Using Morphological and Semantic Features for the Quality Assessment of Russian Wikipedia *

Włodzimierz Lewoniewski[1], Nina Khairova[2], Krzysztof Węcel[1], Nataliia Stratiienko[2], Witold Abramowicz[1]

[1]Poznań University of Economics and Business, Poland
[2]National Technical University "Kharkiv Polytechnic Institute", Ukraine
wlodzimierz.lewoniewski@ue.poznan.pl, khairova@kpi.kharkov.ua,
krzysztof.wecel@ue.poznan.pl, strana@kpi.kharkov.ua,
witold.abramowicz@ue.poznan.pl

Abstract. Nowadays, the assessment of the quality and credibility of Wikipedia articles becomes increasingly important. We propose to use morphological and semantic features to estimate the quality of Wikipedia articles in Russian language. We distinguished over 150 linguistic features and divided them into four groups. In these groups, we considered the features of encyclopedic style, readability and subjectivism of the article's text. Based on Random Forest as a classification algorithm, we show the most importance linguistic features that affect the quality of Russian Wikipedia articles. We compare the classification results of our four linguistic features groups separately. We have achieved the F-measure of 89,75%.

Keywords: quality assessment of texts, morphological and semantics features, Russian Wikipedia articles, random forests classification, encyclopedic, readability, subjectivism

1 Introduction

Nowadays, Wikipedia is the biggest public universal encyclopedia with a free content, which includes over 44 million articles. Most articles in the Wikipedia are comparable in quality to those in the Encyclopedia Britannica [1]. Usually, in order for a Wikipedia article to reach the good quality it must be revised by Wikipedia community many times. This is the main reason for the growing interest and popularity of research on assessment of Wikipedia articles quality.

In 2006, during the Opening plenary at Wikimania, Jimmy Wales suggested concentrating on quality of the articles instead of their number [10]. The best articles of Wikipedia must follow the specific style guidelines. Such guidelines can be quantified in many ways. One of the approaches is to use morphological, syntactic and semantic features of words, which allow evaluating the quality of the Wikipedia articles. Obviously, these features strongly depend on a specific language.

* This is a preprint version. The final publication is available at Springer via https://doi.org/10.1007/978-3-319-67642-5_46

As of April 2017, the Russian-language edition of Wikipedia had more than 1,3 million articles[1] and more than 1 billion page views per month[2]. The Russian Wikipedia subdomain (ru.wikipedia.org) receives approximately 8% of Wikipedia's cumulative traffic, and takes second place after English subdomain (59%, en.wikipedia.org).[3]

There are a lot of articles that study the correlation between English linguistic characteristics and estimating the quality of articles in English Wikipedia. However, studies examining the use of Russian linguistic characteristics to evaluate the quality of texts are very few.

In this paper we focus on using morphological and semantics features of the Russian language to estimate the quality of Russian Wikipedia articles. We suggest applying the Random Forests algorithm of that is based on these features in order to automatically identify quality classes of Wikipedia articles.

2 Related work

All experts admit that there are some difficulties in determining the quality of the Wikipedia articles. Furthermore Wikipedia isn’t a static resource; their amount keeps growing every day. Also that fact that the articles cover different topics complicates the task [12]. It means it requires that experts from different disciplines judge the quality, but such experts are not always available.

Measuring an article’s quality in Wikipedia is not an easy task for human users, complexity of which repeatedly increases in case of the task of automatic evaluation of the article quality. Now there exist enough studies concerning the problems related to automatic estimating the quality of Wikipedia articles. We can divide all research literature into three groups. The first group of researches is based on characteristics related to contributors' reputations and edit network, article status, external factual support and other features [5,18]. However, often such methods require complex calculations and they do not analyze on the content of the article itself.

