wjert, 2017, Vol. 3, Issue 2, 147 -167 *Original Article* ISSN 2454-695X

World Journal of Engineering Research and Technology

WJERT

www.wjert.org SJIF Impact Factor: 4.326

EXTRCTION OF BANGLA ROOT VERBS FROM SENTENCES

Nazmin Akter* and Fabiha Ferdous Chowdhury

Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh.

Article Received on 01/02/2017 Article Revised on 22/02/2017 Article Accepted on 14/03/2017

***Corresponding Author**
Nazmin Akter
Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh.

ABSTRACT

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. NLP is a component of artificial Suffix, intelligence (AI). Morphology is the basement of POS Tagging which plays a significant role in NLP applications such as spell checker, text to speech, English translation etc. Among the all parts of speech I have selected verb tagging from a sentence as the research topics .Our main goal is to tag the verbs along with root and discover verbal diversity of Bengali language. We have considered the limitation and the advantages of suffixes finding the status of the verb. We have found some similar structure on verbal root and we emphasized on it.

KEYWORDS: Natural language processing (NLP).

1. INTRODUCTION

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. NLP is a component of artificial intelligence (AI).

The development of NLP applications is challenging because computers traditionally require humans to ―speak‖ to them in a programming language that is precise, unambiguous and highly structured or, perhaps through a limited number of clearly-enunciated voice commands. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

Current approaches to NLP are based on machine learning, a type of artificial intelligence that examines and uses patterns in data to improve a program's own understanding. Most of the research being done on natural language processing revolves around search, especially enterprise search.

1.1 Application of NLP

- Text-Processing
- Machine Translation
- Speech Recognition
- Speech Analysis
- Document Retrieval
- Question-Answering
- Stemmer
- Sentence Recognition
- Machine Linking
- Entity resolution
- Gist Summarization
- Information Extraction

1.2 Ambition of NLP

In the field of NLP, morphological analysis and POS tagging approach of a language is very much fundamental and significant. Due to structural and morphological complexity Bangla has not met enough research yet.

But it is badly needed to have some to improve the accessibility of Bangla in various fields of internet. Again morphological analysis can be called as the basement of all NLP efforts. As Morphological analysis and POS tagging are very much related to each other, we can Focus on different aspects where both are needed:

- Search Engine
- Spell Checker
- Text to speech
- To English Translation
- Web related task
- NLP

2. BACKGROUND

2.1 Morphology

The term morphology is Greek and is a makeup of Morph- meaning 'shape, form' and Ology which means 'the study of something'. Morphology as a sub-discipline of linguistics was named for the first time in 1859 by the German linguist August Schleicher who used the term for the study of the form of words (2).Now we can summarize the morphology in this words, ―Morphology is the part of linguistics that deals with the study of words, their internal structure and partially their meanings & it refers to identification of a word stem from a full word form &a morpheme in morphology is the smallest units that carry meaning and fulfill some grammatical function. ―In linguistics, morphology is the identification, analysis, and description of the structure of a given language's morphemes and other linguistic units, such as root words, affixes, parts of speech, intonations and stresses, or implied context. ―In Bangla morphology is one of four basic portion of linguistics. Basically it deals with the diversity along with morphemes and inflections. In the journey of thousand years Bangla has been affected or provoked with various inflections.

2.2 Word

Before continuing, as we are mentioning ―word‖, we have to define the actual meaning of word which are being indicated here n texts they are particularly easy to spot since they are divided by white spaces. But how do we identify words in speech? A reliable definition of words is that they are the smallest independent units of language. They are independent in that they do not depend on other words which mean that they can be separated from other units and can change position.

2.3 Morphemes

Although words are the smallest independent units of language, they have an internal structure and are built up by even smaller pieces. There are simple words that don’t have an internal structure and only consist of one piece. There is no way we can divide into smaller parts that carry meaning or function. Complex words however, do have an internal structure and consist of two or more pieces.

2.4 POS Tagging

POS tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition, as well as its context. Grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

2.5 Word Stemming

Word stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form - generally a written word form.

Example: Bariti= Bari+ ti।Here, “Bari”is a root and–ti”is a suffix.

3. CONTENTS OF BANGLA GRAMMAR

3.1 Parts of Speech

In a sentence words are divided into different classes or kinds because of their functions and usage. And it is described by the term of parts of speech.

Dislike English, Bengali parts of speech is categorized into five portion. They are:

- Noun
- Adjective
- Verb
- Pronoun
- Preposition

3.2 Nouns

Nouns are also inflected for case, including nominative, objective, genitive (possessive), and locative. The case marking pattern for each noun being inflected depends on the noun's degree of animacy.

Example: Rahim, Gari, Karim

Two types of Inflection of Noun:

- Singular Noun Inflection
- Plural Noun Inflection

3.3 Adjective Inflection

In Bengali adjective are rarely inflected for the gender, number or person of the nouns or pronouns they qualify. A few adjectives can be inflected to denote the female gender.

3.4 Verb Inflection

Bangla verbs are either finite or non-finite. Non-finite verbs are not inflected for tense or person. Finite verbs are fully inflected for person (first, second, third), tense (present, past, future), aspect (simple, prefect, progressive) and honor (intimate, familiar and formal) but not for number. Each inflection is indicated by a suffix. Additionally, the suffixes indicating tense and aspect can be replaced by conditional, imperative and other special inflection. The number of inflection on many verb roots can total more than 450.

Example:

“Bol, Bole, Bolechi, Bolechilam, Bolbo, Bolbe”।

Here, “Bole, Bolechi, Bolechilam, Bolbo, Bolbe” verb inflectional form of verb ―Bol”.

3.5 Pronouns

Bengali pronouns are somewhat similar to English pronouns, having different words for first, second, and third person, and also for singular and plural.

Example: Singular: Aami, Tumi

Plural: Amra, Tumra

3.6 Verb

Bangla verbs are highly inflected and are regular with only few exceptions. They consist of a stem and an ending; they are traditionally listed in Bangla dictionaries in their "verbal noun" form, which is usually formed by adding –a ( ) to the stem, for instance,(rakha) = "to put or place". The stem can end in either a vowel or a consonant. Verbs are conjugated for tense and person by changing the endings, which are largely the same for all verbs. However, the stem vowel can often change as part of the phenomenon known as ―vowel harmony‖, whereby one vowel can be influenced by other vowels in the word to sound more harmonious.

Example:

would be the verb "to write", with stem lekh-: (tomralikho) meaning ―you (pl.) write‖ but (amralikhi) meaning ―we write‖. Bangla language has more than 30000 verbs. Diversity of verb morphology in Bangla is very significant. For example, if we consider ―‖ (likhmeans write) as a root word then after adding verbal inflexion ―etc‖ (itechhi), we get a word ―׀ (likhitechhi means am writing) which means a work is being doing in present (for first person). Similarly, after adding inflexion ―e(itechhilam) we get the word ―(likhitechhilam means was writing) which means a work was being done in past. Here, one word represents present continuous tense of the root word –‖ (likh) and another represents past continuous tense. Therefore, we get the grammatical attributes of the main word and other attributes.

3.7 Mood

Mood is the manner of verb in which an action in a sentence is represented.

According to mood of verb it can be classified into these 4 types:

- Indicative Mood
- Imperative Mood
- Subjunctive Mood
- Optative Mood

3.8 Tense

A tense is a form taken by a verb to show the time of an action.

For example: English has 3 types of tense along with 12 sub-types. On the same time Bengali has 3 types of tense with 9 sub types. Now we are going to discuss them.

Present Tense

Things that is true when the words are spoken or written.

Example: Rahim Boi Pore.

There are sub types of tense which are:

- Present Indefinite Tense: Aami Porchi.
- Present Continuous Tense: Karim 1990 shale mara jan.
- Present Perfect Continuous Tense: Maherbani kore Bosun

Past Tense

Things that were true before the words were spoken or written.

