VOL. 3, NO.11 Nov, 2012 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

Bangla Academic Dictionaries (BAD) Corpus with some Applications and Statistical Analysis

[1]MD. Abdul Awal Ansary, [2] Mohammad Reza Selim, [3] Muhammad Zafar Iqbal
[1] Assistant Professor, Dept. of CSE, Sylhet International University, Bangladesh
[2] Associate Professor, Dept of CSE, Shah Jalal University of Science & Technology, Bangladesh
[3] Professor, Dept of CSE, Shah Jalal University of Science & Technology, Bangladesh
E-mail :{ [1] awal_sust@yahoo.co.in, [2] selimbd@yahoo.com, [3] mzi@sust.edu }

ABSTRACT
The present era is a globally digital era and information/knowledge sharing is a common scenario now a days. But language difference is a great problem for sharing knowledge among different communities. To overcome this problem, different countries in the world have developed machine translators (MTs) such as English to French, English to Japanese, English to Spanish, English to Chinese, etc. However, there is no sound system that can translate English text into Bengali text. To develop a machine translator for English to Bengali translation and other linguistic purposes, first of all one needs a bilingual corpus between English and Bengali text. In this article, we have described the architecture to develop an English-Bengali bilingual corpus which is based on Bangla Academic Dictionaries (BAD) - English to Bengali (E2B) and Bengali to English (B2E) Dictionaries. We have used Bangla Academic dictionaries (BAD) because vocabularies are available in dictionaries from different genres. We have also mentioned different applications and statistical analysis of this BAD corpus.

Keywords: *Parallel Corpus, POS tagging, Statistical Analysis, Corpus alignment, Machine Translation, NLP.*

1. INTRODUCTION
A corpus in plural corpora or corpuses is a large and organized set of text or speeches, usually electronically stored and processed. It is an important tool of natural language processing (NLP) that can be used as a resource in linguistic research. It can be used for automatic part of speech tagging, spell checking, statistical analysis and hypothesis testing, checking occurrences, enriching the existing dictionaries, spelling verifications, and so on.

Different languages have their own corpuses varying in size, genres and purposes. The English, Japanese, Spanish, Swedish, and many other countries have rich corpuses to develop and analyze their languages from different views. Even our neighbor country India also has its corpuses of different languages for instance Assumi, Hindi, Monipuri etc. But unfortunately we have no any rich corpus in Bangla but it has already started from different views to create corpuses for different purposes, for instance Bangla NEWS corpus, Bangla Prose corpus etc. Comparing with other languages and for the demand of the enrichment of our language we need more research and to make Bangla corpus from different sources and to ensure their effective use. The goal of this article is to discuss about the corpus that is made from BAD and some statistical analysis regarding English words that are available as foreign words in Bangla from different perspectives. Anybody can use this corpus as a part of machine translation from English to Bangla or vice versa. The number of vocabulary of BAD can be increased comparing with larger (in size) corpuses that will make from large part of Bangla texts or voices. The BAD corpus is the Parts of Speech (POS) training corpus and it can be used for any English or Bangla text POS tagging.

2. EXITING WORKS ON BANGLA CORPUSES
Although Bangla language has no so rich corpuses regarding English, French, German, Dutch, Swedish languages, the pioneer named Suniti Kumer Chattapadhyay (1926) had made some quantitative studies on limited sample databases that are collected from Bangla dictionary, selected prose texts of modern Bangali literature. He determined the ratio of occurrence of words of different parts-of-speech. In 1965, the Indian Statistical Institute, Kolkata, Nikhilesh Bhattacharya had made a corpus collecting text samples from novels and other prose texts written by Bankim Chandra Chattopadhyay, Bibhuti Bhusan Bandyopadhyay and others to calculate the frequency of use of various characters of the Bangla language. In 1980s, some scholars at the department of Printing Technology, Jadavpur University, Kolkata had made a corpus to analysis the frequency counts of characters taking from some prose texts of Bengali, Assamese, and Monipuri. The goal of this study was to design the keyboard layout for Bengali type writers. Same type of study also done by Bhakti Prasad Mallick and his group (1994-2000), taking samples from Gitanjali, Sabhyatar Sankat and Shes Lekha written by Rabindranath Tagore. Although all of the above corpuses were created and all findings are generated manually, the Bengali linguistics should express their gratefulness to them for their pioneering contributions.

