"Ruspersonality": a Russian Corpus for Authorship Profiling and Deception Detection

Tatiana Litvinova, Olga Litvinova, Olga Zagorovskaya
Voronezh State Pedagogica1 University
Voronezh, Russia
centr_rus_yaz@mai.ru

Pavel Seredin
Voronezh State University
Voronezh, Russia
paul@phys.vsu.ru

Aleksandr Sboev
Kurchatov Institute
Moscow, Russia
sag111@mail.ru

Olga Romanchenko
Plekhanov Russian University of Economics
Moscow, Russia
ghjd1@mail.ru

Abstract–Authorship profiling which is a process of the extraction of information about the unknown author of a text (demographics, psychological traits, et al.) based on the analysis of linguistic parameters, is a problem of great importance. Research in authorship profiling has always been constrained by the limited availability of training data since collecting textual data with the appropriate metadata (information about authors of texts) requires a lot of effort. We are presenting *RusPersonality* – first Russian-language corpus of written texts labeled with data on their authors. A unique aspect of our corpus is the breadth of the metadata (gender, age, personality, neuropsychological testing data, education level, etc.). Most texts were designed especially for this corpus, do not contain any borrowings and are not edited. The corpus is designed to serve multiple purposes: authorship profiling, authorship attribution, deception detection, genre detection etc. The corpus currently contains over 1850 documents from 1 145 respondents and is currently expanding. The average length of the texts is 230 words. The corpus can freely be used for academic research purposes on demand. The article describes the structure of the corpus and also shows the results of the research performed at our laboratory using its material and analyzes the perspectives for future studies.

I. Introduction

The development of methods of authorship profiling using texts has been gaining momentum worldwide and it involves designing text corpora containing not just texts but also metadata with the details about their authors (gender, age, scores on psychological tests, etc.), labelling of texts using automatic language processing tools, extraction of numerical values of quantifiable text parameters, calculation of correlations between these values and personality traits of authors and designing mathematical models to profile certain characteristics based on them [1].

Designing special text corpora is of growing importance in authorship profiling since collecting textual data with the appropriate metadata (information about authors of texts) requires a lot of effort.

Let us give a brief description of text corpora that are employed in this type of research. As the analysis suggests, these corpora can be divided into two groups – made up of previously written texts (for example, texts from social media) and those especially designed by respondents instructed by a researcher.

The first group contains text corpora with online communication (blogs, chats, tweets, social media messages, etc.) collected using a special software. Besides texts, details about authors are gathered as provided by them in their accounts (normally gender and age). These are corpora used for testing the methods of the annual international conference PAN on new methods of authorship profiling using texts [2]. These corpora have a lot of "noises" in them as they are collected automatically: they might contain borrowings (citations, links, etc.) and as shown in [3] that might make the calculations to classify the authors according to their gender and age almost twice less accurate. Furthermore details about authors might be false, which also undermines the significance of these corpora. What is certain is that this method of collecting material is a large number of these corpora and not so much time involved in compiling them. One of the largest corpora in English with metadata with information about the gender and age of authors contains 140 million words [4].

There are also corpora of online texts containing information about the gender and age of authors, results of online psychological testing. They are myPersonality3, containing texts and information from 250 Facebook users as well as messages and psychological testing data of Twitter users obtained by means of the Big Five Personality Test. It was used by an international team of PAN-2105 participants to test their methods of authorship profiling (English, Spanish, Italian and Danish tweets) (see Table I) [2].

Apart from corpora with online texts, there are also those especially designed for research purposes (see Table II). One of the most popular text corpora for studying language and personality is that of students' English essays compiled under the guidance of the American scientist J. Pennebaker. The corpus took 7 years to put together (from 1997 to 2004) and contains 2469 texts (one from each participant) (1.9 million words with an average of 770 words in a text) and metadata with information about the gender of authors and results of the Big Five Personality Test. This corpus is used by a lot of researchers as well as in the corpus for testing the authorship

ISBN 978-952-68397-6-9 (paperback), ISBN 978-952-68397-7-6 (PDF)

profiling methods as part of the contest in the run up to PAN [5].

