Gender identification in Russian written texts

Tatiana A. Litvinova – Pavel V. Seredin – Olga A. Litvinova – Olga V. Zagorovskaya

DOI: 10.18355/XL.2017.10.03.14

Abstract

This article examines the identification of the gender of authors of Russian written texts using the quantitative parameters analysis approach. Identification of the gender of authors of texts is viewed as part of authorship profiling task.

The material used for the study was a specially designed corpus of Russian texts "RusPersonality" which (along with other Slavic languages) has obtained little attention in authorship profiling studies. We made use of high-frequency text parameters occurring in texts of diffetent topics and genres. The correlation analysis data obtained using Russian texts were compared with those in other languages. The regression analysis was employed. The suggested approach allows one to identify gender as accurately as 64% using only 5 parameters.

Key words: corpus, corpus linguistics, gender attribution, authorship profiling, stylomentry, Russian language

1. Introduction

For decades scientists have studied the differences between male and female speech. These studies indicate a number of differences in the style of writing and highlight the possibility of identifying gender using written texts. However, these studies also argue that all of these differences are not inventory but rather probabilistic, as they manifest themselves in certain features of language use, both qualitatively and quantitatively. However, in order to identify the gender of an author using his/her text, special methods of analysis are necessary, as shown by countless studies. For example, Mulac and Lundell (1994) revealed that people are capable of identifying the gender with a 50% accuracy, i.e. at the level of a random value. Studies concerning the development of methods to identify the gender of a text's author do not only have a practical importance (in marketing, forensics, etc.); indeed, they also have a theoretical significance as they allow one to identify the cognitive activity of males and females which is manifested in their language use. Indeed, this gives a wider insight into human cognitive abilities.

Gender identification is a part of authorship profiling problem which is now gaining momentum as an interdisciplinary subject (Argamon et al., 2009). Authorship profiling is "the task of determining information about the background of the author of an anonymous text based on the language of the text" (Nini, 2014: 13). Some studies report an accuracy as high as 80% and more in identifying the gender of a text's author (see below for more detail). However, there are still many issues which must be addressed. Most of the previous research concerns English texts, although recent times have seen the emergence of studies on other languages. Scientists are still divided on what mathematical methods should be used for this purpose. The main issue is selecting the text parameters to analyse. Content-based features are considered the most effective, although it is obvious that they are consciously controlled. In addition, the obtained mathematical models can be applied only for the text corpora based on which they were originally designed. Studies employing style-based – (lexical, syntactic, character, etc.) parameters do not normally provide any explanation of the correlations between the parameters of texts and the gender of their authors.

The objective of this paper is to design mathematical models to identify the gender of authors of Russian texts. The novelty of the research is found in the following:

1) a specially designed corpus of Russian texts is used, which (along with other Slavic languages) has never been thoroughly investigated as part of authorship profiling studies;

2) only high-frequency text parameters are employed which occur in texts of all topics and genres;

3) a possible explanation of the obtained correlations between the text parameters and an author's gender is provided;

4) an original mathematical solution is set forth.

2. Gender attribution as a task of authorship profiling

Approaches that were later grouped under the name authorship profiling date back to the 2002 seminal paper «Automatically Categorizing Written Texts by Author Gender» (Koppel et al., 2002). This was the first time text parameters had occurred in texts of any topic and genre that proved to be efficient for gender identification (frequencies of 405 most common function words; POS frequencies and frequencies of their most common bigrams and trigrams) had been used to design models for identifying the gender of authors of texts. The study was conducted using the British National Corpus (BNC). For certain genres, the accuracy rate was approximately 80%, while for literary texts the accuracy was 79.4%. Indeed, the frequencies of function words were used even if the number of parameters was reduced to 8 (both for men and women). Data were obtained concerning the frequencies of the use of certain function words (e.g. a and the, which turned out to be men's "favourites", while she, for, with, and not were preferred by women). Also identified were the differences between the use of other POS in the language of men and women in texts of different functional styles and genres. Therefore, it was concluded that genres of texts must be taken into consideration while developing methods of identifying the gender of the texts' authors. This research discovers sound evidence regarding the significance of frequency characteristics of some parameters of written texts in English. These are related to the grammar of the used language units and help to identify the gender of the texts' authors.

