STATISTICAL METHODS FOR VERIFYING THE NATURAL LANGUAGE STATIONARITY BASED ON THE FIRST APPROXIMATION. CASE STUDY: PRINTED ROMANIAN

Adriana VLAD, Adrian MITREA, Mihai MITREA, Dragoş POPA
"Politehnica" University of Bucharest, Electronics and Telecommunications Dept.
email: vadriana@vala.elia.pub.ro

ABSTRACT: This paper contains theoretical and experimental results in verifying the hypothesis of printed language stationarity. These results were carried out in the basis of the first approximation to printed Romanian. The statistical approach here–developed enabled the determination of a *representative* confidence interval for letter probability. The sample which provided this *representative* interval was further used in the comparison among and between natural texts. Several kinds of statistical inferences have been considered: probability estimation with confidence limits, test of the hypothesis that probability belongs to an interval and test of the difference between two probabilities. The first type statistical error of an α probability as well as the second type statistical error of a β probability have been involved in the quantitative results.

KEY WORDS: Natural language stationarity, first approximation to natural language, statistical inferences on probability, second type statistical error, comparison among fields of the language.

I. INTRODUCTION AND SUMMARY

The starting point of this paper is the general assumption according to which natural language (NL) is well approximated by an ergodic Markov chain of a multiplicity order larger than 30, [**1**], [**2**].
The paper FOCUSES ON TWO OBJECTIVES:
1. A fundamental mathematical problem, namely verifying NL stationarity, illustrated on printed Romanian. This problem is complicated, from both the theoretical and the experimental point of view. Therefore, its solving implied successive approximations to NL. The results here presented refer to the first approximation to printed Romanian (*i.e.* statistical structure of letter).
2. A comparison among and between three fields of the language, namely: genuine Romanian literature, scientific and technical works written by Romanian authors, foreign literary works translated into Romanian.
Having these objectives in view, the paper develops a STATISTICAL APPROACH which implies: to estimate the probability with confidence limits; to test the hypothesis that a probability belongs to an interval; to test the difference between two probabilities. The two statistical type errors have been considered. The overall statistical procedure here proposed is general, the Romanian language peculiarities appearing only in the quantitative results.

WHAT IS FIRST APPROXIMATION TO NL? This represents, [**1**], [**2**], a zero memory information source, having the same alphabet and letter probability as the considered NL. Note: The *di*gram (two successive letters), *tri*gram (three successive letters), ... , *n*gram (sequence of *n* letters) probabilities differ from the NL.
The first approximation to a NL can be obtained through a periodical NL sampling having a large enough period as to practically break the dependency among successive letters. In our experiments we have considered that a period of 200 letters satisfies this requirement. (For the sake of clarity in the method presentation, this numerical value of 200 will be furtherly referred to. When doubting about this value of 200, a larger one might be used, if there exists a large enough linguistic corpus). By shifting the sampling origin in the natural text, 200 sets of non-overlapping experimental data are obtained, each of them having the same meaning, see Fig. 1. Note that each of these experimental data sets stands for a first approximation to printed Romanian. **Attention should be granted: these 200 samples are not independent data sets**.

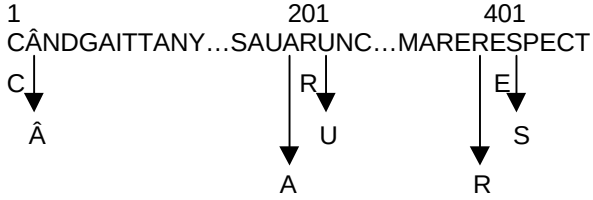

Fig. 1 First approximation to printed Romanian, obtained through a periodical sampling; several experimental data sets can be obtained: **1.** CRE...; **2.** ÂUS...; **200.** AR....

