International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

Frequency Distribution Fitting for Electronic Documents

Arockia David Roy Kulandai[1,2,*]

[1]*Department of Computer Science, Klingler College of Arts and Sciences, Marquette University, Milwaukee, Wisconsin, U.S.A.*
[2]*St. Xavier's College (Autonomous), Ahmedabad, Gujarat, India*
[*]*Corresponding Author: david.roy@marquette.edu*

(Received 19-09-2020; Revised 21-09-2020; Accepted 21-09-2020)

Abstract

Studies of frequency distributions of natural language elements have identified some distributions that offer a good fit. Using electronic documents, we show that some of these distributions cannot be used to model the frequency of bytes in electronic documents even if these documents represent natural language documents.

Keywords: Frequency fitting, quantitative linguistics, phase change memories

1 Introduction

Mathematical linguistics has studied the frequency of phonemes, words, and graphemes in natural languages. At its best, this work is used to obtain linguistical insights or even applications. For example, the Flesch Reading Ease Index [1] uses a combination of average word length and average sentence length. Best [2] still upholds its usefulness but notices that word length and sentence length are only indirectly related to readability. Our own motivation does not stem from linguistics but from the study of new non-volatile memory devices and their integration into future systems. We

Volume 3, Issue 1, pages 1–10
p-ISSN 2655-8564, e-ISSN 2685-9432

are interested in researching how to minimize bitflips in Phase Change Memories (PCM) [3]. PCM are a new non-volatile memory technology that offer bye-addressability, very high density, non-volatility, high retention, and high capacity. Unfortunately, PCM exhibit limited endurance. They use energy only while reading and writing, and usually writing consumes most of the energy. The number of bitflips caused by overwriting electronic documents of one kind by documents of the same kind depends on the encoding. For example, the web-browser cache contains HTML documents which could be placed in the same area of a PCM. To find good encodings, we want to model the frequency of graphemes in these documents [3]. The most frequent encoding for internet documents is UTF-8 so that our graphemes are bytes.

Here, we apply the methods of mathematical linguistics to modelling the frequency of bytes. Linguists are interested in language and graphemes are important as carriers of information on phonems. Unlike linguistics, we are interested in the effects of storing graphemes instead of using them. This makes for important differences. For instance, a linguist is not likely to make a distinction between capital letters and non-capital letters. Similarly, a linguist might conflate equivalent spellings, for example, the English and the US English versions of “tre” and “ter”, the recent abolition of the German letter “ß” in favor of “ss”, or even remove accents in Spanish.

Linguistics has shown that the frequency distribution of graphemes can be modelled by one or two parameter distributions successfully. Our results show that distribution fitting is less successful for bytes than for letters and phonems. Our research has convinced us that modelling a broad category such as text documents using distributions and parameters fitted to one corpus does not translate to another corpus. Evaluation of byte overwrites using these models are dangerous. Fortunately, we did find an encoding strategy that leads to energy savings for a broad class of electronic documents [3].

2 Research Methodology

We observed that encoding, e.g. utf-8, utf-16, ASCII, has a strong impact on the number of bits over-written when string text based electronic documents. This translates immediately into energy savings because each bit over-write costs energy. Also, each

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

bit-write is potentially destructive of the cell. We, therefore concentrated on HTML files stored, for example, in a browser's cache and to a lesser extent on text files. For comparison, with results in linguistics [1], [4], [5], [6] we also extracted pure text content from HTML files by gathering long text between the paragraph meta tag if the text was at least 50 bytes long. This excludes instances where the webpage used a paragraph meta tag only as a structural element. We also only processed letters and did not include punctuations or space. We collected corpora from Internet newspaper articles, Wikipedia, and the Project Gutenberg library of books in four European languages namely, English, German, Spanish and French. Each corpus contained at least 10 MB of raw data. We gathered ten corpora for English and five each for the other languages. For each corpus, we then calculated the frequency of each letter in the language or the frequency of each possible byte. We then fitted various distributions proposed in the linguistics literature to the frequency tables we obtained. For fitting we used Python's SciPy module. We minimized the relative sum of squared differences between the ordered relative frequency of the letters or bytes and the prediction by the distribution. Since the distribution has one, two, or three parameters, this means minimizing a function of one, two, or three variables. For each distribution, and for each of the 25 corpora, we tabulated the best fitting parameters and the goodness of fit for bytes.

