THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 5, Number 3/2004, pp. 000-000

INFORMATION SOURCES APPROXIMATING TO PRINTED ROMANIAN: THE ROLE OF TYPE II STATISTICAL ERROR

Adriana VLAD*, **, Adrian MITREA*, Mihai MITREA*, ***

* Faculty of Electronics and Telecommunications, "POLITEHNICA" University of Bucharest - Romania
** Research Institute for Artificial Intelligence, Romanian Academy, Bucharest – Romania
*** ARTEMIS Project Unit, GET/INT Evry – France
email: adriana_vlad@yahoo.com, avlad@racai.ro

This paper belongs to a larger study the authors dedicated to mathematical knowledge of printed Romanian. It offers theoretical and quantitative results constituting a guide to be used when designing a linguistic corpus for mathematical modelling. Our investigation focuses on zero memory information sources having as symbols the *m*-grams (letters, *di*grams, *tri*grams, *tetra*grams) and the words of the printed Romanian. When obtaining the mathematical models corresponding to these sources, the test on the hypothesis that probability belongs to an interval plays a special role. The type II statistical error involved in this test finally determines the accuracy of the models.

Key Words: information sources approximating to a natural language; test on the hypothesis that probability belongs to an interval; type II statistical error.

1. INTRODUCTION

The study the authors dedicated to printed Romanian modelling by means of statistical methods means a permanent confrontation between several mathematical behaviours, supposed to be true for any natural language (NL) and the reality, here represented by printed Romanian texts, [1-9].

A main direction for this confrontation was the accuracy by which printed Romanian verifies the stationarity hypothesis concerning the *m*-gram and word structures (*m*-gram means a set of *m* successive letters). The stationarity hypothesis is included in the general assumption that any NL is well approximated by a multiple ergodic Markov chain, [10].

By developing an original statistical approach, we provided evidence for the stationarity hypothesis on the basis of *m*-gram ($m \leq 4$) and word structures and we obtained the mathematical models for corresponding information sources. Additionally, the conditional probabilities on one preceding letter were computed (the first order Markov chain approximating to printed Romanian, [1], [9]).

The theoretical and experimental result unifying all these modelled information sources is synthesised in the way in which the *representative* confidence interval for the linguistic entity probability was computed, (1). The *representative* interval is obtained together with a *representative i.i.d.* data set sampled form the NL text in order to apply the statistical inferences on probability. (Note: *i.i.d.* stands for observations coming out from *i*ndependently and *i*dentically *d*istributed random variables).

Although these entities are different (*m*-grams, words) and although they were investigated on various corpora, our study proved that printed Romanian allows *representative* confidence intervals to be built up in a very simple form, suitable for any experimenter:

$$p = p^*(1 \mp \varepsilon_r) \quad , \ \varepsilon_r = z_{\alpha/2}\sqrt{(1-p^*)/Np^*} \tag{1}$$

In Eq. (1), p^* stands for the relative frequency of the investigated entity. As for example, when investigating *m*-grams, p^* is the ratio of the linguistic entity occurrence number to the total number of *m*-grams in the corresponding natural text. Such a result also holds for word structure, when p^* is the ratio

Recommended by Dan TUFIŞ, member of the Romanian Academy

of the occurrence number of the investigated word to the length of the natural text (in words), [6], [9]. The $z_{\alpha/2}$ value is the point value corresponding to the standard Gaussian law, while ε_r represents the experimental relative error in probability estimation.

Next, we briefly present the reason for which formula (1) leads to the *representative* confidence interval for probability.

In order to get to the true probability of the linguistic entity - if there is such a probability - we have to extract experimental data from the natural text which comply with the *i.i.d.* statistical model. There are many ways to sample the natural text in order to obtain such *i.i.d.* data. In our study we used a large enough fixed sampling period (*i.e.* 200 characters for *m*-gram structures or 200 words for word structure) so as to practically eliminate the dependency between successive observations. By shifting the sampling origin, we could obtain 200 *i.i.d.* experimental data sets. Each sample obtained in this way consists of *N* observations (where $N = L/200$ and L is the length of the natural text) and leads to a probability estimate and to confidence limits for probability. Note: the *i*ndependence among the observations is a consequence of the large sampling period and the *i*dentically *d*istribution derives from the stationarity hypothesis assumed for the NL. On the other hand, each and every *i.i.d.* data set brings the same information about the searched probabilities, if the NL features stationarity.

