Olga Vinogradova, Anna Viklova, Darya Overnikova, Anton Buzanov

Genre is a useful concept only when used evaluatively not descriptively[1]: Learner corpus research of the correlation between accuracy/complexity and two task types.

In this paper, we set up the task to evaluate the effects of genre on the accuracy of learner writing. The genre in the case of our learner corpus is understood in its narrow meaning, namely, as the task type of the two in answer to which examination essays were written by learners of English with Russian L1 in their 2nd-year English examination. The task types are the following: Task 1 – description of the graphical material presented in the task, and there are 2,414 texts of this kind under investigation; and Task 2 – arguments proving one of the two positions outlined in the task, 2,345 essays of this type altogether. The learner corpus with those essays (realec.org) has two layers of annotation – automated POS tagging and expert manual annotation of errors.

In prior research of the task effects on learner writing, the following research questions similar to ours have been asked (as shown in Plonsky & Kim (2016)):

- What kind of tasks are the most efficient for learners to develop accuracy and complexity simultaneously? (Loschky and Bley-Vroman, (1993))
- What is better to measure the effects of increasing the complexity of task demands – to look at specific or at general measures of the accuracy and complexity of L2 speech production? (Robinson, Cadierno and Shirai (2009))
- How does the prompt and input of a task and its functional requirements influence task-based linguistic performance? (Alexopoulou et al. (2017))
- How do task design features affect the written language used by second language learners when they try to meet the non-linguistic goal of a task? Specifically, how does task complexity, or task type impact the complexity and accuracy of the language use in global as well as specific features or structures? (Meurers & Dickinson (2017))

Also, SLA linguists focusing on other influences on learner writing have left important observations concerning the role of written tasks. Nina Vyatkina (2012:595), for example, states that task effects constitute a "particularly severe threat to validity in longitudinal designs". Alexopolou et al (2017:181) mention that "learner corpus research has generally not been linked to research investigating effects of task on writing". Back in 1993, Lester Loschky and Robert Bley-Vroman drew the distinction between 'natural', 'useful' and 'essential' structures in learner texts. This distinction paved way to investigating individual language features that may be elicited by the task, such as the vocabulary and grammatical features given in the task prompt. This was initially undertaken by two researchers - Peter Skehan in his central hypothesis of the Limited Attentional Capacity (further on –LAC) model (Skehan, 1998) and Peter Robinson in the Cognition Hypothesis (Robinson, 1995). Skehan states that task complexity (e.g., vocabulary load and variety, clarity and structure of information to process) is bound to cause the linguistic

[1] Stemming from, but contrary to Ursula K. Le Guin's quote "Genre is a useful concept only when used not evaluatively, but descriptively."

complexity and accuracy of the elicited language. However, as shown in the review in Alexopolou et al (2017:182), it has been neither proved nor disproved by further research. Unlike Skehan, Robinson's Cognition Hypothesis assumes that learners can focus on multiple demands requiring their attention. He distinguishes between tasks concerned with the here-and-now and there-and-then, between those tasks that involve reasoning and perspective taking and those that do not, and, finally, Robinson claims that higher reasoning demands will lead to learners' higher linguistic complexity and accuracy. He proposes that in narratives types of essay including argumentative essays there is bound to be a high vocabulary load because all lexis needs to come from the learner. In contrast, the prompts requiring description and comparison of the given materials provide most of the vocabulary needed. On the other hand, narratives and argumentation are characterised with lower structure and low to medium clarity since they are free writing tasks. By contrast, the academic description tasks involve high structure as learners are expected to present certain factual features, trends, and comparisons that can be seen in the prompt. When interpreting a task with "here & now" vs. "there & then" parameters, the descriptions of graphical material are interesting tasks because they are often about there & then, but some include predictions for the future, and besides, a part of the prompt is presented to learners uses the present tense (e.g. *The graphs below* ***show***…), which could be interpreted as "now". Telling a story in the argumentative essay may well involve reasoning, while a lot of reasoning is expected in the academic description tasks since they primarily present facts and statistics and urge learners to draw comparisons wherever possible. The factor "perspective taking" is important for many prompts presenting the sides of the argument, because different points of view need to be taken into account by learners. Yet again, Robinson's predictions related to the task features have not been fully supported by other researchers. Later, the paper by Peter Robinson, Teresa Cadierno and Yasuhiro Shirai showed that there is more complex, developmentally advanced use of tense/aspect morphology on conceptually demanding tasks compared to less demanding tasks, and a trend to more accurate, target-like use of lexicalization patterns for referring to motion on complex tasks. Skehan also considers interdependency between accuracy and complexity and compares his approach with that of Robinson arguing that better accuracy and complexity depend not on task difficulty but on task special characteristics.

