UDC 004.62:004.912

A.O. CHEPOK
Odessa National Academy of Telecommunications n. a. A. Popov
L.V. KALYUZHNYI
Secondary School No 49 of Odessa City

THE STYLISTIC CONTENT-ANALYZER OF TEXTS: PART B – TOWARDS THE AUTOMATIC TEXT STYLE RECOGNAZING

Abstract – We inform about the program «AN-TExp, v.2», which is aimed to determine the textual style of given digitized texts at 95% confidence. The program uses the well-known testing procedures for the proper statistic hypothesis. It was found the series of singularities for statistically significant ($\alpha = 0,05$) style determination of the analyzed text. The average values of singularity factors K_{singul} are shown. This program can be considered as the powerful part of automated classifying/routing systems.

This work is a continuation of earlier studies of computer-aided statistical-and-linguistic analysis of electronic text. The authors believe that the used algorithms are usually enough to automate the "sorting" of 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.

As for the presented values – such as mathematical expectation $< X >$ and factors Fsingul,i of each shown 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.

Keywords: automatic text style estimation, computer-aided text content analysis, quantitative content analysis.

А.О. ЧЕПОК
Одесская национальная академия связи им. А.С. Попова
Л.В. КАЛЮЖНЫЙ
Одесский УВК №49, Украина

СТИЛИСТИЧЕСКИЙ КОНТЕНТ-АНАЛИЗАТОР ТЕКСТОВ: ЧАСТЬ Б – АВТОМАТИЗАЦИЯ РАСПОЗНАВАНИЯ СТИЛЯ ТЕКСТА

Аннотация. Создана программа «AN-TExp, v.2», которая нацелена на определение стиля оцифрованного текста. Программа автоматически определяет стиль исследуемого текста с достоверностью 95%. В программе используются хорошо известные процедуры тестирования определенных статистических гипотез. Найдены совокупности сигнатур для статистически достоверного ($\alpha = 0,05$) определения стилистики анализируемого текста. Приведены значения средних для факторов сингулярности этих сигнатур. Данную программу можно рассматривать как ядро соответствующей автоматической классифицирующей системы для анализа текстового контента.

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]. Anyone can find almost any information in the "space" of Internet, 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.

Computer-assisted textual analysis, in general, has a long and rich history, and for the last two decades the procedures have been widely adopted in contemporary textual content studies. So, any computer which is equipped with the certain program can perform proper statistical analysis of a given electronic text. This makes such computer-aided analysis of the text content especially appropriate and effective for investigating textual differences and similarities as well as textual singularities in order to clear the more detailed structure of the given texts.

Since really huge amount of different kinds of information are born every hour, and it must be stored somewhere and somehow being sorted at first, we can say that the specific task of sorting of the given digitized texts by their style[3] is very actual and "painful" task.

It is rather an applied linguistic task, and it often was solved in this way [2, 3]. For example, the work [4] is concerned with overt devices of textual cohesion in regard to their capacity for acting as stylistic markers. But this is a "pure philological" approach to this problem, and only such the linguistic analysis of the textual links between the given text is still difficult to call automated (computer aided).

Some authors use other varieties of linguistic analysis: e.g. they apply morphological analysis and count the

[3] the common ***textual styles*** such as: Scientific, Official-business, Journalistic and Artistic ones.

frequency of morphemes within the text data [5].

Some successful works [6-9] were carried out with the involvement of statistics and special sections of mathematics, and this approach became more positive. The work [6] is devoted to the lexical diversity and readability values of textual contents of some sorts of documents. In [7] it was presented a system that evaluates the content of essays based on Latent Semantic Analysis (LSA). The system applies LSA to compare the conceptual similarity between the essays and selected text passages from the course material covering the essay assignment-specific subject matter. The work [8] tells about Automatic Essay Assessor (AEA), that is a system that utilizes information retrieval techniques such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA), but it is also aimed to compare the similarities of different documents. Unfortunately, the significant problem of automatic (*unmanned* !) judgment was not solved in the mentioned studies.

The authors of the current work decided to create the program "AN-TExp, v.2" that intended to automate determination of the style of the given electronic text using the standard statistical analysis and verification of several statistical hypotheses about the equality of two samples. The program is based on the algorithm already described [10] with hybrid principles – linguistic as well statistical ones: so it is rather implementation of the idea to combine these both trends of computer aided analyses of texts. The mentioned basic algorithm for the program has been underwent several enhancements for a greater degree of automation, plus some additional singularities were introduced.

1. The research Techniques applied for the Text to be analyzed

It is no secret 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: the mathematical expectation $< X >$ and the standard deviation σ were selected as the main statistical variables. In addition to these "main" statistical indexes it was decided to choose some additional values, 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.