For each investigated entity and for each natural text, one of the 200 probability estimates was practically equal to p^* so that its corresponding confidence interval is practically computed according to (1); it will be further denoted by Δ. Moreover, each and every of the remaining of 199 *i.i.d.* data sets confirmed the hypothesis that the true probability belongs to the Δ interval. (Note: this statistical test is our extension, [2-6], [8], [9] of a similar test applied to the mean in [11]; we had to consider such a test because the 200 *i.i.d.* data sets are not independent. The test is summarised in Section 2.). To conclude with, these results are a proof that each of the 200 *i.i.d.* data sets provides the same information on the probability of the investigated linguistic entity. We stated Δ as *representative* confidence interval for probability. (In fact, there were several confidence intervals among the 200 which were in agreement with the overall natural text but we preferred Δ because it is most easily computed by any experimenter using Eq. (1)).

Hence, Eq. (1) is legitimate for the natural text, as it was validated by all the experiments we carried out: any *m*-gram/word, any corpus, no matter its length (from global corpora of about 50 millions characters to individual books of about 1 million characters), and any type of text (literature, science, mixed). Note: depending on the natural text length, the number of *m*-grams/words which could be investigated was different. We took into consideration only the linguistic entities which complied with the de Moivre - Laplace condition checked up in the experimental form: $Np^*(1-p^*) \geq 20$.

The elements which establish the accuracy of the models we offered for the considered information sources derive from the procedure by which the *representative* interval was computed, (1), namely:

- The confidence level for the probability. In our experiments, we considered $1-\alpha = 0.95$; hence, $z_{\alpha/2} = 1.96$ in (1).
- The size of the *representative* confidence intervals or, equivalently, the ε_r relative error in probability estimation, (1).
- The type II statistical error probability when applying the test on the hypothesis that probability belongs to Δ interval.

The present paper focuses on the β probability of the type II statistical error involved by the test on the hypothesis that probability belongs to an interval. As always, β means to take wrong data for good ones. In our application, β measures how justified our decision is when we state that Δ is *representative*.

Note that type I error probability (which is lower than $\alpha = 0.05$) should not be further discussed here. When these tests are not passed by all (or, at least, almost all) *i.i.d.* data sets sampled from the language, we cannot speak about a mathematical model for the considered information source. As these tests were passed, we could ascertain the very existence of such a model by defining the *representative* confidence intervals but we can still be wrong. The lower the β values, the lower our suspicions concerning the existence and the accuracy of the model we provided.

We shall briefly present the linguistic corpora involved in our experiments (see details in [4], [9]):

- The whole mixed corpus is built up by 93 books, representing various printed Romanian fields: genuine literature and foreign literary works translated into Romanian (novels and short stories), scientific books (law, medicine, forestry, *etc.*) and other types of texts (correspondence, memories, *etc.*). Starting from these books, we shall further consider three types of corpora. 1) **#WMCB** denotes the **W**hole **M**ixed **C**orpus when the alphabet consists of the 31 letters of printed Romanian and of the **B**lank character. 2) **#WMC** corresponds to the alphabet containing only the 31 letters. 3) **#WMCW** considers the natural text as a chain of words. **#WMCB** and **#WMC** are involved in the *m*-gram information source, while **#WMCW** is involved in the word information source investigations. The corresponding lengths are: 51809386 characters for **#WMCB**, 43002953 characters for **#WMC**, and 8806433 words for **#WMCW**.
- The whole literary corpus is built up by 58 books (novels and short stories), out of which 11 books represent genuine Romanian literature and 47 represent foreign literary works translated into Romanian. Here again we shall consider three types of corpora: 1) **#WLCB** (**W**hole **L**iterary **C**orpus in **B**lank case), 2) **#WLC** (**W**hole **L**iterary **C**orpus in no blank case), and 3) **#WLCW** (**W**hole **L**iterary **C**orpus considered as a chain of **W**ords). The corresponding length are: 35548447 characters for **#WLCB**, 29293212 characters for **#WLC**, and 6255235 words for **#WLCW**.

Section 2 highlights the test on the hypothesis that probability belongs to an interval. Section 3 intends to be a guide, by means of β values, in designing a new linguistic corpus aiming at accurate printed Romanian modelling.

