NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Irina M. Panteleeva, Olga N. Lyashevskaya,
Olga I. Vinogradova

INSPECTOR: THE TOOL FOR AUTOMATED ASSESSMENT OF LEARNER TEXT COMPLEXITY

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 79/LNG/2019

Irina M. Panteleeva, Olga N. Lyashevskaya,
Olga I. Vinogradova

INSPECTOR: THE TOOL FOR AUTOMATED ASSESSMENT OF LEARNER TEXT COMPLEXITY

EFL methodology has always recognized the importance of giving student learners of foreign languages regular and quick feedback on student speech production, both written and oral, but over the past two decades there appeared various tools ensuring the provision of automated instant feedback. The presented paper offers such a tool that focuses on measuring text complexity, which will hopefully translate into reasonable feedback about the level of language proficiency when taking into account those text features that are significant for Russian learners of English. The application provides students with advice on how to improve the weaker aspects of the evaluated essay and underlines the relevant linguistic features of the text - for example, the number of adjectival clauses. We point out what text features are more relevant for the assessment of the essays written in English by Russian students. We analyzed 3440 texts from Russian Error-Annotated English Learner Corpus, and for each of them we calculated the text criteria values. Then we used the methods of machine learning and statistical analysis to predict the grade that could be received for the essay.

JEL Classification: E32.

Keywords: text complexity, syntactic complexity, lexical complexity, morphological complexity, discourse-oriented complexity, L1 interference, learner corpora

1. Introduction

Measuring text complexity is considered an integral part of learning foreign languages. The term text complexity is defined as an indicator of how diverse and sophisticated the text is. It is based on linguistic features that can be divided into the following groups: morphological, lexical, syntactic, and discursive. The researchers in this area study which criteria reflect the level of language proficiency better (Lu 2010, Lu 2017, McNamara et al. 2014, Kyle and Crossley 2018, Bulté and Housen 2018, among others).

It is a widely accepted fact that features for measuring the complexity of the text written in the native language differ from those of the text written in the non-native language. Despite a considerable number of tools analyzing the level of foreign language proficiency, it has been proved that they are not particularly reliable for Russian learners of English (Vinogradova et al. 2019, forthcoming). Besides, these applications do not always include feedback understandable to non-professionals - as, for example, the version of feedback for a student essay produced by the system called LexInspector, which was designed for the learner corpus we are working with, but which soon proved to be of almost no use for student authors (see Figure 1).

Figure 1. Automated feedback for the essay in REALEC presented at https://linghub.ru/inspector

Moreover, some of the tools of automated feedback return the values of features without any explanation or suggestions for improving a text - cf. Coh-Metrix online tool (McNamara et al. 2014), L2 Syntactic Complexity Analyzer (Lu 2010), The Biber Tagger (Biber et al. 1999).

This paper presents a user-friendly version of the online application that will outline the direction in which the author of the text can improve some features of the text complexity that have been found wanting. In the first section, we discuss the objectives and goals of this study, and also explain what approaches we used. In Section 2, the dataset for experiments is presented. We show what preprocessing tools are used in this research in Section 3. Then, we discuss what text features are analyzed (Section 4) and present the tool for automated assessment of the essays in English (Section 5).

1. Research objectives

Measuring text complexity is considered an integral part of the evaluation of learners foreign languages' production. The term *text complexity* is defined as an indicator of how diverse and complex the text is. It is based on linguistic features that could be divided into the following groups: morphological, lexical, syntactic, and discursive. The researchers in this area study which criteria reflect more the level of language proficiency.

The study was carried out over the corpus of essays written by the Russian university learners of English in their English examination. The research questions are the following: What criteria reflect the level of language proficiency best? What features are more important when evaluating an essay? Which features are more correlated? How does the genre of the text influence the essay evaluation? Is there a difference between the different parts of the essay? What methods work better in the procedure of estimating essay evaluation automatically?

To answer those questions, we undertook the following research paths:

defined text complexity features - lexical, morphological, syntactical and discursive - that can affect the assessment;

developed a method for automatic evaluation of essays: to identify the accuracy of automatic assessment based on selected features, machine learning models were trained;

created an online application based on the results of this study: after uploading an essay to the site, the user receives a feedback that is clear to everyone who would like to know how well they wrote their English text. It is useful not only to learners, but also to teachers of English as a second/foreign language.

Besides the main goal of offering learners of English automated feedback to their text, our models produced enough data for analyzing L1 interference in learners' texts – a

phenomenon the importance of which has interested researchers in linguistics for a long time now. In this paper we report the results of experiments based on application of various methods, among which were the principal component analysis (PCA) and three approaches to the random forest model. To analyze feature dimensions, we used factor analysis (Biber 1988), which offers to look at the correlation of features and grades. Different machine learning models were tested to choose the best predictor of language proficiency level – namely, random forest, logistic regression, and k-neighbours models.

2. Corpus Data

For the purposes of our analysis – namely, studying the features that most influence the assessment of the texts written by Russian learners of English, – we needed the texts of the essays as well as the grades assigned by experts. All experiments in this research were conducted on the dataset selected from the publicly available corpus REALEC (Vinogradova et al. 2016; Vinogradova et al. 2017). It consists of approximately 14.5 thousand texts of essays (approximately 1.5 million words) written in English by Russian students in preparation for the exam or while completing examination tasks (5.9 thousand essays). This corpus also includes error annotations provided by students training to become EFL professionals, and some optional annotations of the possible L1 interference cases.

The writing tasks are similar to those in IELTS examination. The first task (Task 1) requires an examinee to describe graphical materials (pictures, graphs, etc.) given in the task, and in the second one, examinees are expected to express their opinion on a certain problem. The required size of Task 1 is at least 150 words; the second one, at least 250.

For the sake of having a more homogeneous dataset, we have chosen only essays written in the examination. The essays were evaluated by independent experts with eleven grades (ranging from 0 to 10). Basing on these grades, we additionally divided the essays into two groups: best and non-best. In the first group, we included those essays that are scored higher than or equal to 7, and in the second – the rest of the examination essays.

