iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XI **Month of publication:** November 2017

DOI: http://doi.org/10.22214/ijraset.2017.11038

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XI, November 2017- Available at www.ijraset.com

Creation and Analysis of a New Bangla Text Corpus BDNC01

Dr. Md. Farukuzzaman Khan[1], Afraza Ferdousi, Dr. M. Abdus Sobhan[3]
[1]*Professor Dept. of Computer Science and Technology Islamic University, Kushtia-7003 Bangladesh.*
[2]*Ph.D. Fellow Dept. of Computer Science and Technology Islamic University, Kushtia-7003 Bangladesh.*
[3]*Professor Independent University, Bangladesh*

Abstract: The past three decades have seen a steady growth of interest in corpus-based techniques for speech and natural language processing throughout the world in almost all popular languages. Hence there are continuous research efforts to use corpus-based methods as dominant part in language and speech processing systems. As a part of this continuing effort worldwide we have created a new Bangla text corpus BdNC01, most part of which is collected from web editions of several influential daily newspapers. We have also created a small scale literary corpus from the literary works of old and modern writers. Nearly twelve million word tokens including literary corpus were collected during six years to avoid time dependency and each part of the corpus was manually checked and corrected errors before preserving in final repository as ASCII and Unicode text. This paper includes the compilation processes of the corpus, analysis of results realized from statistical processing, observation of Zipp's law and developed software tools used for various processing. The paper concludes with the properties and significance of the corpus from the analysis of computed statistics.
Keywords: Corpus, BdNC01, Unicode, Unigram, Prior Probability, Zipp's Law.

I. INTRODUCTION

The creation of language related resources and acquiring knowledge from these resources is unavoidable in research for the progressive development of Language related technology. Corpus is such an important resource for any language. A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research [1]. Two main strengths of the corpus-based approach are identified. Among them text corpora provide large databases of naturally-occurring discourse, enabling empirical analyses of the actual patterns of use in a language and when coupled with automatic computational tools, the corpus-based approach enables analyses of a scope not otherwise feasible [2].

The first computerized corpus of transcribed spoken language was constructed in 1971 by the Montreal French Project, containing one million words [3]. Today, while the construction and exploitation of English language still dominate the field of corpus linguistics, corpora of other languages, either monolingual or multilingual also become available [4]. The history of corpus based analysis of Bangla language is not so far. Though Bangla is an important language with a rich heritage and is spoken by approximately 8% of the world population [5], adequate advancement of research in Bangla corpus linguistics is not achieved yet. From the beginning of this millennium a significant research was continued by N. S. Dash and B. B. Chaudhuri [6-11]. Corpus creation, analysis of corpus and finding various linguistic properties from corpus were included in their effort. Over the past few years various research laboratories of Bangladesh, India and also in other countries are also engaged in this field of research. Thus Bangla corpus research includes an incremental number of researchers in a wide area of horizon [12-15]. These efforts in Bangla corpus research are inadequate in this sense that the result is not able to make expected effect in Bangla language structures like English yet. The statistical properties of a language corpus are very important in language modeling and speech related research like speech recognition. Therefore we have started a research effort by creating and analyzing a corpus to find appropriate linguistic rules applicable to structure based language modeling, machine translation, speech recognition and like other sectors in Bangla language.

To collect a corpus any selection must be made on some criteria and the first major step in corpus building is the determination of the criteria on which the texts that form the corpus will be selected. Common criteria include:

A. The mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode.
B. The type of text; for example if written, whether a book, a journal, a notice or a letter.
C. The domain of the text; for example whether academic or popular.

D. The language or languages or language varieties of the corpus.
E. The location of the texts; for example (the English of) UK or Australia.
F. The date or period of the texts.

There are unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web pages as compared with the labor and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence. We have to identify the instances of language that are influential as models for the population. In regard to practical and recent use of words, ease of collection, volume of text, and varieties of contexts, newspaper is most reliable source for corpus collection. Especially for Bangla language, news paper is such standard source because it covered almost all common criteria listed above. This paper describes the processes undertaken to create and process BdNC01 and the software developed as support tools. At the end the paper concludes about the significance of the work.