2. TYPE II STATISTICAL ERROR INVOLVED IN PRINTED ROMANIAN MODELLING

In this section, we briefly present the test on the hypothesis that probability belongs to an interval as it was practically applied for printed Romanian modelling. We shall specify each entity in the test when the investigated event is either an *m*-gram or a word occurrence in printed Romanian.

Be an experimenter who wants to find out whether the probability of a certain event belongs to a fixed $(a;b)$ interval. His analysis is based on a single $\lfloor x_1, x_2, ..., x_N \rfloor$ data sample which complies with the *i.i.d.* statistical model. Be m the number of successes of the event in the N observations. The ratio $\hat{p} = m/N$ is the estimate for the unknown p probability of the event.

The two statistical hypotheses (the null hypothesis H_0 and the alternative hypothesis H_1) are: $H_0: a < p < b$ and $H_1: p \notin (a;b)$. It should be verified, with a chosen α significance level, whether the experimental data are in agreement with H_0 or not. The region meaning that the null hypothesis is accepted is the $(c_1;c_2)$ interval, (2):

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

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

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

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

$\beta(p)$ takes high values when p is very close to $(a;b)$ interval. That means that p should be in the left side of a, but very close to a, or in the right side of b, but very close to b. That is, $p \le (1-\delta)\cdot a$ or $p \ge (1+\delta)\cdot b$, where δ is a small quantity chosen by the experimenter, depending on application.

In this paper, $\lfloor x_1, x_2, ..., x_N \rfloor$ sample is one of the 200 *i.i.d.* data sets, periodically sampled from the natural text. The fixed $(a;b)$ interval for the test procedure is the 95% *representative* confidence interval, computed according to (1).

The N sample size corresponds to a sample period of 200, when considering the natural text either as a chain of characters or as a chain of words. Note that $(a;b)$ interval can be completely determined on the basis of the p^* values measured on the referred corpus.

In our numerical results, in Section 3, we determined the β value for the most disturbing case for us, namely $p=(1-\delta)\cdot a$, where a stands for the left *representative* confidence limit, $a = p^*(1-\varepsilon_r)$, with ε_r according to Eq. (1).

In order to obtain a good accuracy, we first computed N as to ensure a small enough ε_r relative error. However, not all the corresponding β values were low, especially when the experimenter required low δ values in computing $\beta(p)$ according to Eq. (3) (this means that a good accuracy is required). In order to obtain a good accuracy, N size has to be determined to ensure low β values, see Section 3. As a consequence, as regards the accuracy, β is the most important element.

3. A GUIDE IN DESIGNING A LINGUISTIC CORPUS

Here the discussion will follow three directions.

A) An evaluation of the accuracy of the proposed models, [2]-[6]

In our models, the probability of each linguistic entity is computed by means of (1) with the N values corresponding to our linguistic corpora (see Section 1).

For a 95% confidence level in probability estimation and for fixed N values, the model accuracy depends upon p^* relative frequency of the investigated event (measured by the experimenter in the natural text) and also upon the δ quantity chosen by the experimenter (see Section 2). Based on p^* and δ values one can compute ε_r relative error in the probability determination, Eq. (1), and also β values (the size of type two statistical error when verifying that probability belongs to Δ interval).

Fig. 1 presents β as a function of p^* for $\delta = 0.15$, in each of the 4 corpora involved in the study (i.e. **#WMC**, **#WMCB**, **#WLC** and **#WLCB**). Table 1 presents the overall results for different frequency classes.

For example, let us suppose that the experimenter is interested in an accuracy of his measurements expressed by $\beta < 0.2$. In this situation, he can investigate only those linguistic entities with relative frequencies p^* larger than the limited values obtained from Fig. 1 (for *m*-grams). For **#WMCB**, the investigated *m*-grams have to fulfil the condition $p^* > 1.15\times10^{-3}$, according to the plot denoted by "∘" in Fig. 1. Similarly, for **#WMC**, **#WLCB** and **#WLC** (denoted by "×", "◊" and "Δ" respectively), these inferior limits are 1.35×10^{-3}, 1.65×10^{-3} and 2×10^{-3} respectively. If we investigate words instead of *m*-grams, see Fig. 2, (for the same accuracy $\beta < 0.2$ and $\delta = 0.15$) in the whole mixed corpus **#WMCW** one can investigate only words having the relative frequencies $p^* > 7\times10^{-3}$; in the whole literary corpus **#WLCW** one can investigate only the words for which $p^* > 9\times10^{-3}$.

