TWO FREQUENCY-RANK LAWS FOR LETTERS IN PRINTED ROMANIAN

Adriana VLAD, Adrian MITREA, and Mihai MITREA

"POLITEHNICA" University of Bucharest
Faculty of Electronics and Telecommunications
1-3 Iuliu Maniu Bvd., Bucharest, Romania, vadriana@vala.elia.pub.ro

Abstract: This paper investigates the way in which the Romanian language obeys a behaviour considered to be correct in case of several natural written languages. This above–mentioned behaviour is expressed by two frequency–rank laws. The authors advance a method through which to obtain ***representative*** constants of the parameters of the two laws for either one language field or for a language as a whole.

Key words: frequency-rank law, letter probability estimate, multiple confidence intervals.

I. Introduction

The main objective of the paper is to find out how accurately printed Romanian complies with a general behaviour supposed to be correct for other natural languages (NL), [**1**], [**2**]. That is, to study the rank-frequency dependency existing for letters, expressed by means of the two laws in Eqs. (1) and (2).

To carry out this study, the relative frequency of every letter occurrence in the natural text is first to be determined and then the results sorted in a decreasing order, $p(1) \geq p(2) \geq \ldots \geq p(k) \geq \ldots \geq p(q)$, where $p(k)$ stands for the relative frequency of the *k*-rank letter and q is the size of the considered alphabet.

The first law under consideration is:

$$p(k) \cong A - D \ln k \qquad (1)$$

where A and D are constants characterising each NL. In [**1**] is mentioned that such a behaviour holds for over 100 NL where the size of the alphabet ranges between 14 and 60.

The second law under consideration is:

$$p(k) \cong B\, 2^{-C k} \qquad (2)$$

where B and C are constants also characterising each NL. This law was mentioned in [**2**].

To strengthen the meaning of our experimental study, we applied a statistical approach concerning letter-probability, as described in [**3**]. This approach is based on multiple confidence intervals for the same letter probability and considers the test of the hypothesis that probability belongs to an interval. These finally enable the obtaining of representative A, D, B, C constants for the two laws, (1) and (2) in a field of the language (or even in the language as a whole).

All the experiments were carried out by processing natural texts presented in *Appendix*.

In Sec. II, we derived the formulae for the parameters of the two frequency-rank laws. In Sec. III we present the experimental study for printed Romanian, with illustration on a literary corpus of 58 books (novels and short stories) and also on an overall mixed corpus of 93 books. Sec. IV contains supplementary reasoning based on a statistical approach as described in [**3**]. Out of Sec. III and IV, we obtained the A, D, B, C representative constants for the literary Romanian.

II. Formulae for the parameters of the two frequency–rank laws

Let us have the q experimental data $(k;\ p(k)),\ k = \overline{1,q}$ where the pair $(k;\ p(k)),\ k = \overline{1,q}$ stands for both the k rank and relative frequency of the *k*-rank letter. We suppose that these data obey to the law (1) and/or (2). We try to determine the laws parameters so that each of the relations (1) and (2) holds with good accuracy.

Let us consider the relation (1), $p(k) \cong A - D \ln k$. We want to determine the A and D values supposing that this behaviour is correct for printed Romanian, too.

In this paper A and D were calculated to minimise the following function (the sum of error-squares is minimised):

$$f(A,D)=\sum_{k=1}^{q}\left[p(k)-A+D\ln k\right]^2 .$$

Expressing $\frac{\partial f(A,D)}{\partial A}$ and $\frac{\partial f(A,D)}{\partial D}$ and bringing them to the 0 value we obtain *A* and *D*:

$$A=\frac{1}{q}\sum_{k=1}^{q}p(k)+\frac{D}{q}\sum_{k=1}^{q}\ln k \quad (3)$$

$$D=\frac{\sum_{k=1}^{q}p(k)\ln k-\frac{1}{q}\sum_{k=1}^{q}p(k)\sum_{k=1}^{q}\ln k}{\frac{1}{q}\left(\sum_{k=1}^{q}\ln k\right)^2-\sum_{k=1}^{q}(\ln k)^2} \quad (4)$$

Another verified law is the relation (2), $p(k)\cong B2^{-Ck}$, where *B* and *C* are positive constants characterising every NL. Here again we try to determine *B* and *C* supposing that this behaviour is correct for printed Romanian.

