Readability of Open Education Resources (OERs): A key to success of ODL for the young people

Anurag Saxena, Indira Gandhi National Open University (IGNOU), India
s0anurag@hotmail.com

ABSTRACT

According to Baumel (2002), “20 percent of school age kids are poor readers and remain that way through their lifetime”. It has been proved repeatedly that reading is a language-based skill. Poor reading skills thus results in attainment of poor educational levels. It implies that even if a learner is interested in reading a book, he finds it difficult to read it independently. The technology catches the attention of young people and children. The young ones show a positive trend in learning from newer forms of instruction and instructional resources.

A few open resources that have proved their worth in recent times are Project Gutenberg (oldest producer of free e-books on the Internet http://www.gutenberg.org/), public domain electronic texts ((http://www.infomotions.com/etexts/) and open education resources (OERs) of WikiEducator (http://www.wikieducator.org/Main_Page). It is quite possible to provide standardized learning materials to the young learners with the help of above-mentioned resources.

Now the question arises about the readability of these resources. If these resources have a poor readability then they will not be so useful for learners with not so good reading skills and thus resulting in poor education levels. Flesch Readability Index is deemed as a standard as far as the readability of the documents is concerned. It is also said that writers use Zipf’s principle of least effort to simplify communication and that Zipf’s law is applicable in understanding human language. Zipf’s law explains the equilibrium between uniformity and diversity in usage of words.

In this communication, we have taken some sets of texts from aforesaid sources and tried to analyze them to investigate the readability of the document and the Zipf’s coefficients. The paper also discusses the implications of these results for ODL for the young people.

INTRODUCTION

Zipf (1949) in his work, “Human Behavior and the principle of least effort” viewed language as a "tool" that is shaped by its "jobs" in human society. Other works of Zipf were “Selective Studies and the Principle of Relative Frequency in Language” which as published in 1932 and “Psycho-Biology of Languages” which was published in 1935.

Many years after his death linguistics agreed that speakers simplify communication by using a small pool of words that they can retrieve quickly from their memory and listeners simplify communication by preferring words with a single and unambiguous meaning. This proved that Zipf’s law is applicable in understanding human language.

Zipf searched for a principle of least effort that would explain the equilibrium between uniformity and diversity in usage of words. Most others searched for a probabilistic explanation. The burning question still remains- Do we have any new evidence that Zipf’s explanation of principle of least effort is more correct than a statistical explanation?

Flesch Readability Index on other hand has become a sort of a standard as far as the readability of the documents is concerned. At many places, it has become imperative to ascertain that the document/ forms have a high value of Flesch Readability Index, so that it is understood by masses.

Zipf's Law

Zipf formulated a law in 1930 that says frequency count (number of occurrence) of words in any text is inversely proportional to the rank of that word. In other words, the distribution of words adhered to a regular statistical pattern or "The probability of occurrence of words or other items starts high and tapers off exponentially. Thus, a few occur very often while many others occur rarely" (Black, 2000).
To further, explain the basic form of the law,

frequency * rank has a inversely proportional relationship:
frequency * rank = constant or f * r = c

Zipf attributed this law as a consequence of "Principle of Least Effort". The Principle of Least Effort postulates that a person would like to communicate in such a way as to minimize his total effort. Altmann(2002) commented that Zipf's ideas are the foundation stones of modern quantitative linguistics and his influence is not restricted to linguistics but incessantly penetrates other sciences. Mandelbrot (1953) tried to discuss Zipf's law in terms of communication costs and explained that the communication costs increases as the number of words and their length grows. Ferrer-I-Cancho & Sole (2001a) commented that many models of syntactic communication assume this law. It is an obvious ingredient for any theory of language evolution. According to Li (2002), the number of times a word is used in written human languages and the frequency of usage are the variables that indulge in a Zipf's type distribution. Smith & Devine (1985) found that legal texts also follows Zipf's law but in a little different manner. Francis & Kucera (1964) applied the Zipf's law to the Brown corpus of 1 million words of American English. Le Quan Ha et al. (2002) analyzed Zipf's law for large corpora in two languages, English (from the Wall Street journal) and Mandarin (from the People's Daily Newspaper and the Xinhua News Agency. Wang (1989) presented Zipf's distribution of Chinese corpus and Wyllys (1981) took a data set of 3907 English words. Sun et al. (1999) commented, "Studies of word frequency have many interesting and potentially significant applications. For example this model could be used to evaluate a single article or an author's work. Assuming a reasonable level of skill among the writers whose works are the basis for our observations, we can use this model as a benchmark for assessing writer's language skills". Gelbukh and Sidorov (2001) observed that the coefficients of Zipf law are different for different languages. Ferrer-I-Cancho and Sole (2001b) showed that the co-occurrence of words in sentences relies on the network structure of the lexicon. They analyzed the properties in depth and commented that human language can be described in terms of a graph of word interactions.