The second group of the studies focuses on the calculation of volume of different articles components. These studies showed that a better quality article usually are longer, have more images and sections, use bigger number of references [15, 16, 8]. These quantitative features are used in online service WikiRank[4] for the automatic relative assessment of the articles in various language versions of Wikipedia. In some Wikipedia articles we can find special quality flaw templates, which can also help in articles assessment [3].

The third group of the studies concerning the task of automatic estimating the quality based on linguistic characteristics of text in Wikipedia articles [2, 6]. Other studies used linguistic features to examine how density of factual information impact on quality of Wikipedia articles [14, 19]. Such approaches that direct to exploring the

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[2] https://analytics.wikimedia.org

[3] http://www.alexa.com/siteinfo/wikipedia.org

[4] http://wikirank.net

linguistic characteristics of articles might be useful for improvement of the articles quality. For example, it concerns such characteristics as the writing style of an article, the number of verbs, facts, the number of diverse nouns and similar features. However, linguistic characteristics of the text depend on the article’s language. Nowadays, Wikipedia contains articles in approximately 300 languages. One of the main language versions of the online encyclopedia is Russian. There a lot of articles on using linguistic characteristics to estimate the quality of Wikipedia articles in English or Spanish but very few use peculiar properties of Russian linguistic characteristics [11].

This is the first study that use more than 150 features related to Russian language to predict articles quality in Wikipedia. In order to tokenize texts of Russian Wikipedia articles and extract various linguistic features we use own approach. This approach use different open morphological libraries and dictionaries available on the Web. We also add additional rules to this algorithm at the stage of preparation of the text, as well as during the extraction of some features.

3 Description of the experiment

The best Wikipedia articles must be well-written, comprehensive, well-researched, neutral and must follow the specific style guidelines.[5] The main idea of the approach is that the linguistic features of words or sentences of the articles allow evaluating the style of writing, the brevity, correctness, readable and some others of the Wikipedia articles characteristics. In some cases, semantic and syntactic features of the words allow even to evaluate subjectivity of the article authors.

3.1 Linguistic features

We distinguish several groups of linguistics features that can affect the quality of Russian Wikipedia articles. The first group includes **morphological features** such as parts of speech, specific morphological characteristics of a particular part of speech. For instance, we determine the number of verbs and then we determine the number of verb categories - tense, person, etc. Herewith, we use more than 50 similar characteristics. In order to analyze the morphological features, we apply the pymorphy2[6], the library for morphological analysis of the Russian language that is based on the OpenCorpora dictionary[7] which is also used to denote grammatical tags (some of them are presented in Table 1).

The second group of the applicable linguistic features includes some **semantic features**, integral morphological features of the words and even the parameters of word formation. We suppose that the features from the second group can explicitly express the existence of some subjective assessment or opinion of the Wikipedia

[5] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

[6] http://pymorphy2.readthedocs.io

[7] http://opencorpora.org

article authors. Therefore, the presence of these characteristics in the text can affect the quality of the article.

Typically the value judgments are represented by the various linguistic means and characteristics in the text. For example, such morphological features as personal and possessive pronouns of the first and second person can contribute evaluative-expressive shades to a statement. Herewith, one of the main grammatical means of adding of the author's subjectivity and expressiveness in Russian is affectionate diminutive suffixes.

Table 1. Description of some grammatical tags used in the study. Source: http://opencorpora.org/dict.php?act=gram

NOUN	noun	*NUMR*	numeral
ADJF	adjective (full)	*ADVB*	adverb
ADJS	adjective (short)	*NPRO*	pronoun
COMP	comparative	*PRED*	predicative
VERB	verb (personal form)	*PREP*	preposition
INFN	verb (the infinitive)	*CONJ*	conjunction
PRTF	participle (full)	*PRCL*	particle
PRTS	participle (short)	*INTJ*	interjection
GRND	gerund	…	

Moreover, each natural language has a specific vocabulary that expresses emotions, mentality and adds a tinge of author's opinion in the statement. We have created two special vocabularies that express such shade in Russian. The first vocabulary includes more than 300 words and the combination of words (*avt_ocenka*). The second one includes only verbs that have the certain semantic component of subjectivity (*menverb*). It includes 120 speech verbs (such as tell, recall, dictate and others), 154 feelings verbs and 103 emotions verbs (such as wish, rejoice, worry and others) [14]. Additionally, in this group of the features, we use the glossary of introductory turnovers from the Russian National Corpus.[8]

Table 2 shows our full list of the word features that can express some elements of subjective assessment of the Wikipedia article authors.

