Russian text corpora for deception detection studies

Tatiana A. Litvinova, Olga V. Zagorovskaya, and Olga A. Litvinova

***Abstract*—Text-based deception detection is presently on the way to gain even more significance as related studies certainly have both theoretical and practical value and a range of applications for police, security, and customs, as well as predatory communications, e.g. Internet scams). For these studies designing text corpora is essential. Text-based deception detection has been mostly dealt with using English as well as a few other European languages. There is not sufficient research into the problem with the use of Slavic languages, which is mostly due to no corresponding corpora available. In this article we propose an overview of existing text corpora employed in studies of text-based deception detection as well as a detailed description of available Russian corpora specially designed for text-based deception detection.**

***Keywords*—corpus of texts, corpus linguistic, text-based deception detection, automated deception detection.**

I. Introduction

The multidisciplinary field of text-based deception detection is currently gaining momentum.

In recent years deception detection has been commonly addressed as a text classification problem employing the methods of natural language processing and data mining [1; 2; 3; 4; 5]. A surge of interest in the field is due not only to the development and improvement of text categorization technology but also to a growing practical demand. With the advance of Web 2.0, there has been an increasing need for methods of identifying texts containing intentionally deceptive information (news, product/service reviews, dating website profiles, etc.). This has resulted in the progress of domain of automated text-based deception detection aimed to work out means for any type of deceptive information to be recognized.

One of the most important data to be employed in this field are text corpora containing information on the truthfulness/deceptiveness of texts. As Enos states, "one of the primary obstacles to research on the automatic detection of deceptive speech has been the lack of a cleanly-recorded corpus of deceptive and non-deceptive speech for use in training and testing" [6, p. 18]. The major challenge facing most researchers in collecting these corpora is the establishment of correctly labeled datasets: we must make sure we know whether a particular text (sentence) is truthful/deceptive.

Deception, i.e. intentionally deceptive information, might involve factual information from a text as well as its author's personality, i.e. their gender, age, etc. Most research dealing with deception detection has been focused on detecting deception about the content of a message but not its author. According to P. Juola, "another form of "deception" can occur when a speaker or writer offers a statement that he or she does not want to be identified with" [7, p. 92]. Along with Juola, we call this "stylistic deception". Note that there are few studies and text corpora respectively dealing with «stylistic deception»,

Presently, text corpora used in deception detection can be grouped into two major classes:

1) those which contain texts produced according to a researcher's instructions;

2) corpora which contain "real" texts produced in situations where the stakes of deception are middle or high, i.e. when there might be severe consequences in case deception is revealed (loss of a job, imprisonment, etc.).

It should be noted that most studies in text-based deception detection have been performed for English and less frequently for Romance languages (e.g., Spanish [8], Italian [9]). Slavic languages have been entirely left out of consideration with rare exceptions [10; 11]. This is what urged us to start collecting corpora of Russian texts specially designed for studies of deception detection.

This article provides an overview of available texts corpora for deception detection, including stylistic deception. The main outcome of the paper is a description of existing Russian corpora for text-based deception detection including information on their composition, structure and make-up. We hope that such work will encourage innovation and further related studies for Slavic languages.

II. Types of Text Corpora in Deception Detection Studies

A. Text corpora collected according to a researcher's instructions

Most studies in automated text-based deception detection have been conducted using text corpora where participants were instructed to produce truthful and deceptive texts so as to avoid the labeling problem [12; 13; 14]. This problem was thought to have been solved as the production of

Manuscript received October, 20, 2017. This work was supported in part of creation of Russian Deception Bank by the Russian Foundation for Basic Research, N 15-34- 01221 "Lie Detection in a Written Text: A Corpus Study", and by the Russian Science Foundation, project No 16-18-10050 "Identifying the Gender and Age of Online Chatters Using Formal Parameters of their Texts", in part of creation of Gender Imitation Corpus.

T.A. Litvinova is with the Voronezh State Pedagogical University, Voronezh, 394071 Russia, and with the Kurchatov Institute, 123098 Moscow, Russia (+7-980-342-00-73; e-mail: centr_rus_yaz@mail.ru).

O.V. Zagorovskaya is with the Voronezh State Pedagogical University, Voronezh, 394071 Russia (e-mail: olzagor@yandex.ru)

O. A. Litvinova is with the Voronezh State Pedagogical University, Voronezh, 394071 Russia, and with the Kurchatov Institute, 123098 Moscow, Russia (e-mail: olga_litvinova_teacher@mail.ru).

deceptive and truthful texts could be controlled. These corpora can employ written and spoken texts.

