UDC 004.62:004.912

A.O. CHEPOK
Odessa National Academy of Telecommunications n. a. A. Popov, Ukraine
N.I. YEVTUSHENKO
Secondary School No 49 of Odessa City, Ukraine

THE STYLISTIC CONTENT-ANALYZER OF TEXTS: PART A – VALIDATION OF PROPOSED ALGORITHMS

Abstract – For the automation of reliable analysis of content of digitized texts and determining their styles the necessary algorithms have been developed, and based on it was created the program «AN-TExp, v.1», which can be used to implement the raised tasks and to check applicability of basic algorithms. It was found the series of singularities for statistically significant (α = 0,05) style determination of the analyzed text. The average values of singularity factors K_{singul} are shown. Validity of the given methods for the proposed series of singularities is shown.

Keywords: text data content analysis, computer content analysis, text data analysis, quantitative content analysis.

А.О. ЧЕПОК
Одесская национальная академия связи им. А.С. Попова
Н.И. ЕВТУШЕНКО
Одесский УВК №49

СТИЛИСТИЧЕСКИЙ КОНТЕНТ-АНАЛИЗАТОР ТЕКСТОВ: ЧАСТЬ А – АПРОБАЦИЯ ПРЕДЛОЖЕННЫХ АЛГОРИТМОВ

Для автоматизации проведения достоверного контент-анализа оцифрованного текста и определения его стилистики были разработаны необходимые для этого алгоритмы и на их основе была создана программа «AN-TExp, v.1», с помощью которой можно реализовать поставленные задачи и проверить правильность базовых алгоритмов. Найдены совокупности сигнатур для статистически достоверного (α = 0,05) определения стилистики анализируемого текста. Приведены значения средних для факторов сингулярности этих сигнатур. Показана валидность приведенной методики для предложенных серий сигнатур.

Ключевые слова: контент-анализ текста, компьютерный контент-анализ, количественный контент-анализ.

Introduction

For the last two decades of development of information technology (IT) one can see the ever-increasing volumes of stored and transmitted information by countless channels of communications [1]. In the "space" of Internet anyone can find almost any information, and moreover, in any form of. However, due to the history of mankind, most information is stored and is provided in the form of various texts. Still "the printed word" for most of the world population is the most common method of supplying and perception of information.

Naturally, in conditions of avalanche-like accumulation of text information one can see rapidly increasing demand for means of automatic classification of unstructured and structured textual data bases (DB). Of course, such programs are not free.

But Human consciousness is not changed as fast as scientific and technological progress: even at presence of the most advanced technologies the known "old" human challenges, including ones such as plagiarism and generation of meaningless texts (for example, collections of words, which are devoid of meaning) [2, 3].

School Literature teachers in response to questions about plagiarism can definitely tell a lot about how often they have been facing in the last 15-20 years with so-called "works" of students who were "simply borrowed from the Internet". It needs no to say, this is a real problem, moreover, it is dual one because of plagiarism which is not only destroys the individuality of a person, and society as a whole.

Of course, the problem of plagiarism is not new [4]. But for this human "illness" it appears a kind of modern "antidote" – computer programs that can analyze texts [2, 3, 5, 6]. And in this respect it is urgent to develop specialized software, in particular an unstructured text analyzer, which would be able to carry out the desired classification of text content which is analyzed, according to its certain characteristics. Search of this type of analyzer amongst the software accessible to authors did not give positive results.

Since the vast majority of texts being created and processed in common text editors, then the authors when working on the creation of their program named "AN-TExp" (an analyzer of content of unstructured texts) have implemented the idea of computer aided statistical and linguistic analysis of given texts using MS Word and MS Excel software capabilities.

