Differences in type-token ratio and part-of-speech frequencies in male and female Russian written texts

Tatiana Litvinova[1,2], **Pavel Seredin**[1,2,3], **Olga Litvinova**[1,3] **Olga Zagorovskaya**[1,4]
[1]RusProfiling Lab, Voronezh, Russia
[2]The Kurchatov Institute, Moscow, Russia
[3]Voronezh State University, Voronezh, Russia
[4]Voronezh State Pedagogical University, Voronezh, Russia
centr_rus_yaz@mail.ru

Abstract

The differences in the frequencies of some parts of speech (POS), particularly function words, and lexical diversity in male and female speech have been pointed out in a number of papers. The classifiers using exclusively context-independent parameters have proved to be highly effective. However, there are still issues that have to be addressed as a lot of studies are performed for English and the genre and topic of texts is sometimes neglected. The aim of this paper is to investigate the association between context-independent parameters of Russian written texts and the gender of their authors and to design predictive regression models. A number of correlations were found. The obtained data is in good agreement with the results obtained for other languages. The model based on 5 parameters with the highest correlation coefficients was designed.

1 Introduction

Differences in male and female speech have long been of linguists' interest. However, they used to be investigated by means of the qualitative methods and were largely descriptive, whereas these days the quantitative analysis methods are being employed and the goal of the ongoing paper to identify the gender of text authors using numerical values of text parameters. The fundamental paper in the field is the one called "Automatically Categorizing Written Texts by Author Gender" (Koppel et al., 2002). The text parameters were morphological, i.e. context-independent (405 common function words, i.e. pronouns, articles, prepositions, and conjunctions, POS n-grams, n=1,2,3). It was found that even if the number of parameters is reduced to 8 most frequent function words (FW), the classifier shows the accuracy of 80 %. Usefulness of morphological features in gender identification was shown in studies for different European languages (Argamon et al., 2003; Bortolato, 2016; Mikros, 2013; Newman et al., 2008; Rangel and Rosso, 2013; Sarawgi et al., 2011; Schler et al., 2006).

As NLP tools are being employed a lot these days, the list of the text parameters used to identify the gender of text authors has been largely expanded (see Rangel et al. (2016) for review). However, as correctly noted by Company and Wanner (2014), «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 feature». We think that in order to continue improving the gender profiling methods, especially those ones which can be applied in for forensic settings, it is necessary to further explore the associations between text author gender and context independent parameters in different languages, not only Western European ones.

Slavic languages have been underrepresented in authorship profiling studies until now, but recently the problem of gender identification in Slavic languages has been raised. For example, in a recent paper by Sboev et al. (2016) it was shown that using topic independent features gives 86 % accuracy of gender identification, however the paper presents no analysis of the differences between male and female texts.

The aim of this paper is to study the association between topic independent parameters of Russian written texts and the gender of the authors and to design predictive regression models. It should be noted that we deliberately avoid parameters directly indicating author gender (some forms of verbs, etc.) since they are easily imitated.

2 Methods

2.1 Corpus

This study utilised a specially designed corpus designed for authorship profiling studies, *RusPersonality*, which contained, aside from the texts themselves, metadata with information about the authors (gender, age, education, psychological testing data, etc.). All of the texts in the corpus were written in the presence of the researchers in order to prevent borrowings. The texts were manually written and then converted into the digital format preserving the original style. These are all samples of what is called natural written speech. All of the texts contained an average of 130-160 words. The texts are short, which makes the task more daunting, since most stylometric features exhibit authorship quantitative patterns in larger texts (Mikros, 2013) but makes it more similar to those in forensic settings.

Each author was instructed to write one or two texts choosing among topics "A Letter to a Friend", "Description of a Picture", "How I Spent Yesterday", "Why I Am Perfect for this Position (any)", etc. We selected only those authors who chose to write two texts.

All the authors are students of Russia's largest universities and they are all native speakers of Russian. So, it is assumed that participants have similar social and educational background.