Scientists have continued to conduct studies aimed at identifying the gender of an author of a written text based on special mathematical models and a set of formal parameters. In a study by Schler et al. (2006), which analysed texts from 71 000 blogs, the differences between the use of some parts of speech by men and women were confirmed. In order to design predictive models, more than 1000 formal text parameters were employed, including the frequencies of certain content words, the proportions of certain (most frequent) function words, and vocabulary units that are typical of certain genres (lol, haha, ur, etc.). The accuracy of identifying the gender was 80 %. Another interesting study is that of Newman et al. (2008) which examined texts from essays, diaries, descriptions of pictures, and transcripts of everyday conversations; a total of 46 million words were analysed (from 11609 respondents). Indeed, this study also indicated that the major difference between the speech of men and women is the frequency of use of some parts of speech. According to the authors, it is in everyday conversations where speakers are free in their language use that gender differences are most pronounced.

Scientists are currently involved in studying the influence of gender on the characteristics of internet users, and there have been contests to find the most effective methods of identifying gender (Rangel et al., 2015). Different groups of parameters can be retrieved from a text using different NLP tools (content-based features – bag of words, words n-grams, dictionary words, slang words, ironic words, sentiment words, emotional words and style-based features – punctuation, POS, emoticons, etc.). All of these methods were originally developed for English texts. In 2015 PAN shared task about predicting an author's demographics from her writing was organized (Rangel et al., 2015). It was suggested that they were trained and tested on English, Spanish, Italian and Dutch tweets. Teams from 18 countries participated. The method developed by a group of Mexican scientists came first (84% accuracy of gender identification). However, as Company and Wanner (2014) rightfully argue, nearly all state-of-the-art works in the area still very much depend on the datasets they were trained and tested on, since they heavily draw on content features, mostly a large number of recurrent words or combinations of words extracted from the training sets. Generally speaking, as suggested by a thesis paper from A. Nini entitled «Authorship profiling in a forensic context» (2014) and our review of current scientific literature, existing methods of identifying the gender of authors of written texts (mainly tested on English texts) are still not quite efficient in actual practice.

The current, and commonly held belief, is that gender differences in speech are the result of a combination of biological, psychological and social factors (Blum, 1997; Kaiser et al., 2009: Miller and Halpern, 2014). It is obvious that it is impossible to design reliable methods of identifying demographic characteristics of authors of texts to be further used in practice, including in analysis of Internet texts, unless all of these are taken into account.

3. Empirical study: gender attribution in Russian languge written texts

Corpus. This study utilised a specially designed corpus designed for authorship profiling study *RusPersonality* as well as a constantly growing text corpus (Litvinova, 2014; Litvinova et al., 2015), both of which contained, aside from the texts themselves, metadata with information about the authors (gender, age, education, psychological testing data, etc.). The corpus currently contains more than 2000 texts obtained from more than 1 000 respondents, including descriptions of pictures and a letter to a friend. Average text length is 130-160 words.

We selected only those authors who chose to write two texts. All the authors of chosen texts are students of Russia’s largest universities and they are all native speakers of Russian.

Each text from a male author with specific topic and genre should be matched by a text in the same topic and genre from a female author. The total number of texts was 1112 with 112 chosen for testing the models and 1000 for designing them.

Methods. All of the texts were marked using a script based on a morphological analyser pymorphy2 and processed using an online service *istio.com*. The text parameters were only those that were not consciously controlled (finite forms of verbs and other clear indicators of an author’s gender were not considered for the above): indicators of lexical diversity of a text, proportions of POS, and different POS ratios (a total of 78 parameters).

In order to determine the characteristics of and type of connection between the text parameters and individual characteristics of the author, a correlation analysis was performed ($p < 0,05$) using the software SPSS Statistics. A large number of the parameters of the texts and the gender of their authors correlated with the Pearson coefficient r 0,25-0,39. Indeed, this allowed us to design a regression model considering the most significant correlations based on multiparameter linear approximation. However, testing of the quality of the models showed that this type of approximation yields a low level of accuracy as the parameters of texts by individuals of different gender are usually in overlapping ranges (see Fig.). This makes it impossible to design a functional model as part of a multiparameter regression. Therefore, it was decided to use not only a multiparameter regression model as we did in Litvinova (2014) and Litvinova et. al. (2015), but to design a few regression models instead. For each text parameter a regression model was designed. Let us show the suggested approach using an example of 5 parameters with the highest r:

1) TTR (type-token ratio) which is the most commonly used index of lexical diversity of a text (Hardie and McEnery, 2006). Given a text t, let Nt be the number of tokens in t and Vt be the number of types in t, then the simplest measure for the TTR of the text t is:

$$TTRt = Vt/Nt \qquad (1)$$

Note that the measure in eq. (1) is a number defined in [0, 1], since for any text results $1 \leq Vt \leq Nt$. Some interesting attempts to improve the TTR index have been proposed in the literature, although only a few of these variants possess key properties that are essential if they are to be used in our text comparison and these properties are harder to calculate (see Caruso et al., 2014 for details).

