Global Journal of Researches in Engineering: J
General Engineering
Volume 17 Issue 1 Version 1.0 Year 2017
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4596 & Print ISSN: 0975-5861

Annotated Bangla News Corpus and Lexicon Development with POS Tagging and Stemming

Tasnim Haider Chaudhury, Abdul Matin, M.S. Hossain, Asie Uzzaman & Md. Masum

Shahjalal University of Science & Technology

Abstract- In this paper, we have developed a mono-linguistic Bengali news corpus using knowledge based AI (Artificial Intelligence) technique from some widely read Bengali newspapers which will be used as a reference corpus and will be very useful for lexicon development, morphological analysis, and automatic parts of speech detection. The corpus contains 74,698 word forms. The words in the lexicon are annotated with a combination of manual tags addressing Parts-of-Speech, Stemming, Morphemes, and other grammatical features are very important for almost all Natural Language Processing (NLP) applications. The lexicon contains around 14 thousand entries.

Keywords: *corpus, POS, tagging, stemming, lexicon.*

GJRE-J Classification: *FOR Code: 200402, 170203*

Annotated Bangla News Corpus and Lexicon Development with POS Tagging and Stemming

Tasnim Haider Chaudhury [α], Abdul Matin [σ], M.S. Hossain [ρ], Asie Uzzaman [Ѡ] & Md. Masum[¥]

Abstract- In this paper, we have developed a mono-linguistic Bengali news corpus using knowledge based AI (Artificial Intelligence) technique from some widely read Bengali newspapers which will be used as a reference corpus and will be very useful for lexicon development, morphological analysis, and automatic parts of speech detection. The corpus contains 74,698 word forms. The words in the lexicon are annotated with a combination of manual tags addressing Parts-of-Speech, Stemming, Morphemes, and other grammatical features are very important for almost all Natural Language Processing (NLP) applications. The lexicon contains around 14 thousand entries. In this paper we present some statistical analysis on some Bengali newspapers Prothom-Alo, Daily Janakantha, Daily Kalerkantho and Amardesh online from 1st January, 2012 to 31st January, 2012 those are the most popular Bengali newspapers in Bangladesh. We proposed a user friendly software interface to the user to annotate a large existing Bengali word set for the lexicon build up process.

Keywords: corpus, POS, tagging, stemming, lexicon.

I. Introduction

The significance of large annotated corpus is a widely known fact. It is an important tool for researchers in Machine Translation (MT), Information Retrieval (IR), Speech Processing, Knowledge Based Computer System and Natural Language Processing (NLP). But in Bengali language we do not have large annotated corpus. The development of corpus creation and distribution of language resources and its availability is must for enhancing Language processing capabilities and research in this field [1]. A corpus is also an essential language resource for creating automatic dictionary from a huge collection of language text [2]. It is the central repository of data for all language processing applications. Researchers are taking this field as a huge sector of researching. In [3, 4, and 5] they focus on automatic Bangla corpus creation by combination of Bangla font. It contains information for human consumption as well as computer programs. The book of "Corpus linguistics and Language Technology" [6] is a warehouse for corpus related studies with special attention to Bangla, where discussed almost every linguistic features of this language. In this paper, we are trying to present a reference corpus and a new approach into language investigation to understand how a text corpus database is utilized to obtain new result on a language or its properties. Bangla Corpus can be used for several purposes including spell checkers and morphological analysis for Bangla language.

A Bangla corpus can be extracted systematically from a Bangla corpus since it is considered a source of all words which is important for verification of Bangla sentence structure [7, 8]. This paper proposes another process to manually build up a corpus, which is essentially a list of all words in the language and tag the words sufficiently with features such as word meaning, Parts-Of-Speech (POS) and all other grammatical features. All these information need to be stored in a database and properly formatted before display to end users. The aim of the project is to formalize a procedure for a collaborative effort by different individuals or groups towards producing a tagged Bangla corpus. This requires a POS tagging interface, both web based and standalone that would provide a common platform for different contributors to enter tag information, semantic and other grammatical information that is available in a dictionary. So we used this huge and mighty source as our source of our corpus data.