Flesch Readability Index

For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ \ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

According to Wikipedia, the free encyclopedia, a syllable is a unit of organization for a sequence of speech sounds. Syllables are often considered the phonological "building blocks" of words. They can influence the rhythm of a language.

Flesch readability index can be related to the educational level of the audience. For example a score of 91-100 can be easily comprehensible by a 5th grade student, a score of 51-60 understandable by a High School student, a college graduate will be able to comprehend a document with score 31-50 and a document with score less than 0 can be understood by a Law School Graduate only.

RESEARCH QUESTION

If a document has a high Flesch Readability Index, then whether the Zipf's curve will fit this document in a better manner. In other words, if a document is fairly easy to understand, then whether it will follow the Zipfian distribution? Whether Zipf's law is applicable in understanding the human language? And lastly, if there are any implications of these results for ODL for the young people.

DATA

To investigate whether there is a relation in readability of a document and the Zipf's coefficient, we have selected the following sets of text from diverse sources.

- **English:** The Project Gutenberg e-text of "Aladdin and the Wonder Lamp", a "public domain" work distributed by Professor Michael S. Hart through the Project Gutenberg Association. Project Gutenberg is the oldest producer of free e-books on the Internet (http://www.gutenberg.org/).
- **Library Science:** The Project Gutenberg e-text of "The Library", by Andrew Lang #20 in our series by Andrew Lang, December, 1999.
- **E-texts over time:** Public domain electronic texts (e-texts) in the areas of American and English literature as well as Western philosophy are taken in this category. These were "classic" texts that have stood the test of time. They also encompass a huge time period- as far back as 400BC to the present. (http://www.infomotions.com/etexts/)
- **Popular e-texts:** Popular e-texts like "365 Foreign Dishes", "The Arabian Nights Entertainments", "The Arctic Queen" and "The Atomic Bombings of Hiroshima and Nagasaki" were also taken to investigate the relationship.
- **Management:** Few chapters from the Human Resource Management Content on the WikiEducator (http://www.wikieducator.org/Human_Resource_Management)

METHODOLOGY

We have tried many software for calculating the word frequency from the text. We searched the World Wide Web (www) for freeware or shareware, which can do this work. We found four major software in this category. These were Hermetic Word Frequency Counter 5.32, Textanz Word and Phrase Frequency Counter v.1.3, Fore Words Pro 1.2.0.41 and TextSTAT. We tried to analyze various text files with these software. The first three software calculated the frequencies but since we were using the demo version, we faced a major limitation of not been able to transfer the output to a file. We therefore switched to TextSTAT which is completely free software. Thus the Software for calculating the word frequency from the texts used in this work is "Text STAT". Text STAT is a simple program for the analysis of texts made by Free University of Berlin. It produces word frequency lists and concordances from ASCII/ANSI texts, MS Word and HTML files. Text STAT can be downloaded from the website http://www.niederlandistik.fu-berlin.de/textstat/software-en.html. We found the Zipfian data for the various files in order to find the applicability of Zipf-Mandelbrot law. Mandelbrot assumed that the aim of language is to transmit the most information per symbol with the least effort. He proposed the following relationship:

$$f = k(r + c)^{-\theta}$$

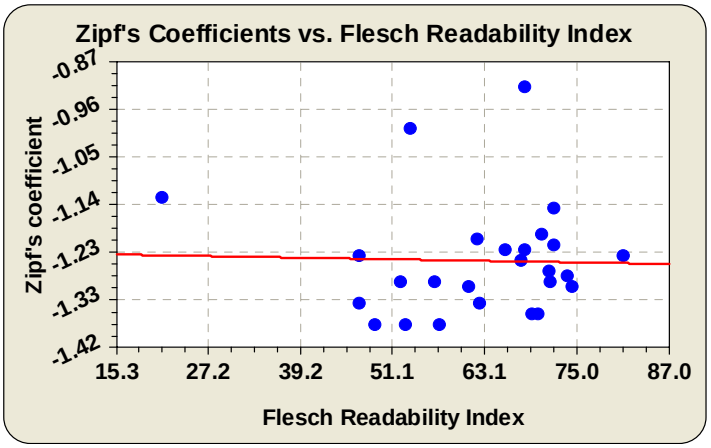
Where, f is the frequency and r is the rank of the word; c and θ are constants. Here, c improves

the fit for small r and the exponent θ improves the fit for large r. A data follows Zipfian distribution if the exponent θ remains close to –1.