B. *Corpora of written texts*

Two methods of collecting texts are used in designing corpora of written texts:
- searching for participants online, most commonly using Mechanical Turk;
- collecting texts from "available" respondents. They are commonly university students.

Most researchers in the field dealing with English texts have employed Mechanical Turk, online survey service by Amazon (www.mturk.com), to collect materials for their corpora (e.g., [13]). It is a fairly quick and convenient way to collect data. However, some researchers making use of this approach to data collection have pointed out a few difficulties associated with this method. E.g., using this service Rubin & Conroy [15] asked each participant to produce a detailed, personal story with some elements of deception in it. There was a common tendency found in the different types of tasks that they reported: it was very difficult to encourage respondents to write long texts.

There is also a crowdsourced deception dataset consisting of short open domain truths and lies from 512 users [16]. Seven lies and seven truths were provided from each user. The dataset also includes user's demographic information, such as gender, age, country of origin, and level of education. However, this collection could hardly be called a text corpus as it only contains individual statements.

Another corpus of written texts which is collected using Mechanical Turk is so-called Cross-Cultural Deception corpus [17]. It contains texts by individuals from different countries: US (English), India (English), Mexico (Spanish). Each dataset consists of short deceptive and truthful essays on three topics: opinions on abortion, opinions on the death penalty, and feelings about a best friend (as well as in the paper [13]). It is of interest that Spanish texts could not be collected with the use of Mechanical Turk and that the authors had to create a separate web interface to collect this data, recruiting participants through contacts of the paper's authors. It is to be noted that for all three cultures, the average number of words for the deceptive statements (62 words) is significantly smaller than for the truthful statements (81 words).

Both corpora are freely available on the website of Language and Information Technologies research group, the University of Michigan[1].

CLiPS Stylometry Investigation (CSI) corpus and Russian Deception Bank have to be mentioned as text corpora designed in "laboratory" conditions and can be used to identify linguistic features of deceptive texts.

CSI corpus is an annually expanded corpus of Dutch written texts by university students produced according to the researchers' instructions [18]. It was not designed specifically for investigating deception but contains a subcorpus of deceptive and truthful texts. Each student was asked to write a convincing review (either positive or negative) about a fictional product, thus pretending to know about the product while actually making up the review. The truthful reviews reflect the author's real opinion on an existing product. Thus the subcorpus contains both truthful and deceptive texts of the same author on the same topic. The corpus is available on the CLiPS website[2] and can freely be used for academic research purposes. Currently the corpus contains 323 truthful and 319 deceptive reviews.

The corpus contains data about the authors to enable it to be used in studies into the effects of personality on deception production (gender, age, personality traits, etc.).

C. *Spoken corpora*

Starting with the study by Newman et. [14], speech recorded in laboratory settings has been used in related research. In this study speech of 101 undergraduates, while discussing both their true and false views on abortion, was recorded. Then it was transcribed, and only transcripts were analyzed with no consideration given to acoustic information.

The CSC Corpus [6] is the first spoken corpus designed and collected for the purpose of deceptive speech detection. The corpus contains interviews with thirty-two individuals speaking Standard American English as their first language. It contains high-quality sound to enable acoustic deception cues to be investigated.

The next step was the creation of Multimodal Dataset for Deception Detection [19] which included physiological, thermal, and visual responses of 30 graduate and undergraduate students (all expressing themselves in English) under three scenarios (mock crime, best friend, abortion). The respondents were instructed to respond either truthfully or deceptively, depending on the scenario being run.

We were not able to find any information as for the access to the materials.

Overall, despite the importance of corpora of texts produced according to the conditions of an experiment, they have certain disadvantages. As pointed out by Rubin & Conroy, "motivating participants to write rich, linguistically diverse descriptions remains a considerable challenge" [15, p. 10]. In different types of tasks and different data collection methods deceptive texts written according to a researcher's instructions were found to be shorter. In addition, it is impossible to confirm the truthfulness/deceptiveness of texts through the use of alternative methods thus being forced rather to put one's trust on the participants.