TABLE I. CORPORA OF ONLINE TEXTS USED IN AUTHORSHIP STUDIES

Name of the Corpus/Characteristics	**Blog Authorship Corpus**	***myPersonality3***	***Twitter corpus***
Language	English	English	English, Spanish, Italian and Dutch
Metadata	Gender, age	Facebook profile, Big Five Personality Test	Gender, age, data of Big Five Personality Test
Number of respondents and texts	19320/ 681.288 posts	250/9900 statuses	726, no data
Length of texts	7250 words (35 posts) From an author	No data	140 characters
Genre	Blog	Facebook statuses	Tweets
Features	Texts are divided into three groups depending on the age of authors, balanced by gender	Data on the number of Facebook contacts as well as their density, connections are available.	Multilingual corpus
Expanded	No	Was last updated in February 2013	No
Access features	Free, online	Free, online	Links to the authors' profiles

This corpus used to be the only one of the type until in 2014 a Danish corpus *Stylometry Investigation (CSI) Corpus* [6] was launched. It contains texts by students of the University of Antwerp and is now in open access. It has 749 texts (305 000 characters) and there are plans to expand it annually with texts by newcomer students. The texts are essays on different topics (the average length of 1126 characters) and product reviews (an average of 128 characters). For each author there is information about the gender, age, birthplace, Big Five Personality Test results. The respondents were also given an option of providing information about their sexual orientation and doing the Myers-Briggs Personality Test.

Therefore according to the analysis, there are no sufficient research text corpora with meta-data providing details about their authors designed considering the genre and topic and used for research of authorship attribution and authorship profiling particularly for languages other than English. So far scientists seeking to monitor a few characteristics of text authors (e.g., gender, age, native language, and neuroticism level) have had to make use of a variety of text corpora instead of one: «Ideally, we would prefer to use a single corpus for all these problems but, unfortunately, there is no single corpus in which the documents are labeled for all four issues we consider» [1].

Up until recently there has been no such a text corpus in Russian. In this paper, we are presenting *RusPersonality*, a freely available Russian corpus that can be used for authorship profiling and many other applications.

TABLE II. CORPORA DESIGNED FOR AUTHORSHIP PROFILING STUDIES

Name of the corpus/Characteristics	**Corpus by J. Pennebaker**	CLiPS	***RusPersonality***
Language	English	Danish	Russian
Metadata	Gender, data of the Big Five Personality Test	Authors' data (gender, age, region, data of the Big Personality Test – Test MBTI, sexual orientation) and texts (for reviews – truthfulness/deceptiveness)	Authors' data (gender, age, region, education level, psychological testing data, type of lateral organization profile) and information about the texts (genre, topic, truthfulness/deceptiveness)
Number of respondents and texts	2469/2469	550, 550/1200, an average of 2.25 texts from an author	1145/1 867, an average of 1.6 texts from an author
Length of texts	770 words	1126/128 characters	From 56 to 230 words, an average of 162 words
Genre	Stream of consciousness essay	Essays and product reviews	Description of a picture, narrative about a day in life, essay
Features	Controlled of the topic and genre, age and education level, all of the texts are written by Psychology students	Controlled topic and genre, age and education level, all of the texts are written by students	A large number of metadata allows a study of mutual effects on linguistic text parameters
Expanded	No	Every year	Every month
Access	On demand	Free, online	On demand

II. CORPUS DESCRIPTION

In available Russian text corpora there is almost no metadata labelling with information about personality traits of their authors. Therefore in 2005 the Regional Centre for the

Russian Language Studies of Voronezh State Pedagogical University launched a corpus of Russian texts named *RusPersonality* containing metadata with information about the authors (gender, age, psychological testing results, education level as well as results of neuropsychological testing and profession in some subcorpora). The entire corpus has been anonymized and all the authors have explicitly given us permission to include their submissions and profile information in a corpus for research purposes.

Let us briefly present the structure and components of the corpus.

