ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 8(S9), 93–97 May 2015*
DOI: 10.17485/ijst/2015/v8iS9/51103

Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study

T. A. Litvinova[1]*, P. V. Seredin[2] and O. A. Litvinova[3]

[1]Regional Centre for Russian Language, Voronezh State Pedagogical University, Voronezh, Russia; centr_rus_yaz@mail.ru
[2]Depatment of Solid State Physics and Nanostructures, Voronezh State University, Voronezh, Russia; paul@phys.vsu.ru
[3]English Language Department of Voronezh State Pedagogical University, Voronezh, Russia; olga_litvinova_teacher@mail.ru

Abstract

Objectives: The aim of the paper is to examine the efficiency of using of parts-of-speech (POS) bigram frequencies to address the problem of personality prediction from text. **Methods/Analysis:** 96 texts were used for the study which were randomly selected from a corpus of Russian students' essays "Personality". Using NLP methods frequencies of POS bigrams in each text (227 types of bigrams were overall identified) were computed, bigrams were then selected which are found in no less than 75% of the analyzed texts. Correlations between POS bigrams frequencies in texts and author gender and personality traits are computed. **Findings:** Some researchers report consistently positive contribution of POS n-grams in construction of models for personality prediction from text. But these conclusions were drawn based on the analysis of English-language texts. Our finding confirms the efficiency of POS bigrams in predicting personality from texts in Russian. Correlations of POS bigrams frequencies and characteristics of the authors of the texts (gender, personality traits measured with McCrae and Costa questionnaire) were computed. Gender was correlated with frequency of prep_noun bigram in the text, males typically score more on this parameter. For neuroticism the correlations were identified with adj-noun, noun-prep, prep-noun bigrams. There were also correlations with the Openness parameter and frequencies of noun-prep bigram. There were also weak correlations between the scores on Extraversion and frequencies of *pers-vfin* bigram, the scores on Agreeableness and frequencies of *pers-vfin* bigrams and *ptcl-vfin* bigrams. The paper points out that the resulting dependencies should be interpreted based on psychology, psycholinguistics and neurolinguistics data. **Novelty of the study:** To our knowledge, the current study has been the first one to deal with the frequencies of certain POS bigrams as parameters for written text author profiling for Russian-language texts. **Conclusion/Application:** The study conducted on Russian texts confirmed early finding in English regarding usefulness of POS n-grams in author profiling task. Further investigations on different text corpora are needed.

Keywords: Author, Authorship Attribution, Corpus Linguistics, Part-of-Speech Bigrams, Personality Prediction from Text, Text

1. Introduction

Due to recent rapid development of automatic language processing (morphological, syntactical parsers, etc.), statistical data processing for studying written text authorship (author profiling based on a text produced), stylometric approach has been extensively used: based on large corpus materials using the methods of statistical data processing, correlations between quantitatively assessed text parameters and their authors' personalities are computed and mathematical models are designed based on the resulting correlations. The ultimate objective of studies of this kind is automatic categorization of texts based on the initial parameters (author's gender, age, etc.) using analyzing software. Author profiling is a problem of growing importance in applications in forensics, security, and marketing. Currently, this field of research is in active progress. For example, author profiling task has recently (in 2013) been proposed in PAN which is a yearly workshop and evaluation lab PAN on uncovering plagiarism, authorship, and social software misuse[1]. Author profiling at PAN-

