International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume–3 No.-3, July 2012 – www.ijais.org

OSERVATIONS FROM STATISTICAL PROCESSING OF BdNC01 CORPUS

Md. Farukuzzaman Khan
Dept. of Computer Science and Technology
Islamic University
Kushtia-7003, Bangladesh.

M. Abdus Sobhan
Dept. of Electrical, Electronics and Telecommunication Engineering , Independent University Bangladesh
Bashundhara, Dhaka-1229, Bangladesh

ABSTRACT

Recent trends in the development of language related technology finds unavoidable requirement of relevant resources and acquiring knowledge from these resources. In this prospect corpus-based methods are getting strong push from various laboratories throughout the world in Bangla language processing. In this paper we have discussed the compilation of BdNC01 corpus and observations from statistical processing of it. BdNC01 is a new Bangla text corpus collected form web edition of several influential Bangla daily newspapers containing more than eleven millions word tokens. Several processing like list and total count of vocabulary, individual word frequencies and prior probabilities were computed and preserved in final repository. The word frequency relation to Zipp's law, time and source dependency of word frequencies and character distribution were also observed. Software support tools required for various processing were implemented using C language. The paper concludes with the usability of the corpus and computed statistical database.

General Terms

Corpus Linguistics, Statistical Processing and Language Technology.

Keywords

Corpus, Vocabulary, word frequency, prior probability, Zipp's law and character distribution.

1. INTRODUCTION

Corpus may be defined as a collection of pieces of language text in electronic form, selected according to some external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research [1]. 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]. 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 [3].

Bangla is one of the most widely spoken languages of about 245 million people around the world, from the United States to the Middle East. The majority of Bangla speakers are located in Bangladesh and west Bengle of India with a diversity of cultural and linguistic traditions. The initial step to produce Bangla corpus was in 1991 when Technology Development for Indian Languages (TDIL) program was started to develop linguistic corpora in almost all Indian languages including Bangla [4]. After then, from the beginning of this millennium a significant research was continued by N. S. Dash and B. B. Chaudhuri [5-12]. 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 [13-18]. These efforts in Bangla corpus research are inadequate in this sense that the result is not able to make expected improvement 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.

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 the mode of texts, the types of texts, the domain of texts, the language or languages or language varieties of the corpus, the location of texts, the date or period of texts etc. 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 in compilation and processing BdNC01. The paper also discusses various observations from the results of statistical processing. At the end the paper concludes about the significance of the work..

2. COMPILATION OF BDNC01

In the compilation of BdNC01 corpus, the major texts were collected manually from web edition of three influential dailies, the daily Ittefak, Vorer Kagoz and Jugantor due to the favor of font matching. Though some contents of Daily Inqilab, Amar Desh and Prothom Alo were also included by manual typing. Table-1 shows the final repository, contains texts preserved as 11 modules, each of different times and sources. Also all the modules were preserved together in one file BdNC01. The repository content includes all the modules as word document, in ASCII with txt extension and in UTF8 format, so that the future workers could reform the corpus as required for further development. The time dependency of text in a corpus have a direct relation to the word frequency and the fact is very influential in newspaper corpus. Words related to an influential fact of a time may appear with high frequency in that time. 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 three.

A standard corpus should include text from various directions. Table-2 shows the percentage of text categories included in the corpus. The modules in the repository were designed so that it can help easy statistical processing in various direction of the corpus, as corpus length requirement for individual researcher may differ to shoot their current problem and it may be helpful to find out time and source dependency of word frequency and other statistical parameters.

Table-1: Structure and properties of corpus BdNC01

File name/type	Sources	No. of tokens	Period of collection
Module01/doc/txt/utf8	Daily Jugantor, Inqilab, Amar Desh, Prothom Alo	34338	October, 2004
Module02/doc/txt/utf8	Daily Jugantor, Ittefak and Vorer Kagoj	993588	Feb., 2005 to March, 2008
Module03/doc/txt/utf8	Daily Ittefaq and Vorer Kagoj	903546	Jun-August, 2009
Module04/doc/txt/utf8	Daily Vorer Kagos	965651	May, 2009
Module05/doc/txt/utf8	Daily Ittefak	966761	August, 2009
Module06/doc/txt/utf8	Daily Jugantar	1080976	July, 2009
Module07/doc/txt/utf8	Daily Vorer Kagos	1038900	July, 2009
Module08/doc/txt/utf8	Daily Jugantar	1095114	January-March, 2010
Module09/doc/txt/utf8	Daily Ittefak	1565904	January-Feb, 2010
Module10/doc/txt/utf8	Daily Vorer Kagos	481590	March, 2010
Module11/doc/txt/utf8	Daily Vorer Kagos, Ittefak and Jugantar	2236156	Jun to October 2010
BdNC01/doc/txt/utf8	All above six dailies	11362524	Oct. 2004 to Oct. 2010