3 Distributions

Zipf is an ancestor of modern quantitative linguistics, but the distribution named after him is also used almost as a default when modelling uneven usage of resources or uneven sizes in Computer Science. He ranked words in descending order of frequency of occurrence and observed that the frequency of the i^{th} word is proportional to $1/i$. Thus, we fit an ordered array of n descending frequencies with an array:

$$[\alpha/1, \alpha/2, \alpha/3, \ldots, \alpha/n]$$

where α is chosen so that the array sums up to one, which means that α is the inverse of the n^{th} harmonic number. Over time, many other distributions have been proposed to

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

model frequency of elements in natural languages. In his later works, Zipf generalized his distribution, matching the ordered array of n descending frequencies with

$$\left[{}^{\alpha}/_{1^{\beta}}, {}^{\alpha}/_{2^{\beta}}, {}^{\alpha}/_{3^{\beta}}, \dots, {}^{\alpha}/_{n^{\beta}}\right],$$

where β is a parameter of the text and α is calculated from β and the length n of the frequency array because the Probability Density Function (PDF) needs to sum up to 1. With other words, the frequency of the i^{th} most frequent item, denoted by f_i, is

$$f_i \sim {}^{1}/_{i^{\beta}},$$

where $\sim$ denotes proportionality. This distribution is also known as the Power Law distribution. Mandelbrot generalized the Zipf distribution by adding a second independent parameter γ so that

$$f_i \sim {}^{1}/_{(i+\gamma)^{\beta}}.$$

The Good distribution [7] is a parameter-less distribution where

$$f_i \sim \sum_{j=i}^{n} \frac{1}{j}.$$

We parameterize the Good distribution by setting

$$f_i \sim \sum_{j=i}^{n} \frac{1}{j^{\alpha}}.$$

In addition, we went through a list of distributions given by Li and Miramontes [5].

Exponential:	$f_i \sim \exp(-\alpha i)$
Logarithmic:	$f_i \sim 1 - \alpha \log(i)$
Quadratic Logarithmic:	$f_i \sim 1 - \alpha \log(i) - \beta (\log(i))^2$
Weibull:	$f_i \sim \log((n+1)/i)^{\alpha}$
Cocho – Beta:	$f_i \sim (n+1-i)^{\beta}/i^{\alpha}$
Frappat:	$f_i \sim \beta i + \exp(-\alpha i)$
Yule:	$f_i \sim \beta^i / i^{\alpha}$
Menzerath-Altmann:	$f_i \sim \exp\left(-\frac{\beta}{i}\right)/i^{\alpha}$

The actual value of the PDF of a distribution with $f_i \sim \psi(i, \alpha, \beta)$ is $c\big(\psi(i, \alpha, \beta)\big)$, where $1/c$ is equal to $\sum_{i=1}^{n} \psi(i, \alpha, \beta)$. The purpose of c is to ensure that the PDF sums up to 1.

International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

n is the number of symbols obtained, namely, $n = 256$ for bytes and $n =$ the number of symbols in a language. We are following the notation of Li and Miramontes [5], which has idiosyncrasies. For some values of β, Yule and Menzerath-Altmann are virtually indistinguishable. What Li and Miramontes call the Yule distribution is in fact not the well-known Yule-Simon distribution. The Yule-Simon distribution would have $f_i \sim \alpha B(i, \alpha + 1)$, where B is the beta function, but is not suited for frequency matching.

4 Results

There are two criteria for a distribution fit for modelling. Most importantly, the distribution should predict the frequency well. We measure this by calculating the sum of the differences squared and dividing it by the number of symbols. The number of symbols n is equal to 256 when we process raw documents, consisting of bytes. For text, it is just the total number of letters that can appear. To allow comparisons between text and raw data, we divide by n.