G. Sources And Period Of Collection

But considering unavoidable realities such as the relative ease of acquiring public produced Bangla electronic texts as compared with the labor and expense of recording and transcribing private conversations or acquiring and keying books, journals, personal handwritten correspondence etc. It is also important to ensure that the selections are to be influential as models for the population using the language. Regarding practical and recent use of words, ease of collection, volume of text, and varieties of contexts, newspaper is most reliable source for corpus collection and also standard source to cover almost all common criteria listed above. In our observations, the most influential source of texts is available in Bangla newspapers and it also represents almost all types of text required for a corpus. In addition with these facts, because of its easy availability as electronic text in web editions, the newspaper was selected as the source of text collection in this work.

The date or period of text in a corpus have a direct relation to the word frequency and the fact is very influential in news paper corpus. Also words related to an influential fact of a time may appear in news with high frequency. Another fact is that text collected from a particular news paper may provide some abnormal statistics because it may be biased with some particular editing style, workers habit may include some word types frequently, it may have a convention to use a language style or it may be biased with a political or social group etc. All of these obviously may lead the various statistical parameters to the wrong direction with special influence on word types and word frequencies. To avoid the time limitation the texts were collected throughout six years, from October 2004 to October 2010. Also to avoid the source limitation of using the texts from a single newspaper, texts were collected from six newspapers with major text from daily Ittefak, Vorer Kagoz and Jugantor

H. Text Collection And Processing

In the first step of compilation of BdNC01 corpus, some selected contents of Daily Inqilab, Amar Desh and Prothom Alo were typed in manually with MS-WORD software using sutonnyMJ font. But it was requiring lot of labor and time. Then in next step, major text was collected from web editions of the daily Ittefak, Vorer Kagoz and Jugantor. With their popularity and influence in Bangla news media, sutonnyMJ font is used by all these three dailies in their web editions. The contents of these newspapers were selected manually, copied and paste in MS-word document (.doc) files. As a result of continuous effort during six years eleven million word tokens were collected from nearly 350 issues of six national dailies of different dates. The collected texts in this way were contained various format, design, graphics, tables, unwanted symbols etc. These unwanted matters were cleared by editing the document files manually. The files were then stored as plain text file type with txt extension and as Unicode text by converting ASCII text to UTF8 text using software program Avro Converter 0.6.0.

I. the repository of bdnc01

Beyond of master module BdNC01, The total corpus was also divided in eleven modules with variations in length, source and period of collection as shown in Table-1. This was done to help easy statistical processing for various requirements, as corpus length requirement for different researcher may differ regarding their specific problem and it may be helpful to find out time and source dependency of word frequency and other statistical parameters. The repository content includes all the modules as word document with doc extension and in ASCII and UTF8 with txt extension. Therefore the future workers could reform the corpus as required for further development. A standard corpus should include texts from various directions. Table-2 shows the text categories in percentage of total text contains in BdNC01 corpus. To estimate the percentage of text categories, seven issues of each major daily were randomly taken as sample. This means that it is estimated by studying (7x3=21) or twenty one issues of three major dailies Jugantor, Ittefak and Vorer Kagoj.

Table-1: Sources and Size of corpus

File name and types (txt in ASCII/ UTF8)	Sources of texts	No. of tokens	Period of collection
Module01.doc/txt	Daily Jugantor, Inqilab, Amar Desh, Prothom Alo	34338	October, 2004
Module02.doc/txt	Daily Jugantor, Ittefak and Vorer Kagoj	993588	February, 2005 to March, 2008
Module03.doc/txt	Daily Ittefaq and Vorer Kagoj	903546	Jun-August, 2009
Module04.doc/txt	Daily Vorer Kagos	965651	May, 2009
Module05.doc/txt	Daily Ittefak	966761	August, 2009
Module06.doc/txt	Daily Jugantar	1080976	July, 2009
Module07.doc/txt	Daily Vorer Kagos	1038900	July, 2009
Module08.doc/txt	Daily Jugantar	1095114	January-March, 2010
Module09.doc/txt	Daily Ittefak	1565904	January-Feb, 2010
Module10.doc/txt	Daily Vorer Kagos	481590	March, 2010
Module11.doc/txt	Daily Vorer Kagos, Ittefak and Jugantar	2236156	Jun to October 2010
BdNC01.doc/txt	All above six dailies	11362524	October 2004 to October 2010