It is obvious that methods of deception detection developed using texts collected in laboratory settings are not quite applicable to those produced in high-stake situations. As correctly pointed out by Fitzpatrick & Bachenko, "high stakes deception cannot be simulated in the laboratory without serious ethics violations" [20, p. 31], and corpora consisting of real-world texts produced in high-stake situations are thus necessary.

D. *Corpora of real-world texts produced in high-stake situations*

Trial records are commonly used as real-world texts produced in high-stake situations. For a long time such materials were not available for wider audiences, but as of

[1] http://lit.eecs.umich.edu/downloads.html#undefined

[2] www.clips.uantwerpen.be/datasets

late, due to the emergence of a variety of Internet resources, there has been a positive increase in availability. However, there are still challenges involved in the collection and labeling of these texts.

Ellen Fitzpatrick and colleagues [20; 21] were one of the few researchers to collect an English text corpus in a natural setting. Narratives were collected from the public domain, e.g. from criminal and legal websites and available police interrogations (almost 35.090 words of narrative), and labeled as deceptive/truthful based on court rulings, etc.

DECOUR [9] is a corpus of hearings held in four Italian Courts where the speakers told lies in front of the judge. As a result, they became the object of particular criminal proceeding for calumny or false testimony where it is shown whether the statements given by the defendant are deceptive. It is due to the final court judgment where lies are specified that each individual utterance of the corpus has been annotated by three coders as true, uncertain or false by the degree of truthfulness (35 hearings by 31 subjects, 6070 utterances in total). We were not able to find any information as for the access to the corpus.

Using deceptive and truthful trial testimonies the first real-life multimodal Deception dataset was designed [22]. The dataset includes 121 short videos (61 deceptive and 60 truthful), along with their transcriptions and gesture annotations. The average length of the videos in the dataset is 28.0 seconds. The data consists of 21 unique female and 35 unique male speakers, with their ages approximately ranging between 16 and 60 years. The corpus is freely available[3].

Apart from forensic data, some researchers also use financial reporting, which can be found for publicly traded companies (at least for the USA) [23]. However, this material is not quite representative in regards to fidelity of the text labeling – there is often not enough information freely available to be able to classify texts as deceptive/truthful.

Public speeches (radio, TV, Internet) by prominent media figures who confessed to lies with unquestionable evidence of their lies are a promising, but an ultimately insufficiently explored data source. We are aware of only one study that uses such data [24].

Hence there is a strong lack of correctly labeled corpora containing real-world texts especially for non-English languages.

III. TYPES OF TEXT CORPORA IN STYLISTIC DECEPTION DETECTION STUDIES

As we have noted, stylistic deception is not sufficiently investigated, which is largely due to the fact that there are no corresponding text corpora available. The only text corpus of the kind is currently that of imitative and obfuscative essays by Brennan-Greenstadt [25]. This dataset contains two types of written samples, regular and adversarial collected from 12 individuals. A regular piece contains about 5000 words of pre-existing writing samples per author. The regular writings are formal, written for business or academic settings. In the adversarial writing samples, participants were instructed to perform two adversarial attacks: obfuscation and imitation. In the obfuscation attack, each of them attempted to conceal his/her identity while writing a 500-word piece describing his/her neighborhood. In the imitation attack, each respondent was instructed to try to hide his/her writing style by imitating Cormac McCarthy's writing style in 'The Road' and as a result, there was a 500-word article with a third person description of a routine day of their life. It was extended by the texts of 56 people using AMT [26].

IV. RUSSIAN CORPORA

A. *Freely available corpora*

Russian Deception Bank is a first corpus of Russian written texts specially designed for text-based deception detection studies [11]. It currently contains truthful and deceptive narratives written by the same individuals on the same topic ("How I spent yesterday" etc.), 113 deceptive texts and 113 truthful texts written by 113 university students. Besides texts, it contains rich metadata (gender, age, self-reported handedness, test results identifying cognitive lateral profile, scores on the Domino's test (for some of the respondents), test result using the questionnaire "Styles of Behavioral Self-Regulation". The above metadata allowed us to identify connections between the linguistic parameters of deceptive texts and their authors' personalities.

The corpus is freely available at RusProfiling Lab website[4].

