Text as Binary Sequence: A Case of Characteristic Constant of Text

Petar Milin

Laboratory of experimental psychology, Faculty of Philosophy, University of Belgrade, Čika Ljubina 18-20
11000 Beograd, YU

Laboratory of experimental psychology, Faculty of Philosophy, University of Novi Sad, Stevana Musića 24
21000 Novi Sad, YU

milinp@ptt.yu

Nada Ilić

Laboratory of experimental psychology, Faculty of Philosophy, University of Belgrade, Čika Ljubina 18-20
11000 Beograd, YU

nsilic@f.bg.ac.yu

Abstract

The relation between vocabulary size (*V(N)*) and the text size (*N*) has been re-examined, where the text has been presented as binary sequence. Six different texts by three authors from different periods were taken from the Corpus of Serbian Language to be analyzed. Statistics included regression analysis, randomness test for binary sequence and stochastic models. Point of equivalence, where number of new and old words is equal, has been proposed as characteristic constant of the text. This constant is independent on *N* and could be used as an index of vocabulary richness.

Key-words: vocabulary size, text size, binary sequence, characteristic constant of the text.

1 Introduction

The text of defined size contains *m* words (tokens) and *n* different word entries (types). The size in word tokens we denote as *N* and refer to as the *text size*, while the number of types in a sample of *N* tokens we denote as *V(N)* and refer to as the *vocabulary size*. It is a well known fact that an increase in *N* is paralleled by a non-linear cumulative increase in *V(N)*. Figure 1 depicts typical cumulative non-linear relation between *N* and *V(N)*. The same type of function is observed in a number of studies related to the statistical analyses of word and/or text data like, for example, Zipf's distribution (Zipf, 1935; 1949; cf. Kunz, 1988), empirical structural type distribution and structural type distribution etc. (cf. Baayen, 2001).

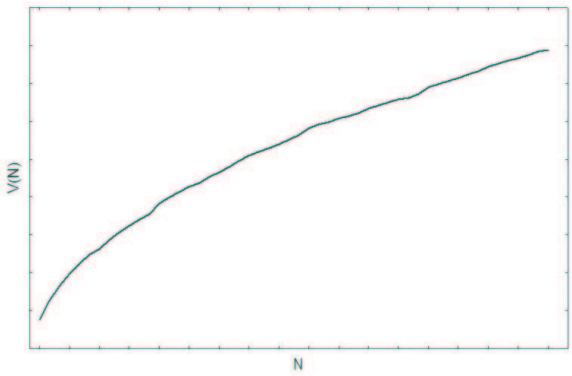

Figure 1. Typical relation of vocabulary size (V(N)) as a function of text size (N)

This study reexamines the relation between the vocabulary size (*V(N)*) and the text size (*N*), questing for the characteristic constant of the text that moderates the relation between *V(N)* and *N*. The relevance of this constant is twofold. On one hand, it could be used for comparing *vocabulary size* of various texts, authors, functional styles etc. On the other hand, it could serve as a useful

tool for designing and building corpora, because it allows for selection of texts with higher number of different lemmata. Also, capturing the possible regularity in relation between *V(N)* and *N* allows for extrapolation which may provide us with prediction of the vocabulary size from the smaller text samples.

Generally, *V(N)* has been considered to be dependent on *N*. Baayen (2001) discusses various text constants which model this regularity like the Herdan's power model (1964), the Guiraud's square-root model as its special case (1954), the Honore's logarithmic model (1979) etc. However, as Baayen and other authors (e.g. Kostić, A., Zec, Milin and Ilić, 2002) concluded, the proposed constants suffer from heavy dependence on *N*, i.e. they vary in a systematic fashion. As a consequence, prediction of *V(N)* from *N* is not reliable.

In their recent study Yang, Gomez and Song (2000) proposed a piecewise curve-fitting approach as a method which presupposes (and thus controls) the dependence of *V(N)* on *N*. They started with the Heaps' power model (1978), identical to Herdan's (1964):

$$V(N) = a \cdot N^b \qquad (1)$$

The curve (*N* vs. *V(N)*) has been cut into pieces, and the constants (*a* and *b*) recalculated for each segment in order to obtain better regressive fit. However, they have left two questions unanswered: (a) in how many pieces should the curve be cut and (b) where should it be cut. The first question is related to the problem of inverse relation between the number of cuts and the degrees of freedom, where the increase in number of pieces increases the goodness-of-fit, but decreases the number of degrees of freedom. The second question appertains the need for external criterion for the points of cuts. Without an external criterion, cuts appear to be arbitrary. As a result, there is an inevitable decrease of interpretability.