FINDINGS

Appendix I illustrates the documents with related statistics on number of words in the document, Flesch Readability Index, Zipf's coefficient, number of sentences, number of syllables per word and the number of words per sentence. Based on this, we tried to analyse documents primarily with respect to values of the Flesch Readability Index and the Zipf's coefficient.

Many documents had excellent value of Zipf's coefficient and also good readability (understandable by a high school level reader). This tend to show that Zipf's law is applicable in documents that have on an average 1.5 syllables per word and have 5-8 words per sentence. Some documents however nullified the claim that was found in documents mentioned above. Almost all these documents have Zipf's coefficient ranging from -1.20 to -1.37, but had variable readability indexes ranging from 46-80. No trend has either been found in the syllables per word and words per sentence. There is one document (from WikiEducator) that has poor readability. This could be an example of a document with bad readability and might not be truw for all the documents.



CONCLUSION

We have tried to make the sample as diverse as possible. We finally found that 23 documents having different readability indices, belong to a different genre and belong to different time periods but have almost similar value for the Zipf's coefficient. This indicates that readability has little to do with the Zipf's coefficients.

This led us to go back to our research question that if a document has a high Flesch Readability Index, then whether the Zipf's curve will fit this document in a better manner. Most documents partially demonstrate this as they have excellent value of Zipf's coefficient and also good readability that varies from 46 to 80.

Coming to the next research question, whether Zipf's law is applicable in understanding the human language? Can it be used as benchmark for assessing a writer's skill? The findings in this communication supports this claim. It is because of the fact that Zipf's principle of least effort says that a writer simplifies communication by using a small pool of words from their memory. This would mean that these communications ought to have good readability indices too. So, all those documents that have good readability coefficients should have good Zipf's coefficient also. This is reflected in the findings. So it is proved that Zipf's law is applicable in understanding human language. This is also in line to Sun et al. (1999) comment that "we can use this model as a benchmark for assessing writer's language skills".

IMPLICATIONS FOR ODL FOR YOUNG PEOPLE

The paper does not solely address the technical aspects of the readability indexes. Since OERs are poised to improve access to education for children and young people, the paper has raised certain vital issues that are crucial for improving access to education for children and young people. We have infact proved more than this. We have proved that OER's should have good readability index (As it will be easier for the learners to follow) and should have good Zipf's coefficient (As the contributors of the content would use least effort in creating them). Sources of OER's should try to ensure readability issues before uploading the documents. We have thus provided a framework for the good OER's.

We wanted to extend our study on issues like gender, intra and inter source variability also. However at this stage it would be too early to consider the implications from a gendered perspective. For example, do boys struggle with reading more than girls? This and other issues can be deliberated in another study.