*Author for correspondence

2013 has been focusing on gender and age identification in social media, both in English and Spanish[2]. With 21 teams, author profiling was one of the most popular tasks at the CLEF conference in 2013. Participants took diverse approaches to the problem: content-based, stylistic-based, n-gram based, etc. Apart from this task, another two tasks were organized in 2013 on predicting different aspects of an author's demographics: specifically, personality traits and native language. This shows the increasing interest of the research community in author profiling[1]. In 2014, PAN was dedicated to the task on author profiling in social media, as well as tasks on author identification and plagiarism detection[3]. Author profiling task in PAN-2015 is about predicting an author's demographics (age, gender and personality traits) from tweets in English, Spanish, Italian and Dutch to predict age, gender and personality traits[4]. Accuracy of author profiling is much lower than 100% but in some research particularly dealing with gender profiling, it can be 80% accurate[5,6]. To our knowledge, no software of this kind has been developed for the Russian language material. In the paper[7] there are the results of a pilot study carried out in order to identify the correlations between formal and grammatical parameters of the text (parts of speech correlations, lengths of words, sentences, etc.) and author's personality (gender; personality tests scores) based on a specially developed corpus of Russian students' essays provided with metatags presenting the information about its authors using statistical data processing methods. The dependencies were obtained that connect the numerical values of formal and grammatical characteristics of the text and its author's personality. To our knowledge, this has been the first attempt in Russia at designing complex prediction models allowing for several written text parameters at a time and employed for gender prediction and some personality traits of the author of a particular text. Overall, our method has proved to be efficient. Our hypothesis on the significance of frequencies of functional words and pronouns for author profiling has also been proved: the analysis of the frequencies of this type of words has been found to be efficient for author profiling. The correlations between the personality traits of the authors and syntactical parameters of the text have also been identified. At the moment, the syntactical level is not capable of being made automatic; therefore there are a limited number of parameters to be studied: 1) number of clauses; 2) number of composite sentences; 3) number of subordinate clauses; 4) number of independent, compound and complex clauses. Note that we intentionally avoided analyzing texts on the vocabulary level as our purpose was to find formal and grammatical parameters of texts correlating with certain personality traits. This As the analysis of scientific literature suggests, these are frequencies of sequences (bigrams) of parts of speech. The research using English-language materials has proved the analysis of the frequencies of different bigrams in texts to be efficient in authorship attribution, see, e.g[8]. It was assumed that N-grams of parts of speech can be "efficient at coding syntactical information and can therefore be used for text classification"[6]. Parts-Of-Speech (POS) sequences as a text parameter are used in forensic author identification using Russian-language materials[9], particularly for the identification of heterogeneous parts which are supposedly by a different author, i.e. borrowings. Dividing texts into homogeneous parts rests on the assumption that there is a kind of a subconscious manner of realizing grammar connections in a speech flow. The authors set forth the algorithm for searching the parts "with a different syntagmatics with a certain sequence of composition elements - words with POS characteristics"[9]. But POS n-grams are used not only for author identification, but also for author profiling. To our knowledge, Oberlander and Gill[10] were the first who introduced features consisting of the POS collocations (mostly 2-grams). Kim Luyckx and her collaborator Walter Daelemans extracted POS n-grams as features to predict Big Five personality traits[11]. They suspect that syntactic features would be more predictive because they are not controlled so consciously as the use of individual words[12]. Wright, Chin[13] trained support vector machine to classify the Five Factor personality. His set of features included bag of words, essay length, word sentiment, negation count and part-of-speech n-grams. He reports of consistently positive contribution of POS n-grams. To our knowledge, the current study has been the first to deal with the frequencies of certain POS bigrams as profiling parameters for written text author profiling for Russian-language texts.

2. Material and Methods

96 texts were used for the study which were randomly selected from a corpus of texts of students' essays (for more on the corpus see[14]). All the texts were examples of natural written speech (composition on the topic "What Would I Do if I Won a Million Dollars?"; description of a picture, etc.). Apart from texts, the corpus also contains the

information about the authors - gender, results of the tests revealing the Big Five personality traits by McCrae and Costa interpreted by A.B. Khromov[15], scales: 1) extraversion; 2) agreeableness; 3) conscientiousness; 4) neuroticism; 5) openness. Natural Language Processing (NLP) methods (free online Xerox morphological analyzer for Russian[16] was used) were employed for computing the frequencies of POS bigrams in each text (227 types of bigrams were overall identified), bigrams were then selected which are found in no less than 75% of the analyzed texts. These are *adj-noun, cm-conj, conj-noun, det-noun, noun-cm, noun-conj, noun-noun, noun-prep, noun-sent, noun-vfin, pers-vfin, pers-pers, prep-adj, prep-noun, prep-pers, ptcl-vfin, sent-noun, sent-pers, vfin-cm, vfin-vfin, vfin-noun, vfin-prep*. The proportion of each bigram in the texts was computed (the number of bigrams was divided into the total number of words in the text). Following that, traditional mathematical methods were applied to identify the correlations between each of the text parameters which were the proportions of the most frequent bigrams and personality traits which were made numerical (gender: female - 0, male - 1, scores on 5 scales of the test). The analysis took several stages. At the first one, by means of the correlation analysis using Pearson's criteria the number of factors connected into a stringently determined system "text parameter-personality parameter" was identified and the credibility of all the characteristics of the correlation connection with the connection p = 0.05. Furthermore, it was assumed that a type of connection (type of analytical function) between numerical values of the text parameters and author's personality traits is linear. At the third stage, by means of regression analysis method of SPSS software, the initial regression equations were identified and the resulting parameters of the equations were analyzed in order to identify the errors of the established laws on a test selection.