When performing of this work such problems were solved:

1. determination of mathematical or statistical characteristics of given texts that can give objective information on their features;

2. development of specific algorithms to determine the stylistic features of given texts that are based on verification of regular statistical hypotheses;

3. development of specific algorithms and methods for analyzing of given texts in order to establish the presence in them of predefined words or phrases and further "reaction" of the program on the words;

4. organization of modular structure for the mentioned program and ensuring interactions between its modules, and ensuring interoperability MS Word and MS Excel via the code during the analysis of the given texts;

5. development of a clear and understandable user interface (UserForm);

6. to test this work authors' assumptions of the ability to have reliable and reasoned judgments about the literary skill of the author of the text being analyzed on the basis of the mathematical (i.e. valid) data which are results of the processing of the given text using the program.

The objective of the work is to test certain algorithms that concerns to analyzing of any texts and development of the author's program – content analyzer of unstructured digitized (i.e. machine-readable) texts provided by the *.txt, *.doc or *.docx files.

The object of study of the work was the process of automated statistical analysis of digitized written texts and determining of some key characteristics and some additional ones of given texts.

The subject of the study of the work was algorithm of textual analysis that implements the technique of computer aided textual content analysis, and further interpretation of the achieved results.

When compiling the mentioned program the following working **hypothesis** were formulated:

1. It is known that the variety or richness of a language for any personality is determined by how many language units (words and phrases) is there in his own vocabulary for use [7, 8]: the richer the language means it may contain more information and more personal opinions, as well as certain attitudes of the author to the subject matter of speech [8]. From here the authors make their assumption that the informational entropy[3] of the analyzed text is clearly linked to the lexical diversity of the person which says or writes [9].

2. There is an opinion that any author eventually makes his certain "author's style" for his own writings that distinguishes his texts from other ones, and the author's texts have individual manner of writing. Surely, the mentioned "author's style" has its own features – peculiarities of the writing manner or so called singularities [10]. From here the authors of this article make their assumption that one can identify these textual peculiarities and use them to further interpretation.

As a result of solving the above problems the program "AN-TExp" was created: it is the Content Analyzer of unstructured texts – namely various texts of general use and literary works.

1. Research technique of the text to be analyzed

It is no secret that some subjective factors can influence the results of any examination and on assessments by expert committees on different issues, for example, concerning to the well-known problems such as plagiarism [4].

It is well known fact that Mathematics is able to remove subjectivity from any problem or question. Therefore, to ensure a really scientific peer review it usually used general and special mathematical procedures in order to get certain key characteristics and some additional characteristics of the text that is under analyze. From the viewpoint of Computer Science, Applied Linguistics and its newest branch – Computational Linguistics – several mathematical values may be considered as such characteristics (*singularities*) [6, 10]. The authors of the work decided that for the analysis of the text one should be chosen exactly these values: entropy of a text H, mathematical expectation $< X >$ and standard deviation σ were selected as the main statistical variables. At the program "AN-TExp" they labeled as **"main Stat. Data**" and appropriately assigned on the UserForm.

In addition to these "main" statistical indexes it was decided to choose some additional values (on the form marked as **"Relat. Frequency of Signes**"), which are based on counting the frequency of certain characteristics of the text. The mentioned values, according to the authors of the work, will help to define the creative style of the author of a text, which is analyzed by the program. Together, these indicators help to form the most objective characteristics of the creative style of the author of the text, which is to be investigated.

To find these values the program "AN-TExp" carries out dual-stage decomposition of the text to be analyzed: at first all of the text is decomposed into separate words, and then these words – into characters, that is, into "their proper" parts. After this procedure the program performs frequency-and-statistical analysis of these characters as already separated elements of the text.

Since entropy is a measure of chaos in any system, then, in terms of computer science, the entropy of the given text can be a measure of its lexical and morphemic diversity: that is a high-quality literary text is able to hold more entropy compared with the entropy of remarks of any "savage" person. So, may be it is the cause that the entropy of works of true masters of literature seems to be exceed the entropy of the well-known speeches of famous "Ellochka-the-Ogress".

The mathematical model of the "AN-TExp" program is based on well-known mathematical and statistical formulas and functions. To calculate the entropy of a text the famous C. Shannon's formula is commonly used [9]: $H = -\sum_i p_i \log_2 p_i$, here p_i – the probability to find i-th character in the text which is being studied.

