Quantitative Analysis of English Corpus in Tourism and Health Domain

Lalit Goyal
Department of Computer Science
DAV College, Jalandhar
goyal_aqua@yahoo.com

Abstract— *Statistical analysis of a language is an essential part of any of the natural language processing activity though it is translation, transliteration, summarization, lexicon formation, keyboard designs and many more. In this paper, a domain specific corpus (health and tourism) of English language provided by Computational Linguistic R&D at Special Centre for Sanskrit Studies J.N.U is analyzed statistically. The frequency analysis and word length analysis of English text is performed. Unigram, bigram, trigram and positional analysis of words has been studied.*

Keywords— *Corpus, English, Statistical Analysis, Quantitative Analysis, unigram, bigram, trigram Introduction.*

I. INTRODUCTION

A corpus or text corpus is a large and structured set of texts which nowadays is stored electronically. The plural of corpus is corpora. The corpus may also be the collection of transcribed speech or videos. It is an essential resource for processing the natural language(so called language engineering). With the recent advancement in computer technology the availability of language corpora and its processing has become even easier and has opened many new areas of research in language processing. 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 machine translation, transliteration and many more [1]. 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 almost all languages such as Chinese, Japanese, Spanish, Hindi, Punjabi, Urdu, Arabic has been compiled and analysed to enrich the language knowledge [2].

English is one of the most widely spoken languages in the world with more than 320 (million) native speakers worldwide [3]. Statistical analysis of English language results in a tremendous progress in different applications like Speech synthesis, Lexicography, Handwritten recognition, Text Summarization, Translation between human languages, natural language database and query answering etc. In this paper, statistical analysis of English corpus by Computational Linguistic R&D at Special Centre for Sanskrit Studies J.N.U has been carried out[4].

II. STATISTICAL ANALYSIS

Statistical analysis[5] of different languages is the foremost requirement to have a comprehensive database for all languages. Various methods are employed for statistical analysis. e.g. Qualitative analysis and Quantitative 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. In the present study, quantitative analysis of English text has been carried out. In quantitative research we classify features, count them, and even construct more complex statistical models in an attempt to explain what is observed.

An English corpus of size 4488312 characters is taken from Computational Linguistic R&D at Special Centre for Sanskrit Studies J.N.U. It was started since 2002 under the supervision of Dr. Girish Nath Jha [4]. The English corpora we are working on consist of Unicode text. The main reason is Unicode is universally accepted script that can be displayed and processed without hassle in all platforms.

In this paper, quantitative analysis of English corpus is performed at both character and word level. The study of the characters constituting the corpus is important for accounting their pattern of use in different context of the texts as well for comprehension of the general characteristics of the language. Thus, multi-layered information of the characters can be important and necessary contribution to Natural Language Processing (NLP), Computational Linguistics (CL), Optical Character Recognition (OCR), key-board design, Word Sense Disambiguation (WSD), Parts-of-speech Tagging, cryptography, language teaching, Machine Translation (MT), besides other applied and interdisciplinary studies. Moreover, it can provide insight about how language is used by different users in different domains of knowledge representation.

To calculate the different frequencies of the characters and performing word length analysis JAVA programming language is used. Usage of JAVA platform has provided a special effect for dealing with a very large database with high computational speed.

III. Results and Analysis

The quantitative analysis of the domain specific English corpus in health and tourism domain has been carried out. The corpus is annotated and consists of grammatically tagged sentences. Each word consists of a tag with a defined meaning e.g. NN stands for singular common noun, NNS stands for plural common noun, NP for proper noun, VB for verb and so on. Our first purpose was to clean the original corpus. All the tags of the words are removed. Even the syntax errors were removed to get the clean corpus on which the analysis was done.

The corpus consists of 50 articles, 25 articles in health domain and 25 articles in tourism domain. it has total of 768448 words with 30071 unique words. The size of the corpus is 7.38MB which after cleaning remains 4.28MB.