To check the consistency of the grades, we asked an EFL expert to re-evaluate 50 essays and then calculated the agreement score between two sets of grades using Cohen's kappa coefficient. The results showed very weak agreement between grades ($0.21 < 0.33 < 0.40$, see (Landis and Koch 1977)). Therefore, the results obtained in predicting a grade should be treated with caution. However, using division into groups of best and non-best, we observed almost a perfect agreement ($0.81 < 0.93 < 1.00$).

The size of the created dataset is 3440 essays: 1699 texts – graph description, 1741 – opinion essays; 385 – best, 3055 non-best. The grades range from 2 to 8. The descriptive statistics of the corpus data is presented in Table 1.

Table 1. Descriptive statistics of the corpus data

	Total number	**Mean in one essay**	**Standard Deviation**
Tokens	883 101	257	78
Sentences	42 987	13	5
Lemmas	392 172	114	33

4. Text Complexity Features

Text complexity is multidimensional feature that consists of absolute or relative components included in language systems (phonological, morphological, syntactic, lexical), see Bultéand & Housen (2012). In our survey, we take into account the following areas of linguistics: lexics, morphology, syntax, and discourse. We also studied features that reflect frequent L1 interference errors.

4. 1. Lexical Complexity

In this section, we consider lexical complexity as a multidimensional phenomenon that consists of three constituents: lexical density, lexical sophistication, and lexical diversity.

4. 1. 1. Diversity

Lexical diversity shows how wide the range of a learner's vocabulary is. The hypothesis is, the more proficient the learner is, the more he or she uses various vocabulary means. There are a lot of methods to measure lexical diversity. All formulas can be found in Table 2.

Table 2. Lexical diversity measures

Name	**Abbreviation**	**Formula**	**Source**
Number of different words	*ndw*	number of lemmas	
Type-token ratio	*ttr*	$\frac{numberofwordtypes}{numberofwords}$	(Templin 1957)
	corrected_ttr	$\frac{numberofwordtypes}{\sqrt{2 * numberofwords}}$	(Carrol 1964)
	root_ttr	$\frac{numberofwordtypes}{\sqrt{numberofwords}}$	(Guiraud, 1960)
	log_ttr	$\frac{Lognumberofwordtypes}{Lognumberofwords}$	(Herdan 1960)

	uber_ttr	$\frac{Log^2 numberofwordtypes}{Log \frac{numberofwords}{numberofwordtypes}}$	(Dugast 1979)
	d	$\frac{D}{N}\left[(1+2*\frac{N}{D})^{\frac{1}{2}}-1\right]$	(Malvern et al. 2004; McKee et al. 2000)
Lexical word variation	*lv*	$\frac{numberoflexicalwordtypes}{numberoflexicalwords}$	(Casanave 1994)
Verb variation	*vvi*	$\frac{numberofverbtypes}{numberofverbs}$	
	squared_vv	$\frac{numberofverbtypes^2}{numberofverbs}$	
	corrected_vv	$\frac{numberofverbtypes}{\sqrt{2*numberofverbs}}$	
	vvii	$\frac{numberofverbtypes}{numberoflexicalwords}$	
Noun variation	*ns*	$\frac{numberofnountypes}{numberoflexicalwords}$	
Adjective variation	*adjv*	$\frac{numberofadjectivetypes}{numberoflexicalwords}$	
Adverb variation	*advv*	$\frac{numberofadverbtypes}{numberoflexicalwords}$	
Modifier variation	*modv*	$adjv + advv$	

4. 1. 2. Density

Lexical density is the ratio of the number of lexical (open-class) words to the total number of words in a text (Ure 1971). The hypothesis is that low-level essays show a lower indicator of lexical density than high-level essays do. There is no established definition of lexical words among researchers of lexical complexity; for example, lexical adverbs could be regarded as adverbs of time, manner, and place (O'Loughlin 1995) or as adverbs with an adjectival base (Engber 1995). In our research, we follow the definition of lexical class proposed by Lu (2012), namely, nouns, adjectives, verbs (except modals and auxiliaries, be, have, and do) and adverbs with an adjectival base. To calculate this feature, we divide the number of tokens with the POS-tag NOUN, VERB, ADV, ADJ and PROPN (the UDPipe parser has a separate tag for proper names) by the total amount of tokens.

4. 1. 3. Sophistication

Lexical sophistication, or lexical rareness, is the ratio of the number of advanced words to the total number of words (Read 2000). The hypothesis is that the lower the learner's

proficiency, the fewer rare vocabulary items there are in his/her essay. There are several ways to calculate this criterion, see Table 3.

Table 3. Lexical sophistication measures

Name	Abbreviation	Formula	Source
Lexical sophistication	*ls*	$\frac{numberofsophisticatedle}{numberoflexicalw}$	(Linnarud 1986), (Hyltenstam 1988)
Lexical frequency profile	*lfp_1000, lfp_2000, lfp_uwl, lfp_rest*	proportion (see below)	(Laufer 1994)
Verb sophistication	*vs*	$\frac{numberofsophisticatedve}{numberofverbs}$	(Harley and King 1989)
Corrected verb sophistication	*corrected_vs*	$\frac{numberofsophisticatedve}{\sqrt{2 * numberofverbs}}$	(Wolfe-Quintero et al. 1998)
Squared verb sophistication	*squared_vs*	$\frac{numberofsophisticatedve}{numberofverbs}$	(Chaudron and Parker's 1990)

M. Linnarud defined sophisticated lexical words as words that are found only at a high level of mastering English (1986). K. Hyltenstam specified that these are words that are not on the list of 7000 frequent words (Hyltenstam 1988). Lexical frequency profile is the proportion of tokens that are included in the list of the first 1000 most frequent words, the list of the second 1000 words, University Word List (Xue and Nation 1989), and the list of those that are not on the lists above (Laufer 1994). B. Harley and M. L. King defined a verb sophistication measure as the ratio of the number of sophisticated verbs to the total number of verbs (Harley and King 1989).

