A MATHEMATICAL COMPARISON BETWEEN SEVERAL SINGLE AUTHOR CORPORA

Stefan CIUCĂ[1], Adriana VLAD[2], and Adrian MITREA[3]

Lucrarea se focalizează pe o comparație matematică între mai multe corpusuri, fiecare corpus provenind din texte literare scrise de un singur autor, urmărind să dea un răspuns unei probleme deschise în literatura de specialitate: dacă şi în ce condiții se poate vorbi de un model lingvistic mai general al limbii române sau dacă variabilitatea este mult prea mare şi se poate vorbi doar de model de autor. Pentru comparații s-a folosit o procedură statistică originală avansată într-o serie de studii precedente, aici extinsă şi adaptată la diversele forme ale corpusurilor. Procedura de comparație implică în primul rând determinarea probabilității evenimentelor lingvistice cu interval de încredere statistică reprezentativ pentru corpus. Această decizie de a determina intervalul reprezentativ se bazează pe estimarea probabilității cu multiple intervale de încredere statistică şi pe repetate teste de apartenență a probabilității la interval. Decizia finală este susținută şi de considerarea acurateței determinărilor, implicând în testele menționate ambele tipuri de erori statistice. Studiul experimental este făcut pe cinci corpusuri construite independent, fiecare corpus este analizat în detaliu pentru fiecare eveniment lingvistic.

The paper focuses on a mathematical comparison between several single author corpora looking to give an answer to an open problem in literature: if and what are the terms one can speak of a general linguistic model or the author variability is too influent and we can only have separate author models. For the comparisons, an original procedure advanced by the authors in some previous studies was used, here extended and adapted for various forms of the corpora. That procedure implies the determination of the probability with a representative confidence interval for every investigated linguistic event in each analyzed corpus. The decision of determining the representative interval for probability is based on the probability estimation with statistical confidence intervals and also on tests verifying the hypothesis that the probability belongs to a certain interval. The final decision is also supported by the accuracy of the results considering the two types of error probability involved in the statistical tests. The experimental study is done on five independently built corpora, each of them being made of novels written by only one author. For each of them a detailed linguistic event analysis was made.

Keywords: literary corpus linguistics, confidence intervals for probability, mathematical comparison between corpora, statistical tests for probability.

[1] Faculty of Electronics, Telecommunications and Information Technology, POLITEHNICA University of Bucharest, Romania, email: istefanciuca@gmail.com

[2] Prof., Faculty of Electronics, Telecommunications and Information Technology, POLITEHNICA University of Bucharest, Romania and The Research Institute for Artificial Intelligence, Romanian Academy , adriana_vlad@yahoo.com, avlad@racai.ro

[3] Faculty of Electronics, Telecommunications and Information Technology, POLITEHNICA University of Bucharest, Romania

1. Introduction

The paper presents a mathematical comparison between five literary corpora built from novels from five different authors, looking to sustain with accurate results a debate whether or not a general statistical model is available for the literary field or the author influences are too strong and we can only consider single, different author models. In order to give some answers for this question and to support them mathematically, a series of statistical procedures and theories developed by the authors in some previous studies were used [1]-[3], [5] and [6]. Some of these studies had as one of the main purposes to determine the impact and role the orthography and punctuation marks had on the general statistical model. However, the impact that orthography and punctuation marks have and also the differences brought by the different average length of words and phrases to the general statistical model, remain still a considerable issue, [4], for which this paper brings some contributions. In the previous studies for printed Romanian [2], [3], [5], the comparison was carried out on quite large concatenated corpora containing various authors so that the discussion was based on an average statistical behavior (not specifically including author models).

To mathematically sustain an answer concerning the language/author models, this paper, for the first time for printed Romanian, brings into comparison single author corpora, large enough to give a certain level of accuracy to the experimental results. The five independent corpora and the five authors are described in the Table 1. These five corpora were, first of all, statistically investigated in great detail using the methods from [1]-[3], [5]-[7]. We started the investigation using the texts in their original forms including orthography and punctuation marks. Thus, the alphabet consists of 47 characters: 31 characters are the letters composing 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), one is the space (blank, denoted by _ character) and 15 represents punctuation marks and orthography. The 31 letters are all upper case symbols. The 15 punctuation and orthography characters are those considered in [2], [3] and [5]: the full stop (a point that marks the end of a sentence); the abbreviation point (a point that marks the shortened form of a word); the ellipsis (a set of three consecutive points indicating that words are deliberately left out of a sentence); the hyphen; the quotation dash; the em dash (a mark introducing an additional text with explanation purposes, somehow replacing the parentheses); the comma; the colon; the semicolon; the question mark; the exclamation mark; the quotation marks; the parentheses; the apostrophe.

