n-Gram Character Analysis of English Text on Domain Specific Corpus

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

Abstract— *Statistical analysis of a language is a vital part of natural language processing. It refers to a collection of methods used to process large amounts of data and report overall trends. In this paper, frequency and word length analysis of individual characters in English text is performed. Unigram, bigram, trigram and positional analysis of characters in the domain specific English corpus in health domain has been studied. Miscellaneous analysis like Percentage occurrence of various numbers of distinct words and their coverage in English Corpus is studied.*

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

I. 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 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 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 Chinese, Japanese, Spanish has been compiled and analyzed to enrich the language knowledge [2].

II. STATISTICAL ANALYSIS

Statistical analysis 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 bytes 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 English corpus we worked on consists of grammatically tagged sentences. The whole corpus consists of 4488312 characters and 768448 words. Each word consists of a tag each with a defined meaning as 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 corpora into following format.

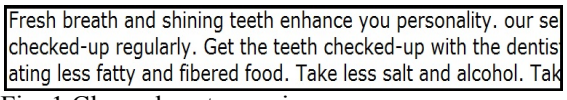
Fresh breath and shining teeth enhance you personality. our se
checked-up regularly. Get the teeth checked-up with the dentis
ating less fatty and fibered food. Take less salt and alcohol. Tak

Fig. 1 Cleaned sentences in corpus

The Corpus when read and analyzed after cleaning, it is found that it contained about 4488312 characters(including space characters) and 3709270 characters(without space), 768448 words, 1393753 vowels and 2204349 consonants. Among vowels, there are about 1370566 vowels with lower case and 23187 with upper case. Among consonants, the upper case consonants are about 88131 and lower case consonants are 2116036. Also it contains 4892 distinct words starting from vowels and 25179 distinct words starting from consonants.

3.1 Unigram Analysis

Unigram analysis is the study of characters taking a single character at a time. Here, in this English corpus, there are total 26 alphabets and 10 digits(digits are being used in the text). So the frequency count of these alphabets is the unigram analysis. The top 10 unigrams are E, A, T, I, O, N, S, R, H, L which covers approximately 73.56% of the corpus. Whereas top 18 characters cover more than 90 % of the corpus.

Table I
Most frequently used characters in English Corpus

Character	Freq.	%age Freq.	Comm. Freq.	%age *c.f.*
E	430653	11.897	430653	11.897
A	320340	8.849	750993	20.746
T	315264	8.709	1066257	29.456
I	280802	7.757	1347059	37.213
O	263430	7.277	1610489	44.491
N	253830	7.012	1864319	51.503
S	240314	6.638	2104633	58.142
R	223863	6.184	2328496	64.327
H	186540	5.153	2515036	69.480
L	147857	4.084	2662893	73.565
D	126986	3.508	2789879	77.073
C	108105	2.986	2897984	80.059
U	98528	2.721	2996512	82.781
F	96836	2.675	3093348	85.457
M	96529	2.666	3189877	88.123
G	76601	2.116	3266478	90.239
P	73163	2.021	3339641	92.261
B	60948	1.683	3400589	93.944
Y	60847	1.680	3461436	95.625
W	48580	1.342	3510016	96.967
K	34611	0.956	3544627	97.924
V	33950	0.937	3578577	98.861
J	8431	0.232	3587008	99.094
X	6180	0.170	3593188	99.265
0	4835	0.133	3598023	99.399
1	4102	0.113	3602125	99.512
Q	2790	0.077	3604915	99.589
2	2643	0.073	3607558	99.662
5	2154	0.059	3609712	99.722
Z	2124	0.058	3611836	99.780
3	1744	0.048	3613580	99.828
4	1438	0.039	3615018	99.868
9	1282	0.035	3616300	99.904
6	1274	0.035	3617574	99.939
8	1186	0.032	3618760	99.972
7	1011	0.027	3619771	100

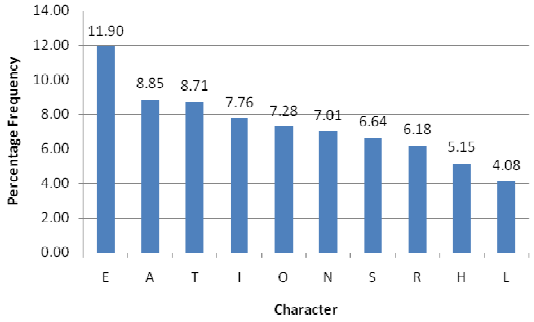

Fig. 2 Top 10 most frequent characters in English

3.2 Bigram Analysis

Bigram analysis is the study of two adjacent characters taken at a time. Here, the frequency of every two adjacent characters is calculated. Bigram analysis is very useful in natural language processing, optical character recognition, text compression, etc. The total number of bigrams found in the corpus is 2914197 where as the distinct number of bigrams are 673. Top 10 bigrams are **th, he, in, an, re, er, is, of, on, es,** which cover about 20% of whole corpus. About 44 bigrams cover approximately 50% of the corpus. 90 bigrams cover around 70% of the corpus. Below is table showing top 20 bigrams with its frequency and commulative frequency.

