ORTHOGRAPHY AND PUNCTUATION MARKS IN PRINTED ROMANIAN MODELLING: A STUDY BASED ON THE LETTER STATISTICAL STRUCTURE IN A LITERARY CORPUS

Adriana VLAD[*,**], Adrian MITREA[*], Mihai MITREA[*]

[*] University "Politehnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology
[**] The Research Institute for Artificial Intelligence, Romanian Academy

Corresponding author: Prof. Adriana VLAD, adriana_vlad@yahoo.com, avlad@racai.ro

This paper is part of a larger study dedicated by the authors to printed Romanian modelling. Here, the statistical investigation establishes whether and how accurately the printed language can be mathematically modelled, when orthography and punctuation marks are considered. In this respect, the authors developed an original statistical approach based on multiple confidence intervals in probability estimation and on two types of statistical tests. The experimental results, which are new for printed Romanian, refer to the letter statistical structure in a literary linguistic corpus of 58 books (novels and short stories).

Key words: Mathematics of natural language, natural language stationarity, orthography and punctuation marks, statistical error control.

1. INTRODUCTION

The present research study belongs to the field of natural language (NL) processing, focusing on the mathematical behaviour of printed Romanian when the alphabet is extended with orthography and punctuation marks. Note that up to now the orthography and punctuation marks have not been too much investigated, at least so as to be included in the mathematical modelling of NL, [1].

By extending a statistical approach developed in some of our previous works, [2]-[10], this paper aims at establishing whether and how accurately the reality (printed Romanian) verifies a theoretical hypothesis, namely the stationarity. This hypothesis is included in the general assumption that the natural languages are well approximated by multiple ergodic Markov chains, [11]. We can speak about a unitary mathematical model if and only if this hypothesis is true.

The investigation in this paper refers to letter statistical structure. Although our initial intention was to include only hyphen and period in the modelling, this study covers more items, namely comma, quotation marks, quotation dash, question mark, exclamation mark, colon, semicolon, parentheses and apostrophe.

Our experimental work processed the literary linguistic corpus previously organised in [4]-[10] and here extended by orthography and punctuation marks. The corpus contains 58 books (novels and short stories) written in the new orthography (after 1993), most of them published by Metropol, Paideia and Univers Publishing Houses (Bucharest, Romania). These books represent genuine Romanian literature (11 books) and foreign literary works translated into Romanian (47 books).

To the initial alphabet consisting of 31 letters of the Romanian alphabet (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 without any distinction between upper and lower case letters) and the blank character (denoted by _), in this study we added 15 orthography and punctuation marks explained below.

- **.** symbol. This symbol is used in texts in three situations (consequently, it brings three elements in the extended alphabet), namely:

- full stop (a point that marks the end of a sentence); in Table 1 it was denoted by the **.** sign
- abbreviation point (a point that marks the shortened form of a word); in Table 1 it was denoted by % sign
- ellipsis (a set of three consecutive points indicating that words are deliberately left out of a sentence); in Table 1 it was considered as a single element and was denoted by } sign

❖ **-** symbol. This symbol is used in texts in three situations (hence, it brings three elements in the extended alphabet), namely:
- hyphen; in Table 1 it was denoted by - sign
- quotation dash; in Table 1 it was denoted by { sign
- em dash (a mark introducing an additional text with explanation purposes, somehow replacing the parentheses); in Table 1 it was denoted by * sign

❖ **,** symbol (comma)

❖ **:** symbol (colon)

❖ **;** symbol (semicolon)

❖ **?** symbol (question mark)

❖ **!** symbol (exclamation mark)

❖ **“** symbol (quotation marks); no distinction between the beginning and the closing quotation marks was made; that means only one new element in the extended alphabet

❖ **(** and **)** symbols (parentheses); that means two elements in the extended alphabet

❖ **’** symbol (apostrophe).

In this paper, we shall denote by *letter* any element from the extended alphabet, namely the 31 Romanian alphabet letters, the blank as well as the orthography and punctuation marks described above. The extended alphabet sums up 47 such *letters*.