Since the texts in the corpus were of a different length, we calculated TTR in the first one hundred words of each text. Indeed, TTR-value is known to depend on the length of the analysed text and therefore the comparison of values makes sense at the same number of tokens (Caruso et al., 2014: 139).

The index was calculated using *istio.com*. The correlation coefficient was r = 0,390.

The resulting regression equation takes the following form:

$$GENDER_1 = -0{,}669 + (2{,}622 * TTR) \qquad (2)$$

2) Formality of a text which was calculated using the following formula (Nini, 2014):

$$F = (noun + adjective + preposition - pronoun - verbs - participles - - adverbs - conjunction - interjections) + 100)/2. \qquad (3)$$

The correlation coefficient with the gender of an author for this parameter is r = 0,315.

The regression equation is as follows:

$$GENDER_2 = -0{,}637 + (0{,}971 * Formality) \qquad (4)$$

3) Percentage of prepositions and modifiers (pronoun-like adjectives) in a text (r = 0,243):

$$GENDER_3 = -0{,}188 + (0{,}0432 * preposition+pronoun\text{-}like\ adjective) \qquad (5)$$

4) Percentage of the 100 most frequent Russian words in a text (Lyashevskaya, 2009), r = -0,322.

The regression equation is as follows:

$$GENDER_4 = 1{,}500 - (0{,}0303 * Frequent\ ones) \qquad (6)$$

5) The index of the lexical density: a ratio of function words to content words in a text multiplied by 100 % (Garcia and Martin, 2007; Nini, 2014), (r= -0,295):

$$GENDER_5 = 1{,}392 - (0{,}0229 * Function) \qquad (7)$$

In order to properly estimate the obtained result, let us determine the average arithmetic values from the solution of the five equations:

$$GENDER = \frac{\sum_{i=1}^{5} GENDER_i}{5}. \qquad (8)$$

Let us assume that a design value in the range [0; 0,499] indicates that the author of a text is female and in the range [0,500; 1] shows that they are male.

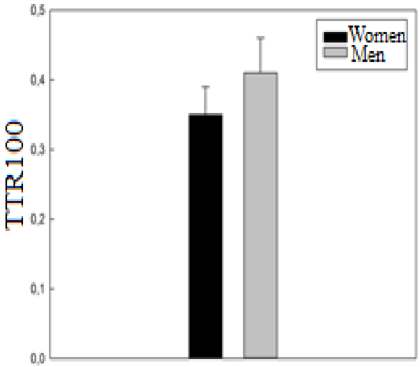
Results. The analysis showed that in Russian written texts by men compared to those by women, the index of lexical diversity and the proportion of prepositions and pronoun-like adjectives are higher as well as level of formality (see Fig.)

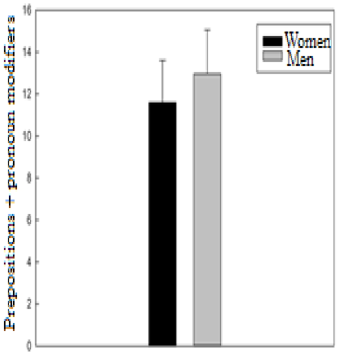
Overall, the data are in good agreement with the results obtained for English texts. Hence, as noted above, many scientists argue that texts by men have on average more nouns and adjectives as well as prepositions and determiners; in contrast, those by women have more verbs and personal pronouns (see a detailed review in Nini (2014)). According to the literature, this is indicative of profound cognitive differences in the linguistic profiles of men and women: reporting is more important for men while rapport is more significant for women; therefore, texts by men seem more “formal” and those by women more “contextual” (see Heylighen and Dewaele, 2002 for more detail). It is interesting to compare this with the paper by Saily et al. (2011), which shows that the prevalence of nouns in texts by men as opposed to pronouns in those by women was common in personal letters written in English from 1415 to 1681. Indeed, this shows that the above gender differences are universal.

In a paper by Nini (2014: 132) it was shown that “the more personal a text becomes the less likely it is to show a gender pattern of the rapport/report type. In other words, in a register in which individuals are already pressed to be Involved and person-centred then there is no room for variation between rapport and report discourse, thus blocking the gender pattern from emerging”. However, this effect, as we suggest, is retained in personal texts such as letters to a friend.

