Tatyana A. Litvinova. Profiling the author of a written text in Russian. Journal of Language and Literature 2014; 5(4), 210-216.
***DOI:** 10.7813/jll.2014/5-4/45*

PROFILING THE AUTHOR OF A WRITTEN TEXT IN RUSSIAN

Tatyana Aleksandrovna Litvinova

Voronezh State University of Architecture and Civil Engineering **(RUSSIA)**
centr_rus_yaz@mail.ru

DOI: 10.7813/jll.2014/5-4/45

Received: 09 Jul, 2014
Accepted: 30 Sept, 2014

ABSTRACT

Statement of the problem. Presently there is no doubt about the fact that the text at its different levels is a reflection of its author's personality (gender, age, psychological profile, etc.). The existing research, which is largely based on the English language materials, shows a promise in studying the uncontrollable parameters of the text mostly at the grammatical level. However, there is a need for further research to be carried out in languages different in their structure from English in order to develop the method of author profiling using the linguistic analysis. **Results and conclusions.** This paper is a pilot study of the relationship between the formal and grammatical parameters of the text and personality traits of the author, i.e. it provides an in-depth insight into written text author profiling. Using the methods of mathematical statistics, regression models were obtained which make a connection between the formal and grammatical characteristics of the text and personality traits of its author. The material of the study is a corpus of texts providing metatags for the information of its authors.

Key words: author profiling, gender attribution, text author profiling, authorship attribution, function words, personality markers in text

1. INTRODUCTION

At present it is considered proved that the text as a product of the individual's speech activities conveys information on its author's personality. However, there is no agreed method for text author profiling based on the linguistic analysis of its parameters.

Linguistic methods of author profiling have been extensively used by the colleagues overseas for about thirty years. As early as in 1979, the German researcher K. Scherer (1) pointed out the possibility of speech not just identifying the social characteristics of its author (social position, education, occupation and social standing) but also some personality traits. Since then the scientists overseas using the materials of text corpora and powerful mathematical tools have identified a number of effective language parameters with a high detection performance for English language texts. The analysis of the papers of foreign linguists in the field suggests two major approaches to text author profiling:

1) content-based;
2) style-based (formal parameters of the text, occurrence of grammatical phenomena).

The first approach suggests that specific lexical-semantic groups of words (positive/negative emotion words, terms of "semantic non-exclusivity" ("none", "always", "not at all", etc. which are all characteristics of emotionally charged speech), words for organizing thoughts) to a certain extent correlate with specific personality parameters (2; 3). Some of the disadvantages of content-based analysis for recognizing psychophysiological characteristics of the author of the text are the following two. First, the topics of the text are controlled by the author. Second, it is unknown which semantic groups need to be studied. Another disadvantage is that topic groups are chosen based on subjective ideas and are in a certain way dependent on the cultural background of the researcher.

The second approach in recognizing psychophysiological characteristics of the author of the text is based on the analysis of consciously controlled predominantly grammatical parameters of the text – morphological and syntactical (e.g., see (4)). The metaanalysis of research papers (predominantly they deal with English language) has shown there are about 1000 different style-based (statistical-stylistic) text parameters. These parameters vary in efficiency and should be used in conjunction. For example, paper (4) shows how the gender of the author of an anonymous text can be determined with a 80% accuracy using "function words" (pronouns, conjunctions, prepositions, particles), sequences of the parts of speech and punctuation features as the parameters. Females use different types of pronouns more frequently than males, while males use more articles and prepositions, which is due to women willing to be involved and men to be informative (see also (5)).

A significant contribution was made by papers on establishing the correlation between language parameters of written texts and an individual's psychological traits. So, the paper (6) shows that extroverts were more inclined to a so-called “implicit language” (more pronouns, adverbs and verbs; fewer nouns, adjectives, prepositions), while introverts displayed a preference to “explicit language”. According to the above scientists, this is caused by extroverts being more involved into space, better knowing their way around, whereas introverts require their current realms to be external rather than individual.

It should be noted that most of author profiling research was performed for English materials, but it seems obvious that in order to introduce effective methods for written text author profiling, other languages need to be included as well as those different in structure from English.