Data and methods

As on the whole the argument between Skehan and Robinson on what in the task is more important for eliciting better learner production has not been resolved, in this research we present the distribution of both Skehan's and Robinson's features for both tasks we have in the learner corpus. These features are presented in Table 1. The opinion that these characteristics capture some core properties of the tasks in question and are bound to produce impact on learners' language use has also been supported by the research team in Alexopoulou et al. (2017).

Table 1. Overview of task characteristics in REALEC

	Graph Description	**Argumentative Essay**
LAC model	Medium vocabulary load	High vocabulary load
	Medium structure	Low structure
	High clarity	Medium clarity
Cognition hypothesis	Medium different elements	Many different elements

	There and then/In the past and in future	Here and now
	Reasoning obligatory	Reasoning obligatory
	Perspective taking likely	Perspective taking obligatory

According to Skehan's criteria, in Task 1 much of the lexis needed is provided in the prompt, so learners have to demonstrate a medium vocabulary load, while in Task 2 much more is to come from the learner than from the prompt. The limitations of the time allotted for Task 1 (20 minutes) and of the length of the answer (about 150 words) account for the fact that the structure of the essay rigidly requires comparisons and descriptions of changes in this task, while in argumentative essay writers can choose the strategy with more freedom. Parameters in Robinson's Cognition Hypothesis for Task 1 show a medium number of elements to be discussed either in the past periods of the research, or the changes and trends from the past to the present or future, while in argumentative essays authors have to talk about many elements related to their argumentation, and mainly the discussion is restricted to here and now, occasionally giving an example from the past. Both tasks have to involve reasoning, while "perspective taking" is a must in giving reasons for thinking differently from the author's standpoint, while it may be expected less in describing the graphical material in Task 1 from the writer's perspective.

To follow the evaluative direction in the title, we bring in statistical data from two stages of learner corpus research. At the first stage, we looked at the difference in numbers of errors annotated manually, which we extracted with Python codes. The average length of the essays in these two genres is 181 vs 282 words, 156%. The first research question was this:

RQ1: Are the frequency numbers of different errors consistant with the 1.5 times difference in length of the essays of two task types?

The answer is clearly negative, so the research naturally took the direction of explaining what specific features of either task type can account for more errors attested for this type than for the other one. For the purposes of this stage in the analysis, we grouped about a hundred error labels used by annotators in the corpus to identify learner errors into 16 clusters, which you can see in Table 2. In order to make the data comparable, we calculated the normalized indices for 1000 words, and compared them rather than the raw total occurrences.

