



<https://doi.org/10.22363/2687-0088-31187>

Research article

Word-formation complexity: a learner corpus-based study

Olga LYASHEVSKAYA^{1,2}  , Julia PYZHAK¹ 
and Olga VINOGRADOVA¹ 

¹National Research University Higher School of Economics, Moscow, Russia

²Vinogradov Russian Language Institute of the Russian Academy of Sciences,
Moscow, Russia

olesar@yandex.ru

Abstract

This article explores the word-formation dimension of learner text complexity which indicates how skilful the non-native speakers are in using more and less complex – and varied – derivational constructions. In order to analyse the association between complexity and writing accuracy in word formation as well as interactive effects of task type, text register, and native language background, we examine the materials of the REALEC corpus of English essays written by university students with Russian L1. We present an approach to measure derivational complexity based on the classification of suffixes offered in Bauer and Nation (1993) and then compare the complexity results and the number of word formation errors annotated in the texts. Starting with the hypothesis that with increasing complexity the number of errors will decrease, we apply statistical analysis to examine the association between complexity and accuracy. We found, first, that the use of more advanced word-formation suffixes affects the number of errors in texts. Second, different levels of suffixes in the hierarchy affect derivation accuracy in different ways. In particular, the use of irregular derivational models is positively associated with the number of errors. Third, the type of examination task and expected format and register of writing should be taken into consideration. The hypothesis holds true for regular but infrequent advanced suffixal models used in more formal descriptive essays associated with an academic register. However, for less formal texts with lower academic register requirements, the hypothesis needs to be amended.

Keywords: *linguistic complexity, morphological complexity, writing accuracy, word formation, English, learner corpora*



For citation:


Lyashevskaya, Olga, Julia Pyzhak & Olga Vinogradova. 2022. Word-formation complexity: a learner corpus-based study. *Russian Journal of Linguistics* 26 (2). 471–492. <https://doi.org/10.22363/2687-0088-31187>

Научная статья

Словообразовательная сложность и ошибки учащихся в экзаменационных эссе

О.Н. ЛЯШЕВСКАЯ^{1,2}  , Ю.В. ПЫЖАК¹ ,
О.И. ВИНОГРАДОВА¹ 

¹Национальный исследовательский университет «Высшая школа экономики»,
Москва, Россия

²Институт русского языка им. В. В. Виноградова РАН, Москва, Россия
olesar@yandex.ru

Аннотация

В статье рассматривается словообразовательная сложность учебных текстов, которая трактуется как система измерений, показывающих разнообразие приемов словообразования разного уровня, от простых до продвинутых, используемых учащимися. Анализируется взаимосвязь между сложностью и ошибками, которые учащиеся допускают в словообразовании. Исследование основано на материалах REALEC – корпуса английских экзаменационных эссе, написанных студентами университета с родным русским языком. Предлагается подход к измерению словообразовательной сложности, основанный на классификации суффиксов Бауэра и Нейшена (Bauer & Nation 1993), и анализируется соответствие между показателями индексов сложности и количеством ошибок словообразования, размеченных в текстах корпуса, с учетом типа экзаменационного задания. Постулируется гипотеза о том, что с увеличением сложности количество ошибок должно уменьшаться, и проводится статистический анализ параметров сложности и безошибочности. В работе показано, во-первых, что использование словообразовательных суффиксов более высокой сложности связано с количеством ошибок в текстах. Во-вторых, разные уровни иерархии сложности оказывают разнонаправленное влияние на точность: в частности, использование нерегулярных словообразовательных моделей положительно связано с количеством ошибок. В-третьих, следует учитывать тип экзаменационного задания, в том числе ожидаемые формально-регистрационные особенности текста. Гипотеза была подтверждена для регулярных, но нечастотных суффиксальных моделей при их использовании в описаниях рисунков и графиков – текстах, следующих определенному формату и включающих элементы академического письма. Однако в случае аргументативных эссе выдвинутая гипотеза требует уточнения.

