УДК 004.942
DOI: 10.25559/SITITO.14.201803.567-572

# LEXICON-BASED APPROACH IN GENERALIZATION EVALUATION IN RUSSIAN LANGUAGE MEDIA

Aleksandr M. Krassovitsky, Irina M. Ualiyeva, Zhazirakhanym Meirambekkyzy, Rustam R. Mussabayev
Institute of Information and Computational Technologies, Almaty, Kazakhstan

# ОСНОВАННЫЙ НА ЛЕКСИКОНЕ ПОДХОД В ОЦЕНКЕ ОБОБЩЕНИЙ В РУССКОЯЗЫЧНЫХ СМИ

А.М. Красовицкий, И.М. Уалиева, Ж. Мейрамбеккизы, Р.Р. Мусабаев
Институт информационных и вычислительных технологий, Алматы, Казахстан



**Keywords**

Official media; lexicon-based approach; generalization evaluation.

**Abstract**

We consider generalization as a property of human thinking to make general conclusion based on authors' own experience and observations and one of the techniques of authors use to manipulate the readership and present an algorithm for evaluation of the generalization in texts. The algorithm is based on the lexicon-based approach. To search the generalization we use ready-made dictionary (KEY-dictionary) and RuSentiLex dictionary. KEY-dictionary contains words and phrases (elements) that express the generalization. In RuSentiLex we take the words and phrases that express opinion and fact. The algorithm searches exact matches the elements from text with the elements from the dictionaries, it is also important that the elements from different dictionaries have their weights. New method is developed for automatic detection of generalization in texts from official media. Numerical calculations of generalization were performed using a special software application. To test the proposed approach the expert estimation were used.

**Ключевые слова**

Официальные СМИ; лексикон-ориентированный подход; оценка обобщения.

**Аннотация**

Статья представляет обобщение как свойство человеческого мышления делать общий вывод на основе собственного опыта и наблюдений авторов, также рассматривается один из приемов, используемых авторами для манипулирования читательской аудиторией, и представлен алгоритм оценки обобщения в текстах. Алгоритм основан на лексиконном подходе. Для поиска обобщения используется готовый словарь (KEY-dictionary)и словарь RuSentiLex. KEY -словарь содержит слова и словосочетания (элементы), выражающие обобщение.
В RuSentiLex мы берем слова и фразы, которые выражают мнение и факт. Алгоритм поиска точно сопоставляет элементы из текста с элементами из словарей, также важно, чтобы элементы из разных словарей имели свои веса. Разработан новый метод автоматического обнаружения обобщения в текстах официальных СМИ. Численные расчеты обобщения выполнены с использованием специального программного приложения. Для проверки предложенного подхода использовалась экспертная оценка.

About the authors:

**Aleksandr M. Krassovitsky**, IICT – Institute of Information and Computational Technologies (125 Pushkin Str., Almaty 050001, Kazakhstan), ORCID: http://orcid.org/0000-0003-2948-374X, akrassovitskiy@gmail.com
**Irina M. Ualiyeva**, Candidate of Physical and Mathematical Sciences, IICT – Institute of Information and Computational Technologies (125 Pushkin Str., Almaty 050001, Kazakhstan), ORCID: http://orcid.org/0000-0003-3853-8896, i.ualiyeva@gmail.com
**Zhazirakhanym Meirambekkyzy**, IICT – Institute of Information and Computational Technologies (125 Pushkin Str., Almaty 050001, Kazakhstan), ORCID: http://orcid.org/0000-0003-2309-7725, jazirahanim@mail.ru
**Rustam R. Mussabayev**, IICT – Institute of Information and Computational Technologies (125 Pushkin Str., Almaty 050001, Kazakhstan), ORCID: http://orcid.org/0000-0001-7283-5144, rmusab@gmail.com

## An overview of the current approaches

The review revealed that G.'s concept has not been seriously analysed in the Russian-language scientific literature yet. The question of the algorithmization of the G. degree's numerical evaluation in the texts has not previously been considered. The works we have studied on G. belong to political scientists and linguists [3, 5, 9]. In order to obtain a numerical estimate of the text G., a number of approaches for sentiment analysis can be applied [10]. Here, the authors suggest to determine the key elements of the text. For example, Vinodhini and Chandrasekaran [11] highlight the following most popular approaches for sentiment analysis:

- subjective lexical approach, i.e., a list of words is compiled, where each word is assigned an estimate indicating the nature of the word: positive, negative or neutral;
- n-grams, i.e., on the basis of training data, unigrams, bigrams, trigrams or combinations thereof are compiled for the further classification of the text;
- machine learning, i.e., algorithms of machine learning are used in order to extract information from the text and train the model.