II. Web as Corpus Source

The use of the web as a corpus for teaching and research on language has been proposed a number of times. There has been a special issue of the Computational Linguistics journal on Web as Corpus. Several studies have used different methods to mine web data. Our attempt was to create an annotated Bangla text corpus which will contain Bangla text from most popular and well read newspapers of Bangladesh based on date (1st January, 2012 to 31st January, 2012) and several categories (Sports, Crime, Editorial, International, National, State, Sports, Business), so as to make it representative of every linguistic phenomena of Bangla. This project was based on huge data or text available in electronic format. As we are lacking good Bangla OCR applications for collecting Bangla text from printed book, journal and newspaper, so we had to restrict our attempt to collect corpus text from

*Auhtor α, ρ, Ѡ, ¥: Dept. of Computer Science and Engineering Shahjalal University of Science & Technology, Sylhet, Bangladesh.
e-mails: shahadat_sust@yahoo.com, wasir.cse@gmail.com, masum-cse@sust.edu, tasnimhaiderchaudhury@gmail.com*

*Auhtor σ: Dept. of Computer Science and Engineering Cox's Bazar International University (CBIU), Cox's Bazar, Bangladesh.
e-mail: matin.cse.pust@gmail.com*

whatever resources we have available mainly from web. We have selected four newspapers that available in online and we used these in order to create a news corpus. These news corpuses contain 74698 word tokens and 13550 distinct words in this corpus.

III. Data Collection

a) *Collection of the Raw Text*

Many newspaper in our country have own online versions, but we choose four newspapers Prothom-Alo, Daily Janakantha, Daily Kalerkantho and Amardesh online among them mainly because they are the widely read newspapers in Bangladesh and with less spelling mistakes. We consider the raw text of those newspaper available in web and download all news from web for collecting corpus The raw text for the corpus was collected from these newspapers through downloading all the news available for the year of 2012 (from 1st January to 31st January) including magazines and periodicals, which were all in html format. The process took about one month to collect all these available data manually. At this point we ended up with news of thirty days with each day having several text files that contained news of different genres. The corpus size is eighty megabytes.

b) *Conversion to UTF-8 Format*

Then we manually convert these entire text file to UTF-8 format to make these data available and correctly readable for our corpus creator program which only allow UTF-8 formatted text for processing.

c) *Classification of Collected Data*

For quick extraction of information of a UTF-8 formatted Bangla text file, we save these text files in some arranged folder format where name of these arranged folders will give information such as date of collection, source and source type, genre of data. Consider the following example like *E:\08102011\news.prothom-alocom\crime\file_001.txt* that shown in Fig. 1.

d) *Database/System Design*

We create a database named Corpus_db for the purpose of lexicon development with a table named Lexicon_entry having following attributes –

WORD, STEM, PREFIX, SUFFIX, POS, COUNT, AMBIGUOUS, DATE, SOURCE, SOURCE_TYPE, CATEGORY, TITLE.

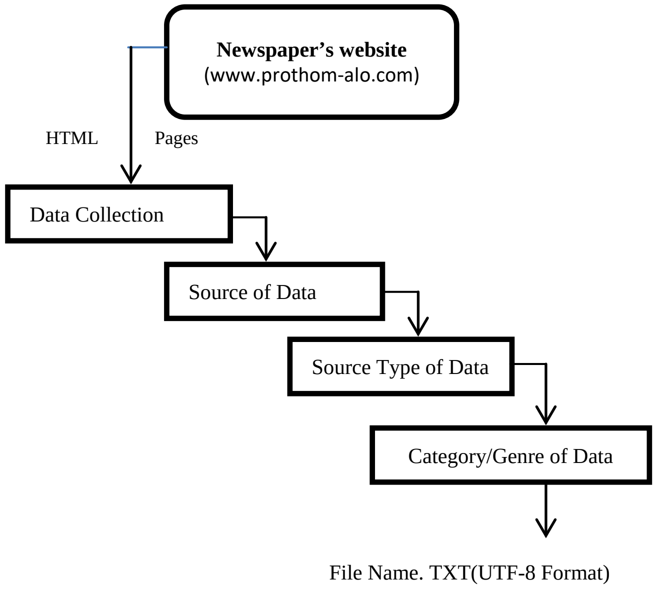

Fig. 1: Definition of data collection.