Thus, our statistical investigation started with the assumption that each of the 200 data sets from Fig. 1 comply with the *i.i.d.* statistical model (*i.e.* observations come out from *i*ndependently and *i*dentically *d*istributed random variable) and convey the same information about letter probability.
By means of the estimation theory each *i.i.d.* data set leads to a confidence interval for the probability; here a statistical confidence of 95% was considered. As a consequence, for one and the same chosen letter, 200 estimates, alongside with their 95% confidence intervals, can be computed in the basis of the 200 *i.i.d.* data sets. Under the stationarity hypothesis, all these 200 confidence intervals have to correspond to the same letter probability. This is, in fact, OUR AIM: **we mean to establish whether the 200 estimates derive from the same**

theoretical probability or not. We also mean to determine which of these confidence intervals is *representative.* We considered that one of the IMPORTANT RESULTS of this paper is the procedure developed with a view to obtain a confidence interval which better suits letter probability. This interval and the data set which provides it are *representative* for the letter and the considered natural text. The qualifier *representative* was assigned to that confidence interval which was statistically validated by each and every (or, at least, by almost every) of the 200 data sets. If the natural text enables the construction of such a *representative* interval for all letters in the alphabet, then the first order language stationarity is confirmed. The procedure to determine the *representative* confidence interval and the interval *per se* can be also used by another experimenter when comparing various natural texts.

The statistical measurements required the organising of a linguistic corpus (such a corpus is not available for printed Romanian). The electronic text was obtained out of processing 21 books, with the new orthography (introduced after 1993), belonging in a quite equal proportion to the three fields: **genuine Romanian literature** (3 990 222 characters from 7 books), **scientific and technical works written by Romanian authors** (4 434 888 characters from 6 books), **foreign literary works translated into Romanian** (4 247 646 characters from 8 books). Blanks, punctuation marks and figures were eliminated; this does not diminish the application area of the method. 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 concatenated text** sums-up to 12 672 756 characters.

II. STATISTICAL BACKGROUND

II.1. Confidence interval for probability

The purpose is to determine letter probability with confidence intervals. Be $\left[x_1, x_2, \ldots, x_N\right]$ the observations which comply with the *i.i.d.* statistical model. In the paper, this experimental data set represents the first approximation to printed Romanian, *i.e.* each of the 200 data sets from Fig. 1. N is the sample size (the natural text length divided by 200). Be m the number of occurrences of the specified letter in the N experimental data; the probability estimate is $\hat{p} = m/N$. The $(1-\alpha)\times 100\%$ confidence interval for the p probability is $I = \left(p_1; p_2\right)$, where:

$$p_{1,2} = \hat{p} \mp z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/N} = \hat{p}(1 \mp \hat{\varepsilon}_r) \qquad (1)$$

Eq. (1) means that in $(1-\alpha)\times 100\%$ of cases the theoretical unknown p value will lie within an $\hat{\varepsilon}_r \cdot \hat{p}$ interval around the $\hat{p}$ estimated value. $\hat{\varepsilon}_r = z_{\alpha/2}\sqrt{(1-\hat{p})/(\hat{p}N)}$ is an experimental relative error between the theoretical unknown p value and the $\hat{p}$ eestimated value. p_1 and p_2 are the $(1-\alpha)\times 100\%$ confidence limits. $z_{\alpha/2}$ is the $\alpha/2$-point value of the normal law of 0 mean and 1 variance. **The values through which we obtained our experimental results are** $1-\alpha = 0.95 \Rightarrow z_{\alpha/2} = 1.96$**.**

The confidence limits in (1) depend on the experimental data due to the $\hat{p}$ estimated value; hence, the confidence interval is random. For the same letter, 200 confidence intervals are obtained, corresponding to the 200 data sets representing the first approximation to NL. A problem of this study is to decide which of these 200 confidence intervals better represents letter probability.

II.2. Test of the hypothesis that the probability belongs to an interval.

Be $(a;b)$ an interval within which we suppose that the p probability of the considered letter lies. Let us have only a single $\left[x_1, x_2, \ldots, x_N\right]$ sample, which complies with the statistical *i.i.d.* model. The aim is to decide whether the $\left[x_1, x_2, \ldots, x_N\right]$ experimental data verify the hypothesis that the p probability belongs to the $(a;b)$ interval, with an α significance level. The procedure follows as such.

The two statistical hypotheses (null hypothesis H_0 and alternative hypothesis H_1) are:

H_0: $a < p < b$ and H_1: $p \notin (a;b)$.