The mathematical expectation $< X >$ is calculated by the well known 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.

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 style of 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. About the Improvements that formed the basis of the new version of the Program

The authors of the work retained the basis of the concept of the first version of the previous version of our program "AN-TExp, v.1" – it produces double decomposition of the given text content using MS Word and MS Excel software capabilities.

The fact of the rigid binding of the studied text to the specific address of this text file and its name has been canceled, and one may consider as the useful innovations to the old version of the program that has already been described (see Table 1). The authors can declare an increase in the number of features that can be used as markers for recognizing of certain textual styles as the nextcoming useful steps in order to improve the basic computational procedure.

To find best ways to achieve the stated goal, the authors fulfilled additional statistical studies of those texts whose styles can be considered as exemplary ones. This work helped to clarify some of the controversial mathematical values (authors call them *singularities*, see [10]).

The presented below Table 2 shows the list of those singularities, which were taken into account by the main module of the program when it runs.

In general, it was processed more than 40 works (texts) which concern to the well-known four basic "writing styles", and this experience allowed the authors to clarify the singularities of the styles.

Table 1

Some innovations that are inherent in this version of the program "AN-TExp, v2":

1	The Program works with any digitized text file, with any file-name and placed (stored) at any storage ("disk")
2	The Program can process the given text by means of single macro only from the Excel master-file
3	After double decomposition of the given text, the Program can statistically analyze the text involving more singularities
4	The Program is supplemented by additional step in finding the final resume concerning to the text style

The program "AN-Texp, v.2" is able to analyze a given text content accordingly to the logic circuit provided in [10]. The program "AN-TExp, v.2" performs the frequency analysis of all keyboard characters involved in the text to be analyzed, and the obtained data will enable to calculate the mentioned above characteristics of the text. Thus the obtained information is *an objective mathematical description of the analyzed text*, and it contains the most detailed information about the studied object.

Table 2

The list of singularities which are used for textual style automated determination

$< X >$	" ! "	" ? "	" : "	" ; "	" , "	" – "	" (&) "	" % "	" < > "	" = "	"0…9"	" … "

3. The main components of the Program and how it works

When creating the program "AN-Texp" – the content analyzer of a digitized text – it was implemented the idea to use potentials of both MS Word and MS Excel tools and their interaction during data inside transfer. The MS Word text editor during preliminary preparation of a 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 all 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 work on the principle of capacity building [10]. 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 compound, which can function on the principle of capacity building.

As a result it was created the multi-modular program "AN-TExp, v.2", which helpes you to determine automatically the style of given textual contents via the desirable range of mathematical parameters that are the basis for objective analysis of the given text.

After start running the program provides calculating the particular values of $\boldsymbol{F}_{\text{singul}}$ (i.e. *singularity factors*) which reflect specific numeric ratios for given singularities of written characters selected by choice of authors. The present list of chosen singularities is presented at the Table 2. And the calculated numerical characteristics for those singularities are shown in the Table 3.

Table 3

The calculated reliable mean values for the Math expectation $< X >$ and the mentioned singularities

the Style :	$< X >$	$\boldsymbol{F}_{\text{singul}}$ " ! "	$\boldsymbol{F}_{\text{singul}}$ " ? "	$\boldsymbol{F}_{\text{singul}}$ " : "	$\boldsymbol{F}_{\text{singul}}$ " ; "	$\boldsymbol{F}_{\text{singul}}$ " , "	$\boldsymbol{F}_{\text{singul}}$ " – "	$\boldsymbol{F}_{\text{singul}}$ " (&)"	$\boldsymbol{F}_{\text{singul}}$ " % "	$\boldsymbol{F}_{\text{singul}}$ " < > "	$\boldsymbol{F}_{\text{singul}}$ " = "	$\boldsymbol{F}_{\text{singul}}$ "0…9"	$\boldsymbol{F}_{\text{singul}}$ " … "
Artistic	2750	0,006	0,006	0,006	0,004	0,131	0,036	0,001	0,0	0,0	0,0	0,0	0,005
Journalistic	668	0,002	0,002	0,007	0,001	0,096	0,026	0,016	0,001	0,0	0,0	0,051	0,0
Scientific	1277	0,0	0,001	0,008	0,004	0,094	0,024	0,032	0,001	0,0	0,003	0,065	0,0
Official-business	408	0,0	0,0	0,007	0,012	0,069	0,048	0,064	0,002	0,002	0,0	0,202	0,0

Note 1: The values of columns of «$\boldsymbol{F}_{\text{singul}}$ "....."» (*singularity factors* $\boldsymbol{F}_{\text{singul}}$) reflect specific numeric ratios for given singularities of written characters selected by choice of authors.