IV. Text Pre-Processing

In this step, remove following information from the chosen data as a pre-processing –

A. First of all, we replace English letters and numeric alphabet both in Bangla and English from the UTF text data by a single space.

```
str=str.replaceAll("[[\u0000-\u007F][\u09E6-
        \u09EF]॥”“&&[^?!।’'-\\.]]"," ");
```

B. Then we replace all punctuation marks (expect Purnacched, Colon, Question Mark, Exclamation mark, Apostrophe, Dash/Hyphen, Dot) by a single space.str=str.replaceAll("[\\p{Punct}&&[^\\.ı!?'-]]"," ");

C. We also replace all Dots by a single space except those which is preceded by alphabet and followed by space.

D. We also replace all Dash/Hyphen by a single space except those which is preceded and followed by alphabet.

```
str = str.replaceAll("-| - |- "," ");
```

অনেকহাসি-কান্না,সুখদুঃখএবংআনন্দবেদনায়মেশানো ঘটনাবহুল২০১১সাল।

(Here Dash/Hyphen has not eliminated)

সমস্যানিরসনেরপথরাজনৈতিকদলগুলোএগিয়েযাবেকিনান তুনবছরেতাদেখারজন্যইদেশবাসীরএখনঅধীরঅপেৰা।

(Here Dash/Hyphen has eliminated)

E. We also replace all Apostrophe/Inverse Comma/Quotation Mark by a single space except those which is preceded and followed by alphabet.

```
str = str.replaceAll(" '| ' |' "," ");
```

'আজিএঊষারপুণ্যলগনে,উদিছেনবীনসূর্যগগনে।'(Here quotation mark has eliminated)

আগেরদু'বছরঅস্বাভাবিকউর্ধগতিরপরবড়ধরনেরপতনেরম ধ্যদিয়েইশেয়ারবাজারে২০১১সালশুরু।

(Here Apostrophe has eliminated)

F. We replace all sequences of single or more spaces by a single space.

```
str = str.replaceAll("(  )+", "  ");
```

G. Then output file will save as “D:\PreProcessed_File.txt” and this pre-processed text will be used for POS tagging, Stemming and Morphological Analysis.

H. This pre-processed text file will be referred to as our Raw Corpus. This corpus will be useful to evaluate the performance of annotated text corpus and also will be used for input text of machine translation system.

V. Develop User Interface

We proposed a user–friendly software interface to the user to annotate a large existing Bangla word set for the lexicon build up process. The effort will be a significant progress towards development of a properly annotated lexicon. This user interface has two distinct parts – one for building corpus and show text information such as source, source type, category, date, title and news content. Another parts for issuing manual addressing Parts-of-Speech, Stemming, Morphemes, issue clitic, ambiguous condition and other grammatical features.

VI. Lexicon Development

A supervised machine learning method has been used for lexicon development from the Bengali news corpus. No extensive knowledge about the language is required except the knowledge of the different inflections that can appear with the different words in Bengali. To make proper annotation of word form, we accomplished each word form with POS tag, stem form, suffix, prefix, ambiguous condition (if exists), statistical counting. Initially, all the words (infected and uninfected) are extracted from the pre-processed text and added to a database with proper POS, stem, prefix, and suffix. The system retrieves the words from the pre-processed text and creates a database of distinct word forms with fully annotation. Here is given the structure of lexicon development process and some sample word forms shown in Fig. 2.