We calculate the estimate $\hat{p} = m/N$, where m is the number of successes of the event (letter occurrence in the N observations). We have to verify, with the chosen α significance level, whether the estimated $\hat{p}$ value is accepted by the test or, on the contrary, belongs to the rejection region. By analogy with [**5**], the region meaning that null hypothesis H_0 is accepted is a larger interval including $(a;b)$, namely $\left(c_1; c_2\right)$, where:

$$1-\alpha = \int_{c_1}^{c_2} \frac{1}{\sqrt{2\pi a(1-a)/N}} \exp\left(-\frac{(x-a)^2}{2a(1-a)/N}\right) dx =$$

$$= \int_{c_1}^{c_2} \frac{1}{\sqrt{2\pi b(1-b)/N}} \exp\left(-\frac{(x-b)^2}{2b(1-b)/N}\right) dx \qquad (2)$$

Two probability density functions are involved in the integrals in (2); they correspond to the normal law:

- with a statistical mean and $a(1-a)/N$ variance;
- with b statistical mean and $b(1-b)/N$ variance.

The null hypothesis H_0 will be accepted if and only if the estimated $\hat{p}$ value falls within the $\left(c_1; c_2\right)$ interval.

Two types of errors might be come across:

Type I error consists in rejecting the null hypothesis H_0 when it is true. This happens when $\hat{p} \notin (c_1; c_2)$ though the

true p probability satisfies $a < p < b$. The probability of this situation is lower then α.

Type II error means not to reject H_0 although it is false. This happens when $c_1 < \hat{p} < c_2$, however the p true probability does not belong to the interval $(a;b)$. The probability of this situation depends on the p value (for fixed α and N). It is denoted by $\beta(p)$, Eq. (3):

$$\beta(p) = \int_{c_1}^{c_2} \frac{1}{\sqrt{2\pi\, p(1-p)/N}} \exp\left(-\frac{(x-p)^2}{2p(1-p)/N} \right) dx \quad (3)$$

$\beta(p)$ takes high values when p is very close to $(a;b)$ interval. In practice, the most disturbing cases are those when $p \le (1-\delta)\cdot a$ or $p \ge (1+\delta)\cdot b$, although the test is passed. The δ quantity is chosen by the experimenter, depending on how disturbing this situation is.

This testing procedure is our extension of a similar test applied to the mean in [5]. We considered such a test absolutely necessary in comparing the 200 samples among themselves in order to check-up the language stationarity. Remember that these experimental data are not independent sets and therefore we could not apply a more usual test, as it is the test of the difference between two probabilities.

II.3. Test concerning the difference between two probabilities

Be there two samples each complying with the statistical *i.i.d.* model, with the sample size N_1 and N_2, respectively. Denoting by m_1 the number of successes of the event (letter occurrence) in the first data sample, the probability estimate is $\hat{p}_1 = (m_1 / N_1)$. Similarly, for m_2 in the second data sample, the probability estimate is $\hat{p}_2 = (m_2 / N_2)$. We mean to establish whether the two estimates $\hat{p}_1$ and $\hat{p}_2$ derive from the same theoretical probability. That is, whether $p_1 = p_2 = p$ or not.

We apply the test based on the z test value, see [**3**], [**4**]:

$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{p_1(1-p_1)/N_1 + p_2(1-p_2)/N_2}$, with:

$$p_1 = p_2 \cong (m_1 + m_2)/(N_1 + N_2). \quad (4)$$

If $|z| \le z_{\alpha/2}$ ($z_{\alpha/2}$ is the same as in (1)), then we shall consider that the two probabilities are equal. Otherwise, *i.e.* $|z| \ge z_{\alpha/2}$, we reject the equality hypothesis at an α significance level. In our experiments we have considered $\alpha = 0.05$, that means $z_{\alpha/2} = 1.96$.

This testing procedure was used for the comparison between different natural texts (texts belonging to the same field or to different fields of the language). A comparison between two texts might suppose 200×200 pairs of *i.i.d.* data sets, and therefore it is difficult to draw a conclusion. In order to surpass this difficulty we considered for each text the *i.i.d.* data set which is *representative*, see Sec. III.

III. STATISTICAL APPROACH TO THE STATIONARITY

Let us consider a natural text and the 200 samples of the first approximation to the NL obtained as in Fig. 1. We compare these 200 samples among themselves trying to find out whether letter probability is the same. As the 200 sets are not statistically independent we can not apply the usual tests of the difference between two probabilities, see Sec. II.3. Therefore we had to develop a new decision procedure in the basis of repeated tests of the hypothesis that probability belongs to an interval, Sec. II.2. This investigation was carried out for each letter of the alphabet, for each field of the NL separately, and for the whole concatenated text, see linguistic corpus, Sec. I (except for the very low frequent letters: K, W, Y, Q).

