Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language

Dina Pisarevskaya
Higher School of Economics
Moscow, Russia
dinabpr@gmail.com

Abstract—**Deceptive news can be used for manipulation of public opinion. As for Natural Language Processing, new tools for automated deception detection and news verification are created to prevent this for different languages, it is highly demanded in our contemporary society. In the research we based on rhetorical structures and their relations in news reports in the Russian language. The texts from our data collection were dedicated to one definite event. We relied on factuality by gathering the data collection. After annotating the texts we used Support Vector Machines method to classify news reports by their rhetorical structures relations features in order to predict if the reports from the test set were truthful or deceptive. We found out that Elaboration and Background relations are more likely to occur in the truthful reports, than deceptive ones. Antithesis and Volitional Cause relations are more often represented in deceptive reports. The model was able to correctly label 60% reports as truthful or deceptive ones. The present research is initial, thus the model should be modified according to larger data collections, combining different topics, and more complex statistical methods. But it can already be used as a preliminary filter for deceptive news detection in the Russian language.**

I. Introduction

In the contemporary world we deal with the large amount of information that we get from different and diverse sources: newspapers and magazines, institutional and non-institutional online media, blogs and social media, TV channels and their websites, radio shows. It is very important to understand the difference between information types and to be able to evaluate the source reliability.

If we speak about texts of news reports, we can say that rumors, deceptive information and deceptive (fake) news can easily be used for manipulation of public opinion, for information warfare. Fake news reports can also be created and published by people who don't take into consideration the necessity of fact checking. News editors usually are aware of some patterns of deceptive news reports [1], but there are still no common criteria for reliable information detection vs deception detection. This is why new tools for automated deception detection and information verification, created for different languages, are required in our society. This issue could be the international challenge for researches from different countries.

Most studies in the field of natural language processing, computational linguistics and information science focus on lexics and semantics and some syntax principles for automated deception detection. Discourse and pragmatics have still rarely been considered [2] due to the complexity of such approach.

II. Rhetorical Structure Theory as a Feature for Deception Detection

A. Literature Review

Problem of deception detection in oral communication has been subject of research interest for a long time. It is an interdisciplinary field which is an interesting issue for computational linguistics, speech processing, psychology (polygraph tests), and physiology studies. In general, in regard to the computational linguistics and natural language processing, different tools and software already exist which help in detecting deception and lie. For instance, it was revealed that people show less negative emotions, use less inconsistencies and modal verbs, use more modificators and speak longer if they tell deceptive information [3]. The records were analyzed according to the Linguistic Inquiry and Word Count (LIWC) software.

Different written texts are also subject of research for studying deception detection methods. Digital texts, online reviews [4], fake social network profiles [5], fake dating profiles [6] etc. were already investigated. Some linguistics research is done to answer the following question: are there any distinctions between the reports about the same event in institutional and social media [7]. But the objective of revealing news verification mechanisms arose rather recently. Some linguistics markers can be found in lexics and semantics level from the Statement Validity Analysis [8]. Existing psycholinguistic lexicons, for instance LIWC [9], can be used in performing binary text classifications for truthful vs deceptive texts (70% accuracy rate) [10].

Speaking about the discourse and pragmatics level, one can face dramatic difficulties in gathering "gold standard" for machine learning and in deciding if certainty markers ("perhaps" ,"I believe" etc.) could help us do identify a report as truthful or fake one [11].

Some studies are focused on creating models that reveal if the described event accords with the facts or not. For example, authors of the article [12] represent a model, which is based on

ISBN 978-952-68397-0-7 (paperback), ISBN 978-952-68397-1-4 (PDF)

grammatical fact description structures in English and the kindred languages. The model lets us consider different sources and contradictory information. It also contains a scale of facts conformity. It is based on linguistics features, such as modal particles, different predicate types etc. This model has been implemented in De Facto, a factuality profiler for eventualities mentioned in text. It is based on specific lexical types (reporting, presuppositional, implicative predicates etc.) and syntax constructions (for instance, conditional structures). It is suitable for English language and, generally, for Germanic and Italic languages, but it should be modified for languages from other families, subfamilies and groups. The researchers also created FactBank - annotated corpus in English. It is mentioned there if events described in text accord with facts. The authors use linguistics markers and constructions listed in the article [13].