It should be noted that Yang, Gomez and Song (2000) had only practical considerations in mind, like corpus predictability, compiling methodology, linguistic comprehensiveness etc. They had no concern about characteristic constant of the text, indirectly showing that the constant is dependent on *N*, because it has to be recalculated for each segment of the text.

In the present study we propose an alternative constant which is not dependent on *N*, and which could be used as an index of vocabulary richness. Some aspects of the problem of extrapolation will be indirectly addressed as well.

2 Method

Instead of analyzing cumulative data, we take binary digits as the simplest, nontrivial way to distinguish the new words from the old words in the text, where 1 stands for a new-word entry, and -1 for an old-word entry. Consequently, the text can be presented as a binary sequence. Note that typical cumulative distribution could easily be reconstructed from this sequence.

2.1 Sample

The sample was taken from the Corpus of Serbian Language (Kostić, Đ., 2001; *http://www.serbian-corpus.edu.yu*). The Corpus consists of 11 million words and spans Serbian language from the 12th century to the contemporary language. Each word in the Corpus is manually annotated up to the level of inflective morphology. The system of annotation distinguishes more than 2000 grammatical forms. In order to test whether the distribution of new lemmata is dependent on the epoch, author or sample size, six different texts by three authors from different periods were analyzed.

2.2 Procedure

Several statistical procedures have been applied in order to explore variation in the textual binary sequence. In addition to usual descriptive statistics and regressive approach, some non-standard procedures have been used too. They derive from different fields of mathematics, like stochastic processes (cf. Cox and Miller, 1978; Grinstead and Snell, 1997) and cryptology (cf. Stevens, 1996).

Author	*Title in Serbian*	*Title in English*	*Period*	*Sample size*
Domentijan	Život Sv. Simeuna	The life of St. Simeun	XIII ct.	26572
	Život Sv. Save	The life of St. Sava		51708
Vuk St. Karadžić	Srpske narodne pripovetke	Serbian National Stories	XIX ct.	98521
	Istorijsko-etnografski spisi	Historical and Ethnographical Writings		205663
Ivo Andrić	Na Drini ćuprija	The Bridge on the Drina	XX ct.	15981
	Travnička hronika	Bosnian Chronicle: The Days of the Consuls		19071

Table 1. Authors, texts, periods and sample size

3 Results

Typical textual binary sequence is presented on Figure 2.

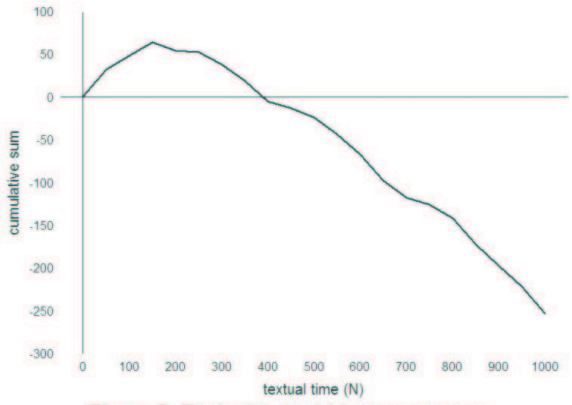

Figure 2. Typical textual binary sequence

The textual time (N) is given on x-axis, while y-axis presents cumulative sum of the textual binary sequence that could be called "text-particle", according to the theory of stochastic processes. A point where the text-particle reaches zero and goes into negative we refer to as the *point of equivalence*. It specifies the position in a given text with equal number of new and old words. Table 2 present values of points of equivalence for six different texts and total number of words in each.

The point of equivalence could also be presented as an intersection of cumulative distributions of new-words and old-words (Figure 3).

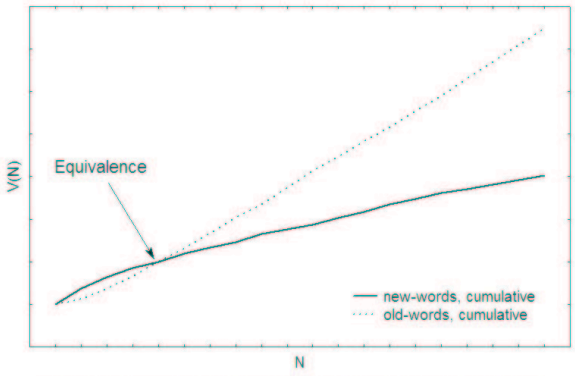

Figure 3. Point of equivalence as the intesection of cumulative function of new and old words

The correlation between the points of equivalence and the vocabulary size ($V(N)$) for a fixed text size ($N = 12000$) is $r = 0.81$; $p \approx 0.05$. This outcome suggests that the point of equivalence may be taken as an index of vocabulary richness. Inspection of Table 2 indicates that all values of the points of equivalence appear within a sample smaller than 1000 words. Hence, 1000 words could be taken as a departure point from which number of new words tends linearly towards negative infinity.