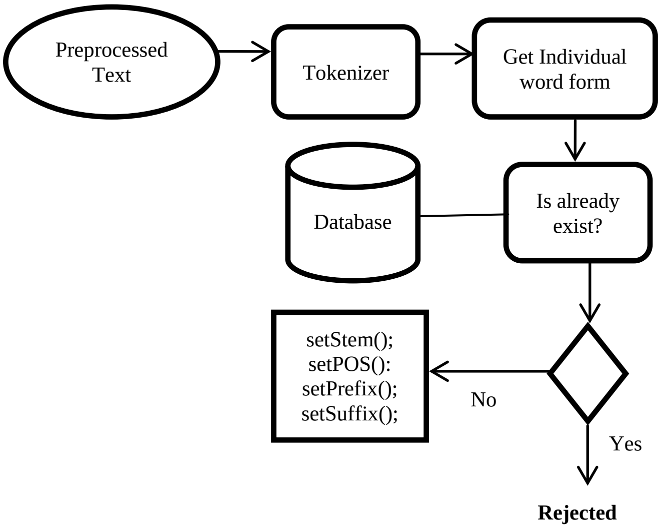

Fig.2: Structure of Lexicon development

A. POS tagging for Lexicon Development

The Part-of-Speech (POS) tagging is the process of assigning each word of a text with an appropriate parts of speech tag. POS tags often signify the morphological [9], phonological and contextual properties of a word, and also provide information about neighboring words.

In Bengali, there are five different POS namely, noun, pronoun, verb, adjective, and indeclinable (prepositions, con-junctions, and interjections). Noun, verb and adjective belong to the open class of POS in Bengali. In this lexicon analysis, we use seven parts-of-speech by extraction main five parts-of-speech in Bangla language. As, we know, there are a lot of word with proper noun are used in Bangla language. So we keep the proper noun distinct from other noun to get us more detail of the word form which is in the range of noun. To handle clitic which is one of the most common ambiguous situation in natural language Processing (NLP), we define a new POS form named clitic. If we add a word form in our lexicon database without setting any POS to that word form, our corpus creator software automatically set its POS as UNKNOWN. Noun and verb words are tagged by looking at their infections. Some infections may be common to some word form. In these cases, more than one POS may be generated for few words form. But here we set the mechanism for only one POS of a word form.

We only suggest a procedure to handle this ambiguity for initial level where POS ambiguity is resolved by checking he number of occurrences of these possible root words along with the POS tags as derived from same word forms. Pronoun and indeclinable are basically closed class of POS [10] in Bengali and these are added to the lexicon manually. It has been observed that adjectives in Bengali generally occur in four different forms based on the suffixes attached. For simplicity of counting or detecting sentence, we propose a user define POS named EOL (end of line). Here we detect POS for *Purnacched, Question Mark, and Exclamation mark* as EOL. The short description of POS categories is given in Table I.

B. Stemming for Lexicon Development

Stemming is an operation that splits a word into the constituent root part and affix without doing complete morphological analysis. It is used to improve the performance of spelling checkers and information retrieval applications, where morphological analysis would be too computationally expensive. Terms with common stems tend to have similar meaning. So it can drastically reduce the dictionary sized used in various NLP applications, especially for highly inflected languages. We handle this stemming process manually like previous POS tagging process. After tokenizing pre-processed text into individual word form we manually set root of that word form by removing prefix and suffix of it. Then this stem form is stored in STEM field of lexicon database.

In our process, we first stripped off the suffix part from Bengali words depending upon the type of suffixes. Then we checked for the validity of the suffix stripped word as root word, using a Bengali dictionary. If it is not sufficient we strip the affix part of the remaining part of the word form. It can bring a set of word with same root form in a series to learn about them easily. We can get almost similar word by retrieving word which has same root/stem form.

C. Setting Morphemes for Lexicon Development

A smallest meaningful linguistic unit is consisting of a word or a word element that can't be divided into smaller meaningful parts. At the time we set stem word for a word from, we also store the stripped part of the word form as morphemes. First of all we split the suffix part of a word and store it in out lexicon database. Then we check out remaining part of our word form whether there is any prefix part of that word. If exist, we strip it from the remaining word part and sore it to database by fully manually.

D. Use this Lexicon for Corpus Creation

A Corpus from linguistic point of view is defined as a collection of transcribed speech or written text compiled mainly to enhance linguistic research. The key resource to any linguistic research is a trained, annotated corpus which can elevate language processing capability such as automatic part of-speech tagging, machine translation, question-answering, stemming etc.