Example: Rahim Schoole Giyechilo.

This category of tense is divided into 4 sub types:

Past Habitual Tense

Past Indefinite Tense

Past continuous Tense

Present Perfect continuous Tense.

Future Tense

Things that will or might be true after the words are spoken or written.

Example: Rahim aagmikal schoole jabe.

This category of tense is divided into 2 subtypes:

- Future Indefinite Tense: Tini collage a jaben.
- Future Perfect tense: Doya kore kal k aasben.

3.9 Number

In linguistics, number (singular) is a grammatical category of nouns, pronouns agreement that expresses count distinctions. It can be applied to both animate and inanimate objects. The number categories are singular (singular) and plural (plural).

Example:

- Ghar, manus। (singular)
- Gharguli, manusra। (plural)

3.10 Gender

In Bangla language, Gender is known as Lingo. Gender is a grammatical distinction in which words such as nouns and adjectives are marked according to distinction between Masculine, Feminine and sometimes Neuter, Common.

3.11 Verb Root

Bangla verbs are highly inflected and are regular with few exceptions. They consist of a stem and an ending; they are traditionally listed in Bangla dictionaries in their –verbal noun‖ form, which is usually formed by adding – a (   ) to the stem, for instance, (rakha) means –to keep‖. The stem can end in either a vowel or a consonant, based on which Bangla verb roots.

3.12 Affix

An affix is a morpheme that is attached to a word stem to form a new word. In bangle language affix has two parts:

- **Primary Affix:** The primary affix is to be added to verb and created a new word.

Example: Kaad, Mil, Chap

- **Secondary Affix:** The Secondary affix is to be added to noun in order to change the meaning of this word in different ways.

Example: Pichon, Dhonachi, Knai.

3.13 Suffix

Suffix has no independent meaning, but it alters, retouches the word. The inflections that are added after Verb Roots to make verbs are called Verbal Inflections.

Example: Ti, Ta, Guli ae, ei u etc.

4. METHODOLOGY

A highly inflectional language has the capability of generating hundreds of words from a single root. No doubt that Bangla has the same variety. We improved the efficiency and accuracy step by step by reducing pitfalls and lacking to get highest optimization.

4.1 Process

In our method first we choose a Bangla file where huge lists of Bangla words are available. Tagging the file according to Bangla tag set and save this file in a text file. First we read a file which we already tagged. Its reads a file correctly and gives us proper POS tag.

Secondly, we called a stringtokenizer which will find out the verb from the tagged file where it find (VRB) and (V).

Finally, those verbs we get from step-2 then we save these verbs in another text file named ―stemming‖. From saved file the verbs will be comparing with the suffix list; which we already declared in the code. When the suffixes will be find in the verb it will trim the suffixes and give the root of these verbs as output.

Example: Pora = Por+aa

Where "Por" is the root word and "aa" is a suffix.

4.2 Code

- load tagged file
- Read file(―tagged.txt‖)
- use stringTokenizer
- String str = st.nextToken()
- compare (str.contains(―VRB‖)||(str.contains(―V‖)))
- print (str)
- load_verb list
- Read file(―stemming.txt‖)
- compare with suffix list

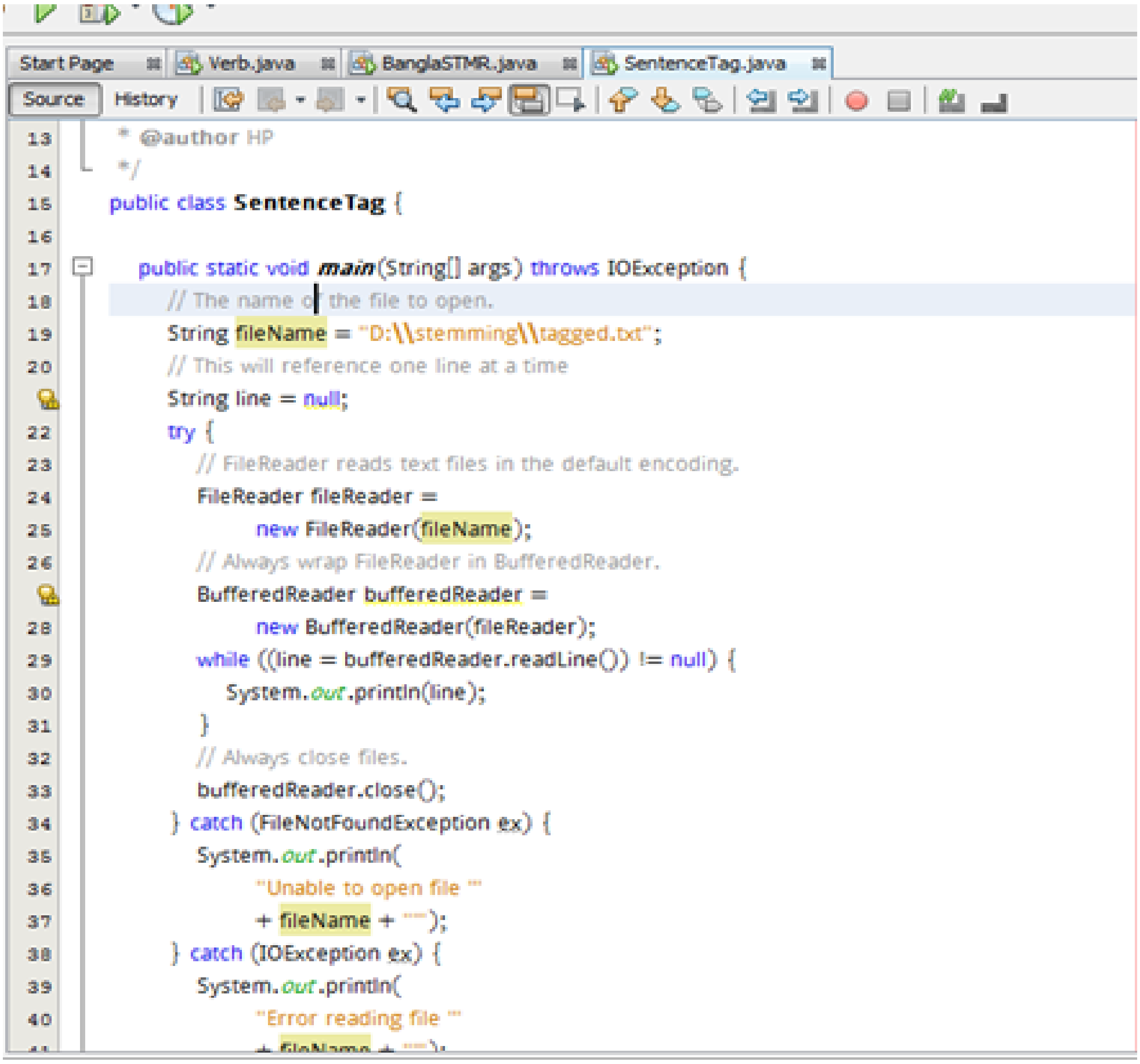
- compare line.endsWith(suffixs[j])
- If found line will be replace by suffix[j] and store in root
- root = line.replace(suffixs[j], "");

4.3 Result and Analysis

We work with tagged file and try to find out the verb to root word .We found 61 words as root words. Almost 60% words have inflection form. We fail to tag some root word correctly because inflectional form of that word not available in our corpus. And also some suffixes are valid for some words at the same time those suffixes are not correct inflectional form of some words.