The mathematical expectation $< X >$ is calculated by the formula $< X >= \sum_i x_i p_i$, when x_i – number of certain characters which are used in the text, and p_i – the probability of their use. In this context, the figure $< X >$ has the meaning of "the cumulative utilization rate of certain characters in the text", and it is one of the unprejudiced "*sensitive-to-person*" characteristics of the text to be analyzed.

[3] **entropy** (*informational*) – it is the logarithmic measure of unpredictability of information content.

Since a mathematical expectation in the statistics is usually accompanied by its standard deviation σ, it is a standard deviation is also computed as an additional parameter: $\sigma = \sqrt{\sum_i x_i^2 p_i - < X >^2}$. A couple of these values is usually written as follows: $< X > \pm \sigma$, that clarifies certain features of the author's style of the person who made the text to be investigated.

Thus it can be considered that the set of mentioned mathematical characteristics is quite informative and objective evidence of the author's style of the person who created the given text. However, it is rather true that this set of characteristics is clearly not to be sufficient to complete the peer review regarding authorship of a text, and this question needs further clarification. Rather, the source of such looks may be some keyboard characters that can be considered as characteristic for the written work of a particular person, such as fingerprints of somebody. Thus, the frequency coefficients of specific use of certain symbols can also be considered as important identifiers in determining the individual author's style.

2. Software implementation of selected methods of research

Since for many years VBA is the standard for controlling over MS Office applications, naturally, it was decided to implement the proposed method of analysis of texts in this programming medium.

When creating the program – content analyzer of a text – it was implemented the idea to use potentials of both MS Word and MS Excel software and their interaction during data inside transfer. The MS Word text editor during preliminary preparation of given text in order to its further analysis will bring it to the standard form (for example, the Font size, the line spacing, etc. – it depends on User preferences).

By itself, the spreadsheet processor provides a wide range of working with tabular data starting from simple calculations for standard functions up to visualization of the obtained results. Thus, one can use the possibility of the "standard" software of highest quality to solve the specific problem by means of a new program, which was conceived and designed like a modular junction, that can function on the principle of capacity building. Thus, the possibility of using the highest quality software to solve the original problem of a specific new program, which was conceived and designed like modular design, functioning on the principle of capacity building.

As a result it was created the multi-modular program "AN-TExp", which enables you to find automatically a desirable range of mathematical parameters that are the basis for objective analysis of the given text. So, such problems have been resolved:

3. obtaining of the series of acceptable objective characteristics (i.e. mathematical or statistical ones) of a given text file using certain processing functions;

4. organization of the modular structure of the program that is created, and coordination between its modules, as well as providing of interoperability between MS Word and MS Excel via the program code when analyzing of the text;

5. it was made a successful attempt to develop some algorithms to analyze text to determine authorship. Further development of these algorithms require the work with a rather large base of texts and as well consultations with experts in linguistics.

6. still forming the developing of some algorithms and methods to analyze texts in case of presence of specific, pre-defined, words or phrases, and "response" of the program for such content;

7. it was organized the convenient user's interface (UserForm), which requires an user's understanding of the obtained results and his ability to interpret them. But the version 2 of the program is designed for trained users (up to for specialists in linguistics), so the interpretation of the obtained data for them is not a problem and it does not require additional knowledge and skills from this category of users.

The principle of capacity building for the "AN-TExp" program as it commonly used for modular-based units will bring it to the level of high-quality programs that one can use to automatically identify the individual author's style of the person who has written this text, and/or it will allow to set the reaction of the program to a specific phrase in the given text on occasion to determine the degree of "common sense" in the text which is analyzed (or so-called "white noise level" in given text).

3. On functional features of the "AN-TExp, v.1" program

The program «AN-Texp, v.1» is able to analyze a given text content accordingly to the logic circuit provided in Fig. 1.

Before starting the program, the user must place the text that is being analyzed, into the "clearly defined" Word-file (i.e. into *.doc- or *.docx-file with a certain name and in specific location on the selected bearer[4]). Then the user has to start the basic program module which is placed in the Excel-master file that will display the UserForm that is aimed to provide the step-by-step works of the analyzer in dialogue mode.