Table-2: Text Categories in percentage of total text

Reports					Others						
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 %

3. VOCABULARY IN FREQUENCY ORDER

A lot of software tools were designed and implemented using C language in various steps of processing the corpus. One of the tools was to clean the collected corpus, parse the text in to words, determine the vocabulary size and count individual words. The input to the program was the ASCII file of main module BdNC01 and output was a list of words with their frequencies. The actual output of the program was contained some error. With the meaningful words, the resulting vocabulary was also contained various marks, space deletions, special characters, wrong spelled words, meaningless or nonwords etc. Therefore a lot of time was given to manual correction of the program output using MSEXCEL with a hope that the resource build in this work may be useful as much as necessary for future research workers. In the actual output, the vocabulary size was 322871 which corrected to 310483 by manual processing. The commonest word, I *(means "and")* has a frequency of 151909, which is lower but not too lower than twice as common as the next one, K‡i "kore" (means "do"), at 95271..

4. PRIOR PROBABILITY

From the corrected output of the program discussed above, top 20 high frequent words are shown in table-3. The developed program also calculates prior probability [19], a very essential parameter for many probabilistic computations especially in works like spelling error correction. 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 N of all words.

Table-3: Words are organized in Frequency order with Prior probabilities

Rank	words	Frequencies or unigram count	Prior Probabilies
1	I	151919	0.0132
2	K‡i	95271	0.0083
3	G	85107	0.0074
4	†_‡K	69838	0.0061
5	Kiv	68858	0.0060
6	bv	67453	0.0059
7	nq	63512	0.0055
8	Ges	59724	0.0052
9	n‡q‡Q	52701	0.0046
10	n‡e	48345	0.0042
11	Rb¨	47501	0.0041
12	GB	43988	0.0038
13	e‡jb	40123	0.0035
14	Ki‡Z	35992	0.0031
15	GKwU	34197	0.0030
16	K‡ib	34172	0.0030
17	GK	32298	0.0028
18	m‡½	29507	0.0026
19	n‡q	29490	0.0026
20	g‡a¨	29223	0.0025

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.

$$P(c) = \frac{C(c) + 0.5}{N + 0.5V}$$

5. ZIPF'S LAW

It is not known why Zipf's law [20] holds for most languages. Explanations may come from the work by Wentian Li [21] in the statistical analysis of randomly-generated texts and another work by Ramon Ferrer i Cancho and Ricard V. Sole [22]. 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. A comparative analysis of high frequent words is presented in table-4. The word frequencies resulting from BdNC01 corpus were shown with Prothom-Alo news corpus, CIIL Bangla corpus [23] and the Brown corpus [24]. The analysis shows that the gradient of word frequencies of BdNC01 is more similar to Brown corpus.