In [10, 12, 13] the authors offer a similar approach to sentiment analysis: (i) machine learning, (ii) lexical approach, and (iii) hybrid approach.

The approach based on machine learning uses classification methods: documents are divided into two sets: teaching and test samples. After all the training samples are used up, testing samples verify how well the classification has been done.

In the lexical approach, researchers distinguish the following methods:

- dictionary-based method: the dictionary of keywords is initially created (manually, based on templates or ready-made dictionaries), then an analytical comparison of the text with dictionaries is performed [14];
- another method based on machine learning [15] combines a linguistic approach and machine learning. In this approach, the linguistic features identified by the researchers are integrated into the algorithms of machine learning;
- method based on textual corpus: words are found from the corpus and a conclusion is made on the basis of these words about the tonality [16];
- ensemble methods [17]: increase the accuracy of classification by combining arrays of individual training samples. Approach techniques: bugging, bootstrapping.

In the hybrid approach, the combination of machine learning and vocabulary-based approach has the potential to improve the quality of G. quantification [18]. The combined approach merges carefully developed vocabulary and machine learning algorithm.

The lexical approach for determining the key elements of a text, with a set of words in dictionaries [14], [19], was applied by the authors for a numerical evaluation of G. [20].

The basic restriction that is placed on the machine learning usage is the lack of a large amount of text corpus to process. Applying a lexical approach, the authors can circumvent this kind of restriction, focusing on the collection of dictionaries, the comprehension of the test data by experts.

## The approach proposed by the authors

Let *T* be an input text in the form of a sequence of words, $K=\{k_1, k_2, ..., k_l\}$ be a set of keywords or phrases that are used to numerically evaluate the generalization of the text. Words or phrases from *K* are called elements of *K*. $F=\{f_1, f_2, ..., f_m\}$ be a set of lexical elements that express the fact, $O=\{o_1, o_2, ..., o_p\}$ be a set of elements expressing the opinion of the author. Note that *O* is represented by lexical introductory terms. The numerical estimate of $gen(T, K)$ is calculated from the frequency of occurrences of text elements entering the set *K*:

$$gen(T,K) = \frac{|T \cap K|}{|T|}, \qquad (1)$$

is a sequence of $T \cap K$ is a sequence of key elements from *T*, $|T|$ is the number of elements in the text.

An extended G. estimate will be calculated from the frequency of occurrences of elements included in the sets K, F, O. It is assumed that the words and phrases from *F* and *O* weaken the generalization rate.

**Generalization weight of the sentence, taking into account the key elements.** It is easy to see that words, phrases or introductory phrases from different dictionaries have their weight. Consider the text as a sequence of the sentences $S = s_1, s_2, ..., s_q$, where *q* is the number of sentences in the text. For each $w_i \in s_r$, $1 \le r \le q$ means $w_i$ is a substring of $s_r$ and for each $w_i \in K$ weight $(w_i) = 1$ for the case the sentence contains only one element from *K*. If the number of elements from *K* is greater than one, then the weight of the generalization of the sentence weight $(s_r)$ is amplified by factor coefficient of generalization:

$$weight^K(s_r) = \begin{cases} weight(w_i), \ w_i \in s_r, w_i \in K \\ \alpha \sum_{i=2}^{l} weight(w_i), \ w_i \in s_r, w_i \in K \end{cases}, \qquad (2)$$

where *l* is the number of words, phrases (in particularly, introductory phrases) in the text from *K* in the sentence .