5. FIGURES

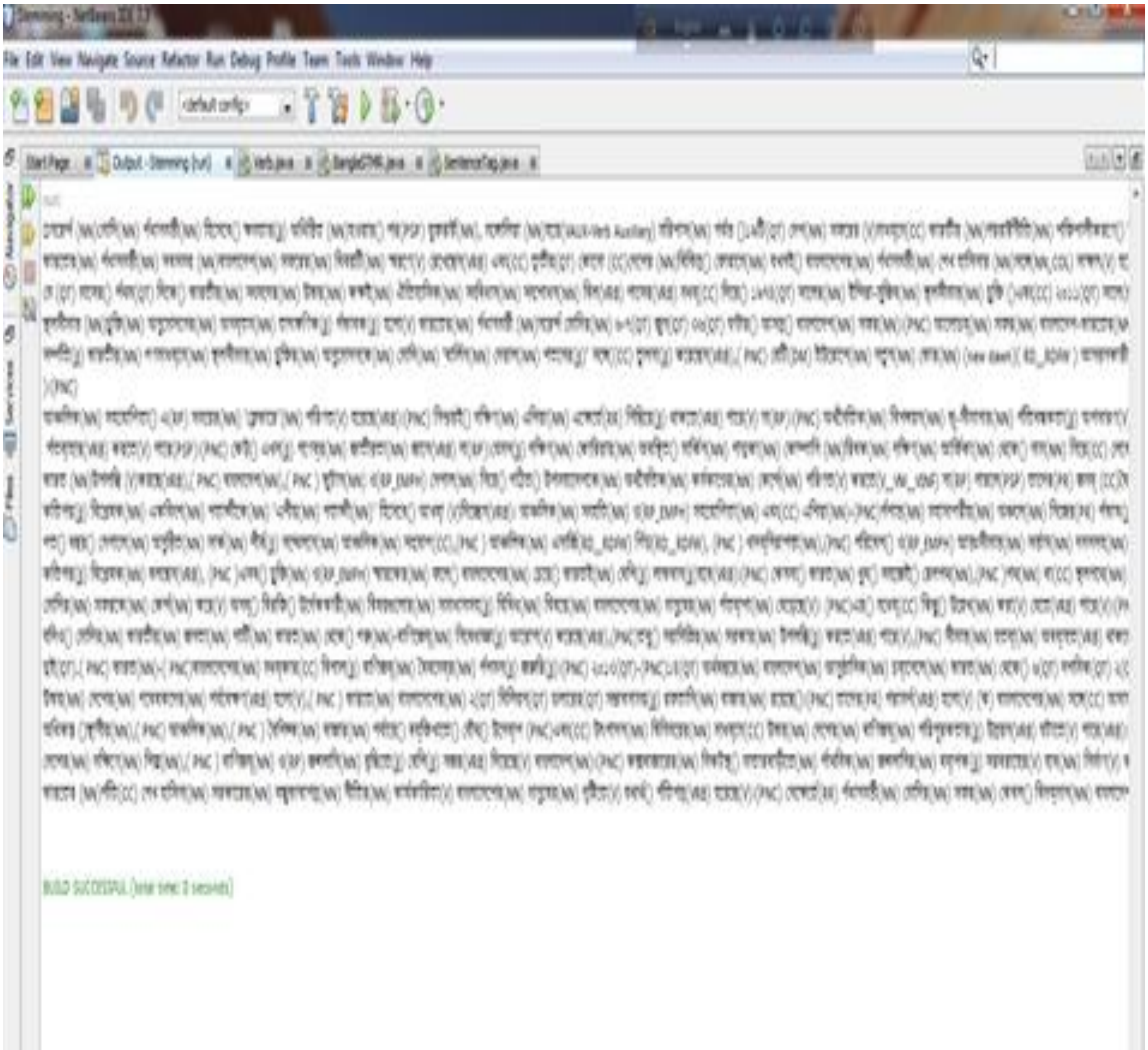
First we tagged a file manually then read a file:



```
 * @author HP
 */
public class SentenceTag {

    public static void main(String[] args) throws IOException {
        // The name of the file to open.
        String fileName = "D:\\stemming\\tagged.txt";
        // This will reference one line at a time
        String line = null;
        try {
            // FileReader reads text files in the default encoding.
            FileReader fileReader =
                    new FileReader(fileName);
            // Always wrap FileReader in BufferedReader.
            BufferedReader bufferedReader =
                    new BufferedReader(fileReader);
            while ((line = bufferedReader.readLine()) != null) {
                System.out.println(line);
            }
            // Always close files.
            bufferedReader.close();
        } catch (FileNotFoundException ex) {
            System.out.println(
                    "Unable to open file '"
                    + fileName + "'");
        } catch (IOException ex) {
            System.out.println(
                    "Error reading file '"
                    + fileName + "'");
```

Figure 1: Input of tagged sentence.

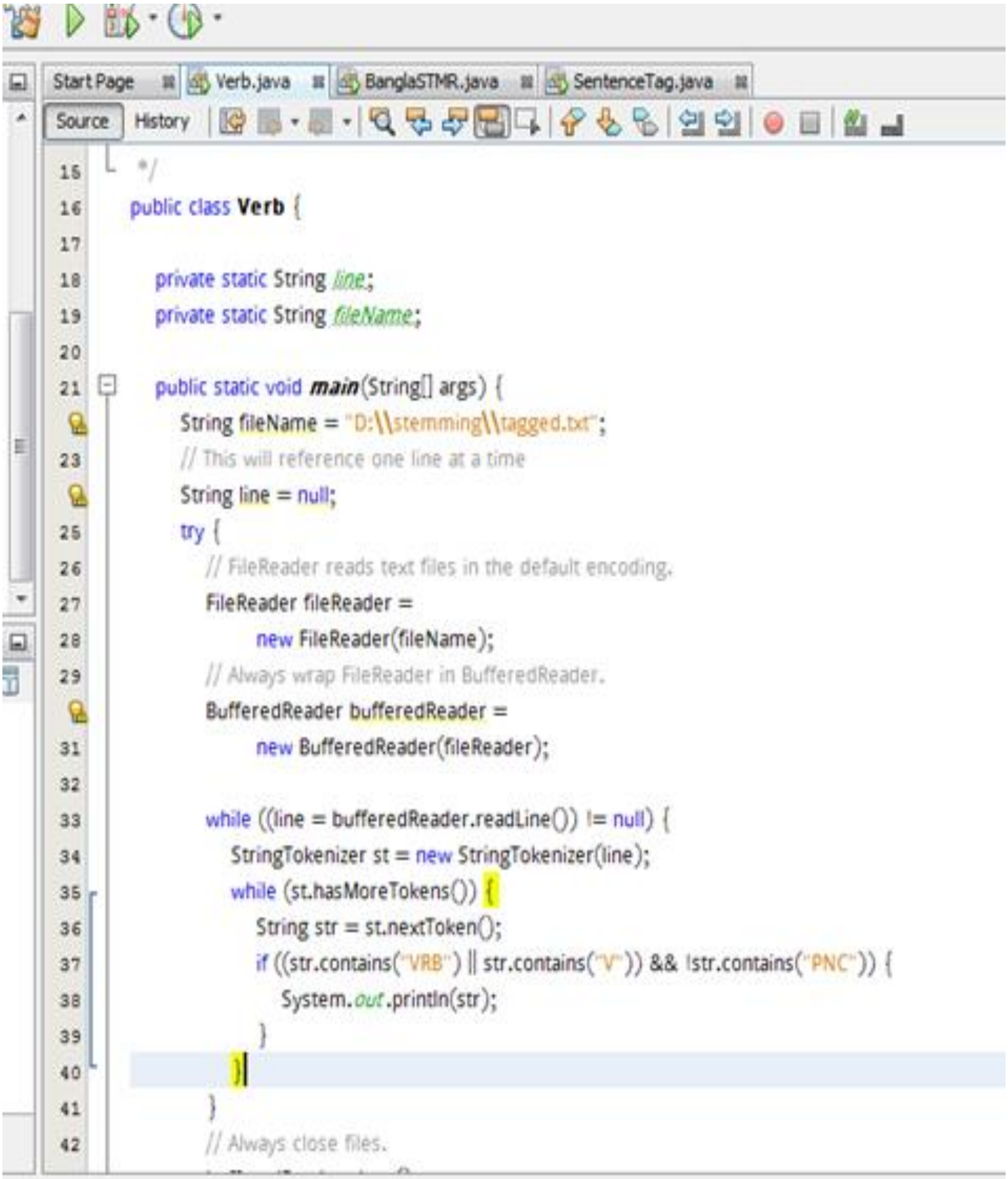
Then we get this output**:**



```
run:
[illegible]
BUILD SUCCESSFUL (total time: 3 seconds)
```

Figure 2: Output of tagged sentence.