The mathematical comparison done using this form of the text brings several differences between the five authors, mainly because of the orthography and punctuation. Due to the need of having large enough corpora in order to conclude various investigations regarding the natural language processing, we decided, for this study, to analyze the text in three different forms:

- *Form 1*: the corpus alphabet consists in 33 characters: the basic 31 letters in the Romanian alphabet, the space character and the hyphen;
- *Form 2*: the corpus alphabet consists in 32 characters: the basic 31 letters in the Romanian alphabet and the space character;
- *Form 3*: the corpus alphabet consists in only the basic 31 letters forming the Romanian alphabet.

In order to carry out the statistical tests on which the mathematical comparisons are based on, we needed to determine for each investigated event, for every one of the five corpora, the *i.i.d. representative* data sample and the *representative* statistical confidence interval for probability. The method for computing these two *representative* elements is the same as the one used in the previous studies [1]-[3], [6] and [7], and it is briefly described in Section 2 of this paper. All the three *forms* of the corpora were brought into comparison and the experimental results are presented in Section 3.

The mathematical comparisons were done considering the rank probabilities of the linguistic entities and also on the probability of the linguistic entities per se.

Table 1.

The five compared corpora in each of the three analyzed forms.

Author		Dumas	Tolkien	Herbert	Asimov	Chirita
Number of books		12	3	8	9	5
Number of characters (Corpus Length)	Form 1	11 165 180	2 661 247	6 707 314	5 063 785	3 389 501
	Form 2	11 010 057	2 634 174	6 649 057	5 010 260	3 354 641
	Form 3	9 064 344	2 153 926	5 520 538	4 126 148	2 747 379
Number of words		1 945 713	480 248	1 128 519	884 112	607 262

The detailed study presented here leads to the idea (Section 3) that, when *Form 3* of the text is used, the differences between authors can be neglected. *Form 3* also mathematically supports the need to concatenate corpora from different authors. Aspects regarding the length of words, which may bring differences between authors, are underlined for *Form 1* and *Form 2*; these two forms also indicate that orthography and punctuation marks may bring a series of differences among author corpora model. Thus, to bring into discussion the concatenating of independently built corpora, it is better to use *Form 3* as a mathematical support. Also different studies from information theory domain use *Form 3* and sometimes *Form 2* for various applications [9].

2. Theoretical background

How to obtain *Representative* Confidence Intervals for the Linguistic Event Probability

As already mentioned, the focus of this paper was to determine if one can speak of a general statistical model for the literary field in printed Romanian or there are different author models. To give an answer, we first discuss if one can speak about the author model and we shall determine the mathematical elements necessary to carry out the comparisons between authors. Therefore, a specific statistical analysis for probability was done for the linguistic event for each of the three corpora *forms* which led to the *representative* confidence interval and the *i.i.d representative* data set for the five analysed corpora, elements which make the comparison possible.

Note: the investigated linguistic is an *m*-gram which represents *m* successive characters; here the linguistic event is for *m=1* (a character), *m=2* (a digram) and *m=3* (a trigram), but the presented illustrations are only for *m=1*.

The reason why we need the two *representative* elements is because from the corpus it can be sampled a high number of *i.i.d* data sets and therefore for the same investigated linguistic event we can have many statistical confidence intervals for probability. An answer is needed if these confidence intervals are compatible among themselves (in fact speaking about the same probability) and, if the answer is positive as expected, one has to be picked as *representative* for the linguistic event in the analysed corpus (this is the statistical *representative* confidence interval). The procedure is explained below when the linguistic event is represented by a *letter* (the procedure is similar for digrams and trigrams).

Although the investigated events are different (letters, digrams, trigrams) and although they were investigated on various corpora (see Table 1), 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 and is the ratio of the number of occurrences of the searched event to the total event occurrences in the corpus (the total event occurrences practically means *L m*-grams, where *L* value is the length of the corpus in characters, given in Table 1). 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; $1-\alpha$ is the confidence level.