The largest corpus we considered in our analysis is the **W**hole **L**iterary **C**orpus with **O**rthography and Punctuation Marks, denoted by **#WLCO**; it contains the 58 books concatenated in a random order and has a total size of $L = 37070049$ *letters*. In our investigations, we also considered smaller corpora obtained by partitioning **#WLCO** in halves, quarters and eighths:

❖ the two halves, denoted by **#1HWLCO** and **#2HWLCO** (with sizes of $L = 18535025$ *letters* and $L = 18535024$ *letters*, respectively);

❖ the four quarters, denoted by **#1QWLCO**, **#2QWLCO**, **#3QWLCO** and **#4QWLCO** (with sizes of $L = 9267513$ *letters* or $L = 9267512$ *letters*);

❖ the eight eighths, denoted by **#1EWLCO**, **#2EWLCO**, ..., **#8EWLCO,** (with sizes of $L = 4633757$ *letters* or $L = 4633756$ *letters*).

The stationarity investigation was organised in two main steps:

A) The first step relies on an algorithm obtained by extending the one considered in [3], [5]-[10], briefly presented in subsection 2.1. It aims at determining *a representative confidence interval for each investigated letter in the analysed corpus*.

If such an interval does exist, we can go deeper in our study, by trying to get to a mathematical model for the letter structure in the literary field.

B) The second step contains a mathematical comparison among and between natural texts, aiming at:

- strengthening printed Romanian stationarity by comparing natural texts belonging to the same field of the language;
- finding out whether we can speak about a model of the literary field or not (on the basis of the *letters* structure).

In order to carry out these mathematical comparisons we reconsidered the procedure we proposed in [5]-[10], briefly presented in subsection 2.2. We compared the smaller corpora obtained from **#WLCO** (halves, quarters, eighths).

Section 3 presents the overall quantitative results. Our experimental study reveals the existence of stationarity concerning *letter* structure in printed Romanian and enables obtaining the mathematical model for *letter* structure in the literary field.

2. THEORETICAL BACKGROUND

2.1. Obtaining *Representative* Confidence Intervals for Probability in NL

The statistical investigation requires to extract from the natural text experimental data sets complying with the *i.i.d.* statistical model (*i.e.* data coming out from *i*ndependently and *i*dentically *d*istributed random variables). There are several ways to obtain such data sets. In our study, we applied a periodical sampling of the natural text, with a large enough period (200 *letters*) so as to destroy the dependence between successive *letters*. By shifting the sampling origin in the natural text, 200 data sets individually complying with *i.i.d.* statistical model are obtained, figure 1. Although these 200 samples are not independent data sets, if the stationarity hypothesis is true, they would convey the same information about the probability of the investigated event (*letter* occurrence). Consequently, in order to find out whether we can speak about a mathematical model of the *letter* structure, we check up whether the 200 *i.i.d.* data sets confirm the same probability or not.

Our investigation begins by computing the $p*$ relative frequency for each *letter* in each analysed text. $p*$ is the ratio of the number of *letter* occurrences to the *L* length of the natural text (the values for *L* are given in Section 1).

Be m_i the occurrence number of the investigated *letter* in the *i*-th sample (one of the 200 *i.i.d.* data sets), $[x_1, x_2, ..., x_N]$. The *N* sample size is equal to the ratio of *L* to 200. By applying the estimation theory, each of the 200 *i.i.d.* data sets leads to an estimated value for the *letter* probability, $\hat{p}_i = m_i / N$, $i = 1 \div 200$, and to the corresponding confidence interval, $I_i = (p_{1,i}; p_{2,i})$. The $p_{1,i}$ and $p_{2,i}$ confidence limits for the *p* true unknown searched probability are:

$$p_{1,i} \cong \hat{p}_i - z_{\alpha/2}\sqrt{\hat{p}_i(1-\hat{p}_i)/N} \qquad p_{2,i} \cong \hat{p}_i + z_{\alpha/2}\sqrt{\hat{p}_i(1-\hat{p}_i)/N} \tag{1}$$

where $z_{\alpha/2}$ is the $\alpha/2$ - point value of the normal (Gaussian) law of 0 mean and 1 variance. In our experimental study we used a statistical confidence level of 95%, corresponding to $z_{\alpha/2} = 1.96$.

```
        1                      201                  401
        CÂND_GAITTANY ... ZÂMBIND_DECONCERTAT ... NU-I_PLĂCEA, ...
1:      C                       _                    _
2:       Â                       D                    P
…
200:                            D                    I
```

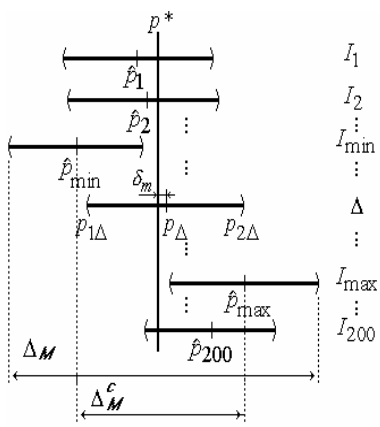
Figure 1. 200 *i.i.d.* experimental data sets obtained through a periodical sampling of the natural text. Each observation is a *letter*.

