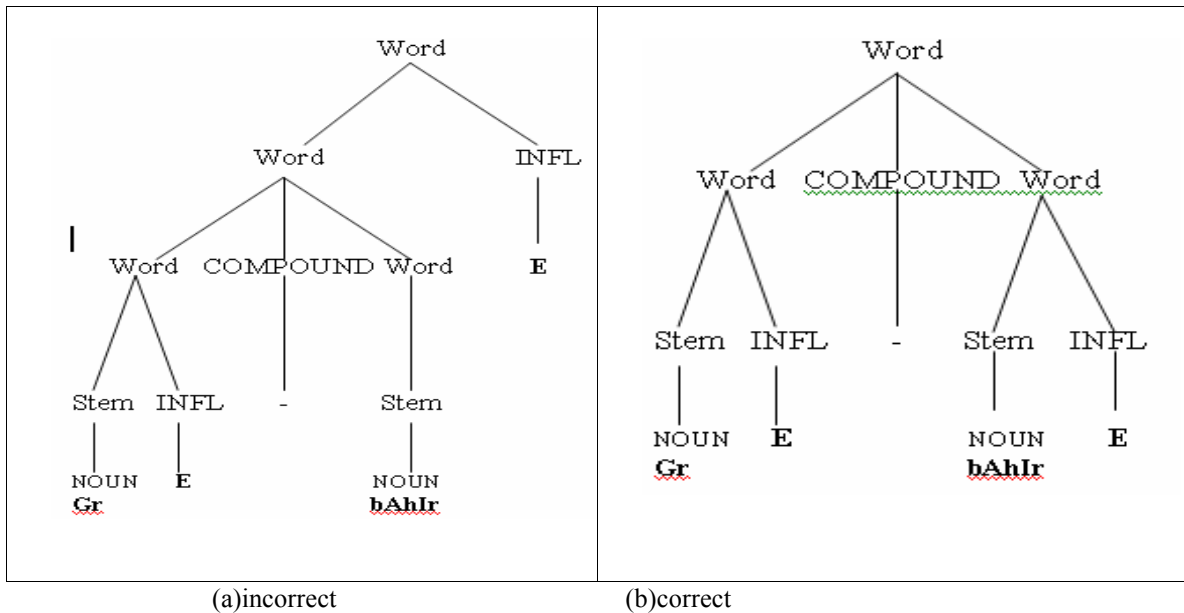


Fig. 7. Two parse trees for the word “GrE-bAhIrE” (ঘর-বাহির)

5.3. Ambiguity Resolution

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

We classify noun and pronominal inflections into 5 categories and define feature variable `inflType` to denote the inflectional categories.

<code>inflType=Ie</code>	[“e” (এ), “yE”(য়), “y”(য়)]
<code>inflType=It</code>	[“tE”(ত), “etE”(এত)]
<code>inflType=Ik</code>	[“kE”(ক)]
<code>inflType=Ire</code>	[“rE”(র), “erE”(এর)]
<code>inflType=Ir</code>	[“r”(র), “er”(এর), “yEr”(য়র)]

There are at most 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” inflectional suffix is added with nouns whose last character is a consonant (e.g. “hAt”, হাত); the “yE” inflectional suffix is added with nouns whose last character is a vowel and has 2 characters (e.g. “pA”, পি); the “y” inflectional suffix is added with nouns whose last character is a vowel and has more than 2 characters (e.g. “kAdA”, কাঁদা) [2,6].

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

<code>rootType=Nc</code>	[noun whose last char is a consonant, e.g., “hAt”]
<code>rootType=Nv</code>	[noun whose last char is a vowel and has two characters, e.g., “pA”]
<code>rootType=Nv2</code>	[noun whose last char is a vowel and has more than two characters, e.g., “kAdA”]

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

Lexicon: NOUN

```
(1) {
    hAt (হাত)
    \feature   Nc
}
```

```
(2) {
    “pA” (পি)
    \feature   Nv
}
```

.....