A periodical sampling of the corpus was applied, with a large enough period (200 *characters*), to destroy the dependence between successive *m*-grams. By shifting the sampling origin in the analysed corpus, 200 data sets, individually complying with the *i.i.d.* statistical model, were obtained.

Using *Eq.* (2), for each *i.i.d.* data set (from the total of 200), a confidence interval for the probability was determined. The $p_{1,i}$ and $p_{2,i}$ are the confidence limits for the *p* true unknown searched probability, [10]:

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

where

- $z_{\alpha/2}$ is the $\alpha/2$ - point value of the normal (Gaussian) law of 0 mean and 1 variance. In the experimental study a 95% statistical confidence level is used, corresponding to $z_{\alpha/2} = 1.96$.
- $\hat{p}_i = m_i/N$, $i = 1 \div 200$ is an estimate in the *i*-th data set. m_i is the occurrence number of the investigated event in the *i*-th sample. The *N* sample size is practically equal to the ratio of L to 200.

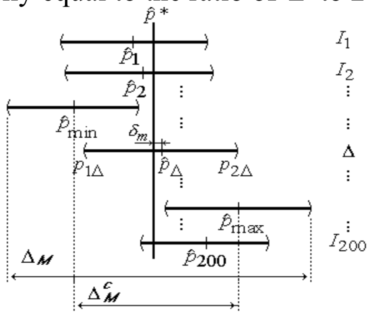

Fig. 1. Entities which point to language event probability

From the 200 $\hat{p}_i$ ($i = 1 \div 200$) estimates, the one nearest to $p*$ was selected, alongside with the respective *i.i.d.* data set; its confidence interval was denoted by Δ. In all the experimental results it was found that Δ confidence interval is practically centred on $p*$, thus it can be computed by means of Eq. (3):

$$\Delta = (p_{1\Delta}\,; p_{2\Delta})\,, \cong p*(1 \mp 1.96 \times \sqrt{(1-p*)/Np*}) \tag{3}$$

By using the procedure advanced by the authors in the previous mentioned studies, see [5] - [7], based on the test on the hypothesis that the probability belongs to an interval described in this Section, this Δ interval was proved to be in agreement with the entire analysed corpus. Therefore the Δ interval was decided to be the *representative* confidence interval for the probability of the investigated event in the analysed corpus. Also, the *i.i.d.* data set which provided Δ confidence interval becomes the *representative* experimental data set further used in the mathematical comparisons.

In conclusion, the *p* true unknown probability of the investigated event lies within the *representative* Δ interval with a confidence level of 95%, see Eq. (1).

Note that the investigated event (*m*-gram) has to fulfil de Moivre-Laplace condition, checked up under the form $N\,p*(1-p*)>20$, where N is the *i.i.d.* data set size.

To give an example for the taken decisions, presented in Table 2 the results of the detailed statistical analysis (see also Fig. 1) referring to the letter structure for the Dumas corpus; these experimental results are obtained using *Form* 2 of the Dumas corpus (illustration only for the first 20 ranks - the most frequent 20 characters). For letter E, on column 1 it is shown the $p*$ value ($p*$=9.39%) of the investigated event. The column 2 shows the number of I_i confidence intervals (out of a total of 200) that contain the $p*$ value (192 intervals for the E letter probability). The 3rd and 4th columns give the borders of the Δ interval computed by use of Eq. (3) and the 5th and 6th are the borders of the (c_1;c_2) extended interval for Δ, computed using Eq. (4). The method on how this extended interval is computed is described in the test on the hypothesis that the probability belongs to a certain interval. On column 7 is the relative error ε_r (0.025) computed by means of Eq. (1) and on columns 8 through 10 we have the β type two error probability (for different values of δ) assigned to the test of the hypothesis that the probability belongs to Δ confidence interval, computed using Eq. (5).

Test of the probability belonging to a certain interval

Be *I=(a;b)* an interval which presumably contains the probability *p* of an investigated linguistic event. The test is based on an experimental *i.i.d* data set of *N* size (of the type from Fig. 1). We are interested if the *i.i.d* experimental data set confirms the hypothesis that the searched *p* probability belongs to the interval *(a;b)* for a chosen *α* significance level; in our case *α*=0.05.