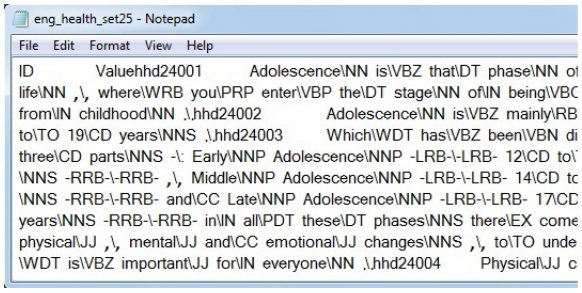

Fig. 1: Annotated sentences in English corpus

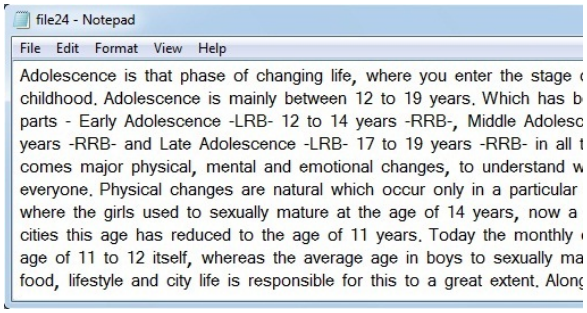

Fig. 2: Cleaned sentences in English corpus

3.1 Unigram Analysis

Unigram analysis is the study of text based on single word. The results shows that the top 5 words of the corpus are **the, of, in, is, and** which covers approximately 20% of the whole corpus. Top 118 words occupy 50% and top 3737 words occupy 90% of corpus. This means, the English language has approximately 4000 words which are commonly used irrespective of the total 30071 unique words.

TABLE I
Top 10 most frequently used unigram words

Rank	Word	Freq.	%age	Comm. Freq.	Comm. %age
1	the	51284	6.6737	51284	6.6737
2	of	41322	5.3773	92606	12.051
3	in	24293	3.1613	116899	15.212
4	is	20457	2.6621	137356	17.874
5	and	17970	2.3384	155326	20.212
6	to	13142	1.7102	168468	21.923
7	a	11125	1.4477	179593	23.370
8	this	8441	1.0984	188034	24.469
9	on	6934	0.9023	194968	25.371
10	are	6666	0.8674	201634	26.239

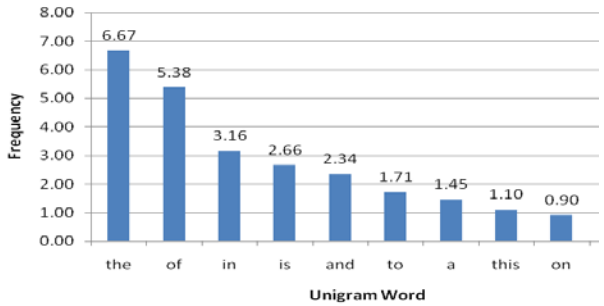

Fig. 3 Top 10 most frequent words

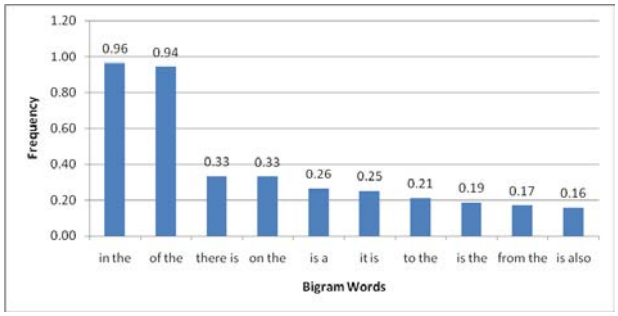

Fig. 4 Top 10 most frequent bigram words

3.2 Bigram Analysis

A bigram is every sequence of two adjacent elements in a string of tokens, which are typically letters, syllables, or words. Bigram Analysis is very useful in Natural language processing and in post processing in OCR and text compression. Here, we find the frequency occurrences of each adjacent pair of words in the text. Total number of bigrams present in the corpus is 768398 and there are 278228 unique bigrams. The results shows that the top 5 bigram words of the corpus are **in the, of the, there is, on the, is a**. These five bigrams covers approximately 2.9% of the whole corpus. Top 13443 bigram words occupy 50% and top 201389 bigram words occupy 90% of corpus.

Table II

Top 10 most frequently used bigram words

Rank	Word	Freq.	%age	Comm. Freq.	Comm. %age
1	in the	7409	0.964	7409	0.964
2	of the	7253	0.943	14662	1.908
3	there is	2550	0.331	17212	2.239
4	on the	2536	0.330	19748	2.570
5	is a	2020	0.262	21768	2.832
6	it is	1943	0.252	23711	3.085
7	to the	1609	0.209	25320	3.295
8	is the	1445	0.188	26765	3.483
9	from the	1310	0.170	28075	3.653
10	is also	1197	0.155	29272	3.809

3.3 Trigram Analysis

Trigram is a group or sequence of three adjacent letters or symbols or words. Trigram analysis can be used for machine translation between the natural languages. The total number of trigrams calculated in English corpus is 768348 out of which the unique trigrams are 565597. The results shows that the top 5 trigram words of the corpus are **there is a, a distance of, at a distance, the name of, the form of**. These five trigrams cover approximately 0.3% of the whole corpus. Top 181423 trigram words occupy 50%.