Now-a-days, the corpuses of Bangla Texts are in digital forms that means text are collected and statistical findings are also generated using computers and effective software of different algorithms. In this case, Dr. Niladri Sekhar Dash at the Indian Statistical Institute, Kolkata has done a lot of works including frequency count processes that are applied on the corpus to draw information, the

comparison between the new findings with the observations made by scholars, frequency of occurrence of basic character types, occurrence of vowel allograph, counted frequency of use of characters at word-initial position, global occurrence of consonant clusters, consonants graphic variants, the number of clusters in words, counted the frequency of use of punctuation marks, and so on.

In Bangladesh, at the Center for Research on Bangla Language Processing (CRBLP), BRAC University, Dhaka has also done some works on Bangla corpus from newspaper 'Prothom-Alo' with some statistical findings including average word length, word level and character level frequency analysis, English words finding in "Prothom-Alo" newspaper as foreign word. Besides these, some academic institutes and universities have also done works on Bangla. Besides this, University of Dhaka, Bangladesh, Shah Jalal University of Science and Technology, Bangladesh, Rajshahi University, Bangladesh also have research-works on Bangla Language Processing including Bangla corpuses.

3. OUR RESEARCH METHODOLOGY

First of all our dictionary corpus is approximately 100% balanced corpus because it is created from general purpose BAD dictionaries, which accommodated texts of different genres of science, medical science, humanitarian, commerce, literatures including novels, stories, text books, sports, journals of different fields, etc. We have collected the text from Bangla to English and English to Bangla dictionaries of Bangla Academy, Dhaka and both of them have the Bengali font and English font all together. So it is a big problem to collect the Bangla and English words at a time because of absent of expert OCR that can recognize the Bangla and English characters simultaneously. Bangla Academic dictionaries have a great acceptance to the Bengali researchers and linguistics. But there is no e-version of these dictionaries. So we were bound to collect the corpus texts manually instead of totally automatically. There are about 150 thousand words are available in our corpus. Since it is a