As of 23.04.2016 the corpus *RusPersonality* contains texts by **1 145** respondents with the total of **1 867** texts (depending on the type of the tasks the respondents were instructed to write one or two texts). In total there are around 300 000 word uses or around 1 800 000 characters.

A. Characteristics of authors

A unique aspect of our corpus is the breadth of the metadata. There is metadata available on both the authors and the documents included in the corpus.

For each author, we have information about the age, gender, language, education, personality scores on different tests. For some respondents data on their neuropsychological assessment are available.

GENDER. The authors were asked to indicate their gender, choosing 'male' or 'female'. The corpus contains texts by 386 males and 751 females. 8 respondents chose not to specify their gender (Fig. 1).

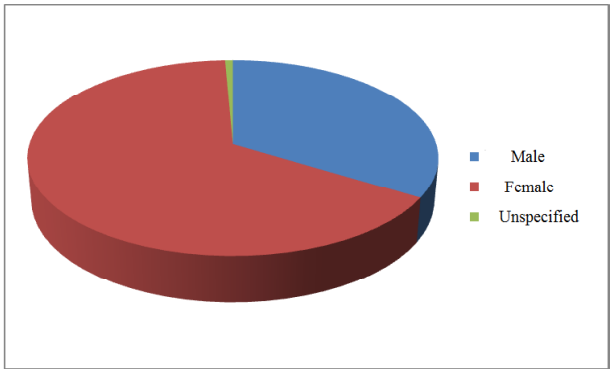

Fig. 1. Author distribution by gender

AGE. The authors provided us with their dates of birth. Given the timestamp of a document, we can compute the age of the author at the moment of writing. Individuals aged from 12 to 79 participated with the corpus (1519 texts) mostly made up of texts from individuals from 18 to 22 – students of the universities of Voronezh, Rostov-on-Don and Moscow (Fig. 2).

NATIVE LANGUAGE. Most respondents are native speakers of Russian but at the moment a subcorpus of texts by individuals speaking Russian as their second language who are expected to have a proficiency level to enable them to do psychological testing as well as to produce a text is being designed. At the moment of writing there are 48 texts by 37 individuals speaking Russian as their second language (Fig. 3).

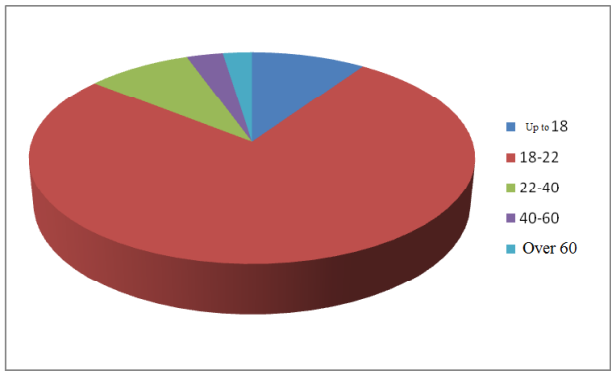

Fig. 2. Author distribution by age

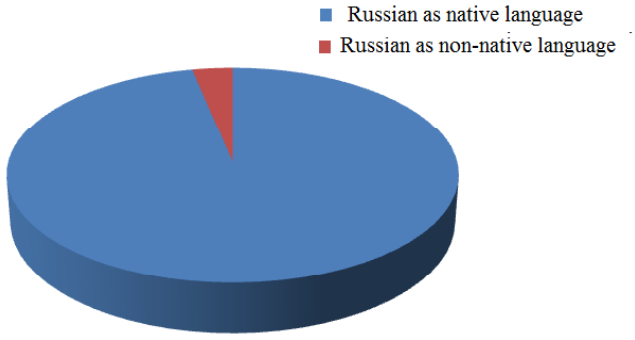

Fig. 3. Author distribution by native language

EDUCATION LEVEL. The corpus contains texts by individuals who have not completed school (i.e. school students, 96), university (i.e. students, 954), with a university degree (95) (Fig. 4).