Table III

Top 10 most frequently used trigram words

Rank	Word	Freq.	%age	Comm. Freq.	Comm. %age
1	there is a	651	0.0847	651	0.0847
2	a distance of	473	0.0615	1124	0.1462
3	at a distance	463	0.0602	1587	0.2065
4	the name of	402	0.0523	1989	0.2588
5	the form of	312	0.0406	2301	0.2994
6	there is no	310	0.0403	2611	0.3398
7	in the form	271	0.0352	2882	0.3750
8	in the morning	228	0.0296	3110	0.4047
9	a height of	224	0.0291	3334	0.4339
10	of the body	219	0.0285	3553	0.4624

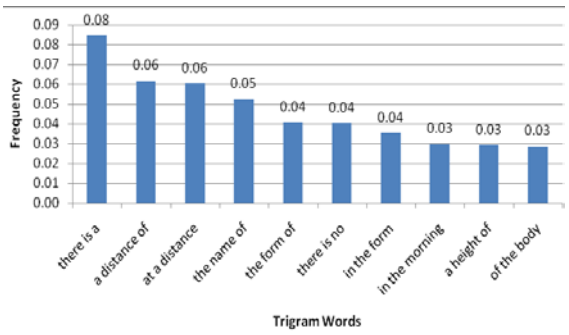

Fig. 5 Top 10 most frequent trigram words

Four grams has also been calculated. Total number of 4-grams are 768298 out of which unique 4-grams are 700360. The most frequent five 4-grams are **at a distance of, in the form of, at a height of, is at a distance, a distance of kilometers**. These five 4-grams cover approximately 0.16% of whole corpus.

3.4 Word Length Analysis

We have studied word length analysis of English corpus. It is very useful in the area like information storage and retrieval, automatic correction. The two largest word of length 21 characters found in the corpus are **hastpadangushthhasana** and **tatvavibhanjanshastra**. Both these words are transliterated words of Hindi language. The largest English word(length 20) found in the corpus is **paraphenylenediamine** which is the name of the compound in chemistry. A study on Wall Street Journal (WSJ) shows that for English the average word length is 5.04 at character level [20]. In our corpus the average word length has come out to be 4.82 when space is not considered in the character set. This increase in word length might be due to the fact that the type of corpus we have analyzed is in tourism and health domain. In both these domains, the terms used are of length more than the normal English terms.

Table IV
Frequency of Words with different Length

Word Length	Total Words	%age	Comm. %age
1	14097	2.20	2.20
2	14270	2.23	4.43
3	140216	21.89	26.32
4	132759	20.73	47.05
5	97996	15.30	62.35
6	68945	10.76	73.11
7	63163	9.86	82.98
8	43237	6.75	89.73
9	31090	4.85	94.58
10	17761	2.77	97.35
11	9311	1.45	98.81
12	4437	0.69	99.50
13	2110	0.33	99.83
14	703	0.11	99.94
15	190	0.03	99.97
16	99	0.02	99.98
17	48	0.01	99.99
18	36	0.01	100.00
19	6	0.00	100.00
20	3	0.00	100.00

English words up to length 8 covers approximately 90% of the corpus. Maximum number of words in the corpus is the words with length 3.

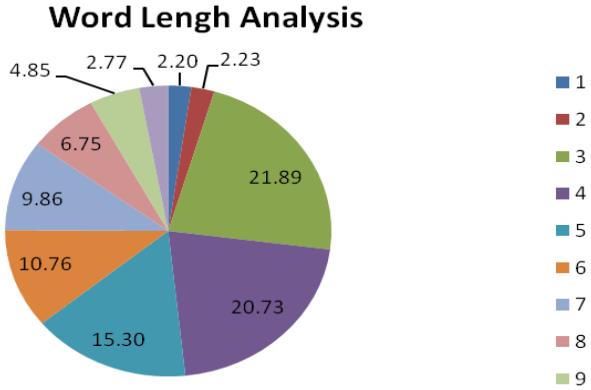

Fig. 6 Percentage of words with different length.

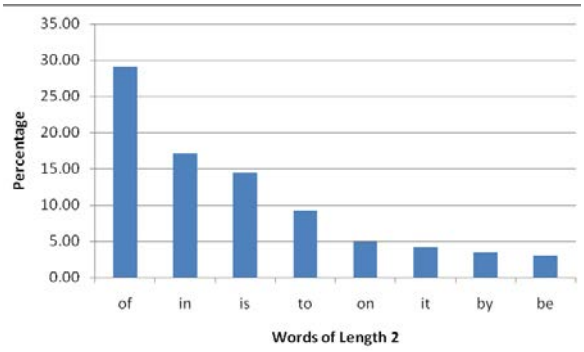

Fig. 7 Words with Length 2.

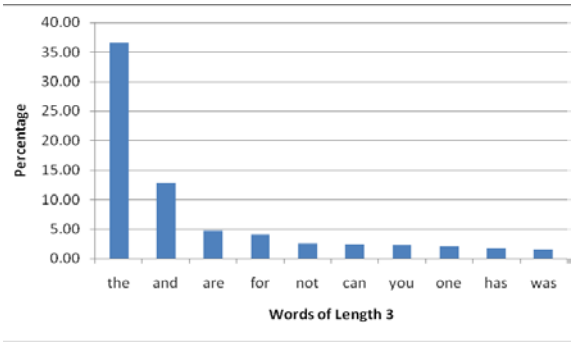

Fig. 8 Words with Length 3.