Table 3 summarizes the results of regression analysis for all six samples, when $N \geq 1000$ and $N \to \infty$.

Author	*Title*	*Point of Equivalence*	*Sample size*
Domentijan	Život Sv. Simeuna	392	26572
	Život Sv. Save	452	51708
Vuk St. Karadzic	Srpske narodne pripovetke	226	98521
	Istorijsko-etnografski spisi	422	205663
Ivo Andric	Na Drini ćuprija	562	15981
	Travnička hronika	644	19071

Table 2. The value of equivalence for six different novels

Title	$V(N) = a + b \times N$	$R^2_{Adjusted}$
Život Sv. Simeuna	V(N) = 1438.681 - 0.824 × N	0.998365
Život Sv. Save	V(N) = 1985.739 - 0.878 × N	0.999329
Srpske narodne pripovetke	V(N) = 1990.213 - 0.868 × N	0.999735
Istorijsko-etnografski spisi	V(N) = 2683.640 - 0.852 × N	0.999825
Na Drini ćuprija	V(N) = 928.481 - 0.601 × N	0.996017
Travnička hronika	V(N) = 1256.476 - 0.651 × N	0.994427

Table 3. The results of regression analysis for six different samples

In order to see whether extrapolation is possible in principle, non-randomness of the textual binary sequences, as necessary logical condition, has been tested. Clearly, the infinite linear increase of new words is not realistic because there is a limited number of different word entries in language. Therefore, possible extrapolation should be taken as being tentative. The outcome of the analyses indicates that none of the six textual binaries had satisfied the first and the weakest of the three Golomb's randomness postulates for the binary sequence of period p (cf. Stevens, 1996): *If p is even then the length of p shall contain equal number of ones and zeros. If p is odd, the number of zeros shall be one more or one less than the number of ones*. As expected, the texts contain many more old-words than new-words. Hence, according to Golomb, it could not be random. Additional statistical tests for randomness of the binary sequence – Frequency test, Serial test, Runs test and Poker test, were applied (Stevens, 1996). All of them confirmed that textual binaries are not random.

Although the tests for randomness proved that the text is not "white noise", i.e. that there is some structure in relation between ($V(N)$) and N, they tell nothing about the possibility to extrapolate the vocabulary size from the given text size. To address this question directly, simple random walk model has been applied (Cox and Miller, 1978; Grinstead and Snell, 1997). Probability that the text-particle is in state j at time n has been calculated. In other words, the probability that in the text new-words *(a)* and old-words *(b)* have been realized at time n, where $j = a - b$, with fixed probabilities of new-word (p) and old-word (q) is to occur:

$$\Pr(X_n \geq j) \approx 1 - \Phi\left(\frac{j - 1 - N\mu}{\sigma\sqrt{N}}\right) \qquad (2)$$

where $\mu = p - q$ is the mean, $\sigma = p + q - (p - q)^2$ is the standard deviation of a jump, $N\mu$ is the ex-

pectation for X_n ($E(X_n)$) and $\Phi(y)$ is the standard normal distribution function.

The outcome of this analysis indicated that the text, specified as a binary sequence, does not belong to a class of random walk stochastic processes. Basically, probabilities of new-word (p) and old-word (q) to occur are constantly changing through the textual time (N). At the beginning $p > q$, followed by a period of N where $p \approx q$, and finally $p < q$, when $N \geq 1000$ and $N \rightarrow \infty$.

4 Discussion

The point of equivalence, i.e. the text size where the number of new-words and old-words is equal, could be used as a constant that is not dependent on the text size (N). The respective point indicates an influx of old-words: the slower the influx, the higher the point of equivalence. This, on the other hand, implies higher vocabulary richness, which was empirically confirmed. However, the point of equivalence as a measure of vocabulary richness has one serious deficiency. Since it is reached very early in the textual time, it could be insensitive to the late changes in the text and, particularly, to the influx of new-words late in the text. Therefore, complementary measure is required.

