Analysis of and Observations from a Bangla News Corpus

Khair Md. Yeasir Arafat Majumder, Md. Zahurul Islam, Naushad UzZaman and Mumit Khan
Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh
tanipotro@yahoo.com, zahurul@bracu.ac.bd, naushad@bracu.ac.bd, mumit@bracu.ac.bd

Abstract

In this paper we present the compilation methodology and some statistical analysis on a Bangla news corpus-"Prothom-Alo", which is the first of its kind for Bangla. We compare some of the statistics with the CIIL Bangla corpus and also present our observation of atypical behavior of Zipf's curve for Prothom-Alo corpus.

1. Introduction

A Corpus from linguistic point of view is defined as a collection of transcribed speech or written text compiled mainly to enhance linguistic research. It is as important a resource as any other in the field of language engineering. With the recent advancement in computer technology the availability of language corpora (by corpora we mean corpus) and its processing has become even easier and has opened many new areas of research in language processing. The intuitive language study was put against strong challenge by the results obtained from the analysis of a corpus and in many cases the intuitive study was proven wrong [1]. A corpus can be the best resource to study many different linguistic phenomena such as the spelling variations, morphological structure, and word sense analysis and how the language has evolved over the time and many more. The lack of corpus data severely limits the ability of the language engineers to develop language processing applications. 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, ma-chine translation, question-answering, stemming etc. Thus the well proven utility of the corpus has made many languages to create one of their own. The first ever corpus is the Brown corpus of American English which was created by W. Nelson Francis and Henry Kucera (1964) and since then many English corpus as well as corpus for Chinese, Japanese, Spanish has been compiled and analyzed to enrich the language knowledge [2]. Bangla is one of the 7th most widely spoken languages in the world with more than 180 (million) native speakers worldwide [3]. Bangla is unique in its characteristics and diverse in its grammatical constructions and phonetic variations it possesses as well. One of the motivations of this work was the fact that even being a language of so many people and a having a rich literature history Bangla was lacking the most important re-source for language engineering tasks - mainly because very few resources are available electronically.

2. Previous work

In comparison to other major languages like English, Chinese, Spanish etc. research on Bangla is a journey to the recent past. Natural Language engineering is much more of a new story in this part of the world. This is because advancement in computer technology took longer to have noticeable effect in this region than the western world. The project for first Bangla corpus generation was initiated in 1991 and closed in 1995 by department of electronics (DOE), Govt. of India [1] and was created by Central Institute of Indian languages (CIIL) and by then the first electronic corpus, i.e.; the Brown corpus ages more than 25 years. Since then this corpus of three million words has been delivering much of the linguistic data required by the scholars working on Bangla. The book of "Corpus linguistics and Language Technology" by N. S. Dash is a warehouse for corpus related studies with special attention to Bangla, where he has discussed almost every linguistic features of this language and the study is supported by data from the CIIL corpus. Bharati, Sangal and Bendre (1998) analyzed frequency distribution, common word comparison between Bangla and other seven Indian languages [4]. But due to the differences in the writing style as well as the phonetic structure between Indian and Bangladeshi Bangla we decided to compile a Bangla corpus, "Prothom-Alo" news corpus which is the first of its kind for Bangladeshi Bangla. In section 3 we shall study briefly why we came up with a news corpus and also the methodology of compilation. But we shall shade light more on some basic analysis on our corpus, which is discussed in section 4. Then in

section 5 we study the behavior of Zipf's curve for the Prothom-Alo news corpus.

3. Compilation of Prothom-Alo corpus

Our first thought was to design a balanced corpus for Bangla, which will accommodate texts of different genres like scientific, medical, humanitarian and newspaper articles, samples from novels, stories, textbooks as well as transcribed speech, so as to make it representative of every linguistic phenomena of Bangla. This was an ambitious project mainly because of the scarcity of texts available in electronic format. Then again we are lacking good Bangla OCR applications as well as sophisticated, state of the art tools and applications for transcribing speech. For these reasons we turned our attention to create a corpus from whatever resources we have, mainly we focused on the texts available in the web. Unfortunately, texts in Bangla scripts are very rare in the web, although the need for it is increasing day by day. However, we have a couple of newspapers that have their web versions and we decided to collect these in order to create a news corpus. This decision reduced the complexity against all the odds and made life much easier. As one can smell a sense of urgency here and surely there was because we were so much in need of a corpus to take our research on language processing tasks further. There are 18,067,470 (eighteen million plus) word tokens and 386,639 distinct word types in this corpus. The corpus has been created in two phases: collection of the raw text from the Prothom-Alo website, and the conversion to Unicode.