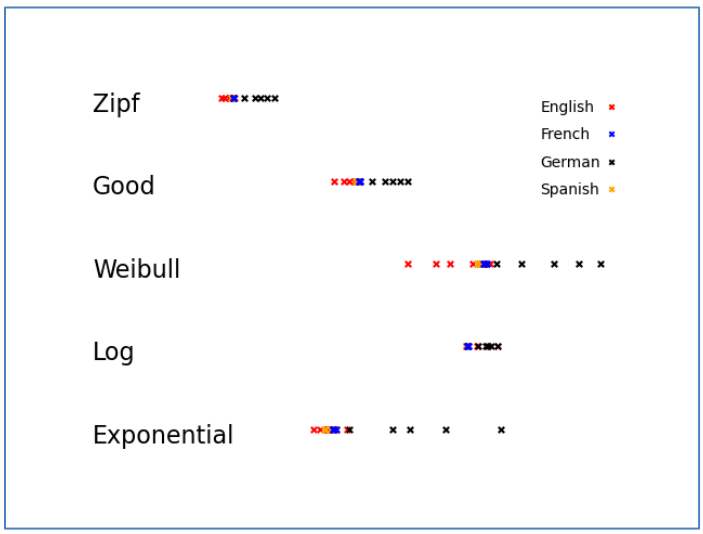

Figure 1. Distribution of the parameter for fitted one parameter distributions.

The second criterion is good clustering of the parameters. If two different corpora can be fitted well to the same distribution but with widely different parameters, then either we have too many parameters or the parameters are specific to one corpus. In the first case, we are better off with a distribution with less parameters and in the second case

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

the distribution with these parameters does not generalize and is not suitable for modelling.

For one parameter distributions, the fitted parameters lie close together and often in bands determined by the language, Figure 1. Only the parameters for German raw documents are more spread out in the case of the Weibull distribution and the Exponential distribution. In Figure 1, we plotted the sole parameter along the x-axis multiplying the parameter for the Logarithmic distribution by 10 and the parameter for the Exponential distribution by 20. Because the best fitting parameters in general appear in small ranges with sometimes differences between the languages, we conclude that modeling byte distribution with a single parameter will apply across a broad spectrum of corpora as long as they are in the same language.

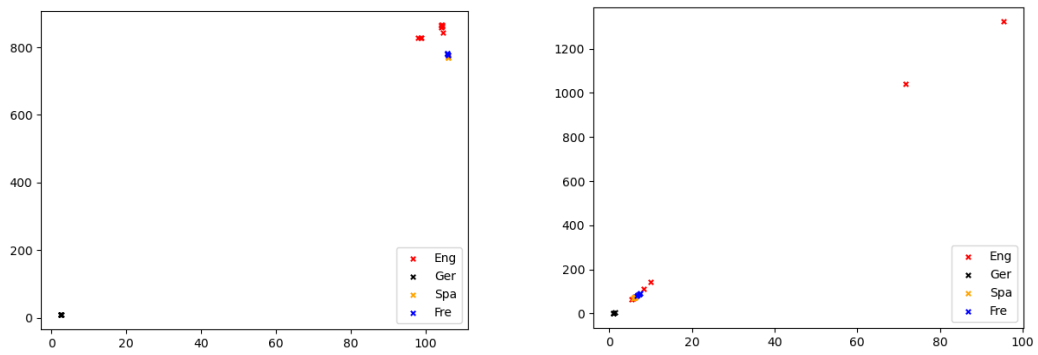

Figure 2. Parameters for Zipf Mandelbrot for text (left) and raw HTML (right) corpora.

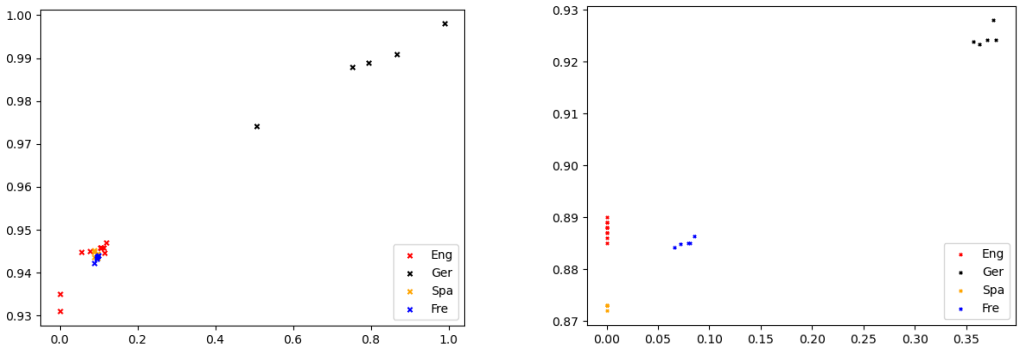

Figure 3. Parameters for Yule for text (left) and raw HTML (right) corpora.