Gender Imitation Corpus is the first Russian corpus for studies of stylistic deception. Each respondent (n=142) was instructed to write 3 texts on the same topic (from a list). Let us provide an example of the task: "Last summer you bought a package tour from a travel agency, but you were not at all pleased with your experience with that company and the trip was not worth the price. You are about to ask for a refund. Write three texts describing your negative experience providing a detailed account of it. Give a warning that you are intending to sue the company". The first text is supposed to be written in a way usual for whoever writes it (without any deception), the second one should be written as if by someone of the opposite gender ("imitation"); the third one should be as if one by another individual of the same gender so that their personal writing style will not be recognized (what is referred to as "obfuscation"). Most of the texts are 80-150 words long.

All of the respondents are students of Russian universities. Besides the texts, the corpus includes metadata with the authors' characteristics: gender, age, native language, handedness, psychological gender (femininity/masculinity). Therefore the corpus provides countless opportunities for investigating problems arising in imitating properties of the written speech in different aspects as well as gender (biological and psychological) imitation in texts. To the best of our knowledge, this is the first corpus of the kind globally. Presently, the corpus is being prepared to be made available on the RusProfiling Lab website.

Examples of the texts in Russian Deception Bank and Gender Imitation Corpus are given in Table.

[3]http://lit.eecs.umich.edu/downloads.html#Open-Domain Deception

[4] http://en.rusprofilinglab.ru/korpus-tekstov/

Russian Deception Bank (topic – How I spent yesterday)	
Truthful	**Deceptive**
Вчера я проснулся около 11 часов утра. Это было не лучшее время в моей жизни. Было холодно. Пожелал доброго утра родителям, сестре и брату. Поел очень вкусный плов, попил чай с печеньем и отправился в комнату брата делать презентацию по английскому языку. Дело шло не очень хорошо. В процессе поиска информации я изменил тему моей презентации что не очень то мне помогло. Написав несколько предложений я спустился на кухню и пообедал. По моему я ел борщ и кашу какую-то. Вскоре всей семьей поехали в магазин откуда отправился в общежитие. Здесь я встретился с друзьями. Ближе к вечеру нагрянул еще один друг и вместе с ним мы пошли играть в шанчнин z. Спустя час я отвалился от этой компании, и пошел готовить себе на ужин вареники. Плотно поужинав я вместе с соседом посмотрел фильм, написал еще несколько предложений к презентации и отправился спать.	Проснувшись и плотно позавтракав я пошел в фитнес клуб где поплавав полчаса в бассейне отправился в тренажерный зал. После легкой утренней тренировки я с хорошим настроением поехал в университет где меня ждали ну очень интересные лекции. После пар в универе я пошел обедать в столовую. После чего отправился домой. Дома я сделал работу на следующий день. И стал собираться на прогулку со своей девушкой. Мы встретились было 18-00 в парке. Сходили в кино, погуляли, я проводил ее домой. И с хорошим настроением отправился домой. Вот такой продуктивный денек.
Gender Imitation Corpus (**topic -** Ask for a refund**)**	
Without deception (female author)	Здравствуйте! Прошлым летом я через Ваше агентство ездил в страну Н. и заплатил за это немалые деньги! Однако я едва ли доволен поездкой, потому что практически ничего из того, что было прописано в туре, исполнено не было! Вместо 4-х звездочного отеля меня поселили в отель 3 звезды, без завтрака, хотя оплачивал я 4 звезды! Две экскурсии были попросту отменены из-за плохого самочувствия экскурсовода! Я требую компенсации за все эти неудобства!!! Или я буду вынужден в противном случае обратиться в суд!
Gender imitation	Как можно так обманывать людей!? Вы вообще понимаете, что НИЧЕГО из того, что вы прописали в туре, не соответствовало реальности? 3 звезды вместо 4-х-это что такое вообще? Отмена экскурсий, отсутствие трансфера до отеля из Аэрофлота- вы вообще туристическая фирма или кто? Я вас по судам затаскаю! Немедленно выплатите мне компенсацию!
Style imitation	Ну это уже ни в какие ворота! Почему я вынужден жить не там, где я планировал, при этом заплатив больше? Как устроены ваши дурацкие механизмы, если я не увидел ДАЖЕ главной достопримечательности? Ни одной экскурсии! Пешком от отеля, который располагается в 30 минутах ходьбы от центра, до самостоятельной «прогулки», якобы экскурсии! Браво! Требую компенсации за все ваши ужасные выходки, или все дела будем решать с вами в суде, не иначе!

B. Corpus in progress

We are currently in the process of working on the first corpus of “real” deceptive texts in Russian. They are taken from video recordings of police interrogations and job interviews. The records are 19 hours long.