For fixed p^* and δ, β takes different values corresponding to the different sizes of the referred linguistic corpora: a larger size of the linguistic corpus leads to a smaller β value (the probability of an unjustified joy decreases when the corpus size increases).

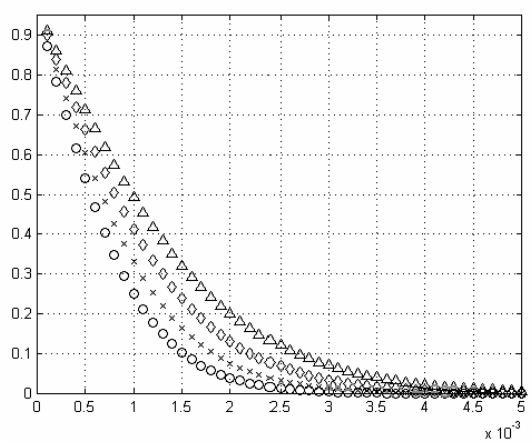

Fig. 1 β **- type II statistical error probability – corresponding to *m*-gram structure in printed Romanian, computed for** $\delta = 0.15$ **.** Horizontal axis – p^* values; vertical axis – β values. The 4 plots correspond to the following corpora: **#WMCB** "o", **#WMC** "x", **#WLCB** " ◊ " and **#WLC** " Δ ".

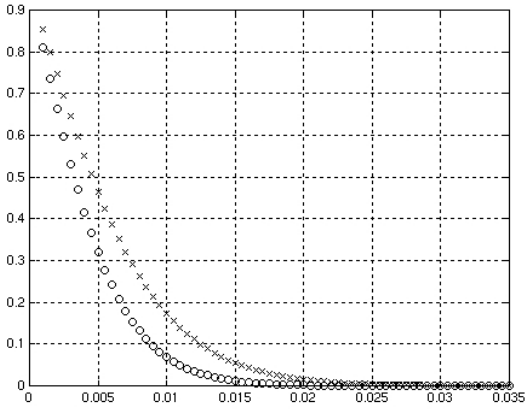

Fig. 2 β **- type II statistical error probability – corresponding to word structure in printed Romanian, computed for** $\delta = 0.15$ **.** Horizontal axis – p^* values; vertical axis – β values. The 2 plots correspond to the following corpora: **#WMCW** "o" and **#WLCW** "x".

Concerning the β values, Table 1 presents rough information (no values with a lot of digits). We present numerical results only for the mixed and literary corpora in case with blank. The frequency classes are organised by splitting the range values in a logarithmic way. The inferior limits are: 0.1; 0.05; 0.02; 0.01; 0.005; 0.002; 0.001; 0.0005.

For example, for the *m*-grams from **#WMCB** for which the relative frequencies are in the range 0.002 and 0.005, the Δ *representative* confidence interval is determined with a relative error $\varepsilon_r < 10\%$ (column 2), see (1). For $\delta = 0.1$, the type II statistical error probability is $\beta < 0.3$ (column 3). This upper limit for β decreases a lot, *i.e.* $\beta < 0.05$ (column 4), when the experimenter accepts $\delta = 0.15$; it becomes practically zero (column 5) when he accepts $\delta = 0.2$.

Based on our study on the *m*-gram structure, [2-5], we can state the following:

i) The first 6 rows in Tab. 1 $\left(p^* > 0.2 \times 10^{-2}\right)$ *correspond to:*

- letters that cover more that 99.5% of the size of **#WMCB**;
- digrams that cover more that 80% of the size of **#WMCB**;
- trigrams that cover more that 25% of the size of **#WMCB**;
- tetragrams that cover more that 5% of the size of **#WMCB**.

In **#WMCB** the *m*-gram probabilities for which $p^* > 0.2 \times 10^{-2}$ were computed with a 95% statistical confidence level and with an ε_r relative error lower than 10%. The β values for which we validated these results are lower than 0.05 for $\delta = 0.15$ and a statistical significance level $\alpha = 0.05$. We can notice a very good accuracy for letters and digrams.

ii) The first 8 rows in Tab. 1 $\left(p^* > 0.05 \times 10^{-2}\right)$ *correspond to:*

- practically all the investigated letters in **#WMCB**;
- digrams that cover more that 95% of the size of **#WMCB**;
- trigrams that cover more that 60% of the size of **#WMCB**;
- tetragrams that cover more that 20% of the size of **#WMCB**.