3.1. Collection of text

Prothom-Alo is currently the most widely read newspaper in Bangladesh. Although there are other newspapers with daily web versions, we choose Prothom-Alo mainly for one reason- this is one with less spelling mistakes and with conventional spelling of Bangla words. The diagram in Figure 1 shows the structure of Prothom-Alo corpus.

The raw text for the corpus was collected from the Prothom-Alo home page www.prothom-alo.com. This was done using a web crawler program that surfed through the website of Prothom-Alo and downloaded all the news available for the year of 2005 (from 1st January to 31st December) - including magazines and periodicals, which were all in html format. The process took about twelve hours. Then using a Linux shell scripts with library reference to Lynx, the texts were extracted from the HTML files. At this point we ended up with news of three hundred and sixty five days with each day having several text files that contained news of different genres. For better management and research needs we merged the news of same category of the whole year in one text file. The end product was twenty-seven text documents making up one big news corpus. This made each document of the corpus big enough to represent a particular news category to help research on applications like automatic text categorization yet small enough to be processed efficiently. The corpus size is three hundred and eighteen mega bytes and it was made available as one single text file also.

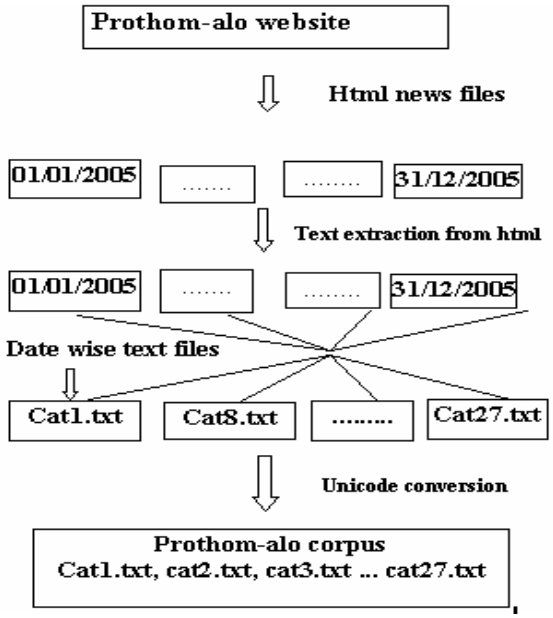

Figure 1: Compilation steps of Prothom-Alo corpus

3.2. Unicode Conversion

The second phase in creating the corpus was to convert all the texts to large body of Unicode. There are several reasons behind this. The main reason is Unicode is universally excepted script that can be displayed and processed without hassle in all platforms. The second problem comes with fonts. It is possible to find fonts in al-most all languages now, even African and Indian languages [3]. A good-quality font, used with a word processor such as Microsoft Word is a general solution. But font-specific encodings are not desirable choices because different

fonts may encode the same language differently, may limit the scope for multilingual documents, and typically requires the use of specialized keymaps. The large number of diacritics and consonant clusters in Indic languages pose a challenge as well. To overcome all these problems decision was made to convert the texts collected from the web to Unicode. Prothom-Alo uses two fonts, namely "Bangsee Alpona" and "Prothoma", both true type fonts (TTF), for the on line version of the newspaper [5]. The previous was in use up to the May of 2005 and after that latter has been in use. The encoding of all these fonts are maintained in different files, which are later required. Two font specific converters developed in Java were used to convert the collected texts files to Unicode texts.

4. Statistical analysis

Regardless of the size of the corpus, it may be subjected to both qualitative as well as quantitative analysis using various methods of statistics [1]. Both these types of corpus analysis have different perspectives. Quantitative analysis focuses on classifying different linguistic properties where as qualitative analysis aims to give some complete and detailed description of the observed phenomena. We wish to focus on some simple quantitative analysis.