The confidence interval in equation (1) depends on the experimental data; hence, it is random. In our study, for the same *letter*, 200 confidence intervals are obtained. Our problem is to decide which of these 200 intervals better represents the *letter* probability. The following entities and questions guided our investigation, see figure 2.

1. How large the Δ_M^c, δ_M, Δ_M and δ_m values are? Δ_M^c and δ_M are important when analysing the spread of the estimates around $p*$. Δ_M gives an idea about the largest interval where *p* (the true *letter* probability) could lie. δ_m leads to that $\hat{p}_i$ estimate which is nearest to $p*$; this estimate will be further denoted by p_Δ. We denoted by Δ the confidence interval corresponding to the p_Δ estimate.

2. How many I_i confidence intervals cover $p*$? Under the assumption of stationarity we expect that a large number of intervals overlap while including $p*$ ($p*$ is the average of the 200 estimates).

3. Can we find a confidence interval for the p probability in agreement with each of the 200 experimental data sets? Are there several such intervals? If the language features stationarity such intervals should exist. To answer these questions, we successively applied a procedure where each I_i, $i = 1 \div 200$, was considered to be a fixed interval and 199 tests of the hypothesis that p probability belongs to this mentioned interval were applied, see A1, Appendix. (We consider this procedure to be absolutely necessary when comparing the 200 data samples among themselves in order to check up the language stationarity. That is because these data sets are not independent sets and therefore we could not apply the more usual test, A2 from Appendix.) If, for a certain interval, all these 199 tests (or, at least, almost all) are passed, then we shall consider that the searched probability really exists. All the confidence intervals which are validated by this procedure are equally good in representing the *letter* probability. However, if Δ interval is one of them, it will be preferred and stated as *representative* because it is the easiest to be obtained and dealt with by any experimenter.



$i = 1 \div 200$

- $\hat{p}_{\min} = \min_i \hat{p}_i$, minimum estimated value
- $\hat{p}_{\max} = \max_i \hat{p}_i$, maximum estimated value
- $\Delta_M = \max_i p_{2,i} - \min_i p_{1,i}$, union of the 200 confidence intervals
- $\Delta_M^c = \max_i \hat{p}_i - \min_i \hat{p}_i$, maximum difference between two estimates
- $\delta_M = \max_i |\hat{p}_i - p^*|$, maximum difference between the $\hat{p}_i$ estimates and p^*
- $\delta_m = \min_i |\hat{p}_i - p^*|$, minimum difference between the $\hat{p}_i$ estimates and p^*.

Figure 2. The 200 statistical confidence intervals for *letter* probability and the associated entities.

2.2. A Mathematical Comparison between Natural Texts

The above stationarity investigation was completed by a mathematical comparison between natural texts, aiming at strengthening the NL stationarity and finding out whether we can speak about a model for the investigated field (literature) on the basis of the *letter* structure.

We compared natural texts (corpora from Section 1) having in view the following: (1) to compare the *letters per se*, that is to see whether a certain *letter* has the same probability in the two compared texts;
(2) to compare *ranks* instead of *letters*, that is to see whether the *letters* on the same frequency rank in the two texts have the same probability. For example, in the two halves, **#1HWLCO** and **#2HWLCO,** *letters* A and I correspond to rank 3; we shall check up whether *letters* A and I have the same probability or not.

All the mathematical comparisons were carried out on the basis of the two statistical tests from Appendix, namely the test of the hypothesis that probability belongs to an interval, *A1,* and the test of the equality between two probabilities, A2.

For each of the two natural texts and for each investigated *letter*, we first determined the Δ *representative* confidence interval alongside with its *i.i.d. representative* data set.