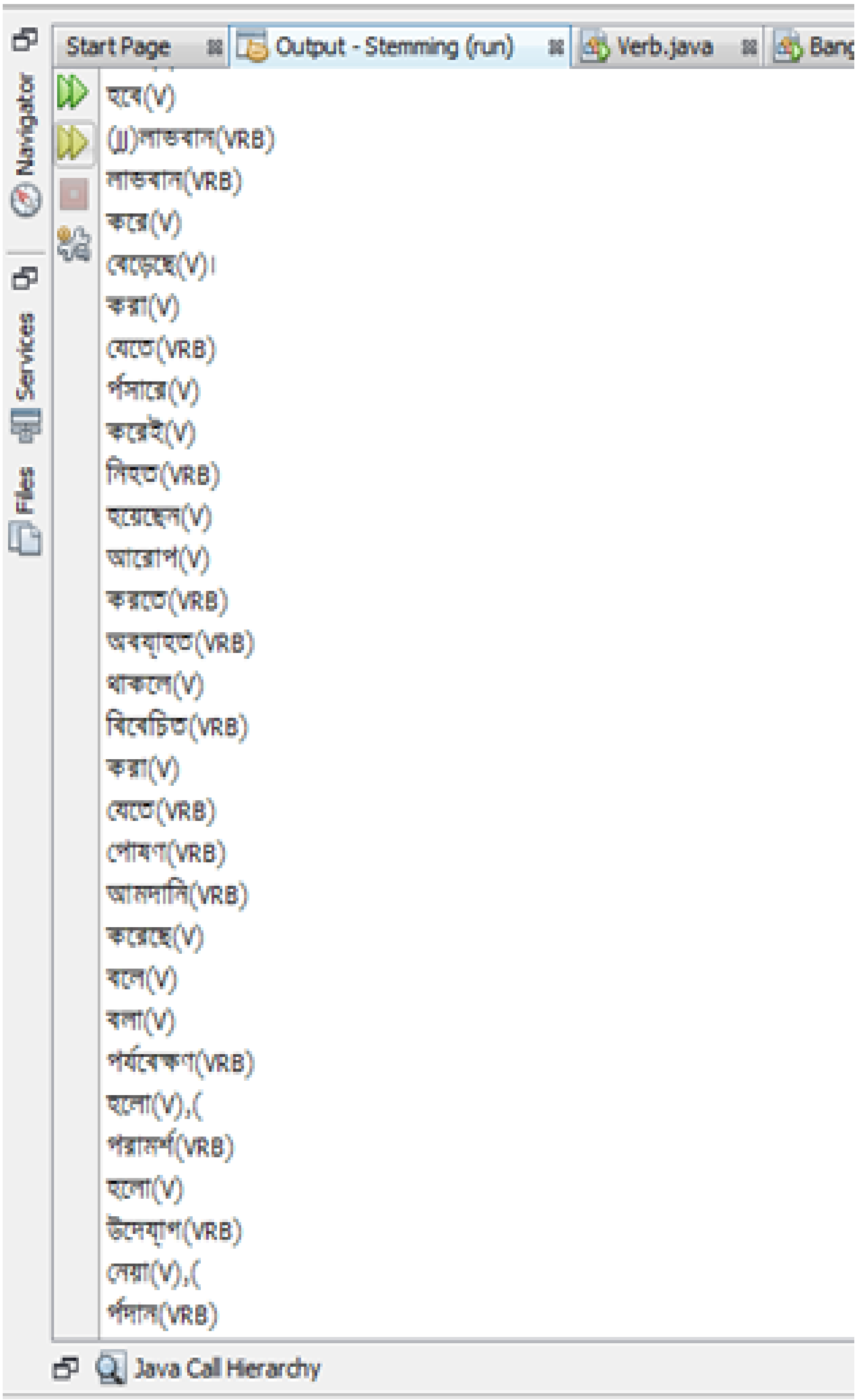
Secondly, we find out the verb from this tag file:



```
 */
public class Verb {

    private static String line;
    private static String fileName;

    public static void main(String[] args) {
        String fileName = "D:\\stemming\\tagged.txt";
        // This will reference one line at a time
        String line = null;
        try {
            // FileReader reads text files in the default encoding.
            FileReader fileReader =
                    new FileReader(fileName);
            // Always wrap FileReader in BufferedReader.
            BufferedReader bufferedReader =
                    new BufferedReader(fileReader);

            while ((line = bufferedReader.readLine()) != null) {
                StringTokenizer st = new StringTokenizer(line);
                while (st.hasMoreTokens()) {
                    String str = st.nextToken();
                    if ((str.contains("VRB") || str.contains("V")) && !str.contains("PNC")) {
                        System.out.println(str);
                    }
                }
            }
            // Always close files.
```

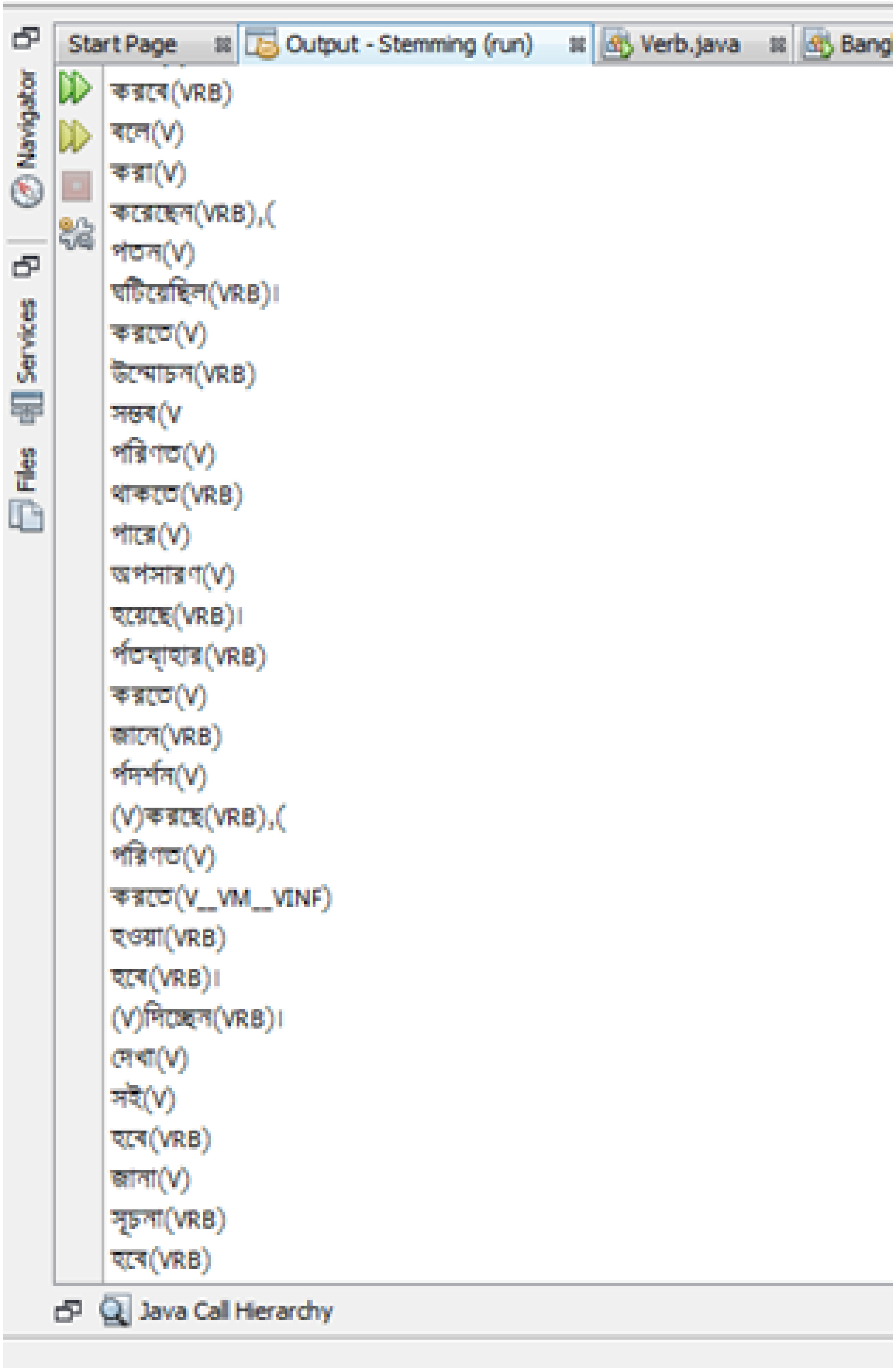
Figure 3: Input of verb tagging.