Table II
Top 20 frequently used bigram characters

Bigram Char	Freq.	%age Freq.	Comm. Freq.	%age *c.f.*
th	102059	3.502	102059	3.502
he	83678	2.871	185737	6.373
in	81453	2.795	267190	9.168
an	54881	1.883	322071	11.05
re	53836	1.847	375907	12.89
er	49823	1.709	425730	14.60
is	45662	1.566	471392	16.17
of	42282	1.450	513674	17.62
on	40922	1.404	554596	19.03
es	40070	1.374	594666	20.40
ar	39588	1.358	634254	21.76
at	37874	1.299	672128	23.06
en	36043	1.236	708171	24.30
ng	34673	1.189	742844	25.49
al	33769	1.158	776613	26.64
nd	32905	1.129	809518	27.77
ti	31035	1.064	840553	28.84
or	30605	1.050	871158	29.893
te	30070	1.031	901228	30.925
it	28916	0.992	930144	31.917

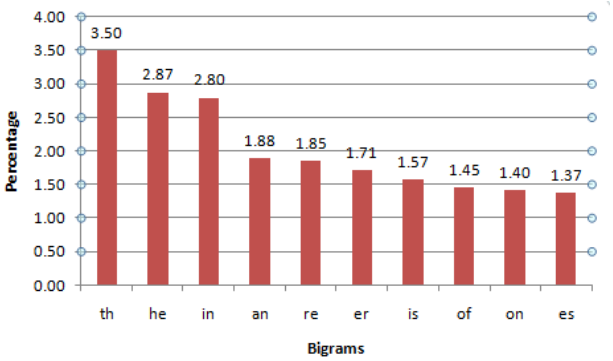

Fig. 3 Top 10 most frequent bigrams

3.3 Trigram Analysis

Trigram analysis is the study of three adjacent characters taken at a time. Here, the frequency of every three adjacent characters is calculated. Trigram analysis is very useful in machine translation applications. The total number of trigrams found in the corpus is 2164695 where as the distinct number of bigrams are 6832. Top 10 trigrams are **the, ing, and, her, ent, ion, ere, tio, thi, for,** which cover about 9% of whole corpus. About 47 trigrams cover approximately 20% of the corpus. 279 trigrams cover around 50% of the corpus. About 621 trigrams cover approximately 70% of the corpus. Below is table showing top 20 bigrams with its frequency and commulative frequency.

Table III
Top 20 frequently used trigram characters

Trigram Char	Freq.	%age Freq.	Comm. Freq.	%age *c.f.*
the	68443	3.161	68443	3.161
ing	27382	1.264	95825	4.426
and	22822	1.054	118647	5.481
her	13301	0.614	131948	6.095
ent	12414	0.573	144362	6.668
ion	12166	0.562	156528	7.230
ere	11288	0.521	167816	7.752
tio	10712	0.494	178528	8.247

thi	10654	0.492	189182	8.739
for	10104	0.466	199286	9.206
his	9984	0.461	209270	9.667
ter	9899	0.457	219169	10.12
ati	9558	0.441	228727	10.56
are	9308	0.429	238035	10.99
ate	8864	0.409	246899	11.40
rea	7585	0.350	254484	11.75
ome	7117	0.328	261601	12.08
all	7031	0.324	268632	12.40
rom	7020	0.324	275652	12.73
fro	6932	0.320	282584	13.05