The test procedure

The two hypotheses of the test are

- H_0 : p belongs to the *(a;b)* interval;
- H_1 : p does not belong to the *(a;b)* interval.

We compute the probability estimated value from the experimental data set as $\hat{p}$ *=m/N*, where *m* is the number of occurrences of the investigated event in the data sample. We verify if $\hat{p}$ estimated value is in the *(c_1;c_2)* interval which means the accepted region for this test. The *(c_1;c_2)* interval includes *(a;b)* and is computed using the formula:

$$\int_{c_1}^{c_2}\frac{1}{\sqrt{2\pi a(1-a)/N}}\exp\left(-\frac{(x-a)^2}{2a(1-a)/N}\right)dx=\int_{c_1}^{c_2}\frac{1}{\sqrt{2\pi b(1-b)/N}}\exp\left(-\frac{(x-b)^2}{2b(1-b)/N}\right)dx=1-\alpha \quad (4)$$

In *Eq*. (4) we have two normal probability density functions: the first has the mean equal to *a* and the variance equal to *a(1-a)/N* and the second has the mean equal to *b* and the variance equal to *b(1-b)/N*.

The H$_0$ hypothesis will be accepted if the $\hat{p}$ estimated value is in the *(c$_1$;c$_2$)* interval. If this is not the case, we accept the H$_1$ hypothesis (we reject H$_0$ for the α significance level).

As in any other statistical test there can be two types of errors:

- Type I statistical error: this is the error of rejecting valid data. This is the case when the $\hat{p}$ estimated value does not belong to the *(c$_1$;c$_2$)* interval, even though the true probability belongs to the *(a;b)* interval. This error is lower than α - the chosen significance level.
- Type II statistical error: the error of accepting false data as being valid data. This is the case when $\hat{p}$ belongs to the *(c$_1$;c$_2$)* interval even though the true probability of the linguistic event does not belong to the *(a;b)* interval. For given values for α and *N*, the probability of this type of error depends on the true unknown *p* probability and is computed with *Eq*. (5):

$$\beta(p)=\int_{c_1}^{c_2}\frac{1}{\sqrt{2\pi p(1-p)/N}}\exp\left(-\frac{(x-p)^2}{2p(1-p)/N}\right)dx \qquad (5)$$

$\beta(p)$ takes high values when p is to the left of *a* or to the right of *b* but very close to their values, *i.e.* when $p=(1-\delta)a$ or $p=(1+\delta)b$ where the δ is a small quantity. The experimenter is to decide upon the δ value, depending on the particular constraints of the targeted application.

This test of the hypothesis of the probability belonging to a certain interval was essential in establishing the *representative* items (the representative confidence interval and the *representative i.i.d.* data set assigned to the investigated linguistic event and the analysed corpus). We explain that for letter E in Table 2.

Table 2

Experimental results: 0. Letter; 1. *p relative frequency; 2. Number of confidence intervals that include *p**; 3.4. The Δ confidence interval; 5.6. (c$_1$;c$_2$) extended interval; 7. ε_r - relative error; 8-10. $\beta(p)$ values for δ equal to 0.1, 0.15 and 0.2. *Values from columns 1.,3,.4.,5.,6. are multiplied by 100.***

0	1	2	3	4	5	6	7	8	9	10
_	17.67	196	17.35	17.99	17.09	18.26	0.018	0	0	0
E	9.39	192	9.15	9.64	8.95	9.85	0.025	0	0	0
A	7.95	194	7.73	8.18	7.54	8.37	0.028	0	0	0
I	7.76	195	7.54	7.99	7.35	8.17	0.028	0	0	0