In this paper *B* and *C* were calculated to minimise the following function (the sum of error-squares is minimised):

$$g(B,C)=\sum_{k=1}^{q}\left[p(k)-B2^{-Ck}\right]^2 .$$

Therefore *B* and *C* are the solution of the system:

$$\frac{\partial g(B,C)}{\partial B}=0 \text{ and } \frac{\partial g(B,C)}{\partial C}=0 .$$

The following equation in *C* (which will be numerically solved) results:

$$\sum_{k=1}^{q}\left[kp(k)2^{-Ck}\right]=\frac{\sum_{k=1}^{q}\left[p(k)2^{-Ck}\right]}{\sum_{k=1}^{q}2^{-2Ck}}\sum_{k=1}^{q}\left(k2^{-2Ck}\right) \quad (5)$$

The *C* value yields to *B* according to formula:

$$B=\left(\sum_{k=1}^{q}\left[p(k)2^{-Ck}\right]\right)\Big/\left(\sum_{k=1}^{q}2^{-2Ck}\right). \quad (6)$$

III. An experimental study for printed Romanian

All the experiments were carried out by processing natural texts presented in *Appendix*.

As a first step we computed the relative frequency of occurrence of each and every letter in the natural text and then sorted these units in a decreasing order:

$p(1)\geq p(2)\geq\ldots\geq p(k)\geq\ldots\geq p(q)$.

Note: At this moment $p(k)$ is just a ratio between the occurrence number of the letter and the length of the text (in letters). As a result of a stationarity study as described in [**3**], $p(k)$ will get the meaning of probability – *i.e.*, the probability of the *k*-rank letter, see Sec. IV.

We calculated the *A*, *D*, *B*, and *C* constant values according to relations (3) – (6) for all of the natural texts considered in *Appendix*. The results are presented in Tables 1 and 2.

To evaluate how correct the behaviours expressed by (1) and (2) are in printed Romanian, we define the quantities $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ and the normed entities $\varepsilon_r^{(1)}$ and $\varepsilon_r^{(2)}$. Namely, $\varepsilon^{(1)}$ concerning law (1), and $\varepsilon^{(2)}$ concerning law (2), are sums of squares of the errors in the analysed text, see (7) and (8):

$$\varepsilon^{(1)}=\sum_{k=1}^{q}\left[p(k)-(A-D\ln k)\right]^2 \quad (7)$$

$$\varepsilon^{(2)}=\sum_{k=1}^{q}\left(p(k)-B2^{-Ck}\right)^2 \quad (8)$$

The $\varepsilon_r^{(1)}$ and $\varepsilon_r^{(2)}$ normed values are obtained by dividing $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ to the values $\sigma^{(1)}$ and $\sigma^{(2)}$ respectively:

$$\sigma^{(1)}=\sum_{k=1}^{q}(A-D\ln k)^2 \qquad \varepsilon_r^{(1)}=\frac{\varepsilon^{(1)}}{\sigma^{(1)}} \quad (9)$$

$$\sigma^{(2)}=\sum_{k=1}^{q}\left(B2^{-Ck}\right)^2 \qquad \varepsilon_r^{(2)}=\frac{\varepsilon^{(2)}}{\sigma^{(2)}} \quad (10)$$

Experimental study on a literary corpus

The literary corpus was obtained by randomly concatenating 58 books (novels and short stories, written by Romanian authors or translated into Romanian, see *Appendix*). The first row in Tab. 1 refers to this whole literary corpus, #WLC. The length – in characters – is $L=29293213$. The parameters of the two frequency-rank laws calculated by means of (3)–(6) are: $A=12.66\times10^{-2}$, $D=3.75\times10^{-2}$ and $B=12.71\times10^{-2}$, $C=16.46\times10^{-2}$. These constants will be further considered as ***representative*** for the literary field. (The

qualifier of ***representative*** will be emphasised in Sec. IV.)

Further we applied (3)–(6) on the two halves of the whole literary corpus, denoted by #1HWLC and #2HWLC. It resulted the *A, D, B* and *C* parameters given in the rows 2 and 3 of Tab. 1.