Table 2. Linguistic features of the words that can express some elements of subjective assessment of the Wikipedia article authors

lichprit	– personal and possessive pronouns of the first and second person
formal_priz	– dative case with a preposition
ocen	– affectionate diminutive suffixes
avt_ocenka	– the special vocabulary
ruscorp_parenth	– the glossary of introductory turnovers from Russian National Corpus
sl_by	– the use of the subjunctive

[8] http://www.ruscorpora.ru/en/

menverb	– the special vocabulary of the verbs that have the certain semantic component of subjectivity
VERB_wmv	– the verb that does not have the semantic component of subjectivity

The third group of the applicable linguistic features allows making exploratory conclusions about the **readability of the texts**. We have included in this group both characteristics that are commonly used to assess the complexity of texts as well as new characteristics based on dictionaries of the Russian National Corpus, the Russian Internet corps I-RU [13] and the Open Corpora. Traditionally the estimation of readability is based on features such as the statistical average word length (in characters and in syllables), the sentence length, the maximum number of words in a sentence, the number of unique words (*uslov*) and some others [12].

In addition to the listed characteristics, we also highlight the following statistical indicators: the number of words having 3 syllables and more (*slog3*), the number of words having 4 syllables and more (*slog4*), the number of words having 5 syllables and more (*slog5*), the number of unique words of specific parts of speech (*uverb*, *unoun*, *uadj*).

Furthermore, we assume that the frequency of word usage in texts correlates with their comprehensibility and readability. Therefore, we can include the lists of the most frequent words in the Russian language in the third group of the linguistic features that affect the readability of the texts. Table 3 shows these features that take into account different lists of the most frequent words in the Russian language.

Table 3. Features that take into account different lists of the most frequent words in the Russian language.

frec100 (...500, ...1000, ...5000)	– the 100 (500, 1 000, 5 000) first most common words in the Russian Internet corps I-RU
slovoformy100 (...500, ...1000, ...5000, ...10000)	– the 100 (500, 1 000, 5 000, 10 000) first most common words in the Russian National Corpus
2grammy100 (...500, 1000, ...5000, ...10000)	– the 100 (500, 1 000, 5 000, 10 000) first most common bigrams in the Russian National Corpus
3grammy100	– the 100 first most common 3-grams in the Russian National Corpus
4grammy100	– the 100 first most common 4-grams in the Russian National Corpus
5grammy100	– the 100 first most common 5-grams in the Russian National Corpus
oc100un (oc500un, oc1000un, oc5000un, oc10000un	– the 100 (500, 1 000, 5 000, 10 000) first most common unigrams in Open Corpora.
oc100bi (oc500bi, oc1000bi, oc5000bi, oc1000bi)	– the 100 (500, 1 000, 5 000, 10 000) first most common bigrams in Open Corpora.

oc100tri	– the 100 first most common 3-grams in Open Corpora

The total number of the third group of the applicable linguistic categories reaches 40.

The fourth group of the applicable linguistic features characterizes an **encyclopedic style** of an article. An encyclopedia-style article should display a comprehensive view of the subject matter in a simple and understandable manner. In the general case, such style means the condensed presentation of material, which identifies the subject sufficiently, completely, naturally and authentically.