The 200 data sets are of the type denoted by $[x_1, x_2, \ldots, x_N]$, see Sec. II. **Now we investigate whether there exists a *p* letter probability to be the same in all the 200 data sets.** In order to determine the p probability of the investigated letter the estimation theory should be applied to each of these data sets. Denoting by m_i the occurrences of the letter in the N observations of the i-th sample, the probability estimate is $\hat{p}_i = (m_i / N), i = 1 \div 200$. To compute the 95% confidence interval $I_i = (p_{1,i}\,;\, p_{2,i})$, in the basis of the i-th sample, we used (1).

Note: The average of the 200 estimates will be here denoted by $\hat{p}^*$ and will represent the relative frequency of the letter in the considered NL text, (before sampling). *E.g.*, for the A letter in the whole text, $\hat{p}^*$ is the ratio between the A letter occurrences and the size of NL text. That is, $\hat{p}^*$ is computed from dependent data and therefore we can not use (1). We emphasise that $\hat{p}^*$ is an important entity for any experimenter and will be referred to in what follows.

Verifying stationarity is a difficult task, involving various criteria and tests. We shall further consider the following entities, see Fig. 2:

- $\hat{p}_{\min} = \min_{i=1\div 200} \hat{p}_i$, - minimum estimate value;
- $\hat{p}_{\max} = \max_{i=1\div 200} \hat{p}_i$, - maximum estimate value;
- $\Delta_M = \max_{i=1\div 200} p_{2,i} - \min_{i=1\div 200} p_{1,i}$, - length of reunion of the 200 confidence intervals;
- $\Delta_M^c = \max_{i=1\div 200} \hat{p}_i - \min_{i=1\div 200} \hat{p}_i$, - maximum difference between two estimates;
- $\delta_M = \max_{i=1\div 200} |\hat{p}_i - \hat{p}^*|$, - maximum difference between the $\hat{p}_i$ estimates and the relative $\hat{p}^*$ frequency;
- $\delta_m = \min_{i=1\div 200} |\hat{p}_i - \hat{p}^*|$, - minimum difference between the $\hat{p}_i$ estimates and the relative $\hat{p}^*$ frequency;

Our theoretical and experimental investigation mainly relies on the following questions:

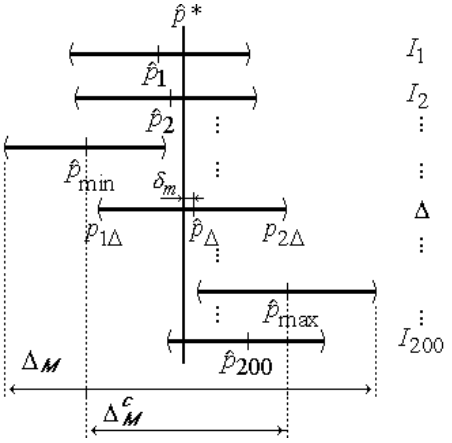

Fig. 2 Entities which point to language stationarity

1. How large the Δ_M, δ_M and Δ_M^c values are? These are important when analysing the estimates spread around the *p* theoretical value (in case *p* exists). *For the whole concatenated natural text, $\hat{p}*$, Δ_M, δ_M and Δ_M^c values can be noticed in Table 1.*

2. Are $\hat{p}_i$ estimates very close to $\hat{p}*$ and how close? To answer this question the δ_m values were obtained.

Experimentally, for all letters, in all fields of the language and in the whole concatenated text small values for δ_m are obtained. The worst case, i.e. the maximum value of the ratio $\delta_m/\hat{p}*$ is 0.014 *and corresponds to the J letter in the text translated into Romanian. $\delta_m/\hat{p}*$ values corresponding to the whole concatenated text can be noticed in Table 1.*

3. How many I_i confidence intervals cover $\hat{p}*$? Under the assumption of stationarity we expect that a large number of confidence intervals from the 200 overlap while including $\hat{p}*$.

Experimentally, for all letters, in each considered NL text at least 185 confidence intervals include the $\hat{p}*$ value. The worst case stands for the Ş letter in scientific and technical texts. At the same time, there is not a case where all the 200 I_i confidence intervals include $\hat{p}*$.

The best result consists of 198 confidence intervals of I_i type which include $\hat{p}*$; this situation stands for the E letter in the literary translated text.

4. Can we find a confidence interval for the *p* letter probability in agreement with each of the 200 experimental data sets? If the language presents stationarity such an interval should exists. That would be a *representative* confidence interval for the investigated letter and the considered NL text.