R	5.57	195	5.39	5.77	5.23	5.93	0.034	0	0	0
N	5.47	189	5.28	5.66	5.13	5.83	0.034	0	0	0
U	5.29	198	5.11	5.48	4.95	5.64	0.035	0	0	0
T	4.84	187	4.66	5.02	4.52	5.17	0.037	0	0	0
C	4.30	191	4.13	4.47	3.99	4.62	0.039	0	0	0
S	3.82	196	3.66	3.98	3.53	4.12	0.041	0	0	0
Ă	3.70	193	3.55	3.86	3.42	4.00	0.042	0	0	0
L	3.57	188	3.42	3.73	3.29	3.86	0.043	0	0	0
O	3.27	189	3.13	3.42	3.00	3.55	0.045	0	0	0
D	3.09	191	2.95	3.24	2.83	3.36	0.046	0	0	0
M	2.76	190	2.63	2.91	2.52	3.02	0.049	0.010	0	0
P	2.47	193	2.35	2.61	2.24	2.72	0.052	0.020	0	0
Ş	1.21	185	1.13	1.31	1.05	1.39	0.075	0.185	0.010	0
Î	1.17	189	1.08	1.26	1.01	1.34	0.076	0.198	0.014	0
V	1.10	187	1.01	1.19	0.94	1.26	0.079	0.225	0.025	0
F	0.91	188	0.83	0.99	0.77	1.06	0.087	0.300	0.045	0
Coverage	95.4%	-	-	-	-	-	-	-	-	-

This test was applied 199 times for the case when (a;b) is the Δ interval given in columns 3 and 4, namely (9.15%;9.64%), and the *i.i.d.* sample submitted to the test was successively each of the 199 *i.i.d* data sets. 199 means 200 minus the set that produced the Δ confidence interval. As all the 199 tests were passed we could consider that Δ confidence interval was confirmed by the entire corpus so that we claimed that it is the *representative* confidence interval. The very small values for $\beta(p)$ supports the claim that Δ is the *representative* confidence interval for the letter E probability in the analysed corpus.

Alongside with this statement comes out the *representative i.i.d.* data set from the Dumas corpus: thus, for the E letter, the *representative i.i.d.* set is the one that provided Δ confidence interval.

3. A Mathematical Comparison between Corpora and Experimental Results

The obtaining of the *representative* interval for probability for all the investigated events, with the accuracy provided by the ε_r relative error and the $\beta(p)$ type two error probability, enabled us to make a clear statement about the author models. These author models are compared against each other to see if the differences which appear between them are significant enough to dismiss the idea of a literary field statistical model or not.

The mathematical comparisons use the *representative* elements (the *i.i.d. representative* data sample and the *representative* confidence interval).

All the comparisons are carried out by using two criteria:

(1) *the linguistic entity probability criterion*, which verifies if a certain linguistic entity has the same probability in the two compared corpora; *e.g.* whether the same letter has the same probability in the two compared corpora;
(2) *the rank probability criterion*, which verifies whether the linguistic entities placed on the same rank in the two compared corpora have the same probability; *e.g.* whether the letter placed on the second rank in the first corpus has the same probability as the letter placed on the second rank in the second corpus.

The mathematical comparisons were carried out on the basis of two tests: the test of the equality between two probabilities [10], and the test on the hypothesis if the probability belongs to a certain interval (described in the previous section). The tests were applied considering $\alpha = 0.05$ significance level, *i.e.* the probability of rejecting good data was 0.05. The test was applied on pairs of two *representative i.i.d.* data samples extracted from the five compared corpora for every investigated event.

Note that otherwise, a comparison between the five texts would suppose 200×200 pairs of *i.i.d.* data sets, and therefore it would be difficult to draw a conclusion. In order to surmount this difficulty, only the *representative* above-mentioned experimental data sets were used for each corpus.

Hypothesis Test for Comparing Probabilities

Be there two samples each complying with the *i.i.d.* statistical model, with the sample size N_1 and N_2. Denoting by m_1 the number of occurrences of the event in the first data sample, the probability estimate is $\hat{p}_1 = m_1 / N_1$. Similarly, 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 population meaning $p_1 = p_2$. We apply the test based on the z test value:

$$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{p_1(1-p_1)/N_1 + p_2(1-p_2)/N_2} \quad p_1 = p_2 \cong (m_1 + m_2)/(N_1 + N_2) \tag{6}$$

If $| \; z \; | \leq z_{\alpha/2}$, we shall consider that the two probabilities are equal. Otherwise, we reject the equality hypothesis at an α significance level.

In this paper, we compare in pairs the five corpora for all the three forms mentioned in Section 1. In all these cases we applied the test of equality between two probabilities based on the *representative i.i.d.* set assigned to the investigated event.

The experimental results presented below refer to the letter statistical structure. The Tables 3 to 8 present the results for the test. Table 3 presents the comparison based on the rank probability criterion and Table 4 presents the comparison based on the entity probability criterion, for the *Form* 1 of the text (31 letters, the space character and the hyphen character).