When applying the test A1 from Appendix, the $(a;b)$ interval is the Δ *representative* interval from the first natural text involved in the comparison and the $[x_1, x_2, ..., x_N]$ *i.i.d.* set is the *representative* sample from the second text. The test was applied in the two situations: corpus1 *vs.* corpus2 and corpus2 *vs.* corpus1.

When applying the test A2 from Appendix we considered the two *representative i.i.d.* data sets from the two natural texts involved in the comparison.

Both tests were applied for $\alpha = 0.05$ statistical significance level; that is, the probability of rejecting good data is lower than 0.05. For the test A1 we also evaluated the β probability of the type II error (to accept wrong data as good ones).

It is possible that the mathematical comparisons of the corpora point to some differences concerning *letter* probability. However, if the rank based comparisons do not present any differences, we can still sustain the idea of a unique model for the investigated field. If the rank based comparisons indicate important differences, then we can no longer speak about the same model in the compared corpora, and we should consider the possibility of different models (a model characterizing an author or a group of authors). To conclude with, the rank based comparison is absolutely necessary to check up the very existence of a model for the investigated field.

Note: The key idea in our procedure to mathematically compare the two natural texts was to first identify the *representative* entities (namely, the confidence interval and the *i.i.d.* data set in each of the two compared texts). A direct comparison between two texts might suppose 200×200 pairs of *i.i.d.* data sets, and therefore it would be difficult to draw a conclusion.

3. EXPERIMENTAL RESULTS

We applied our statistical approach to each linguistic corpus mentioned in Section 1. We computed the confidence interval for probability, equation (1), only for those *letters* which fulfilled the de Moivre - Laplace condition, checked up in the form $N\,p^*(1-p^*) > 20$, where N is the *i.i.d.* data size, ($N = L/200$).

We have to mention an overall important result we obtained: when applying the statistical procedure described in subsection 2.1, we could determine a *representative* confidence interval of Δ type and a *representative i.i.d.* data set for each investigated *letter* and each analysed corpus, even for extended alphabet case. Each time, the p^* relative frequency was practically the center of the *representative* confidence interval. That is, each time, one of the 200 $\hat{p}_i$ estimated values, denoted by p_Δ, was practically equal to p^*; p_Δ led to δ_m entity in figure 2 (see the very low values for the ratio δ_m / p^* in Table 1).

The extended alphabet sorted in decreasing order of the p^* values in **#WLCO** is:

_(17.14); E(9.06); I(7.87); A(7.87); R(5.39); N(5.11); U(4.90); T(4.77); C(4.17); L(3.54); S(3.48); O(3.22); Ă(3.21); D(2.72); P(2.51); M(2.45); ,(1.43); Ş(1.22); Î(1.11); V(0.97); F(0.93); .(0.93); B(0.85); Ţ(0.79); G(0.78); Â(0.72); -(0.61); Z(0.56); H(0.37); J(0.19); {(0.18); "(0.14); ?(0.12); !(0.10); X(0.09); :(0.09); K(0.09); *(0.07); }(0.06); Y(0.06); ;(0.04); ((0.03);)(0.03); W(0.03); %(0.01); '(0.01); Q(0.00)

(in this hierarchy the numerical values for p^* are multiplied by 100).

Having p^* and N we can compute the *representative* confidence interval for the probability of each *letter* in **#WLCO,** Eq. (1), where $\hat{p}_i$ is replaced by p^*.

Table 1 contains a selection of the experimental results covering all the orthography and punctuation marks in **#WLCO** (abbreviation point and apostrophe *letters* don't fulfill de Moivre - Laplace condition) and the A *letter*. A detailed analysis was made to indicate the existence of the *representative* confidence interval for *letter* probability in each linguistic corpus (**#WLCO**, **#1HWLCO**, **#2HWLCO,** *etc*.).