The weight of the generalization of the sentence, taking into account the word-modifiers: In the case of weakening the generalization, we also consider the text as a sequence of sentences $S = s_1, s_2, ..., s_q$, where *q* is the number of sentences in the text. Calculate weight $weight^M(s_r)$ of $s_r$ as follows:

$$weight^M(s_r) = \prod_{i=1}^{m} weight(w_i), \qquad (3) \ w_i \in F \cup O, w_i \in s_r$$

where $m=|F \cup O|$. The initial weight $(w_i)$ is set to 0.75 so the generalization rate is weakened for the case a single word-modifier presents in the sentence. Thus, the weight of generalization of each sentence in the text, taking into account the inclusion of dictionaries of key words and dictionaries of generalization weakening, will be calculated by the following formula:

$$weight(s_r) = \begin{cases} weight^K(s_r), \ w_i \in K, s_r \cap (F \cup A \cup O) = \emptyset \\ weight^K(s_r) * \ weight^M(s_r), w_i \in K, F, O \\ 0, w_i \in F, O, s_r \cap K = \emptyset \end{cases}, \qquad (4)$$

for each $w_i \in s_r$.

G. rate with intensification gen(S, K, F, O)is calculated as a sum of weights of the generalization for each sentence in S:

$$gen(S,K,F,O) = \frac{\sum_r weight(s_r)}{|S|}. \qquad (5)$$

The veracity of the numerical of generalization rate evaluation is based on manually expert assessments.

Optimization is due to the adjustment of the weights of key elements, as well as due to the weights of the dictionaries.

Let us consider some examples. We will calculate the weight for the following syntactic sentence "*The overall standard of living in Kazakhstan is falling: the quality of education, health services and culture are declining*" – «*В результате – науки нет, больницы не лечат, школы не учат, вузы не дают должного образования, театры не востребованы, телевидение бездарно, искусство вторично и низкопробно».* In the sentence, three words and a phrase from the set of keywords *K*: "*Kazakhstan*", "*falls*", "*the general standard of living,*" "*decline*." The generalization coefficient α is set equal to 3 (accessors got it by empirical way). We calculate the weight of the generalization of the sentence by formula (2). Hence, the weight is equal to 3 * (1 + 1 + 1 + 1) = 12.

Another sentence "*All as one: adults and children, observed strict discipline, called for strengthening the unity of the group*" – «*Все как один: взрослые и дети, соблюдали жесткую дисциплину, призывали укреплять единство групп*» contains three key words from the set *K*: "*all as one*", "*adults*", "*children*" and two word-modifier expressing the author's opinion from the set O: "*hard*" and "*unity of the group*". Word-modifiers that express the author's opinion weaken the weight of the sentence as their initial weight is 0.75. Then the total weight of these words is equal to 0.75 * 0.75. Three keywords give a weight equal to 9 (nine). As a result, according to the above formula (5), the total weight of generalization will be 9 * 0.56 = 5.04. The example clearly shows that the presence of word-modifiers significantly weakens generalization.

## Filling in dictionaries

The main difficulty of lexical approach is the formation of dictionaries. Researchers [10] distinguish the following techniques for a set of dictionaries: (i) manual collection of dictionaries, (ii) using a textual body, (iii) using templates of dictionaries or ready-made dictionaries, with this approach, a set of words by hand, supplement it by searching for new words in the dictionaries of WordNet or RussNet, or to take already ready dictionaries, for example, synonyms and antonyms. For the collection of dictionaries, the authors used manual collection to form a dictionary of basic words, ready dictionaries from the vocabulary of sentiment RuSentiLex by N.V. Lukashevich and A.V. Levchik [21].

The dictionary of basic elements has formed a set of basic elements of generalization $K=\{k_1, k_2, ..., k_l\}$, where l is the dictionary size. Categories of basic elements were formed on the basis of the work of researchers Dankova [5], Smith [9], Orlova [8], Frolova [4]. The set of basic elements of K includes the following categories of words for the generalization evaluation:

- non-specific verbs that allow you to talk about some process indefinitely, to draw some conclusion, not showing how you came to it. For example, "by doing so, you will understand that this is correct";
- non-specific nouns: "people", "citizens", "many". those nouns that try to generalize a single phenomenon to the general;
- non-specific pronouns: "all", "we", "they", "this", "those." For example, "everyone understands that schools do not teach" or "this is what we call" confidence in the future;
- universal generalists are words that do not allow exceptions: "all", "every", "never", "always";
- lexical units with the value of regularity/irregularity, such as "usually", "rarely", "infrequently", "spontaneous", "never";
- quantitative indicators, such as "every tenth", "mass", "grown up" lexical units with semantics of universality. Examples: "known", "indicative";
- the stylistic devices containing metaphors with a generalizing meaning. Examples: "the so-called quality of life," "poke your nose."