The formula proposed by B. Harley and M. L. King was criticized, as the ratio decreases if the sample size increases (Arnaud 1992; Richards 1987). The changes in this formula were proposed in (Wolfe-Quintero et al. 1998) and (Chaudron and Parker 1990) to reduce sensitivity to the text size.

In our research, we check all three methods to define lexical sophistication. We consider tokens sophisticated if they are not present on the list of most frequent verbs extracted from the list of frequent words generated from the British National Corpus (Leech et al. 2001). There are no big differences between best and non-best essays by the features that reflect the sophistication of the essay: the average, maximum and minimum values are practically the same. With this result, the hypothesis proposed above could not be confirmed.

4. 2. Syntactic Complexity

The syntactic complexity of the L2 texts reflects how productive various grammatical structures are (Foster and Skehan 1996). The more accurate definition of syntactic complexity is given in (Lu and Ai 2015). They determine it as a part of linguistic complexity that studies the number and variety of syntactic structures and the degree of sophistication of those structures.

There are a lot of research articles in which the authors discuss which of the criteria are more reliable syntactic features in predicting the level of language proficiency (Lu 2010, Lu 2017, McNamara et al. 2014, Panteleeva 2018, Kyle and Crossley 2018, Bulté and Housen 2018, among others).

In (Panteleeva 2018), some of the criteria replicating the features that are used in L2 Syntactic Complexity Analyzer (Lu 2010) were applied, as well as The Biber Tagger (Biber et al. 1999), and Coh-Metrix (McNamara et al. 2014). There were also proposed new features that could be used in the analysis of the syntactic complexity, namely: maximum, minimum, and average depth of the sentence, the number of adverbial, relative, and adjectival clauses, the number of constructions like noun + infinitive. One of the research aims in (Panteleeva 2018) was to outline which values – relative or absolute – better differentiate the worst essays from the best ones. It was concluded that absolute values help to predict the level of language proficiency better.

The cluster of the criteria with absolute values includes the number of tokens (num_tokens); the maximum (max_depth) and minimum (min_depth) depth of sentence, the number of relative (num_acl_relcl), adverbial (num_advcl), and adjectival (num_acl) clauses; the number of sentences (num_sents), clauses (num_cl), T-units (num_tu), complex T-units (num_ctu), coordinated phrases (num_coord), nominal phrases (possessive constructions (poss); constructions where two nouns are related by a preposition (prep_ph); infinitive or gerund as an object / subject (ger_inf); constructions like adjective + noun(adj_n), participle + noun(part_n), noun + infinitive(num_n_inf)); the total number of complex nominal phrases (num_np), the number of verb phrases (num_vp).

The cluster of the criteria with relative values includes the following measures: the average depth of the sentence (av_depth), the variety of structures (mean_l_sim, mean_p_sim is the Levenshtein distance between lemmatized and POS-tagged sentences respectively (each with each), mean_l_sim_nei, mean_p_sim_nei (only neighboring)), the average number of tokens before the main word in the sentence (mean_tokens_root), an average sentence length (mean_length_s), an average clause length (mean_length_c); the number of clauses per sentence (c_s), the number of clauses per T-unit (c_t), the number of dependent clauses per T-unit (acl_t, acl_relcl_t, advcl_t), the number of dependent clauses per clause (acl_cl, acl_relcl_cl, advcl_cl),

the number of clauses (coord_cl), the number of T-units per sentence (t_s), the number of possessive constructions (poss_s), constructions with prepositions (prep_s), constructions like adjective + noun(adj_n_s), participle + noun(part_n_s), noun + infinitive(n_inf_s), infinitives or gerunds as an object / subject of a sentence (ger_inf_ s) per sentence; the number of verb phrases per sentence (vp_s).

In this research, we take into account features proposed in all discussed articles and follow the conclusion from (Panteleeva 2018) that absolute values are better than relative ones while assessing essays.

The criteria that we examine are proposed in Table 4:

Table 4. Syntactic measures

Name	Abbreviation	Formula
Average tree depth	*av_depth*	$\frac{sum(depthofsentences)}{num(sentences)}$
Maximum tree depth	*max_depth*	$max(depthsofsentences)$
Minimum tree depth	*min_depth*	$min(depthsofsentences)$
Adjective clause modifier	*num_acl*	$num(acl)$
Adverbial clause modifier	*num_advcl*	$num(advcl)$
Relative clause modifier	*num_rel_cl*	$num(rel\_cl)$
Number of sentences	*num_sent*	$num(sentences)$
Number of tokens	*num_tok*	$num(tokens)$
Average number of tokens before root	*av_tok_before_root*	$\frac{num(tokensbeforeroot)}{num(sentences)}$
Average length of the sentence	*av_len_sent*	$\frac{num(tokens)}{num(sentences)}$
Number of clauses	*num_cl*	$num(finiteforms)$
Number of T-units	*num_tu*	$num\_cl - num\_advcl - num([iforwhen])$
Number of complex T-units	*num_compl_tu*	$num\_cl - num\_tu$
Number of coordinated phrases	*num_coord*	
Number of possessives	*num_pos*	
Number of prepositional phrases	*num_prep*	
Number of adj+noun	*num_adj_noun*	
Number of participle+noun	*num_part_noun*	
Number of noun+infinitive	*num_noun_inf*	

Levenshtein distance between pos-tagged sentences (between neighbour sentences)	*pos_sim_nei*	see below
Levenshtein distance between lemmatized sentences (between neighbour sentences)	*lemma_sim_nei*	see below
Levenshtein distance between pos-tagged sentences (between all sentences)	*pos_sim_all*	see below
Levenshtein distance between lemmatized sentences (between all sentences)	*lemma_sim_all*	see below

Following (McNamara et al. 2014) and (Panteleeva 2018), we defined the syntactic diversity (*pos_sim_nei*, *lemma_sim_nei*, *pos_sim_all*, *lemma_sim_all*) of the sentences in the essay in this way: at first, each sentence was lemmatized and labeled with the parts of speech. Each sentence was transformed into the chain of lemmas or parts of speech tags. Then the Levenshtein distance was calculated between all the transformed sentences in the text and between the adjacent ones. The first reflects how diverse the essay is in general, and the second reflects how different the closest constructions are.