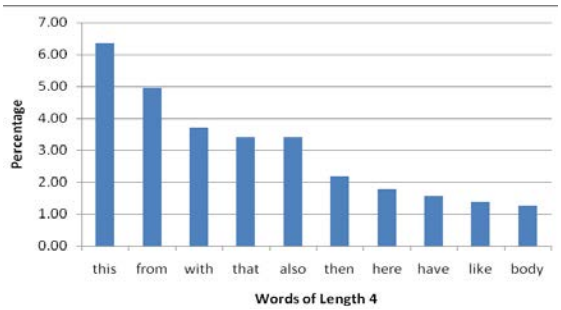

Fig. 9 Words with Length 4.

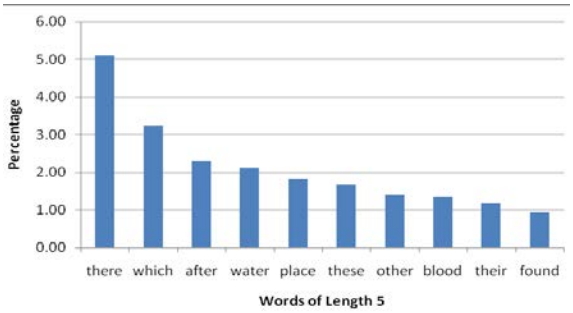

Fig. 10 Words with Length 5.

3.5 Words with Initial Character

Findings show that 4892 words distinct words start with its initial character as vowel. Similarly, there are 25179 words that start with a consonant. Approximately, 282 words with initial character as vowel cover 90% of the whole corpus. Only 6 unique words with initial character as vowel cover 50% of the corpus. In terms of words with initial character as consonant, about 3855 words cover 90% of the whole corpus where as about 185 words cover 50 % of the whole corpus.

The top 6 words with its initial character as vowel are **of, in, is, and, a, on**. The top 6 characters with its initial character as consonant are **the, to, this, from, for, there.**

Table V
Coverage of corpus with words having initial character as vowel or consonant

%age of Corpus	No. of Unique Words starting with vowels	No. of Unique Words starting with consonants
50%	6	185
60%	10	369
70%	23	728
80%	72	1512
90%	282	3855
100%	4892	25179

IV. CONCLUSION

The statistical analysis is needed for research community in the areas like automatic correction of misspellings, speech synthesis, and information storage and retrieval, machine translation, summarization and many more. The quantitative analysis of English text shows many aspects of this language. In English language, top 5 most frequent words occupy more than 20% of the whole corpus where as 185 words covers more than 50% of whole corpus. The average word length in the studied corpus comes out to be 7.43. With initial character as vowel, only 6 words covers more than 50% of the whole corpus. The results of these investigations can be applied to the processing of written English for various applications as stated.

REFERENCES

1. Majumder, Khair Md, and Yasir Arafat. "Analysis of and observations from a Bangla News Corpus." (2006).

2. Meyer, Charles F., ed. English corpus linguistics: An introduction. Cambridge University Press, 2002.

3. http://www2.ignatius.edu/faculty/turner/languages.htm

4. http://sanskrit.jnu.ac.in/index.jsp

5. Agrawal, Shyam S., Shweta Bansal Abhimanue, and Minakshi Mahajan. "Statistical Analysis of Multilingual Text Corpus and Development of Language Models."

6. Lalit Goyal: Comparitive analysis of Printed Hindi text and Punjabi text based on statistical parameters. Communications in Computer and Information Science, Volume 139, 2011, pp 209-213

7. Bharti, A., Sangal, R., Bender, S.M.: Some observation regarding corpora of some Indian languages. In: Proceedings KBCS 1998,pp. 203-213

8. Bharati, Akshar, Rajeev Sangal, and Sushma M. Bendre. "Some Observations Regarding Corpora of Some Indian Languages." Proc. Intl. Conf. Knowledge Based Computer Systems (KBCS-98). 1998.

9. Irene Langkilde: Generation that exploits corpus-based statistical knowledge. In proceeding COLING '98: 704-710, 1998.

10. V Goyal, GS Lehal: Advances in Machine Translation System. Language in India 2009.

11. Niladri S. Dash, "A Corpus Based Computational Analysis of the Bangla Language: A Step Towards Natural Language Processing". PhD Thesis submitted to ISI Calcutta, 2000.

12. K.W Church and R.L. Mercer, "Introduction to the special issue on computational linguistic using large corpora." Computational linguistic 19(1) : 1-23, March 1993.