The authors put forward the hypothesis that the author's own opinion in the form of words, phrases or introductory speech or a confirmed event, phenomenon, incident (*fact*) weakens generalization. Fact in our research means concrete event or words from RuSentiLex, which marked as *fact*. For example, in the following sentences:

(1) "All as one: adults and children, observed a strict discipline, called for strengthening the unity of the group". –

«Все как один: взрослые и дети, соблюдали жесткую дисциплину, призывали укреплять единство групп»

In the first sentence, the phrase "*discipline*" from the dictionary RuSentiLex, marked in as *fact*, weakens the generalization of the basic words " *adults and children*". The opinion words "*strict*", expressing the author's opinion, also weaken the proposal's generalization "*all as one*". We call such elements, weakening generalization, modifier elements. Dictionaries of modifier elements were formed on the basis of ready-made dictionaries of sentiment.

(2) "Overall level of unprofessionalism and just amateurism is growing in all spheres of life". –

«Растет общий уровень непрофессионализма и просто дилетантства во всех сферах жизни»

G. in (2) is formed by phraseological combinations (*Overall level is growing, in all spheres of life*) and words with negative semantics (*unprofessionalism, amateurism*). The author's reasoning is based on generalization, the transition from the particular to the general. Thus, the situation is absolutized. Properties of *unprofessionalism and amateurism* are attributed by the author as permanent, extending to all spheres. The negative processes' scale meaning is affirmed. This sentence is very strong of G. and there no *facts*, which weakens the generalization.

## Experiments

To test the proposed approach for automatic evaluation of generalization in texts, a special software was developed in C# witch uses SQLite for using dictionaries efficiently and Regular Expressions for detection the similar language forms expressed by diversity in Russian language. Databases were added including lemmatization and morphology dictionaries using recognizing rules and regular expressions for automatic searching, extraction of generalization expressions and numerical estimation of the G. Examples of the regular expressions' use in dictionaries (fig.1):



Fig.1. Examples of the regular expressions
Рис. 1. Примеры регулярных выражений

| | |
|---|---|
| \d+\.?\d{1,4}миллионов | \d+\.?\d* millions |
| привычны\w{1,4} | regular\w{1,4} |
| кажд\w{1,4} | each\w{1,4} |
| кажд\w{1,4} втор\w{1,4} | each\w{1,4} second\w{1,4} |
| кажд\w{1,4} трет\w{1,4} | each\w{1,4} third\w{1,4} |

The collected dictionary of key elements has a hierarchical structure, namely, for phrases there is a basic form, for example, “each” and its specification “every second”. The latter has a greater weight of generalization.

In order to minimize the difference between the expert evaluation of text generalization and predicted programmatically, we adjusted the parameters of weighting factors, such as weights of words, phrases, introductory turns from the key dictionary, weights of modifier elements, G. coefficient α. The validation of the words included in the dictionary is checked by the experts themselves, on the basis of cross-assessments. In the future, the validity of the dictionaries will be further verified taking into account the importance of the experts themselves based on the Delphi method. Weights parameters, rules, dictionary elements will be adjusted until the difference between the expert evaluation and the program evaluation becomes minimal.

To carry out experiments with numerical evaluation of G. and further comparative analysis of evaluations obtained by expert and programmatic methods, the authors used publications of official and semi-official media of the Republic of Kazakhstan, which are publicly available. The sample included 30 articles, divided by experts on the degree of generalization. In total, experts determined five degrees of generalization by analogy with the correlation coefficients: very weak generalization, weak generalization, medium generalization, strong and very strong generalization.

Tables 1 and 2 describe the total number of proposals in publications, the number of proposals with generalization, the numerical rating of the publication’s generalization for a sample of five publications. The publications were selected taking into account the expert evaluation of the degree of generalization.