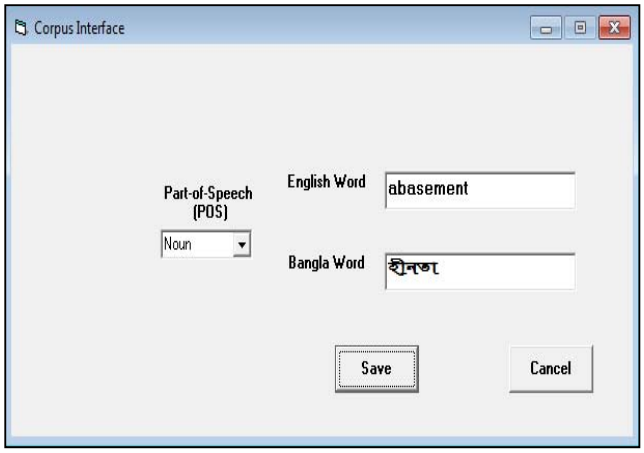


Figure 1: UI of BAD Corpus Compiling

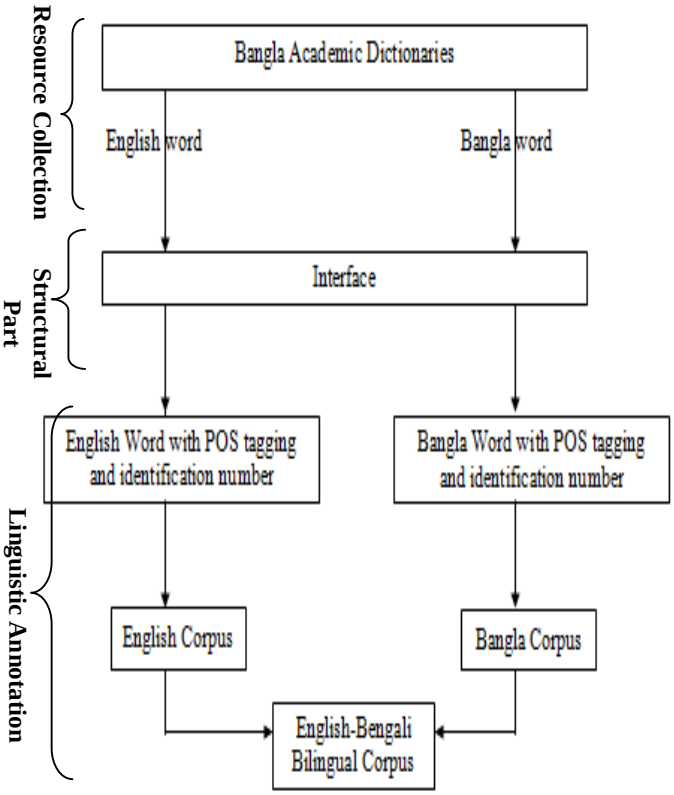
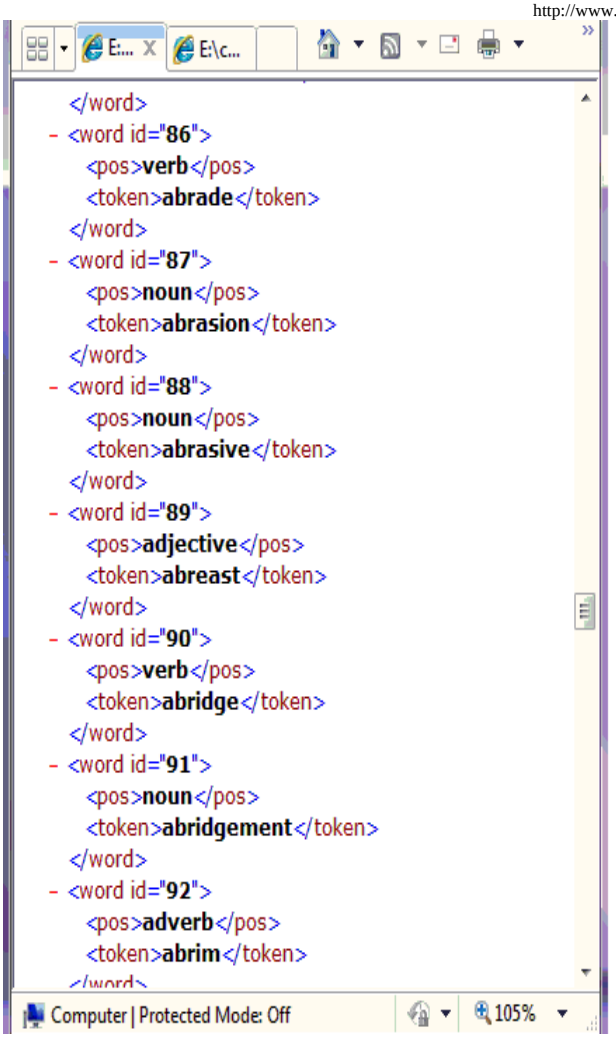


Figure 2: Architecture of BAD Corpus

lexicon corpus, the number of words is not small because the repetition of words is very rare and it is occurred in case of same word spelling for several meanings. For instance, sometimes **sound** in English is in Bangla **শব্দ** or sometimes **sound** in English is in Bangla **সুন্দর** or sometimes **sound** in English is in Bangla **ভালো** etc.. We have also a future plan to enrich the corpus by increasing the number of new words.

4. SYSTEM ARCHITECTURE

Since there is a problem of compiling the BAD corpus due to the absent of smart OCR of extracting English and Bengali texts simultaneously, so we had to compile our corpus using a simple user interface (UI) where we just code the English word and corresponding Bangla word with POS tag. The UI automatically store the English word and Bangla word into two separate files with identification numbers and POS tags. Though the total work was too much time consuming, we had no any alternative way. The corpus compiling process is illustrated in figure 1. The architecture of the BAD corpus is in figure 2. The snapshots of the file formats (XML) are given in figure 3. One of XML files is for English texts and other one is Bangla texts which are shown in alongside.