After calculating the frequency characteristics of the content of a given text file, the program compares the resulting coefficients with the corresponding singularity factors $\boldsymbol{F}_{\text{singul}}$ of the given text and the values from the database obtained by the authors (see the specific DB in [10] as the example). This step-by-step comparison of the obtained data with the "reference norms" occurs as comparison of statistical sampling characteristics with a

reliability of 95%.

So, the program "AN-TExp, v.2" realizes the procedure of defining a textual content style, and the procedure is similar to filtering routine of all the various information about the given text content (see Fig. 1), and such filtering is performed according to the criteria chosen by the authors – in terms of $< X > \pm \sigma$ and for selected singularities (see Table 2).

As for the presented values – such as mathematical expectation $< X >$ and factors $\boldsymbol{F}_{\text{singul},i}$ of each shown 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.

To answer the question about the possibility of presented algorithms "to feel" the characteristics of different text styles, it was decided to 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 for testing 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 the mean values), and as well as two-sample F-test for the variances.

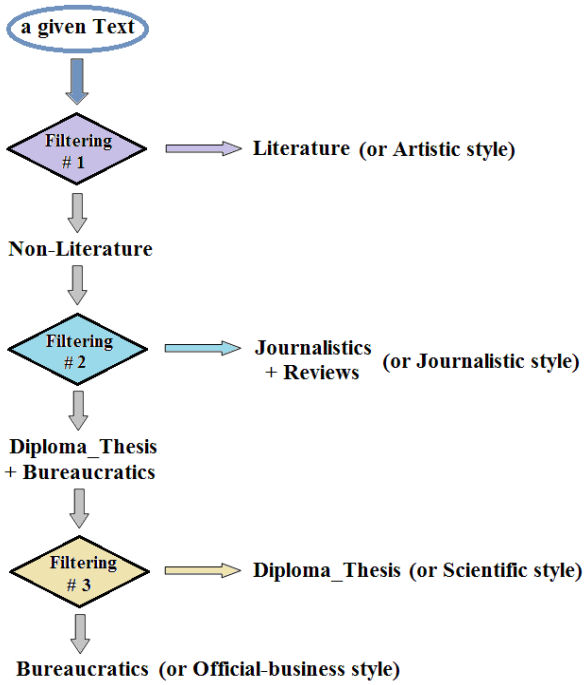

Fig. 1 : The basic logic for determining the style of a given text content

The program does its further conclusions about the textual content style based on the following "*table of matches*" (see Table 4).

4. Results and discussion

With the mentioned above program it has been analyzed more than 40 examples of the text of different styles (all in Russian), amongst which there were the real literary works (10 pcs.), essays and reviews of columnists and analysts from the serious electronic mass media of Ukraine (11 pcs.), graduate works of scientific and technical direction (10 pcs.), different administrative instructions or directives (11 pcs.). The shortened list of these texts is shown in the [10]. According to the authors, this study allowed to obtain the minimum reliable information by means of which it was possible to make correct judgments about those numerical characteristics that are associated with the style of text content.

After a three-level cross-checking of statistical hypotheses, the authors managed to compile the following coincidence table (see Table 4), which makes sense of a logical key for automatically determining the style of the text being examined.

Table 4

The obtained results of statistical hypotheses testing on equality of mean values (the Table of Matches):

	$< X >$	" ! "	" ? "	" : "	" ; "	" , "	" – "	" (&)"	" % "	" < > "	" = "	"0…9"	" … "
Artistic vs. non-Artistic	"diff. rel."	"diff. rel."	"diff. rel."		**0**			**0**			**0**		"diff. rel."
Journalistic vs. (Scientific + Official)			"diff. rel."		**0**		**0**	**0**	"diff. rel."	**0**	**0**	**0**	"diff. rel."
Scientific vs. Official	"diff. rel."			**0**	**0**	"diff. rel."	**0**		"diff. rel."	**0**	"diff. rel."	**0**	"diff. rel."

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

It means that all of these singularities in the table above are significant indicators which may help to reliably distinguish the text styles when examining a given text. So, using the Table 3 together with Table 4, one can reliably determine the text styles and finally include such a module in the overall algorithm of the entire procedure in order to automate it.

Note 3. The mentioned program "AN-TExp, v.2" was presented at the following intellectual competitions:

1) at the All-Ukrainian Championship Information Technology "Ecosoft-2017", the category "Programming" – the 3rd degree Diploma;

2) the XXI Belorussian (open) competition of students' research works, the category "Computer Science" – the Encouraging Diploma (Fig. 2).