Table 1: Results of comparison of the algorithm and the expert evaluation obtained using the dictionary according to Dankova [5]
Таблица 1. Результаты сравнения алгоритма и экспертной оценки, полученные с использованием словаря по Данковой [5]

| Ref. | Num. of sentences | Generalization evaluation | Expert estimate |
|---|---|---|---|
| (Duvanov, 2018)[1] | 45 | 1.0 | very strong |
| (Bompiyeva, 2018)[2] | 70 | 0 | very weak |
| (Satpayev D, 2018)[3] | 46 | 0.04 | average |
| (Tukpiyev, 2018)[4] | 157 | 0.08 | weak |

Table 2: Results of the algorithm and expert evaluation using the dictionary according to Dankova [5], Smith [9], Orlova [8], Frolova [4]
Таблица 2. Результаты алгоритма и экспертной оценки с использованием словаря по Данковой [5], Смиту [9], Орловой [8], Фроловой [4]

| Ref. | Num. of sentences | Generalization evaluation | Expert estimate |
|---|---|---|---|
| Duvanov | 45 | 1.24 | very strong |
| Bompiyeva | 70 | 0.31 | very weak |
| Dosymov | 46 | 0.8 | average |
| Tukpiyev | 157 | 0.5 | weak |
| Agentstvo[5] | 12 | 0.04 | very weak |
| Isabayeva[6] | 104 | 0.64 | average |

## Conclusion and Future Work

This paper introduces an algorithm for numerical evaluation of generalization, based on a lexicon-based approach, and its comparison with expert evaluation. The minimization of the difference between the expert evaluation and the algorithmic estimation is achieved by optimizing the parameters of the weightings of the key dictionary, elements-modifiers and the generalization coefficient α.

Words, phrases, opening phrases of the generalization selected according to the researchers in psychology and applied computer linguistics were selected to the dictionary. The dictionaries were formed manually and based on the ready-made dictionaries and opening phrases of the Russian National Corpus.

Numerical calculations were performed using a special software application on C#, with the connection of SQLite, Regular Expressions. The algorithm of numerical estimation was tested on a small text-based corpus. Numerical estimates are obtained, the accuracy of the text recognition and the correlation between the algorithmic estimation and the expert one are observed. The experiment showed that the numerical estimate has an average degree of correlation between the algorithmic estimate of the publication’s generalization and its expert evaluation.

According to our research program, we will test the proposed algorithm on a large text-based corpus with its volume over 30 million words, seven years of official media of the Republic of Kazakhstan, articles to be collected from open sources will be stored.

In the future works, the sentiment classification, the joint influence of the generalization and sentiment of the publication on the audience will be considered, and extend the research for the Kazakh language.

## Acknowledgments

This research is conducted within the framework of the grant num. BR05236839 «Development of information technologies and systems for stimulation of personality’s sustainable development as one of the bases of development of digital Kazakhstan»

[1]Duvanov S. How to stop the «brain drain» from Kazakhstan [Электронный ресурс] // Ca-portal.ru, November 21, 2017. URL: http ://www.ca-portal.ru/article:39057 (дата обращения: 01.06.2018). (In Russian)
[2]Bompiyeva Zh. Kazakhstan needs serious results in the field of professional education [Электронный ресурс] // Tengrinews.kz, February 28, 2018. URL: https://tengrinews.kz/conference/240/ (дата обращения: 01.06.2018). (In Russian)
[3]Satpayev D. Linguistic decolonization [Электронный ресурс] // Forbes.kz, February 21, 2018. URL: https://forbes.kz/life/opinion/dosyim_satpaev_lingvisticheskaya_dekolonizatsiya/ (дата обращения: 01.06.2018). (In Russian)
[4]Tukpiyev Zh. There is nothing to breathe! In which cities of Kazakhstan is the most dangerous air? [Электронный ресурс] // Kazakhstanskaya Pravda, March 16, 2018. URL: http://www.kazpravda.kz/news/obshchestvo/nechem-dishat-v-kakih-gorodah-kazahstana-samii-opasnii-vozduh/ (дата обращения: 01.06.2018). (In Russian)
[5]In Uzbekistan, because of the twos on the foreheads of first-graders, the entire school leadership was fired [Электронный ресурс] // Informburo.kz, May 16, 2018. URL: https://informburo.kz/novosti/v-uzbekistane-iz-za-dvoek-na-lbah-pervoklassnikov-uvolili-vsyo-rukovodstvo-shkoly.html (дата обращения: 01.06.2018). (In Russian)
[6]Isabayeva S. Religious tomorrow. Is an “Islamic revival” threatening Kazakhstan and what is it fraught with? [Электронный ресурс] // Central Asia Monitor, May 16, 2018. URL: https://camonitor.kz/31118-religioznoe-zavtra-grozit-li-kazahstanu-islamskoe-vozrozhdenie-i-chem-eto-chrevato.html (дата обращения: 01.06.2018). (In Russian)

## References

[1] Teun A. van Dijk. Discourse, Ideology and Context. *Folia Linguistica*. 2001; 35(1-2):11–40. DOI: 10.1515/flin.2001.35.1-2.11

[2] Teun A. van Dijk. Prejudice in Discourse. An Analysis of Ethnic Prejudice in Cognition and Conversation. Amsterdam: Benjamins. 1984. 172 p.