Recent research projects are dedicated to discourse differences between deceptive (fabricated) and truthful (authentic) news, specifically in terms of their rhetorical structures and coherence relation patterns [14]. Vector space modeling application lets us predict whether a particular news report is truthful or deceptive (63% accuracy). We see that rhetorical structures and discourse constituent parts and their coherence relations are already reviewed as possible deception detection markers in the English news reviews. If we review international deception detection methods, we also should keep in mind the cultural considerations [15].

There are no research articles about automated deception detection for the Russian language in Russia. There is also a significant lack of linguistics tool for natural language processing. It seems to be a theoretical and methodological challenge.

RST (Rhetorical Structure Theory) framework [16] is addressed to the discourse level of text. It represents text as an hierarchical tree. Some parts are more essential (nucleus) than others (satellite). Elementary discourse units are connected to each other according to relations: elaboration, justify, contrast, antithesis, volitional result etc. The theory pretends to be universal for all languages, that's why we chose it in our research. It is used in the Russian computational linguistics [17][18]. Nevertheless automated parser was never worked out specially for the Russian language.

Support Vector Machines (SVM) method can be grasped as supervised learning models for machine learning with associated algorithms that analyze data and recognize patterns; they are used for classification and regression analysis. Text classifiers learn from examples [19]. In our case, news reports are shown as vectors in n-dimensional space. After classification based on training set, a new report is placed in one of two groups, deceptive or truthful.

B. Research objective

Our hypothesis is that there are significant differences between structures of truthful news reports and structures of deceptive (fake) ones. This is based on some peculiarities of RST relations among discourse parts in these texts. Our aim is to reveal the differences using RST relations as deception detection markers. Firstly, we would like to find out the differences among types of RST relations and their frequencies in our two news groups (truthful news reports and deceptive news reports). Then we shall classify them with the use of Support vector machines (SVMs) methodology, based on the RST relations labeling, and we shall do our best to predict if news reports are truthful or deceptive.

This model can be useful for news verification, in detecting and filtering deceptive (fake) news. Especially it is of vital necessity for the Russian language, because news reports in Russian nowadays often contain deceptive information and deliberate misinformation, and there is no way how to check it excepting manual. Our research is based on the methodology of the news reports research for the English language [20], but it also takes into consideration some features of this research field for the Russian language.

C. Analysis details

1) Data Collection Principles: The main difficulty of collecting data set for deception detection is the lack of websites, newspaper columns, TV programs or radio shows in Russian that contain verified samples of fake and truthful news. For example, radio show in English "Wait, Wait, Don't Tell Me" with its "Bluff the Listener" segments was used for researching RST analysis method in deception detection in English [21]. Each show contains three thematically-linked news reports, one of which is truthful and the other two are fake (deceptive). We deal with the absence of such shows or other sources of news in the Russian language. In addition, there are no Factbanks in the Russian language, as it is in English.

There are several fact checking websites for English or German that contain the reports of investigative journalism dedicated to different news stories in media [22]. They are based on the analytical, non-automated work of journalists and volunteers. Similar reports in Russian could be useful in collecting authentic and deceptive news as well. However, two instances of such sites in Russian [23] are biased, subjective and politically motivated, that is why their collections of fake news cannot be used for gathering the data set.

The only way out in solving the problem was the reliance on the presented facts, on the factuality. We decided to exclude political and military news from our monitoring for avoiding mistakes because it is very difficult to check if such news articles are deceptive or truthful. We also did not take into consideration scientific news, because news stories posted on the Russian science web portals are usually very carefully verified.