We continued our experimental study determining *A, D, B* and *C* parameters for various parts of the literary corpus, meaning both individual books (#1, #2, #9 and #10) and groups of books written by the same authors (#Author_Radu_Anton_Roman, #Author_John_Le_Carré, #Author_Alexandr_Soljeniţîn). The lengths – in characters – of these analysed parts of the corpus are shown in column 2.

In columns $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ we evaluated the errors by using (7) and (8). For all the rows in Tab. 1 (*i.e.* for all the natural texts analysed) the errors $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ were calculated considering the representative values: $A=12.66\times10^{-2}$, $D=3.75\times10^{-2}$ and $B=12.71\times10^{-2}$, $C=16.46\times10^{-2}$.

In order to get the relative errors $\varepsilon_r^{(1)}$ and $\varepsilon_r^{(2)}$ defined in (9) and (10), all the values in column $\varepsilon^{(1)}$ should be divided by $\sigma^{(1)}=0.062913$ and those in column $\varepsilon^{(2)}$ by $\sigma^{(2)}=0.062974$. The $\sigma^{(1)}$ and $\sigma^{(2)}$ numerical values were obtained by considering in (9) and (10) the ***representative*** $A=12.66\times10^{-2}$, $D=3.75\times10^{-2}$ and $B=12.71\times10^{-2}$, $C=16.46\times10^{-2}$ values. For example, for the first half of the whole literary corpus, #1HWLC: $\varepsilon^{(1)}=0.0799\times10^{-2} \Rightarrow \varepsilon_r^{(1)}=0.0127$ and $\varepsilon^{(2)}=0.0740\times10^{-2} \Rightarrow \varepsilon_r^{(2)}=0.0118$.

Rows #2, #9 and #10 from Tab. 1 refer to those books (out of 58) which mostly differ from the representative parameters.

Overlooking Tab. 1, we may say that all the numerical results sustain the qualifier *representative* assigned to $A=12.66\times10^{-2}$, $D=3.75\times10^{-2}$, $B=12.71\times10^{-2}$, and $C=16.46\times10^{-2}$ parameters in the first row. This conclusion is based on:

1. **The numerical values obtained for the two halves #1HWLC and #2HWLC are practically the same and quite equal with those obtained for the #WLC. (Note that the two halves are composed by sorting the books according to a random rule.)**
2. **The accuracy expressed by $\varepsilon^{(1)}, \varepsilon^{(2)}$, $\varepsilon_r^{(1)}$ and $\varepsilon_r^{(2)}$ is good enough.**

In Sec. IV, we shall show that $p(k)$ is a very good estimate for the probability of the k – rank letter.

		The law in (1)			The law in (2)		
The analysed text	***L***	***A***	***D***	ε (1)	***B***	***C***	ε (2)
#WLC Whole Literary Corpus	29293213	**12.66**	**3.75**	0.0791	**12.71**	**16.46**	0.0729
#1HWLC First Half of Whole Literary Corpus	14646607	12.67	3.75	0.0799	12.71	16.46	0.0740
#2HWLC Second Half of Whole Literary Corpus	14646606	12.66	3.74	0.0772	12.71	16.47	0.0711
#1 *Precum fumul*, see *Appendix*	551989	12.38	3.63	0.0765	12.44	15.98	0.0758
#2 *Zile de pescuit*, see *Appendix*	405664	12.17	3.53	0.0800	12.19	15.65	0.0690
#9 *Canettis Angst*, see *Appendix*	309436	13.00	3.89	0.0808	13.17	17.02	0.0566
#10 *O último cais*, see *Appendix*	278578	12.97	3.87	0.0650	13.14	17.07	0.0709
#Author_Radu_Anton_Roman	957653	12.31	3.61	0.0774	12.08	15.46	0.0728
#Author_John_Le_Carré	1874166	12.74	3.78	0.0631	12.83	16.65	0.0655
#Author_Alexandr_Soljeniţîn	3115634	12.86	3.83	0.0791	12.94	16.69	0.0598

Table 1: Verifying frequency–rank laws for literary Romanian field. All the numerical values in Tab.1. – except column *L* which represents the length of text – have to be divided by 100. **The *representative* values are: $A=12.66\times10^{-2}$, $D=3.75\times10^{-2}$, $B=12.71\times10^{-2}$ and $C=16.46\times10^{-2}$.**

Experimental study on a mixed corpus

The whole mixed corpus, denoted by #WMC, consists of 93 books summing-up 43002954 characters (the 58 literary books included). These 93 books are concatenated according to a random rule.