[3] Ajiboye E. Ideological Discourse Analysis of the Functions of Feedback Comments on Online Reports of Socio-political Crises in Nigeria. *Covenant Journal of Language Studies*. 2013. 1(2):128-147. Available at: https://journals.covenantuniversity.edu.ng/index.php/cjls/article/view/43/38 (accessed 01.06.2018).

[4] Frolova I.V. On expression of subjectivity in analytical articles of the British and Russian quality press. *Political Linguistics*. 2015; 1(51):138-145. Available at: https://elibrary.ru/item.asp?id=23590795 (accessed 01.06.2018). (In Russian)

[5] Dankova N.S. Ggeneralization strategy as a means of representing judicial power (on the material of Russian and English print media). *Political Linguistics*. 2016; 1(55):73–81. Available at: https://elibrary.ru/item.asp?id=25718823 (accessed 01.06.2018). (In Russian)

[6] Recasens M., Danescu-Niculescu-Mizil C., Jurafsky D. Linguistic Models for Analyzing and Detecting Biased Language. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 2013. Vol. 1. Pp. 1650–1659.

[7] Morstatter F., Wu L., Yavanoglu U., Corman S.R., Liu H. Identifying Framing Bias in Online News. *ACM Transactions on Social Computing*. 2018; 1(2): Article 5. DOI: 10.1145/3204948

[8] Orlova O.G. Russia stereotypes glossary: manner and matter. *Political Linguistics*. 2013; 3(45):175–182. Available at: https://elibrary.ru/item.asp?id=20378624 (accessed 01.06.2018). (In Russian)

[9] Smit S. Master the power of suggestion! Achieve all you want! [Ovladejte siloj vnushenija! Dobivajtes' vsego chego hotite!]. M.: Izdatel'stvo AST, 2014. 448 p. (In Russian)

[10] D'Andrea A., Ferri F., Grifoni P., Guzzo T. Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications*. 2015; 125(3):26-33. DOI: 10.5120/ijca2015905866

[11] Vinodhini G., Chandrasekaran R.M. A sampling based sentiment mining approach for e-commerce applications. *Information Processing & Management*. 2017; 53(1): 223-236. DOI: 10.1016/j.ipm.2016.08.003

[12] Maynard D., Funk A. Automatic Detection of Political Opinions in Tweets. R. García-Castro, D. Fensel D., G. Antoniou G. (Eds.) *The Semantic Web: ESWC 2011 Workshops. ESWC 2011*. Lecture Notes in Computer Science, Vol. 7117. Springer, Berlin, Heidelberg, 2012. Pp. 88-99. DOI: 10.1007/978-3-642-25953-1_8

[13] Thakkar H., Patel D. Approaches for Sentiment Analysis on Twitter: A State-of-Art study. CoRR. Vol. abs/1512.01043. 2015. Available at: http://arxiv.org/abs/1512.01043 (accessed 01.06.2018).

[14] Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*. 2011; 37(2):267–307. DOI: 10.1162/COLI_a_00049

[15] Repaka R., Pallelra R.R., Koppula A.R., Movva V.S. UMDuluth-CS8761-12: A Novel Machine Learning Approach for Aspect Based Sentiment Analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, 2015. Pp. 742-747. Available at: http://aclweb.org/anthology/S/S15/S15-2126.pdf (accessed 01.06.2018).

[16] Rice D., Zorn C. Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Proceedings of NDATAD*. 2013. Pp. 98-115. Available at: https://static1.squarespace.com/static/5557a550e4b0443afbea6783/t/5b569390758d4635216e5a73/1532400530505/Rice-Zorn-DictionaryPaper-PS-RM-R%26R.pdf (accessed 01.06.2018).