4.1. Word level frequency analysis

Study of frequency calculation can provide important information about the usage of words in a text. Although it depends on the domain of the text, but given a balanced corpus it can be figure out which of the words are generally most frequent and which not. But before discussing our findings of the Prothom-Alo corpus we need to define tokens. A token is defined as a sequence of characters which has a white space character, brackets, braces or punctuation marks as the token boundary. In our calculation we have excluded punctuation marks, quotation marks, braces, brackets etc. to be counted in or as a token. So the Bangla sentence - ("তোমাকেই খুঁজছে বাংলাদেশ"), প্রতিযোগিতার গানগুলো এখন বাজারে। - has 7 word tokens namely তোমাকেই, খুঁজছে, বাংলাদেশ, প্রতিযোগিতার, গানগুলো, এখন and বাজারে. After the initial calculations we found that there are 18100378 (eighteen million plus) word tokens and 384048 (three lacks and eighty four thousand and eighty-four) distinct word types in Prothom-Alo corpus. Table 1 and Table 2 show the ten most frequent words in the Prothom-Alo corpus and the CIIL Bangla corpus respectively [1].

Table 1: Top ten most frequent words in the Prothom-Alo corpus

Word	Percentage	Word	Percentage
ও	1.23%	**হয়**	0.57%
এ	0.92%	**করা**	0.52%
করে	.084%	**তার**	0.49%
না	.072%	**এবং**	0.46%
থেকে	0.62%	**হয়েছে**	0.43%

Table 2: Top ten most frequent words in the CIIL corpus

Word	Percentage	Word	Percentage
না	1.15%	**এবং**	0.65%
করে	0.99%	এই	0.65%
এই	0.94%	**থেকে**	0.55%
ও	0.91%	আর	0.51%
হয়	0.76%	**তার**	0.5%

4.1.1. Behavior of function words. Function words are the small and closed set of words that mark grammatical structure rather than referring to something concrete [6]. While the functions words are not interesting in the context of information retrieval because of their relative high frequencies, these words can provide important information while assessing the quality of a corpus. As A. Sarkar, A. De Roeck and P. Garthwaite (2005) discuss that, in a balanced corpus function words will appear to be most frequent and are likely to be distributed more homogeneously than the content words [7], whose occurrence is "bursty" (Katz 1996, Church 2000) - that is if reasonably enough sample is taken from a corpus and the most frequent words in that body of text are found not to be function words than that corpora may need closer inspection in order to make it balanced and representative. Based on this theory, we investigated the behavior of very frequent words. As mentioned earlier, Prothom-Alo has news of twenty-seven distinct categories. To divide the corpus in to two chunks, we randomly choose fourteen categories for the first chunk and the rest made up the second. At this point a frequency analysis over the two chunks reveals that in both the part the ten most frequent words are almost same. The only differences were in their ranks. The results are shown in Table 3.

Table 3: Numerically sorted word list of two equal sized chunks of Prothom-Alo corpus

Chunk1		*Chunk2*	
Word	%	Word	%
ও	1.24	ও	1.23
এ	0.99	করে	0.86
করে	0.81	এ	0.84
টা	0.8	না	0.7
না	0.75	থেকে	0.6
থেকে	0.64	হয়	0.59
করা	0.56	তার	0.53
হয়	0.56	করা	0.48
এবং	0.5	হয়েছে	0.44
হবে	0.47	এবং	0.42

4.1.2. Type-to-token ratio (TTR). Type-to-token ratio is the measurement of how many times "old" words repeat themselves before a "new" word makes it appearance in a body of text [7]. The ratio is calculated by dividing total number of word tokens by the total number of distinct words. This measurement varies with the sample size as well as language. Texts with a high proportion of distinct words are likely to have low type-to-token ratio. In comparison to a bigger sample small samples will experience low TTR and will be sparser. This is because initially the corpus size being small the data in the text can not cover up all the words of that language and new words frequently enter the text, making the TTR low. But as corpus size grows, almost all the words are likely to have made their entry and new words are less frequent to enter the text, which gradually increases the TTR.

TTR depends on language in a sense that one language is different from another in terms of morphosyntactic features and orthographic variations [7]. Language with case system will have comparatively low TTR. Arabic being a language with highly inflective morphology has a very low TTR compared to English (Yahya 1998). So, different languages will show different TTR for equal text lengths in comparable domain. Table 4 is a comparison between our findings of Prothom-Alo to other corpora [7].