To date there has been a considerable amount of Russian language research that examines the interrelationships between individual traits of the author and parameters of the text produced by individual. However, not much progress has been made in the field (7) because no statistical models have been used on a large Russian language corpus (8), and predominantly such studies deal with content-based features.

2. MATERIALS AND METHODS

The literature review concludes that recently due to a rapid development of automatic language processing (morphological and syntactic parsers), statistical data processing software, stylometric approach to text author profiling emerges into the spotlight. The idea of the approach is that using an extensive corpus material correlations are found between quantitative parameters of texts and characteristics of their authors by means of statistical data processing methods. Therefore, in order to address this problem, three essential approaches need to be employed:

- a text corpus designed for the task containing metatags as social biographical information of its authors (gender, age, education, profession, psychological tests data, etc.). It should be noted that creating this kind of a corpus is a daunting task, since there are no widely accessed Russian corpora like that (for more details see (9));
- *a set of text parameters* which can be informative for profiling a specific characteristics of the individual producing it. As the current research suggests, the author’s personality comes through at all the levels of the text, however, a quantitative analysis of semantics and vocabulary is time and effort consuming and cannot be made fully automatic at this point of the research and this is why a large share of attention is paid to morphological and to a certain extent syntactical parameters of the text;
- *mathematical methods* of establishing the correlations of numerical values of text parameters and the author’s personality traits (scores on psychological tests). Statistical data processing and machine learning, etc. methods are applied for the task (10).

The problem we are seeking to address in this paper is the description of the results of the experiment conducted by the team of authors to establish the correlation between formal and grammatical quantifiable text parameters and the author’s personality traits (gender and psychological traits) based on the text corpus using statistical data processing methods.

Let us look in more detail at the text corpus used in the study, a set of formalized text parameters, mathematical methods.

Methods for designing a corpus for author profiling are detailed in (9). To the best of our knowledge, the corpus used is one of its kind. At the moment the corpus contains 1025 texts gathered from 586 individuals (they were instructed to write two texts on a specific topic but some of them chose to write only one), information on the author’s gender and their psychological testing data. The participants were first- to fifth-year students of Voronezh and Moscow universities (science and humanities students). The topics to choose from were “A Letter to a Friend”; “Describe a Picture”; “What would I do if I had a million US dollars?”, “Convince the employer that you are perfect for this job”, etc.

For a pilot study, 150 texts from 75 participants (26 males, 49 females) were selected with the average number of words being 166. The participants were also asked to fill in their gender, major (science/humanities) and filled in the questionnaires of two psychological tests – Big-Five personality test by McCrae and Costa (traditionally used for author profiling in English-language research, we used Russian version adapted by V. E. Orlov in collaboration with A. A. Rukavishnikov and I. G. Senin) that helps to measure one of the key five factors (extraversion, agreeableness, conscientiousness, neuroticism, openness) and questionnaire “Methods of Recognizing a Communicative Intention” by V.V. Boyko. Individuals with high scores on the test show low level of tolerance towards other people in the course of communication, and vice versa, those who have low score on the test are highly tolerant.

Automatic text processing (TAP) (using morphological analyzers, word frequency counter software) extracted numerical values of formal and grammatical parameters of the text which were listed according to the data obtained in the review of Russian and English language literature as well as the author’s research prior to the project. There was a total of 75 text parameters all of which are relatives values that is correlations of numerical values of different text parameters (part-of-speech correlations, e.g. (vfin+vinf)/noun, adj/(adv+pronadv), correlations of the number of types of various syntactic structures and so on). Ratios, i.e. relative frequencies, were used as the parameters in order to refrain from the dependence on the length of the text.

We are building on the assumption that *function words* (pronouns, prepositions, conjunctions, particles, auxiliary verbs, deictic adverbs) are most relevant in author profiling. They do not have a nominative function, are morphologically inseparable, semantically and syntactically dependent. Function words are less consciously controlled and given less attention in speech than content words and processed by the brain in a different way than content words as suggested by the studies of aphasia.