How to read table 1

For the A *letter*, the relative frequency in **#WLCO** is $p^* = 7.87 \times 10^{-2}$. There are 191 I_i confidence intervals (out of 200) which contain p^*. There is an estimated value $\hat{p}_i$ practically equal to p^*: the ratio of δ_m to p^* is about 0 (this means $p_\Delta = 7.87 \times 10^{-2}$). The minimum value among 200 estimates is $\hat{p}_{\min} = 7.70 \times 10^{-2}$ and the maximum value is $\hat{p}_{\max} = 8.06 \times 10^{-2}$. The spread of the 200 estimates around p^*

is indicated by the ratio $\Delta_M^c / p^* = 4.68\times10^{-2}$. The ratio of Δ_M to p^* is equal to 7.93×10^{-2}; this gives an idea about the largest interval where the true searched *letter* A probability could lie. The ratio of δ_M to p^* is 2.51×10^{-2}, where δ_M is the maximum distance between the p_i estimates and p^*. The Δ confidence interval (the *representative* one) for A *letter* probability in **#WLCO** is $(7.74; 8.00)\times10^{-2}$. Column 13 presents information concerning the relative error in determining *letter* A probability: $\Delta / p^* = 3.25\times10^{-2}$ (about twice the relative error). There were 111 I_i (out of 200 intervals) in total compatibility with **#WLCO** (*i.e.* each one of these 111 intervals was confirmed by all the 199 tests of the hypothesis that the probability belongs to it); Δ is one of these 111 intervals and was decided to be the *representative* one (Δ is the easiest to be used by any experimenter).

Table 1. Obtaining 95% *representative* confidence intervals for probability in #WLCO.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
_	1	17.14	196	0.00	16.82	17.39	3.28	5.48	1.84	16.95	17.33	2.20	82
A	4	7.87	191	0.00	7.70	8.06	4.68	7.93	2.51	7.74	8.00	3.25	111
,	17	1.43	190	0.01	1.37	1.50	9.41	17.02	4.88	1.38	1.49	7.61	157
.	22	0.93	197	0.00	0.88	0.97	9.50	18.97	4.84	0.88	0.97	9.47	188
-	27	0.61	195	0.02	0.56	0.65	14.77	26.45	7.63	0.57	0.64	11.70	150
{	31	0.18	193	0.07	0.15	0.21	30.52	52.01	16.09	0.16	0.20	21.46	124
"	32	0.14	192	0.13	0.12	0.17	35.16	59.49	18.86	0.13	0.16	24.29	114
?	33	0.12	192	0.19	0.09	0.13	38.78	64.80	23.35	0.10	0.13	26.70	74
!	34	0.10	193	0.19	0.08	0.11	38.28	67.73	19.89	0.08	0.11	29.50	146
:	36	0.09	188	0.02	0.07	0.10	38.12	69.28	20.61	0.07	0.10	31.06	164
*	38	0.07	190	0.07	0.05	0.09	46.54	80.82	23.98	0.06	0.08	34.65	119
}	39	0.06	187	0.15	0.05	0.08	50.07	85.91	26.02	0.05	0.08	35.97	128
;	41	0.04	189	0.38	0.03	0.06	60.03	104.72	32.83	0.03	0.05	44.48	136
(	42	0.03	187	0.67	0.02	0.04	69.90	122.00	37.46	0.02	0.04	53.45	130
)	43	0.03	191	0.08	0.02	0.04	74.48	127.56	39.89	0.02	0.04	53.26	114
%	45	0.01	-										
'	46	0.01	-										

Note: 1. *Letter*; 2. *Letter* relative frequency rank in the corpus; 3. p^* – *letter* relative frequency in the corpus; 4. The number of I_i intervals containing p^*; 5. Ratio of δ_m to p^*; 6. $\hat{p}_{\min} = \min_i \hat{p}_i$; 7. $\hat{p}_{\max} = \max_i \hat{p}_i$; 8. Ratio of Δ_M^c to p^*; 9. Ratio of Δ_M to p^*; 10. Ratio of δ_M to p^*; 11.,12. Confidence limits for Δ; 13. Ratio of Δ to p^*; 14. The number of I_i intervals in total compatibility with the corpus. *All the values (except for columns 1, 2, 4 and 14) are multiplied by 100.*

Numerical results when comparing various corpora are presented in table 2.

Some details concerning Table 2 are presented below. The first two columns contain the two compared corpora (note that when comparing the eighths, only some particular cases were reported; however, the other cases – *e.g.* #2EWLCO *vs.* #3EWLCO – feature a similar behaviour). Column 3 gives the number of investigated *letters* (which fulfilled the $N p^*(1-p^*) > 20$ condition on the two *representative* data sets involved in the comparison). Columns 4 and 5 show how many *letters* did not pass the test of the hypothesis that the probability belongs to an interval. Column 4 refers to the situation when the $(a;b)$ interval is the *representative* confidence interval for corpus 1 and the $[x_1, x_2, ..., x_N]$ sample is the *i.i.d. representative* data set for corpus 2. Similarly, column 5 checks up whether the probability from corpus 1 belongs to the *representative* confidence interval from corpus 2. Column 6 gives the number of *letters* which are rejected by test *A2*, *Appendix*. Columns 7-9 contain the same type of information as columns 4-6 with the only difference that this time the comparisons considered the ranks instead of *letters* themselves.