One of the new criteria proposed in (Panteleeva 2018) is the depth of the sentence. It was presented as the distance from the root of the tree to its farthest descendant. For example, the depth of sentence (1) is 5 (the path from the root *America* to the node *than*):

(1) The most attractive for tourists in 1995 was North America, which was visited by more than 70 million people. (REALEC)

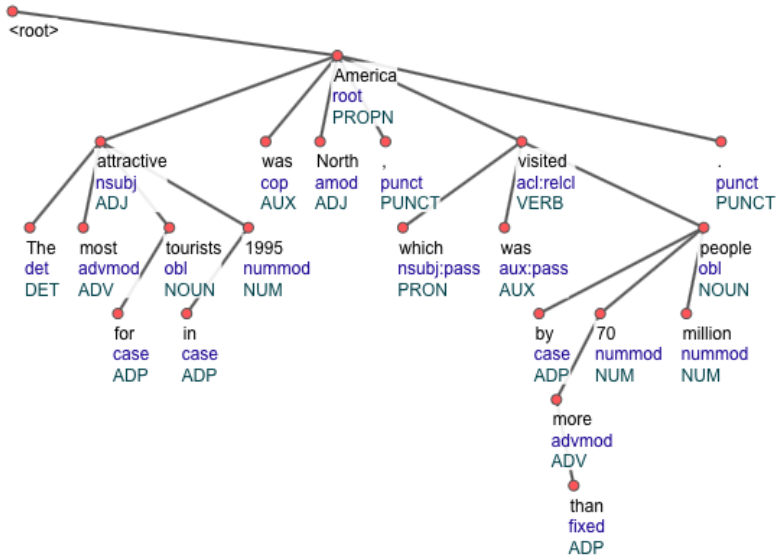


Figure 2. Syntactic tree of sentence (1)

Among all syntactic features only the number of T-units turned out to be the feature showing that there is no difference between best and non-best essays. Analyzing the average, minimum, and maximum values of other criteria, we have been able to draw the following conclusions:

(A) the more dependent clauses, sentences, words before the root of the sentence, coordinational phrases, and complex noun phrases there are, the higher grade the essay is assigned;

(B) the longer the sentences in the essay are, the higher grade the essay is assigned.

4. 3. Morphological Complexity

In this section we discuss features which can be measured at the word level. The criteria of morphological complexity are divided into two groups: derivational and inflexional (Bulté and Housen 2012). The first type refers to the measure of affixation, the second one – to the frequency of tense forms, frequency of modals, the number of different verb forms, variety of past tense forms, and MCI (Morphological Complexity Index) (Brezina and Pallotti 2019).

4. 3. 1. Derivational Features

L. Bauer and P. Nation (Bauer and Nation 1993) proposed a scale for categorizing English suffixes. Their research was based on productivity, frequency, and predictability of the meaning, regularity of the spelling form of affix, regularity of its function, and regularity of the written form of the base (Leontjev 2016). There are 7 levels in this scale:

Table 5. Levels of English affixes

Level	Examples
3	*-able*, *-er*, *-ish*, *-less*, *-ly*, ...
4	*-al*, *-ation*, *-ess*, *-ful*, *-ism*, *-ist*, *-ity*, ...
5	*-age* (*percentage*), *-al* (*approval*), *-ally* (*idiotically*), *-an* (*Russian*), *-ance*, ...
6	*-ee*, *-ic*, *-ify*, ...

In our research we counted the total number of affixes of each level (see Table 5) in every text and also the relative value of this criterion: the number of affixes on n's level to the total number of affixes.

4. 3. 2. Inflectional Features

MCI represents an average inflectional diversity for the parts of speech in the sample. The hypothesis is that essays that have more different forms (show, showing, shows– 3 forms) will be considered to be more complex than those that have fewer forms (show, show, show– 1 form).

Counting MCI, we are limited only to verb forms. The process of finding this measure is described below (Brezina and Pallotti 2019):

1. Extract all inflexional suffixes of verbs in the sample (for irregular verbs, we still label their forms of past tense as -ed): *-ed, -ed, -Ø, -ed, -ed, -ed, -ed, -ed, -ed, -Ø, -ed, -ed, -Ø, -ed, -ed, -Ø, -s, -ing, -ing, -Ø, -ed, -ed, -Ø, -ed, -ed, -Ø*
2. Form two random groups of 10 inflexional suffixes:
    a. *-ed, -ed, -Ø, -ed, -ed, -ed, -ed, -ed, -ed, -Ø* diversity=2
    b. *-ed, -ed, -Ø, -ed, -ed, -Ø, -s, -ing, -ing, -Ø* diversity=4

    mean diversity = 2+4= 3
3. Count the index of unique value: IUV = number of unique affixes / 2

    There is the affix *-s* in the second group, the first one does not have it: IUV=0.5
4. Count MCI: MCI = mean diversity + IUV / 2 - 1

    MCI = 3 + 0.25-1=2.25

In contrast to MCI, other inflectional criteria (frequency of tense forms, frequency of modals, the number of different verb forms, variety of past tense forms) are measured by the category itself, and not by affixes:

- Frequency of tensed(finite) forms
- Frequency of auxiliaries
- The number of infinitives
- Frequency of gerunds
- Frequency of verbs in present simple (sing)
- Frequency of verbs in present simple (plur)
- Frequency of past participle
- Frequency of verbs in past simple
- Average inflectional diversity
- The number of derivational suffixes (level 3)
- The number of derivational suffixes (level 4)
- The number of derivational suffixes (level 5)
- The number of derivational suffixes (level 6)

4. 4. Discursive Complexity

4. 4. 1. Discourse-Organizing Nouns

Discourse-Organizing Nouns (DONs, also known as shell nouns) are semantically unspecific abstract nouns, such as fact, thing, issue, problem and argument, which play an organisational role in the text (Tåqvist 2016). These nouns refer to information given in the preceding (2) or following (3) parts of the discourse:

(2) *I like my subject and that is probably the most important thing when choosing a topic; it has to be a subject which interests you.* (ICLE-SW: SWUL8027)

(3) *Admittedly, there is a risk that our modern society becomes too concentrated on the fact that we should be efficient and deal with our duties at a rapid pace.* (ICLE-SW: SWUL3013)

To identify DONs, we trained the logistic regression model using the dataset from (Roussel 2018). In this work, the authors proposed the reliable annotation of the Europarl Corpus (Koehn, 2005). We selected 993 sentences in English in which each candidate word was annotated either as a shell noun or not. Our training data includes 50 different discourse-organizing nouns. We analyze only those words that can be shell nouns. After that our task has narrowed down to binary classification. We vectorized the left and the right context using TF-IDF approach to transform text data for training the chosen model. The logistic regression model showed the accuracy on 10-fold cross-validation –0.78 (precision – 0.68, recall – 0.82).

4. 4. 2. Functional N-grams

Following the approach in (Chen and Baker 2014), only four-token lexical bundles were investigated. These combinations were extracted from the whole corpus REALEC. Besides, only collocations that occurred more than 280 times in the corpus were chosen. Then context-dependent and task-dependent bundles, usually containing proper names, were manually excluded from the list, as these phrases are unlikely to be discourse features. The bundles that are in top-10 are the following (Table 6):

Table 6. top-10 4-grams in REALEC

№	Collocation	Number	Frequency
1	*to sum up ,*	2306	0.003
2	*on the other hand*	2278	0.003
3	*the other hand ,*	2076	0.002
4	*first of all,*	1872	0.002
5	*i would like to*	1312	0.002
6	*are a lot of*	1312	0.002
7	*there are a lot*	1250	0.001
8	*in my opinion ,*	1113	0.001
9	*on the one hand*	944	0.001
10	*at the same time*	942	0.001

We also added collocations from the list presented in (Chen and Baker 2014), namely, 106 bundles. The number of intersections between the list of n-grams from REALEC and Chen and Baker's list is 17 collocations.

For linguistic interpretation, all word combinations should be divided into groups that would represent different functions of discourse. We took the classification suggested by Y.-H. Chen and P. Baker as the basis and added our examples. The classification is based on functional characteristics. Functions and examples are presented in Table 7:

Table 7. Classification of lexical bundles based on functional characteristics

Function	Subfunction	Example
Referential	Quantifying	*most of the people,* *the rest of the world, ...*
	Time/place/text deixis	*at the beginning of the,* *in the following paragraphs, ...*
	Framing	*the reason is that,* *as a result of, ...*
Stance	Epistemic	*it is believed that,* *some people think that, ...*
	Attitudinal/modality	*I hope I can,* *it is not easy, ...*
Discourse organizers	Topic elaboration/clarification	*is more important than,* *it is because the, ...*
	Identification/focus	*my point of view,* *is the most important, ...*
	Topic introduction	*I would like to,* *I am going to, ...*

After extracting these collocations, we counted their total number in each essay and got two features: the number of 4-grams (collection of two lists) and the number of functional 4-grams (Chen and Baker 2014). ANOVA showed that neither of these metrics is informative for essay evaluation (p-value > 0.05). Practically, a half of the essays do not include functional n-grams. 4-grams from REALEC, on the contrary, are present in most of the texts from the data, compare Figure 3:

Figure 3. N-gram distribution: left – the number of functional n-grams, right – the number of frequent REALEC n-grams 4. 4. 3. Linking Tools

From each essay we extracted linking words that were presented in (Swales and Feak 2009), where all collocations were divided into the following groups:

- Addition (*Moreover, In addition*)
- Adversativity (*However, despite*)
- Cause and effect (*Because, consequently*)
- Clarification (*i.e., that is*)
- Contrast (*whereas, Conversely*)
- Illustration (*For example, For instance*)
- Intensification (*On the contrary, In fact*)

Analyzing the corpus dataset, we found out that the most frequent linking expressions belong to the clarification group: the average number per essay reaches 10 collocations in best essays, whereas in non-best ones – 7, see Figure 4. Students also often use phrases that help to emphasize contrast or cause: on average there are 2 collocations of these groups per essay. In Figure 4 we see a similar situation as in the case of n-grams. Some students whose essays were rated low used as many templates as possible: in 86 out of 3442 essays students use more than 20 linking units of clarification (the maximum number is 32).