3. Results

The only bigram found to have a significant correlation with the gender of the author of the text is *prep_noun* bigram. Its Pearson's correlation coefficient is 0.215. Considering that the selection is N = 96, a number of degrees of freedom is N-2 = 94. The critical correlation coefficient for a level of significance p = 0.05 is 0.205. Therefore, it can be stated that there is weak linear connection between the proportions of prep_noun bigrams in the text and the gender of its author, males typically score more on this parameter. Selecting different types of linear functions revealed that this dependence is most accurately described by a four-parameter logical regression as f = D + (A-D)/(1+10^((x-logC)*B)), where A = 1.0378, B = 101.8044, logC = 0.0750, D = 0.3728, x is the proportion of prep_noun in a text. The model was tested on test set (texts not used for designing the model, 10 written by males, 10 written by females, mean length = 161 word). If calculated values were in range (0, 0.5), we concluded the author was female, if calculated values were in range (0.5, 1), we concluded the author was male. The model was found to be 65% accurate, error rate = 0.27. It also should be noted that the model was considerably better at distinguishing females than males. For neuroticism (test scores) the correlations were identified with *adj-noun* (-0.405; 0.00354), *noun-prep* (-0.414; 0.00282), *prep-noun* bigrams (-0.322; 0.0225). A regression model was designed where y is a predicted number of scores on the neuroticism scale:
y = 65.086 - (26.071•*adj-noun*) - (123.534•*noun-prep*) - (108.884•*prep-noun*). The results of the tests of the model on the test selection indicated its high level of accuracy: accuracy was 79%, error rate was 4 points. There were also correlations with the Openness parameter and frequencies of noun-prep bigram (-0.506; 0.000178). The following model was designed: y = 58.264 - (281.926•*noun-prep*).
According to the results of its tests, accuracy was 88 %, error rate was 2.5 points. There were also weak correlations between the scores on Extraversion and frequencies of *pers-vfin* bigram (0.304; 0.0320), the scores on Agreeableness and frequencies of *pers-vfin* bigrams (0.297; 0.0359) and *ptcl-vfin* bigrams (-0.321; 0.0229). However, these correlations are not strong to be used for designing regression models. The results of the study are summarized in Table 1.

4. Discussion

The efficiency of POS bigrams as the parameters used for written text author profiling has for the first time been proved for Russian-language materials. It still remains unclear, though, how certain correlations between the frequencies of particular language tools and personality traits are accounted for. In our opinion, a globally developing method for author profiling making use of the latest statistical methods and automatic language processing on large corpus materials should keep pace with the latest achievements in psychology, cognitive science and neurolinguistics in order for the resulting correlations to

Table 1. Correlations of POS-bigrams and personality traits

Characteristics of an Author	Bigrams	Correlation Coefficient	Accuracy of Regression Model	Error Rate
Gender	prep_noun	0.215	65 %	0.27
Neuroticism	adj-noun noun-prep prep-noun	-0.405 -0.414 -0.322	79%	4 points
Openness	noun-prep	-0.506	88 %	2.5 points
Extraversion	*pers-vfin*	0.304	-	-
Agreeableness	*pers-vfin* *ptcl-vfin*	0.297 0.321	-	-

be given a theoretical explanation. The current research employing the methods of neurovisualization of the brain in addressing certain issues found that taxonomic composition of words (i.e. the quantitative correlation of parts of speech in a text) is influenced by the activities in the right or left hemisphere. Particularly, as Sedov[17] reported, the deterioration of the left hemisphere and the activity of the right one respectively cause a reduction in the number of functional words as well as verbs and pronouns, while there is an increasing number of nouns and adjectives. The right hemisphere was also found to be responsible for reference functions and language awareness, it "stores" deictic elements: pronouns (particularly, demonstrative), adverbs (there, here, etc.), particles (here), for more see[17]. At the other end of the scale, there is currently enough data to suggest that a lot of psychological states have a neurobiological origin, e.g., are associated with the activities of the right or left hemisphere. A shift in the balance of interhemispheric activation for the right hemisphere is known to be caused by negative emotional environments, for more see[18].

That would be logical to assume that in order to design more efficient complex models for written text author profiling it is necessary that as early as when correlations are being identified, text parameters are selected not merely based on the intuition but the latest data of domestic and foreign neurolinguistics which specifies what parts of the brain are associated with particular elements of a text on one hand and the latest developments of cognitive science on neurobiological origins of certain personality traits on the other.