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. Then 1000 texts were used to design two subcorpora. In the first one ("joined") made up by texts written by the same author, they were both joined into one and processed as one text (500 texts in total). In the second subcorpus ("separate") each text was processed individually (1000 texts in total). Both subcorpora were processed individually.

2.2 Text processing

All of the texts were processed using morphological analyzer for the Russian language pymorphy2 (https://pymorphy2.readthedocs.org/en/latest/) able to normalize, decline and conjugate words, provide analyses or give predictions for unknown word. Also all of the texts were processed using an online service *istio.com.* The text parameters were only those that were not consciously controlled: indicators of lexical diversity of a text, POS (17 broad categories, see https://pymorphy2.readthedocs.io/en/latest/user/grammemes.html for tagset), different ratios of POS (a total of 78 parameters). While choosing the parameters we stuck to the criteria set forth by Oakes (2014) Firstly, the parameter should be frequent enough so that the results are statistically reliable (we chose only the parameters with the frequency more than 0 in no less than 50 % of the texts). Secondly, the parameter needs to be objectively countable.

2.3 Mathematical analysis

To estimate the association between gender and text parameters, we calculated Pearson's correlation coefficient r (t-tailed) using SPSS Statistics software.

3 Results

A large number of the parameters of the texts were correlated with the gender of their authors with *r* in the range 0.25-0.39 ($p < 0.05$; they are not presented due to lack of space). We have chosen only the parameters that were shown to correlate with the gender of authors in the joined and separate subcorpora and then 5 of them that had the highest averaged r were selected.

1. Type-token ratio (TTR). This 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:

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

Note that the measure in eq. (1) is a number defined in [0, 1], since for any text results $1 \leq Vt \leq Nt$.

Since the texts in subcorpora 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 averaged correlation coefficient r = 0.39.

2. Percentage of the 100 most frequent Russian words divided by text length in words (aver-

aged r = -0.322). The list of the words was taken from Lyashevskaya, Sharov, 2009.

3. The index of formality. It was calculated using the following formula (Nini, 2014):

$$F = (\text{noun} + \text{adjective} + \text{preposition} - \text{pronoun} - \text{verbs} - \text{participles} - \text{adverbs} - \text{interjections} + 100)/2 \quad (2)$$

Averaged r = 0.315.

4. The index of the lexical density. It was calculated as a ratio of function words to content words multiplied by 100 % in a text. It is also known as an index of functional density (Nini, 2014), averaged r= -0.295.

5. Percentage of prepositions and modifiers (so called pronoun-like adjectives, i.e. такой "such", какой "what", всякий "any", мой "my", наш "our", ваш "your", тот "that", этот "this", etc.) (averaged r = 0.243).

For each text parameter a linear regression model was designed. In order to properly estimate the obtained result, let us determine the average arithmetic values from the solution of the five equations:

$$GENDER = \frac{\sum_{i=1}^{5} GENDER_i}{5} \quad (3)$$

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. According to our experiments, this approach proved to be more accurate than using single linear regression model over all of the features in combination.

Let us determine the accuracy of the model. Accuracy, in this context, is the ratio of the number of texts that were correctly classified according to the author gender to the total number of texts. The calculations suggest that gender was correctly identified in 65% of women and 63% of men. Thus, the accuracy of the approach was 64% (averaged accuracy for "joined" and "separate" subcorpora).

4 Discussion

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 modifiers are higher; their texts are more formal (see Figure 1 for details).

Overall, the data are in good agreement with the results obtained for other languages.

A high degree of lexical diversity in male texts was pointed out by Argamon et al. (2003) as well as significantly higher mean word lengths, which was also identified in the study performed by Oschepkova (2003) using Russian texts by different social groups (students and prisoners). Fewer clichés were also found in Russian male speech. 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.