Similar results are computed for *Form 2* and *Form 3* of the text.

In the Table 3 the comparison included all the 33 ranks but only the first 20 were considered as being important in the study. When comparing Dumas corpus with Tolkien corpus all the ranks passed the test. When comparing Dumas corpus with Herbert corpus there are five ranks that failed the test and we present in the assigned parentheses the exact test values. When comparing Dumas corpus with Asimov corpus all the 4 ranks that fail the test are not among the first 20 ranks, so we consider that Dumas and Asimov are in good agreement. Tables 4-6 are similarly organised.

Table 3.

Rank comparison, Form 1 – 33 characters

	Dumas	Tolkien	Herbert	Asimov	Chirita
Dumas		-	1(3.71), 2(3,37), 25(2.55), 29(3.00), 30(3.01)	24(2.52), 27(3.00), 29(3.72), 30(2.98)	14(2.60),
Tolkien			1(3.92), 2(3.87)	2(2.84), 17(2.76),	9(2.69)
Herbert				1(2.62), 11(2.75)	1(3.16), 2(3.45), 9(2.68),
Asimov					9(2.55), 17(2.55),

When the comparison between the five corpora is done using *Form 1* of the text, we can expect to have the lowest compatibility of the three forms. Some clear differences can be seen when comparing the Herbert corpus with the rest because we have clear differences on the first two ranks. The general compatibility between Herbert and the other corpora is worse for *Form 1* of the text than for *Form 3* where we do not consider the blank character (which is on the first rank). However, the compatibility for *Form 1* between the Dumas corpus and the Herbert one is 71.3% (based on the ranks that passed the tests), which is still high enough and does not represent a problem when considering a concatenation of the two corpora.

Table 4.

Letter comparison, *Form* 1 – 33 characters

	Dumas	Tolkien	Herbert	Asimov	Chirita
Dumas		-	_(3.71), Ş(2.87), Ă(3.83), Â(2.78), -(2.55), E(3.37), L(2.66), M(3.85), P(2.60),X(3.00)	H(3.00), Ă(3.34), Â(3.29), X(3.72)	D(3.05),
Tolkien			_(2.92), -(2.88), Ş(4.04), Ă(3.75), Â(3.29), E(3.87)	-(2.77), Ş(2.76), Ă(3.47), Â(3,72), E(2.84),T(2.87)	C(2.69)
Herbert				_(2.62), L(2.75)	_(3.16), Ş(3.58), Ă(2.52), C(2.68), E(3.45),
Asimov					-

In the Table 4 we can acknowledge exactly which are the linguistic entities that failed the test for *Form 1*. Because the blank character is one of the differences between the Herbert corpus and the rest we can state that there is a difference between the average lengths of the words. For instance the *p** for the space character for the Herbert corpus is 17.04% and for Dumas is 17.98%. As a conclusion, we can state that on average the words in the Herbert corpus are longer than in the other four corpora.

Another important difference is the letter E which appears when comparing the Herbert corpus with the Dumas, Tolkien and Chirita corpora. When analyzing why this difference occurred we noticed a series of frequent words that contain the E character which had a much higher relative frequency in the three corpora than in the Herbert one and also the number of total words that contains the E character cover less than 2% in the Herbert corpus than in the Dumas one. These cumulated differences are enough to make the E character fail the test.

When considering for comparison *Form 3* of the corpora, see the Tables 5-6, we obtain a clear idea that sustains the general field statistical model. We still have the lowest compatibility when comparing the Herbert corpus with the other four corpora, but they are close to the 90% range for all of them. *Form 3* clearly gives the best support for a concatenation between independently built corpora.

Table 5.

Rank comparison, *Form* 3 – 31 characters

	Dumas	Tolkien	Herbert	Asimov	Chirita
Dumas		-	1(2.76), 13(2.71), 27(2.95), 28(2.96)	23(2.56), 25(3.05), 26(3.65), 27(2.92),	13(2.58)
Tolkien			1(3.40), 16(2.59)	1(2.67), 16(2.85)	8(2.71)
Herbert				10(2.54)	1(3.98), 8(3.07),
Asimov					-

Table 6.