The *m*-gram probabilities in these 8 frequency classes in **#WMCB** were computed with a 95% statistical confidence level and an ε_r relative error lower than 20%. However β is large for those *m*-grams belonging to the last two frequency classes (*i.e.* the last 2 rows in Tab. 1): $\beta < 0.6$ for $\delta = 0.15$ and $\beta < 0.3$ for $\delta = 0.2$.

Table 1 β **- type II statistical error probability – corresponding to *m*-gram structure in printed Romanian, case with blank. 1.** Limits of frequency classes; **2.** and **6.** ε_r , the relative errors for **#WMCB** and **#WLCB** respectively; **3.**, **4.** and **5.** β values in **#WMCB**, computed for $\delta = 0.10$, $\delta = 0.15$ and $\delta = 0.20$; **7.**, **8.** and **9.** β values in **#WLCB**, computed for $\delta = 0.10$, $\delta = 0.15$ and $\delta = 0.20$. All the values in the table are multiplied by 100.

p^*	#WMCB				#WLCB			
	ε_r	β			ε_r	β		
		$\delta = 10$	$\delta = 15$	$\delta = 20$		$\delta = 10$	$\delta = 15$	$\delta = 20$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$(10;20)$	< 2	$\cong 0$	$\cong 0$	$\cong 0$	< 2	$\cong 0$	$\cong 0$	$\cong 0$
$(5;10)$	< 2	$\cong 0$	$\cong 0$	$\cong 0$	< 5	$\cong 0$	$\cong 0$	$\cong 0$
$(2;5)$	< 5	$\cong 0$	$\cong 0$	$\cong 0$	< 5	$\cong 0$	$\cong 0$	$\cong 0$
$(1;2)$	< 5	$\cong 0$	$\cong 0$	$\cong 0$	< 5	$\cong 0$	$\cong 0$	$\cong 0$
$(0.5;1)$	<10	< 5	$\cong 0$	$\cong 0$	<10	<10	$\cong 0$	$\cong 0$
$(0.2;0.5)$	<10	< 30	< 5	$\cong 0$	<15	< 50	<15	$\cong 0$
$(0.1;0.2)$	<15	< 60	< 25	<10	< 20	< 70	< 50	< 20
$(0.05;0.1)$	< 20	< 75	< 60	< 30	< 30	< 80	< 70	< 50

B) The case of a new experimenter who accepts relation (1), but makes a parallel study on different corpora

This situation is analysed in Tab. 2. The table brings into discussion the linguistic entities that can be investigated applying our method when the experimenter has a given text at his disposal: a book, a group of books written by the same author, different linguistic corpora. For example, when the experimenter has at his disposal one book with $5000 \times 200 = 1000000$ characters (this is quite a long book), he will use the first row from Tab. 2. If he is interested in measurements with a very good accuracy (this means $\beta \le 0.10$ for $\delta = 0.10$ - see column 2), then he can investigate only *m*-grams with $p^* \ge 16.10 \times 10^{-2}$. Practically, this means that he can investigate nothing but the blank (the rest of investigated *m*-grams having relative frequencies less than this value).

Table 2 Inferior limits for p^* **relative frequencies that can be investigated with a desired accuracy. 1.** The size of *i.i.d.* data sample; **2.** and **3.** Inferior limits for p^* corresponding for $\beta = 0.1$ and $\delta = 0.10$ and $\delta = 0.15$ respectively ; **4.** and **5.** Inferior limits for p^* corresponding to $\beta = 0.2$ and $\delta = 0.10$ and $\delta = 0.15$ respectively; **6.** and **7.** Inferior limits for p^* corresponding to $\beta = 0.3$ and $\delta = 0.10$ and $\delta = 0.15$ respectively. All the values from columns **2.** – **7.** are multiplied by 100.