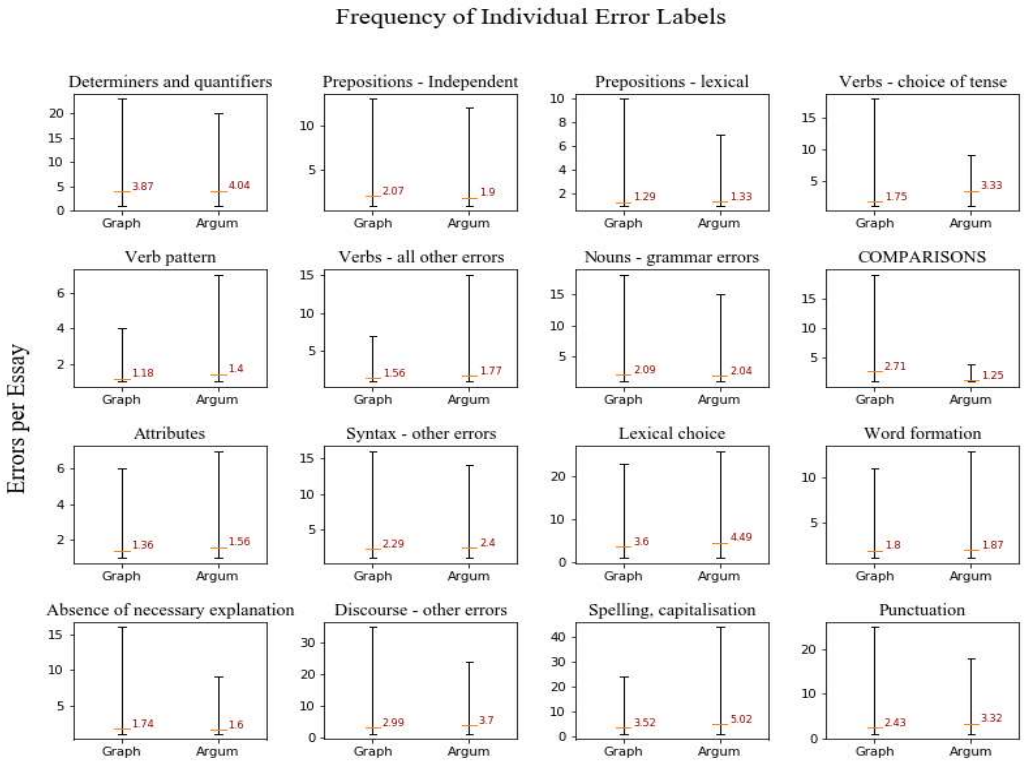
Table 2. Distribution of error tags in student essays of two tasks.

	Task1	Task2	Task1	Task2	Task2/Task1
			Normalized by 1000 words		
Determiners and quantifiers	6,580	7,081	15.05	10.70	71%
Prepositions - Independent	2,416	2,112	5.53	3.19	58%
Prepositions - lexical	340	641	0.78	0.97	125%
Verbs - choice of tense	3,575	1,504	8.18	2.27	28%
Verb pattern	97	710	0.22	1.07	484%

Verbs - all other errors	953	1,615	2.18	2.44	112%
Nouns - grammar errors	1,881	2,323	4.30	3.51	82%
COMPARISONS	1,363	149	3.12	0.23	7%
Attributes	273	610	0.62	0.92	148%
Syntax - other errors	3,107	3,662	7.11	5.54	78%
Lexical choice	6,114	8,123	13.98	12.28	88%
Word formation	1,395	1,817	3.19	2.75	86%
Absence of necessary explanation or detail	1,269	956	2.90	1.44	50%
Discourse - other errors	4,609	6,480	10.54	9.79	93%
Spelling, capitalisation	6,515	10,480	14.90	15.84	106%
Punctuation	3,513	5,701	8.03	8.62	107%

The lines without highlighting show "standard" distribution – i.e. the normalized number of errors in Task 2 almost equal to that of Task 1 (in the range from 80% to 120%). Blue lines show lower percentage or errors in Task 2 than the "standard" (below 80%), and the darker blue, the lower the level; the green lines, on the contrary, show the percentage of errors in Task 2 exceeding the expected "standard" above 120%.

Figure 1. Error frequency per task.



The second stage in the research was devoted to investigating the correlation between accuracy and complexity, which was the focus of many interesting papers dealing with specific areas or cases (Larsen-Freeman (2006); Alexopoulou et al. (2017)). We also decided to look at one specific area – namely, errors and complexity parameters related to syntax in learners' written production. At this stage our research hypothesis was the following:

RQ2: Is there a correlation between syntactic complexity and the number of syntactic errors?

To test this hypothesis, essays with one and more than five syntactic errors - 1,173 and 480 essays, respectively, - were elicited from REALEC with the Python code searching for annotator error tags. Syntax errors at this stage included the rest of the violations of the rules of syntax except Relative clauses, namely: agreement errors (*this problems; he insist)*, word order errors (*I don't know could I do this*), lack of parallelism of forms in coordination (*decided to go to him and asking about it)*, wrong forms in comparative constructions (*on two times more than before*), and, the last but very important, confusion of structures (for example, *it is* instead of *there is* - or the other way round (as in *It is a big poster above his bed in his room; There is important for him to be there);* or the possessive determiner instead of genitive *of (The USA's President*).