The video records are manually transcribed and personal data have to be removed using identifiers in order to make it impossible for subjects to be identified or linked to the subjects. As the data is not made publicly available, extra discretion has to be taken while working with it.

All of the recordings contain high quality recorded speech so that they are possible to be employed for a multimodal analysis. The transcripts of the video records will be accessed along with the data about the authors and the relevant text labeling.

Their truthfulness/deceptiveness is confirmed by means

of a series of inspections, investigations, interrogations of individuals involved, etc. Ultimately, the evidence was the narrative itself – the narrator contradicting a claim previously made. For example, one narrator, after denying a theft throughout the interview, went on to say ―All right, I did it, hence allowing his previous denials to be marked as False.

The corpus will be made available on request following the signing of the license agreement.

V. Conclusion

It is beyond doubt that in order for further progress to be made on text-based deception detection, there should be special text corpora in place. However, collecting such corpora is challenging, time-consuming and labor-intensive. It is important that corresponding corpora are continued to be collected for as many languages as possible to allow cross-linguistic studies of deception including Slavic ones that are unfortunately currently beyond the scope of any related investigations.

One of the promising directions is creating a corpus of deceptive texts by individuals speaking Russian as their second language.

Hopefully the attempts we have been making to collect corpus of deceptive Russian texts are going to pave the way for more studies of deception detection and relevant corpora of texts in other Slavic languages.

References

[1] Fuller, Ch. M., Biros, D. P., Dursun, D. 2008. Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, IEEE. Waikoloa, HI, USA. DOI= 10.1109/HICSS.2008.158.

[2] Hirschberg, J., Benus, S., Brenier, J., Enos, F., Friedman, S., Gilman, S., Gir, C., Graciarena, G., Kathol, A., Michaelis, L. 2005. Distinguishing deceptive from non-deceptive speech. In Proceedings of Interspeech 2005, Lisbon, Portugal, ACM, 1833–1836.

[3] Mihalcea, R., Strapparava, C. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In Proceedings of the Association for Computational Linguistics (ACL-IJCNLP 2009), ACM, 309-312.

[4] Zhang, H., Wei, S., Tan, H., Zheng, J. 2009. Deception detection based on SVM for Chinese text in CMC. In Proceedings of Sixth International Conference on Information Technology: New Generations (ITNG '09) (Las Vegas, NV, USA, April 27-29, 2009), IEEE, 481–486. DOI=10.1109/ITNG.2009.66.

[5] Zhou, L., Burgoon, J., Nunamaker, J., Twitchell, D. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group Decision & Negotiation, 13(1), 81–106. DOI=10.1023/B:GRUP.0000011944.62889.6f

[6] Zhou, L., Burgoon, J., Nunamaker, J., Twitchell, D. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group Decision & Negotiation, 13(1), 81–106. DOI=10.1023/B:GRUP.0000011944.62889.6f

[7] Enos, F. 2009. Detecting Deception in Speech. Doctoral thesis. Publication Number: 3348430. Columbia University.

[8] Juola, P.: Detecting stylistic deception. In: Proceedings of the Workshop on Computational Approaches to Deception Detection, Avignon, pp. 91–96 (2012)

[9] Almela, A., Valencia-Garc´ıa, R., Cantos, P. 2012. Seeing through deception: A computational approach to deceit detection in written communication. In Proceedings of the Workshop on Computational Approaches to Deception Detection. ACM. Avignon, France, 15–22.

[10] Fornaciari, T., Poesio, M. 2013. Automatic deception detection in Italian court cases. Artificial Intelligence and Law, 21(3), 303–340. DOI=10.1007/s10506-013-9140-4.

[11] Litvinova, O., Litvinova, T. Seredin, P., Lyell, J. 2017. Deception Detection in Russian Texts. In Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics (Valencia, Spain, April 3-7 2017), ACM. 43–52.

[12] Litvinova, T., Litvinova, O. 2016. Russian Deception Bank: A Corpus for Automated Deception Detection in Text. In Proceedings of CBBLR 2016, Tribun EU, 1–7.

[13] Hirschberg, J., Benus, S., Brenier, J., Enos, F., Friedman, S., Gilman, S., Gir, C., Graciarena, G., Kathol, A., Michaelis, L. 2005. Distinguishing deceptive from non-deceptive speech. In Proceedings of Interspeech 2005, Lisbon, Portugal, ACM, 1833–1836.