Next, the form offers to register the text to be analyzed by the program. In order to form the "List of Records" (registration list of works) the registration procedure for a text that will be analyzed occurs in two stages: firstly input its author's name, then – input the title of the text.

[4] this is true for the version №1 of the program "AN-TExp": "AN-TExp, v.1".

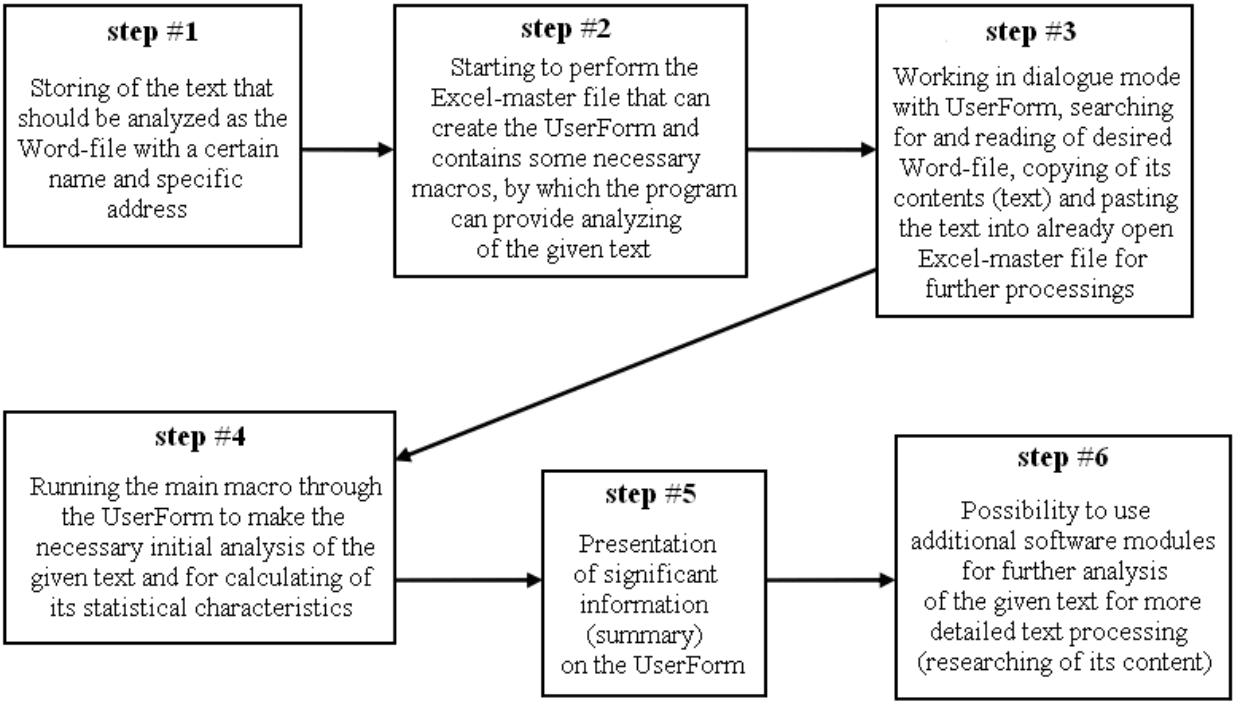

Fig. 1: The sequence of basic operations which are performed by the program "AN-TExp, v.1" accordingly to the algorithms that are mentioned above

The program "watches" for user's errors, e.g. if not made any records about the investigational text and/or its author, the appropriate message occurs as a reminder. After successful completion of the registration procedure of the specified text the next phase of mathematical analysis of the text comes.

The program "AN-TExp, v.1" performs frequency analysis of all involved keyboard characters in the text to be analyzed, and the obtained data will enable to calculate the entropy of the text and other mentioned above characteristics of the text. Thus the obtained information is *an objective mathematical description of the analyzed text* and contains the most detailed information about the studied object.

After following the instructions of the basic module of the program the obtained research results will be available for users. A user is able to govern the completeness of output data onto Userform by using of switch controls.