[17] Whitehead M., Yaeger L. Sentiment Mining Using Ensemble Classification Models. T. Sobh (Ed.) *Innovations and Advances in Computer Sciences and Engineering*. Springer, Dordrecht, 2010. Pp. 509-514. DOI: 10.1007/978-90-481-3658-2_89

[18] Mudinas A., Zhang D., Levene M. Combining lexicon and learning based approaches for concept-level sentiment analysis. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'12).* ACM, New York, NY, USA, 2012. Article 5, 8 pages. DOI: 10.1145/2346676.2346681

[19] Turney P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2002. Pp. 417-424. DOI: 10.3115/1073083.1073153

[20] Ualiyeva I., Krassovitsky A., Mussabayev R. Generalization rate evaluation in open publication materials in Russian language. *Proceedings of the 16th International Scientific Conference Information Technologies and Management.* April 26-27, 2018, ISMA University, Riga, Latvia, 2018. Pp. 81-82. Available at: https://www.isma.lv/FILES/SCIENCE/IT&M2018_THESES/02_CMIT/30_IT&M2018_Ualieva_Krassovitskiy_Mussabayev.pdf (accessed 01.06.2018).

[21] Loukachevitch N.V., Levchik A.V. Creating Russian Sentiment Lexicon. *Open Semantic Technologies for Intelligent System*. 2016; 6:377-382. Available at: https://elibrary.ru/item.asp?id=30080327 (accessed 01.06.2018). (In Russian)

Submitted 01.06.2018; revised 20.08.2018;
published online 30.09.2018.

## Список использованных источников

[1] *Teun A. van Dijk.* Discourse, Ideology and Context // Folia Linguistica. 2001. Vol. 35, issue 1-2. Pp. 11–40. DOI: 10.1515/flin.2001.35.1-2.11

[2] *Teun A. van Dijk.* Prejudice in Discourse. An Analysis of Ethnic Prejudice in Cognition and Conversation. Amsterdam: Benjamins. 1984. 172 p.

[3] *Ajiboye E.* Ideological Discourse Analysis of the Functions of Feedback Comments on Online Reports of Socio-political Crises in Nigeria // Covenant Journal of Language Studies. 2013. Vol. 1, issue 2. Pp. 128-147. URL: https://journals.covenantuniversity.edu.ng/index.php/cjls/article/view/43/38 (дата обращения: 01.06.2018).

[4] *Фролова И.В.* О выражении субъективности в аналитических статьях качественной британской и российской прессы // Политическая лингвистика. 2015. № 1(51). С. 138-

145. URL: https://elibrary.ru/item.asp?id=23590795 (дата обращения: 01.06.2018).

[5] *Данкова Н.С.* Стратегия генерализации как средство репрезентации судебной власти (на материале российских и английских печатных СМИ // Политическая лингвистика. 2016. № 1(55). С. 73–81. URL: https://elibrary.ru/item.asp?id=25718823 (дата обращения: 01.06.2018).

[6] *Recasens M., Danescu-Niculescu-Mizil C., Jurafsky D.* Linguistic Models for Analyzing and Detecting Biased Language // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013. Vol. 1. Pp. 1650–1659.

[7] *Morstatter F., Wu L., Yavanoglu U., Corman S.R., Liu H.* Identifying Framing Bias in Online News // ACM Transactions on Social Computing. 2018. Vol. 1, issue 2, Article 5, pp. 1–18. DOI: 10.1145/3204948

[8] *Орлова О.Г.* Составление словаря стереотипов о русских: содержание и форма // Политическая лингвистика. 2013. № 3(45). С. 175–182. URL: https://elibrary.ru/item.asp?id=20378624 (дата обращения: 01.06.2018).

[9] *Смит С.* Овладейте силой внушения – добивайтесь всего, чего хотите. М.: Издательство АСТ, 2014. 448 с.

[10] *D'Andrea A., Ferri F., Grifoni P., Guzzo T.* Approaches, Tools and Applications for Sentiment Analysis Implementation // International Journal of Computer Applications. 2015. Vol. 125, issue 3. Pp. 26-33. DOI: 10.5120/ijca2015905866

[11] *Vinodhini G., Chandrasekaran R.M.* A sampling based sentiment mining approach for e-commerce applications // Information Processing & Management. 2017. Vol. 53, issue 1. Pp. 223-236. DOI: 10.1016/j.ipm.2016.08.003

[12] *Maynard D., Funk A.* Automatic Detection of Political Opinions in Tweets / R. García-Castro, D. Fensel D., G. Antoniou G. (Eds.) // The Semantic Web: ESWC 2011 Workshops. ESWC 2011. Lecture Notes in Computer Science, Vol. 7117. Springer, Berlin, Heidelberg, 2012. Pp. 88-99. DOI: 10.1007/978-3-642-25953-1_8

[13] *Thakkar H., Patel D.* Approaches for Sentiment Analysis on Twitter: A State-of-Art study // CoRR. Vol. abs/1512.01043. 2015. URL: http://arxiv.org/abs/1512.01043 (дата обращения: 01.06.2018).