Table 4: Type to-token ratio for texts of varying size on corpora of different language

Text Length (words)	Bengali (Prothom-Alo)	English (Brown)	Arabic (Al-Hayat)
100	1.136	1.449	1.190
1600	1.984	2.576	1.774
6400	2.385	4.702	2.357
16000	3.135	5.928	2.771
20000	3.366	6.341	2.875
1000000	14.855	20.408	8.252

4.1.3. Appearance of non-Bangla words. A corpus is the best source to develop a vocabulary or lexicon for use with other language processing applications. We developed a vocabulary file consisting of the distinct words in the corpus. Manual sampling of the vocabulary list shows that a significant number of English words appear in Bangla script. While finding out all of them manually from a list of four lacks words is a mammoth and impossible task, Table 5 lists a few of them common to the news corpus. This observation of non-Bangla words tells us the story of how in the course of time words from other languages made their permanent residence in Bangla. This can well be the story for many other languages and a diachronic corpus can provide important evidence of this incredible nature of language.

Table 5: A few English words that appear in Prothom-Alo corpus

অ্যাশট্রে	আইসোলেশন	আউটপুট	ইকনমিক
একসেন্ট	ইনডেমনিটি	ক্যাশিয়ার	একসেপ্টেট্যান্স
ক্যালেন্ডার	একসেসরিজ	লিবারেশন	ক্যালকুলেশন
কার্ডিনাল	ওরিয়েন্টশন	ওয়ালপেপার	একস্ট্রাঅর্ডিনারি

4.1.4. Average word length. We have studied average word length of other language and thought it would be interesting to see what the number for Bangla looked like. Elderton (1949), Herden (1956) and many others studied for word length distribution of English. Dewey (1950) found that average word length at character level is 4.38 [1]. Other studies show that the average word length in Chinese is as little as 1.60 characters. Thai, which has an alphabetic writing system, shows an average word length of 5.1 [12]. An-other study on Wall Street Journal (WSJ) shows that for English the

average word length is 5.04 at character level [8]. For Bangla there have been very few studies. Dash (2005) shows that for Bangla the average word length is 5.12, where he studied the CIIL corpus [1]. However, in our analysis we found it to be 8.62, which is surprisingly high and does not go with the findings from CIIL corpus. Further inspection suggests that one reason that might have very little and almost negligible contribution to this high number is the presence of conjunct words. And there exists a large number of hyphenated words, which has been marked as conjunct as well- a flaw in our tokenization. In Prothom-Alo no distinction was made between a hyphen and a dash. In texts, especially newspaper ones, hyphenated words are numerous. While tokenizing, we used punctuations, white space, brackets, braces, and numbers etc. as word boundaries. This does not include the special character hyphen, i.e.: "-", because doing so will cause conjunct words to split around the hyphen and form two distinct words, which we do not want. One solution to this problem will come from a tagged corpus in a way that our manual inspection revealed that in case of conjunct words, both the words around the hyphen are nouns, which is not the case for the hyphenated words. Since we do not have the tagged version of the corpus yet, we could not come up with a solution.

Table 6: Character level length of words in Prothom-Alo corpus

Word length (char)	%	Word length (char)	%
9	12.35	15	2.99
8	12.21	4	2.13
10	11.2	16	2.04
7	10.97	17	1.47
11	9.73	18	1.02
6	8.36	19	0.7
12	7.48	3	0.48
13	5.91	20	0.42
5	5.18	21	0.27
14	4.29	22	0.15

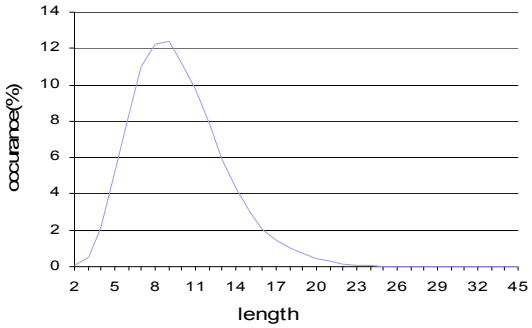

Figure 2: Usage of words vs. word length in Prothom-Alo corpus

In our study (Table 6) we found that words with nine characters are most frequent followed by eight and ten. Figure 2 graphs the usage of words against word length. Words as long as forty-five characters are also present. Figure 3 is a list containing samples of words that are more than twenty characters in length. While carefully inspection makes it clear that most of these words suffer from spelling mistakes, missing word boundary such as a space character or are hyphenated words as stated earlier, these words are outnumbered-only about one thou-sand. However at this point we do not have enough knowledge to come with an answer what is the reason of such big word length- if this is not what it should be.