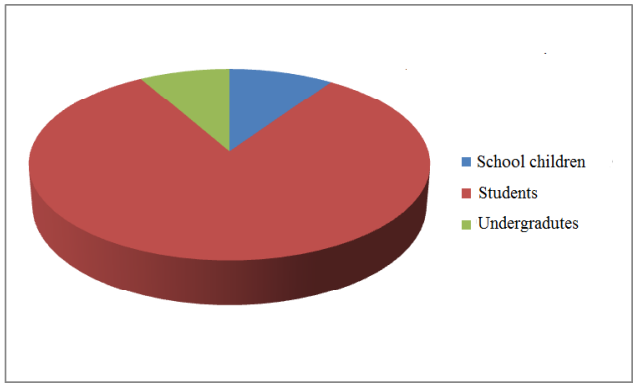

Fig. 4. Author distribution by education level

PSYCHOLOGICAL TRAITS. All the respondents except those with mental disorders did a series of psychological tests. The Big Five Personality Test (192 respondents), methods of communicative motivation by V. Boiko (75 respondents), Freiburg Personality Inventory Test (550 respondents), method by A. Belov for the dominant temperament type (124), accentuation type test by Leonhard -. Schmischeck (124), test "Domino" (54), "Style of Self-Regulation of Behaviour" (51), Hospital Anxiety and Depression Scale (HADS) (117), *The* Junior Eysenck Personality Questionnaire (158) (Fig. 5).

NEUROPSYCHOLOGICAL ASSESSMENT. One of the most important neuropsychological characteristics reflecting

individual differences in the joint operation of the human brain hemispheres (asymmetry) is the lateral brain organization of functions (LBOF) [7]. It is considered the foundation for the typology of individual differences of the mental condition of healthy individuals as part of a study in neuropsychology of individual differences. LBOF has an influence on the characteristics of the speech production as well [8], but this problem has not been sufficiently studied. It is considered the foundation for the typology of individual differences of the mental condition of healthy individuals as part of a study in neuropsychology of individual differences. As of now, 447 individuals have been tested for the motor, sensor and cognitive lateral profile.

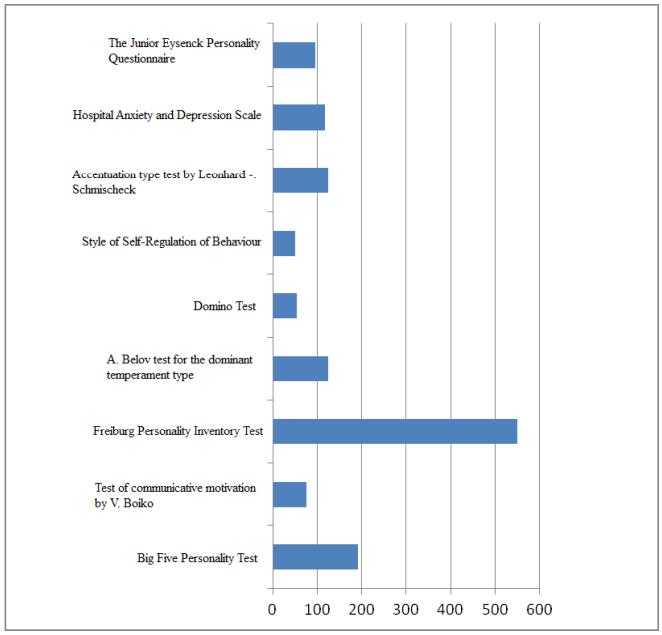

Fig. 5. Author distribution by personality tests

MENTAL DISORDER. There is also a corpus being designed for individuals with mental disorders (it currently contains 25 texts by individuals suffering from schizophrenia and 2 texts by those with bipolar personality disorder, the total of 27 respondents).

SUICIDE SUBCORPUS. Online texts (posts, diaries) of suicidal individuals are being collected. This is the only subcorpus of *RusPersonality* containing texts that were not especially designed but gathered from public sources. The subcorpus contains texts by 20 individuals with the total of 350 000 words.

This information can be used for a number of interesting experiments. For example, we can investigate the influence of someone's masculinity/femininity (FPI test) on the accuracy of gender identification using a linguistic analysis of their texts. Having both the psychological data and information about the type of lateral brain organization of functions allows us to compute the relation between these dimensions.