Table-2: Text Categories in percentage of total texts

Reports (In percentage of total texts)					Others (In percentage of total texts)						
General	International	Business	Sports	Culture	Letters	Article	Academics	Science	Literature	Medical	Religion
47 %	5.4 %	4 %	10.4 %	12 %	3.5 %	6.4 %	2 %	2.7 %	4.4 %	1 %	1.2 %

J. Statistical Computations

Several program modules were developed for corpus processing using C language to clean unwanted marks or characters, parsing and sorting alphabetic characters and words, prior probability calculation etc. To find a list of words from the corpus is to delete unwanted marks and characters in the words. In this process, words are parsed and scanned one after another from the input text file and after cancelling any unwanted character the clean words are stored in a new output file as shown in the following C program segment.

```
While(!feof(fp1))     /*fp1 is the file pointer to read BdNC01.txt*/
{
fscanf(fp1,"%s",word1); /*word1 is string variable*/
n=strlen(word1);
for(i=0;i<n;i++)
if(word1[i]!='>'&&word1[i]!='<'&&word1[i]!='.'&&word1[i]!=','&&word1[i]!=';'&&word1[i]!='('&&word1[i]!=')'&&word1[i]!='|'
&&word1[i]!='?'&&word1[i]!=':'&&word1[i]!='*'&&word1[i]!='!'&&word1[i]!='&')
fprintf(fp2,"%c",word1[i]); /*fp2 is the FILE pointer to write
 clean and parsed words in a new file*/
```

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XI, November 2017- Available at www.ijraset.com

Word Frequency is a measure of how often a word type is encountered in text corpus. A recursive search algorithm finds the word types and count each word type from the text file (corpus) contains the clean text generated by above program segment. The result of the program is a vocabulary list with frequency of each type stored in an output text file. This is also the unigram count for the word type and thus usable for computing N-gram probabilities.

Prior probability is an important statistical parameter extensively used in almost all classification tasks like speech and image processing, DNA typing, remote sensing etc. In natural language processing, the prior probability is frequently used parameter in spelling error correction. Let t represent the misspelled word, and let c range over the set C of candidate corrections. The most likely correction is then:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(t|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}} \qquad \text{Eq. - 1}$$

Prior probability of each correction P(c) can be estimated by counting how often the word c occurs in some corpus, and then normalizing these counts by the total count of all words [16]. So the probability of a particular correction word c is computed by dividing the count of c by the number N of words in the corpus. To avoid zero counts, 0.5 is added to all the counts. Here V represents the vocabulary size.

$$P(c) = \frac{C(c) + 0.5}{N + 0.5V} \qquad \text{Eq. - 2}$$

A program module using above theory was designed and implemented to compute prior probability *P(c)* for each word of the determined vocabulary. The input to the program is a text file contained unigram count and output file stores prior probability together with unigram count as in table-3.

RESULTS AND DISCUSSION

For any corpus one of the first and simplest queries is a list of word forms, which can be organized in frequency order. The frequencies follow Zipf's Law[17], It states that given some corpus of natural language, the frequency of any word is inversely proportional to its rank. It is not known why Zipf's law holds for most languages.