[14] Mihalcea, R., Strapparava, C. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In Proceedings of the Association for Computational Linguistics (ACL-IJCNLP 2009), ACM, 309-312.

[15] Newman, M., Pennebaker, J., Berry, D., Richards, J. 2003. Lying words: Predicting deception from linguistic style. Personality and Social Psychology Bulletin, 29(5), 665-675. DOI=10.1177/0146167203029005010.

[16] Rubin, V. L., Conroy, N. J. 2012. The art of creating an informative data collection for automated deception detection: A corpus of truths and lies. In Proc. Am. Soc. Info. Sci. Tech., 49, John Wiley & Sons, Inc., 1–11. DOI= 10.1002/meet.14504901045.

[17] Perez-Rosas, V., Mihalcea, R. 2015. Experiments in Open Domain Deception Detection. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2015) (Lisbon, Portugal, September 17-21, 2015), ACM, 1120–1125.

[18] Perez-Rosas, V., Mihalcea, R. 2014. Cross-cultural Deception Detection. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), (Baltimore, Maryland, USA, June 23-25, 2014). ACM, 440–445.

[19] Verhoeven, B., Daelemans, W. 2014. CLiPS Stylometry Investigation (CSI) corpus: a Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In Proceedings of the 9th International Conference on Language Resources and Evaluation (Reykjavik, Iceland, May 2014).

[20] Perez-Rosas, V., Mihalcea, R., Narvaez, A., Burzo, M. 2014. A multimodal dataset for deception detection. In Proceedings of the Conference on Language Resources and Evaluations (LREC 2014) (Reykjavik, Iceland, May 2014).

[21] Fitzpatrick, E., & Bachenko, J. 2012. Building a data collection for deception research. In Proceedings of the 13th Conference of the European Chapter for the Association for Computational Linguistics: Computational Approached to Deception Detection Workshop (EACL 2012), ACM, Avignon, France, 31-38.

[22] Fitzpatrick, E., Bachenko, J. 2009. Building a Forensic Corpus to Test Language-based Indicators of Deception. Corpus Linguistics. In Corpus-linguistic applications Current studies, new directions. Edited by Stefan Th. Gries Stefanie Wulff Mark Davies, Series in Language and Computers. Rodopi, 183-196.

[23] Perez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M. 2015. Deception Detection using Real-life Trial Data. In Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2015) (Seattle, Washington, USA — November 09-13, 2015), ACM, 59-66. DOI=10.1145/2818346.2820758.

[24] Larcker, D. F., Zakolyukina, A. A. 2010. Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50: 495–540. DOI=10.1111/j.1475-679X.2012.00450.x.

[25] Koper, R. J., Sahlman, J. M. 1991. The behavioral correlates of real-world deceptive communication Paper presented at the Annual Meeting of the International Communication Association (41st, Chicago, IL, May 23-27, 1991). Distributed by ERIC Clearinghouse, 1991.

[26] Michael Brennan and Rachel Greenstadt. 2009. Practical attacks against authorship recognition techniques. In Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence (IAAI), Pasadena, CA.

[27] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy, pages=To appear. IEEE

Tatiana A. Litvinova received her PhD from Voronezh State University, Voronezh, Russia. She is a founder and head of RusProfiling Lab, Voronezh State University, Voronezh, Russia. She is also a researcher in Kurchatov Institute, Moscow, Russia. She and her lab team are involved in the study of author profiling in Russian texts, text-based deception detection, author gender imitation, text-based suicide behavior prediction, etc. Tatiana Litvinova is in charge of collecting "RusPersonality" which is the largest Russian text corpus with rich metadata about their authors (gender, age, education, psychological traits, etc.). She is a member of the Russian Cognitive Linguists Association.

Olga V. Zagorovskaya is a Doctor of Science in Philology (1991), Professor (since 1993). She is a professor of the Department of the Russian Language, Modern Russian and Foreign Literature of Voronezh State Pedagogical University and a head of the Scientific Linguistic School of Voronezh State Pedagogical University exploring the issues and development of the Russian language at the turn of the 21st century and ecolinguistics.

Olga A. Litvinova is a PhD student at the RusProfiling Lab, Voronezh State University, and lecturer at Department of English Language, Voronezh State Pedagogical University. She is also a researcher at Kurchatov Institute, Moscow.
Her research interests are text-based deception detection, authorship profiling in non-native speaker texts.