B. Characteristics of Texts

All the texts in the corpus are samples of natural written speech. The respondents were instructed to write a description of a picture (540 texts), any day of their lives (228), write a letter to a friend (734 texts), essays on "What would I Spend a Million US Dollars on?" (75), "What is the Meaning of Life?" (124) and also to describe their best assets to a potential employer (166) (Fig. 6).

The average length of a text ranged from 56 (a subcorpus of texts by individuals with mental disorders) to 230 words (a subcorpus of texts "What is the Meaning of Life?").

The text corpus *RusPersonality* also contains a subcorpus of "truthful" and "deceptive" texts. 114 respondents have participated as of now. In the first task they were asked to describe any day of their lives (e.g., a day before writing or a particular one) as it happened and in the second task they were to consciously twist the facts but so that a potential reader would find it hard to tell which text is "truthful" and which one is "deceptive".

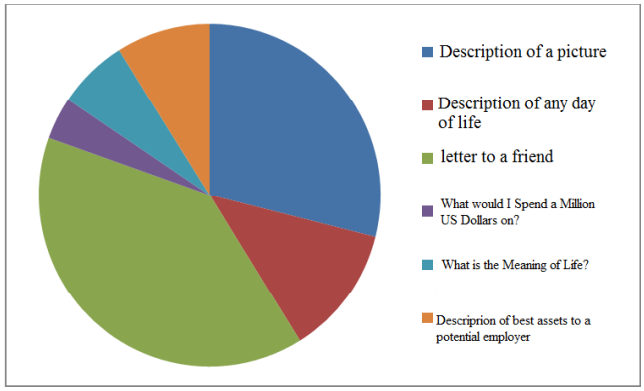

Fig. 6. Text distribution by topics

III. RESULTS OF THE RESEARCH ON THE CORPUS MATERIAL

To illustrate the usefulness of this corpus, let us present the results of a number of studies performed on the corpus material.

GENDER. We used a range of context-independent parameters:

- *Morphological features* – POS tag features, which mainly represent a particular part of speech for every word in a given text. These are the number of nouns; the number of numerals; the number of adjectives; the number of prepositions; the number of verbs; the number of pronouns; the number of interjections; the number of adverbs; the number of particles, the number of conjunctions, the number of participles, the number of infinitives, the number of finite verbs.
- *Syntactical parameters*– syntactic relations of different types (groups from 1 to 4 from [9]).
- *Derivative coefficients* which are different ratios of parts of speech (Treiger index, dynamics coefficient etc.);
- The number of exclamatory marks, the number of question marks; the number of dots; the number of emoticons;

- The number of words pertaining to a particular group "Emotion" (e.g., "Anxiety", "Discontent", the total of 37 categories)

as features and mathematical methods as logistic regression and different machine learning algorithms of text classification (see Table 3 and [10, 11] for details).

TABLE 3. RESULTS OF PREDICTING AUTHOR GENDER USING MACHINE LEARNING ALGORITHMS

Model	Feature selection techniques	Mean F1-score (25 cycles)
Gradient Boosting	imp_quarter	0.72
adaBoosting	imp20	0.71
ExtraTrees	imp10	0.7
adaBoosting	common	0.7
Random Forest	imp10	0.7
PNN(sigma = 0.1)	imp10	0.68
SVM	PCA (30)	0.66
ReLU (1 Hidden Layer with 26 neurons)	imp10	0.74

PERSONALITY. Using morphological (POS, POS bigrams, POS ratios) and syntactical features (a number of clauses, simple sentences, complex and compound sentences), we have designed regression models to detect personality traits from the Big Five Personality Test. The accuracy of the model was from 60 to 65 % depending on a personality trait [11, 12].

Besides we have set forth a method for detecting not only individual personality traits but also a set of personality traits and self-destructive tendencies in particular [13].