We argue that such style can be represented explicitly by the various linguistic means and characteristics in the text. We have included in this group such proper names as the first name of the person (*name*), the last name of the person (*surn*), the middle name of the person (*patr*), a name (*orgn*), and a trademark (*Trad*) of the organisation and toponyms (*Geox*). We also believe that the list of the most popular words of Russian Wikipedia can represent the encyclopedic style of the article (*250wiki*)

Additionally, we have included amounts of simple and complex facts of the article to the fourth group of the applicable linguistic features. According to the logical-linguistic model of fact extraction from English [7] or Russian Texts [14], the simple fact (*fact*) in a Russian sentence is the smallest grammatical clause that includes a verb and a noun; the complex fact (*FactPlus1*, *FactPlus2*) in Russian texts is a grammatical sentence that includes a verb and a few nouns. Among these nouns, one has to play the semantic role of the Subject (*FactPlus1*) and the other has to be the Object (*FactPlus2*)[9].

3.2 Source Data

Our dataset includes all articles from Russian Wikipedia that have manual evaluation of their quality, i.e. about 130,000 (April 2017). According to the previous studies [15, 16], we distinguish two quality classes of the Russian Wikipedia articles. We called the first class GoodEnough: it includes articles that are evaluated by the Wikipedia community as Featured and Good. The second class is called NeedsWork; it includes I, II, III and IV level (stub) articles. One of the peculiarities of Russian Wikipedia is the availability of such an assessment of the quality of the article as Solid. According to the binary classification, this grade can be classified either as GoodEnough or NeedsWork. In order to show peculiarity of the group of articles that are evaluated as Solid, we consider three versions of the classification. They are *FG-standard*, *FGS-standard* and *FG-S standard*.

Table 4 shows the distributions of the analyzed articles according to the grade of assessment quality.

[9] Detailed definitions of the simple and complex facts are given in [14]

Table 4. The distributions of the analyzed articles according to the grade of assessment quality.

Quality Grade	Number of articles	FG standard	FGS standard	FG-S standard
Featured	997	GoodEnough	GoodEnough	GoodEnough
Good	2738			
Solid	3927	NeedsWork		Disabled
I level	2516		NeedsWork	NeedsWork
II level	9978			
III level	48183			
IV level (stub)	61711			

4 Implementation aspects and experimental results

Analysis has shown that usually, articles with high-quality grades have the higher value of a particular feature. On figure 1 is shown the distribution of some features among different quality grades in Russian Wikipedia.The used Random Forests classifier determines the probability that an article belongs to one of the two classes. The classifier allows us to use the specific analytical methods to explore hidden patterns, rules and dependencies between different linguistic features. At the same time, the Random Forests classifier allows calculating the predictive power of the different features and every group of the applicable linguistic features.

As already mentioned before, better articles usually have more text (including characters, words, sentences). So we can expect that the value of a majority of the considered linguistic characteristics is more in articles with better quality. Therefore, we decided to normalize all features by word count, sentences count and character count (without spaces) separately. On figure 2 it is shown distribution of some features normalized by words.

Typically, the encyclopedic style of a Wikipedia article requires that the article