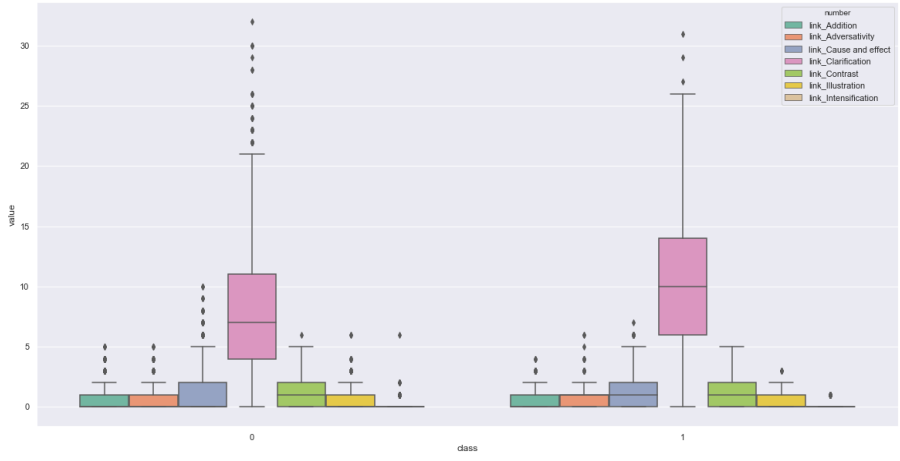


Figure 4. Group distribution of linking tools: 0 – non-best, 1 – best

4. 5. Alert of L1 Interference Mistakes

We chose 100 random sentences which included tags "L1 interference mistakes" and grouped them by the type of mistakes: Word choice (*The percentage of people with qualifications with undergraduate diploma and Bachelor's degree and Master's degree are quite* ***middle*** (*-> medium*)), Absence of a component or a redundant component (*The argument in favor of the opinion that for children* ***is better*** *(-> **it is better**) to*), Spelling (***english*** *(-> **English**)*), Relative clauses (*This number is still less than tourists**, who** (-> who) visited France.*), Verb (*Each of the factors have greatly* ***influenced to*** *(-> **influenced**) the loss of experts in this area.*), Standard Word Order (***For undergraduate diploma the situation changes.*** *(-> **The situation changes for undergraduate diploma.**)*), Discourse (*be useful for* ***him*** *(**him or her**)*). Even though we identified the most common L1 interference mistakes, we did not add them to the feature dimension due to the difficulty of their automatic extraction. At this stage of the research we decided to extract simple L1-induced mistakes, namely, those that can be found using rule-based approach. From the special dataset of sentences with errors that student annotators marked as "L1 interference mistakes" we chose the following features to predict language proficiency: the number of tokens misspelled in such a way that they resemble Russian orthographic norm; the number of punctuation mistakes in participial phrases, the latter being always separated with commas in Russian, while their English equivalents have commas only in the position before the noun head; the number of punctuation mistakes connected with the conjunction than, as, again, its Russian equivalent is always used with a comma in front of it; the number of wrong plural forms of nouns after the number within a numeral (like *5 millions passenger*).

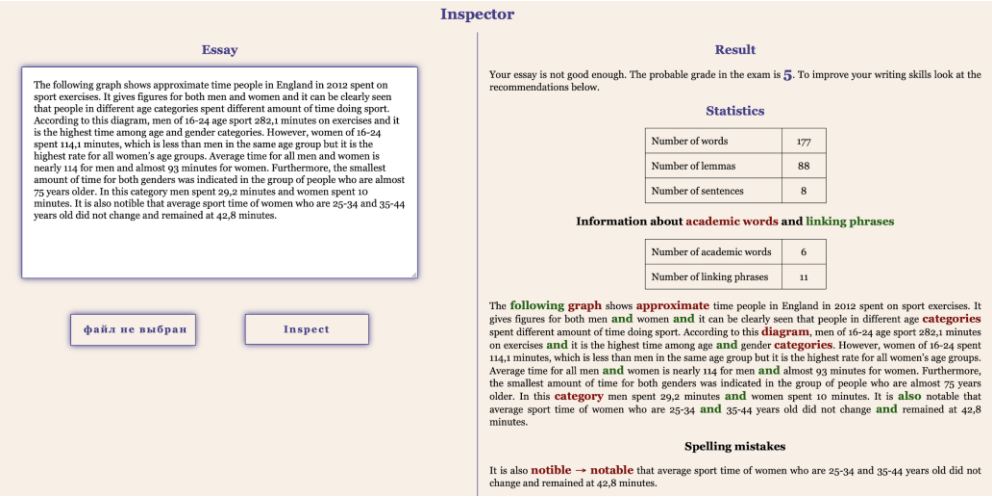
A striking peculiarity found in REALEC is that in forming a numeral, Russian students use a plural noun (*thousands*, *millions*, etc.) up to 14 times as often as a singular form of the noun in the numeral in their essays - evidently, under the influence of the plural form of the equivalent noun in Russian numerals.

While counting the number of punctuation mistakes in participial phrases, we regarded the sentence as correct if the participle group before the noun head it attributes was separated by commas, or if the participle phrase following the noun head was not separated by commas. Firstly, to identify a participle phrase, we found the participle with the UDPipe POS tag Ger – active participle or Part – passive participle and relation-tag advcl – adverbial clause modifier. Secondly, we defined the location of the participle phrase with respect to the noun head using UDPipe head-tags. Finally, we checked for commas around the participle phrase and its head.

5. Application: CompInspector

We created an online application based on the results obtained in this research. We hope that users will find it friendly enough and that everyone - professional linguists, English teachers and professors, and learners of English – will make use of it.

The application called CompInspector provides the user with the automatic assessment on the basis of the comparison of the author's and the mean (across the best essays in the corpus) values for features that proved to be important. A user can upload his/her essay from the file or type it in the text window. After pushing the button "Inspect," he/she gets the feedback on the essay. Two models have been chosen to be applied in the tool. The first one is a multiclass random forest classifier model trained only on the important features with the help of the drop-column approach and using only the least collinear metrics. The second one is a binary random forest classifier model that predicts if the essay is similar to the best essays or non-best ones. The data for the experiment can be found at https://github.com/panteleeva48/REALEC_Inspector/. The sample feedback is presented in Figure 5. The first sentence is a result of the binary classification (the general information on the essay states what group the essay is closer to - best or non-best), the second part shows the multiclass classification (the general sentence is detailised, namely, the grade predicted by the model is suggested to the user).



Recommendations

The model that predicted your grade is based on the features presented in the table below. Having compared the mean values of these features, we highlight the features that you should improve (**green** - good indicators; **red** - features to be improved). All in all, in your essay you should use more different verbs; linking phrases; gerunds; morphologically complex words; infinitives; different words; functional n-grams; complex sentences; relative clauses; words before the main predicate; possessive constructions. You used enough long sentences; academic words; different nouns; verbs in past simple; auxiliary verbs; sophisticated verbs; nouns, verbs, adverbs, and adjectives; coordinate constructions.