Ключевые слова: лингвистическая сложность, морфологическая сложность, безошибочность письма, словообразование, английский язык как иностранный, учебные корпуса

Для цитирования:

Lyashevskaya O.I., Pyzhak J.V., Vinogradova O.N. Word-formation complexity: a learner corpus-based study. *Russian Journal of Linguistics*. 2022. Vol. 26. № 2. P. 471–492. <https://doi.org/10.22363/2687-0088-31187>

1. Introduction

Text complexity is one indicator that can be used to assess authors' written and spoken skills and how varied and complex are the means of linguistic expression

they apply. Starting with the work of Norris and Ortega (2009) and two publications by Bulté and Housen (2012, 2014), researchers agree that complexity is a multifaceted phenomenon, consisting of several sub-constructs, dimensions, levels, and components, and many of these constructs and sub-areas have been independently evaluated. Consequently, the term “complexity” has been widely applied to manifestation of objective properties of linguistic production. Bulté and Housen (2012) also draw the distinction between propositional complexity, discourse-interactive complexity, and linguistic complexity. Of these three, linguistic complexity of written production, will be the target of the present article.

In addition to the general consent on including lexical, syntactic, discursive, and morphological parameters of texts when looking at text complexity, it soon became clear that the majority of researchers focus on the first three areas, while morphological and phonological complexity or complexity phenomena at the interfaces between the traditional levels of linguistic analysis were largely ignored, which resulted in some appeals to the linguistic community to expand the construct of, and research on, complexity beyond the syntactic and lexical levels. In 2019, Centre for English Corpus Linguistics at UCLouvain hosted the colloquium, Broadening the Scope of L2 Complexity Research, and in a way the research in this paper is our contribution to bridging this gap. More specifically, we set aside the topic of inflectional complexity, which has already gained considerable attention in learner data analysis and has acquired its own methodology (Brezina & Pallotti 2016, Yoon 2017, Tywoniw & Crossley 2020), and focus exclusively on the diversity of word-formation models.

Roots and affixes are the building blocks of morphological competence. Acquiring the morphology of a second language can be seen as not only learning strategies for composing and decomposing forms from and to the building blocks and enhancing vocabulary. It is also learning relations between the stored word forms that facilitates processing at the morphological level (Baerman et al. 2015). Word-formation skills give L2 learners flexibility when modifying and adapting word meaning to the context, coercing a word’s syntactic class to fit into an available grammatical construction, choosing the right vocabulary for a given type of discourse, or in constructing new words. Inappropriate use of derivational models may result in communication fallacy and other losses in social interactions, just as any kind of erroneous linguistic behaviour may do.

Recent discussion on the status of word(-formation) families has brought to light many derivation-related issues important for L2 research and education (Brown et al. 2020, Laufer et al. 2021, Nation 2021). To put it in a nutshell, the researchers emphasise that the derived forms cover a significant portion of texts and hence learners’ word-formation competence impacts text comprehension. Affix knowledge develops with general proficiency and facilitates vocabulary learning. However, many advanced learners have limited or patchy knowledge of affixes and find it challenging to identify derivational forms of known headwords given in context, even in structures with top-frequent affixes. Brown et al. (2020) present some evidence that the learners are concentrated on the recognition of the affix

meaning rather than its grammatical function. Although the evidence provided largely concerns the receptive aspects of acquisition, the discussion has important implications for the theories of L2 production.

In the task of examination writing, learners are expected to achieve a balance in the interaction of the performance areas such as complexity, accuracy, and fluency in choosing preferable derivational strategies. Although the lack of high-level and complex skills is not the only source of word-formation errors and that low writing quality may also be related to, for example, time and stress management, genre of the task and familiarity of the topic, and learners' individual effects, learner corpora and annotated research datasets provide the basis for empirical studies focused on the relationship between complexity and accuracy (Skehan 2009, Bardovi-Harlig & Bofman 1989, Plakans et al. 2019, Lahuerta 2018, among others).