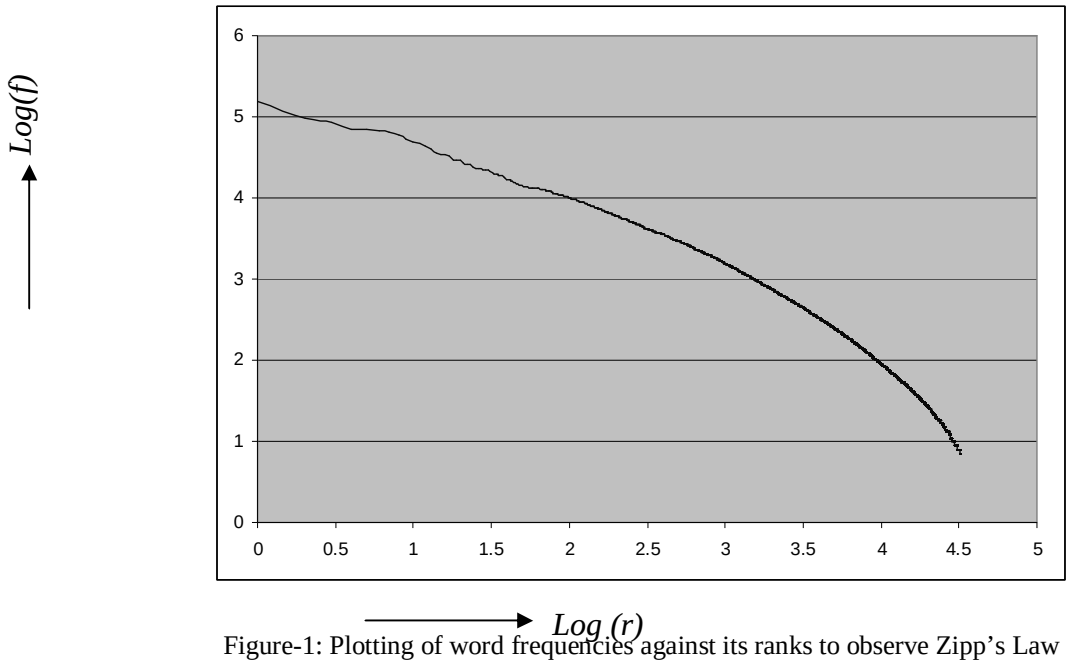


Figure-1: Plotting of word frequencies against its ranks to observe Zipp's Law

Explanations from the work by Wentian Li [18] in the statistical analysis of randomly-generated texts and another work by Ramon Ferrer I. Cancho and Ricard V. Sole [19] may be accepted as standard. But the fact is that the Zipf's curve appears approximately linear on log-log plot for most of the standard corpora. Figure-1 shows the Zipf's curve for BdNC01 corpus and it is almost linear. Zipp's law stated also in other way as about half the word forms in a corpus occur once only, a quarter twice only, and so on [20]. Thus the most frequent word will occur approximately twice as often as the second most frequent word. The final vocabulary size becomes 310483 after manual correction of the program output as it was contained shorthand, wrong spelled words, nonwords, repeated words etc. The commonest word, ও *("o" means "and")* has a frequency of 151919, the next frequent word ("kore" means "do") appears with a frequency of 98271. The top 20 frequent words of frequency-prior probability database are shown in table 3.

Table-3: A part of Frequency-Prior probability database

Rank	words	Frequencies	Prior Probabilies
1	ও	151919	0.0132
2		98271	0.0086
3	এ	85107	0.0074
4		69838	0.0061
5		68858	0.0060
6		67453	0.0059
7	হয়	63512	0.0055
8		59724	0.0052
9		52701	0.0046
10		48345	0.0042
11	জ	47501	0.0041
12	এই	43988	0.0038
13		40123	0.0035
14		35992	0.0031
15		34197	0.0030
16		34172	0.0030
17	এক	32298	0.0028
18		29507	0.0026
19		29490	0.0026
20		29223	0.0025

Table 4: Comparative result of word frequencies

	BdNC01		Prothom-Alo		CIIL		Brown	
Ranks	Words	Percentage	Words	Percentage	Words	Percentage	Words	Percentage
1	ও	1.337%	ও	1.23%		1.15%	The	6.887%
2		0.865%	এ	0.92%		0.99%	of	3.584%
3	এ	0.749%		0.84%	এই	0.94%	and	2.840%
4		0.615%		0.72%	ও	0.91%	to	2.574%
5		0.606%		0.62%	হয়	0.76%	A	2.299%
6		0.594%	হয়	0.57%		0.65%	in	2.101%
7	হয়	0.559%		0.52%	যে	0.65%	that	1.043%
8		0.526%	তার	0.49%		0.55%	is	0.994%
9		0.464%		0.46%	আর	0.51%	was	0.966%
10		0.425%		0.43%	তার	0.50%	He	0.939%

A comparative result is shown for 10 top frequent words from four corpora in table 4.