	$\beta = 10$		$\beta = 20$		$\beta = 30$	
	$\delta = 10$	$\delta = 15$	$\delta = 10$	$\delta = 15$	$\delta = 10$	$\delta = 15$
1	**2**	**3**	**4**	**5**	**6**	**7**
$N = 5000$ (for example, a book)	16.10	8.15	12.21	6.08	9.61	4.73
$N = 15000$	5.94	2.85	4.39	2.10	3.39	1.62
$N = 25000$	3.64	1.72	2.67	1.27	2.06	0.97
$N = 35000$	2.62	1.23	1.92	0.91	1.48	0.70
$N = 75000$	1.24	0.58	0.90	0.42	0.69	0.32
$N = 150000$ (for example, **#WLC**)	0.62	0.29	0.45	0.21	0.34	0.16
$N = 1000000$ (a very large corpus)	0.08	0.04	0.06	0.03	0.05	0.02

If he has a longer text of about $15000 \times 200 = 3000000$ characters (a group of books written by a same author), for the same accuracy ($\beta \le 0.10$ and $\delta = 0.10$) he can investigate only the *m*-grams for which $p^* \ge 5.94 \times 10^{-2}$. Practically, this means that he can investigate only the blank and the most frequent 3 letters (A, E and I).

Aiming at an accurate model for the natural language (even limiting the study to the *m*-gram and word statistical structures) needs investigation on quite large linguistic corpora (at least of the size of our literary

corpus). On the other hand, if the experimenter investigates relative small corpora, he can take advantage of our NL stationarity study and he can compute the probability of the searched entity using relation (1) with the corresponding p^* and N values (from his corpora).

C) Designing a new linguistic corpus to ensure the desired accuracy of the model

In fact, the problem is how long should be a new corpus an experimenter has to use in order to accurately check up the concordance between his numerical results and our reference quantities. The referred quantities consist of the mathematical model corresponding to the *m*-gram and word structure, already obtained by us for printed Romanian. That is, for any linguistic entity (*m*-grams or word) we determine the *representative* confidence for probability by means of (1) where p^* and N values correspond to our referred corpora, see Section 1. (We suppose that the new linguistic corpus is organised in the same manner as our referred corpora.)

The experimenter has to check up if the probability of any searched linguistic entity investigated on the new corpus is contained by the corresponding *representative* confidence interval from our model. For example, if there is a new literary corpus of novels and short stories, he wants to verify if the A letter occurrences in his experimental data confirms the fact that the true unknown letter A probability is contained in the *representative* confidence interval computed on our referred literary corpus. Let us consider he applies the test on the hypothesis that probability belongs to an $(a;b)$ interval, where $(a;b)$ is the referred 95% *representative* confidence interval, and $\hat{p}$ from Section 3 is p^* relative frequency from his corpus. If the test is passed, the question is how large the new corpus should be so as to ensure low β values.

For fixed α and δ, β depends both on the length of the new natural text and on the p^* value (also measured on the new corpus). Note that the accuracy has to be ensured for each and every *m*-gram/word in printed Romanian or, at least, for some frequency classes, see Fig. 3.

Fig. 3 presents the N size of the i.i.d. data sample as a function of p^*, computed for $\delta = 0.1$, $\alpha = 0.05$ and a fixed β. The *representative* confidence intervals were computed for *m*-grams in our mixed corpus - **#WLMB**. There are three plots, for the three β values: $\beta = 0.05$ – plotted in 'o', $\beta = 0.1$ – plotted in 'x' and $\beta = 0.2$ – plotted in '◊'. We can notice the steep slopes of the curves: this requires much larger corpora when we have to consider frequency classes with $p^* < 2 \times 10^{-2}$.

Other information needed for the design of the new corpus, can be seen in Tab. 2. The accurate value for N can be obtained by numerically solving Eq. (3) (also involves Eq. (2)).

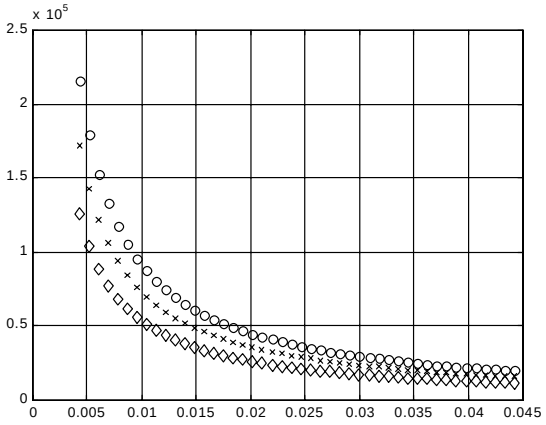

Fig. 3. The dependency between N size of the *i.i.d.* data sample and p^*.