Some modern humanities, for example, applied linguistics and psychology of personality, believe that in this way you can with high probability to set the individual author's style. So after researching a sufficiently large amount of text – it can be a few fragments[5] of texts, "which came from the pen" of the same person – one can collect the textual personal database (PDB) as a kind of "*dossier*", and the appropriate results of mathematical processing of primary information. At least, one can refine the file depending on the circumstances, and based on the obtained information one can identify the author of the given text with some degree of probability. The basis of this procedure is at least convergence of the main statistical data and other singularities taken from the existing PDB of certain person with the same data that can be obtained by examining any given text. But in order that a decision on authorship of the analyzed text was more significant, one has to provide proper validation for appropriate statistical hypotheses about relevant significance of the differences between the given samples, as well as to provide the variance and correlation analysis to identify relationships between the sample data.

4. Results and discussion

With the mentioned above program it has been analyzed 23 examples of the text of different style, amongst which there were the real literary works (9 pcs.), essays and reviews of columnists and analysts from the serious electronic mass media of Ukraine (6 pcs.), graduate works of scientific and technical direction (4 pcs.), different administrative instructions or directives (4 pcs.). The list of these texts is shown in the Table #1.

Here, in the Table #1, the texts taken for analysis have been collected in the respective categories – "**L**iterature", "**R**eviews & **A**nalytics", "**D**iploma thesis" and "**B**ureaucratic writing". The volume of each of the texts was not less than 65 KB.

The main results of the calculations by described algorithms presented in the Table 2. The last line of obtained data in the Table shows the relative errors ε_t calculated according to a standard procedure for small

[5] as the authors' experience shows, the given text that contains about 10,000 words or more may be considered as the reliable text fragment.

samples: $\varepsilon_t = \frac{\Delta_t}{\bar{x}} = \frac{t_{\alpha,n} \cdot s}{\bar{x}}$ (α=0.05) for each of these given quantities. The data from the line show clearly that all the values presented here, except for entropy, have enough variability, i.e., in principle, it may contain information about the features of the text contents taken into consideration.

Table 1

The list of texts which have been taken for the content analysis using the program «AN-Texp, v.1»

text #	the text names	the text Author	year of origin	category of the text
1	«Нос»	Гоголь Н.В.	1834	(masterpiece of) Literature
2	«Портрет»	Гоголь Н.В.	1834	(masterpiece of) Literature
3	«Коляска»	Гоголь Н.В.	1835	(masterpiece of) Literature
4	«Шинель»	Гоголь Н.В.	1841	(masterpiece of) Literature
5	«Дубровский», (1-5 гл.)	Пушкин А.С.	1841	(masterpiece of) Literature
6	«Антоновские яблоки»	Бунин И.А.	1900	(masterpiece of) Literature
7	«Одесские рассказы»	Бабель И.Э.	1931	(masterpiece of) Literature
8	«Журавлиный крик»	Быков Василь	1959	(modern) Literature
9	«Утро вечера мудренее»	Быков Василь	1967	(modern) Literature
10	«Диссергейт»	Суржик Лидия	2016	Reviews & Analytics
11	«Налоговики и коррупция»	Верстюк Иван	2015	Reviews & Analytics
12	«Борец с франкоманией»	Шама Олег	2015	Reviews & Analytics
13	«Урок экономики от Швейцарии»	Перегонцев Илья, Половец Ирина	2016	Reviews & Analytics
14	«Украина в Первой мировой: между Сциллой и Харибдой»	Гусев Виктор	2016	Reviews & Analytics
15	«Талантливые дети Украины: будут ли оправданы надежды?»	Молодцова Елена	2016	Reviews & Analytics
16	«Программная визуализация работы ЦАП»	Гордиенко Алексей	2015	Diploma thesis
17	«Анализ и оценка типовых топологий вычислительных сетей»	Берш Евгений	2008	Diploma thesis
18	«Земля как планета Солнечной системы»	Паничева Евгения	2007	Diploma thesis
19	«Зонная концепция систем молниезащиты зданий»	Головач Татьяна	2015	Diploma thesis
20	Приказ: «О порядке работы секции «Информатика»	---	2016	Bureaucratic writing
21	Приказ: «Об утверждении порядка проведения олимпиад школьников	---	2015	Bureaucratic writing
22	Приказ: «О формировании платежной ведомости»	---	2014	Bureaucratic writing
23	Приказ: «Об учете кадров предприятий системы образования и науки»	---	2014	Bureaucratic writing

On the main statistical indicators.