And Output is:



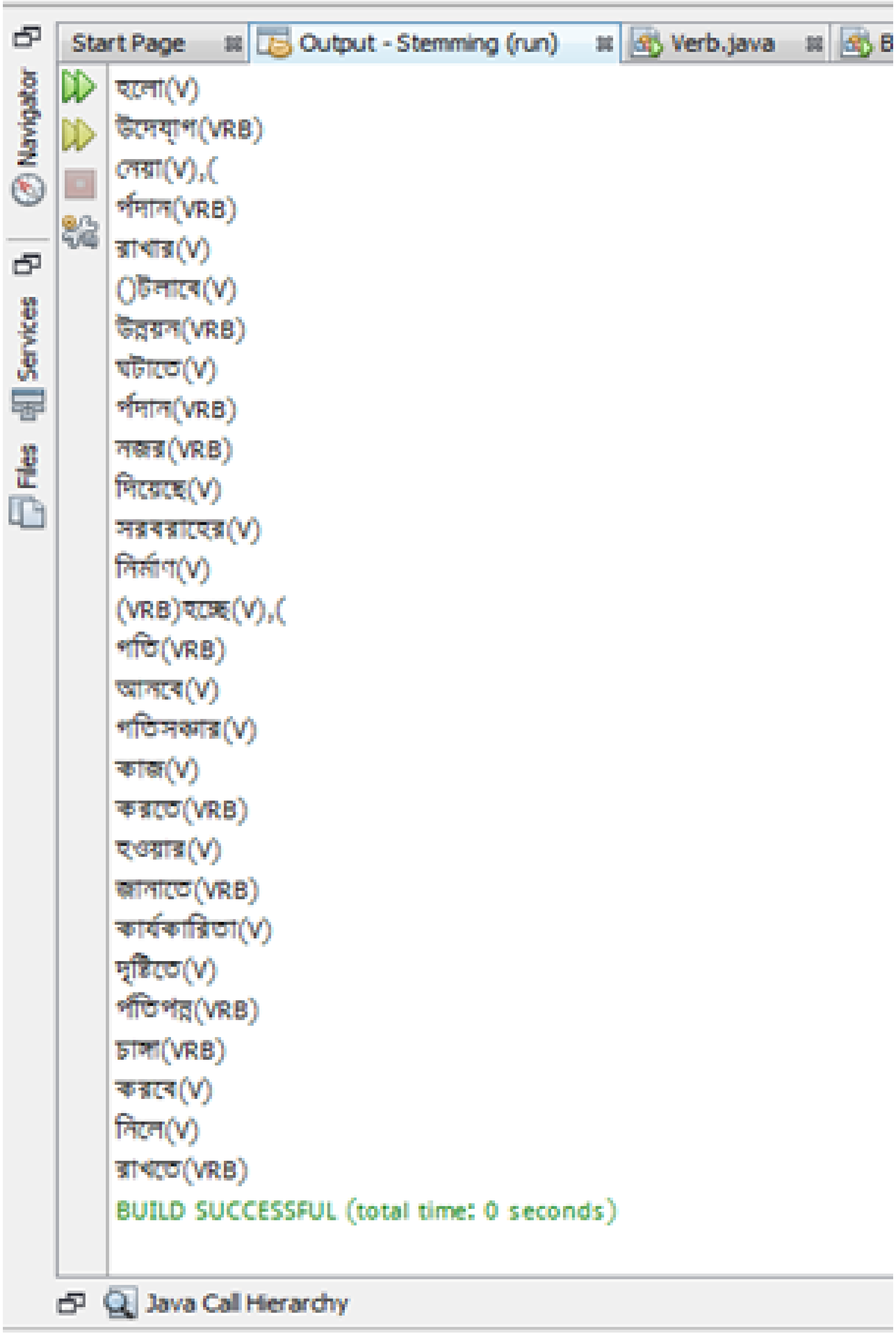
```
হবে(V)
(JJ)লাভবান(VRB)
লাভবান(VRB)
করে(V)
বেড়েছে(V)।
করা(V)
যেতে(VRB)
পারে(V)
করেই(V)
নিহত(VRB)
হয়েছেল(V)
আরোপ(V)
করতে(VRB)
অব্যাহত(VRB)
থাকলে(V)
বিবেচিত(VRB)
করা(V)
যেতে(VRB)
পোষণ(VRB)
আমদানি(VRB)
করেছে(V)
বলে(V)
বলা(V)
পর্যবেক্ষণ(VRB)
হলো(V),(
পরামর্শ(VRB)
হলো(V)
উদ্যোগ(VRB)
নেয়া(V),(
প্রদান(VRB)
```

Figure 4: Output of verb tagging file (a).



```
করবে(VRB)
বলে(V)
করা(V)
করেছেন(VRB),(
পতন(V)
ঘটিয়েছিল(VRB)।
করতে(V)
উন্মোচন(VRB)
সম্ভব(V
পরিণত(V)
থাকতে(VRB)
পারে(V)
অপসারণ(V)
হয়েছে(VRB)।
প্রত্যাহার(VRB)
করতে(V)
জানে(VRB)
প্রদর্শন(V)
(V)করছে(VRB),(
পরিণত(V)
করতে(V_VM_VINF)
হওয়া(VRB)
হবে(VRB)।
(V)দিয়েছেন(VRB)।
দেখা(V)
সই(V)
হবে(VRB)
জানা(V)
সূচনা(VRB)
হবে(VRB)
```