We design and develop a view of annotated corpus which is mainly based on knowledge based representation (Knowledge based AI technique). Here we used our Lexicon as knowledge reference for our corpus. Our lexicon is the collection of word forms with fully annotated where each word form is accomplished with parts-of-speech, stem, Morphemes (suffix, prefix) and statistical counting. When we add a word form to corpus, we bring all the morphological and grammatical information from lexicon and add this information with that word form. Corpus procedure and flow are shown in Fig. 3 that we used for creation of corpus by following steps-

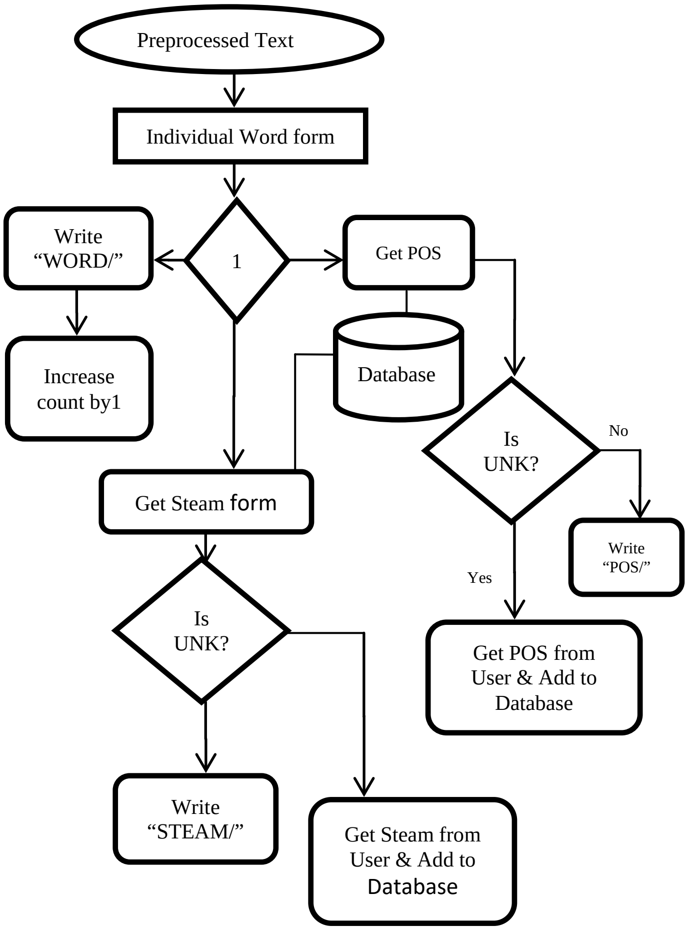

Fig.3: Process flow diagram of Corpus

Table I: Description of POS categories

Tag Description	Tag Label	Examples
Noun(Except Proper Noun)	NN	অতীত,আন্দোলন, কমিটি, গাড়ি, তথ্য
Proper Noun	PN	মোহাম্মদ,বাংলাদেশ, টাঙ্গাইল
Adjective	ADJ	অনেক, অধিকাংশ,উচ্চ, সাধারণ, ভালো
Adverb	ADV	তাহলে,বর্তমানে,সাধারণত, অবৈধভাবে
Verb	VRB	করেছেন,যাওয়া, পারবেন, রাখা, আসা
Pronoun	PRO	আমি,আমার,নিজেদের,অনেকের
Indeclinable(Preposition, Conjunction & Interjection)	IND	এবং,যেমন,হতে,থেকে, কিন্তু, মধ্যেই

1. First of all we take pre-processed text as an input of our corpus creation.
2. Then tokenize this text into word forms. Then all these word forms is stored in an iterative list.
3. This iterative list is looped and gets each word forms as a sequence they were in pre-processed text. Then for each word, POS, Stem is brought from lexicon database and adds this information following the word form separating with for slash (/). Then we make a small change in lexicon, we increase the count value of lexicon by 1 of a word from each time we find this word form. This helps us to find the number of occurrence of a word form.
4. For defining end of a sentence, we use EOL as word form and EOL as POS.

- If sentence is an Assertive sentence, we use as stem.
- If sentence is an Interrogative sentence, we use as stem.
- If sentence is an exclamatory sentence, we use ES as stem.