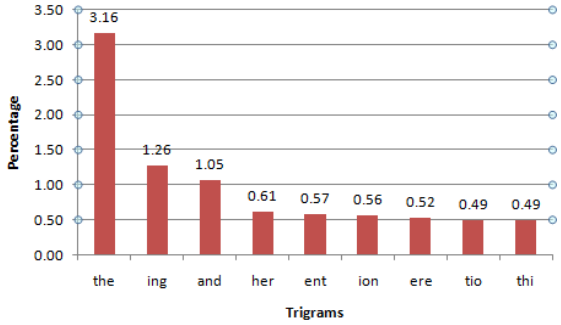

Fig. 4 Top 10 most frequent trigrams

3.4 Average Word Length Analysis

We have studied average word length of English corpus. It is very useful in the area like information storage and retrieval. We have considered maximum 18 length words covering nearly the whole corpus. A study on Wall Street Journal (WSJ) shows that for English the average word length is 5.04 at character level [20]. However, in our analysis we found it to be 4.82, here, we have not considered the space character in our character set. The following table shows the number of different length words with cumulative frequency:

Table IV
Words of different length in English text

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

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.

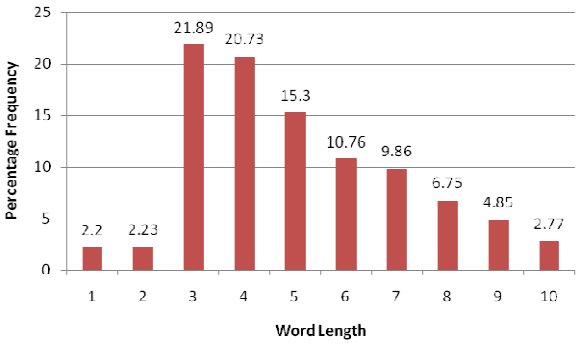

Fig. 5 Percentage of words with different Length

In our study (Table 2) we found that words with three characters are most frequent followed by words with four characters.
The graph below shows the usage of words against word length.

Figure below is a list containing samples of words that are more than fifteen characters in length. While carefully inspection makes it clear that most of these words suffer from missing word boundary such as a space character or are hyphenated words.

blood-circulation, non-governmental, reading-writing, tra
Glaxosmithkline, Non-communicable, excommunication,
post-childbirth, anti-tuberculosis, multi-blistered,
gastrointestinal, medical-philosophy, unicompartmental,
grandmastership, Pharmaceuticals, unconsciousness, co
neurobiological, Gastroenteritis, knot-endometriosis,

Fig. 6 Words of length more than 15 characters

IV. CONCLUSION

The statistical analysis is needed for research community in the areas like automatic correction of misspellings, speech synthesis, and information retrieval. The quantitative analysis of English text shows many aspects of this language. In English language, top 10 most frequent characters occupy more than 73% of the whole corpus, e is most frequently used character and further its most common position in the word is third. Top 30 bigrams of English text at character level occupy almost 40% of the whole text. As large number of trigrams is calculated at character level thus top 30 trigrams occupy about 15% of whole corpus. The results of these investigations can be applied to the processing of written English for practical purpose.

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. Yannakoudakis, E.J., Tsomokos, I., Hutton, P.J. :N-grams and their implication to natural language understanding. Pattern Recognition 23(5), 509 – 528.
10. Church, K.W., Mercer, R.L.: Introduction to the special issue on computational linguistic using large corpora. Computational Linguistic 19(1), 1 – 23.
11. Gurpreet Singh Lehal: "Corpus Based Statistical Analysis of Punjabi Syllables for Preparation of Punjabi Speech Database". International Journal of Intelligent Computing Research (IJICR), Volume 1, Issue 3, June 2010.
12. Irene Langkilde: Generation that exploits corpus-based statistical knowledge. In proceeding COLING' 98: 704-710, 1998.
13. Douglas Biber: Methodological issues regarding Corpus-based analysis of linguistic variations. Lit Linguist Computing (1990) 5 (4): 257-269.
14. V Goyal, GS Lehal: Advances in Machine Translation System. Language in India 2009.
15. Niladri S. Dash, "A Corpus Based Computational Analysis of the Bangla Language: A Step Towards Natural Language Processing". Ph.D. Thesis submitted to ISI Calcutta, 2000.
16. E.J. Yannakoudakis, I. Tsomokos and P.J.Hutton," N-grams and their implication to natural language understanding". Pattern recognition, 23(5) : 509-528, 1990.
17. 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.
18. Tao Hong, "Degrading text recognition using visual and linguistic contexts ", A Ph.D. dissertation submitted to the faculty of Graduate School of State University of New York at Buffalo.(Sept. 1995).
19. R.C. Eldridge, "Six Thousand common English Words", Nigeria Falls, New York, 1911.
20. J. Karlgren, "Stylistic Experiments in Information Retrieval". In Proceedings NeMLaP 2, Bilkent University, Ankara, September, 1996.