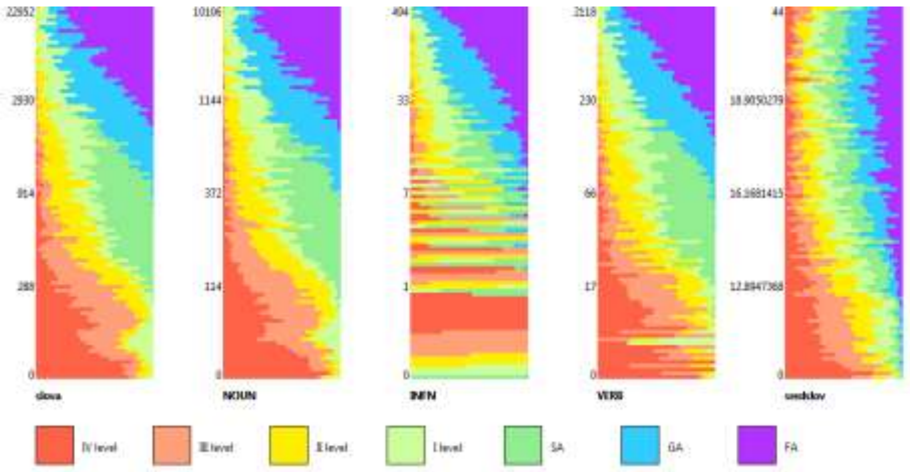

Figure 1. Value distribution of 5 articles features (from left to right: number of words, nouns, infinitives, verbs, avg. number of words in a sentence) among different quality grades in Russian Wikipedia. Source: own calculation.

includes a brief definition or description of the assigned subject, which is called "The lead section" followed by a broad examination of the topic, which is called "The 1st section" followed by a number of sub-sections. We have evaluated the precision, recall and F-Measure for three way of the normalization and for three analysed areas: the lead section, the 1st section, the whole article's text.

Table 5 shows that the evaluation of the linguistic parameters of the whole article is more significant than the evaluation of the linguistic parameters of the lead section and the 1st section only. According to the table, there is not much difference in F-measure between the various way of the normalization. We decided to normalize our features by the number of words based on the research of corpus linguistics [17].

Table 5. Classication results using various types of the normalisation and three versions of the classification standards.

	FGS standard		
Normalize by	**characters**	**words**	**sentence**
Lead section	75,24%	75,04%	75,59%
1st section	75,82%	75,38%	75,89%
Article text	81,47%	81,05%	80,76%

	FG standard		
Normalize by	**characters**	**words**	**sentence**
Lead section	81,68%	81,49%	81,44%
1st section	78,78%	78,74%	78,90%
Article text	89,54%	89,75%	89,50%

	FG-S standard		
Normalize by	**characters**	**words**	**sentence**
Lead section	81,98%	82,01%	82,03%
1st section	79,93%	79,85%	80,40%
Article text	88,81%	89,14%	88,85%

The Random Forest classifier can show the importance of features in the model. It provides two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy. Table 6 shows 30 most importance features, which are based on average impurity decrease. Table 7 shows 30 most important features based

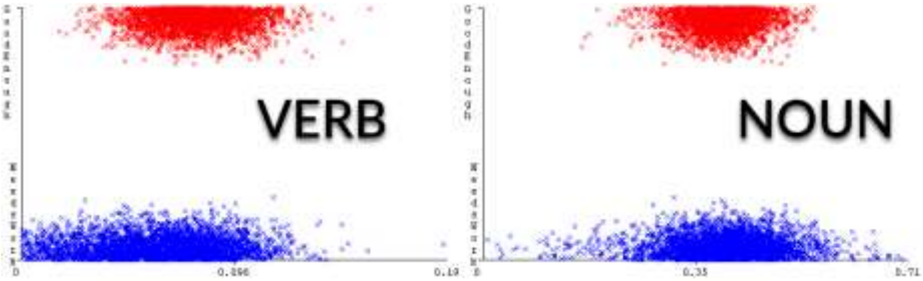

Figure 2. Distribution of normalized features (by the number of words) in quality classes. Source: own calculations in Weka.

on number of nodes using that attribute. Every feature is normalized by the number of words of the corpus class. Additionally, as was mentioned before, the linguistic parameters correspond to the whole article.