We decided to consider as *representative* that I_i confidence interval (one from the 200) corresponding to that $\hat{p}_i$ estimate which is nearest to $\hat{p}*$. This interval will be further denoted by Δ and the corresponding estimate by $\hat{p}_\Delta$ ($\hat{p}_\Delta$ leads to δ_m value) , see Fig. 2 and Table 1. *First, the reason of our decision was experimental, namely for each letter of the alphabet (in all types of the text) we could find an estimate $\hat{p}_i$ very close to $\hat{p}*$, and, moreover, $\hat{p}*$ belongs to Δ. Further on, the statistical tests of the hypothesis that the true p letter probability belongs to Δ strengthen our decision. Practically, we applied such tests (see Sec. II.2) separately for each of the 200 data samples, in all cases, namely: for each letter of the alphabet and for all considered natural texts. We found out that all the tests were passed with a 95% confidence level, with very few exceptions, discussed in Sec. IV.* **The Δ interval, alongside with the data set which produced it, enables the comparison among and between several natural texts, in the basis of the statistical test concerning the difference between two probabilities.**

5. How many confidence intervals as strong as Δ exist? The way to determine all the *strong* intervals is illustrated by Table 2 for the probability of letter A. The 200 rows correspond to the 200 estimates $\hat{p}_i$ or equivalently to their confidence intervals I_i, $i = 1 \div 200$. Successively each I_i was considered to be a *reference* interval and we applied 199 tests of the hypothesis that the probability belongs to this mentioned interval. Let us take $I_1 = (9.81;10.30)\cdot 10^{-2}$ as a reference. We verify the hypothesis that the probability of letter A belongs to I_1 based on a single data sample; *i.e.* separately each of the 199 experimental data is checked up (200 sets minus the set which produced I_1). The acceptance of the hypothesis is marked with "Yes" in Table 2, row 1 for the respective column. Otherwise, we write "No" in the respective location. The total number of samples which passed the test was filled in the last right column (here 192). Then we take into account only the rows having 199 in the last column. Experimentally we could see that there were many confidence intervals compatible with

Table 1. Experimental values multiplied by 100 for the whole text.

Letter – frequency class	$\hat{p}^*$	$\Delta_M/\hat{p}^*$	$\Delta_M^c/\hat{p}^*$	$\delta_M/\hat{p}^*$	$\delta_m/\hat{p}*$	$\Delta/\hat{p}^*$
High frequency: E, I, A	9.8÷12.3	9÷12	5÷7	2.7÷3.6	< 0.01	4.4÷5
Medium frequency: R, N, T, U, C, L, O, S, Ă, D, P, M	2.8÷7.5	13÷24	7÷15	3.8÷8.4	< 0.05	5.7÷9.3
Low frequency: Ş, Î, F, T, V, G, B, Z, Â, H	0.4÷1.3	30÷61	17÷37	8.6÷21	< 0.15	13÷25
Very low frequency: J, X	0.2÷0.25	85÷87	50÷55	27÷31	0.21÷0.23	33÷35

the rest of the 199 samples (in the last column, 199 occurred many times). *The worst case was for the L letter in literary texts namely 30 confidence intervals as strong as* Δ. *The best situation was for the T letter, in literary text, namely 185 confidence intervals* ***as*** *strong as* Δ. **The existence of many 95% confidence intervals in good agreement with all the 200 data sets points to the stationarity of the language.**

Conclusions to Sec. III: **To determine the *representative* confidence interval for the letter probability and natural text, an experimenter should fulfil the following two steps:**

- **select among the 200 I_i confidence intervals the one corresponding to that $\hat{p}_i$ which is nearest to $\hat{p}^*$;**
- **verify whether the rest of the 199 experimental data sets validates the hypothesis that the true *p* letter probability belongs to that interval;**

If each and every (or, at least, almost each) data set passes the test, then the interval becomes the Δ *representative* interval.