First, the distribution of syntactic errors in the two sets of essays by the type task was counted. It shows that there are more errors of these types in graph descriptions than in argumentative essays in both sets of essays – with 1 syntactic error and with 5 or more of them - relative to the essay length (to remind – argumentative essays are 1.5 times as long on average as graph descriptions):

Table 3.

Number of syntactic errors in graph descriptions with 1 syntactic error	Number of syntactic errors in argumentative essays with 1 syntactic error	Number of syntactic errors in graph descriptions with 5 or more syntactic errors	Number of syntactic errors in argumentative essays with 5 or more syntactic errors
597	**576**	**224**	**255**

Second, we received the parameters of syntactic complexity for all the essays under consideration. We applied to the dataset the following parameters of syntactic complexity out of 59 that had been chosen by Irina Panteleeva for the online system of text complexity analysis called Inspector (Olga Vinogradova et al. 2020):

1. average tree depth (av_depth)
2. average length of the sentence (av_len_sent)
3. average number of tokens before root (av_tok_before_root)
4. Levenshtein distance between lemmatized sentences - between all sentences (lemma_sim_all)
5. Levenshtein distance between lemmatized sentences - between neighbour sentences (lemma_sim_nei)
6. maximum tree depth (max_depth)
7. minimum tree depth (min_depth)

8. adjective clause modifier (num_acl)
9. number of adj+noun constructions (num_adj_noun)
10. adverbial clause modifier (num_advcl)
11. number of clauses (num_cl)
12. number of complex T-units (num_compl_tu)
13. number of coordinate phrases (num_coord)
14. number of noun+infinitive (num_noun_inf)
15. number of participle+noun (num_part_noun)
16. number of sentences (num_sent)
17. number of tokens (num_tok)
18. number of T-units (num_tu)
19. Levenshtein distance between pos-tagged sentences - between all sentences (pos_sim_all)
20. Levenshtein distance between lemmatized sentences - between neighbour sentences (pos_sim_nei)

Our research hypothesis includes as a part of it the surmise that different parameters will be significant in showing syntactic complexity for the two task types. To interpret the difference between the two tasks, two methods were applied: a *t*-test and the logistic regression. First, the table with the evaluation of every feature in every essay was made (Table 4 below). Then a *t*-test was applied. Since the number of essays is large, it is possible that the test statistics would follow a normal distribution if the value of a scaling term in the test statistics were known. However, scaling terms are not known, and must be replaced by an estimate based on the data. The independent two-sample *t*-test was applied. According to the null hypothesis, the mean values per feature between two samples were not significantly different. As has been mentioned, we had to estimate scaling terms, so, we used the following formula for that:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ , where } n \text{ is the number of essays, } \bar{x} \text{ is a mean of the sample.}$$

Then test statistics for all features were calculated, and *p*-values were received, which are given in Table 4 for graph descriptions with 1 syntactic error (_1) and with 5 and more syntactic errors (_5+), and in Table 5 for argumentative essays with the same numbers of syntactic error identified by the annotators.

Table 4. T-test for graph descriptions.

feature	**mean_1**	**Dispers_1**	**mean _5+**	**Dispers _5+**	**t-stats**	**p-val**	**1 – p-val, %**
av_depth	4.67	0.76	4.55	0.62	2.32	0.02	97.85
av_len_sent	22.57	28.31	21.78	24.54	0.39	0.70	30.48
av_tok_before_root	6.49	5.54	6.24	5.14	0.61	0.54	46.09
lemma_sim_all	24.30	40.67	23.61	32.84	0.25	0.80	19.60
lemma_sim_nei	23.84	39.24	23.12	31.91	0.27	0.79	21.20