Two other Bangla corpora the Prothom-Alo corpus [21] and the CIIL Bangla corpus [22] are compared with our BdNC01 corpus. An English corpus, the Brown corpus [23] is also compared for its world acceptance in standardization. The main cause of variation of word frequencies is that these three corpora are designed differently: CIIL Bangla corpus contains every type of texts from various sources to make it balanced, Prothom Alo is a news paper corpus with text collected from only a single newspaper and our corpus is also a newspaper corpus but with text from three major source collected through six years.

Table 5: Effect of sources of texts on word frequencies

Source: Daily Vorer Kagos, Tokens: 965651 (May, 2009)		Source: Daily Ittefak, Tokens: 966761 (August, 2009)		Source: Daily Jugantar, Tokens: 1080976 (July, 2009)		Literary Corpus, Tokens: 476165 (Rabindranath to Humaun Ahmed)	
ও	13187	ও	13284	ও	13137		9571
এ	8609		7721		9188	গানটা	5154
	7492		6025	এ	8668	আমি	3492
	6203		5956		7059	আমার	3334
	5960	এ	5785		6646	আর	3113
	5500	হয়	5597		6387	এই	2919
হয়	5374		5387	হয়	5918	ছে·াট্ট	2789
	4636		5150		5236	দিও	2541
	4010	এই	4637		5110	এসছেলিাম	2392
	3997		4363	তার	4871	ভাবি	2366
তার	3831		4149		4663	কথা	2162
এই	3773		4124		4330	সদেনি	2084
	3735		3433		4293	যতে	2060
	3374		3211	ন	3691	অবসর	2008
	3241	তার	3106		3598	মনে	1954
	3219		2834		3476	সন্দশে	1634
	2933		2766	wনয়িে	3383	করনি	1630
নয়িে	2926	এক	2753	কে·ান	3368	অর্ধাঙ্গনী	1553
কে·ানে·া	2713		2706		3331	দনি	1547
	2675	এর	2517	এক	3205	ভবদীয়	1444

To find the cause of frequency variations of the words in different Bangla corpus a small scale literary corpus was compilled from the literary works of old and modern writers. Table 5 shows the effect of source, type, amount and domain of texts on frequency of words. It is surprising that word using convention in literary corpora is clearly different from newspaper corpora and in this literary corpus the maximum frequent word is as in CIIL. Hence the deficiency of literary texts in BdNC01 and Prothom Alo corpus is a major cause of variation in rank of high frequent words with CIIL as shown in table 5.