Structure:

WORD/SETM/POS

Example:

আইনি/আইন/ADJ/পদক্ষেপের/পদক্ষেপ/NN/প্রসঙ্গে/প্রসঙ্গ/NN/তিনি/তিনি/PRO/কিছু/কিছু/ADJ/বলার/বলা/RB/আগেই/আগে/ADV/উপস্থিত/উপস্থিত/NN/বিভিন্,বিভিন্ন/ADJ/বাস-ট্রাক/বাস ট্রাক/NN/সমিতির/সমিতি/NN/প্রতিনিধিদের/প্রতিনিধি/NN/সম্মিলিত/সম্মিলিত/ADJ/প্রতিবাদে/প্রতিবাদ/NN/তিনি/তিনি/PRO/চুপ/চুপ/NN/করে/করা/VRB/যেতে/যতে/VRB/বাধ্য/বাধ্য/ADJ/হন/হন/VRB/EOL/AS/EOL

5. If we do not get the POS and Stem information of a word form we simply assign it's POS and Stem as UNK. But we steel make change in lexicon database by increasing its counting value by 1.

   Example: বিআরটিএ/UNK/UNK

6. Before adding tokenized word form of pre-processed text to corpus, we add the entire information associate with this new text with some predefine TAG to raw corpus. Tag format of our news corpus has given in Table II.

VII. Experiment and Data Analysis

a) Statistical Analysis

Regardless of the size of the corpus, it may subjected to both qualitative as well as quantitative analysis using various methods of statistics . Both these types of corpus analysis have different perspectives. Quantitative analysis focuses classifying different linguistic properties where qualitative analysis aims to give some complete and detailed description of the observed phenomena. We wish to focus on some simple quantitative analysis using U-Gram model.

We develop our corpus development program in such efficient away where researcher can easily get a lot of common and most focused perspective statistical output without any further processing. Here also some user define output generator where user can get output with is desire requirement.

Here we divide our statistical output generator procedure in two distinct parts:

- One for automated query based information.
- Another for user defines query based information.

As result of automated query base perspective statistical output, we provide twelve statistical counting results. This type of statistical counting will be very helpful for linguistic analysis, machine translation, Morphological analysis, spelling variations, morphological structure, and word sense analysis. These statistical counting are,

- Number of source from where this corpus data collected and there list.
- Number of source type and their list of this source of data.

Table II: Tag Formats of News Corpus

Tag Name	Tag Description / Purpose
<ENTRY>	To define start of a new news information/data.
<SOURCE></SOURCE>	Source of data. (www.prothom-alo.com)
<TYPE></TYPE>	Source type of data (news, blog)
<DATE></DATE>	Date of collection of data (11-01-12)
<CATAGORY></CATAGORY>	Genres of that data (sports, crime)
<TITLE></TITLE>	Title of news/data
<CONTENT></CONTENT>	Main content of the news.
</ENTRY>	To define end of this news information/data.

- Number of fields/genres of data collection.
- Number of Raw word/Number occurrence of word stored in corpus.
- Number of Unique word in Corpus.
- Number of Unique Stem word of these raw words.
- Total Number of Sentence.
- Total Number of Assertive Sentence.
- Total Number of Interrogative Sentence.
- Total Number of Exclamatory Sentence.
- Total Number of Clitic.
- Total Number of occurrence of Clitics.

Statistical counting of our annotated Bangla text corpus is shown in Table III. Our Corpus program also provide some qualitative analysis aims to give some complete and detailed description of the observed phenomena which include word level frequency analysis, behavior of bangle word, use of non-Bangla word etc. These type of information can be get by using user defines query based annotated text corpus program interface.

b) Word frequency Analysis

Study of frequency calculation can provide important information about the usage of words in a text. Using above query based information retrieval system, we can be figure out which of the words are generally most frequent given in Table IV to Table VII.