International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

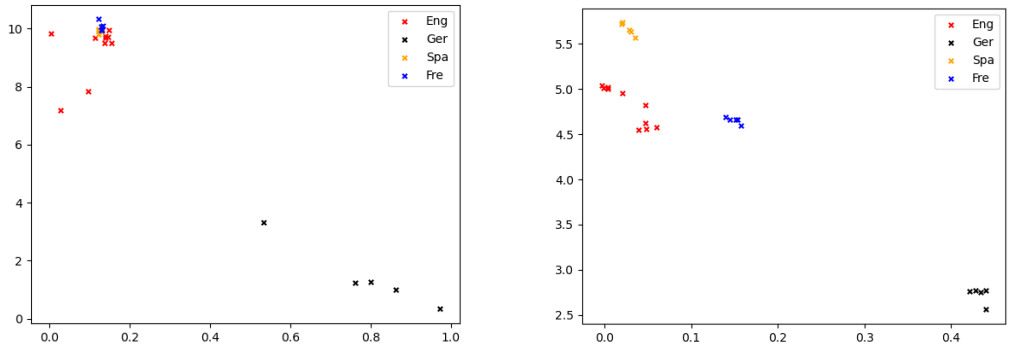

Figure 4. Parameters for Cocho-Beta for text (left) and raw HTML (right) corpora.

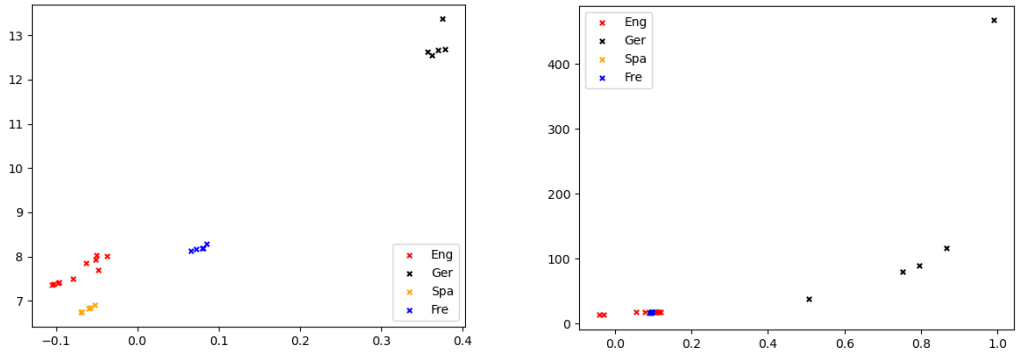

Figure 5. Parameters for Menzerath-Altmann for text (left) and raw HTML (right) corpora.

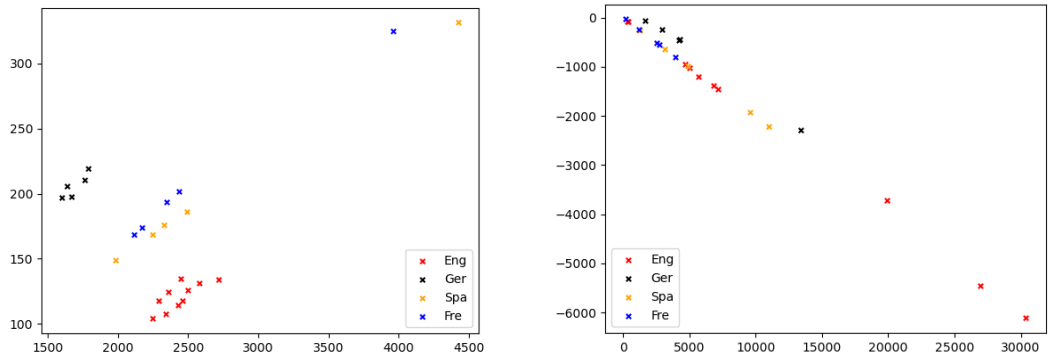

Figure 6. Parameters for Quadratic logarithmic for text (left) and raw HTML (right) corpora.

International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

Table 1. Range and average of goodness of fits for distributions and language corpora.