When accepting a lower α significance level for test *A2*, columns 6 and 9 from Table 2 reveal fewer non-zero values. For example, if $z_{\alpha/2} = 2.48$, there is only one non-zero value in the columns 6 and 9 from Table 2 (when comparing the ellipses from the two halves, the test value is 2.74).

The accuracy of the test concerning the hypothesis that probability belongs to an interval, *Appendix A1*, is expressed by the following items: the length of the referred $(a;b)$ interval, the α and β probabilities of the two types of statistical errors and the chosen δ value.

For instance, when considering a corpus of $L = 200 \times 180000$ *letters* (practically as long as the **#WLCO**) and $\delta = 0.15$, the β size of the type II probability error depends on the investigated *letter* and is the following (the β numerical values here after are multiplied by 100):

_(0); E(0); I(0); A(0); R(0); N(0); U(0); T(0); C(0); L(0); S(0); O(0); Ă(0); D(0); P(0); M(0); ,(0); Ș(0); Î(0); V(0); F(0); .(0); B(0); Ț(0); G(0); Â(0); -(0.02); Z(0.06); H(1.19); J(14.38); {(16.28); "(26.25); ?(32.88); !(40.75); X(45.16); :(45.16); K(45.16); *(54.94); }(60.28); Y(60.28); ;(71.67); ((77.64);)(77.64); W(77.64); %(89.93); '(89.93); Q(95.00)

As it can be seen, an acceptable accuracy ($\beta \le 0.4$ and $\delta = 0.15$) cannot be obtained for very low frequent *letters*, *i.e.* for the *letters* with $p^* \le 0.001$, namely ! X : K * } Y ; ( ) W % ' Q.

A smaller corpus of about $L = 200 \times 90000$ *letters* (*i.e.* as long as the **#1HWLCO**) enables the same accuracy ($\delta = 0.15$ and $\beta \le 0.4$) for entities having their $p^* \ge 0.002$, *i.e.* for all the *letters* except for J { " ? ! X : K * } Y ; ( ) W % ' Q. When considering just quarters of the literary corpus (*e.g.* **#1QWLCO**) with length of about $L = 200 \times 45000$ *letters*, it can be seen that the accuracy decreases: $\delta = 0.15$ and $\beta \le 0.4$ allows us to investigate just the *letters* with $p^* \ge 0.004$.

When comparing parts from the literary field, practically no differences were noticed. There were few exceptions, which disappeared when the comparison was carried out according to the *rank* instead of the *letters per-se*.

Table 2. Comparing various texts on the basis of the *letter* statistical structure (Columns 4-9 contain the number of letters which are rejected by the test).

Compared texts		No.	Comparison according to *letter*			Comparison according to rank		
Corpus 1	Corpus 2		Test *A1* 1 vs. 2	Test *A1* 2 vs. 1	Test *A2*	Test *A1* 1 vs. 2	Test *A1* 2 vs. 1	Test *A2*
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
#1HWLCO	#2HWLCO	44	1	1	2	0	0	2
#1QWLCO	#2QWLCO	39	1	1	2	0	0	0
#1QWLCO	#3QWLCO	40	0	0	3	0	0	0
#1QWLCO	#4QWLCO	41	0	0	0	0	0	0
#2QWLCO	#3QWLCO	39	0	0	2	0	0	2
#2QWLCO	#4QWLCO	39	2	2	2	0	0	5
#3QWLCO	#4QWLCO	40	1	1	3	0	0	0
#1EWLCO	#2EWLCO	35	0	0	1	0	0	0
#1EWLCO	#3EWLCO	33	0	0	0	0	0	0
#1EWLCO	#4EWLCO	34	0	0	2	0	0	0
#1EWLCO	#5EWLCO	36	0	0	0	0	0	0
#1EWLCO	#6EWLCO	34	0	0	1	0	0	0
#1EWLCO	#7EWLCO	36	0	0	0	0	0	0
#1EWLCO	#8EWLCO	35	0	0	0	0	0	0

Note: Part of the experimental results concerning p^* and the corresponding Δ *representative* confidence interval for probability are given in [12], where a larger mixt corpus (including **#WLCO**) was considered. In this paper we investigated the **#WLCO** going deeper in verifying printed Romanian language stationarity. The type II statistical errors in granting Δ interval as *representative* are evaluated; also mathematical comparisons between natural texts are added.