We argue that a higher index of lexical diversity in texts by men is due to the above differences: in “male” texts there are fewer most frequent words, the majority of which are function words; in addition, there are fewer repetitions, and more unique vocabulary units occurring in a text at one time.

Let us determine the accuracy of the approach. Accuracy, in this context, is the ratio of the number of test documents that were correctly predicted to the total number of test documents. The calculations suggest that gender was correctly identified in 65% of women and 63% of men. Thus, the accuracy of the approach was 64%.





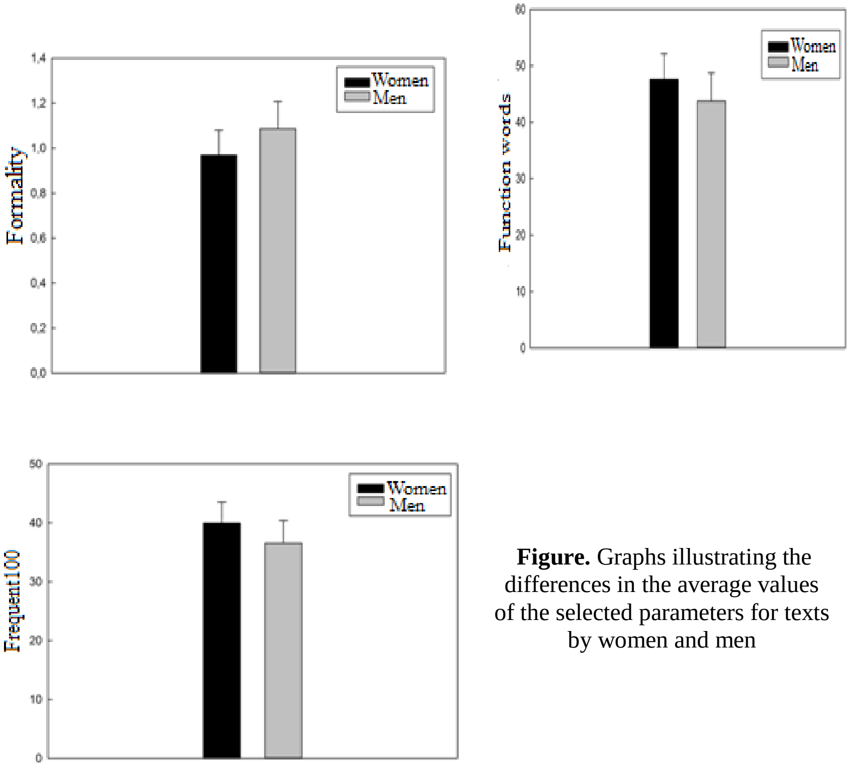


Figure. Graphs illustrating the differences in the average values of the selected parameters for texts by women and men

3. Conclusions

This is most certainly a pilot study, and its conclusions are not final. The suggested approach is considered to be improved by expanding a list of text parameters and more in-depth optimisation of selections used in designing regression models.

In accordance with Chambers (1992), in the future sociological gender as well as biological gender should be considered as independent variables. In addition, it is essential to analyse the gender characteristics of authors of texts depending on a range of personality traits and femininity/masculinity, profile functional cerebral asymmetry, etc. As correctly pointed out by Nini (2014: 34), it can be assumed that “the real differences in the linguistic patterns adopted by people depend on their personality and/or hormone levels and that genders are different to the extent that on average different genders are prone to different personality orientations and/or hormone levels”. This analysis to be conducted during further research would allow one to develop a more current and deeper insight into the way gender is manifested in written texts and to develop more accurate methods of identifying the gender of individuals based on the quantitative parameters of their texts.

Acknowledgment

This research 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”.

Bibliographic references

ARGAMON, S. – KOPPEL, M. – PENNEBAKER, J. – SCHLER, J. 2009. Automatically profiling the author of an anonymous text. In: Communications of the ACM, vol. 52, n. 2, pp. 119–123. ISSN 0001-0782.

BLUM, D. 1997. Sex on the Brain: The Biological Differences between Men and Women. NY: Viking Press. ISBN 030647770.

CARUSO, A. – FOLINO, A. – PARISI, F. – TRUNFIO, R. 2014. A statistical method for minimum corpus size determination. In: Proceedings of JADT 2014, pp. 135-146. ISBN 978-2-9547781-1-2.

CHAMBERS, J. K., 1992. Linguistic correlates of gender and sex. In: English World-Wide, vol. 13, n. 2, pp. 173–218. ISSN 0172-8865.