The parameters of the two laws, obtained for the #WMC, are given in Tab.2, row 1. The numerical values $A=12.88\times10^{-2}$, $D=3.83\times10^{-2}$, $B=13.07\times10^{-2}$, and $C=16.97\times10^{-2}$ will be considered as ***reference values*** and the other rows in Tab. 2 will refer to them.

The rows 2 and 3 in Tab. 2 (#1HWMC and #2HWMC) correspond to the two halves of the mixed corpus.

The following rows in Tab. 2 contain the *A, D, B* and *C* values obtained for several parts of #WMC: literature, law, medicine (separately processed) and science at large (#WSC: law, medicine, forestry, history, sociology, *etc*.).

The *L* column stands for the lengths – in number of characters – of the analysed texts.

In the columns $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ we evaluated the errors by using (7) and (8). For all the rows in Tab. 2 (*i.e.* for all the natural texts here analysed) the errors $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ were calculated considering the ***reference*** values: $A=12.88\times10^{-2}$, $D=3.83\times10^{-2}$ and $B=13.07\times10^{-2}$, $C=16.97\times10^{-2}$.

In order to get the relative errors $\varepsilon_r^{(1)}$ and $\varepsilon_r^{(2)}$ defined in (9) and (10), all the values in column $\varepsilon^{(1)}$ should be divided by $\sigma^{(1)}=0.064337$ and those in column $\varepsilon^{(2)}$ by $\sigma^{(2)}=0.064363$. The $\sigma^{(1)}$ and $\sigma^{(2)}$ numerical values were obtained by considering in (9) and (10) the ***reference*** $A=12.88\times10^{-2}$, $D=3.83\times10^{-2}$ and respectively $B=13.07\times10^{-2}$, $C=16.97\times10^{-2}$ values. For example for the first half of the whole mixed corpus, #1HWMC: $\varepsilon^{(1)}=0.0635\times10^{-2}\Rightarrow\varepsilon_r^{(1)}=0.0099$ and $\varepsilon^{(2)}=0.0641\times10^{-2}\Rightarrow\varepsilon_r^{(2)}=0.01$.

		The law in (1)			The law in (2)		
The analysed text	***L***	***A***	***D***	$\varepsilon(1)$	***B***	***C***	$\varepsilon(2)$
#WMC Whole Mixed Corpus	43002954	**12.88**	**3.83**	0.0657	**13.07**	**16.97**	0.0631
#1HMC First Half of Whole Mixed Corpus	21501477	12.91	3.84	0.0635	13.12	17.04	0.0641
#2HMC Second Half of Whole Mixed Corpus	21501477	12.85	3.82	0.0683	13.03	16.91	0.0625
#WLC Whole Literary Corpus	29293213	12.66	3.75	0.0806	12.71	16.46	0.0752
#Law	1824035	13.73	4.19	0.0739	14.22	18.44	0.0756
#Medicine	1510708	13.45	4.06	0.1004	14.02	18.21	0.0794
#WSC Whole Scientific Corpus	5936496	13.41	4.04	0.0724	13.99	18.22	0.0620

Table 2: Verifying frequency–rank laws for the whole mixed corpus. All the numerical values in Tab. 2 – except column *L* which represents the length of text – have to be divided by 100. **The *reference* values are:** $A=12.88\times10^{-2}$, $D=3.83\times10^{-2}$, $B=13.07\times10^{-2}$ **and** $C=16.97\times10^{-2}$.