Table 6. 30 most important linguistic features based on average impurity decrease

0,52	VERB_wmv	0,44	PRTF	0,41	GRND
0,5	Fact	0,44	INFN	0,41	ADVB
0,49	FactPlus1	0,44	menverb	0,41	CONJ
0,48	FactPlus2	0,43	PREP	0,4	inan
0,47	FactPlus2_wmv	0,43	COMP	0,4	PRCL
0,47	Fact_wmv	0,43	PRTS	0,4	anim
0,47	FactPlus1_wmv	0,43	sred_dlin_slov	0,39	GNdr
0,46	ADJF	0,42	NUMR	0,39	voct
0,46	NOUN	0,42	ADJS	0,39	INTJ
0,46	VERB	0,42	PRED	0,39	NPRO

Table 7. 30 most important linguistic features based on number of nodes using that features

856	sred_dlin_slov	650	ADJS	588	sing
744	FactPlus1	645	INFN	587	PRTF
738	menverb	635	FactPlus1_wmv	577	anim
733	FactPlus2	621	nomn	576	PREP
723	VERB_wmv	618	NPRO	574	gent
690	makslov	617	FactPlus2_wmv	559	GRND
680	ADJF	608	PRTS	558	VERB
654	sredslov	607	Fact_wmv	551	inan
652	Fact	606	ADVB	538	Sgtm
651	NOUN	590	NUMR	537	PRCL

We found that except for the morphological categories the main features affecting the quality of Russian Wikipedia articles are such semantic characters as the simple fact or the complex fact [14], and such characters of the subjective assessment as a verb that have the certain semantic component of subjectivity. Moreover, one of the main feature to classify the Russian Wikipedia article are correlated features of the number of the verbs that do not have the semantic component of subjectivity and the number of the facts that do not have the semantic component of subjectivity.

We also analyzed the classification efficiency using separate parameters for each of our four linguistic features groups. The results reported in Table 8 were obtained using the random forest classifier with features of the encyclopedic, morphological, readability, subjectivism groups separately.

Additionally, we analyzed classification results using two versions of the classification standards. They are FGS standard and FG standard.

There are significant differences of results between the FGS version of classification and FG classification. The precision, recall and F-measure are significantly higher when Solid articles are referred to the class NeedsWork articles.

Table 8. Classication results using the encyclopedic, morphological, readability, subjectivism features groups separately.

Features group	FGS standard			FG standard		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Encyclopedic	76,7%	76,5%	76,6%	82,4%	82,4%	82,4%
Morphological	80,7%	80,6%	80,7%	87,9%	87,6%	87,7%
Readability	79,8%	79,7%	79,7%	88,4%	88,0%	88,1%
Subjectivism	76,5%	76,4%	76,4%	85,3%	84,8%	85,0%
All groups	**81,2%**	**81,0%**	**81,1%**	**89,9%**	**89,7%**	**89,8%**

5 Conclusions and Future Works

In this work, we proposed to exploit linguistic features of an article for assessing Wikipedia content quality. We distinguished and categorized over 150 linguistic features of Russian Wikipedia articles. We divided all the linguistic characteristics into four groups: morphological features, semantic features that can explicitly express the existence of some subjective assessment or opinion of the authors, the features that are exploratory conclusions about the readability of the text and the features that characterize the encyclopedic style of the article.

We found that the most important groups of linguistic characteristics that affect the quality of Russian Wikipedia articles are the parts of speech and semantic features of the simple fact and the complex fact. Moreover, such correlated features as the number of the verbs and the number of the facts that do not have the semantic component of subjectivity possess the great predictive power of classification of the quality of the articles. Our experiments on a subset of the Russian Wikipedia revealed that frequency dictionaries are poorly effective in the problem of classifying the quality of articles.

Our experiments showed that the evaluation of the linguistic features of the whole article is more significant than the evaluation of them for some sections of the text. We also investigated the use of three versions of the articles classification standards depending on the position of Solid Articles. Using FG schema allowed achieving the F-measure of the classification results of 89,75%.

While the initial results are very promising, more in-depth investigations of these linguistic features are needed. We guess that the most effective way is to apply our linguistic features with others parameters that affect the Wikipedia articles quality.

In future work, we plan to conduct similar experiments for other languages to analyze how linguistic features of different languages affects the quality of Wikipedia articles. Additionally, we are going to expand the list of semantic variables and also consider the quality of the articles in a more complex categorization.