The total data set consists of 30 news reports, which are related to the same event. Russian bikers from the motorcycle club "The Night Wolves" ("Nochnie volki"), members of which are on the US and Canadian sanctions lists, planned to celebrate the "Victory Day" on May 9 (2015) in a trip from Moscow to Berlin. But the government of Poland refused them entry, and German authorities cancelled the bikers' Schengen visas. On April 26, some online newspapers stated that motorcycle club members crossed the Polish border despite the diplomatic note of the Ministry of Foreign Affairs of the Republic of Poland on April 24, which had forbidden club members to enter the country. On the same day and on April 27, online newspapers clarified that it was another group of bikers, who wanted to take part in the memorial ceremony, which was dedicated to the 68-th anniversary of World War II ending. It took place at Red Army cemetery in Braniewo, near the Russian's Kaliningrad Region border. Finally, on April 27, it was reported that motorcyclists from the club "The Night Wolves" were stopped and turned back by Polish Border Guards. It means that on the same day (April 26) and actually at the same period of time we can see news articles arguing that "The Night Wolves" have crossed the border and asserting that the bikers who have crossed the border do not belong to the rally group which is en route Berlin to mark the anniversary of the Second World War. Therefore, we can conclude that we study here deceptive news reports and truthful news reports, and not just ordinary news reports and their refutation. For example, on April 26, at 8:00 p.m., the opposite news reports were published: reports of crossings the border by "The Night Wolves" as well as reports denying the fact or stating that it was another group.

Thanks to the Russian search engine Yandex (the section news.yandex.ru), we have found approximately 140 news reports dedicated to this event. But they quoted each other, and most texts were almost similar. Thus, we chose 30 reports, as much as possible without repetitions, but the sources (online newspapers in Russian) were selected randomly: this means that if we found resembling text in 10 newspapers, we randomly chose one of them. We took news reports in the Russian language from Russian, Ukrainian and Moldavian online newspapers.

2) Corpus Details and Data Analysis: As our research is trial and initial, we would like to discover some patterns that would be examined, affirmed or corrected, and elaborated later, during the further research, according to the larger data set and, thus, to get better results of the machine learning. In the present research we have a sample of 30 news reports. 15 news reports are truthful, 15 news reports are deceptive. 20 reports belonged to the training set (10 truthful reports and 10 deceptive reports), the test set consisted of 10 reports (5 truthful reports and 5 deceptive reports).

Annotator A analyzed 15 randomly selected reports (5 truthful news reports from the training set, 5 deceptive news reports from the training set, 5 reports from the test set), annotator B analyzed the remaining reports.

We used RSTTool for discourse-level annotation (the unicode version for the Russian language). It is grounded in the Rhetorical Structure Theory (RST) framework and subsequently extended relation sets – for instance, we took the relational category "Unconditional" from the B.Mann's set (2000): the satellite could influence on the action in its nucleus, but nucleus does not finally depend on its satellite. There are 5 different sets in RSTTool, but news reports usually have a definite template, thus, we used a relatively small number of different relational categories. Some relation categories assigned in the "classic" set by Mann and Thompson [24] were not used in our research at all – such as Motivation and Solutionhood. It is because they are rarely applied in news reports in Russian.

A sample of RST relations assignments to Antithesis, Purpose, Volitional Result is presented in Fig. 1.

Each news report was first segmented into elemental discourse units. Clauses were taken as elementary discourse units - although there are several approaches for determining elementary discourse unit: clause; prosodic unit; turns of talk; sentences; intentionally defined discourse segments etc. [25]. There are two strategies of annotator's work in RST analysis [26]. An annotator could apply relations to the segments sequentially, from one segment to another, connecting current node to the previous node. This method is suitable for short texts, such as news reports, but even in such texts there is a risk of overlooking important relations. The other method is more flexible: the annotator segments multiple units simultaneously, then builds discourse sub-trees for each sentence, links nearby sentences and builds firstly larger sub-trees and after that the final tree, linking key parts of the discourse structure. We combined these two strategies. At first the annotators applied relations sequentially to the top-level unit which is usually at the beginning of the news report in Russian. Then they labeled other relations in text using the second strategy.

The annotators re-read each news report several times to understand better the logic of the story. The checking steps involved making sure that the tree has a single root node and there are no missing fragments from the text.