As shown earlier, when $V(N)$ is regressed on N (for $N \geq 1000$), almost perfect correlation has been observed for the six samples. Theoretically, if all words for $N \geq 1000$ were old-words, $R^2 = 1$ and slope $b = -1$. Hence, smaller b indicates higher influx of new-words late in the text. Therewith, b is not used in a standard manner, but rather as an index of the late-text vocabulary richness. Table 4 summarizes values of the total text size (N) and the vocabulary size ($V(N)$), the points of equivalence, R^2 and b, for the six samples.

The correlation between the point of equivalence and b is $r = 0.79$; $p \approx 0.05$. This suggests that they could be taken as the complement measures of vocabulary richness.

Still, further analyses are mandatory, and three things need to be done: (1) Sample size should be increased in order to enhance the reliability of findings; (2) Number of samples should be increased too, in order to increase the test-power in regression analysis; (3) It should be examined to what extent slope (b) is dependent on N. In addition, it would be interesting to examine relation between the peak ($X_n = max$) of the textual binary function, the point of equivalence and b. If the peak and the point of equivalence determine b, then the vocabulary richness of the late text could be expressed by those two points.

While the point of equivalence could be taken as a coarse index of the vocabulary richness which needs further examination, extrapolation, i.e. prediction of the vocabulary richness based on the textual binary sequence does not seem possible. Although the textual binary sequence is not random, probabilities of new-words and old-words are changing across textual time, indicating that they are dependent on N. This brings us back to the conclusions made by Baayen (2001) and also Kostić, A. et al. (2002) about various models based on cumulative distribution, i.e. that there is substantial variability in all of the proposed constants of the text. This variability could be inherent to a given measure, or it can be related to the discourse organization of the text. Finally, it can be even a combination of foregoing factors (Baayen, 2001).

Title	*N*	*V(N)*	*Point of Equivalence*	R^2	*b*
Život Sv. Simeuna	26572	2853	392	0.998365	-0.824
Život Sv. Save	51708	3944	452	0.999329	-0.878
Srpske narodne pripovetke	98521	8015	226	0.999735	-0.868
Istorijsko-etnografski spisi	205663	16291	422	0.999825	-0.852
Na Drini ćuprija	15981	3479	562	0.996017	-0.601
Travnička hronika	19071	3724	644	0.994427	-0.651

Table 4: Various measures of text for six samples

To summarize, there are serious doubts about a possibility to predict vocabulary richness from the text size. On one hand, approaches based on the cumulative distribution have failed. They are erroneous, even the most sophisticated ones like parametric models (Baayen, 2001), while the piecewise curve-fitting approach suffers from idiosyncrasy. On the other hand, the results obtained on the random walk model suggest that more powerful stochastic models, like Markov's chains, would probably also fail.

References

Baayen, R. H. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Cox, D. R. and Miller, H. D. 1978. *The Theory of Stochastic Processes*. Chapman and Hall, London.

Grinstead, C. M. and Snell, J. L. 1997. *Introduction to Probability*. American Mathematical Society, Providence.

Guiraud, H. 1954. *Les Caracteres Statistiques du Vocabulaire*. Presses Universitaires de France, Paris.

Heaps, H. S. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, New York.

Herdan, G. 1964. *Quantitative Linguistics*. Butterworths, London.

Honore, A. 1979. Some Simple Measures of Richness of Vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 172-179.

Kostić, A., Zec, D., Milin, P. and Ilić, N. 2002. Distribution of Lemmata as a Measure of the Functional Differentiation. *32nd International Conference Meeting of Slavicists in Honor of Vuk St. Karadzic*, Belgrade (conference paper, in Serbian).

Kostić, Đ. 2001. *Quantitative Description of Serbian Language Structure – Corpus of Serbian Language*. Institute for Experimental Phonetic and Speech Pathology and Laboratory for Experimental Psychology, Belgrade (in Serbian).

Kunz, M. 1988. Lotka and Zipf: Paper Dragons with Fuzzy Tails. *Scientometrics*, 13 (5-6): 289-297.

Stevens, C. C. 1996. *Detecting Non "Randomness" for Cryptographic Purposes*. ftp://ftp.ox.ac.uk/pub/cryptanalysis/alltest.tar.gz, (15 Dec 2002).

Yang, D. H., Gomez, P. C. and Song, M. 2000. An Algorithm for the Predicting Relationship between Lemmas and Corpus Size. *ETRI Journal*, 22 (2): 20-31.

Zipf, G.K. 1935. *The Psycho-Biology of Language*. Houghton Mifflin, Boston.

Zipf, G.K. 1949. Human Behaviour and the Principle of the Least Effort, An Introduction to Human Ecology. Hafner, NY.