	Method	Range of Fits (Text)	Text Avg	Range of fits (Raw)	Raw Avg
English	Zipf	0.008573 - 0.009917	0.009379	0.005736 - 0.008065	0.006442
	Good	0.004025 - 0.004987	0.004601	0.003446 - 0.005398	0.004305
	Logarithmic	0.002510 - 0.002940	0.002745	0.002630 - 0.005295	0.004440
	Weibull	0.002129 - 0.002758	0.002481	0.001313 - 0.002383	0.001534
	Exponential	0.000649 - 0.000908	**0.000795**	0.000109 - 0.000258	**0.000208**
	Zipf-Mandelbrot	0.000664 - 0.000901	0.000801	0.000115 - 0.000175	0.000143
	Yule	0.000649 - 0.000908	0.000795	0.000109 - 0.000167	0.000134
	Cocho-Beta	0.000507 - 0.000620	**0.000568**	0.000109 - 0.000173	0.000146
	Quadratic log	0.060180 - 0.064961	0.062800	0.015871 - 0.023709	0.019974
	Menzerath-Altmann	0.000627 - 0.000838	0.009440	0.000102 - 0.000160	**0.000132**
	Frappat	0.000651 - 0.000814	0.009440	0.000109 - 0.016701	0.004980

	Method	Range of Fits (Text)	Text Avg	Range of fits (Raw)	Raw Avg
German	Zipf	0.003868 - 0.004299	0.004149	0.001605 - 0.002684	0.002095
	Good	0.001315 - 0.001661	0.001509	0.000785 - 0.001904	**0.001209**
	Logarithmic	0.004030 - 0.004748	0.004310	0.005799 - 0.021800	0.013460
	Weibull	0.000444 - 0.000653	**0.000521**	0.000511 - 0.006085	0.003124
	Exponential	0.001871 - 0.002235	0.002009	0.003151 - 0.016770	0.009970
	Zipf-Mandelbrot	0.001061 - 0.001414	0.001194	0.001236 - 0.002684	0.001882
	Yule	0.000431 - 0.000649	0.000515	0.000401 - 0.002671	**0.001493**
	Cocho-Beta	0.000323 - 0.000500	**0.000392**	0.000344 - 0.002625	0.001416
	Quadratic log	0.067545 - 0.071953	0.069897	0.028288 - 0.059766	0.044443
	Menzerath-Altmann	0.000431 - 0.000649	0.009389	0.000401 - 0.002671	**0.001493**
	Frappat	0.001353 - 0.001763	0.009389	0.002006 - 0.004427	0.003596

	Method	Range of Fits (Text)	Text Avg	Range of fits (Raw)	Raw Avg
Spanish	Zipf	0.011111 - 0.011337	0.011283	0.006152 - 0.006370	0.006254
	Good	0.006357 - 0.006546	0.006494	0.00422 - 0.004456	0.004333
	Logarithmic	0.004410 - 0.004520	0.004474	0.004648 - 0.005131	0.004877
	Weibull	0.003246 - 0.003374	0.003314	0.001400- 0.001472	0.001435
	Exponential	0.001116 - 0.001281	**0.001193**	0.000163 - 0.000208	**0.000184**
	Zipf-Mandelbrot	0.001123 - 0.001297	0.001195	0.000063 - 0.000114	0.000085
	Yule	0.001116 - 0.001281	**0.001193**	0.000095 - 0.000126	**0.000111**
	Cocho-Beta	0.000817 - 0.000972	0.000891	0.000135 - 0.000159	0.000147
	Quadratic log	0.074113 - 0.074588	0.074396	0.018778 - 0.020212	0.019447
	Menzerath-Altmann	0.001068 - 0.001248	0.011604	0.000095 - 0.000126	**0.000110**
	Frappat	0.001021 - 0.001183	0.011604	0.000094 - 0.000149	0.000117

	Method	Range of Fits (Text)	Text Avg	Range of fits (Raw)	Raw Avg
French	Zipf	0.010303 - 0.010482	0.010381	0.006304 - 0.006504	0.006358
	Good	0.006042 - 0.006157	0.006093	0.004383 - 0.004586	0.004442
	Logarithmic	0.005219 - 0.005289	0.005244	0.005038 - 0.005300	0.005148
	Weibull	0.003713 - 0.003848	0.003778	0.001451 - 0.001534	0.001473
	Exponential	0.002739 - 0.002869	**0.002795**	0.000165 - 0.000203	**0.000181**
	Zipf-Mandelbrot	0.002720 - 0.002860	0.002776	0.000104 - 0.000128	0.000113
	Yule	0.002677 - 0.002797	0.002736	0.000093 - 0.000120	**0.000102**
	Cocho-Beta	0.002216 - 0.002328	**0.002270**	0.000109 - 0.000141	0.000120
	Quadratic log	0.074867 - 0.075727	0.075307	0.020078 - 0.020616	0.020275
	Menzerath-Altmann	0.002677 - 0.002797	0.012376	0.000093 - 0.000120	**0.000102**
	Frappat	0.002656 - 0.002783	0.012376	0.000129 - 0.016198	0.003352