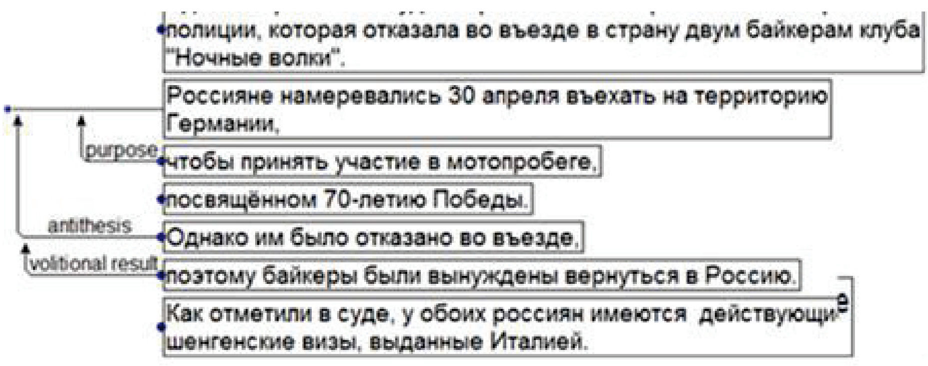

Fig. 1. Sample of RST relations assignments (visualization in UAM CorpusTool)

After that annotators reviewed nucleus-satellite assignments and the choice of relations types. There are no discourse parsers for Russian, that's why we could not detect possible tree errors with them. Tagging and validation were made manually.

3) Inter-annotator consistency: Subjectivity of applying RST relations is a declared critique. Therefore, the team of annotators tried to reach a significant level of consistency, even though they faced some challenges that reflected differences in their tagging. During the first step, the annotators chose randomly one news report from the training set, segmented it and assigned RST relations. Then they discussed the variations and tried to reach consensus. We decided that we use 25 relation types in our work, and reached agreement on the segmentation into elementary discourse unites per RST. The interface of the RSTTool also allows to create relation types, so we decided to make Evidence 1 (the source of information, the speaker, is mentioned precisely) and Evidence 2 (the source is mentioned imprecisely). Hence, finally we had 26 relation types. Two days later the annotators chose randomly two texts from the training set, processed them, and then they compared their assignments for the RST relations. According to our confusion matrix, the level of inter-annotator agreement was more than 70%: 22 agreed upon assignments out of 29 (75.86%). The results were discussed and two days later two texts were also chosen randomly, and the agreement was improved by 2%.

There is the automatic identification method of the best tag [27] which is suitable for two annotators, if they deal with choices between two characteristics. While in some cases we had to choose one of three and more characteristics. As the choice made on the first step influences the further choices during the next steps, we decided to decline this method. Instead of it we tried to reach consensus in disputable issues and to write a guideline. We faced the following discrepancies during our preliminary tagging work: Conjunction/Sequence/Elaboration, Antithesis/Contrast/Unconditional, Circumstance/Volitional cause, Background/Elaboration/Volitional result. We prepared a short guideline for these cases.

We also decided not to tag Evidence relations and to miss the relevant clauses, if the name of a speaking person is repeated more than twice in the text. It means that there could be only one or two Evidence relations according to the same person. We did so because some news reports included the references of one definite person literally in every sentence after the direct speech, it can be often seen in online media in Russian. If the motorcyclists' plans to visit different cities on their way to Berlin were described in the clauses, we assigned such relations as Circumstance (the satellite sets a framework and describes an unrealized matter within which the reader is intended to interpret the situation presented in the nucleus) instead of Purpose.

Finally we calculated the Cohen's kappa Weighted for two random labelled news reports from our data collection. It was 0.75, while values between 0.6 and 0.8 reflect the good agreement.

4) Statistical procedures: We aimed to find out if there is any difference between truthful news reports and deceptive ones. It concerns types of RST relations and their frequencies. We tried to reveal possible dependences. If the relations of discourse units in deceptive reports differ from the ones in truthful reports, then the research on such relations could be useful in detecting and filtering deceptive news.

Firstly we got a statistics file for each news report in the training set. These files contained information about relation types in text and the numbers of each type. The relation types were marked as text features. Then we integrated all these files into the common statistics file. The table was subdivided into two parts: features (with relevant numbers) in truthful reports and features (with their numbers) in deceptive reports. We calculated average value for each feature in the set of truthful reports. Then we calculated standard deviation for each feature in the set of truthful reports. After that we did both procedures for each feature in the set of deceptive news reports.

We shall conclude that Elaboration and Background relations are more likely to occur in truthful reports. While Antithesis and Volitional Cause relations are more represented in deceptive reports.

As our data collection is relatively limited, we can immediately see the patterns after the simple calculations above (Fig. 2).