The Diploma of 3rd degree of the All-Ukrainian Championship Information Technology "Ecosoft-2017", the category "Programming", for the program "AN-Texp, v.2"; the Nominee is Kalyuzhnyi Leonid – the student of 11-B form (school #49, Odessa) (Kyiv, feb.'2017)

The Encouraging Diploma of the XXI Belorussian (open) competition of students' research works, the category "Computer Science", for the program "AN-Texp, v.2"; the Nominee is Kalyuzhnyi Leonid – the student of 11-B form (school #49, Odessa) (Minsk, feb.'2017)

Fig. 2 : The Diplomas (2017) for the Program "AN-TExp, v.2"

Summary

We have presented our program "AN-TExp, v.2" for automatically text style recognizing of different digitized texts written in the Russian language. The results obtained show that the program (as well the basic algorithm [10]) can be successfully applied to Russian texts. The program-analyzer "AN-TExp, v.2" 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.

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 $\boldsymbol{F}_{\text{singul},i}$ of the texts) one can automate analyzing procedure of textual contents style in order to separate all the entries of the available textual DB according to style of their text contents: i.e. the program helps us in further routing of given texts accordingly to the style that one had identified.

The authors believe that it was calculated the reliable mean values for mathematical expectation $< X >$ and singularities (see Table 3), which will be useful for computer aided analysis of electronic texts.

The proposed algorithms of text contents analysis that were implemented by the program "AN-Texp, v.2", allow to use it for successful applying for both professionals and for interested users:

- when one has to provide automatic evaluation of vocabulary diversity of certain person based on his written works;
- when one has to provide automatic analysis of given posts on social networks and SMS according to certain criteria;
- when one has to provide automatic textual 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 style of the studied texts with a high degree of reliability (the statistical hypotheses about the equality or difference in the sample characteristics were tested with a reliability level of 95%).

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.

Acknowledgments

I would like to thank my colleague and partner, Mrs. Natalia Yevtushenko, for her thoughtful comments and suggestions.

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. Boytchev, P., Kamenova, S., Sendova, E., Stefanova, E., Kovatcheva, E., Nikolova, N., "IT for Innovative Educational Environments: Exploring, Authoring and Programming", In Proceedings of International Conference for Interactive Computer Aided learning – The Challenges of Life Long Learning ICL 2009, Ed. Michael E. Auer, Kassel University Press, Villach, Austria, 2009, pp. 434-444, ISBN 978-3-89958-481-3.

3. Naidenova X.A. and Shagalov V.L., "Diagnostic test machine" (2009). In M. Auer (Ed.), Proceedings of the ICL'2009 – Interactive Computer Aided Learning Conference, Austria, CD, (pp. 505-507). Kassel University Press, ISBN: 978-3-89958-481-3.

4. L. Dušková, "Textual links as indicators of different functional styles", Acta Universitatis Carolinae – Philologica 2, Prague Studies in English XXI (1996), pp. 113-123.

5. T. Suzuki, E. Kanou, Y. Arakawa, "Comparative Analyses of Textual Contents and Styles of Five Major Japanese Newspapers" (2013).

6. Kakkonen, T. "TexComp – A Text Complexity Analyzer for Student Texts" / Proceedings of the 12th International Conference on Interactive Computer-aided Learning, Villach, Austria, 2009 (8 pp).

7. Kakkonen, T., Myller, N., Sutinen, E.: "Semi-Automatic Evaluation Features in Computer-Assisted Essay Assessment" / Proceedings of the 7th IASTED International Conference on Computers and Advanced Technology in Education, Kauai, Hawaii, USA, 2004 (7 pp).

8. T. Kakkonen, N. Myller, E. Sutinen and J. Timonen, "Comparison of Dimension Reduction Methods for Automated Essay Grading", Educational Technology & Society, 11(3), 2008, pp. 275–288.

9. T. Kakkonen, E. Sutinen, "Semi-automatic Assessment Model of Student Texts – Pedagogical Foundations" / Conference ICL2009, September 23 -25, 2009 Villach, Austria (8 p.).

10. Chepok A.O., Yevtushenko N.I., "The Stylistic content-analyzer of texts: Part A – Validation of proposed Algorithms" // International scientific & technical magazine "Measuring and Computing Devices in Technological Processes", – 2016. – №3. – pp. 110-116. (available at: http://journals.khnu.km.ua/vottp/pdf/pdf_full/2016/vottp-2016-3.pdf)

Рецензія/Peer review : 24.5.2017 р. Надрукована/Printed :27.6.2017 р.
Стаття рецензована редакційною колегією