The data in Table 2 shows that the values of Entropy H_i (i=1…23) for all treated 23 examples of different texts are about the same: $H = 4.627 \pm 0.062$.

Variation of information entropy H for the studied texts of different styles is small: the relative error of entropy ε_H for this sample is not more than 1.5% ($\varepsilon_H \approx 1.35\%$): with such a small variation of information entropy H can hardly be considered as an indicator of the style of the given text.

As for the other presented values – such as mathematical expectation $< X >$ and factors $\boldsymbol{F}_{\text{singul},i}$ of singularity, which can carry some information about the frequency of involved keyboard characters, the variation of these parameters is significant and it depends on the style of the given text. And the values with significant variations can contain certain data which can be interpreted as indicators that may be useful in automation of sorting of analyzed texts by their style.

Consider the data from Table 2 as the sampled data from a variety of general populations. To answer the question about the possibility of presented algorithms "to feel" characteristics of different text styles, it was decided to divide all the processed texts into 2 parts, then compare the statistical characteristics of the samples, and then carry out the two-sample tests for the mean values with different variances (dispersions), which are commonly used

to test the hypothesis of equality of means of two random variables (paired two-sample T-test for the mean values assuming unequal variances, two sample Z-test for mean values), and as well as two-sample F-test for the variances.

First, the all text database was divided into “Literature” (i.e. real literary works, see Tab. 1, positions №1-№9) and “Non-Literature” (see Tab. 1, positions №10-№23). In order to identify (detecting) significance of differences for these two categories was performed standard analysis of the data using of Student's T-test, two-sample F-test for the variances, two-sample Z-test for mean values.

Table 2

The main results of the calculations of basic statistical characteristics and singularities for the texts by different authors (accordingly to results of running of the program «AN-Texp, v.1»)

# текста	энтропия H_i, bit	$< X >$	F_{singul} ” ! ”	F_{singul} ” ? ”	F_{singul} ” : ”	F_{singul} ” ; ”	F_{singul} ” … ”	F_{singul} “ (&) ”	F_{singul} ” , ”	F_{singul} ” – ”
1	4,651	2115	0,0149	0,0072	0,0106	0,0095	0,0060	0,0018	0,1258	0,0393
2	4,601	4962	0,0047	0,0037	0,0060	0,0100	0,0022	0,0004	0,1325	0,0248
3	4,671	999	0,0037	0,0084	0,0052	0,0008	0,0062	0,0001	0,1272	0,0420
4	4,593	2883	0,0030	0,0032	0,0084	0,0080	0,0022	0,0006	0,1465	0,0252
5	4,595	2157	0,0034	0,0027	0,0040	0,0067	0,0038	0,0011	0,1127	0,0234
6	4,655	1211	0,0078	0,0021	0,0071	0,0021	0,0108	0,0005	0,1402	0,0415
7	4,662	1176	0,0016	0,0096	0,0036	0,0007	0,0060	0,0000	0,1213	0,0442
8	4,611	7546	0,0053	0,0070	0,0060	0,0007	0,0023	0,0002	0,1432	0,0412
9	4,628	1698	0,0081	0,0093	0,0036	0,0010	0,0026	0,0007	0,1326	0,0439
10	4,599	861	0,0026	0,0061	0,0105	0,0026	0,0035	0,0257	0,0963	0,0279
11	4,579	842	0,0000	0,0000	0,0089	0,0000	0,0000	0,0042	0,0839	0,0228
12	4,594	710	0,0010	0,0000	0,0077	0,0005	0,0000	0,0000	0,0820	0,0197
13	4,398	540	0,0000	0,0055	0,0033	0,0000	0,0000	0,0044	0,1139	0,0175
14	4,632	1509	0,0000	0,0007	0,0012	0,0002	0,0012	0,0048	0,0853	0,0198
15	4,639	670	0,0099	0,0018	0,0128	0,0041	0,0035	0,0367	0,0624	0,0350
16	4,646	975	0,0007	0,0025	0,0047	0,0004	0,0000	0,0363	0,0799	0,0202
17	4,708	1080	0,0000	0,0002	0,0001	0,0000	0,0000	0,0002	0,0030	0,0006
18	4,731	974	0,0000	0,0002	0,0001	0,0000	0,0000	0,0002	0,0027	0,0005
19	4,726	1003	0,0000	0,0002	0,0001	0,0000	0,0002	0,0002	0,0020	0,0004
20	4,446	550	0,0000	0,0000	0,0041	0,0082	0,0000	0,0428	0,0693	0,0224
21	4,898	721	0,0000	0,0007	0,0003	0,0000	0,0000	0,0007	0,0087	0,0017
22	4,310	410	0,0000	0,0017	0,0009	0,0000	0,0000	0,0017	0,0226	0,0043
23	4,838	704	0,0000	0,0007	0,0003	0,0000	0,0000	0,0007	0,0090	0,0017
ε_t	1,35%	52,06%	68,92%	50,93%	40,07%	71,92%	65,09%	95,72%	31,08%	34,11%