Overlooking Tab. 2, some differences between #WLC and #WSC can be noticed. The numerical values might hint to a certain difference between the mathematical models corresponding literary and scientific fields. However, when considering only the first approximation of the language (the statistical structure of letters) this difference is not too large. The whole mixed corpus averages these models. Note: the scientific corpus (#WSC) is varied (including medicine, law, forestry, sociology, *etc*.) and quite small when compared with the literary one. Therefore, we could not determine letter–probabilities with the same accuracy we did for the literary corpus.

IV. Obtaining the representative constants for the two laws in literary printed Romanian

For every type of natural text analysed in Tab. 1, we carried out a study concerning the

language stationarity in the basis of the first order approximation of the language. This study follows the procedure from [**3**]. The first NL approximation was defined by Shannon in [**4**].

We shall briefly describe the procedure we applied. All the illustrations in Tab. 3 stand for the whole literary corpus, see Tab. 3.

We periodically sampled the NL (with a period of 200 letters) to obtain the first approximation of the language. By shifting the sampling origin within the natural text, we obtained 200 sets of non–overlaping experimental data, each of them having the same meaning. The 200 data sets are not independent data sets. However, each of them complies with the *i.i.d.* statistical model (*i.e.* observations came out from *i*ndependently and *i*dentically *d*istributed random variables).

For every letter of the alphabet we applied the following steps (we shall exemplify for the *E* letter in the whole literary corpus, see first row in Tab. 3).

1. We calculated the relative frequency p^* for the considered letter; p^* is the ratio between the *E* letter occurrences and the length of the natural text (here the length is *L* = 29293213). The *E* letter is on the first rank, hence $p^* \equiv p(1) = 11.47 \times 10^{-2}$. Generally, for any letter $p^* = p(k)$, where *k* is the rank of the respective letter.
2. Out of the 200 sets of experimental data, 200 estimates alongside with 200 confidence intervals were obtained.
3. We selected among the 200 probability estimates that (estimated) value which is nearest to p^*. This selected estimate will be further denoted by p_Δ and the corresponding confidence interval by Δ. It results:

   $$p_\Delta = \frac{m}{N},$$

   where *m* is the number of occurrences of the *E* letter in the selected experimental data set and *N* is the length of the *i.i.d.* data set (here $N = L/200$).

   The Δ confidence interval is $\Delta = (p_1 ; p_2)$, where the p_1 and p_2 confidence limits are:

   $$p_{1,2} \cong p_\Delta \mp z_{\alpha/2} \sqrt{\frac{p_\Delta (1 - p_\Delta)}{N}}.$$

   $z_{\alpha/2}$ is the $\alpha/2$ point value corresponding to the standard Gaussian law. In our experimental study we considered an $1 - \alpha = 0.95$ statistical confidence level, hence $z_{\alpha/2} = 1.96$.
4. We applied the statistical test of the hypothesis that the *E* letter probability belongs to the Δ interval, [**3**], separately for each of the 199 experimental data (200 minus the one which produced the Δ interval).

The steps $1 \div 4$ were resumed for every letter of the alphabet. **The stationarity is accepted if –for every letter of the alphabet– each and every (or, at least, almost each and every) test out of the total of 199 is passed. In our literary Romanian corpus (#WLC) practically all the tests were passed, at an $\alpha = 0.05$ statistical significance level and with small β sizes of the II type statistical error, see Tab. 3.**

In Tab. 3 there are details concerning the verifying of the frequency–rank laws and also concerning the meaning of $p(k)$.

Columns 1 and 2 contain the ranks and the corresponding letters. For example, the *E* letter is on the rank $k = 1$. Its relative frequency is $p^* \equiv p(1) = 11.47 \times 10^{-2}$ (column 3). The estimate $p_\Delta = 11.47 \times 10^{-2}$ (it is practically equal with p^*). The two confidence limits (column 5 and 6) are $p_1 = 11.30 \times 10^{-2}$ and $p_2 = 11.63 \times 10^{-2}$. We may say that in 95% of cases the true letter *E* probability lies within $\Delta = (p_1 ; p_2) = (11.47 ; 11.63) \times 10^{-2}$. Further, the Δ interval was validated by every of the 199 *i.i.d.* experimental data sets. The second type statistical error is the error to accept wrong data as good ones (*i.e.* to enjoy for nothing when the statistical test is passed). The β size of this error is large when the true letter probability is very close to the bounds of the Δ interval, but outside the Δ interval. In Tab. 3, β is calculated for the cases when the letter probability we test is on the left–side of Δ interval, namely equal to $0.95 p_1$ (column 7) or $0.9 p_1$ (column 8). That is, the true letter probability differs with 5% or 10% from the p_1 left bound. Column 9 was filled in by