5. Conclusion

The research conducted for author profiling on the corpus of Russian language written texts shows that there are statistically significant stable correlations between the frequencies of POS bigrams and personality traits. Although obtained regression models are dependent on dataset, they show that using POS bigrams in author profiling task can be useful and increase accuracy of prediction models. Further investigations on different text corpora are needed.

Obtained correlations can be accounted for applying the latest neurolinguistics and neuropsychological data. An approach to text author profiling can therefore be called neurocognitive and in conjunction with the study of large corpus materials using the methods of language processing and mathematical statistics can yield new data on the connection of language and mind and in particular the way certain personality traits are reflected in a text.

6. Acknowledgement

The study was supported by the grant of Russian Fund of Fundamental Research, project N 13-06-00016 "Modeling the Personality of the Author of a Written Text", grant of Voronezh State Pedagogical University, and project N 11 (2014).

7. References

1. Rosso P, Rangel F. Uncovering Plagiarism - Author Profiling at PAN. Available from: http://ercim-news.ercim.eu/en96/ri/uncovering-plagiarism-author-profiling-at-pan
2. Rangel F, Rosso P, Koppel M, et al. Overview of the Author Profiling Task at PAN 2013. In: Forner P, Navigli R, Tufis D, editors. Working Notes Papers of the CLEF 2013 Evaluation Labs. 2013. Available from: http://www.clef-initiative.eu/documents/71612/2e4a4d3a-bae2-47f9-ba3c-552ec66b3e04

3. Rangel F, et al. Overview of the 2nd Author Profiling task at PAN 2014. Available from: http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-papers-final/pan14-author-profiling/rangel14-overview.pdf
4. PAN 2015. Available from: http://pan.webis.de/
5. Argamon SH, Koppel M, Pennebaker JW, Schler J. Automatically profiling the author of an anonymous text. Comm ACM. 2009; 52(2):119–23.
6. Argamon SH, Koppel M, Fine J, Shimoni AR. Gender, genre, and writing style in formal written texts. Text. 2003; 23(3): 321–46.
7. Litvinova TA. Profiling the author of a written text in Russian. J Lang Lit. 2014; 5(4):210–6.
8. Keselj V, Peng F, Cercone N, Thomas C. N-gram-based author profiles for authorship attribution. Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING'03). 2003. Available from: http://web.cs.dal.ca/~vlado/papers/pacling03.pdf
9. Sedov AV, Rogov AA. Analiz neodnorodnostej v tekste na osnove posledovatel'nostej chastej rechi [Analysis of Inconsistences in a Text Based on POS Sequences]. Sovremennye problemy nauki i obrazovanija. Filologicheskie nauki [Modern Problems of Science and Education. Philology]. 2013. Available from: http://www.science-education.ru/107-r8339.
10. Oberlander J, Gill AJ. Language with character: A stratified corpus comparison of individual differences in e-mail communication. Discourse Process. 2006; 42(3):239–70.
11. Luyckx K, Daelemans W. Using syntactic features to predict author personality from text. Proceedings of Digital Humanities 2008 (DH 2008). 2008. p. 146–9.
12. Stamatatos E, Fakotakis N, Kokkinakis G. Computer-based authorship attribution without lexical measures. Comput Humanit. 2001; 35(2):193–214.
13. Wright WR, Chin DN. Personality profiling from text: introducing part-of-speech n-grams, user modeling, adaptation, and personalization. Lect Notes Comput Sci. 2014; 8538:502–7.
14. Zagorovskaya OV, Litvinova TA, Litvinova OA. 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]. 2012; 3(34):387–9.
15. Khromov AB. Piatifaktornyi oprosnik lichnosti [The Five-Factors Personal Inventory]. Kursk, 2000.
16. Xerox morphological analyzer. Available from: https://open.xerox.com/Services/fst-nlp-tools/Consume/Part%20of%20Speech%20Tagging%20(Standard)-178 17. Sedov KF. Neiropsikholingvistika [Neurolinguistics], Moscow: Labirint; 2007.
18. Yegorov AY. O narushenii mezhpolusharnogo vzaimodejstvija pri psihopatologicheskih sostojanijah [On the anomalies of interhemispheric relations associated with psychopathology], Zhurnal Evolyutsionnoi Biokhimii i Fiziologii [J Evol Biochem Physiol]. 2003; 39(1):41–52.