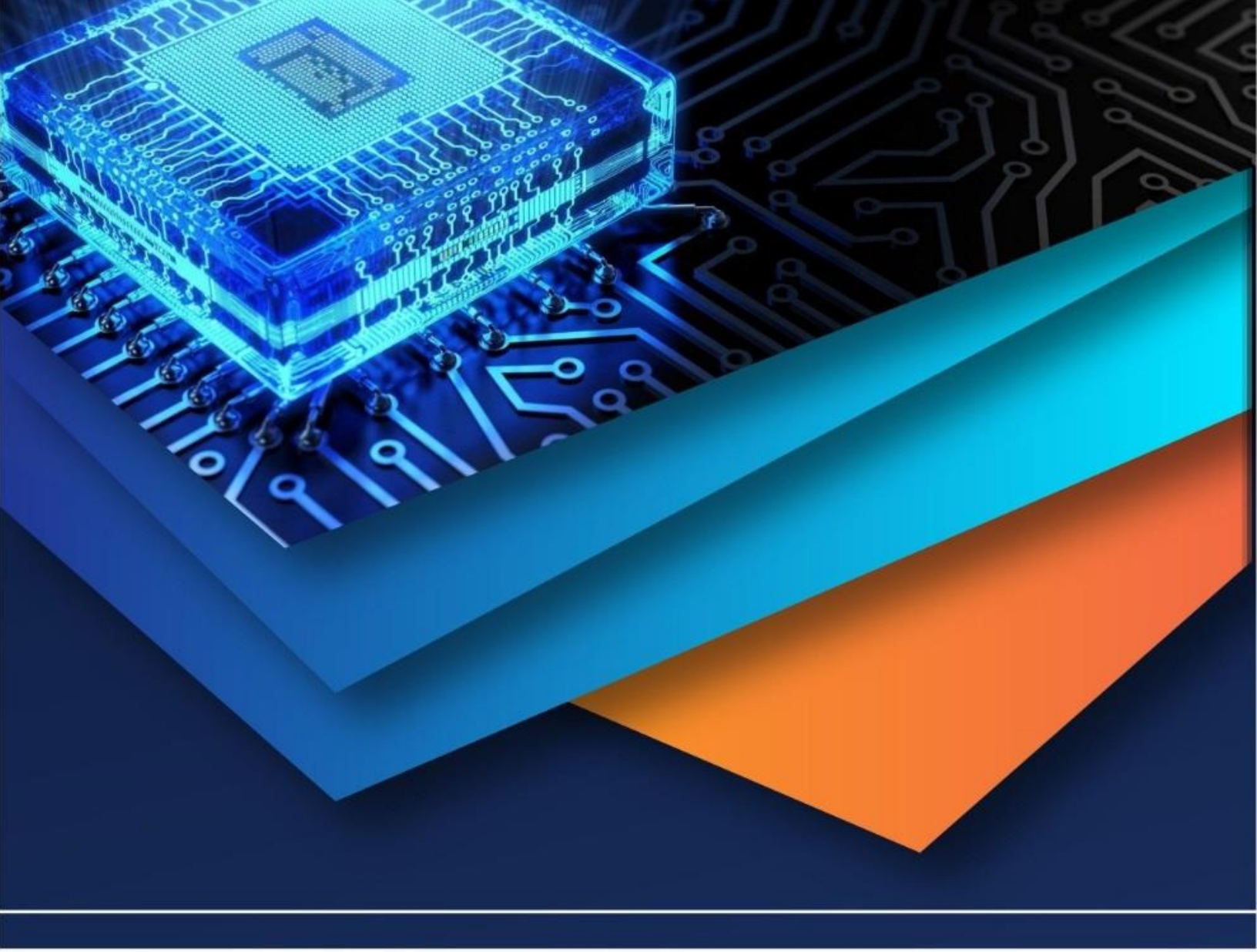
II. CONCLUSION

We have collected more than eleven million word text corpus collected from web edition of Bangla newspapers with an advantage that it reflects the current tradition of a language. Some Statistical processing produced a database which contains unigram count

and prior probabilities that are very useful in many probabilistic computations for language modeling. The database is also useful in works like spelling and structural detection and correction [24]. It is very interesting that the gradient of word frequencies of BdNC01 is more similar to Brown corpus than other similar corpora.