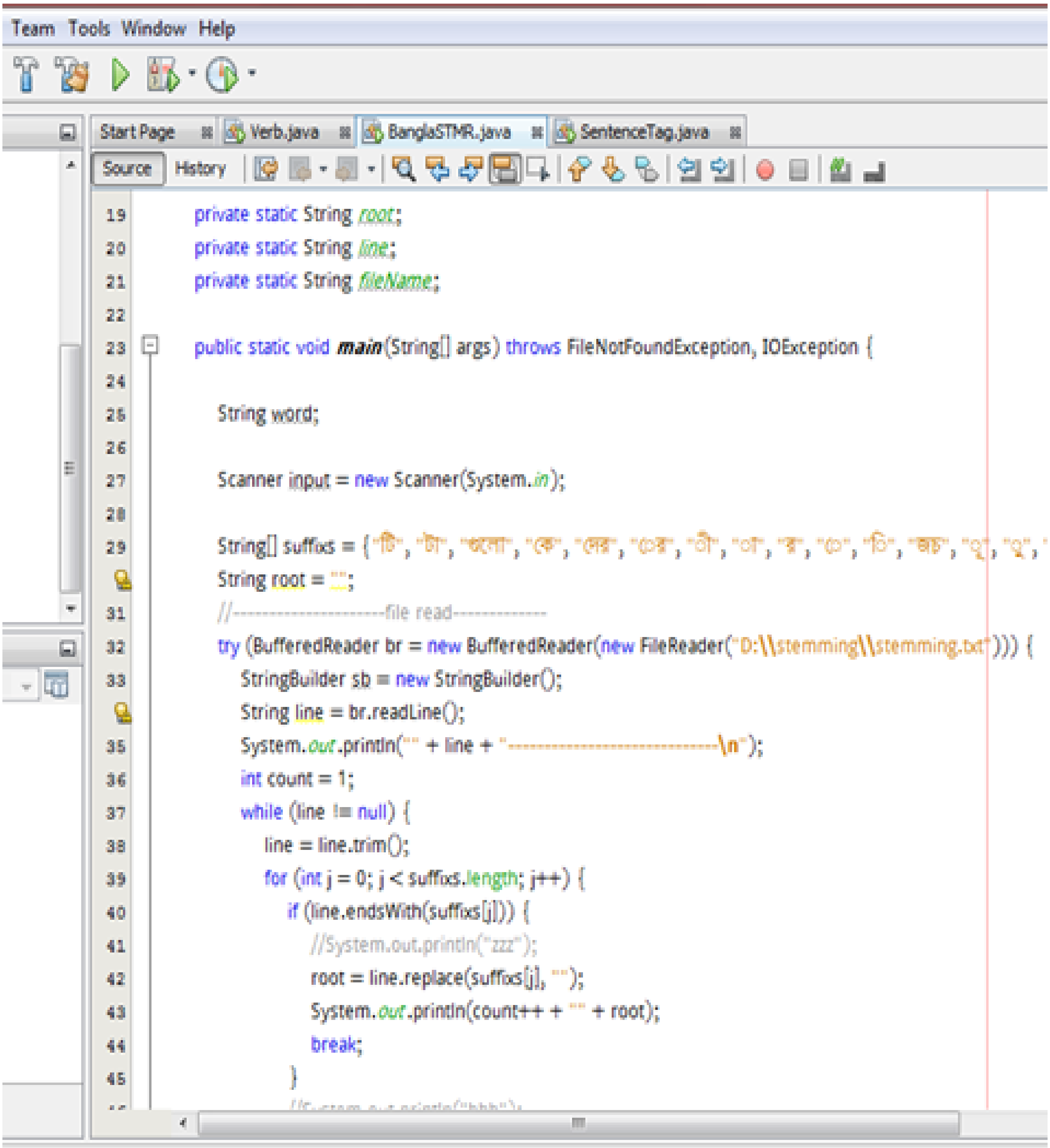
Figure 5: Output of verb tagging file (b).



```
হলো(V)
উদেযাগ(VRB)
নেয়া(V),(
র্পালন(VRB)
রাখার(V)
()টিলাবে(V)
উন্নয়ন(VRB)
ঘটাতে(V)
র্পালন(VRB)
নজর(VRB)
দিয়েছে(V)
সরবরাহের(V)
নির্মাণ(V)
(VRB)হচ্ছে(V),(
গতি(VRB)
আসবে(V)
গতিসঞ্চার(V)
কাজ(V)
করতে(VRB)
হওয়ার(V)
জাগাতে(VRB)
কার্যকারিতা(V)
দৃষ্টিতে(V)
গতিপথ(VRB)
চালনা(VRB)
করবে(V)
নিলে(V)
রাখতে(VRB)
BUILD SUCCESSFUL (total time: 0 seconds)
```

Figure 6: Output of verb tagging file (c).

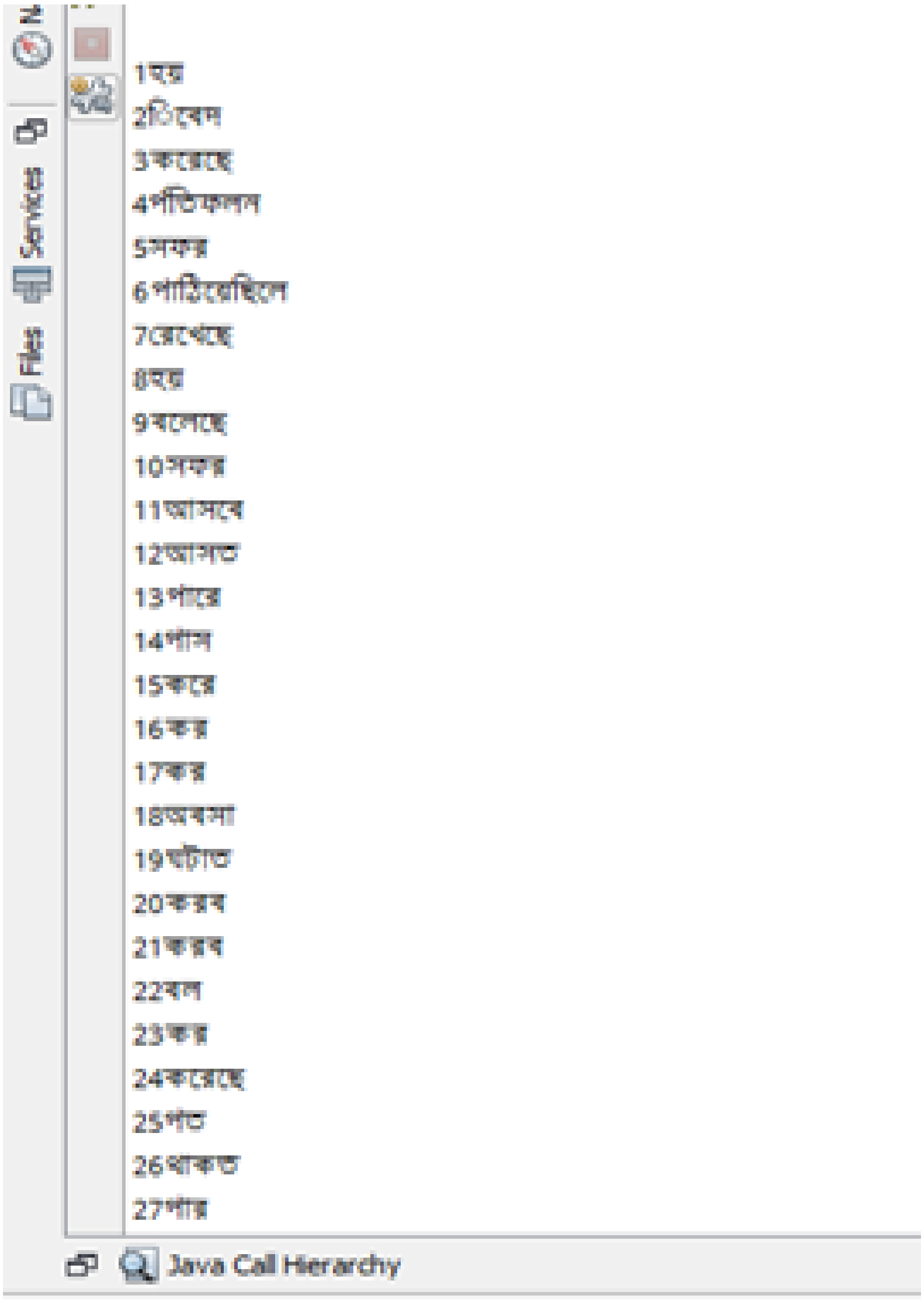
Finally, when we found suffix with word we trim these suffixes in order to find the root of verb.



```
    private static String root;
    private static String line;
    private static String fileName;

    public static void main(String[] args) throws FileNotFoundException, IOException {

        String word;

        Scanner input = new Scanner(System.in);

        String[] suffixs = {"চি", "চা", "অলো", "কে", "লের", "চের", "তী", "তা", "স", "ে", "তি", "অছ", "ু", "ূ", "
        String root = "";
        //--------------------file read-------------
        try (BufferedReader br = new BufferedReader(new FileReader("D:\\stemming\\stemming.txt"))) {
            StringBuilder sb = new StringBuilder();
            String line = br.readLine();
            System.out.println("" + line + "-----------------------------\n");
            int count = 1;
            while (line != null) {
                line = line.trim();
                for (int j = 0; j < suffixs.length; j++) {
                    if (line.endsWith(suffixs[j])) {
                        //System.out.println("zzz");
                        root = line.replace(suffixs[j], "");
                        System.out.println(count++ + "" + root);
                        break;
                    }
```