অংশগ্রহণকারীপ্রতিষ্ঠানের, অগ্নিপ্রতিরোধ-ব্যবস্থাহীন, অজান্তে-
অনিচ্ছাকৃতভাবে, অধিনায়কত্ব-ক্যারিয়ারেরই, অনার্স-মাস্টার্সধারীদেরই,
অনিশ্চিত-আনপ্রেডিকটেবল, অনুভূতি-রবীন্দ্রনাথের,
অনুযায়ীছাত্রছাত্রীভর্তিকরা, অনুষ্ঠান-দলমতনির্বিশেষে, অনুষ্ঠান-সর্বক্ষেত্রে,
অন্তর্দর্ৃষ্টিসম্পন্ন

Figure 3: Few words of more than twenty characters long in corpus

4.2. Character level analysis

At the character level we analyzed global occurrence of character and words with particular characters at the first position.

4.2.1. Overall character frequency. Error! Reference source not found. shows the global percentage of use of characters in the Prothom-Alo news corpus. Top thirty are presented here. The results are more or less similar. Because both "r" and "t" has graphic variants, the uses of these characters in the corpus are higher. The variants of "r" are *raphala and reph* while "t"'s variant is *khandata (half-t)*. This

result shows that use of consonant graphic variant strongly influences the use of characters in Bangla.

Table 7: Overall usage of characters in Prothom-Alo corpus (top 30 entries)

Ch.	%	Ch	%	Ch	%
া	10.6	ল	3.01	ো	1.42
র	8.55	ম	3.01	শ	1.33
ে	8.15	প	2.5	গ	1.29
্	6.55	দ	2.23	এ	1.03
ি	5.74	য়	2.22	ী	1.03
ন	5.3	ু	1.83	আ	1.02
ক	4.59	য	1.61	ছ	0.99
ব	3.84	হ	1.51	ই	0.92
ত	3.75	ট	1.51	চ	0.88
স	3.15	জ	1.49	ও	0.74

4.2.2. Initial characters. Table 8 shows what we found by analyzing the characters that start a word, the word initial characters. This type of analysis can suggest about the preferences of the native language users. Consonants dominating the vowels show that the presence of large number of consonant in a language has a great effect in formation of words.

Table 8: Usage of word initial characters in Prothom-Alo corpus (top 30 entries)

Ch.	%	Ch.	%	Ch.	%
ক	9.06	ত	3.64	ট	1.44
স	8.74	অ	3.3	ভ	1.35
ব	8.64	জ	3.21	ল	1.15
প	7.73	গ	2.59	ছ	1.11
আ	5.32	র	2.36	খ	1.01
ম	5.23	শ	2.34	ই	1
এ	5.09	চ	1.93	ফ	0.99
হ	4.7	য	1.92	ঘ	0.6
ন	4.62	ও	1.91	ড	0.59
দ	4.25	উ	1.57	থ	0.57

5. Zipf's curve for Prothom-Alo corpus

Zipf's law, discovered empirically by Zipf in 1949, states that the frequency of word tokens in a large corpus of natural language is inversely proportional to the rank [9]. That is, if f is the frequency of a word in a corpus and r is the rank, then: f = k / r, where k is a constant. If log(f) is drawn against log(r) in a graph (which is often known as Zipf's curve), a straight line with a slope of -1 is obtained [10]. Experiments confirm that while the law was correct for small corpora, the curve dropped below that of Zipf's straight line at about rank 5000 for large ones [11]. To draw Zipf's curve for Prothom-Alo corpus, we need to rank the words in the corpus, using one of the two common approaches: words that have the same frequency are assigned the same rank or, words with same frequency are given different ranks – one that has higher precedence in lexicographic order will get the higher rank. The sampling method may vary; we sampled at every 500th rank. Figure 4 shows the results.