COMPANY, J.S. – WANNER, L. 2014. How to Use Less Features and Reach Better Performance in Author Gender Identification. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, pp. 1315-1319. ISBN 9782951740884.

GARCIA, A. M. – MARTIN, J. C. 2007. Function words in authorship attribution studies. In: Literary and Linguistic Computing, vol. 22, n. 1, pp. 49-66. ISSN 1477-4615.

HARDIE, A. – McENERY, T., 2006. Statistics. In: BROWN K. (ed.). Encyclopedia of Language and Linguistics, 2nd edition. Amsterdam: Elsevier, pp. 138-146. ISBN 9780080448541.

HEYLIGHEN, F. – DEWAELE, J. 2002. Variation in the contextuality of language: an empirical measure. In: Foundations of Science, vol. 6, pp. 293-340. ISSN 1233-1821.

KAISER, A. – HALLER, S. – SCHMITZ, S. – NITSCH, C. 2009. On sex/gender related similarities and differences in fMRI language research. In: Brain Research Reviews, vol. 61, n. 2, pp. 49–59. ISSN 0165-0173.

KOPPEL, M. – ARGAMON, S. – SHIMONI, A. 2002. Automatically categorizing written texts by author gender. In: Literary and Linguistic Computing, vol. 17, n. 4, pp. 401-412. ISSN 1477-4615.

LITVINOVA, T. A. – SEREDIN, P. V. – LITVINOVA, O. A. 2015. Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study. In: Indian Journal of Science and Technology, vol. 8, n. 9 [s. l.], pp. 93-97. ISSN 0974-6846.

LITVINOVA, T. A. 2014. Profiling the author of a written text in Russian. In: Journal of Language and Literature, vol. 5, n. 4, pp. 210-216. ISSN 2078-0303.

LYASHEVSKAYA, O. – SHAROV, S.A. 2009. Frequency Dictionary of Modern Russian language (on materials of the Russian National Corpus). Moscow: Azbukovnik. ISBN 978-5-91172-024-7.

MILLER, D. I. – HALPERN, D. F. 2014. The new science of cognitive sex differences. In: Trends Cogn Sci., vol. 18, n. 1, pp. 37-45. ISSN 1364-6613.

Morphological analyzer pymorphy2. URL: https://pymorphy2.readthedocs.io/en/latest/

MULAC, A. – LUNDELL, T. L. 1994. Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. In: Language & Communication, vol. 14, n. 3, pp. 299–309. ISSN 0271-5309.

NEWMAN, L. M. – GROOM, C. J. – HANDELMAN, L. D. – PENNEBAKER, J. W. 2008. Gender differences in language use: An analysis of 14,000 text samples. In: Discourse Processes, vol. 45, n. 3, pp. 211–236. ISSN 0163-853X.

NINI, A. 2014. Authorship Profiling in a Forensic Context. PhD thesis. Aston Uni.

RANGEL, F. – CELLI, F. – ROSSO, P. – POTTHAST, M. – STEIN, B. – DAELEMANS, W. 2015. Overview of the 3rd Author Profiling Task at PAN 2015.

In: CLEF 2015 Labs and Workshops: notebook papers. Toulouse, France. ISSN 1613-0073.

SAILY, T. – SIIRTOLA, H. – NEVALAINEN, T. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. In: Literary and Linguistic Computing, vol. 26, n. 2, pp. 167-188. ISSN 1477-4615.

SCHLER, J. – KOPPEL, M. – ARGAMON, S. – PENNEBAKER, J. 2006. Effects of Age and Gender on Blogging. In: Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, vol. 6, pp. 199-205. ISBN 1577352645.

Words:3518
Characters:22 281 (12,4 standard pages)

Researcher Tatiana A. Litvinova, Ph.D.
The Kurchtov Institute
1 Kurchatov Square
123182 Moscow
Russia
centr_rus_yaz@mail.ru

Researcher Pavel V. Seredin, Doctor of Physics and Mathematics
The Kurchtov Institute
1 Kurchatov Square
123182 Moscow
Russia
paul@phys.vsu.ru

Researcher Olga A. Litvinova
The Kurchtov Institute
1 Kurchatov Square
123182 Moscow
Russia
olga_litvinova_teacher@mail.ru

Professor Olga V. Zagorovskaya, Doctor of Philology
Voronezh State Pedagogical University
86 ul. Lenina
394043 Voronezh
Russia
olzagor@yandex.ru