[14] *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* Lexicon-Based Methods for Sentiment Analysis // Computational Linguistics. 2011. Vol. 37, issue 2. Pp. 267–307. DOI: 10.1162/COLI_a_00049

[15] *Repaka R., Pallelra R.R., Koppula A.R., Movva V.S.* UMDuluth-CS8761-12: A Novel Machine Learning Approach for Aspect Based Sentiment Analysis // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, Denver, Colorado, 2015. Pp. 742-747. URL: http://aclweb.org/anthology/S/S15/S15-2126.pdf (дата обращения: 01.06.2018).

[16] *Rice D., Zorn C.* Corpus-based dictionaries for sentiment analysis of specialized vocabularies. Proceedings of NDATAD. 2013. Pp. 98-115. URL: https://static1.squarespace.com/static/5557a550e4b0443afbea6783/t/5b569390758d4635216e5a73/1532400530505/Rice-Zorn-DictionaryPaper-PS-RM-R%26R.pdf (дата обращения: 01.06.2018).

[17] *Whitehead M., Yaeger L.* Sentiment Mining Using Ensemble Classification Models / T. Sobh (Ed.) // Innovations and Advances in Computer Sciences and Engineering. Springer, Dordrecht, 2010. Pp. 509-514. DOI: 10.1007/978-90-481-3658-2_89

[18] *Mudinas A., Zhang D., Levene M.* Combining lexicon and learning based approaches for concept-level sentiment analysis // Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'12). ACM, New York, NY, USA, 2012. Article 5, 8 pages. DOI: 10.1145/2346676.2346681

[19] *Turney P.D.* Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 2002. Pp. 417-424. DOI: 10.3115/1073083.1073153

[20] *Ualiyeva I., Krassovitsky A., Mussabayev R.* Generalization rate evaluation in open publication materials in Russian language // Proceedings of the 16th International Scientific Conference Information Technologies and Management. April 26-27, 2018, ISMA University, Riga, Latvia, 2018. Pp. 81-82. URL: https://www.isma.lv/FILES/SCIENCE/IT&M2018_THESES/02_CMIT/30_IT&M2018_Ualieva_Krassovitskiy_Mussabayev.pdf (дата обращения: 01.06.2018).

[21] *Лукашевич Н.В., Левчик А.В.* Создание лексикона оценочных слов русского языка РуСентилекс // Открытые семантические технологии проектирования интеллектуальных систем. 2016. № 6. С. 377–382. URL: https://elibrary.ru/item.asp?id=30080327 (дата обращения: 01.06.2018).

Поступила 01.06.2018; принята в печать 20.08.2018;
опубликована онлайн 30.09.2018.

Об авторах:

**Красовицкий Александр Михайлович**, Институт информационных и вычислительных технологий (050001, Казахстан, г. Алматы, ул. Пушкина, д. 125), ORCID: http://orcid.org/0000-0003-2948-374X, akrassovitskiy@gmail.com

**Уалиева Ирина Маратовна**, кандидат физико-математических наук, ведущий научный сотрудник, лаборатория анализа и моделирования информационных процессов, Институт информационных и вычислительных технологий (050001, Казахстан, г. Алматы, ул. Пушкина, д. 125), ORCID: http://orcid.org/0000-0003-3853-8896, i.ualiyeva@gmail.com

**Мейрамбеккызы Жазираханым**, Институт информационных и вычислительных технологий (050001, Казахстан, г. Алматы, ул. Пушкина, д. 125), ORCID: http://orcid.org/0000-0003-2309-7725, jazirahanim@mail.ru

**Мусабаев Рустам Рафикович**, Институт информационных и вычислительных технологий (050001, Казахстан, г. Алматы, ул. Пушкина, д. 125), ORCID: http://orcid.org/0000-0001-7283-5144, rmusab@gmail.com