considering $A = 12.66 \times 10^{-2}$ and $D = 3.75 \times 10^{-2}$ representative values from Tab. 1. Similarly, column 11 is filled in considering $B = 12.71 \times 10^{-2}$ and $C = 16.46 \times 10^{-2}$ representative constants. Columns 10 and 12 present the relative errors among the experimental and theoretical values, according to relations:

- $\varepsilon_{r,1}(k) = \dfrac{p(k) - (A - D \ln k)}{A - D \ln k}$ for the law in (1);
- $\varepsilon_{r,2}(k) = \dfrac{p(k) - B2^{-Ck}}{B2^{-Ck}}$ for the law in (2).

k	x_k	$p^* \equiv p(k)$	p_Δ	Δ		β_1	β_2	$A - D \ln k$	$\varepsilon_{r,1}(k)$	$B2^{-Ck}$	$\varepsilon_{r,2}(k)$
				p_1	p_2						
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	E	11.47	11.47	11.30	11.63	0.00	0.00	12.66	-9.45	11.34	1.12
2	I	9.96	9.96	9.80	10.11	0.00	0.00	10.07	-1.10	10.12	-1.59
3	A	9.95	9.96	9.80	10.11	0.00	0.00	8.55	16.47	9.03	10.30
4	R	6.82	6.82	6.69	6.95	0.02	0.00	7.47	-8.73	8.05	-15.34
5	N	6.47	6.47	6.35	6.60	0.03	0.00	6.63	-2.45	7.18	-9.93
6	U	6.20	6.20	6.08	6.32	0.05	0.00	5.95	4.16	6.41	-3.30
7	T	6.04	6.04	5.92	6.16	0.06	0.00	5.37	12.37	5.72	5.59
8	C	5.28	5.28	5.17	5.39	0.19	0.00	4.87	8.33	5.10	3.47
9	L	4.48	4.48	4.38	4.59	0.60	0.00	4.43	1.14	4.55	-1.53
10	S	4.40	4.41	4.30	4.51	0.67	0.00	4.04	9.11	4.06	8.47
11	O	4.07	4.07	3.97	4.17	1.07	0.00	3.68	10.65	3.62	12.38
12	Ă	4.06	4.06	3.96	4.17	1.08	0.00	3.35	21.18	3.23	25.73
13	D	3.45	3.45	3.35	3.54	2.47	0.00	3.05	12.80	2.88	19.45
14	P	3.18	3.18	3.09	3.27	3.46	0.00	2.78	14.59	2.57	23.65
15	M	3.10	3.10	3.01	3.19	3.84	0.00	2.52	22.99	2.30	34.90
16	Ş	1.55	1.55	1.48	1.61	23.62	0.07	2.28	-32.07	2.05	-24.50
17	Î	1.40	1.40	1.34	1.46	27.42	0.15	2.05	-31.63	1.83	-23.33
18	V	1.23	1.22	1.17	1.28	32.69	0.40	1.84	-33.25	1.63	-24.87
19	F	1.18	1.18	1.13	1.24	34.05	0.50	1.63	-27.77	1.45	-18.93
20	B	1.07	1.07	1.02	1.12	38.01	0.91	1.44	-25.67	1.30	-17.50
21	Ţ	1.00	1.00	0.95	1.05	40.71	1.33	1.26	-20.41	1.16	-13.55
22	G	0.99	0.99	0.94	1.04	41.11	1.41	1.08	-8.61	1.03	-4.16
23	Â	0.91	0.91	0.86	0.96	44.40	2.16	0.92	-0.58	0.92	-1.09
24	Z	0.71	0.71	0.67	0.75	53.00	5.72	0.76	-6.24	0.82	-13.64
25	H	0.47	0.47	0.44	0.51	64.89	17.16	0.60	-21.60	0.73	-35.40
26	J	0.24	0.24	0.21	0.26	78.43	45.35	0.46	-47.91	0.65	-63.59
27	X	0.11	0.11	0.10	0.13	85.69	67.63	0.32	-64.57	0.58	-80.82
28	K	0.11	0.11	0.09	0.13	86.40	69.95	0.18	-39.62	0.52	-79.16
29	Y	0.07	0.07	0.06	0.09	88.62	77.20	0.05	45.49	0.46	-84.87
30	W	0.03	0.03	0.02	0.04	92.07	87.76	**-0.08**	–	0.41	-92.10
31	Q	0.00	–	–	–	**–**	**–**	**-0.20**	–	0.37	-98.89