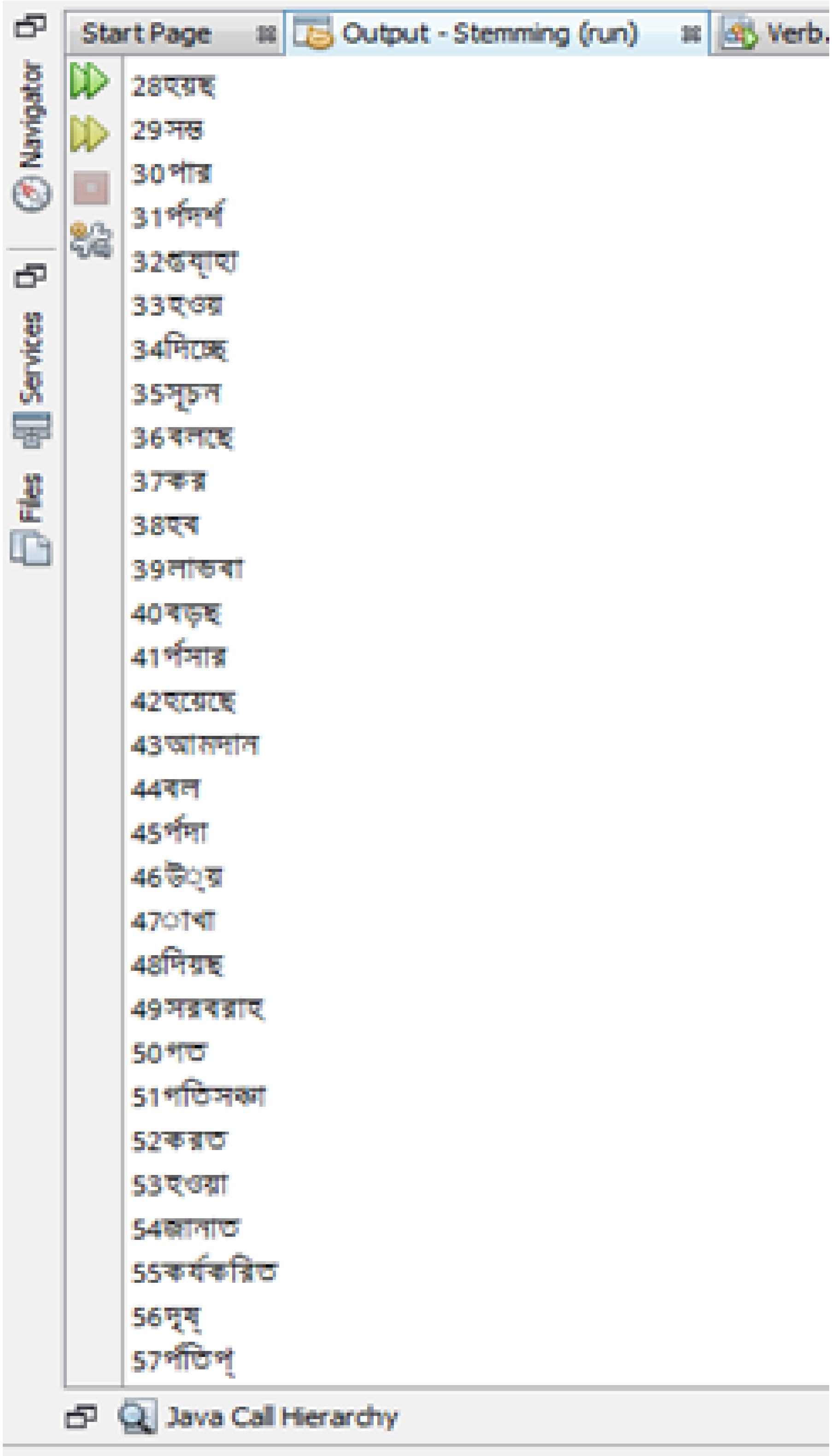
Figure 7: Input verb to root word.

Output is:



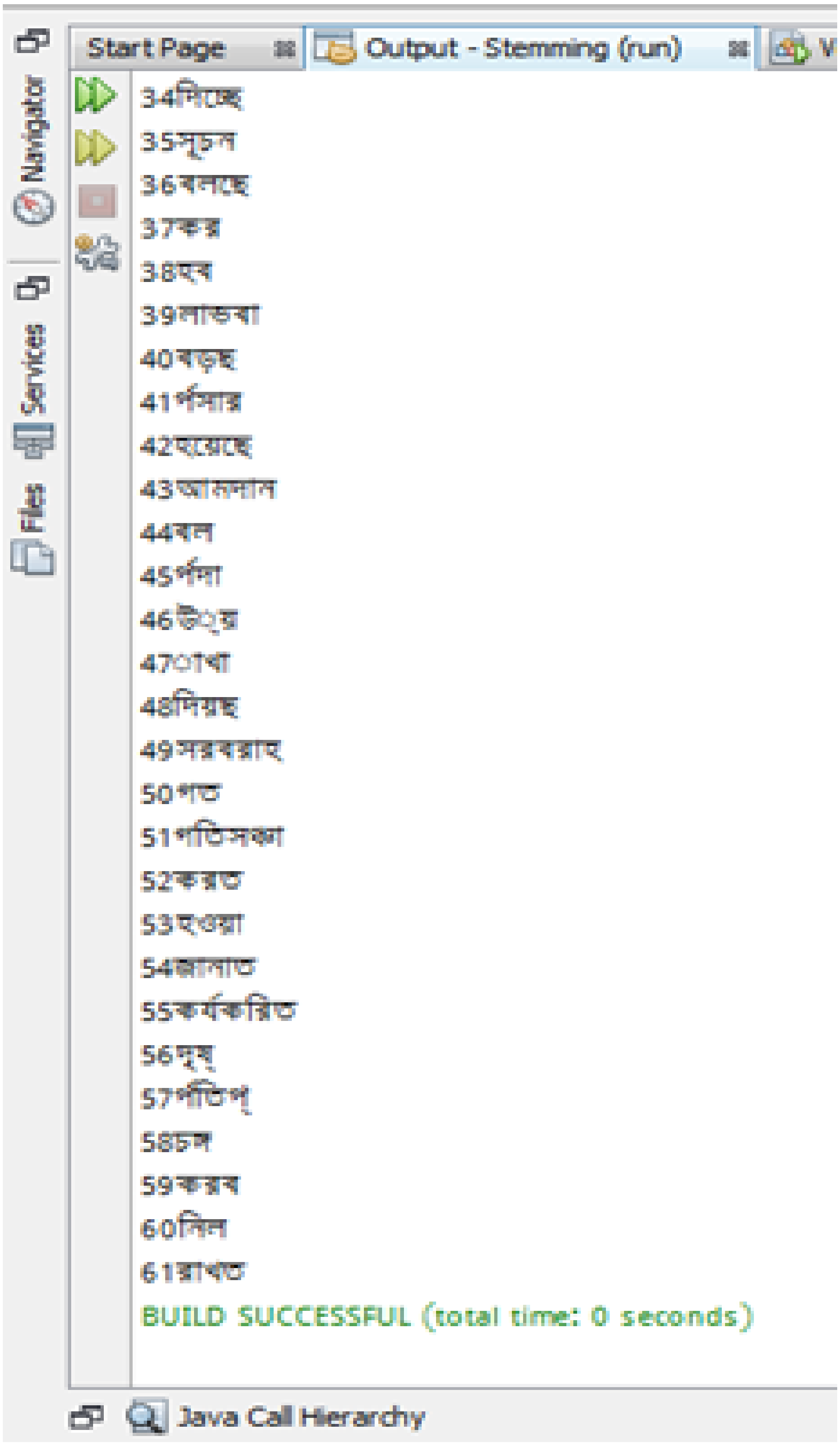
```
1হয়
2িদেবন
3করেছে
4পতিফলন
5সফর
6পািঠেয়িছেল
7রেখেছ
8হয়
9বলেছে
10সফর
11আসেব
12আসেত
13পারে
14পান
15করে
16কর
17কর
18অবসা
19ঘটাত
20করব
21করব
22বল
23কর
24করেছে
25পেত
26থাকত
27পার
```