Your essay	Feature	Best
9	Number of coordinate phrases	6
0.47	Lexical density	0
0.59	Verb sophistication	0.31
0.85	Number of auxilaries	0.61
6	Number of verbs in past simple	4
0.1	Noun variation	0
0.22	Lexical sophistication	0.12
24	Average length of the sentence	23
2.16	Verb variation	2.82
13	Number of linking phrases	14
1	Number of gerunds	3
0.03	Derivational level 3	0.05
0.0	Derivational level 4	0.07
2	Number of infinitives	11
88	Number of lemmas	131
0.0	Derivational level 6	0.04
0	Number of functional n-grams	1
17	Number of clauses	25
1	Number of adjective clause	5
5.12	Average number before root	6.4
8	Number of sentences	13
0.05	Derivational level 5	0.06
15	Number of possessives	16

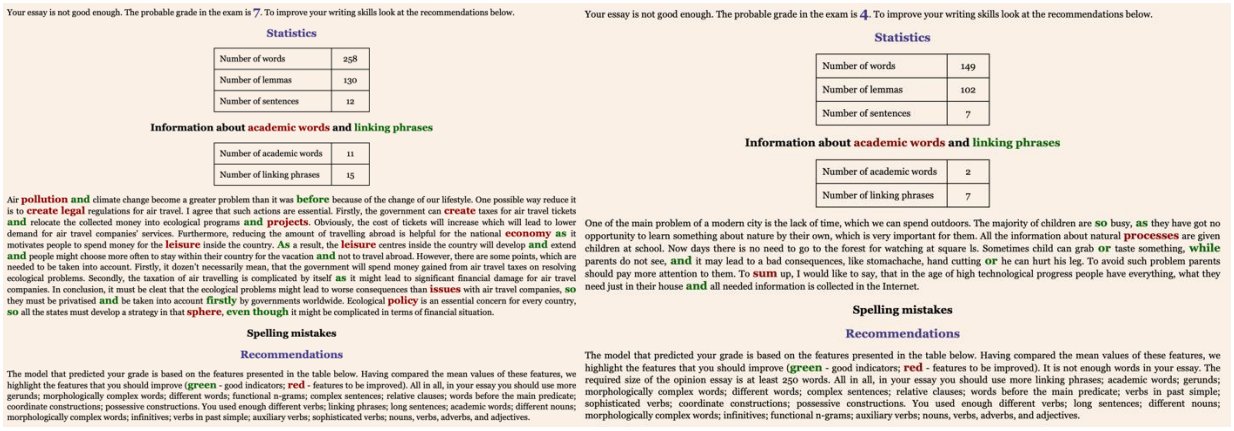
Figure 5. CompInspector

After the results of the classification are shown, CompInspector presents general statistics on the essay, namely: the number of tokens, lemmas, sentences, academic words, and linking phrases in the text. Academic words and organizing phrases in the text are highlighted with different colour schemes.

The result of spell checking is presented in the next section. A correction (or corrections) for a misspelled word is (are) suggested. The tool chosen for this purpose - pyenchant - does not always suggest the best corrections, but we hope that we will improve spell checking in future work by training our model to identify spelling errors and to choose the appropriate corrections of the misspelled words.

The last section in the automated feedback is recommendations. The values of the analyzed text are compared with the mean values of the best and non-best essays. If the indicator of the feature is higher than or equal to the mean value of the best essays, this feature is considered to be the predictor of the better grade for the essay. The result of such comparison is to be given both as a worded-out feedback and as the table with all the data (good values in green, and worse, in red).

In the Figure 6 two essay are presented – the first one is a well-written essay (mark is 7) and the second one refers to worst (mark is 4). In the first (best) essay you can see more academic and linking phrases and the indicators of the important features are higher than the indicators of the second (non-best) essay:



Your essay	Feature	Best
0.49	Lexical density	0
0.12	Derivational level 5	0.06
0.38	Verb sophistication	0.31
0.74	Number of auxilaries	0.61
5	Number of verbs in past simple	4
0.06	Derivational level 6	0.04
20	Number of infinitives	11
0.1	Derivational level 4	0.07
0.17	Noun variation	0
0.15	Lexical sophistication	0.12
23.58	Average length of the sentence	23
14	Number of linking phrases	14
2.92	Verb variation	2.82
1	Number of gerunds	3
0.03	Derivational level 3	0.05
130	Number of lemmas	131
0	Number of functional n-grams	1
24	Number of clauses	25
4	Number of adjectival clauses	5
5.17	Average number of words before root	6.4
12	Number of sentences	13
5	Number of coordinated phrases	6
15	Number of possessives	16

Your essay	Feature	Best
0.43	Lexical density	0
0.07	Derivational level 5	0.06
0.78	Number of auxilaries	0.61
2	Number of functional n-grams	1
11	Number of infinitives	11
0.17	Derivational level 4	0.07
0.18	Noun variation	0
23.71	Average length of the sentence	23
2.83	Verb variation	2.82
7	Number of linking phrases	14
0.08	Lexical sophistication	0.12
1	Number of gerunds	3
0.0	Derivational level 3	0.05
102	Number of lemmas	131
0.03	Derivational level 6	0.04
17	Number of clauses	25
3	Number of adjectival clauses	5
6.14	Average number of words before root	6.4
7	Number of sentences	13
3	Number of verbs in past simple	4
0.17	Verb sophistication	0.31
2	Number of coordinated phrases	6
11	Number of possessives	16

Figure 6. The best and non-best essays

We consider the feedback to be useful for the English instructors, who could easily see what could be improved in each student's writing in order to obtain better results in the exam. Instructors can also get the summary spreadsheet for the group with highlighted values that need working on in the process of preparing for the exam. A student trying to improve their English writing proficiency by him- or herself can with such feedback check which parts of the uploaded text could be improved and can thus correct them, practicing in this way self-editing skills.

6. Conclusion

In the course of our research 59 features were identified as demonstrating written text complexity, and they were divided into 5 groups: lexical, morphological, syntactic, discursive, and raising alert of possible L1 interference.

We presented an online application that is based on the results of our research. This tool is aimed at helping English instructors and university students preparing for the examination in English.

In future, we hope to fix the weak agreement between grades and the imbalance in the data. To improve the sample, we could 1) increase the sample itself so that the grades for essays have a wider range than it is now and 2) apply data balancing methods.

Model prediction can also be improved: better results can be obtained by 1) expanding the number of features, primarily - those related to L1 interference errors; 2) changing procedures at the stage of checking the spelling; 3) using another models to predict grade 0; and applying some new techniques. Hopefully, measurements of student text complexity will become even more useful to the three target groups - EFL instructors, students, and linguists.