Volume 3, Issue 1, pages 1–10
p-ISSN 2655-8564, e-ISSN 2685-9432

For two parameter distributions, the situation is more difficult. In some cases, such as the Zipf-Mandelbrot distribution, Figure 2, language specific parameters are nicely clustered by language if we only look at text. If, however, we look at raw text, then the English cluster dissolves. For the five German corpora, the parameters are too widely distributed for text and raw files. We attribute this to over-fitting, a phenomenon well known from machine learning. Fitting Zipf-Mandelbrot "learns" the corpus but not the general category. In addition, we observe that the parameters for raw HTML lie along a line, indicating a linear relationship between the two parameters. This indicates that the distribution should be made into a one-parameter distribution. In fact, as can be seen from Table 1, the goodness of fits for Zipf-Mandelbrot is better than for the Zipf distribution but still at the worst range of two parameter distributions. Similarly, the parameters for Menzerath-Altmann are nicely clustered for text but lie on a one-dimensional curve for the raw corpus. For the Quadratic logarithmic distribution, again the results differ between text and raw corpora. For this reason alone, a number of distributions suitable in linguistics are not suitable to model byte frequencies. We refer readers to see Figures 2-6 for the details of our illustration results.

5 Discussion

Our interest is not in linguistics but modelling the overwriting of non-volatile memory. Therefore, our frequency tables make a distinction between capital and non-capital letters. For a linguist, this distinction is probably artificial. Also, unlike for example, Li and Miramontes [5], we do not conflate the letters that differ only in an accent or *umlaut* because they are encoded differently even though they can be considered the same letter. We gave results for texts as a comparison point for raw data. For example, we learned that some distributions such as Zipf-Mandelbrot overfit for raw data and are therefore probably useless for analytics while this does not happen for text. Overall, just as in the work of Li and Miramontes, the Cocho-Beta distribution and the Yule distribution allow best fits without the overfitting phenomenon. Among single parameter distributions the Zipf or Power Law distribution does not fare so well as it is outperformed by the Exponential distribution and by the parametrized Good distribution.

Volume 3, Issue 1, pages 1–10

p-ISSN 2655-8564, e-ISSN 2685-9432

6 Conclusion

Frequency modelling of bytes in electronic documents can be done with the Exponential distribution. While a better fit can be achieved with the Menzerath Altmann distribution or the Cocho-Beta distribution, their parameter range is not only language but also corpus specific. It is hard to see how scientific conclusions can be obtained with such variety. When restricted to text, our observation is not valid.

References

[1] R. Flesch. "A new readability yardstick." *Journal of Applied Psychology,* **32** (3), 221, 1948.

[2] K. H. Best. "*Sind Wort-und Satzlänge brauchbare Kriterien zur Bestimmung der Lesbarkeit von Texten? In: Wichter, Sigurd/Busch, Albert (eds.) Wissenstransfer Erfolgskontrolle und Rückmeldungen aus der Praxis*." Peter Lang Verl, Frankfurt, 2006.

[3] A. D. R. Kulandai and T. Schwarz. "Content-Aware Reduction of Bit Flips in Phase Change Memory." *IEEE Letters of the Computer Society*, 2020.

[4] B. Krevitt and B. Griffith. "A Comparison of Several Zipf-Type Distributions in Their Goodness of Fit to Language Data." J*ournal of the American Society for Information Science*, **23** (3), 220, 1972.

[5] W. Li and P. Miramontes. "Fitting Ranked English and Spanish Letter Distribution in U.S and Mexican Presidential Speeches." *Journal of Quantitative Linguistics,* **18** (4), 359–380, 2011.

[6] C. Manning and H. Schütze. "*Foundations of Statistical Natural Language Processing.*" MIT Press, 2003.

[7] H. Pande and H.S. Dhami. "Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language." *SKASE Journal of Theoretical Linguistics*, **7** (2), 2010.