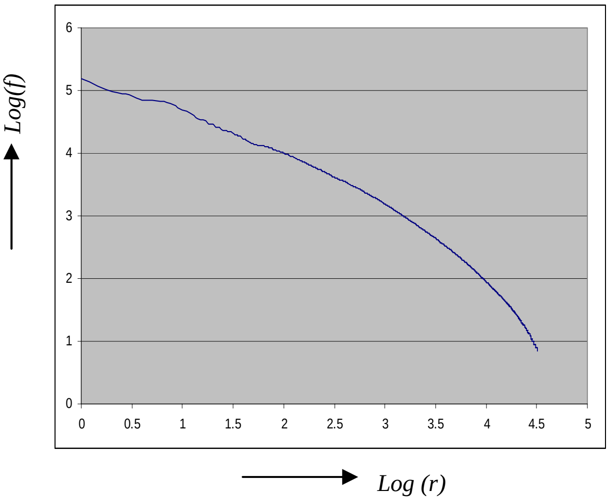

Figure-1: Plotting of word frequencies against its ranks to observe Zipp's Law

6. WORD FREQUENCY VARIATIONS

The word frequencies as percentage of total texts for some selected words are shown in table-5 to demonstrate the effect of time and source variation on word frequency. Six modules including the main module BdNC01 with different source, duration and volume of texts were taken into consideration. The result is also shown as bar diagram in figure-2. It is seen in figure-2 that two words Dc‡`óv (means advisor) and `bxwZ (means corruption) appears higher and other two words gšx (means minister) and gnv‡RvU (means big-alliance) appears lower than the other words in Module02. The reason is that these four words was get influence from political situation in the duration of text collection for Module02, February 2005 to March 2008. In that time, Bangladesh was ruled by a caretaker government composed of ten advisors and thus there were no ministers in the country. The major activity of the government was against the corruption. The word gnv‡RvU is the name of a big political alliance in Bangladesh and it was not influential before the national pole on December 2008. The word Kwiqv (means after doing) is from a style of Bangla language called shadhu which is now obsolete. One of our sources daily Ittefaq was published using this style from its beginning and stopped before 2005. But it continues using shadhu only in their editorials. Other news papers in 21st century are not using shadhu style. Therefore the word Kwiqv is source dependent and appeared with considerable frequency only in the modules where the Ittefaq is included. The frequency of other two words mKvj (means morning) and †i‡L (means after putting) appears as almost stable against time and source variation.

Table-4: Comparative result of word frequencies

	BdNC01		Prothom-Alo		CIIL		Brown	
Ranks	Words	Percentage	Words	Percentage	Words	Percentage	Words	Percentage
1	I	1.337%	I	1.23%	bv	1.15%	The	6.887%
2	K‡i	0.838%	G	0.92%	K‡i	0.99%	of	3.584%
3	G	0.749%	K‡i	0.84%	G	0.94%	and	2.840%
4	†_‡K	0.615%	bv	0.72%	I	0.91%	to	2.574%
5	Kiv	0.606%	†_‡K	0.62%	nq	0.76%	A	2.299%
6	bv	0.594%	nq	0.57%	Ges	0.65%	in	2.101%
7	nq	0.559%	Kiv	0.52%	GB	0.65%	that	1.043%
8	Ges	0.526%	Zvi	0.49%	†_‡K	0.55%	is	0.994%
9	n‡q‡Q	0.464%	Ges	0.46%	Avq	0.51%	was	0.966%
10	n‡e	0.425%	n‡q‡Q	0.43%	Zvi	0.50%	He	0.939%

Table-5: Time and Source dependency of word frequencies

	BdNC01 All sources, Oct. 2004 to Oct. 2010, 11362524 tokens	**Module02** Jugantor, Ittefak and V. Kagoj, Feb. 2005 to March 2008, 993588 tokens	**Module04** V. Kagos, May 2009, 965651 tokens	**Module05** Ittefak, August 2009, 966761 tokens	**Module06** Jugantar, July 2009, 1080976 tokens	**Module11** V. Kagos, Ittefak and Jugantar, Jun to Oct. 2010, 2236156 tokens
Dc‡`óv	0.0218	0.0682	0.0191	0.0177	0.0216	0.0125
`bxwZ	0.0090	0.0243	0.0065	0.0076	0.0158	0.0048
gšx	0.0275	0.0188	0.0294	0.0374	0.0247	0.0280
gnv‡RvU	0.0088	0.0002	0.0084	0.0043	0.0095	0.0089
Kwiqv	0.0067	0.0085	0.0013	0.0191	0.0009	0.0088
mKvj	0.0370	0.0313	0.0362	0.0389	0.0353	0.0386
†i‡L	0.0369	0.0376	0.0374	0.0430	0.0290	0.0379

7. CHARACTER DISTRIBUTION

The character level analysis results of the corpus are presented in tables-6 and 7 from the output of a C program. Table-6 shows of use of characters in the BdNC01 corpus. Total no. of character positions in the corpus is more than 70 millions and maximum count is 10.93% for the character 'v'. Distribution of all of the characters usually used to create Bangla words are presented here. The usual presentation of Bangla alphabet includes from A to , but from to in table-6 has an effect of moderation of the utterances mostly of consonants. The character is usually used to nasalize the utterance of any letter but others have the effect of moderation of only the consonant letters. For example, the utterance of letter K is K+A = K (k+ə), but when w is added with K then its utterance becomes K+B = wK (k+i). Thus the characters from to may be called moderators. Thus the corpus contents 39.47% letters and 28.15% moderators. Within the letters, 4.01% are vowels and 35.46% are consonants. Table-7 shows what we found by analyzing the characters that start a word. This type of analysis can suggest about the preferences of the native language users and that the number of word class will be 43 in designing a lexical classifier depending on initial character. From the results in both tables, it is seen that consonants dominating the vowels and the presence of large number of consonants has a great effect in formation of words in Bangla.