Table III: Statistical counting of annotated Bangla text corpus

Serial No	Information	Count
1	Number of source	4
2	Number of source type	1
3	Number of fields/genres	19
4	Number of Raw word/Number	74698
5	Number of Unique word	13550
6	Number of Unique Stem word	1423
7	Total Number of Sentence	5472
8	Number of Assertive Sentence	5377
9	Number of Interrogative Sentence.	72
10	Number of Exclamatory Sentence.	23
11	Number of Clitic	3
12	Number of occurrence of Clitics	136

Table IV: Most frequently used RAW words in corpus

Word	Percentage	Word	Percentage
ও	1.78	এ	0.39
না	1.34	এর	0.30
হবে	1.21	করে	0.30
হয়	1.15	তিনি	0.26
করা	0.95	আর	0.17
জন্য	0.52	তাঁর	0.08
এই	0.47	মধ্যে	0.069

Table V: Most frequently used STEM words in corpus

Word	Percentage	Word	Percentage
করা	0.4	বছর	0.16
জন	0.25	সে	0.15
দিন	0.23	জন্ম	0.13
জানা	0.20	হয়	0.13
পারা	0.18	দল	0.13
থাকা	0.16	আসা	0.13
দেশ	0.16	কম	0.10
এক	0.16	কাজ	0.10

Table VI: Most frequently used POS words in corpus.

POS Name	Percentage
NN	56.43
VRB	20.53
ADJ	16.41
PN	13.71
ADV	5.94
PRO	3.39
IND	1.98
CLK	0.104
UNK	1.35

Table VII: Most frequently used PREFIX and SUFFIX words in corpus

Prefix	Percentage	Suffix	Percentage
অ	9.30	এ	15.70
নব	4.07	এর	15.43
উপ	2.23	র	8.53
প্রতি	2.23	য়	4.72
এ	1.74	ই	4.63

VIII. Conclusion

Lexicon development, Part of Speech (POS) tagging and stemming are very important for almost all Natural Language Processing (NLP) applications. Corpus is considered as basic resource for language analysis and research for many foreign languages. This reflects both ideological and technological change in the area of language research. The effort will be a significant progress towards development of a properly annotated lexicon. The outcome of the research will significantly be helpful for future analyzer in the processes of Morphological Analysis, Automatic grammar Extraction and Machine Translation for Bangla.

References Références Referencias

1. C. Cieri and M. Liberman, "Issues in Corpus Creation and Distribution: The Evolution of the Linguistic Data Consortium, University of Pennsylvania and Linguistic Data Consortium Philadelphia," Pennsylvania, USA.
2. J. Hasan. "Automatic dictionary construction from large collections of text," Master's thesis, School of Computer Science and Information Technology, RMIT University, 2001.
3. Asif Iqbal Sarkar, Dewan Shahriar Hossain Pavel and Mumit Khan, "Automatic Bangla Corpus Creation," BRAC University, Dhaka, Bangladesh.
4. Khair Md. Yeasir Arafat Majumder, Md. Zahurul Islam, Naushad Uz Zaman and Mumit Khan, "Analysis of and Observations from a Bangla News Corpus," Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh.
5. A. Bharati, R. Sangal and S.M. Bendre, "Some Observations Regarding Corpora of Some Indian Languages," Proc. Intl. Conf. Knowledge Based Computer Systems (KBCS98), NCST, Mumbai, 17-19 Dec. 1998.
6. N.S. Dash, "Corpus Linguistics and Language Technology," Mittal, New Delhi, 2005
7. Nur Hossain Khan, Md. Farukuzzaman Khan, Md. Mojahidul Islam, Md. Habibur Rahman and Bappa Sarker, "Verification of Bangla Sentence Structure using N-Gram," Global Journal of Computer Science and Technology, Volume 14 Issue 1 Version 1.0 Year 2014.
8. Md Hanif Seddiqui, AKMS Rana, Abdullah Al Mahmud, Taufique Sayeed, "Parts of speech tagging using morphological analysis in bangla," Proceeding of the 6th International Conference on Computer and Information Technology (ICCIT), Bangladesh.
9. M M Asaduzzaman and Muhammad Masroor Ali, "Morphological Analysis of Bangla Words for Automatic Machine Translation," 6th International Conference on Computer and Information Technology (ICCIT) 2003. Jahangirnagar University, Dhaka, Bangladesh, pp.265-270,2003
10. Kristina Toutanova and Colin Cherry, "A global model for joint lemmatization and part-of-speech prediction," Proceeding on ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1, Pages 486-494.