In order to estimate the closeness and direction of the linkage between the parameters of the text and personality and to establish the analytical expression (form), correlation and regression analysis was used based on modern statistical data visualization software. The main aim of the study was to establish a function dependence of a conditional mean of the result property (Y) (gender, scores on psychological tests) on the factor properties (x_1, x_2, …, x_k), which are the parameters of the text. Therefore the initial regression equation, or a statistical model of the relationship between the author's personality traits and quantitative parameters of the text is given by the function

$$Y_x = f(x_1, x_2, \dots, x_n),$$

where n is a number of factors included in the model; x_i are the factors that influence the result Y.

Correlation regression analysis for profiling the author of a written text was conducted in several stages. The first stage was the statement of the problem of the study, the selection of the computing methods and collection of the data as well as the identification of the factors connected as a rigid system "text parameters – author parameters" using the Pearson correlation coefficient and assessment of credibility of all the characteristics of the correlation linkage with the linkage p=0.05. It was then assumed that a linkage (type of an analytic function) between the chosen parameters of the text and author and their personality traits should be linear. The third stage sought to identify the initial regression equations using regression analysis methods of the SPSS software and analyze the obtained equations in order to identify the errors of the resulting laws in the test group.

For the logical regression (gender) it was assumed that 1 = male, 0 = female. As for an ordinary multilinear regression (results of the psychological tests), the number calculated using the equation ranges from 0 to 100 as well as the test scores. The deviation, or error, was assumed to be averaged in 75 parameters prior to the calculation of the mean deviation of a specific result followed by averaging the result.

3. RESULTS

Below are the obtained regressions describing the relationship between the numerical values of the text parameters and personality traits.

3.1. Gender

Hence in order to identify the gender of the author of the text, the following correlations are found (Table 1).

Table 1. Correlations for gender

Index number	1	2	3	4	5	6	7	8	9
Gender	the number of content words / the number of function words	the number of nouns / the total of words	the number of function words / the number of nouns	(DET+IREL+PERS+PRONADV)/ the total of words	(pronouns of all types + prepositions + pronominal adverbs) / the total of words;	asyndetic composite sentences / the total of the composite sentences;	(pronouns + conjunctions + particles) / the total of words	the number of adverbial participles / the total of words;	personal pronouns / the total of words
Correlation coefficient	0.258	0.252	-0.297	-0.325	-0.269	0.253	-0.286	-0.272	-0.274
Pearson correlation coefficient	0.0285	0.0327	0.0114	0.00528	0.0223	0.0322	0.0148	0.0209	0.0198

Following regression is obtained based on the correlations above:

$$Y = -0.231 - (0.0395\cdot(\mathbf{1})) + (2.681\cdot(\mathbf{2})) + (0.204\cdot(\mathbf{14})) - (1.301\cdot(\mathbf{20}) -$$
$$- (0.658\cdot(\mathbf{21})) + (0.466\cdot(\mathbf{25})) - (2.214\cdot(\mathbf{48})) + (1.173\cdot(\mathbf{55})) - (1.832\cdot(\mathbf{59})),$$

where numbers in bold are index number of the parameters.

The accuracy of the model assessed on test corpus is **~60%.**

3.2. Scores on the test "methods of recognizing a communicative intention" by V.V. Boyko

For the parameter "Scores on Boyko's test" the following correlations are found (Table 2).

Table 2. Correlations for the scores on Boyko's test

Index number	1	2	3
Scores on Boyko's test	compound sentences / the total of composite sentences	proper names / the total of words	proper names / (the total of nouns + personal pronouns)
Correlation coefficient	-0.255	0.341	0.339
Pearson correlation coefficient	0.0358	0.00448	0.00472

$$Y = 65.263 - (13.116\cdot(\mathbf{1})) - (18.872\cdot(\mathbf{2})) + (86.626\cdot(\mathbf{3})).$$

The deviation from the actual result is ~ **10 %.**

3.3. Extraversion

For this personality trait the following text parameters were found to be relevant (Table 3).