Table 2. *Strong* confidence intervals for letter A probability (the probability values are multiplied by 100).

i	$\hat{p}_i$ (%)	I_i (%)		Sample *i*								Total of Yes
		$p_1^{(i)}$	$p_2^{(i)}$	1	2	...	74	...	90	...	200	
1	10.05	9.81	10.30		Yes		No		Yes		Yes	192
...												
74	9.48	9.25	9.72	No	Yes				No		Yes	168
...												
90	10.15	9.90	10.40	Yes	Yes		No				Yes	159
...												
200	9.74	9.50	9.98	Yes	Yes		Yes		Yes			199

IV. EXPERIMENTAL RESULTS. THE RELEVANCE OF THE STUDY

1. THE MAIN QUESTION THIS STUDY DEALS WITH is whether the 200 data sets obtained through the periodical sampling from Fig. 1 confirm the same theoretical *p* letter probability. Experimentally, that meant the determination of the Δ type *representative* intervals, separately for each letter of the alphabet, through repeated tests of the hypothesis that probability belongs to an interval. Table 3 contains some of the experimental results. Each selected letter in Table 3 stands for a frequency class of the letters (see also Table 1). From Table 3, one can notice that for each letter and for each considered type of natural text there is an estimate $\hat{p}_\Delta$ very close to the relative frequency $\hat{p}^*$ (see also the comments concerning δ_m entity, Sec. III). The confidence interval corresponding to $\hat{p}_\Delta$ – Eq. (1) – was denoted by $\Delta=(p_{1\Delta}; p_{2\Delta})$. This interval becomes *representative* because for each letter, in each considered text, all the 199 data sets passed the test of the hypothesis that probability lies within the Δ; the significance level tests was $\alpha = 0.05$. The only exceptions were for: the M letter in the whole concatenated text; the D, G, A, and the L letter in the literary text; the R and the V letter in the literary text translated into Romanian. For each of these exceptions, only one sample (from 199) did not pass the test for the $\alpha = 0.05$ significance level. When considering $\alpha = 0.02$ there are no more exceptions left. The relative error $\hat{\varepsilon}_r$, Eq. (1), represents the accuracy by which we determined letter probability. For the high and medium frequency letter, the $\hat{\varepsilon}_r$ value was less than 0.1 (note that for the whole text $\hat{\varepsilon}_r$ was less than 0.05). We also wondered whether our satisfaction concerning the hypothesis $p \in \Delta$ (p is the true letter probability) is

Table 3. Quantitative results multiply by 100, pointing to the language stationarity

Letter	Field	$\hat{p}^*$	$\hat{p}_\Delta$	$\Delta = (p_{1\Delta}; p_{2\Delta})$	$\hat{\varepsilon}_r$	β_1	β_2
A	Whole	9.80	9.80	9.56 ; 10.05	2.36	0.00	0.00
	literature	9.83	9.83	9.41 ; 10.28	4.20	0.00	0.00
	science	10.15	10.15	9.73 ; 10.57	3.92	0.00	0.00
	translation	9.42	9.42	9.01 ; 9.84	4.17	0.00	0.00
M	whole	2.85	2.85	2.72 ; 2.99	4.54	0.00	0.00
	literature	3.22	3.22	2.98 ; 3.48	7.61	0.01	1.26
	science	2.39	2.39	2.19 ; 2.60	8.43	0.09	3.30
	translation	3.00	3.00	2.77 ; 3.24	7.65	0.01	1.34
H	whole	0.40	0.40	0.36 ; 0.45	12.25	6.13	24.91
	literature	0.43	0.43	0.34 ; 0.53	21.26	49.77	67.06
	science	0.36	0.36	0.29 ; 0.44	22.05	51.73	68.35
	translation	0.42	0.42	0.34 ; 0.52	20.66	47.44	65.51
J	whole	0.22	0.22	0.19 ; 0.26	16.62	26.87	49.89
	literature	0.20	0.20	0.15 ; 0.27	30.99	73.06	81.41
	science	0.29	0.29	0.23 ; 0.37	24.31	59.49	73.26
	translation	0.16	0.16	0.11 ; 0.22	34.12	77.56	84.07

justified or not, and therefore we calculated the size of type II statistical error denoted by β, see Sec. II.2, Eq. (3). We carried out the experiments for each letter of the alphabet and for each considered natural text. The quantitative results are presented in the last two columns of Table 3. The probabilities β_1 and β_2 represent β calculated for two values $\delta_1 = 0.15$ and $\delta_2 = 0.2$, respectively. That is: $\beta_1 \equiv \beta((1-0.15)a)$; $\beta_2 \equiv \beta((1-0.2)a)$; a stands for the lower confidence limit of Δ, *i.e.* $a = p_{1\Delta}$. **The determining of the Δ *representative* intervals with such a good statistical error control, *i.e.* $\hat{\varepsilon}_r < 0.1$; $\alpha = 0.05$; $\beta < 0.05$ for the high and medium frequency letters points to the Romanian language stationarity.**