Lexicon: INFL

```
(3) {
    “e”
    \feature   Ie, Nc
}
```

```
(4) {
    “yE”
    \feature   Ie, Nv
}
```


So, the grammar incorporating 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
```

Category 4:

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

```
"mAmAr-bArI" (মামর-বাড়ি) = ("mAmA" + "r") - ("bArI")
"mAmAr-kArA" (মামর-করা) = ("mAmA" + "r") - ("kArA")
"tOmAr-dEs" (তামর-দশ) = ("tOmA" + "r") - ("dEs")
```

So, the grammar incorporating 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" (মামর-বাড়িত) and "GrE-bAhIrE" (ঘর-বাহির) 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" )
```

Final Grammar:

```
Word=Word_1 INFL
  <Word inflType> = <INFL inflType>
```

```

Word=Word_1 COMPOUND Word_2 //category 1
  <Word_1 inflType> = < Word_2 inflType> = Ie
  <Word_1 pos> = < Word_2 pos > = N
  <Word inflType> = Ie

Word=Word_1 COMPOUND Word_2 //category 2
  <Word_1 inflType> = < Word_2 inflType>
  <Word_1 pos> = < Word_2 pos > = Pr
  <Word inflType> = <Word_1 inflType>

Word=Word_1 COMPOUND Word_2 //category 3
  <Word_1 inflType> = 0
  <Word_2 inflType> != 0
  <Word_2 pos> =Adj
  <Word inflType>=0

Word=Word_1 COMPOUND Word_2 //category 4
  <Word_1 inflType> = Ir
  <Word_2 inflType> = 0
  <Word inflType>=0

Word=Word_1 COMPOUND Word_2 //No inflections
  <Word_1 inflType> = < Word_2 inflType> = 0

Word= Stem
  <Word inflType> = 0

Word= Stem INFL
  <Stem rootType> = <INFL rootType> //check
  <Word inflType> = <INFL inflType>

Word=Stem PREFIX
Stem=Stem TADHIT_SUFFIX
Stem=NOUN
Stem=ADJECTIVE
Stem=VERB_ROOT KRIT_SUFFIX
  [Here we have shown only those feature unifications which 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 used the compound-words found from the Bangla grammar books [2,6,7] to produce our test cases and got 100% correct result. The word-grammar we proposed here is generalized one and incorporates almost all possible compound word combination. Therefore it will work for any given inflectional compound word whether it is in our test cases or not.

7. Conclusion

We present a morphological parser for Bangla compound words, handling 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 inflectional 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 implementing a complete-morphological parser for Bangla in future.

8. References

- [1] Comrie, Bernard, ed., "The World's Major Languages", Oxford University, New York, 1987.
- [2] Suniti Kumar Chottopaday, "Vasha Prokash Bangla Bakaran", May 1989.
- [3] Samit Bhattacharya, Monojit Choudhury, Sudeshna Sarkar and Anupam 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 Mumit 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] Pabitra Sarkar, "Bangla Rupthatter Bhumica", 1997.
- [7] Rameshar S., "Sadaran Vhasabiggan and bangla vhasa", Ananda press, 1996.
- [8] Koskenniemi, Kimmo., "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] Andrew Spencer and Arnold M. Zwicky, "The Handbook of Morphology", Blackwell Publishers, 2001.
- [10] Daniel Jurafsky and James H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Prentice Hall, (2000).
- [11] Shieber, Stuart M. 1986. "An introduction to unification-based approaches to grammar.", CSLI Lecture Notes No. 4. Stanford, CA: Center for the Study of Language and Information.
- [12] Antworth, Evan L. "Morphological Parsing with Unification-based Word Grammar.", A paper presented at North Texas Natural Language Processing Workshoup (May 23, 1994).
- [13] Antworth, Evan L. "PC-KIMMO: a two-level processor for morphological analysis.", Occasional Publications in Academic Computing No. 16. Dallas, TX: Summer Institute of Linguistics (1990).
- [14] PC-KIMMO, available at www.sil.org/pckimmo.