max_depth	7.51	3.10	7.42	2.99	0.40	0.69	31.43
min_depth	2.50	1.04	2.37	0.97	1.76	0.08	92.09
num_acl	3.14	7.74	2.71	5.63	0.87	0.39	61.39
num_adj_noun	10.68	20.47	10.09	22.85	0.34	0.73	26.55
num_advcl	2.58	4.35	2.09	3.21	1.74	0.08	91.86
num_cl	15.30	20.97	15.21	18.79	0.06	0.95	4.57
num_compl_tu	3.69	7.65	3.37	5.43	0.67	0.50	49.74
num_coord	4.63	6.21	4.46	6.35	0.34	0.74	26.39
num_noun_inf	0.55	0.79	0.54	0.74	0.23	0.82	18.23
num_part_noun	2.45	6.23	1.90	4.32	1.44	0.15	85.03
num_sent	9.71	7.49	9.70	6.13	0.03	0.98	2.15
num_tok	210.35	2 198.48	204.02	1 938.09	0.04	0.97	3.20
num_tu	11.61	13.70	11.84	13.24	0.22	0.82	17.68
pos_sim_all	19.20	28.47	18.73	22.56	0.25	0.80	19.63
pos_sim_nei	18.77	27.63	18.28	21.92	0.26	0.79	20.61

Table 5. T-test for argumentative essays.

feature	**mean_1**	**Dispers_1**	**mean _5+**	**Dispers _5+**	**t-stats**	**p-val**	**1 – p-val, %**
av_depth	4.68	0.69	4.56	0.78	2.12	0.03	96.50
av_tok_before_root	5.93	3.11	5.60	2.61	1.61	0.11	89.29
num_advcl	7.45	13.10	6.26	10.43	1.40	0.16	83.84
min_depth	1.64	1.64	1.53	1.54	0.94	0.35	65.22
num_part_noun	2.07	4.12	1.85	3.95	0.74	0.46	53.94
av_len_sent	21.73	21.21	20.71	20.34	0.66	0.51	49.01
pos_sim_all	18.95	21.17	18.01	20.35	0.61	0.54	45.62
max_depth	8.59	4.33	8.42	3.95	0.56	0.57	42.66
num_coord	8.01	15.92	7.34	16.12	0.55	0.58	42.07
num_adj_noun	19.35	41.16	17.50	46.68	0.55	0.59	41.44
pos_sim_nei	18.14	16.76	17.40	18.92	0.54	0.59	41.31

num_compl_tu	10.99	26.50	9.94	26.06	0.53	0.59	40.69
lemma_sim_all	24.82	34.36	23.55	31.96	0.52	0.61	39.35
lemma_sim_nei	23.83	26.83	22.77	30.02	0.48	0.63	37.11
num_acl	4.52	7.28	4.29	7.77	0.41	0.69	31.48
num_noun_inf	3.07	3.89	2.96	4.21	0.33	0.74	26.06
num_cl	30.86	61.16	29.36	66.74	0.31	0.76	23.99
num_sent	15.33	15.60	15.02	18.02	0.24	0.81	18.86
num_tu	19.86	35.77	19.42	39.03	0.15	0.88	12.26
num_tok	323.22	4 694.18	302.80	6 086.86	0.05	0.96	3.81

The values in the last column of both tables show the probability that the means of two samples (with one error and with five and more errors) are significantly different, i.e. the difference is not accidental and must be explained by some external factor. We can be confident that the difference is due to the number of errors because the essays were initially divided according to that factor. However, it is possible, but highly unlikely, that the number of errors is not the external factor we are looking for.

The next step was the use of the logistic regression. The logistic regression is a statistical model that allows getting weights of every feature. We used Python module sklearn[2] to apply the regression.

First, the scaling of data was made using MinMaxScaler from sklearn.preprocessing[3]. This method provided the most accurate model results with the mean error 0.291. Then the model was trained using cross-validation (the sample was divided into five parts, and the model was trained five times, and every time the test sample was different). Then the coefficients of every feature were averaged, and results are presented in Table 6 for graph descriptions, and in Table 7 for argumentative essays.

Table 6. Coefficients for graph descriptions.

av_depth	0.18
av_len_sent	0.32
av_tok_before_root	0.46
lemma_sim_all	0.08
lemma_sim_nei	0.10
max_depth	0.23
min_depth	0.40
num_acl	0.40
num_adj_noun	0.48
num_advcl	0.90

Table 7. Coefficients for argumentative essays.

av_depth	0.14
av_len_sent	0.33
av_tok_before_root	0.56
lemma_sim_all	0.33
lemma_sim_nei	0.12
max_depth	0.24
min_depth	0.14
num_acl	0.42
num_adj_noun	0.89
num_advcl	1.37