Note 1: The values of columns of «F_{singul} "....."» (*singularity factors* F_{singul}) reflect specific numeric ratios for given singularities of written characters by choice of authors.

So, after all the three two-sample tests for the categories of "Literature" vs "Non-Literature", as well as for the other “paired” categories correspondently, and the following results were obtained (see Table 3):

Table 3

The obtained results of statistical hypotheses testing on equality of mean values:

H	< X >	” ! ”	” ? ”	” : ”	” ; ”	” … ”	“ (&) ”	” , ”	” – ”
the categories of "Literature" vs *"Non-Literature"*									
0	diff. rel.	diff. rel.	diff. rel.	0	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.
the categories of "Literature" vs *"Reviews & Analytics"*									
0	diff. rel.	0	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.
the categories of "Bureaucracy" & "Diploma" vs *"Reviews & Analytics"*									
diff. rel.	0	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.	diff. rel.
the categories of "Bureaucracy" vs *"Diploma thesis"*									
0	diff. rel.	0	0	0	0	0	0	0	0

Note 2: here – "0" means “*no differences*” between the mean values; “diff. rel.” means “*differences reliable*”.

Table 4

The calculated reliable mean values for mathematical expectation $< X >$ and singularities

The text styles	$< X >$	F_{singul} ” ! ”	F_{singul} ” ? ”	F_{singul} ” : ”	F_{singul} ” ; ”	F_{singul} ” … ”	F_{singul} “ (&) ”	F_{singul} ” , ”	F_{singul} ” – ”
"Literature"	2750	0,0058	0,0059	0,0060	0,0044	0,0047	0,0006	0,1313	0,0362
"Non-Literature"	825	0,0010	0,0014	0,0039	0,0011	0,0006	0,0113	0,0515	0,0139
"Reviews & Analytics"	855	0,0022	0,0023	0,0074	0,0012	0,0014	0,0126	0,0873	0,0238
"Diploma thesis"	1008	0,0002	0,0008	0,0012	0,0001	0,0000	0,0092	0,0219	0,0054
"Bureaucratic writing"	596	0,0000	0,0008	0,0014	0,0020	0,0000	0,0115	0,0274	0,0075

It means that almost all of these singularities in the table (except the colon ":" !) are reliable indicators at the program separation of the text content on the basis of "Literature" vs "Non-Literature".

The following categories of the text content, which were subjected to the mentioned two-sample tests, were texts of categories "Literature" and "Reviews & Analytics".

Summary

The program-analyzer «AN-TExp» conceived and made as a modular constructor, which operates on the principle of capacity building, and certainly such structure of the program may be regarded as its advantage.