2. TO ALSO VERIFY ROMANIAN LANGUAGE STATIONARITY we divided each natural text into two parts and we compared the obtained halves by means of the statistical test of difference between two probabilities, see Sec. II.3. We applied this procedure to each type of text. **All these tests were passed**. The *i.i.d.* data sets sets involved in the comparison were the *representative* samples (*i.e.* those which produced the Δ confidence intervals). Note: even for half-texts we could determine the Δ type intervals, but with lower accuracy.

3. We also used THE Δ *REPRESENTATIVE* INTERVALS ALONGSIDE WITH THEIR *I.I.D.* DATA SETS IN COMPARING different NL fields. This comparison was carried out in the basis of the test of difference between two probabilities, Sec. II.3. Such tests were applied to each letter and to representative samples from the following types of texts: genuine Romanian literature vs. scientific and technical; genuine Romanian literature vs. translation; scientific and technical vs. translation. Table 4 gives the z test values, see (4). The situations when the test failed are written in bold characters, *i.e.* $z > 1.96$, meaning different probabilities for the respective letter in the two compared fields (for example, see letter A in scientific and technical vs. translation test). From Table 4 we can see that there are differences concerning probabilities in the three fields.

Table 4. The *z* test value in the comparison among and between probabilities, according to relation (4)

Compared fields	A	M	H	J
Genuine Romanian literature vs. scientific and technical	1.07	**5.19**	1.15	1.90
Genuine Romanian literature vs. translation	1.43	1.28	0.04	1.09
Scientific and technical vs. translation	**2.56**	**3.95**	1.13	**3.02**

4. We further wondered about THE RELEVANCE OF OUR QUANTITATIVE RESULTS for another experimenter, who disposes of another similar linguistic corpus for printed Romanian. Suppose that he wants to verify whether the new corpus confirms the letter probability obtained by us. How should he proceed?

- He may use the same procedure we did when we compared the fields of the NL using the test of the difference between two probabilities, see point 3, above. For that, he has to determine the *representative i.i.d.* data set and the corresponding $\hat{p}_i$ estimate, for his NL text; then, he may use (4), considering our *representative* values given in Table 3. For the size of our corpus, see Sec. 1.
- He may apply a test of the hypothesis that the probability belongs to the Δ interval given by us. He may consider his *representative* data set. If the test is passed, our results will be confirmed. He also can evaluate the β probability, that is the size of type II error, by means of (3).

V. FINAL REMARKS

The main conclusions obtained in this paper are:

- **A fundamental result:** there is a very nice mathematical behaviour of printed Romanian, concerning the first approximation to NL. All the measurements concerning probability were carried out with a good statistical error control; that means low values for the $\hat{\varepsilon}_r$ relative error and for the sizes of the two types of statistical errors (α and β) involved in the tests by which we validated the *representative* confidence intervals for letter probability.
- The Δ representative confidence intervals alongside with their *i.i.d.* data sets are **useful elements when comparing different natural texts.**

A new experimenter who wants to continue verifying printed Romanian stationarity can find in the paper a guide to his experiments. He may also design, by means of (3), the length of his linguistic corpus in order to obtain low values for the size of type II error, when comparing it to our results.

ACKNOWLEDGEMENT The authors are grateful to Dr. Dan Tufiş, Member of the Romanian Academy and acknowledge for his scientific help in this paper.

REFERENCES

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

[**2**] Adriana Vlad, and A. Mitrea, "*Estimating conditional probabilities and digram statistical structure in printed Romanian*", in "Recent Advances in Romanian Language Technology, Dan Tufis & Poul Andersen Editors, Ed. Academiei Române, ISBN 973-27-0626-0, Bucharest, 1997, pp. 57-72.

[**3**] J. Devore, *Probability and Statistics for Engineering and the Sciences*, second edition, Brooks/Cole Publishing Company, Monterey, California, 1987.

[**4**] B.R. Frieden, *Probability, Statistical Optics and Data Testing*, Springer-Verlag, Berlin, Heildelberg, New York, 1983.

[**5**] A. Mood, F. Graybill, and D. Boes, *Introduction to the Theory on Statistics*, third edition, McGraw-Hill Book Company, 1974, pp. 427-428.