References:

Arnaud, P. J. L. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In Arnaud, P. J. L. & Béjoint, H. (eds.). Vocabulary and applied linguistics. London: MacMillan, 133–145.

Bauer, L., & Nation, I. S. P. (1993). Word families. International Journal of Lexicography. 6 (4), 253–279.

Biber, D. (1988). Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. London: Longman.

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. Second Language Research, 35 (1), 99–119.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. Amsterdam: John Benjamins Publishing Company, 21–46.

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and group trends. International Journal of Applied Linguistics, 28, 147–164.

Casanave, C. (1994). Language development in students' journals. Journal of Second Language Writing, 3, 179–201.

Chaudron, C., & Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. Studies in Second Language Acquisition, 12, 43–64.

Chen, Y–H., & Baker, P. (2014). Investigating criterial discourse features across second language development: lexical bundles in rated learner essays, CEFR B1, B2 and C1. Applied Linguistics, 37 (6), 849–880.

Crossley, S. A., Allen L. K., & McNamara, D. S. (2014). A Multi-Dimensional analysis of essay writing. Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber. 197–237.

Dugast, D. (1979). Vocabulaire et stylistique. I Théâtre et dialogue. Travaux de linguistique quantitative. Geneva: Slatkine-Champion.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. Journal of Second Language Writing, 4, 139–155.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. Studies in Second Language Acquisition, 18, 299–323.

Guiraud, P. (1960). Problèmes et méthodes de la statistique linguistique. Dordrecht: Reidel, D.

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. Studies in Second Language Acquisition, 11, 415–440.

Herdan, G. (1960). Quantitative linguistics. Butterworth, London.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. Journal of Multilingual and Multicultural Development, 9, 67–84.

Koehn Ph. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

Kyle, K., & Crossley, A. S. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. The Modern Language Journal, 102, 333–349.

Landis, J. R., & Koch, G. G. (1977) The Measurement of Observer Agreement for Categorical Data. Biometric, 33, 159–174.

Laufery, B. (1994). The lexical profile of second language writing: Does it change over time? RELC Journal, 25, 21–33.

Leech, G., Rayson, P., & Wilson, A. (2001). Word frequencies in written and spoken English: Based on the British National Corpus. London: Longman.

Leontjev, D. (2016). L2 English derivational knowledge: Which affixes are learners more likely to recognise? Studies in Second Language Learning and Teaching 6 (2), 225–248.

Linnarud, M. (1986). Lexis in composition: A performance analysis of Swedish learners' written English. Lund: CWK Gleerup.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15, 474–496.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. The Modern Language Journal, 96, 190–208.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. Language Testing, 34 (4), 493–511.

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. Journal of Second Language Writing, 29, 16–27.

Lyashevskaya, O. N & Panteleeva, I. M. (2018). REALEC learner treebank: annotation principles and evaluation of automatic parsing. Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), Prague, Czech Republic, 80–87.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). Lexical diversity and language development: Quantification and assessment. Houndmills, UK: Palgrave MacMillan.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. Literary and Linguistic Computing, 15, 323–337.

McNamara, D. S. , Graesser, A. C., McCarthy, P., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge: Cambridge University Press.

Nivre, J., Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. , McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association, Portorož, Slovenia, 1659–1666.

O'Loughlin, K. (1995). Lexical density in candidate output of direct and semi-direct versions of an oral proficiency test. Language Testing, 12, 217-237.

Panteleeva, I. M. (2018). Sintaksicheskij parsing korpusa oshibok [Dependency Parsing of a Learner Corpus] (course paper).

Read, J. (2000). Assessing vocabulary. Oxford: Oxford University Press.

Richards, B. (1987). Type/token ratios: What do they really tell us? Journal of Child Language, 14, 201-209.

Roussel, A. (2018). Detecting and Resolving Shell Nouns in German. Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference, New Orleans, Louisiana, 61–67.

Straka, M. (2017). UDPipe baseline models and supplementary materials. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Straka, M., Jan Hajič, and Jana Straková. (2016). UD-Pipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association, Portorož, Slovenia.

Swales, J. M., & Feak, C. B. (2009). Academic writing for graduate students: Essential tasks and skills (2nd ed.). Ann Arbor: The University of Michigan Press.

Tåqvist, M–K. (2016) "Another thing": Discourse–organising nouns in advanced learner English. (Doctoral dissertation), Karlstad University, Karlstad.

Templin, M. (1957). Certain language skills in children: Their development and interrelationships. Minneapolis: The University of Minnesota Press.

Ure, J. (1971). Lexical density: a computational technique and some findings. In M. Coultard (Ed.), Talking about text. Birmingham: English Language Research, University of Birmingham, 27–48.

Vinogradova, O. (2016). The role and applications of expert error annotation in a corpus of English learner texts. In Computational Linguistics and Intellectual Technologies. Proceedings of Dialog 2016, 15, 740–751.

Vinogradova, O. I., Lyashevskaya, O. N., & Panteleeva, I. M. (2017). Multi-level Student Essay Feedback in a Learner Corpus. In Proceedings of the International Conference «Dialogue 2017», Moscow, 2017.

Vinogradova, O., Ershova, E., Sergienko, A., Generalova, S. (2019). AWARL (Automated Writing Assistant for Russian Learners) As a Computer-Assisted Language Learning Tool. Paper to be presented at Eurocall 2019, Louvain-la-Neuv, August, 2019.

Wolfe-Quintero, K., Inagaki, K., S., & Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity (Report No. 17). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Xue, G., & Nation, I. S. P. (1984). A university word list. Language Learning and Communication, 3, 215–229.

Olga I. Vinogradova
National Research University Higher School of Economics (Moscow, Russia).
E-mail: olgavinogr@gmail.com
Olga N. Lyashevskaya
National Research University Higher School of Economics (Moscow, Russia).
E-mail: olesar@yandex.ru
Irina M. Panteleeva
National Research University Higher School of Economics (Moscow, Russia).
E-mail: impanteleyeva@gmail.com