Using the program «AN-Texp, v.1», some works of famous writers of XIX and XX centuries, as well as other texts of different styles were analyzed accordingly to the described above methodology (see Table 1). Thus, it is formed to the present the database of the texts which allows to confirm or to remove some hypotheses of the authors of the work.

The assumption of the authors of that information entropy of the text content is uniquely associated with a variety of a person's vocabulary has not found explicit confirmation: this item in the research methodology requires considerable revision.

The authors of the present work believe that based on the totality of number of objective characteristics of the given texts (e.g. the calculated value of mathematical expectation $< X >$ and some singularity factors $F_{\text{singul},i}$ of the texts) one can automate analyzing procedure of text contents in order to separate all the entries of the available textual DB according to style of their text contents.

The authors believe that it was calculated the reliable mean values for mathematical expectation $< X >$ and singularities (see Table 4), which will be useful for computer aided analysis of a written texts.

The proposed algorithms of text contents analysis that were implemented by the program «AN-Texp, v.1», allow to use it for successful applying for both professionals and for interested users:

8. when one has to provide automatic evaluation of vocabulary diversity of certain person based on his written works;

9. when one has to provide automatic analysis of given posts on social networks and SMS according to certain criteria;

10. when one has to provide automatic style definition of given texts (e.g., identification of modern and ancient artifacts, etc.).

The value of the work done is that they obtained the multi-module program – the text analyzer, which uses original algorithms and techniques capable to calculate the necessary statistical characteristics of the given texts in order to form qualified judgments about the writer's style of the studied texts with a high degree of reliability.

The authors believe that the used algorithms are usually enough to automate the "sorting" the given texts that being analyzed according to the certain or general criteria. Further study of existing methods of linguistic analysis of text contents and creation of suitable mathematical models will complement the basic version of the program with new modules and it may expand its functionality.

In developing this program, one can compile it as a cross-platformed App. In addition, we can offer users some versions of the program in different languages.

References

1. UNPAN : Rapid Development of Information Technology in the 20th Century. E-Government – What a Government Leader Should Know // https://publicadministration.un.org/published/courses/1343/Course2451/v2010_11_2_16_5_13/media/content/Part1_68335.pdf

2. A.S. Romanov, R.V. Mescherjakov, Z.I. Rezanova, «Metodika proverki odnorodnosti teksta i vyjavlenija plagiata» / Doklady TUSUR, № 2 (32), june 2014, serija «Upravlenie, vychislitel'naja tehnika i informatika», 264-269 p.

3. Bubnov V.A., Anufriev S.V., Kazakova I.S. Analiz poeticheskih tekstov na urokah Literatury s pomoschiju informatsionnyh tehnologij // Informatsionnye tehnologii v predmetnoj oblasti. – M.: MGPU, 2002.

4. Mazur V. Plagiat ili vorovstvo», – «Zerkalo nedeli. Ukraina» №47, 2015 // http://gazeta.zn.ua/science/plagiat-ili-vorovstvo-_.html

5. Gal'perin I.R. Text kak ob'ekt lingvisticheskogo issledovanija, Izd. 4-e. M: KomKniga, 2006. – 144 p.

6. Vershik A.M. Informatsija, entropija, dinamika / «Markov Processes and related fields» (2010), p. 47-76.

7. Golovin B.N. Jazyk i stilistika, M.: «Prosveschenie», 1970. – 190 p.

8. Golovin B.N. Osnovy kul'tury rechi; Uchebnik. – 2-e izd. – M.: Vysshaja shkola, 1988. – 320 p. – ISBN 5-06-001165-8.

9. Shannon, Claude E. "A Mathematical Theory of Communication" // Reprinted with corrections from The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

10. Brilluin L., «Termodinamika, Statistika i Informacija» / UFN, vol. LXXVII, №2, 1962, 337-352 p.

Рецензія/Peer review : 25.9.2016 p. Надрукована/Printed : 8.11.2016 p.
Стаття рецензована редакційною колегією