At the next step we evaluated the predictive power of the model. We used scikit-learn library. It contains several tools for machine learning for the Python programming language [28]. In the first place we selected Support vector machines (SVMs) algorithm. It is a set of supervised learning methods used for classification, regression and outliers detection. We used the class sklearn.svm.SVC (Support Vector Classification). It is capable of performing multi-class classification on a dataset. We had two classes: truthful news (1) and deceptive news (0). The model was based on the training dataset. We tested it on our test set and got score 0.6. The model was able to correctly label 60% reports as truthful or deceptive ones.

Test set: 1 1 1 1 1 0 0 0 0 0

Prediction: 0 0 1 0 1 1 0 0 0 0

Precision for truthful news is 2/3, precision for deceptive news is 4/7, recall for truthful news is 2/5, recall for deceptive news is 4/5. False Positive Rate is 1/5. We can conclude that the current model could separate the set of definitely truthful news, herewith it identifies some truthful news as deceptive ones.

Scikit-learn also includes the Support Vector Regression method, which is the extension of the Support Vector Classification. In addition to the deceptive news vs truthful news identification, it also shows how near the news report is to the truth vs deception center. We got the following results: [0.35, 0.42, 0.74, 0.41, 0.52, 0.53, 0.32, 0.02, 0.22, 0.32]. It is visible that some figures are almost in the middle of the scale. The difference between deceptive report and truthful report is not always significant. A cutoff is needed. The model should be modified in further research.

D. Discussion

Further measures should be taken to find features for deception/truth detection in automated news verification model for the Russian language. It should be learned and tested on a larger data collection (minimum 300 news reports), and news reported should be dedicated to different events. In addition, we should use more complex statistical methods, such as logistic regression, chi-square etc.

According to the results of our current research, Elaboration and Background relations are more likely to occur in the truthful reports. Antithesis and Volitional Cause relations are more likely to occur in the deceptive ones. Probably this could be explained as follows. Authors of deceptive (fake) news pay more attention to the causation, because they want to represent the required point of view and to explain all events with the internal logic of their position, without any internal inconsistencies and conflicts.

The present research is initial. The existing model should be modified. Its extrapolation to all possible news reports in Russian, dedicated to different topics, would be totally incorrect. But despite this fact, it can already be used as a preliminary filter for deceptive (fake) news detection. It should leave truthful news in the set. At the same time the results of its work should be double-checked and refined, especially for suspicious instances fact checking.

A statistics file for each news report was translated to a multi-dimensional vector representing RST relation types and their frequencies. In fact, the model dealt with "the bag of RST relations". One of the purposes of our further research by modifying the model is to take into consideration "the trees" - hierarchies of RST relation types in texts and dependences between relation types, which are located close to each other in news reports.

The last but not the least, the assignment of RST relations to news report could be connected with the subjectivity of annotators' interpretations. This problem can be partly solved by preparing more precise manuals for tagging and by developing consensus-building procedures.

III. Conclusion

News verification tends to be a very important issue in our actual world, with its information wars, information warfare and propaganda methods. RST basis seems to be a promising and methodologically challenging field for automated deception detection and deceptive (fake) news filtering. We researched this issue according to the features of the Russian language and the templates of news reports in Russian. We faced lack of computational linguistics tools for the Russian language, e.g. discourse parsers.

We also dealt with some difficulties during our data set selection, because there were no websites, TV programs, radio shows and newspaper columns, that contain absolutely credible and checked analysis of deceptive news reports. Despite of these problems, we collected a corpus based on the presented facts.

We chose 30 news reports focused on one topic (from online newspapers from three different counties). At first glance, it was not clear if Russian motorcyclists from the club "The Night Wolves" had crossed the Polish border on April 26, 2015 or not. We found out that deceptive news stated information about the successful border crossing, and truthful reports disproved it.