Letter comparison, *Form* 3 – 31 characters

	Dumas	Tolkien	Herbert	Asimov	Chirita
Dumas		-	Ş(3.08), Ă(4.21), Â(2.93), D(2.71), E(2.76), M(4.18), X(2.95)	H(3.05), Ă(3.45), Â(3.33), X(3.65)	D(3.03),
Tolkien			Ş(4.23), Ă(4.08), Â(3.46), E(3.40)	Ş(2.85),Ă(3.60), Â(3,81), E(2.67), T(2.75)	C(2.71)
Herbert				L(2.54)	Ş(3.75), Ă(2.83), C(3.07), E(2.98)
Asimov					U(2.53)

Test on the Hypothesis that the Probability Belongs to an Interval – Experimental Results

We applied the test described in Section 2. We present only the results for the rank comparison (we compare the characters situated on the same ranks from all the corpora). Because the Dumas corpus was the largest one we used it as reference in all the comparisons for the simple reason that the experimental results extracted from it have the best accuracy. The test is applied in both directions. For example: when we compared the Dumas corpus with the Tolkien corpus, the test was applied using for *(a;b)* the representative confidence interval for the compared rank from Dumas and the *i.i.d* data set was the *representative* data set from Tolkien for the same rank. When we compared Tolkien versus Dumas, (a;b) was the *representative* confidence interval assigned to the considered rank from the Tolkien corpus and the *i.i.d.* set was the *representative* data sample from the Dumas corpus (the first comparison is denoted Dumas-Tolkien and the second is Tolkien-Dumas, see Table 7.

The results in Table 7 are obtained using *Form 2* of the corpora. Also, for every successful result which we marked with ok we displayed below it the $\beta(p)$ type two error probability for $\delta=0.15$, to give an idea about the accuracy of the results. $\beta(p)$ is computed with Eq. (5).

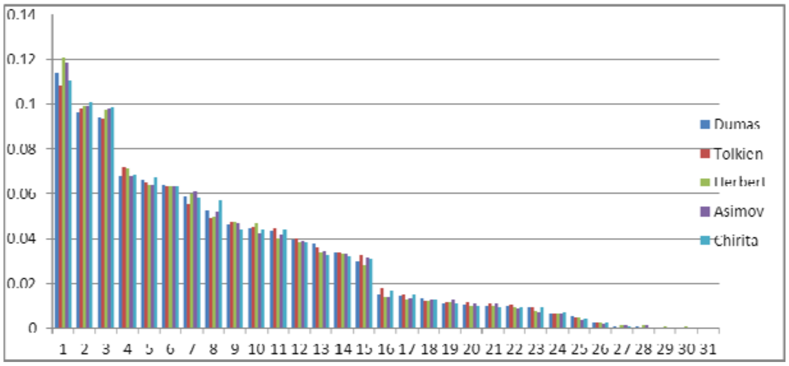

Fig. 2. The letter relative frequencies of the five compared corpora (*Form 3*)

Fig. 2 presents the p^* values for the five corpora for each rank. It can be is easily noticed the 4 zones of character distribution: the zone 1-3 which in all five corpora contains the same 3 characters (E, A, I); the zone 4-7 where for all the five corpora there are the same characters (R, N, U, T); the zone 8-15 where for all the five corpora there are the same characters (C, S, Ă, L, O, D, M, P) and zone 16-31 where, obviously, there are the same remaining characters for all the five corpora brought into comparison.

Table 7.

Results for the test of the hypothesis that the probability belongs to an interval, for *Form* 2, rank comparison for the letter structure

Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Dumas-Herbert	fail	fail	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.08	0.09	0.11	0.17
Herbert-Dumas	fail	fail	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.016	0.025	0.034	0.04
Dumas-Asimov	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.014	0.168	0.184	0.206	0.28
Asimov-Dumas	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.022	0.023	0.03	0.04
Dumas-Tolkien	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
	0	0	0	0	0	0	0	0	0.02	0.03	0.036	0.04	0.06	0.07	0.11	0.154	0.417	0.432	0.457	0.51
Tolkien-Dumas	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.016	0.05	0.05
Dumas-Chirita	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	fail	ok	ok	ok	ok	ok	ok
	0	0	0	0	0	0	0	0	0	0	0.012	0.01	0.02	0.03	0.05	0.071	0.326	0.337	0.364	0.5
Chirita-Dumas	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	fail	ok	ok	ok	ok	ok	ok
	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0.014	0.026	0.044	0.08

This is again a strong argument which sustains the idea of a general statistical field model for printed Romanian.