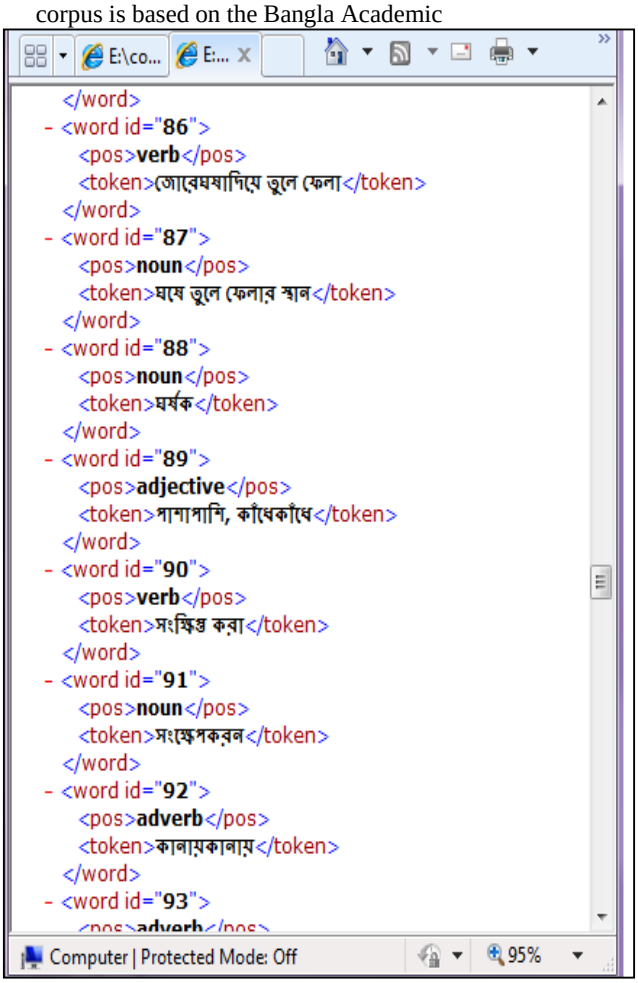


Figure 3: After Compiling, the BAD Corpus in XML Format

5. APPLICATIONS OF BAD CORPUS

5.1. Part-of-speech tagger We can use the BAD corpus to assigns the parts of speech to each word (and other token) of any English and Bangla text, such as noun, verb, adjective, etc.. In our BAD corpus the POS of each token is coded in full form as Noun, Adverb, etc.

5.2. English to Bengali machine translation: Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another. It is a complex task to get the proper translation of target language. One of the important tools of machine translation is a corpus. Our BAD corpus is a word level corpus and each of the English and Bangla word has a corresponding identification number. So we can use the BAD corpus for English to Bangla and Bangla to English machine translation in Rule-Based/Transfer-Based Approach.

5.3. Spell checking and correction Since our BAD corpus is based on the Bangla Academic dictionaries and it has a universal acceptance to Bangla linguistic, so it contains most of the Bangla words of Bangla language. Besides this the academy always research to find new words and their proper spellings. So we can use the BAD corpus for Bangla word spell checking and correction.

5.4. Lexicon dictionary: Since our BAD corpus is a word level corpus and each English word has a corresponding Bangla meaning and POS, so we can also use the BAD corpus as a dictionary.

5.5. Enrich the Bangla Academic Dictionaries using BAD corpus: Our BAD corpus includes about 150 thousand words. But a lot of words of Bangla language which are not included in Bangla Academic dictionaries. One of the ways to include the new words in Bangla Academic dictionaries comparing the BAD corpus with any other Bangla corpuses.

5.6. Finding the numbers of English words in any Bangla written text: Our Bengali language is enriched by the words of different foreign languages including English language. So using the BAD corpus we can find the number of English words that is available in any text document written in Bangla.

6. STATISTICAL ANALYSIS

Every corpus has two types of statistical analysis according to the various methods of statistics. One of them is quantitative analysis and the other is qualitative. Quantitative analysis defines the different linguistic properties and qualitative analysis represents some complete and detailed description of the findings phenomena. Some of the analyses are given below.

6.1. Number of English words in both dictionaries is given separately below in Table 1 with percentage and it is remarkable that the number of English words in both dictionaries is almost same.

Table 1

Dictionary	No of English words	Percentage regarding total words of the corpus
English to Bengali dictionary	692	0.46
Bengali to English dictionary	750	0.5

6.2. Comparing with different English corpuses, the percentage of English words in our corpus are given below in Table 2.