Table 3. Verifying the frequency–rank laws in #WLC. Arguments for ***representative*** constants. The numerical values in columns 3÷12 have to be divided by 100

To conclude with, it can be noticed that $p^* = p(k) \cong p_\Delta$; these values approximate the true letter probability with a good accuracy. The accuracy is expressed both by the length of the Δ interval (which was determined with a statistical confidence level of 95%) and by the β size of type II statistical error. Note that the ratio $(p_2 - p_1)/p^* < 0.06$ for all high and medium frequency letters (up to $k = 15$). On the other hand, the β values are also very small,

up to rank 15.

Other commentaries upon Tab. 3

1. The law in (1) gives negative values for $k = 30$ and $k = 31$, see column 9. Certainly, these can not be considered probability estimates. Therefore, column 10 was not filled in for $k = 30$ and $k = 31$. However, when we evaluated the $\varepsilon^{(1)}$ error in Tab. 1 and Tab. 2, we considered all the ranks, $k = \overline{1,31}$.

2. For a meaning of probability estimates, when summing–up the quantities $A - D\ln k$ (respectively $B2^{-Ck}$) we have to obtain a value very close to 1. Skipping over the letters of very low probability in Tab. 3 (X, K, Y, W, Q) we obtained: $\sum_{k=1}^{26} p(k) = 0{,}9968$, $\sum_{k=1}^{26}(A - D\ln k) = 0{,}9960$, $\sum_{k=1}^{26} B2^{-Ck} = 0{,}9986$.

Tab. 4 enables a comparison between the two laws (1) and (2) on the whole literary corpus (#WLC).

	$l=3$	$l=10$	$l=15$	$l=25$
$\sum_{k=1}^{l} p(k)$	31.38	71.07	88.93	99.44
$\sum_{k=1}^{l}(A - D\ln k)$	31.28	70.05	85.43	99.28
$\max_{k=1,l} \varepsilon_{r,1}(k)$	16.47	16.47	22.99	33.25
$\sum_{k=1}^{l}[p(k) - (A - D\ln k)]^2$	0.0343	0.0468	0.0599	0.0776
$\sum_{k=1}^{l}\left(B2^{-Ck}\right)$	30.48	71.57	86.17	99.10
$\max_{k=1,l} \varepsilon_{r,2}(k)$	10.30	15.34	34.90	35.40
$\sum_{k=1}^{l}\left[p(k) - B2^{-Ck}\right]^2$	0.0091	0.0324	0.0546	0.0629

Table 4. A comparison between the two laws in (1) and (2), in the basis of #WLC

How to read Tab. 4 (*e. g.* for the letters on the first 3 ranks)

The letters on the first three ranks cover $\sum_{k=1}^{3} p(k) = 31.38\%$ of the total length of the natural text. By computing this summation in the basis of the law (1) we obtained 31.28% and in the basis of (2) we obtained 30.48%. The maximum relative error between the experimental values $p(k)$ and the corresponding values obtained with the laws (1) or (2) appears for $k = 3$, namely: $\varepsilon_{r,1}(3) = 16.47 \times 10^{-2}$ and $\varepsilon_{r,2}(3) = 10.30 \times 10^{-2}$. The sums of squares of errors up to rank $k = 3$, for the two laws are 0.0343×10^{-2} and 0.0091×10^{-2}.

To conclude with, for the first three ranks, the two laws are comparable, the second being slightly better. When the comparison is carried out in the basis of larger rank values, ($k = 10, 15, 25$), the two laws are equally good.