REFERENCES

[1] John Sinclair, Corpus and Text: Basic Principle, Tuscan Word Center, 2004, http:// www.ahds.ca.uk/litangling.
[2] Anthony McEnery and Richard Xiao "Developing Linguistic Corpora: a Guide to Good Practice", Lancaster University, 2004,
[3] Sankoff, D. & Sankoff, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Darnell R. (ed.) Canadian Languages in their Social Context Edmonton: Linguistic Research Incorporated. 1973. 7-64.
[4] Douglas Biber, Susan Conrad And Randi Reppen, "Corpus-based Approaches to Issues in Applied Linguistics", Oxford Journals Humanities Applied Linguistics Volume15, Issue2, Pp. 169-189, Oxford University Press, 1994.
[5] Abul Hasanat, Md. Rezaul Karim, Md. Shahidur Rahman and Md. Zafar Iqbal, "Recognition of Spoken letters in Bangla", 5th ICCIT 2002, East West University, Dhaka, Bangladesh, 27-28 December 2002.
[6] Dash, N.S. (1999) "Corpus oriented Bangla language processing". Jadavpur Journal of Philosophy. 11(1): 1-28.
[7] Dash, N.S. and B.B. Chaudhuri (2001) "A corpus based study of the Bangla language". Indian Journal of Linguistics. 20: 19-40.
[8] Dash, N.S. and B.B. Chaudhuri (2001) "Corpus-based empirical analysis of form, function and frequency of characters used in Bangla". The Proceedings of the Corpus Linguistics 2001 Conference, Lancaster: Lancaster University Press. UK. 13: 144-157. 2001.
[9] Dash, N.S. and B.B. Chaudhuri (2002) "Corpus generation and text processing". Inter ational Journal of Dravidian Linguistics. 31(1): 25-44.
[10] Dash, N.S. and B.B. Chaudhuri "Using Text Corpora for Understanding Polysemy in Bangla". Procedings of the Language Engineering Conference (LEC'02) IEEE, 2002.
[11] Niladri Sekhar Dash, Methods in Madness of Bengali Spelling: A Corpus-based Investigation", South Asian Language Rewiew, Vol. XV, No. 2, June 2005
[12] Md. Jahangir Alam, Naushad UzZaman and Mumit Khan "N-gram based Statistical Grammar Checker for Bangla and English", 9th International Conference on Computer and Information Technology (ICCIT) 2006, Bangladesh, 2006.
[13] Samit Bhattacharya, Monojit Choudhury, Sudeshna, Sarkar, and Anupam Basu. 2005. Inflectional Morphology, Synthesis for Bangla Noun, Pronoun and Verb Systems. In Proc. of the National Conference on Computer Processing of Bangla (NCCPB 05), pages 34 - 43.
[14] Sajib Dasgupta and Vincent Ng, "Unsupervised Word Segmentation for Bangla", Human Language Technology Research Institute, University of Texus, TX 75083,
[15] "The EMILLE Corpus", University of Lancaster and University of Sheffield in collaboration with The Central Institute of Indian Languages, India, 2003, Available at:. http://www.lancs.ac.uk/fass/projects/corpus/emille/, retrieved on 14th January, 2011.
[16] Daniel Jurafsky and James H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall USA, September 28, 1999, pp 139-232.
[17] Wikipedia, "Zipfs Law", available at: http://en.wikipedia.org/wiki/Zipf%27s_law, retrieved on 21th March., 2011.
[18] Wentian Li, "Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution", IEEE Transactions on Information Theory **38** (6): 1842–1845, Website: http://www.nslij-genetics.org/wli/pub/ieee92_pre.pdf., Retrieved on 1st May 2012.
[19] Ramon Ferrer i Cancho and Ricard V. Sole (2003), "Least effort and the origins of scaling in human language", Proceedings of the National Academy of Sciences of the United States of America **100** (3): 788-791, Website: http://www.pnas.org/content/100/3/788.abstract?sid=cc7fae18-87c9-4b67-863a-4195bb47c1d1 , Retrieved on 1st May 2012.
[20] John Sinclair, "Corpus and Text: Basic Principle", Tuscan Word Center, 2004, http:// www.ahds.ca.uk/litangling, retrieved on 14th January, 2011.
[21] Khair Md. Yeasir Arafat Majumder, Md. Zahurul Islam, and Mumit Khan, "Analysis of and Observations from a Bangla News Corpus", Website: http://www.panl10n.net/english/final%20reports/pdf%20files/Bangladesh/BAN03.pdf, Retrieve on 1st May 2012.
[22] Dash, N. S., Corpus linguistics and Language Technology, pp 94, Mittal Publications, New Delhi, India, 2005.
[23] "The first 2000 most frequent words from the Brown Corpus", Website: http://www.edict.biz/lexiconindex/frequencylists/words2000.htm, Retrieve on 1st May 2012.
[24] Nur Hossain Khan, Md. Farukuzzaman Khan, Md. Mojahidul Islam, Md. Habibur Rahman, Bappa Sarker, "Verification of Bangla Sentence Structure using N-Gram", Global Journal of Computer Science and Technology, Vol 14, No 1-A (2014).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY