SUMono: A Representative Modern Bengali Corpus

Article · January 2014

CITATIONS
7

READS
299

5 authors, including:

Mohammad Abdullah Al Mumin
Shahjalal University of Science and Technology
13 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Mohammad Reza Selim
Shahjalal University of Science and Technology
5 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Muhammed Zafar Iqbal
Shahjalal University of Science and Technology
21 PUBLICATIONS **50** CITATIONS
SEE PROFILE

SUST Journal of Science and Technology, Vol. 21, No.1, 2014; P: 78-86

SUMono: A Representative Modern Bengali Corpus

(Submitted: November 11, 2013**; Accepted for Publication:** January 21, 2014**)**

Md. Abdullah Al Mumin[1], Abu Awal Md. Shoeb[2], Mohammad Reza Selim[3] and M. Zafar Iqbal[4]
[1, 2, 3, 4]*Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Bangladesh*
Email: *mumin-cse@sust.edu[1], shoeb-cse@sust.edu[2], selim@sust.edu[3], mzi@sust.edu[4]*

Abstract

The development of Language Engineering applications requires availability of sizable, reliable and representative corpora. However, such corpora are not routinely available for Bengali language. This paper introduces *Shahjalal University Monolingual (SUMono)* corpus, a representative modern Bengali corpus consisting of more than 27 million words, which is the largest of its kind. This paper describes how we have constructed SUMono corpus from available online and offline Bengali texts, with articles tagged as belonging to 6 domains: Natural Science, Social Science, Computer and IT, Literature, Mass Media and Blogs. We show some characteristics of Bengali language based upon the statistical analysis of this corpus. We also compare the 'inherent sparseness' of Bengali with English and Arabic by observing Type-to-Token ratio of the languages. We assess our corpus in terms of its representativeness, homogeneity and vocabulary growth rate using established techniques like Zipf's law, distribution of function words and Baayen's equation, respectively. We found that our corpus is balanced with respect to the frequency distribution as well as to the range of idiosyncratic phenomena.

Key Words: monolingual corpora; representative corpus; modern Bengali; Bengali corpus; Zipf's law;

1. Introduction

A corpus can be defined as 'a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis. [1]

The importance of corpora to linguistic study is appreciated. A corpus to a linguist is very valuable because it allows statements to be made about language in very convincing fashion. The actual use of the corpus includes computational linguistics as well as studies in the grammar, lexicography, language variations, historical linguistics, language acquisition, and language pedagogy.

It is now widely recognized that for most applications, a sufficiently large corpus reflecting the full range of domains and usage is essential. However, for Bengali, freely available corpora that meet these requirements do not exist. British National Corpus (BNC), the corpus for British English, is the first corpus that has been constructed keeping these requirements in mind. Later, the BNC model has been followed in the construction of the American National Corpus, the Korean National Corpus, the Polish National Corpus, and the Russian Reference Corpus [2].

In this paper, we introduce a large-scale representative Bengali corpus, the SUMono corpus. The format and contents of the SUMono corpus follows the framework of the American National Corpus (ANC) [3]. Like ANC, the SUMono corpus does exhibit two criteria: First, it is broad, i.e., both large and well balanced. Second, it is available for the entire research community. These two properties make the SUMono corpus first of its kind for the Bengali language.

The organization of the rest of the paper is as follows: section 2 reviews on the previous monolingual corpora for Bengali and justify the necessity of developing another Bengali corpus. Section 3 describes the development of the SUMono corpus. Section 4 focuses on some characteristics of Bengali language by analyzing various statistics obtained from the corpus. Section 5 performs some established experiments to assess the quality of the corpus and finally, section 6 concludes the paper.

2. Why a Bengali Corpus?

The need for large scale representative corpus for natural language and speech is well established. There are many such corpora for English and many other European and Asian languages. However, such collections have not been constructed systematically for Bengali language. Most researchers in NLP and IR construct their own corpora that are usually small, special purpose, not representative and not publicly available.

Central Institute of Indian Languages (CIIL) first introduce a Bengali corpus along with corpus of other nine Indian languages. The CIIL [4] corpus is a three million words corpus. Bharati et al. [5] analyzed and compared the data between Bengali and other Indian languages using the CIIL corpus. Although it has been designed to make sufficiently representative, the small size of the CIIL corpus is not sufficient for today's large-scale applications. Moreover, the differences in the writing style as well as the phonetic structure between Indian and Bangladeshi Bengali languages also show the necessity of developing our own corpus.

'Prothom-alo' [6] news corpus has been developed by collecting data from a Bangladeshi daily news paper, the 'Prothom-Alo', for the year 2005. Although the corpus contains a moderate size of more than 18 million words, the corpus is not representative of Bengali language. As they cited, Prothom-Alo being a news corpus is biased to some particular editing style while flexible in terms of new word type usage. This corpus may also not be a good source to create a language model. Moreover, the corpus is not available for the research community.

Islam et al. [7] propose a method for building an effective corpus which can be used only for evaluation of Bengali text compression. Shamshed et al. [8] propose a method for building Bengali text corpus which is only designed for information retrieval system.

In our expeiments, we develop the SUMono corpus as a large scale in size and sufficiently representative for Bengali language. Table 1 depicts a comparison in size between SUMono corpus and the other Bengali corpora whose corpus statistics are available.

Table 1: Comparison in Size Between SUMono and other Bengali Corpora

	SUMono	'Prothom-alo'	CIIL
Corpus size (in words)	27,118,025	18,100,378	3,044,573
Vocabulary size (no. of unique words)	571,572	384,048	190,841

3. Development of the SUMono Corpus

The SUMono corpus project was initiated in 2010 with the aim of building a carefully designed corpus of 100 million words of Bangladeshi written and spoken Bengali language that generally follows the framework of the ANC. However, the first release of the SUMono corpus contains only written texts of more than 27 million words. In this section, we describe various aspects of design and construction of the corpus.

3.1 Representativeness

The major issue that is addressed in design of SUMono corpus is its representativeness. According to Biber et al. [9], "representativeness refers to the extent to which a sample includes the full range of variability in a population." In other words, representativeness can be achieved through balancing and sampling of language or language variety presented in a corpus. SUMono corpus contains roughly 3,691 articles covering 6 broad subject categories. In addition, the articles are written by many authors from a variety of backgrounds and contain texts of different types (e.g., quantum mechanics vs fine arts). Besides, it also contains real life text in everyday use of Bengali that implies it has the sampling and representativeness property. Table 2 shows the category wise summary of the dataset[1]. Lexical diversity score (i.e., token/type ratios) refer to the number of times each vocabulary item appears in the text on average.

Table 2: Summary of the SUMono Dataset

Subject Category	No. of Articles	Total Words		Number of Distinct Words	Lexical Diversity
		Number	%		
Natural Science	683	1,711,179	6.31	101,088	16.93
Social Science	1,208	8,780,323	32.38	278,466	31.53

[1] according to the data on November 1, 2013.

Computer and IT	248	975,112	3.60	57,034	17.10
Literature	446	6,777,650	24.99	259,954	26.07
Mass Media	1,094	7,846,419	28.93	221,076	35.49
Blogs	12	1,027,342	3.79	79,002	13.00
-The Whole Dataset	3,691	27,118,025	100	571,572	47.44

3.2 Data Sources

We have used texts from the following sources that are either publicly available or granted permission from respective copyright holders.

- Books written in Bengali like 'Quantum Mechanics', 'Relativity Theory', 'Science and Math collections', 'Hundred interesting game of Science' and many others by Muhammed Zafar Iqbal; 'Some Questions about Function' by Dr. Rashed Talukder; Translated version of 'A Brief History of time'; Bengali version of NCTB books.
- Online version of newspapers like Prothom-Alo, BDNews24.com, Bangladesh Pratidin, Daily JaiJaiDin, Daily Inqilab, Shaptahik, Shaptahik 2000.
- Websites like comjagat.com, computerbarta.com, bigganschool.org, biggani.org, at-tahreek.com, natunpata.com, golpokobita.com, kaliokalam.com, wikipedia.com/bn
- Social science articles usually written in Bengali from 'SUST Studies', a journal published by Shahjalal University of Science and Technology.
- Bengali part of the SUPara [10] corpus.

3.3 Preprocessing

Since the individual sources of collected texts differ in many aspects, a lot of effort was required to integrate them into a common framework. The following steps have been applied as preprocessing on the documents.

Cleaning: We start by cleaning up the original material that we collected from the different sources. This cleaning up means that the various formats, for example rtf, doc and pdf, are converted to plain text files. Tagged files like html and php files are normalized by deleting tags and then converted to plain text files.

Encoding: We use simple principles for the encoding of documents in our corpus. The texts are encoded according to international standards by using UTF8 (Unicode). We have used Nikosh[2] converter to encode all formats into Unicode.

3.3 Availability

The corpus is available free of charge for educational and research purposes. However, the license agreement requires that the use of any statistical data must include a citation. The corpus is distributed through the Computer Science and Engineering (CSE) department of Shahjalal University of Science and Technology (SUST)[3].

4. Statistical Properties of Bengali

Statistical inference allows the linguists to generalize from properties observed in a specific sample (corpus) to the same properties in the language as a whole. Statistical inference requires that the problem at hand is operationalized in quantitative terms, typically in the form of units that can be counted in the available samples [11]. This is the case we will concentrate on here now. Using the corpus of 442 MB size, we analyze some simple characteristics of Bengali initially.

Character Level Analysis

We begin with computing relative usage of Bengali characters. Table 3 shows the percentage of occurrence of each letter in the corpus. There are about 139,689,873 characters excluding spaces and punctuations in the corpus with the average of 5.15 letters per word. In ordinary English text, there are on the average about 4.5 letters per word [12]. English words form with only 5 vowels and 21 consonants whereas Bengali words form with 12 vowels, 20 allographs and 39 consonants making the Bengali word length longer.

We see from the data that the first two mostly used letters are vowel allographs. The next most frequently used letter is the consonant 'র'. The reason is besides its usual use in texts, 'র' also used in cluster formation texts as

[2] http://www.ecs.gov.bd/nikosh
[3] http://www.sust.edu/

Table 3: Percentage of Occurrence of Each Letter in the Corpus

letter	%	letter	%	letter	%	letter	%	letter	%
া	10.514	য	02.210	ই	01.098	ড	00.388	ৌ	00.066
ে	08.767	দ	02.157	ী	00.944	ক্ষ	00.331	ঃ	00.041
র	08.346	ু	01.922	চ	00.906	ড়	00.312	ঢ	00.041
়	05.953	য়	01.666	থ	00.760	ঁ	00.232	ঈ	00.015
ি	05.538	হ	01.492	ঘ	00.740	ঠ	00.224	ঐ	00.015
ন	05.238	ো	01.439	ষ	00.739	ূ	00.211	ঋ	00.008
ক	04.776	ট	01.384	ভ	00.692	ৃ	00.197	ঊ	00.004
ত	03.908	জ	01.295	ধ	00.690	ঙ	00.171	ঢ়	00.004
ব	03.901	শ	01.246	ও	00.664	ঘ	00.163	ঔ	00.002
ম	02.983	আ	01.179	অ	00.629	ঞ	00.094		
স	02.954	এ	01.146	ং	00.465	ৈ	00.088		
ল	02.868	গ	01.138	ণ	00.444	্	00.086		
প	02.356	ছ	01.110	উ	00.401	ঝ	00.079		

‘□‘ (*reph*) and ‘ ‘ (*ro-phola*). Surprising to our intuition, the next most frequently used letter is ‘্’ (*hoshonto*). While writing Bengali texts in paper, we barely use *hoshonto* but we use many clustered texts. We are not used to see or think *hoshonto* in those clustered texts while writing in paper. But in computation, each cluster form includes a *hoshonto* in its formation, which makes its count high.

Table 4 shows the percentage of occurrence of each letter that start a word, i.e., the word initial letter. It seems that most of the words starts with a consonant with ‘ক’ having the most of the times. Among the vowels ‘আ’ and ‘এ’ are used most of the times as the word initial letter.

Table 5 shows the frequency of top n-grams (sequences of letters) in the corpus. Space characters have been converted to ‘◊’ for legibility. Again, just n-grams up to 5 letters are shown.

Table 4: Percentage of “Occurrence of Initial Letter of Words in the Corpus

letter	%	letter	%	letter	%	letter	%	letter	%
ক	09.927	দ	04.369	উ	01.436	ড	00.486	ঔ	00.010
ব	08.526	অ	03.213	থ	01.255	ঢ	00.198	ে	00.006
স	08.275	য	02.760	ছ	01.189	ঠ	00.168	ূ	00.006
প	07.330	জ	02.662	ল	01.154	ঝ	00.131	া	00.005
আ	05.949	র	02.142	খ	01.042	ঐ	00.076	ড়	00.005
এ	05.647	শ	02.127	ক্ষ	00.950	ঈ	00.056	ঙ	00.004
ন	04.918	গ	02.083	ই	00.848	ঋ	00.034	ি	00.004
ম	04.899	চ	01.933	ট	00.661	ষ	00.026	য়	00.003
হ	04.645	ভ	01.677	ধ	00.644	ঊ	00.015	ঞ	00.002
ত	04.424	ও	01.480	ঘ	00.587	ী	00.011	ঢ়	00.002

Table 5: The top 10 Frequent n-gram Lletter in the SUMono Corpus

1-gram	Freq.	2-gram	Freq.	3-gram	Freq.	4-gram	Freq.	5-gram	Freq.
া	14686368	ে ◊	4804831	ে র ◊	1846104	◊ প ্ র	612648	◊ ক র ে ◊	207833
ে	12246681	র ◊	4409020	া র ◊	1308857	◊ ক র ে	407788	া দ ে র ◊	181371
র	11658807	া র	2677875	◊ ক র	956265	দ ে র ◊	371488	◊ খ ে ক ে	175475
়	8315941	◊ ক	2605799	প ্ র	714765	ন ে র ◊	256857	খ ে ক ে ◊	152824
ি	7735977	া ◊	2418849	ত ে ◊	651088	ি য় ে ◊	246167	◊ এ ব ং ◊	146044
ন	7316659	◊ ব	2211503	◊ প ্	631644	◊ ত া র	230413	র ◊ প ্ র	144208
ক	6671400	◊ স	2128266	ক ে ◊	624038	ত া র ◊	220896	◊ প ্ র ত	143449
ত	5459011	ে র	2009212	◊ ত া	591799	ে র ◊ স	220828	প ্ র ত ি	136476
ব	5449812	◊ প	1912624	় য া	540597	র ে র ◊	214613	◊ ত া র ◊	135278
ম	4166883	় র	1580520	র ে ◊	529583	ক র ে ◊	209045	ত ে ◊ প া	126155

Word Level Analysis

Table 6 shows n-letter high frequency words in the corpus. All words shorter than 20 letters were extracted for further calculations. However, just up to 5-letter words are depicted in the table. 1-letter words such as 'ক', 'খ', 'গ', 'ঘ' etc. come from when we use these letters for numbering or indexing the texts in documents. Most of the valid long words are foreign words borrowed from English scientific terms such as হাইড্রোপারোক্সিকোসাটেট্রানায়েক (hydroperoxitetranoyek), এসটিমাইক্রোইলেক্ট্রনিক্স(stmicroelectronics). Table 7 shows the top 50 frequent words in the SUMono corpus.

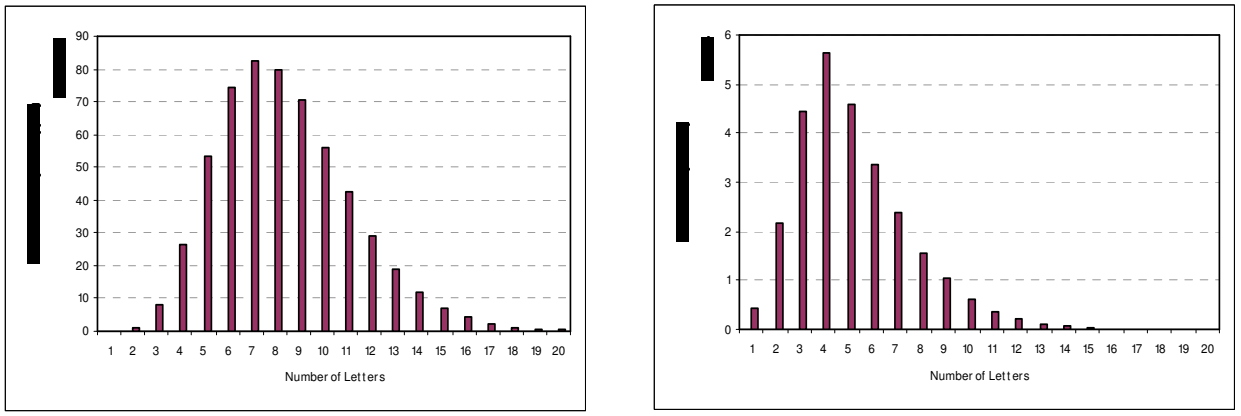
Table 6: The top 10 Frequent n-Letter Words in the SUMono Corpus

1-letter	Freq.	2-letter	Freq.	3-letter	Freq.	4-letter	Freq.	5-letter	Freq.
ও	236784	না	295522	করে	258957	থেকে	160359	সঙ্গে	75159
এ	133844	এই	179105	এবং	147558	জন্য	108655	তাদের	66305
ই	9739	আর	127794	তার	140863	আমার	98923	হয়েছে	64892
ঐ	9207	যে	125377	করা	120207	করতে	98062	মধ্যে	64683
ক	5871	হয়	124109	হবে	111020	একটি	90443	হচ্ছে	51519
খ	5497	এর	88853	আমি	105736	তিনি	82045	দেশের	34607
গ	4750	এক	84180	কথা	78874	কোনো	79926	প্রথম	33542
র	3910	সে	78758	ছিল	73121	নিয়ে	73669	মানুষ	33258
ঘ	3662	কি	72067	হয়ে	70458	একটা	69369	সরকার	32661
আ	3009	বা	65258	আছে	70055	আমরা	67591	করেছে	30314

Table 7: The top 50 Frequent words in the SUMono Corpus

Word	%	Word	%	Word	%	Word	%	Word	%
না	1.09	হয়	0.46	এক	0.31	হয়ে	0.26	তা	0.23
করে	0.95	করা	0.44	তিনি	0.30	আছে	0.26	যায়	0.22
ও	0.87	হবে	0.41	কোনো	0.29	একটা	0.26	দিয়ে	0.21
এই	0.66	জন্য	0.40	কথা	0.29	আমরা	0.25	সেই	0.21
থেকে	0.59	আমি	0.39	সে	0.29	তাদের	0.24	কিছু	0.21
এবং	0.54	আমার	0.36	আমাদের	0.28	বা	0.24	অনেক	0.20
তার	0.52	করতে	0.36	সঙ্গে	0.28	হয়েছে	0.24	তারা	0.20
এ	0.49	একটি	0.33	নিয়ে	0.27	বলে	0.24	করেন	0.19
আর	0.47	এর	0.33	ছিল	0.27	মধ্যে	0.24	তো	0.19
যে	0.46	কিছু	0.32	কি	0.27	মনে	0.23	নেই	0.19

Figure 1a shows the number of distinct words (types) recognized in the corpus. As it is seen, most of the word types are 7 letters long in Bengali. Figure 1b depicts the total occurrence of n-letter words (tokens). From both figures, we see that though the number of 4-letter word type is quite low, we use them most often in Bengali.



(a) The number of distinct n-letter words (types) (b) The umber of total n-letter words (tokens)

Figure 1: Distribution of Usage of n-Letter Words in SUMono Corpus

The above statistics may have different applications in different contexts. For example, post processing in Bengali OCR, Speech-to-Text, spell checker, etc. applications may employ above information to build a probabilistic model of the language and guessing words and letters in case of ambiguity in recognition.

Inherent Sparseness

'Inherent sparseness' of a language compared to other languages can be measured by observing Type-to-Token Ratio (TTR) of the languages for identical text lengths in comparable genres. TTR measures the number of 'old' words we expect to see in running text before coming across a 'new' one. The ratio is easily calculated by dividing the total number of terms in a fragment by the number of distinct terms. From the perspective of statistical language processing, it is important to note that different languages appear to display different TTRs of what could be called 'inherent sparseness' [13].

In order to verify inherent sparseness of Bengali compared to English and Arabic, we picked sample sizes of 1 million words that allow us to compare Bengali and Arabic results with data reported for English on the Brown corpus. Table 8 shows the TTRs for fragments of different lengths from corpora of different language. The English Brown corpus data and Arabic Al-Hayat corpus data are taken from Sarkar et al. [13]. The TTR for the one-million English Brown corpus approximately equals 20.408 and for the Arabic Al-Hayat corpus of the same text length equals 8.252 whereas for Bengali it is 15.859. The finding invites the conclusion that Bengali textual data may be inherently sparser than English and quite inherently denser than Arabic. This suggests that, for some statistical applications, Bengali corpora may need to be significantly larger than English ones and significantly smaller than Arabic ones for similar effect.

Table 8: Type-to-Token Ratios for Corpora Fragments of Different Lengths of Different Language

Text Length	Bengali (SUMono)	English (Brown)	Arabic (Al-Hayat)
100	1.204	1.449	1.190
1600	1.913	2.576	1.774
6400	2.455	4.702	2.357
16000	2.985	5.928	2.771
20000	3.244	6.341	2.875
1000000	15.859	20.408	8.252

5. Assessment of the Corpus

In this section, we adopt two rough, but computationally cheap techniques [14] for a-priori profiling of corpus quality. First, we check for obvious imbalances by tracking term distribution patterns against Zipf's law. Second, we trace the behavior of the function words to measure homogeneity of the corpus. In addition, we study the vocabulary growth rate of the corpus.

5.1 Zipf's Distribution

Zipf's law is useful as a rough description of the frequency distribution of words in human languages [15]. Set against Zipf's law, frequency distribution in an actual dataset is also a reasonable way to gauge data sparseness, and can provide evidence of imbalance in a sample.

Zipf's law draws a relationship between the frequency of a word f and its position in the list, known as its rank r. The law states that: $r.f = c$, where r is the rank of a word, f is the frequency of occurrence of the word, and c is a constant that depends on the text being analyzed.

Word frequencies have been counted for all the domains separately, and for the whole dataset. In all, seven lists of word frequencies were created. Each was sorted in descending order of frequency. Rank was assigned and the sorted lists were plotted against rank. Table 2 shown in section 3 is a summary of all the data used in our experiments. Figure 2 shows the results of the plots on logarithmic scale.

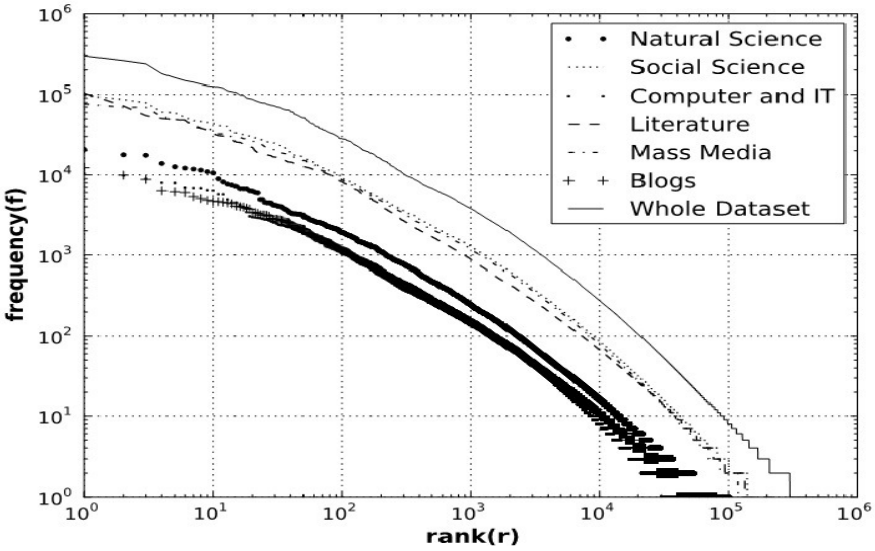


Figure 2: Zipf's Curve for All Six Domains and Whole Dataset

According to Zipf's law, for a representative sample the graphs should be a straight line with slope -1. In practice, this may not be the case because many words will have the same frequency but be assigned different rank. As expected, graphs improved as the size of data increased, and the proportion of rare words declined. The analysis of the graph shows that term distribution in whole dataset or in each subject area fits Zipf's law comfortably. As a result, we can believe that the dataset is balanced, either overall, or for each subject area.

5.2 Behavior of Function Words

Function words are words whose purpose is more to signal grammatical relationship in a sentence than to convey lexical meaning. In the context of information retrieval, function words are not so informative because of their very frequent occurrence in all documents. However, the occurrence and distribution of frequent words has some value in assessing corpus quality. In a balanced collection, the function words will tend to distribute more homogeneously than content words, whose occurence is "bursty" [16].

Hence, we investigate the distribution of very frequent terms in SUMono corpus by dividing the corpus into three chunks and observing whether the function words occur very frequently in each chunk. We take two domains for each chunk and make a frequency analysis for each chunk. Table 9 shows the top 10 frequent words for three chunks of SUMono corpus. We observe that most 10-20 frequent words are same for each chunk and also same for the whole dataset (Table 7) only with difference in their rank. Thus we conclude that, in this corpus, very frequent terms distribute more homogeneously than less frequent terms.

Table 9: The Dstribution of Function Words in the SUMono Corpus

CHUNK1 (Natural Science, Mass Media)		CHUNK2 (Social Science, Blogs)		CHUNK3 (Computer and IT, Literature)	
Word	%	Word	%	Word	%
ও	0.95	ও	1.08	না	1.40
না	0.91	না	1.01	করে	1.07
করে	0.91	করে	0.91	আর	0.74
এই	0.70	এই	0.67	তার	0.67
থেকে	0.62	এবং	0.67	আমি	0.64
এ	0.60	থেকে	0.60	আমার	0.63
এবং	0.58	যে	0.53	এই	0.60
হয়	0.57	করা	0.52	থেকে	0.55
করা	0.51	এ	0.52	ও	0.53
তার	0.45	হয়	0.50	সে	0.51

5.3 Vocabulary Growth

The statistical models of Baayen [17] link the degree of productivity of a morphological process to the rate of vocabulary growth, i.e., to how frequently new word types that are formed by the process are encountered when an increasing amount of text is sampled. If the degree of productivity changes over time, there should be a corresponding changes in the vocabulary growth rate [18].

Baayen shows that the growth rate of the vocabulary, the rate at which the vocabulary size increases as sample size increases, can be estimated as follows: $G = V(1) / N$. In this equation, $V(1)$ is the number of words occurring once (*hapax legomena*) in a sample size N. In the Brown corpus, G = 24375/996883 = 0.024, indicating that the vocabulary size is still growing at a relatively fast pace. The vocabulary is still growing (although at a slower pace) in much larger corpora, such as the written section of the BNC (G = 0.003) [19]. Figure 3 shows the vocabulary growth curve for the SUMono corpus. The vocabulary growth rate for SUMono corpus is, G = 273617/27118025 = 0.01, indicating that the vocabulary size in SUMono corpus is still growing at a relatively medium pace.

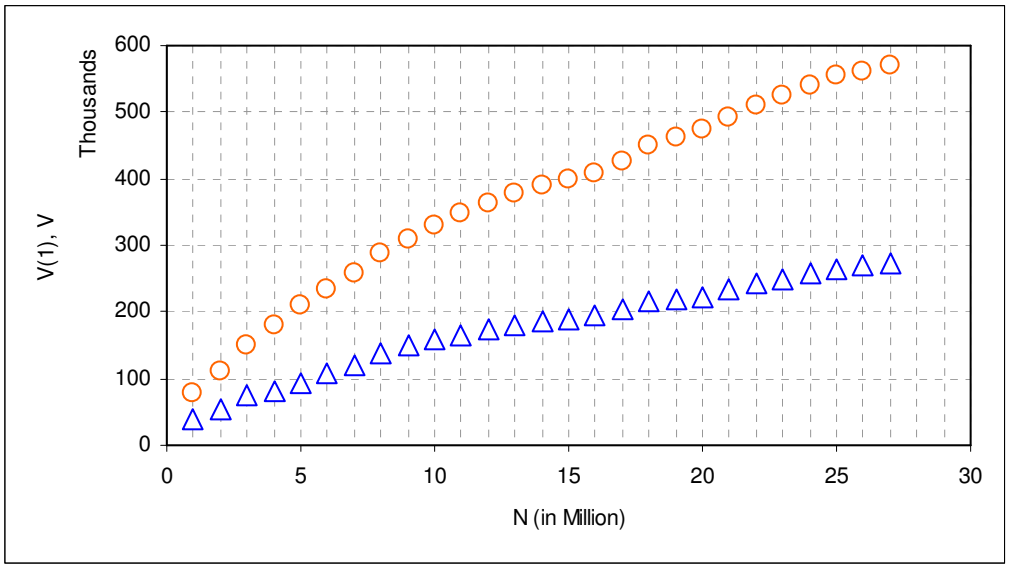


Figure 3: The SUMono Corpus Vocabulary Growth Curve: Number of Types (circles) and Hapax Legomena (triangles) for 27 Increasingly Larger Token Samples (N)

6. Conclusion and Future Work

In this paper, we have presented the SUMono corpus, a large-scale collection of representative Bengali texts. The corpus which consists of 27,118,025 words in Bengali is the largest available Bengali corpus. We have presented statistics of SUMono corpus which help us to study some properties of Bengali language. Findings from these corpora-based studies on Bengali will help develop more Bengali friendly and efficient word processors, OCR systems, search engines and similar other widely used applications. We have compared inherent sparseness of Bengali with English and Arabic and concluded that Bengali data is sparser than English and much denser than Arabic.

In its design, SUMono corpus is made representative by integrating a variety of text materials from different domain. We have investigated the balance of the corpus by checking Zipf distribution, over each of the sample domains as well as over the dataset as a whole. We have also investigated homogeneity by checking distribution of the function words in the corpus and observed the vocabulary growth rate using Baayen's equation. On the whole, we can suggest that the dataset is significantly balanced either with respect to frequency distribution, or with respect to the range of idiosyncratic phenomena. In this sense, the corpus is useful as a background for the development of techniques.

In future, we plan to integrate spoken data as well as enlarge the corpus further to achieve a corpus of 100 million words of written and spoken language. We would like to annotate SUMono corpus on various levels up to deep syntactic layer. We hope that the SUMono corpus will function as the basic source of reference for both national and international researchers who are willing to do their computational research on Bengali language processing.

References

[1] Tognini-Bonelli, E., 2001. *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

[2] McEnery, Tony, Xiao, R. and Tono, Y., 2006. *Corpus-based Language Studies,* Routledge.

[3] American National Corpus website: *http://americannationalcorpus.org*

[4] Dash, N. S., and Chaudhuri, B. B., 2001. *Corpus based Empirical Analysis of Form, Function and Frequency of Characters used in Bangla.* Special Issue of the Proceedings of the Corpus Linguistics Conference; 13:144-157.

[5] Bharati, A., Sangal, R. and Bendre, S.M., 1998. *Some Observations Regarding Corpora of Some Indian Languages.* In Proceedings of the International Conference on Knowledge Based Computer System (KBCS-98), NCST, Mumbai.

[6] Majumder, K.M.Y., Islam, M.Z. and Khan, M., 2006. *Analysis of and Observations from a Bangla News Corpus.* In Proceedings of 9th International Conference on Computer and Information Technology, ICCIT 2006, pp. 520-525.

[7] Islam, M.R. and Rajon, S.A.A., 2010. *Design and Analysis of an Effective Corpus for Evaluation of Bengali Text Compression Schemes.* Journal of Computers, Vol. 5, No. 1.

[8] Shamshed, J. and Karim, S.M.M., 2010. *Novel Bangla Text Corpus Building Method for Efficient Information Retrieval.* JCIT, ISSN 2218-5224, Vol. 1, Issue 1.

[9] Biber and Douglas, 1993. *Representativeness in corpus design.* Literary and Linguistic Computing 8: 243-257.

[10] Mumin, M.A.A., Shoeb, A.A.M., Selim, M.R. and Iqbal, M.Z., 2012. *SUPara: A Balanced English-Bengali Parallel Corpus.* SUST Journal of Science and Technology, Vol. 16, No.2; pp. 46-51.

[11] Trento, M.B. and Osnabr\v{u}ck, S.E. *Statistical Methods for Corpus Exploitation.* Corpus Linguistics: An International Handbook, pp. 777-803.

[12] Pierce, J.B., 1980. *An Introduction to Information Theory: Symbols, Signals and Noise.* Dover Publications

[13] Sarkar, A., De Roeck, A. and Garthwaite, P. 2004. *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.

[14] Goweder, A. and De Roeck, A., 2001. *Assessment of a significant Arabic corpus.* Proceedings Workshop on Arabic Language Processing, 39th ACL. Toulouse.

[15] Manning, C. and Schuetze, H., 1999. *Foundations of Statistical Natural Language Processing.* MIT Press. Cambridge, MA.

[16] Katz, S., 1996. *Distribution of content words and phrases in text and language modeling.* Natural Language Engineering, 2(1):15-59

[17] Baayen, R.H., 2001. *Word Frequency Distributions.* Dordrecht: Kluwer

[18] *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives.* edited by Stephan Kepser, Marga Reis, pp-357

[19] Trento, M.B. *Distributions in text.* Corpus Linguistics: An International Handbook, pp. 803-822.

[20] Darrudi, E. et al., 2004. *Assessment of a Modern Farsi Corpus.* In Proceedings of the 2nd Workshop on Information Technology and Its Disciplines, pp.73-77, Kish Island, Iran.