V. Final remarks

As a final remark, we can say that the general behaviour expressed by (1) and (2) is quite correct for printed Romanian, too.

A problem we consider to be general (beyond the printed Romanian peculiarities) is the way to decide which law parameter values

are representative for a field of the language or for the NL *per–se*. Our study offer a solution to this problem, as illustrated for a literary corpus of novels and short stories.

References

[**1**] I. Kanter, D. A.Kessler, "*Markov Processes: Linguistics and Zipf's Law",* Physical Review Letters, Volume 74, Number 22, (May 1995).

[**2**] S. Marcus, Ed.Nicolau, S. Stati, *Introducere în lingvistica matematică,* Ed. Ştiinţifică, Bucureşti, 1966. (Also *Introduction en la linguistica matematica*, Editorial Teide, Barcelona, 1978.)

[**3**] Adriana Vlad, A. Mitrea, M. Mitrea, and D. Popa, "*Statistical methods for verifying the natural language stationarity based on the first approximation.Case study: Printed Romanian*" in Vol. **VEXTAL'99** Ed. Unipress, ISBN 88-8098-112-9, pp. 127-132, Nov. 22-24, 1999, Venice; http://byron. cgm.unive.it/events/papers/vlad.pdf

[**4**] Shannon C. E., "*Prediction and Entropy of Printed English*", Bell Syst. Tech. J., vol. 30, pp. 50-64, January 1951.

Appendix

In order to carry out this study we first elaborated a corpus for printed Romanian literature in the basis of 93 books with the new orthography (introduced after 1993). Blanks, punctuation marks and figures were eliminated. The alphabet thus obtained consists of 31 letters: A Ă Â B C D E F G H I Î J K L M N O P Q R S Ş T Ţ U V W X Y Z. **The whole global concatenated corpus (#WMC)** sums-up about 43 million characters.

These books represent genuine Romanian literature (11 novels and short stories); foreign literary works translated into Romanian (47 novels and short stories) and scientific texts (books from law, medicine, forestry, history, sociology, *etc.*).

The books used in Tab. 1:

#1. Radu Anton Roman, *Precum fumul*, Ed. Cartea Românească, Bucureşti, 1996, ISBN 973-23-0274-7.

#2. Radu Anton Roman, *Zile de pescuit*, Ed. Metropol, Bucureşti, 1996, ISBN 973-562-073-1.

#3. John Le Carré, *Spionul care venea din frig (The spy who came in from the cold),* Ed. Univers, Bucureşti, 1996, ISBN 973-34-0355-5.

#4. John Le Carré, *Casa Rusia (The Russia House),* Ed. Univers, Bucureşti, 1997, ISBN 973-34-0457-8.

#5. John Le Carré, *Micuţa toboşăreasă (The Little Drummer Girl),* Ed. Univers, Bucureşti, 1998, ISBN 973-34-0430-6.

#6. Alexandr Soljeniţîn, *Arhipelagul Gulag vol. I, (Arhipelag GULag I-II),* Ed. Univers, Bucureşti, 1997, ISBN 973-34-0454-3.

#7. Alexandr Soljeniţîn, *Arhipelagul Gulag vol. II, (Arhipelag GULag III-IV),* Ed. Univers, Bucureşti, 1997, ISBN 973-34-0480-2.

#8. Alexandr Soljeniţîn, *Arhipelagul Gulag vol. III, (Arhipelag GULag V-VI-VII),* Ed. Univers, Bucureşti, 1998, ISBN 973-34-0497-7.

#9. Rüdiger Wischenbart, *Frica lui Canetti (Canettis Angst),* Univers, Bucureşti, 1997, ISBN 973-34-0501-9.

#10. Helena Marques, *Ultimul chei (O último cais),* Univers, Bucureşti, 1997, ISBN 973-34-0424-1.

In our illustrations (Table 1) we also used some natural texts obtained by linking books written by the same authors:

#Author_Radu_Anton_Roman is composed of **#1** and **#2** (in this order);

#Author_John_Le_Carré is composed of **#3, #4** and **#5** (in this order);

#Author_Alexandr_Soljeniţîn is composed of **#6, #7** and **#8** (in this order).