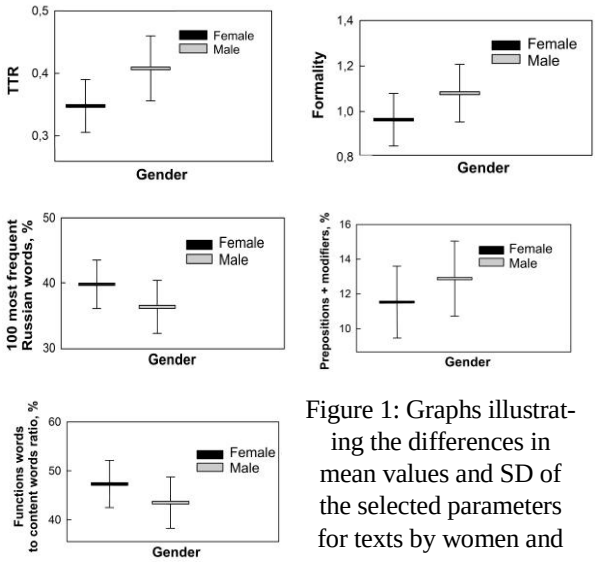

Figure 1: Graphs illustrating the differences in mean values and SD of the selected parameters for texts by women and men

Argamon et al. (2003) found that males use the informational features attributive adjectives and prepositions significantly more often and had significantly higher mean word lengths in nonfiction texts. In fiction texts, men used significantly more nouns and prepositions.

Rangel and Rosso (2013) also observed male preference for prepositions and female preference for pronouns and interjections. A high level of "formality" in male texts was also reported in a large number of studies (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 Säily 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 seem to be universal (see also Johannsen et. al., 2015).

In a paper by Nini (2014) 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" (p. 132). However, this effect is retained in Russian personal texts such as letters to a friend.

As for the ratio of function and content words, it is not commonly employed in studies related to gender identification but is used in other sorts of analysis (García and Martín, 2007). E.g., it was shown to be significant in distinguishing Alzheimer's patients and healthy individuals, i.e. it is indicative of some personal cognitive features (Kernot et al., 2017). As far as gender identification is concerned, using Italian literary texts Bortolato (2016) showed that this parameter is more informative than frequencies of function words (particularly, conjunctions and pronouns) individually.

5 Conclusions

In this paper we have proved that there are differences between male and female texts in a number of morphological indices and TTR level. Some of these differences are in agreement with the previous findings for other languages, which suggests that they are universal. We argue that it is necessary that a list of context-independent text parameters is expanded and Russian texts of other genres are explored.

There are currently plans to account for the relations between the text parameters selected for analysis as well as to apply other methods of statistical analysis.

It is also essential that the parameters that are easily to imitate while pretending to be someone of the opposite sex are investigated. Therefore we have collected a text corpus named Russian Gender Imitation Corpus. Each author was instructed to write three texts on the same topic (out of a list of five) in their natural style, as someone of the opposite sex, someone else of the same sex. Studies of the corpus would enable us to identify which parameters changed while taking on the role of the other gender and which ones persist even during conscious imitation.

In addition, it is essential to analyse the gender characteristics of authors of texts with respect to their personality traits and femininity/masculinity, laterality, etc. As correctly pointed out by Nini (2014, p. 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". Taking this into account, in future it will be useful to treat gender as non-binary category.

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 for forensic settings.

Acknowledgments. This work was supported by the grant of the Russian Science Foundation, project No 16-18-10050 "Identifying the Gender and Age of Online Chatters Using Formal Parameters of their Texts".

The authors would like to thank four anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

Aleksandr Sboev, Tatiana Litvinova, Dmitry Gudovskikh, Roman Rybka, Ivan Moloshnikov. 2016. Machine Learning Models of Text Categorization by Author Gender Using Topic-independent Features. *Procedia Computer Science*, 101, 135-142. https://doi.org/10.1016/j.procs.2016.11.017

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing*, 26(2): 167-188. https://doi.org/10.1093/llc/fqr004

Andrea Nini. 2014. *Authorship Profiling in a Forensic Context.* PhD thesis. Aston Uni. http://publications.aston.ac.uk/25337/1/Nini_Andrea_2015.pdf

Andrew Hardie, Tony McEnery. 2006. Statistics. In: BROWN K. (ed.). *Encyclopedia of Language and Linguistics, 2nd edition*. Amsterdam: Elsevier, pp. 138-146.