REFERENCES

1. Altmann, Gabriel (2002), "Zipfian linguistics", *Glottometrics* 3, pp 19-26, 2002
2. Black Paul E. (2000), "Zipf's Law: Definition", at http://hissa.nist.gov/dads/HTML/zipfslaw.html site accessed on 1/29/01.
3. Cancho Ferrer Ramon, and Sole Richard V. (2001a). "Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited". *Journal of Quantitative Linguistics* 8:165-174.
4. Cancho Ramon Ferrer I and Sole Ricard V (2001b), "The small world of human language", *Proc. R. Soc. Lond.* B (2001) 268, 2261-2265
5. Francis, W. N. & Kucera, H. (1964). "Manual of Information to accompany a standard corpus of present-day edited American English for use with digital computers" Department of Linguistics, Brown University, Providence, Rhode Island
6. Gelbukh Alexander, Sidorov Grigori (2001), "Zipf and Heaps Laws' Coefficients Depend on Language", *Proc. CICLing-2001*, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.
7. http://users.dickinson.edu/~braught/courses/cs132f02/labs/lab07.html for information about Flesch Index. Site accessed on 02/06/2006.
8. Jan Baumel, M.S.,(2002) Learning to Read — Research Informs Us, Charles and Helen Schwab Foundation, http://www.schwablearning.org/articles.aspx?r=22 site accesses 18/11/07
9. Le Quan Ha, E. I. Sicilia-Garcia, J. Ming and F. J. Smith. 2002. Extension of Zipf's Law to Words and Phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, pages 315-320, Taipei, Taiwan.
10. Li, Wentian (2002), "Zipf's Law Everywhere", *Glottometrics,* 5, 2002, 14-21
11. Mandelbrot B.B. (1953), "An informational theory of the statistical structure of languages", in *Communication Theory*, ed. W. Jackson (Butterworth, 1953) , pp. 486-502.
12. Smith, F. J. & Devine K. (1985), "Storing and Retrieving Word Phrases" *Information Processing & Management*, Vol. 21, No. 3, pp 215-224.
13. Sun Qinglam, Shaw D. , Davis C.H. (1999), " A model for estimating the occurrence of same frequency word and the boundary between the high and low frequency words in texts", *Journal of the American Society for Information Science*, Mar 1999: 50, 3
14. Wang, C.(1989), " Zipf's distribution of Chinese corpus", *Information Sciences*, 10, 1-8
15. Wyllys, R.E. (1981), "Empirical and theoretical bases of Zipf's law", *Library Trends*, 30, 53-64
16. Zipf G.K. (1932), Selective Studies and the Principle of Relative Frequency in Language.

17. Zipf G.K. (1935), Psychobiology of Languages, Houghton-Mifflin, 1935; MIT Press.
18. Zipf G.K. (1949), *Human Behavior and the principle of least effort,* Cambridge, MA: Addison-Wesley Press
19. Zipf G.K., *Human Behavior and the principle of least effort: An Introduction to Human Ecology*, New York, 1965.

Appendix I: Document Statistics and Zipf's Law

S.no	File Name	No- of Words	Flesch Index	Zipf's Coefficient	No- of Sentences	Syllables per word	Word per sentence
1	aladdin eng.txt	5319	68.23	-0.92	661	1.54	8.05
2	librarys.txt	37498	53.33	-1	5037	1.73	7.44
3	jefferson-autobiography-73.txt	40648	48.76	-1.37	5326	1.78	7.63
4	wollstonecraft-maria-196.txt	45874	52.75	-1.37	5426	1.72	8.45
5	franklin-autobiography-244.txt	68157	57.27	-1.37	7270	1.66	9.38
6	chaucer-canterbury-102.txt	99403	69.18	-1.35	13578	1.54	7.32
7	augustine-confessions-276.txt	176014	69.99	-1.35	22974	1.53	7.66
8	mill-subjection-217.txt	45240	46.75	-1.33	5108	1.79	8.86
9	Arabian nights entertainments.txt	90768	62.41	-1.33	10672	1.61	8.51
10	aristotle-meteorology-80.txt	43470	60.99	-1.3	5030	1.62	8.64
11	freud-young-763.txt	72133	74.4	-1.3	11160	1.49	6.46
12	berkeley-treatise-177.txt	36342	52.17	-1.29	4115	1.72	8.83
13	locke-concerning-111.txt	53786	56.56	-1.29	5732	1.66	9.38
14	barrie-peter-277.txt	47885	71.56	-1.29	6906	1.52	6.93
15	bunyan-pilgrims-304.txt	57122	73.73	-1.28	7241	1.48	7.89
16	anonymous-beowulf-543.txt	27129	71.35	-1.27	4173	1.52	6.5
17	dickens-christmas-125.txt	21818	67.75	-1.25	3301	1.56	6.61
18	hiroshima nagasaki.txt	25341	46.75	-1.24	3313	1.8	7.65
19	twain-tom-40.txt	24486	80.99	-1.24	3564	1.41	6.87
20	lucretius-on-395.txt	75386	65.78	-1.23	10549	1.58	7.15
21	keats-endymion-484.txt	31962	68.19	-1.23	4847	1.56	6.59
22	365 foriegn dishes.txt	27891	72.03	-1.22	4424	1.52	6.3
23	The arctic queen.txt	16703	62.09	-1.21	2451	1.63	6.81
24	shakespeare-hamlet-25.txt	33098	70.42	-1.2	4931	1.53	6.71
25	shakespeare-romeo-48.txt	26784	71.97	-1.15	3854	1.51	6.95
26	HRM from WikiEducator	1627	21.24	-1.13	202	2.10	8.05