We selected from *RusPersonality* texts by individuals with high (those scoring high (7-9) on 3 of 12 scales of Freiburg Personality Inventory: "Spontaneous Aggressiveness" (individuals scoring high on this, display high psychotisation levels resulting in growing impulsive behaviour risks), "Depressiveness" (individuals scoring high on psychopathological depressive syndrome), "Emotional Lability" (high scores are indicative of an unstable emotional condition with affective reactions), and low (1-3) on "Composedness" (low scores are indicative of low stress resistance), N = 33 (16 females, 17 males, average age is 20, SD = 2.3), and low risks of self-destructive behaviour (i.e. those scoring low (1-3) on 3 scales of FPI: "Spontaneous Aggressiveness", "Depressiveness", "Emotional Lability", and high (7-9) on "Composedness", N = 27 (13 females, 14 males, average age is 19.5, SD = 2.2) and labelled it according to the list of parameters chosen based on the neurolinguistics data on speech production in the brain. Each participant was asked to produce two texts which were then analyzed as one text: a letter to a friend about things happening lately, and one to an imaginary employer explaining why they (the respondents) were good for a particular job. Respondents were instructed to write as much as possible: whatever first came into their minds. There was a time limit of 40 minutes. An average text was 176 words long, SD = 54 words.

In order to detect self-destructive tendencies (as noted above, a set of personality traits) by means of the obtained correlation coefficients considering multicollinearity, a regression model, which was a system of linear equations (for each personality trait associated with self-destructive behaviour), was designed. The accuracy of the model was 80 %.

As correlation-regression analysis ($p<0.05$) show, texts produced by individuals with a greater likelihood of self-destructive behaviour typically show less lexical diversity, fewer prepositions, more pronouns overall (and particularly personal ones), a higher coefficient of coherence (due to more conjunctions and deictic particles), and higher average sentence lengths as compared to texts produced by people with less likelihood of self-destructive behaviour.

DECEPTION DETECTION. We have conducted a pilot experiment to detect intentionally deceptive information in written texts [14]. Although this field is currently gaining momentum, the majority of past studies have concentrated on English-language texts. We use LIWC feature [15]. LIWC processes text based on 4 main dimensions: standard linguistic dimensions, psychosocial processes, relativity and personal concerns. Within each dimension, a number of variable are presented, for example, the psychosocial processes dimension contains variable sets representing affective and emotional processes, cognitive processes and so forth. The analysis was performed along 32 parameters used to distinguish "truthful" and "deceptive" texts.

Deception cues were identified using mathematical statistics in three stages. During the first stage, a coefficient of variation for each of the parameters of "truthful" and "deceptive" texts was determined. This was done in order to establish which linguistic characteristics remain stable in texts by the same author and which vary. In order to achieve this, we calculated the deviation of each text parameter from its average value for a particular individual. Furthermore, we averaged a deviation for each parameter in all of the texts and determined a coefficient of variation to enable us to evaluate a range of text parameters and see how large it is in relation to an average value.

Proposed approach for detection deception in written texts has exhibited an accuracy of ~ 72 %. Our results are comparable with the state-of-the-art results of for English opinion texts.

DETECTION OF THE LATERAL BRAIN ORGANIZATION OF FUNCTIONS (LBOF). Currently, we are working on designing models for identifying lateral brain organization of functions using the analysis of written texts. In [16] the results of a correlation analysis are shown which identified a connection ($p < 0.05$) between text parameters (index of lexical diversity TTR was identified and a negative one with a proportion of function words + pronouns; proportion of function words; proportion of cognitive words; proportion of full stops; proportion of 100 most frequent Russian words) and the type of LBOF of their authors. The index of the lateral brain organization (motor, cognitive, sensory as well as individually for hands, legs, eyes, ears) was calculated as the difference between the number of "right", "left" and "symmetrical" answers divided into the number of tests. The largest number

of correlations of the parameters (r = 0.27-0.41) of texts were found with right/left handedness of the author of a text.

A large number of correlating parameters with the Pearson coefficient enabled a regression model to be designed based on a multi-parameter linear approximation considering the most significant correlations. However a test of the models suggests that this type of approximation yielded low accuracy making it hardly possible to detect individuals with left-side profiles. It is easy to explain if we note a small proportion (<10%) of individuals with such profiles in the training set in particular. I.e. for low correlation coefficients as well as a small proportion of left-side profiles there are more errors in designing a model which has to do with a rather large range of values of an analyzed personality trait and a set of text parameters and thus an optimal selection of an optimal regression being impossible to design.