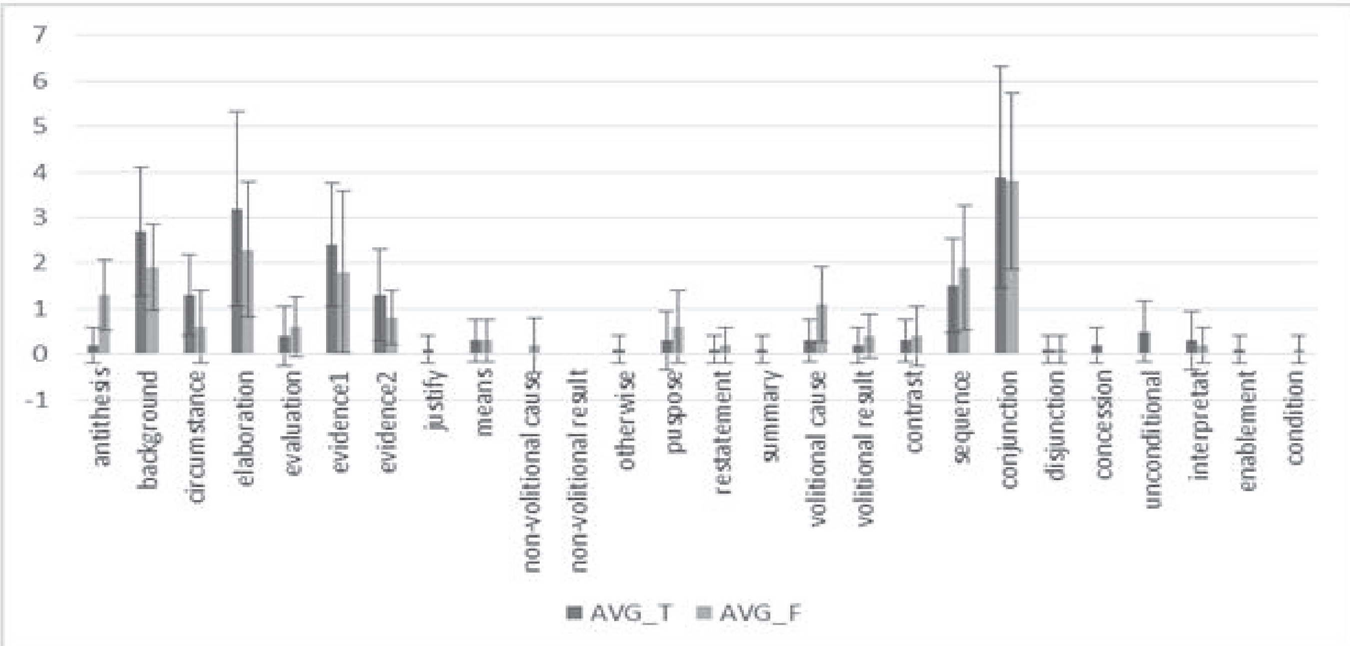

Fig 2. RST relation types in truthful and deceptive news reports. (AVG_T – average value for a feature in the data set of truthful news, AVG_F – average value for a feature in the data set of deceptive (fake) news. Standard deviation is marked)

At the following step we segmented the texts and applied RST relations tagging to them. The Cohen's kappa Weighted (as a measure of agreement between two annotators) is 0.75.

We revealed that Elaboration and Background relations are more likely to occur in the truthful reports. Antithesis and Volitional Cause relations are more represented in the deceptive ones. Then we applied Support vector machines (SVMs) algorithm to classify the news reports. Relation types were accepted as features for machine learning. The model was able to correctly label 60% reports as truthful or deceptive ones. The present research is initial, and the model should be modified according to larger data collections, dedicated to different topics, and more complex statistical methods. But it can be already used as a preliminary filter for deceptive (fake) news detection in the Russian language. The modified model could combine RST relations markers with other deception detection markers in order to make a predictive model.

IV. References

[1] D.L. Lasorsa and J. Dai, "Newsroom's Normal Accident?", *Journalism Practice*, 1 (2), 2007, pp. 159-174.

[2] V.L. Rubin, N.J. Conroy and Y. Chen, "Towards News Verification: Deception Detection Methods for News Discourse", in *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, January 5-8, 2015, Grand Hyatt, Kauai, p. 3.

[3] M. Ali and T.R. Levine, "The language of truthful & deceptive denials & confessions", in *Communication Reports*, vol. 21, 2008, pp. 82-91.

[4] A. Mukherjee et al, *Fake Review Detection: Classification & Analysis of Real & Pseudo Reviews. Technical Report*. Department of Computer Science, University of Illinois at Chicago, & Google Inc,.2013.