Figure 8: Verb to root word file (a).



```
28হয়েছ
29সহ
30পার
31র্নর্ম
32ভুবাহ
33হওয়
34দিচ্ছ
35সূচন
36বলছ
37কর
38হব
39লাভবা
40বড়হ
41র্পমার
42হয়েছে
43আমমান
44বল
45র্পমা
46উঠ্র
47◌াখা
48দিয়হ
49সরবরাহ
50গত
51গতিসকা
52করত
53হওয়া
54জাগাত
55কর্যকরিত
56দূর্
57র্পতিপ্
```

Figure 9: Verb to root word file (b).



```
34গিয়েছ
35সূচন
36বলেছে
37কর
38হব
39লাভবা
40বড়হ
41র্পসার
42হয়েছে
43আসমান
44বল
45র্পলা
46উ্র
47াধা
48গিয়হ
49সরবরাহ
50গত
51গতিসক্ষা
52করত
53হওয়া
54জাগাত
55কর্যকরিত
56দূষ্
57গতিপ্
58চল
59করব
60নিল
61রাখত
BUILD SUCCESSFUL (total time: 0 seconds)
```

Figure 10: Verb to root word file (c).

6. LIMITATION AND FUTURE WORKS

6.1 Limitations

6.1.1 Listing exception

It’s really hard to handle a language like Bangla full of inflections and wide ranged variations. And secondly verb is the most responsive portion of the Bangla language. It doesn’t likely to follow the rules all time.

6.1.2 Similarity of verb, Noun and Adjective

In Bangla a single word may have multiple meaning by acting verb/noun/adjectives. More often we notice that same word is being used as verb and noun. And it is hard to differentiate between verb and noun for the foreigners

6.1.3 Manual correction

To make an efficient verb tagger we tried to remove some nonverbal from the ―verb-list‖ satisfying a condition. Hence it’s a tedious, time consuming & tiring approach yet can’t be done on a large scale.

6.2 About the Future Research Chapter

Though it is our first footstep we are planning some more aspects on parts of speech tagging of verb in Bengali. Some are mentioned below:

6.2.1 Structure of Changing

We just gave some sorts of example but there are lots of verbal root in Bengali. Some of them are conventional, some of them are obsolete and some of are no use today. They get changed with the change of suffix. The thing we want that all of the changing should be counted. And that is one of our goals!

6.2.2 Verb tagging with tense

Yes! Bengali is a rich morphological language and with the same line it has large diversity of changing along with person, mood and tense. We are intend to count all the changes with the change of tense that means we are planning to tag the verb with tense.

6.2.3 Building an archive of Bengali roots

As we said before Bengali is a rich language which is actually combination of various roots come from different races such as Mundari, Sanskrit etc. Unfortunately we don’t have enough referential data on verb and verbal root in online. Lack of this people often faces enormous difficulty who works on Bengali grammatical topics. For this reason we are encouraged to make an online archive of root especially verbal root of Bengali words.

6.2.4 Dealing with exceptional verb

We often see some weird word in our day to day life when we talk in our mother tongue. We have said it several times that Bengali has strong diversity of verb along with other parts of speech. And we are willing to deal with this exceptional verb.

6.2.5 Differentiate between verb and noun

More often we notice that same word is being used as verb and noun. And it is hard to differentiate between verb and noun for the foreigners. But when we decided this topic as our thesis research we made our mind not only to grab some CGPA but also to contribute in easy comprehensibility to our mother tongue.

6.2.6 Extinct root

There are some roots in Bengali lexicon which are not used in our day to day conversation. As a result they are going to extinct. For this purpose we are also determined to make a list all of obsolete roots.

6.2.7 Sentence type

It is an easy observation that placing of verb determines the type of sentence mostly both in English and Bengali. For example- if the verb sits at the beginning of the word it might be optative or imperative. So if we spend some of our time with intellectual effort we can easily determine the mood of sentences.

7. CONCLUSION

As natural language processing is a new field and a small amount of work is done on Bengali language, in this paper we have described an approach for verb tagging of natural language text for Bengali. More specifically, I would like to mention that we emphasized on the verb root rather than tagging at this time but definitely our future plan is to tag the verbs to root word correctly and verbal words. And in the coding section we didn't use a strong and established Algorithm based model such as MM (Hidden Markov Method) we tried to find verbs with the Structure-based approach. So, the performance of the current system is not as good as that of the contemporary POS taggers available for English and other European languages. The best performance is achieved for the supervised learning model along with suffix information and morphological restriction on the possible grammatical categories of a word. We conclude that the use of morphological features is especially helpful to develop a reasonable POS tagger when tagged resources are limited. Above all, as we are on the door step of our research, we are very much optimistic to find out an accurate root of these verbs along with some extra beneficiary sides of our research.

REFERENCES

1. Monir Chowdhury, Mofazzal Haider Chowdhury ―Bangla Vasar Bayakoron" Bangla Bhasar Bayakaron, Dhaka, June 2012.
2. Dr. Hayat Mamud, "Uccothoro Sonirvor Bishudo Vasha Shikka" Uccotoroshonirbhor Bishuddhovashashikkha, Dhaka, March 2001.
3. Humayun Azad ―Bakyatatta. The University of Dhaka, 1994.
4. UzZaman, Naushad, and Mumit Khan. "A double metaphone encoding for Bangla and its application in spelling checker." *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*. IEEE, 2005.
5. Md. Shahnur Azad Chowdhury, Nahid Mohammad Minhaz Uddin, Mohammad Imran, Mohammad Mahadi Hassan and Md. Emdadul Haque, Department of Computer Science & Engineering International Islamic University Chittagong - Parts of Speech Tagging of Bangla Sentence‖.
6. Himanshu Agrawal and Anirudh Mani, ―Part of Speech Tagging and Chunking with Conditional Random Fields‖, In Proceedings of the NLPAI Machine Learning Competition, 2006.
7. A Part of Speech Tagger for Indian Languages (POS tagger)‖, In Guidelines of the Workshop on Shallow Parsing in South Asian Languages (SPSAL) 2007.
8. Bangla Newspaper, Prothom-Alo. Online version available online at: www.prothom-alo.net
9. Dasgupta, Sajib, and Dr. Mumit Khan. "Feature unification for morphological parsing in Bangla," 2004.
10. Alam, Firoj, S. M. Habib, DilAfroza Sultana, and Mumit Khan. "Development of annotated Bangla speech corpora," 2010.