[2] The documentation is available on https://scikit-learn.org/stable/

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

num_cl	0.24
num_compl_tu	0.17
num_coord	0.17
num_noun_inf	0.15
num_part_noun	1.14
num_sent	0.31
num_tok	0.25
num_tu	0.36
pos_sim_all	0.16
pos_sim_nei	0.12

num_cl	0.13
num_compl_tu	0.21
num_coord	0.25
num_noun_inf	0.15
num_part_noun	0.40
num_sent	0.07
num_tok	0.42
num_tu	0.29
pos_sim_all	0.25
pos_sim_nei	0.13

These numbers are relative, i.e. they do not refer to an outer scale, they themselves form the scale. So, to compare the results of two methods, again MinMaxScaler was applied. Then, numbers from the two tables were summed, and the table was sorted. The results are shown in Table 8 giving values for three sets - graph descriptions with 1 syntactic error and with 5 and more syntactic errors, argumentative essays with 1 syntactic error and with 5 and more syntactic errors, and both sets of essays together.

Table 8. Pivot table for graph descriptions, argumentative essays, and both sets of essays.

complexity parameters for graph descriptions with 1/5+ syntactic errors	t	LR	Sum	complexity parameters for argumentative essays with 1/5+ syntactic errors	t	LR	Sum	complexity parameters for all essays with 1/5+syntactic errors	t	LR	Sum
num_part_ noun	0.87	1.00	1.87	num_advcl	0.86	1.00	1.86	num_part_ noun	0.90	0.90	1.80
num_advcl	0.94	0.77	1.71	av_tok_before_ root	0.92	0.38	1.30	num_advcl	0.67	1.00	1.67
min_depth	0.94	0.30	1.24	av_depth	1.00	0.05	1.05	av_tok_before_ root	0.86	0.49	1.34
av_depth	1.00	0.09	1.09	num_adj_ noun	0.41	0.63	1.04	min_depth	0.95	0.27	1.22
num_acl	0.62	0.30	0.92	num_part_ noun	0.54	0.25	0.79	av_depth	1.00	0.07	1.07
av_tok_before_ root	0.46	0.36	0.82	min_depth	0.66	0.05	0.72	av_len_sent	0.54	0.44	0.98
num_adj_ noun	0.26	0.38	0.63	av_len_sent	0.49	0.20	0.69	num_acl	0.48	0.49	0.96
num_compl_ tu	0.50	0.08	0.58	pos_sim_all	0.45	0.14	0.59	num_adj_ noun	0.25	0.44	0.69
av_len_sent	0.30	0.23	0.52	lemma_sim_ all	0.38	0.20	0.58	max_depth	0.35	0.22	0.57
max_depth	0.31	0.14	0.45	num_acl	0.30	0.27	0.57	pos_sim_all	0.44	0.08	0.52
num_tu	0.16	0.26	0.43	num_coord	0.41	0.14	0.55	num_coord	0.30	0.16	0.46
num_coord	0.25	0.08	0.34	max_depth	0.42	0.13	0.55	lemma_sim_ all	0.40	0.06	0.45

pos_sim_all	0.18	0.08	0.26	num_compl_tu	0.40	0.11	0.51	num_compl_tu	0.20	0.24	0.44
num_noun_inf	0.17	0.07	0.23	pos_sim_nei	0.40	0.05	0.45	pos_sim_nei	0.42	0.00	0.42
pos_sim_nei	0.19	0.04	0.23	lemma_sim_nei	0.36	0.04	0.40	lemma_sim_nei	0.39	0.02	0.42
lemma_sim_nei	0.20	0.02	0.22	num_noun_inf	0.24	0.06	0.30	num_tok	0.00	0.39	0.39
num_sent	0.00	0.22	0.22	num_tok	0.00	0.27	0.27	num_noun_inf	0.14	0.21	0.35
lemma_sim_all	0.18	0.00	0.18	num_cl	0.22	0.05	0.26	num_cl	0.01	0.20	0.21
num_cl	0.03	0.15	0.18	num_tu	0.09	0.17	0.26	num_tu	0.05	0.11	0.16
num_tok	0.01	0.16	0.17	num_sent	0.16	0.00	0.16	num_sent	0.02	0.04	0.06