Table 3. Correlations for extraversion

Index number	1	2	3	4	5	6	7
Extraversion	the number of clauses / the total of the sentences	the total of participles and adverbial participles / the total of words	the number of conjunctions / the number of prepositions	the number of demonstrative + relative interrogative pronouns / the total of words	prepositions / the number of words	the number of adverbial participle groups + the number of participle groups / the number of isolations	the number of adverbial participle groups + the number of participle groups / the number of isolations
Correlation coefficient	0.232	-0.245	0.257	0.33	-0.232	-0.236	-0.351
Pearson correlation coefficient	0.0454	0.0343	0.0258	0.00386	0.0449	0.0414	0.00201

$$Y = 63.740 - (0.107\cdot(\mathbf{1})) - (39.485\cdot(\mathbf{2})) - (1.499\cdot(\mathbf{3})) +$$
$$+ (10.665\cdot(\mathbf{4})) - (120.792\cdot(\mathbf{5})) - (3.899\cdot(\mathbf{6})) - (623.818\cdot(\mathbf{7})).$$

The deviation from the actual results is ~**13-14%.**

3.4. Agreeableness

The relevant parameters for agreeableness are as follows (Table 4).

Table 4. Correlations for agreeableness

Index number	1	2	3	4	5	6
Agreeableness	the number of words / the number of clauses	the number of prepositions / the number of function words	(the total of pronouns + prepositions + pronominal adverbs) / (conjunctions + pronouns + interjections + prepositions + pronominal adverbs + particle + auxiliary verbs);	the number of conjunctions / the number of prepositions	prepositions / the total of words	adverbial participles / the total of words
Correlation coefficient	-0.246	-0.257	-0.23	0.276	-0.267	-0.347
Pearson correlation coefficient	0.035	0.027	0.0491	0.0174	0.0217	0.00247

$$Y = 80.427 - (1.227\cdot(\mathbf{1})) - (15.140\cdot(\mathbf{2})) - (12.020\cdot(\mathbf{3})) - (1.452\cdot(\mathbf{4})) - \\ - (51.413\cdot(\mathbf{5})) - (766.367\cdot(\mathbf{6})).$$

The deviation from the actual result is ~**15%.**

3.5. Conscientiousness

The following parameters of texts are found to correlate with scores on conscientiousness (Table 5).

Table 5. Correlations for conscientiousness

Index number	1	2	3	4
Conscientiousness	adjectives / (adverbs + pronominal adverbs)	the total of the number of participles and adverbial participles / the total of words;	the number of demonstrative + interrogative relative pronouns / the total of words;	the number of the adverbial participles / the total of words
Correlation coefficient	-0.267	-0.242	0.233	-0.329
Pearson correlation coefficient	0.0215	0.0376	0.0454	0.00421

$$Y = 55.472 - (2.689\cdot(\mathbf{1})) - (55.871\cdot(\mathbf{2})) + (8.077\cdot(\mathbf{3})) - (546.071\cdot(\mathbf{4})).$$

The deviation from the actual result is **~18%.**

3.6. Neuroticism

The relevant parameters for neuroticism are as follows (Table 6).

Table 6. Correlations for neuroticism

Index number	1	2
Neuroticism	adjectives / (adverbs + pronominal adverbs)	the number of adverbial participles / the total of words
Correlation coefficient	-0.287	-0.272
Pearson correlation coefficient	0.0131	0.0192

$$Y = 55.201 - (2.697\cdot(\mathbf{19})) - (521.891\cdot(\mathbf{54})).$$

The deviation from the actual result is ~**18%.**

3.7. Openness

The following parameters were found to be relevant (Table 7).

Table 7. Correlations for openness

Index number	1	2	3	4	5	6	7
Openness	the number of conjunctions / the number of function word frequencies	the total of particles / (conjunctions + pronouns + interjunctions + prepositions + pronominal adverbs + particles + auxiliary verbs)	the total of participles and adverbial participles / the total of words;	(demonstrative pronouns + interrogative relative pronouns) / the total of words;	the number of nouns / the total of pronouns	particles / the total of words	adverbial participles / the number of words.
Correlation coefficient	0.237	-0.285	-0.33	0.268	-0.25	-0.294	-0.417
Pearson correlation coefficient	0.0417	0.014	0.00405	0.021	0.032	0.0109	0.000217

$$Y = 60.238 + (14.697\cdot(\mathbf{1})) + (227.831\cdot(\mathbf{2})) - (76.134\cdot(\mathbf{3})) + (5.893\cdot(\mathbf{4})) - \\ - (2.515\cdot(\mathbf{5})) - (576.137\cdot(\mathbf{6})) - (580.465\cdot(\mathbf{7})).$$

The deviation from the actual result is ~**9%**.