4. CONCLUSIONS

Modelling the information sources approximating the printed Romanian means to determine the probabilities of the corresponding linguistic entities, in case such probabilities do exist. Our study brought into evidence the existence of a 95% *representative* confidence interval for each investigated linguistic entity (*m*-gram/word), denoted by Δ. The *representative* qualifier was granted to Δ out of a statistical approach in which the test on the hypothesis that probability belongs to an interval plays a special role. Hence, we should take into account the β probability corresponding to this test. First of all, this β value measures how much you can enjoy the models you obtained. Secondly, β determines the length of a linguistic corpus which can afford a sound mathematical investigation. Note that we started the investigation by designing our corpus based on the ε_r constraints. Although this condition upon ε_r is very important, in order to answer whether there is a mathematical model for the natural language and how accurately such a model can be obtained, β is the decisive element.

On the other hand, note that Eq. (1) is remarkable for printed Romanian enabling a *representative* confidence interval to be computed on p^*; thus, a mathematical meaning for the relative frequency measured on natural text is revealed. Hence, we may say that our study can be also useful for an experimenter who is not interested in an accurate mathematical investigation: he can now just take advantage on p^*. Of course, for a mathematical investigation, large corpora are required, but good practical results can be also obtained on small ones. This derives from the very good concordance between the reality and the mathematical model, *i.e.* this derives from the printed Romanian stationarity as revealed by Eq. (1).

ACKNOWLEDGEMENT

The authors would like to acknowledge the continuous encouragement and scientific support of Prof. Dr. Dan Tufiş, Corresponding Member of the Romanian Academy, in the development of their study dedicated to printed Romanian.

REFERENCES

1. VLAD, A., MITREA, A., *Estimating conditional probabilities and digram statistical structure in printed Romanian*, Recent Advances in Romanian Language Technology, Dan Tufis & Poul Andersen Editors, Academiei Publishing House, Bucharest, ISBN 973-27-0626-0, pp.57-72, 1997. http://www.racai.ro/books/awde/vlad.html.
2. VLAD, A., MITREA, A., MITREA, M., POPA, D., *Statistical methods for verifying the natural language stationarity based on the first approximation. Case study: printed Romanian*, Proc. VEXTAL'99, Nov. 22-24, 1999, Venice-Italy, Ed. Unipress, ISBN 88-8098-112-9, pp. 127-132; http://byron.cgm.unive.it/events/vlad.pdf.
3. VLAD A., MITREA A., MITREA M., *Verifying Printed Romanian Language Stationarity Based on the Digram Statistical Structure*, Proceedings of the Romanian Academy, Series A, Vol. **I**, No. 2/2000, pp. 129-139.
4. VLAD A., MITREA A., MITREA M., *The trigram statistical structure in printed Romanian*, ROMJIST (Romanian Journal of Information Science and Technology), Vol. **4**, No. 3/2001, pp. 353-372.
5. VLAD A., MITREA A., MITREA M., *Estimating tetragram probabilities by using multiple data samples from a natural text. Case study: printed Romanian*, Proc. IPMU2002, July 2002, Annecy-France, pp. 1285–1292.
6. VLAD A., MITREA A., *Contribuții privind structura statistică de cuvinte în limba română scrisă*, Limba Română în Societatea Informațională - Societatea Cunoașterii, D. Tufiș and F. G. Filip Editors, Romanian Academy, Ed. Expert, Bucharest, pp. 207-233, 2002.
7. VLAD A., MITREA A., MITREA M., *Two frequency–rank laws for letters in printed Romanian*, Procesamiento del Languaje Natural, Revista N^o **24**, ISSN 1135-5948, pp. 153-160, Septiembre de 2000.
8. VLAD A., MITREA A., MITREA M., *A Corpus - based Analysis of how Accurately Printed Romanian Obeys Some Universal Laws*", Chap. 15 in "A Rainbow of Corpora: Corpus Linguistics and the Languages of the World", A. Wilson, P. Rayson, T. McEnery Editors, Lincom-Europa Publishing House, ISBN 3895868728, Munich, pp. 153-165, 2003.
9. VLAD A., MITREA A., MITREA M., *Limba română scrisă ca sursă de informație*, Paideia Publishing House, Bucharest, 2003.
10. SHANNON, C. E., *Prediction and Entropy of Printed English*, Bell Syst. Tech. J., vol. **30**, pp. 50-64, January 1951.
11. MOOD, A., GRAYBILL, F., BOES, D., *Introduction to the Theory on Statistics*, third edition, McGraw-Hill Book Company, pp. 427-428, 1974.

Received March 3, 2004