4. FINAL REMARKS

Let us now summarise the main steps of our study.

Considering successive approximations to NL is compulsory in order to get to a mathematical model for a printed natural language (*i.e.* in order to verify whether and how accurately the printed language can be approximated by an ergodic Markov chain with a multiplicity order larger than 30). Note that taking into account orthography and punctuation marks raises additional suspicions concerning the validity of this model.

Obtaining *representative* intervals in all analysed cases (*i.e.* for each *letter* in each corpus) is the first important result bringing evidence in favour of a mathematical model for the *letter* statistical structure. Note that it would not have been possible to speak about *letter* probability in a NL field if various *i.i.d.* data sets had not been in agreement among themselves. This result also encourages us to continue this type of investigation for higher order structures (*e.g.* digrams, trigrams, tetragrams, words when orthography and punctuation marks are included).

The study is further completed by mathematical comparisons between natural texts in order to strengthen the stationarity hypothesis and to point to a mathematical model for the literary field. Note that even if the comparisons had indicated differences between the probabilities of the two compared corpora (but it was not the case here), the *representative* intervals are still important. In case such differences would exist, we could think about different models by authors or group of authors (this last aspect is due to the existence of the *representative* interval). We expect this to be the case when we will resume this study on scientific corpora.

It should be emphasised that our study provides additional relevance to the usual p^* relative frequency: it becomes the centre of the *representative* confidence interval for probability. This makes it possible for any NL experimenter to investigate and take advantage of the connection between the meaning conveyed by the natural text and its mathematical description. After all, aren't relative frequencies (for letters, digrams, …, words, conditional events) such powerful tools in cryptanalysis (code breaking)?

The fact that the punctuation and orthography marks are subject to the same mathematical regularities as the letters from the alphabet (as it results from our study) would support the idea that these marks are not mere conventional signs, but rather related to the meaning conveyed by the natural text (at least for the literary field).

ACKNOWLEDGEMENT

The authors want to acknowledge the continuous encouragement and scientific support from Prof. Dan Tufiş, corresponding member of the Romanian Academy, in the development of their study dedicated to printed Romanian modelling. This work is part of a larger research theme included in the program of The Research Institute for Artificial Intelligence, Romanian Academy.

REFERENCES

1. SAY B., AKMAN A., *Current approaches to punctuation in computational linguistics*. Computer and the Humanities, **Vol. 30** (1997), pp. 457–469.
2. VLAD A., MITREA A., *Estimating conditional probabilities and digram statistical structure in printed Romanian*. In "Recent Advances in Romanian Language Technology", Tufiş D. and Andersen P. editors, Romanian Academy Publishing House, Bucharest, (1997), pp. 57–72; http://www.racai.ro/books/awde/vlad.html.
3. 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. of the VEXTAL'99, Venice-Italy, pp. 127–132, (Nov. 1999); http://byron.cgm.unive.it/events/vlad.pdf.
4. VLAD A., MITREA A., MITREA M., *Two frequency–rank laws for letters in printed Romanian*. SEPLN Procesamiento del Lenguaje Natural, **Revista No. 26** (2000), pp. 153–160, http://www.sepln.org/revistaSEPLN/revista/26/index.html.
5. VLAD A., MITREA A., MITREA M., *Verifying printed Romanian language stationarity based on the digram statistical structure*. Proc. of the Romanian Academy, Series A, **Vol. I**, **No. 2** (2000), pp. 129–139
6. 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.