Therefore we made a decision to make use of not one multi-parameter regression model but to design a few regression equations for each of the text parameters most correlating and with certain lateral organization profile based on optimal selections for each text parameter considering the sign of a correlation coefficient and without statistical outliers. Let us demonstrate the suggested approach using an example of lateral profile organization of hands as one of the most significant indices of lateral organization profile.

The text parameters most significantly correlating with lateral profile organization of hands are TTR, a proportion of function words + pronouns in a text, proportion of common words, proportion of the conjunction “but”, proportion of the preposition “with”. For each of the parameters regression equations were designed with the training selections for each parameter considering a range of its values and a diversity of lateral organization profile of hands.

The designed linear regression equations for detecting handiness are as follows:

$$Y = -2.230 + (3.538 * TTR) \quad (1)$$

$$Y = 2.886 - (7.824 * (\text{proportion of function words+ pronouns in a text})) \quad (2)$$

$$Y = 3.723 - (0.0976 * \text{common}) \quad (3)$$

$$Y = 1.143 - (0.0190 * \text{proportion of pronominal nouns}) \quad (4)$$

$$Y = 1.405 - (0.872 * \text{proportion of “but”}) \quad (5)$$

$$Y = -0.275 + (0.370 * \text{proportion of “with”}) \quad (6)$$

For effective evaluation of the results we are suggesting that the following approach be employed. Let us determine how many times six of the obtained equations the design value of handiness is within the following ranges: (-1; -0.33), [-0.33; 0.33], (0.33; 1). Then by estimating the likelihood of that, we can make conclusions about right/left handedness of the author of a text.

The test of this approach and the obtained model for handiness showed that the likelihood of identifying right/left handedness by means of this approach is 67 %.

We believe that this approach is promising and might yield higher accuracy provided it is used on a large corpus.

VII. Conclusion

Hence we have employed a personality profiling approach that is gaining momentum abroad for Russian texts. It proved to be efficient, however we tend to think that an approach to personality profiling using texts that involves a psycholinguistic analysis of different levels of texts as well as their coherence and cohesion is effective and well-grounded.

We are planning to expand the text corpus by getting more participants involved as well as by including online texts where there is information about authors (normally gender and age). However, we think that manual processing should be used for collecting this material: there can be citations, links and other information that can be deleted. Furthermore, as was previously said, there can also be intentionally deceptive information and thus special considerations should be made while designing this type of corpora.

Thus the corpus of Russian written texts *RusPersonality* can presently be used in developing and testing the methods for addressing the following issues facing forensic authorship analysis:

1) authorship attribution;
2) authorship profiling;
3) deception detection etc.

Corpus studies of written texts enable further research into variations of written text units and development of effective methods of their formal linguistic description in order to identify a set of objective stylometric parameters of written Russian texts, which is a pressing issue facing forensic authorship analysis.

Acknowledgment

The study of gender identification of the author of a written text is financially supported by the Russian Science Foundation, project No 16-18-10050 “Identifying the Gender and Age of Online Chatters Using Formal Parameters of their Texts”. The study of the identification of intentionally deceptive information in written text is supported by a grant from the Russian Foundation for Humanities N 15-34-01221 “Lie Detection in a Written Text: A Corpus Study”. The study of the connection between the human lateral organization profile and typological characteristics of their written texts is financially supported by the grant of RFBR “Linguistic Parameters of a Written Text and Neuropsychological Characteristics of its Author: A Corpus Study”, project number 16-36-00036. The study of texts by suicidal individuals was performed as part of the project “Predicting the probability of suicide behavior based on speech analysis” from RF President's grants for young scientists (grant agreement N° MK-4633.2016.6).

REFERENCES

[1] S Argamon., M Koppel., J. Pennebaker, J. Schler, "Automatically profiling the author of an anonymous text", *Communications of the ACM*, vol. 52, 2009, pp. 119–123.