6 References.

1. Michael B., (2015), *Wikipedia Or Encyclopædia Britannica: Which Has More Bias?,* Forbes, URL: http://www.forbes.com/sites/hbsworkingknowledge/2015/01/20/wikipedia-or-encyclopaedia-britannica-which-has-more-bias (access date: 15.06.2017)
2. Xu, Y., & Luo, T., (2011), *Measuring article quality in Wikipedia: Lexical clue model. In Web Society (SWS)*, 2011 3rd Symposium on IEEE, pp. 141-146.
3. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany (2013).
4. Kittur, A., Kraut, R. E., (2008), *Harnessing the wisdom of crowds in wikipedia: quality through coordination*, Proceedings of the 2008 ACM conference on Computer supported cooperative work, ACM, pp. 37-46
5. Velázquez, C. G., Cagnina, L. C., & Errecalde, M. L., (2017), *On the Feasibility of External Factual Support as Wikipedia's Quality Metric*, Procesamiento del Lenguaje Natural, 58, pp. 93-100.
6. Lipka, N., Stein, B., (2010), *Identifying Featured Articles in Wikipedia: Writing Style Matters*, Proceedings of the 19th International Conference on World Wide Web, pp. 1147–1148.
7. Khairova, N., Petrasova, S., Gautam, A., (2016), *The Logical-Linguistic Model of Fact Extraction from English Texts*. International Conference on Information and Software Technologies. CCIS 2016: Communications in Computer and Information Science, pp. 625-635.
8. Warncke-Wang, M., Cosley, D., & Riedl, J., (2013), *Tell me more: an actionable quality model for Wikipedia*, Proceedings of the 9th International Symposium on Open Collaboration.
9. Tausczik., Y., Pennebaker, J., (2010), *The psychological meaning of words: Liwc and computerized text analysis methods*, Journal of language and social psychology, vol. 29, no. 1, pp. 24–54.
10. Giles, G., *Internet encyclopaedias go head to head*, Nature, 438 (2005), pp. 900-901.
11. Panicheva, P., Ledovaya, Y., Bogolyubova, O., (2016), *Lexical, morphological and semantic correlates of the dark triad personality traits in russian facebook texts*. Artificial Intelligence and Natural Language Conference (AINL), IEEE, 2016, pp. 1-8.
12. Lenzner, T., (2014), *Are readability formulas valid tools for assessing survey question difficulty?*, Sociological Methods & Research 43 (4), p. 677-698.
13. Sharoff S., Umanskaya E., Wilson J., (2014), A frequency dictionary of Russian: Core vocabulary for learners, Routledge.
14. Khairova N., Lewoniewski W., Wecel K., (2017), Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction, International Conference on Business Information Systems (pp. 28-42)
15. Węcel, K., Lewoniewski, W., (2015), *Modelling the Quality of Attributes in Wikipedia Infoboxes*, Business Information Systems Workshops, Volume 228 of Lecture Notes in Business Information Processing. Springer International Publishing, pp. 308–320
16. Lewoniewski,W., Węcel, K., Abramowicz,W., (2016), *Quality and importance of Wikipedia articles in different languages*, Information and Software Technologies: 22nd International Conference, Druskininkai, Lithuania, October 13-15, 2016, Proceedings. Springer International Publishing, pp. 613–624.

17. Rebuschat, P. E., Detmar, M., McEnery T., (2017), *Language learning research at the intersection of experimental, computational and corpus-based approaches*, Language Learning.

18. Wu G., Harrigan M., Cunningham P., (2011), *Characterizing wikipedia pages using edit network motif profiles*, Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM,. pp. 45-52.

19. Lex, E., Voelske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Granitzer, M., (2012), *Measuring the quality of web content using factual information*, Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality, ACM, pp. 7-10.