Colour legend for cells: *green – value for graph descriptions is close to value for all essays; blue - value for argumentative essays is close to value for all essays; yellow –value for both task types are lower than value for all essays; grey – values for either task type and for all essays are very close; pink –value for one task type is higher than both the value for the other type and that of all essays; purple – value for one task type is lower than both the value for the other type and that of all essays.*

Results and discussion

The data reflecting error frequencies may seem controversial at a first glance, so we will try to account for the differences within and between the classes. Some deviations from the "standard" can be explained by the specific features of the tasks. Comparative constructions, errors in which are much more frequent in the descriptions of graphical materials than the same errors in argumentative essays, are explicitly required in Task 1: "…***make comparisons where relevant."*** One more prevalence of error type in graph descriptions over argumentative essays can be explained by the special requirement of the task – examinees are expected to provide all the details presented in the task in every statement (when/where/brand name/age group/numbers or some other details) while describing the graphical material, so when they fail to include these details, annotators mark it with a special discourse tag (Absence of necessary explanation or detail).For some other features the only possible reason for the discrepancy between Task 1 and Task 2 can be found in the phenomenon of interference with the native language – a factor mentioned in many papers from SLA research. Such is the case with frequent appearance of present tense forms (Present Simple or Present Perfect) in the context with the clear definition of time in past in graph descriptions written by Russian L1 students, even those with a very high level of proficiency. The plausible explanation is that such uses are commonly found in Russian academic discourse reflecting past events or phenomena. Compared to English, a very different convention for the choice of tense can be observed in Russian, as illustrated by the following statistics: about 10% of 10 most frequent verbs used in academic texts in the National Corpus of the Russian language are used in the present tense form in the contexts with clear references to the past events. As native speakers of Russian, the authors confirm that the existence of both past and present forms of verbs in describing past actions, sometimes even in one and the same sentence, do not seem wrong in academic discourse. To prove it, we carried out the experiment across a database of academic papers in sociology written in Russian, from which we extracted descriptions of data from past years presented in the papers with some graphical material. Both manual and automated extraction of sentences talking about past events or phenomena was performed, and only those sentences were included in the research dataset that include the time

in the past stated or implied from the nearest context. As a result, out of 221 verbs only 130 were forms of the Russian past tense, while 91 were forms of the Russian present tense, participles with the omitted present tense auxiliary verb among them. So, it is no wonder that learners even at an advanced level of English proficiency demonstrate a huge deviation from consistent use of Past tense when they look at past events while describing some academic research. Correspondingly, in Task 1 essays, the level of erroneous uses of tenses is much higher than that in Task 2 essays, where authors talk predominantly about here and now, even though they sometimes add a perspective or some comparison with the past. At the same time all other types of errors in using verbs (use of modals, voice, different verb patterns, non-finite constructions) are within standard distribution between the two task types in line with the difference in their length, i.e. about 1,5 times as many errors in argumentative essays as in graph descriptions.

Errors in the use of prepositions in our research are difficult to get their correct number, but we designed a method to combine search for error tags with search for special syntactic and lexical prepositional constructions, and in the end the picture was very different for errors in the prepositions used in adverbial modifiers (called "Independent" in Table 2), where the errors in graph descriptions were much more numerous than those in argumentative essays, and in prepositional collocations – prepositional nouns (*need **for***), or prepositional adjectives (*typical **of***), or parts of idioms (*see eye **to** eye **with** smb*) – the frequency of errors in argumentative essays was slightly higher than "standard", that is a little more than 1.5 times as high as that of the graph descriptions. It is clear that the need to describe many factual details in Task 1 elicits plethora of adverbial modifiers with prepositions as their part – they are a must in practically every sentence; nevertheless, the narrative, as stated in (2017:13)), is supposed to elicit a lot of temporal and locative adverbial modifiers, also often requiring prepositions, and narrative is not a must, but is often present in argumentative essays.

In the area much less often covered in research papers – misuses of punctuation by learners of English – and also in the frequency of spelling and capitalisation errors we observed standard distribution of errors between descriptions of graphical materials and argumentative essays consistent with the difference in the length of the essays written in answer to the two tasks.