4. Conclusions

This paper, based on single author corpora comparisons, brings new results for the printed Romanian language. Other attempts in this matter were done before and some results were presented, but they did not serve well enough the main goal because they were obtained from multiple author concatenated corpora [2], [3], [5] and [6]. Those results were obtained from averaging probability estimates so they could not give a clear answer for the debate between single author model versus a general language model, therefore this paper fulfils very well that need.

In our opinion, at least for *Form 3*, the corpora do not show major differences as to discourage the idea of a general field model for printed Romanian. As a final answer for this debate we can state that a general field statistical model for printed Romanian can be considered.

Form 2 of the corpora brings into discussion the average length of the words which can be an element that is influenced by the author's personality [4]. Regarding this aspect, there are some differences, however the compatibility of the author model remains good.

The experimental results are supported by a good accuracy of the language modelling because all corpora had a large enough length thus giving a valid support to this debate when considering the letter structure. In our study, we also considered the digram and trigram structures but the detailed investigation is not presented in this paper. The length of the five corpora was not large enough to support the comparisons results with a satisfying accuracy for digrams and trigrams, even though in almost all the cases the tests used in comparisons were passed and no important differences appeared.

Acknowledgment

The authors acknowledge the continuous encouragement and scientific support from Prof. Dan Tufiş, corresponding member of the Romanian Academy. This work is part of an ongoing research theme at The Research Institute for Artificial Intelligence, Romanian Academy.

The work has been funded by the Sectorial Operational Program Human Resources Development 2007-2013 of the Romanian Ministry of Labor, Family and Social Protection through the Financial Agreements POSDRU/6/1.5/S/19.

REFERENCES

[1] *Adriana Vlad, A. Mitrea, St. Ciuca, A. Luca*, A Study on the Statistical Structure of Words and of Word Digrams in a Literary Romanian Corpus, SPED2011, ISBN 978-1-4577-0440-6, Brasov

[2] *Adriana Vlad, A. Mitrea, M. Mitrea, Şt. Ciucă*, Enriching Printed Romanian Statistical Description: an Approach by Mathematically Comparing Two Independent Literary Corpora, Dan Tufiş, Corina Forăscu (eds.) (2010), Multilinguality and Interoperability in Language Processing with Emphasis on Romanian, Editura Academiei, 2010, pp. 245-271

[3] *S.Ciuca, A.Vlad, A.*: *Mitrea,*A Comparison Between Two Literary Printed Romanian Corpora Based on the Statistical Letter Structure with Orthography and Punctuation Marks, Proc. of the IEEE Intl. Conf. Communications'2010 Bucharest, pp. 119-122

[4] *P. Grzybeck, E. Kelih, E. Stadlober*, The relationship between word length and sentence length: an intra-systemic perspective in the core data structure, Glotometrics 16, 2008, pp. 111-121

[5] *Adriana Vlad, A.Mitrea, M. Mitrea*, Printed Romanian Modelling: A Corpus Linguistic Based Study With Orthography And Punctuation Marks Included, Lecture Notes in Computer Science, vol. 4705 (ICCSA 2007), Springer Verlag, Berlin Heidelberg, 2007, pp. 409-423, ISSN 0302-9743.

[6] *A.Vlad, A.Mitrea, M.Mitrea,*: Limba română scrisă ca sursă de informație (Printed Romanian Language as an Information Source, textbook in Romanian). Ed. Paideia, Bucharest (2003).

[7] *A.Vlad, A.Mitrea, M.Mitrea,* A Corpus – based Analysis of how Accurately Printed Romanian Obeys Some Universal Laws. Wilson A., Rayson P., McEnery T. (eds): A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, Chapter 15, Lincom-Europa Publishing House, Munich (2003) 153–165.

[8] *B.Say, A.Akman,* Current Approaches to Punctuation in Computational Linguistics. Computer and the Humanities, Vol. 30 (1997) 457–469

[9] *C.E.Shannon,* Prediction and Entropy of Printed English. Bell Syst. Tech. J., vol. 30 (1951) 50–64.

[10] *R.E.Walpole, R.H.Myers,* Probability and Statistics for Engineers and Scientists. 4th edn., MacMillan Publishing Comp., New York (1989).