4. DISCUSSION

It is therefore clear that this approach proved to be effective overall. The models were acquired that yield a high accuracy except the logical regression (authorship gender attribution), which might be due to gender attribution being insufficiently balanced. The assumption on the significance of function words and pronouns for gender attribution was also proved: specific relationships between these parts of speech come in handy in identifying most personality traits.

A large contribution was made by the analysis of the syntactic level of the texts and the structure of sentences in particular, however, it has not been made automatic enough so far and there is therefore a limited number of text parameters to be investigated at the syntactic level. These are the number of simple sentences; the number of compound sentences; the number of simple sentences in the compound one; the number of composite sentences according to the connection between its parts (asyndetic, compound, complex). All these parameters of the text are crucial in author profiling.

It should be noted that we intentionally avoided a content analysis of the specified texts since we made it our aim to search for formal grammatical parameters of texts that would correlate with personality traits. Our research showed that this correlation does exist and more research needs to be undertaken in the field and a single theoretical conception employing not only linguistic but also psychology data needs to be developed in order to account for the effective performance of specific formal and grammatical parameters in personality attribution. The creation of this conception will inevitably result in new relevant text parameters and thus more effective models.

It also should be noted that the current study is pilot and is merely an outline of the tasks to be addressed in author profiling based on formalized, consciously uncontrolled parameters of a written text. There needs to be a more relevant and correct selection procedure in logical regression; an analysis of the behavior of a specific text parameter as correlating with a certain characteristics of the individual producing it; a study on a larger corpus material. However, the assumption has already been proved that the most relevant parameters for automatic author profiling are indicative of the frequencies of function words and pronouns.

ACKNOWLEDGEMENTS AND FUNDING

The study was supported by grants of Russian Foundation for Basic Research (RFBR) «Profiling of the author of the written text» [N 13-06-00016].

REFERENCES

1. K.R. Scherer and H. Giles (eds). Social Markers in Speech. Cambridge University Press, 1979, pp. 148-151.

2. J. Pennebaker. Secret Life of Pronouns: What Our Words Say About Us. Reprint edition. Bloomsbury Press, 2011. 368 pp.
3. Sh. Argamon, S. Dhawle, M. Koppel and J.W. Pennebaker. Lexical predictors of personality type. Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America, Theme: Clustering and Classification (8-12 June 2005. St. Louis, MO). http://eprints.pascal-network.org/archive/00001492/01/argamon-etal-csna.pdf.
4. Sh. Argamon, M. Koppel, J.W. Pennebaker and Jonathan Schler. Automatically profiling the author of an anonymous text. Commun. ACM, 52(2): 119-123 (2009).
5. S. Argamon and S. Levitan. Measuring the usefulness of function words for author-ship attribution. Proc. of the 2005 ACH/ALLC Conference, pp. 23-31 (2005).
6. J-M. Dewaele and A. Furnham. Extraversion: The unloved variable in applied linguistic research. Language Learning, 49 (3): 509-544 (1999).
7. A.M. Stolyarenko (ed). Prikladnaya Yuridicheskaya Psikhologiya [Forensic Psychology]. Moscow, YuNITI-DANA, 2001, pp. 399-406.
8. E.I. Galyashina. Osnovy Sudebnogo Rechevedeniya [The principals of Forensic Linguistics]. Moscow, STENSI, 2003, p. 32.
9. O.V. Zagorovskaya, T.A. Litvinova, O.A. Litvinova. Elektronnyy korpus studencheskikh esse na russkom yazyke i ego vozmozhnosti dlya sovremennykh gumanitarnykh issledovaniy [Electronic corpus of student essays and its applications in modern humanity studies]. Mir nauki, kul'tury i obrazovaniya [World of science, culture and education], 3(34): 387-389 (2012).
10. I.M. Luyckx and W.Daelemans. Shallow text analysis and machine learning for authorship attribution. Computational Linguistics in the Netherlands 2004. Selected papers from the Fifteenth CLIN Meeting. pp. 149-160 (2005).