[2] F. Rangel, C. Fabio, P. Rosso, M. Potthast, B.Stein, W.Daelemans, Overview of the 3rd Author Profiling Task at PAN 2015, in *Linda Cappellato and Nicola Ferro and Gareth Jones and Eric San Juan (eds.): CEUR Workshop Proceedings*. Toulouse, France (2015), Web: http://www.sensei-conversation.eu/wp-content/uploads/2015/09/15-pan@clef.pdf

[3] M. Villegas, M. Ucelay, M.Errecalde, L. Cagnina, "A Spanish Text Corpus for the Author Profiling Task", in *Proc. of XX Congreso Argentino de Ciencias dela Computación (CACIC 2014).* San Justo, Buenos Aires, Argentina, 2014, Web: http://sedici.unlp.edu.ar/bitstream/handle/10915/42290/Documento_completo.pdf?sequence=1

[4] J. Schler , M. Koppel , S. Argamon, J. Pennebaker, Effects of Age and Gender on Blogging, in *Proc. of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs* (March 2006), Web: http://u.cs.biu.ac.il/~koppel/papers/springsymp-blogs-07.10.05-final.pdf

[5] F. Celli, F. Pianesi, M. Stillwell, D. Kosinski, "Workshop on computational personality recognition (shared task)", Web: http://clic.cimec.unitn.it/fabio/wcpr13/celli_wcpr13.pdf

[6] B. Verhoeven, W. Daelemans, "CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection", in Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). Reykjavik, Iceland (May, 2014), Web: http://www.clips.ua.ac.be/bibliography/clips-stylometry-investigation-csi-corpus-a-dutch-corpus-for-the-detection-of-age-gende

[7] E.D.Khomskaya, I.V.Yefimova, E.V. Budyka, E.V. Enikolopova, Neuropsychology of Individual Differences (Left-Right Brain and Mental Condition). Moscow: Russian Pedagogical Agency, 1997.

[8] A.V. Shubin, E.I. Serpionova, "Brain Asymmetry and Features of Verbal Creativity", *Voprosy psikhologii*, vol. 4, 2007, pp. 89-97.

[9] Russian language text corpus with syntactic markup: user information, Web: http://www.ruscorpora.ru/instruction-syntax.html#Синтаксическая

[10] T. Litvinova, P. Seredin, O. Litvinova A. Sboev, O. Zagorovskaya, D. Gudovskikh, I. Moloshnikov, R. Rybka, "Predicting The Gender of an Author of a Russian Text Using Regression and Classification Techniques", in *Proc. of CDUD 2016* (in press).

[11] T. A. Litvinova, P. V. Seredin, O. A. Litvinova, "Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study", Indian Journal of Science and Technology, vol. 8, N 9 [S. l.], 2015, pp. 93-97.

[12] T. Litvinova, "Profiling the author of a written text in Russian. Journal of Language and Literature", vol. 5(4), 2014, pp. 210-216.

[13] T. Litvinova, O. Litvinova, "Authorship Profiling in Russian-Language Texts", in *JADT 2016: 13ème Journées internationales d'Analyse statistique des Données Textuelles* (in press).

[14] T.A. Litvinova, Litvinova O.A., "A study of linguistic characteristics of texts that contain deliberately distorted information by program Linguistic Inquiry and Word Count", Bulletin of Moscow State Open University. Series: Linguistics vol. 4, 2015, pp. 71-77.

[15] J. W. Pennebaker, C. K. Chung, M. E. Ireland, A. L. Gonzales, R. J. Booth, *The Development and Psychometric Properties of LIWC2007*. Web: http://www.liwc.net/LIWC2007LanguageManual.pdf

[16] T. Litvinova, E. Ryzhkova, O. Litvinova, "Features of Written Texts of People with Different Profiles of the Lateral Brain Organization of Functions (on the Basis of *RusNeuroPsych* Corpus)", in *Proc. of the International Conference of Experimental Linguistics ExLing 2016*, Saint Petersburg, 2016 (in press).