8. CONCLUSION

A new text corpus is created with more than eleven million words from web edition of Bangla newspapers that follow Zipf's law. Though text from academic, literature, medical, science, religion etc. were included in the corpus but the amounts were not satisfactorily sufficient as compared to the general reports. If the amount of these text increases rather than reports, the influence of words from political and other popular issues of news media will be decreased to make the corpus more standard. The statistical processing of the corpus produced a database and observations from the processing prove its usability in technology development in Bangla language. Various language properties, Parameters for many probabilistic computations like language modeling or error correction, selection of a representative list of words for construction of speech corpus are examples of its feasibilities in technology applications.

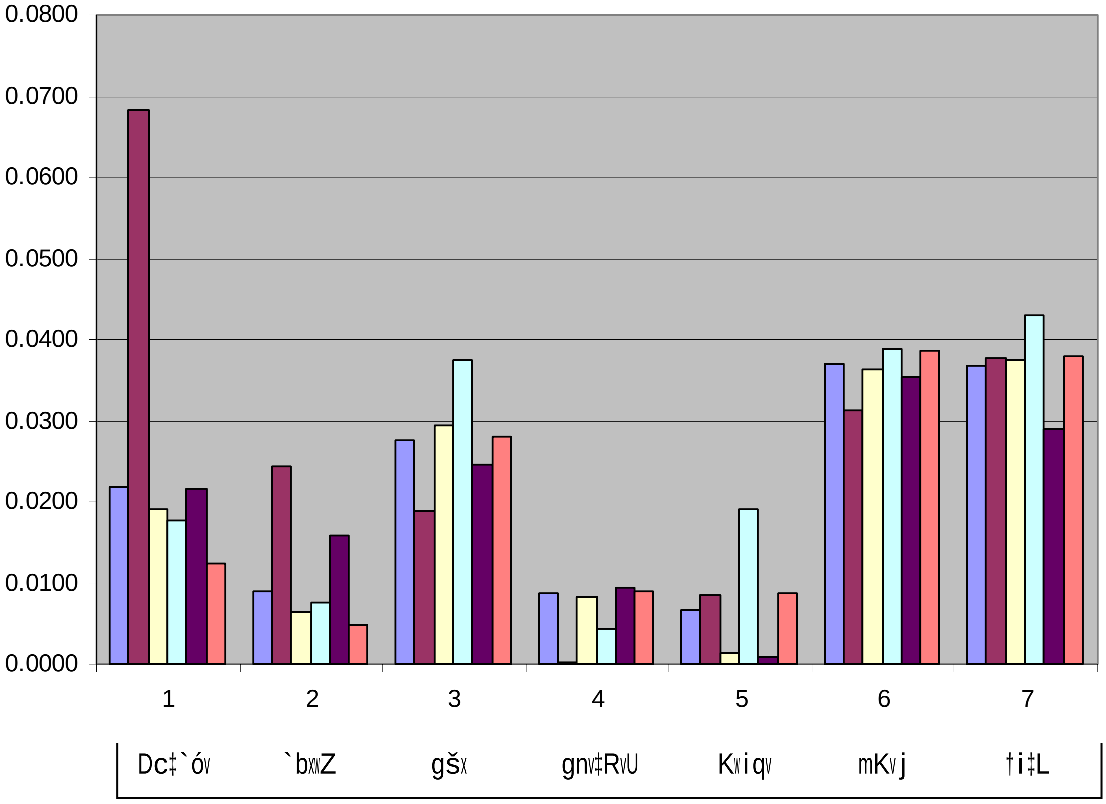

Figure-2: Time and Source dependency of word frequencies

Table 6: Character distribution in BdNC01 Corpus.