Antonio M. García, Javier C. Martín. 2007. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1): 49-66. https://doi.org/10.1093/llc/fql048

Assunta Caruso, Antonietta Folino, Francesca Parisi, Roberto Trunfio. 2014. A statistical method for minimum corpus size determination. In *Proceedings of Proceedings of the Twelfth International Conference*

on Textual Data Statistical Analysis (JADT 2014). JADT.org, pages 135-146.

Claudia Bortolato. 2016. Intertextual Distance of Function Words as a Tool to Detect an Author's Gender: A Corpus-Based Study on Contemporary Italian Literature. *Glottometrics*, 34: 28-43.

David Kernot, Terry Bossomaier, Roger Bradbury. 2017. The Impact of Depression and Apathy on Sensory Language. *Open Journal of Modern Linguistics*, 7: 8-32. https://doi.org/10.4236/ojml.2017.71002

Ekaterina S. Oschepkova. 2003. Written Text Author Identification: Lexicogrammatical aspect. PhD thesis. Moscow State Linguistic Uni. (in Russian).

Francis Heylighen, Jean-Marc Dewaele. 2002. Variation in the contextuality of language: an empirical measure. *Foundations of Science*, 7: 293-340. doi:10.1023/A:1019661126744

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, Benno Stein. 2016. 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In Proceedings of CLEF (Working Notes), ceur-ws.org, pages 750-784. Available at: http://www.clips.ua.ac.be/~walter/papers/2016/rrvdps16.pdf

Francisco Rangel, Paolo Rosso. 2013. Use of language and author profiling: identification of gender and age. In *Proceeding of the 10th workshop on natural language processing and cognitive science* (NLPCS 2013). Marseille, France. Available at: http://users.dsic.upv.es/~prosso/resources/RangelRosso_NLPCS13.pdf

George K. Mikros. 2013. Systematic stylometric differences in men and women authors: a corpus-based study. In Köhler, R. and Altmann, G. (eds.), *Issues in Quantitative Linguistics*, 3, pages 206-223. Lüdenscheid: RAM – Verlag. http://users.uoa.gr/~gmikros/Pdf/Systematic%20stylometric%20differences%20in%20men%20and%20women%20authors.pdf

Jonathan Schler, Moshe Koppel, Shlomo Argamon, James W. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199-205.

Juan S. Company, Leo Wanner. 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, pages 1315-1319.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14.000 text samples. *Discourse Processes*, 45(3): 211–236. http://dx.doi.org/10.1080/01638530802073712

Michael P. Oakes. 2014. *Literary Detective Work on the Computer.* Amsterdam/Philadelphia: John Benjamins Publishing.

Morphological analyzer pymorphy2. URL: https://pymorphy2.readthedocs.io/en/latest/ (in Russian)

Moshe Koppel, Shlomo Argamon, Anat R. Shimoni. 2002. Automatically categorizing written texts by author gender. Lit Linguist Computing, 17(4): 401-412. https://doi.org/10.1093/llc/17.4.401

Olga Lyashevskaya, Sergei Sharov. 2009. *Frequency Dictionary of Modern Russian language (on materials of the Russian National Corpus)*. Moscow, Azbukovnik. URL: http://dict.ruslang.ru/freq.php (in Russian).

Ruchita Sarawgi, Kailash Gajulapalli, Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the fifteenth conference on computational natural language learning (CoNLL '11)*, Association for Computational Linguistics, pages 78-86.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, Anat R. Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Interdisciplinary Journal for the Study of Discourse*, 23(3): 321–346. https://doi.org/10.1515/text.2003.014

Tanja Säily, Terttu Nevalainen, Harri Siirtola. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing*, 26(2): 167-188. https://doi.org/10.1093/llc/fqr004