Table 2

Name of English corpuses	No of words	% of English words in E2B dictionary	% of English words in B2E dictionary	% of English words in both dictionaries
American national Corpus	22 million	0.031	0.034	0.066
Collins Word Web	550 million	0.0013	0.0014	0.0026
British National Corpus (BNC)	100 million	0.0069	0.0075	0.0144
Helsinki Corpus	1,572,800	0.044	0.048	0.092
Oxford English Corpus	2 billion	0.00035	0.00038	0.00072
Corpus of Contemporary American English (COCA)	425 million	0.0016	0.0018	0.0034
Bank of English	525 million	0.0013	0.0014	0.0027
Brown Corpus	two million	0.35	0.38	0.72

6.3. Since our corpus is dictionary based, there is a little repetition of words in the whole corpus. But the same object or entity in Bangla has different spelling regarding different factors like utterance variations, foreign utterance, print/electric media, etc. For instance, Acid in English is in Bangla এসিড or অ্যাসিড or এ্যাসিড and such kind of spelling variation is very common in Bangla language.

Some more examples of English words in Bangla spelling in Table3 and this are because of foreign utterance:

Table 3

Words in English	Different Spelling in Bangla
Apple	আপল / অ্যাপল / এ্যাপল
Alcohol	এলকোহল / অ্যাকোহল / আলকোহল

7. CONCLUSION

Since a parallel corpus is an important tool to develop a machine translation system, we should enrich our BAD corpus regarding different perspectives including training with elimination of ambiguity in Bangla language, increasing the vocabulary, etc. A single parallel corpus is not sufficient for a machine translation. So our future plan is to develop and research more and more on corpuses and final goal is to develop a machine translation between English and Bangla languages.

8. REFERENCES

[1]. Niladri Sekhar Dash "Modern Bengali Script: An Introduction",ISBN:81-89803-07-7, India, 2010

[2]. Niladri Sekhar Dash "Corpus Linguistics: An Introduction",ISBN: 8131716031, 9788131716038, India, 2008

[3]. Grinder John T. and Elgin S. Haden "Guide to transformational Grammar: history, Theory and practice." Holt, Rinehart and Winston, Inc., USA, 1973.

[4]. Patterson Dan W. "Introduction to Artificial Intelligence and Expert System"

[5]. Elaine Rich, Kevin Knight "Artificial Intelligence"

[6]. Selim Mohammad Reza and Zafar Muhammad Iqbal "Syntax Analysis of Phrases and Different types of sentences in Bangla." International Conference on Computer and Information technology, SUST, Sylhet, Bangladesh, 1999.

[7]. Azad Humayuun "Baccaw Tatha." Publisher – Dhaka University, December 1994.

[8]. Alfred V. Aho and Jeffrey D. Ullman "Principal of Compiler Design", Narosa Publishing House, India, 1997.

[9].http:/en.wikipedia.org/wiki/transfer-based_machine_translation

[10].http://en.wikipedia.org/wiki/Bengali_language#Vocabulary

[11]. Atkins, Sue; Jereme Clear; and Nicholas Ostler (1992) "Corpus design criteria", Literary and Linguistic Computing, 7(1); 1-16.

[12]. Naushad UzZaman and Mumit Khan "A Double Metaphone Encoding for Bangla and its Application in Spelling Checker", Center for Research on Bangla Language Processing, BRAC University Dhaka, Bangladesh

[13]. Md. Abul Hasnat "Research Report on Bangla OCR Training and Testing Methods", BRAC University, Dhaka, Bangladesh.

[14]. Khair Md. Yeasir Arafat Majumder, Md. Zahurul Islam, Naushad UzZaman and Mumit Khan "Analysis of and Observations from a Bangla News Corpus"

[15]. Asif Iqbal Sarkar, Dewan Shahriar Hossain Pavel and Mumit Khan "Automatic Bangla Corpus Creation", BRAC University, Dhaka, Bangladesh

[16]. British National Corpus, www.natcorp.ox.ac.uk/

[17]. Altaf Mahmud and Mumit Khan "Research Report on Bangla Tagset", Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh.

[18].http://www.lrec_conf.org/proceedings/lrec2010/pdf/13_Paper.pdf

[19].http://aclweb.org/anthology-new/Y/Y04/Y04-1030.pdf.

[20].http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Mamitimin_and_Dawut.pdf