Character	Percentage	Character	Percentage	Character	Percentage	Character	Percentage
A	0.5359	Av	0.8018	B	0.7587	C	0.0062
D	0.4056	E	0.0039	F	0.0088	G	0.8885
H	0.0128	I	0.5855	J	0.0004	K	3.4464
L	0.4821	M	0.8731	N	0.1418	O	0.0167
P	0.6130	Q	0.8089	R	1.2028	S	0.0379
T	0.0011	U	0.9763	V	0.1215	W	0.2438
X	0.0419	Y	0.3705	Z	2.3963	_	0.4277
a	0.5027	b	3.9287	c	1.9856	d	0.3327
e	3.0643	f	0.5593	g	2.3062	h	0.4943
j	2.3793	k	1.0701	l	0.2447	m	2.2425
n	1.3130	o	0.3147	p	0.0035	q	1.9314
r	0.0745	s	0.4700	t	0.0418		0.1075
v	10.9279	w	4.9642	x	0.9319	† ‡	8.1798
	0.2419		0.1456		0.0063	¨	0.9417
	0.7742	Š	0.0700	ˆ ‰	0.0761		0.1847
	0.0067						

Table 7: Distribution of words according to starting character in BdNC01 Corpus.

Character	Percentage	Character	Percentage	Character	Percentage	Character	Percentage
A	3.234%	Av	4.706%	B	0.969%	C	0.021%
D	1.704%	E	0.022%	F	0.049%	G	4.871%
H	0.077%	I	1.852%	J	0.006%	K	6.889%
L	0.675%	M	1.406%	N	0.449%	P	1.261%
Q	0.436%	R	2.661%	S	0.081%	U	0.529%
V	0.020%	W	0.315%	X	0.229%	Z	2.366%
_	0.479%	`	1.856%	a	0.498%	B	2.015%
C	6.886%	D	0.570%	e	5.330%	F	0.945%
G	3.810%	H	1.220%	i	1.841%	J	0.703%
K	0.983%	L	0.024%	m	6.568%	n	6.568%
w	9.765%	†	11.201%	ˆ	0.320%		

9. REFERENCES

[1] John Sinclair, Corpus and Text: Basic Principle, Tuscan Word Center, 2004, http:// www.ahds.ca.uk/litangling, retrieved on 6th Jan., 2011

[2] Anthony McEnery and Richard Xiao "Developing Linguistic Corpora: a Guide to Good Practice", Lancaster University, 2004,

[3] 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.

[4] Niladri Sekhar Dash, Language Corpora: Present Indian Need, Indian Statistical Institute, Kolkata, available at: http://www.elda.org/en/proj/scalla/SCALLA2004/dash.pdf, retrieved on 6th Jan., 2011.

[5] Dash, N.S. (1999) "Corpus oriented Bangla language processing". Jadavpur Journal of Philosophy. 11(1): 1-28.

[6] Dash, N.S. (2000) "Bangla pronouns - a corpus based study". Literary and Linguistic Computing. 15(4): 433-444.

[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". Published in Rayson, P., Wilson, A., McEnery, T., Hardie, A., and Khoja, S., (eds.) Special issue of 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 (2002) "Spelling variation of words in Bangla: a corpus-based study". To appear in International Journal of Dravidian Linguistics.

[11] 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.

[12] Niladri Sekhar Dash, Methods in Madness of Bengali Spelling: A Corpus-based Investigation", South Asian Language Rewiew, Vol. XV, No. 2, June 2005

[13] 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

[14] M S A Chowdhury, N M M Uddin, M Imran, M M Hassan and M. E. Haque, "Part of Speech Tagging of Bangla Sentence", 7th International Conference on Computer and Information Technology (ICCIT) 2004, Bangladesh, 2004.

[15] 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.

[16] 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.

[17] Niladri Sekhar Dash. 2006. The Morphodynamics of Bengali Compounds decomposing them for lexical processing. In *Language in India (www.languageageinindia.com)*, Vol 6:7.

[18] Sajib Dasgupta and Vincent Ng, "Unsupervised Word Segmentation for Bangla", Human Language Technology Research Institute, University of Texus, TX 75083,

[19] 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.

[20] Christopher D. Manning, Hinrich Schütze "Foundations of Statistical Natural Language Processing", MIT Press (1999), ISBN 978-0262133609, p. 24

[21] Wentian Li (1992). "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 at 8:30 AM.

[22] 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 at 8:30 AM.

[23] 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, at 8:30 AM

[24] The first 2000 most frequent words from the Brown Corpus, Website: http://www.edict.biz/lexiconindex/frequencylists/words2000.htm, Retrieve on 1st May 2012 at 8:30 AM.