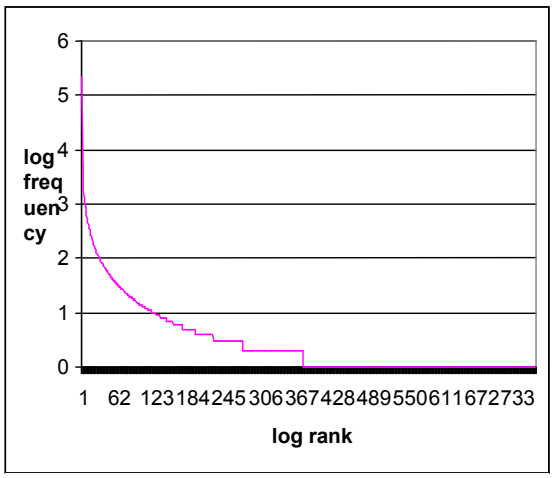

Figure 1: Zipf's curve for samples at every 500th rank

The discrepancy from the expected behavior is obvious, but we have yet to come up with a proper explanation. This may indicate excessive sparseness or idiosyncratic term distribution patterns in the corpus [7].

6. Conclusion

We have described our methodology of creating Prothom-Alo corpus. By analyzing the data we have shown that there are some differences between the

statistics from CIIL corpus which is reflected in the results for the average word length of CIIL and Prothom-Alo corpus. While Zipf's curves for English Brown corpus and WSJ (Wall Street Journal) show almost same characteristics [11], this is not the case for Bangla. We conclude that this abnormal behavior is due to the fact that Prothom-Alo being a news corpus is biased to some particular editing style while flexible in terms of new word type usage, which is obvious from the number of words with a frequency of one. This corpus may also not be a good source to create a language model from.

7. Acknowledgement

This work has been supported in part by the PAN Localization Project (www.panl10n.net) grant from the Inter-national Development Research Center (IDRC), Ottawa, Canada. The authors would like to thank Naira Khan and Arnab Zaheen of CRBLP for their support.

8. References

[1] N.S. Dash, *Corpus Linguistics and Language Technology*, Mittal, New Delhi, 2005.

[2] C.F. Mayer, *English Corpus Linguistics - An Introduction*, 2002.

[3] www2.ignatius.edu/faculty/turner/languages.htm

[4] A. Bharati, R. Sangal and S.M. Bendre, "Some Observations Regarding Corpora of Some Indian Languages", *Proc. Intl. Conf. Knowledge Based Computer Systems (KBCS- 98)*, NCST, Mumbai , 17-19 Dec. 1998.

[5] "Prothom-Alo" website: www.prothom-alo.com.

[6] Definition of function word available at: web.cn.edu/kwheeler/lit_terms_F.html.

[7] A. Sarkar, A. De Roeck and P. Garthwaite, "Easy Measures for Evaluating non-English Corpora for Language Engineering: Some Lessons from Arabic and Bengali", *Technical Report No: 2004/05*, Open University – Department of Computing, 16th February, 2004.

[8] J. Karlgren, "Stylistic Experiments in Information Retrieval". In *Proceedings NeMLaP 2*, Bilkent University, Ankara, September, 1996.

[9] D.M.W. Powers, "Applications and Explanations of Zipf's Law". *In Proceedings of the Second Conference on Computational Language Learning (CoNLL-98)*, Sydney, Australia, January 22 - 24, 1998.

[10] L.Q. Ha, E. I. Sicilia-Garcia, J. Ming, F.J. Smith, "Extension of Zipf's Law to Words and Phrases", *In Proceedings of the 19th International Conference on Computational Linguistics*, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002.

[11] L.Q. Ha, E. I. Sicilia-Garcia, J. Ming, F.J. Smith, "Extension of Zipf's Law to Word and Character N-grams for English and Chinese", *Intl. Journal of Computational Linguistics & Chinese Language Processing Vol. 8, No. 1*, February 2003.

[12] D.D. Palmer, "A Trainable Rule-Based Algorithm for Word Segmentation", *Proc. 35th annual ACL meeting*, 1997, pp. 321-328.