The situation contrary to the described in the paragraph above was attested in the errors in Relative clauses, with the much higher level of them in argumentative essays than just 1,5 times as many. Errors in the use of relative clauses, as they are annotated in REALEC, are in fact of different nature: first, the attempt to make a distinction between defining and non-defining types of relative clauses is so much of a challenge with no such distinction in Russian that even EFL professionals are prone to confusion; second, one and the same Russian equivalent is used in Russian to the three English relativisers – *who*, *that* and *which*, and it leads to many erroneous uses of those three English relativisers by Russian learners of English; finally, Russian learners, having developed an almost automatic urge to use commas for the corresponding construction in Russian, tend to apply commas to defining relative clauses, where they cannot be used (2nd most frequent misuse of commas in the corpus). Nevertheless, all types of such errors, no matter of what nature, are annotated with the tag "Relative clauses". In view of all these factors, we

decided not to include the cases annotated with relative clause error tags in the class of syntactic errors, nor consider the number of relative clauses as the parameter of syntactic complexity.

The thorough analysis of the 20 parameters of syntactic complexity implemented in the system of automated text complexity inspection revealed that the **number of participle + noun** constructions, the **number of adjective clause modifiers** and **minimum tree depth** are the best predictors of difference between almost no syntactic errors and many of them in the graph descriptions, while the **number of adverbial modifiers**, the **number of adjective + noun** constructions, **average number of tokens before root**, **number of coordinate phrases** and **maximum tree depth** are the best predictors of difference between better and worse essays from the point of view of syntactical accuracy. Such parameter as **average tree depth** is a good predictor for both task types – graph descriptions and argumentative essays. Thus, all these features allow us to predict the number of syntactic errors, relying on the parameters of syntactic complexity, and the features are different for the two task types - graph descriptions and argumentative essays.

Conclusions and further research

Before drawing the conclusions, we have to admit that this research aimed at exploring the possibility to apply Natural Language Processing techniques within the automated system of text complexity and linguistic tools available in the learner corpus to test hypothesis about the correlations between task types and error frequency, on the one hand, and task types and syntactic complexity, on the other. The exploratory nature of the research may account for the tentativeness of our results, which are likely to show only some of the many factors needed for the wider-scope modelling both task-based language teaching for SLA and developmental studies for learner corpus research.

Obviously, influence of every feature on the overall performance of students requires further study, and even in our learner corpus, similar research has to be carried out to test the correlation between frequency of other types of errors and parameters of other domains of texts complexity. However, now we know how to narrow down the list of features to focus on and how to design further experiments accordingly. One more direction for further studies is to calculate target-like use in obligatory contexts in order to calculate accuracy and not just error frequency. As suggested by Meurers and Dickinson (2017), we will need to calculate overuse/underuse of certain patterns and constructions separately for Task 1 and Task 2.

Nevertheless, even at this stage, the presented corpus research yields useful knowledge for English instructors preparing their students for the English examination with the two tasks under consideration - from the task-based insights to the automated tracing of specific syntactic features in student essays of either type in the attempt to prevent specific errors related to syntax.

References

Alexopoulou, Th., Michel, M., Murakami, A. & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language

Processing techniques. *Language Learning*, 67 (S1), 2017, 180–208. https://doi.org/10.1111/lang.12232.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4), 590–619. https://doi.org/10.1093/applin/aml029

Meurers, D. & Dickinson, M. (2017). Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. *Language Learning, 67(2).*

Plonsky, L., & Kim, Y. (2016). Task-Based Learner Production: A Substantive and Methodological Review. *Annual Review of Applied Linguistics, 36*: 73-97. doi:10.1017/S0267190516000015

Robinson, P. (1995). Task complexity and second language narrative discourse. *Learning Language,* 45(1):99–140.

Robinson, P. & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. IRAL, 45(3), 161–176.

Robinson, P., Cadierno, T. & Shirai, Y. (2009). Time and Motion: Measuring the Effects of the Conceptual Demands of Tasks on Second Language Speech Production. *Applied Linguistics*, 28(4):533-554.

Skehan, P. (1998). A cognitive approach to language learning. Oxford University Press, Oxford.

Skehan, P. (2003). Task-based instruction. *Language teaching*, 36(1): 1–14.

Olga Vinogradova, Olga Lyashevskaya, Irina Panteleeva (2020). Automated Assessment of Learner Text Complexity. To appear in *Assessing Writing* (Manuscript Number: ASW-D-20-00017).