7. VLAD A., MITREA A., MITREA M., *Estimating tetragram probabilities by using multiple data samples from a natural text. Case study: printed Romanian*. Proc. of the 9th Intl. Conf. on Information Processing and Management of Uncertainty in Knowledge–Based Systems - IPMU2002, Annecy-France, pp. 1285–1292, (July 2002).
8. VLAD A., MITREA A., MITREA M., *A corpus – based analysis of how accurately printed Romanian obeys some universal laws*. In "A Rainbow of Corpora: Corpus Linguistics and the Languages of the World", Wilson, A., Rayson, P., McEnery, T. editors, Lincom-Europa Publishing House, Munich (2003), pp. 153–165.
9. VLAD A., MITREA A., MITREA M., *Limba română scrisă ca sursă de informație* (*Printed Romanian Language as an Information Source*), Paideia Publishing House, Bucharest, (2003).
10. VLAD A., MITREA A., MITREA M., *Printed Romanian modelling: the m-grams and the word information sources*. In "Speech Techonology and Human-Computer Dialogue", Burileanu C. editor, Romanian Academy Publishing House, Bucharest (2003) pp. 79–98.
11. SHANNON C.E., *Prediction and entropy of printed English*, Bell Syst. Tech. J., **vol. 30** (1951), pp. 50–64.
12. VLAD A., MITREA A., MITREA M., *Letter statistical structure in Printed Romanian language when orthography and punctuation marks are included*. Proc. of the IEEE Intl. Conf. Communications'2006, Bucharest,, pp. 127–130, (June 2006).
13. MOOD A., GRAYBILL F., BOES D., *Introduction to the theory on statistics*, 3rd edn., McGraw-Hill Book Company (1974).
14. WALPOLE R.E., MYERS R.H., *Probability and statistics for engineers and scientists*, 4th edn., MacMillan Publishing Comp., New York (1989).

APPENDIX

A1. Test of the Hypothesis that the Probability Belongs to an Interval

This testing procedure is our extension of a similar test applied to mean in [13].

The present test is to decide whether the probability of a certain event (*letter* occurrence) belongs to a fixed $(a;b)$ interval, based on a single $[x_1, x_2, ..., x_N]$ data sample which complies with the *i.i.d.* statistical model. In this paper, $[x_1, x_2, ..., x_N]$ can be any of the 200 experimental data sets, sampled from NL according to figure 1. Be m the number of successes of the searched event in the N observations. The $\hat{p} = m/N$ ratio is the estimate for the p unknown probability of the event. The testing procedure follows as such:

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

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

We have to verify, with a chosen α significance level, whether the experimental data are in agreement with H_0 or not. The null hypothesis H_0 will be accepted if and only if the $\hat{p}$ estimate falls within the $(c_1;c_2)$ interval:

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

Two types of errors might be encountered:

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 to accept H_0 although it is false. This happens when $c_1 < \hat{p} < c_2$, although 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 $\beta(p)$:

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

$\beta(p)$ takes high values when p is very close to $(a;b)$ interval, *i.e.* when $p=(1-\delta)\cdot a$ or $p=(1+\delta)\cdot b$, where the δ is a small quantity, see figure 3. The experimenter is to decide upon the δ value, depending on the application.

This test enabled our stationarity investigation: we had to decide whether two estimates computed on two *i.i.d.* samples *which are not independent sets* derive from the same theoretical probability or not.

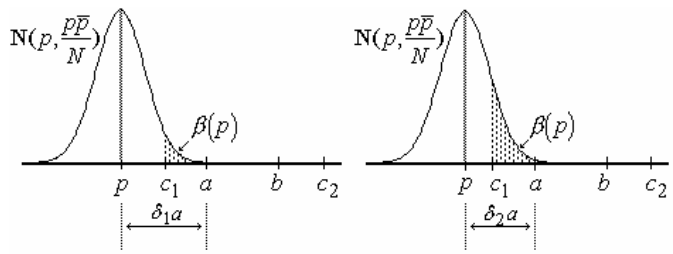


Figure 3. The type II statistical error in the test on the hypothesis that probability belongs to an interval. The $\beta(p)$ probability (the shadowed area) is computed in two situations: $p=(1-\delta_1)a$ and $p=(1-\delta_2)a$. If $\delta_1>\delta_2$, then $\beta((1-\delta_1)a)<\beta((1-\delta_2)a)$.

A2. Test of the Equality Between two Probabilities

Be there two samples each complying with the *i.i.d.* statistical 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 want 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, [14]:

$$z=(\hat{p}_1-\hat{p}_2)/\sqrt{p_1(1-p_1)/N_1+p_2(1-p_2)/N_2} \tag{4}$$

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

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

This testing procedure was used for the comparison between different natural texts.