[5] N. Kumar and R.N. Reddy. *Automatic Detection of Fake Profiles in Online Social Networks*. BTech Thesis, 2012.

[6] C.L. Toma and J.T. Hancock, " What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles", *Journal of Communication*, vol. 62, no. 1, 2012, pp. 78-97.

[7] C. Zimmerman et al, "Marius, the Giraffe: A Comparative Informatics Case Study of Linguistic Features of the Social Media Discourse", in *CABS'14,* August 20–22, Kyoto, Japan, 2014.

[8] S. Porter and J.C. Yuille, "The language of deceit: An investigation of the verbal clues to deception in the interrogation context", *Law and Human Behavior*, vol. 20, no. 4, 1996, pp. 443-458.

[9] J. Pennebaker and M. Francis, *Linguistic inquiry and word count: LIWC.* Erlbaum Publishers, 1999.

[10] R. Mihalcea and C. Strapparava, "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language", in *Proc. 47th Annual Meeting of the Association for Computational Linguistics*, Singapore, 1999, pp. 309-312.

[11] V.L. Rubin, N.J. Conroy and Y. Chen, "Towards News Verification: Deception Detection Methods for News Discourse", in *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, January 5-8, 2015, Grand Hyatt, Kauai, p. 3.

[12] R. Sauri and J. Pustejovsky, "Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text", *Computational Linguistics*, 2012, pp. 1-39.

[13] R. Sauri and J. Pustejovsky, " FactBank: a corpus annotated with event factuality*", Language Resources & Evaluation*, vol. 43, no. 3, 2009, pp. 227-268.

[14] V.L. Rubin, N.J. Conroy and Y. Chen, "Towards News Verification: Deception Detection Methods for News Discourse", in *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, January 5-8, 2015, Grand Hyatt, Kauai, 11 pages.

[15] V.L. Rubin, "Pragmatic and cultural considerations for deception detection in Asian languages", in *ACM Trans. Asian Lang. Inform. Process*. 13, 2, Article 10 (June 2014), 8 pages. DOI = http://dx.doi.org/10.1145/2605292.

[16] W.C. Mann and S.A. Thompson, "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization", *Text*, vol. 8, no. 3, 1988, pp. 243-281.

[17] А.А. Кибрик, *Анализ дискурса в когнитивной перспективе.* Дис. … д-ра филол. наук. М.: Ин-т языкознания РАН, 2003.

[18] *Рассказы о сновидениях: Корпусное исследование устного русского дискурса* / Под ред. А.А. Кибрика, В.И. Подлесской. М.: Языки славянских культур, 2009.

[19] T. Joachims, "Text Categorization with Support Vector Machines Learning with Many Relevant Features", in *European Conference on Machine Learning (ECML),* Chemnitz, Germany: [s.n.], 1998, pp. 137-142.

[20] V.L. Rubin, N.J. Conroy and Y. Chen, "Towards News Verification: Deception Detection Methods for News Discourse", in *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, January 5-8, 2015, Grand Hyatt, Kauai, 11 pages.

[21] V.L. Rubin, N.J. Conroy and Y. Chen, "Towards News Verification: Deception Detection Methods for News Discourse", in *Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, January 5-8, 2015, Grand Hyatt, Kauai, p. 4.

[22] http://www.correctiv.org, https://theconversation.com/uk/topics/fact-check, http://www.politifact.com, etc.

[23] http://maxkatz.livejournal.com/237980.html ("Antipropaganda" project), http://www.stopfake.org.

[24] W.C. Mann and S.A. Thompson, "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization", *Text*, vol. 8, no. 3, 1988, pp. 243-281.

[25] L. Carlson, D. Marcu and M.E. Okurowski, "Building a Discourse Tagged Corpus in the Framework of Rhetorical Structure Theory", in *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Aalborg, 2001, p. 4.

[26] L. Carlson, D. Marcu and M.E. Okurowski, "Building a Discourse Tagged Corpus in the Framework of Rhetorical Structure Theory", in *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Aalborg, 2001, p. 6.

[27] J. Wiebe, R Bruce and T. O'Hara, "Development and Use of a Goldstandard Data Set for Subjectivity Classifications", in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, 1999, pp. 246-253.

[28] http://scikit-learn.org.