Our study is based on the learner corpus REALEC (Vinogradova et al. 2017), which includes examination writings of students with Russian L1. We examine the use of word-formation constructions with the focus on suffixal ones, since their number in the examination texts significantly exceeds the number of prefixes and other word-formation units. We will test the following hypothesis: the higher the parameter of derivational morphological complexity, the less often errors occur in word formation, since the scale for measuring morphological complexity is based on the order in which students studying English as a foreign language learn derivational affixes. Accordingly, the more the student uses advanced suffixes, the higher their language proficiency.

2. Word-formation complexity

There are several ways to define complexity, among which relative complexity (difficulty) and absolute (structural) complexity are most discussed (de la Torre García et al. 2021). Relative complexity refers to the cognitive difficulty of the task, the amount of effort and resources that a speaker has to employ in order to process and make use of a linguistic structure. In contrast to this, absolute complexity is defined as the numerical characteristics of a text based on the quantity of encoded and encoding linguistic units and the number of connections between these components. Morphological complexity belongs to the area of formal parameters of absolute complexity, according to the classification of parameters critical for measuring the acquisition of a target language (Bulté & Housen 2012), see Fig. 1. These are features which can be measured objectively at the word level in a text.

Figure 1 shows that the criteria of morphological complexity are divided into two groups: inflectional and derivational. The first type refers to the use of grammatical forms, for example, the frequency of tense forms, the frequency of modals, the number of different verb forms, the variety of past tense forms, and MCI (Morphological Complexity Index) (Brezina & Pallotti 2019). The second group of criteria deals with the use of derivational affixes, composites (multi-root words), and conversives, and refers to the variability of word-formation models and the size of word families.

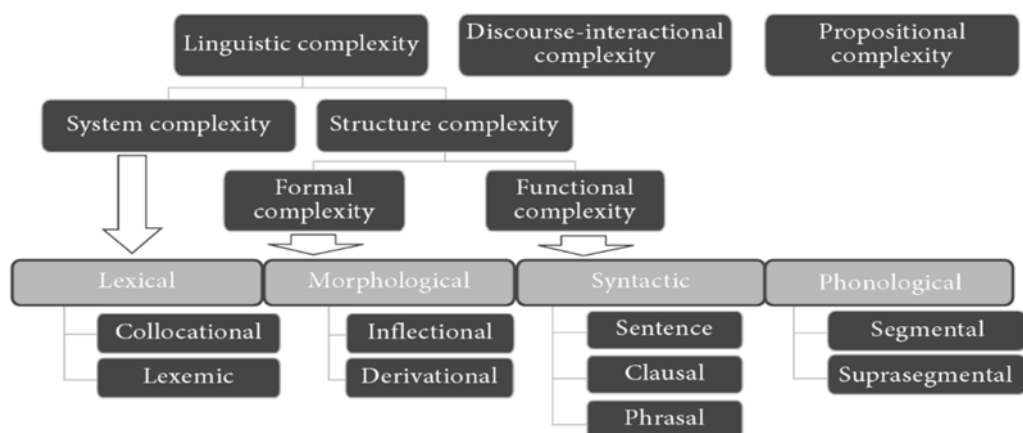


Fig. 1. A taxonomy of absolute complexity criteria (Bulté & Housen 2012: 23)

Word-formation complexity criteria have gained much less attention compared to inflectional complexity criteria. A possible explanation for this is that there is low agreement among theoretical approaches to this level of linguistic description, the coverage of derivational models, affixes, and that their taxonomy in lexical resources is sketchy and inconsistent (cf. Table 1), and it is problematic to use (a very few) tools for automatic morpheme segmentation of non-standard texts. It is indicative that Biber used only one derivation-related feature Nominalizations (words ending in *-tion*, *-ment*, *-ness*, *-ity*) among the other 67 linguistic features in his multidimensional analysis of English register (Biber 1988). A slightly longer list of regular affixes (*-able*, *-er*, *-ish*, *-less*, *-ly*, *-ness*, *-th*, *-y*, *non-*, and *un-*) was used in the calculation of a types-per-family ratio in a study of vocabulary structure (Horst & Collins 2006), yet no distinction was drawn between inflection and derivation in the construct of word families.

Recently, Tywoniw and Crossley (2020) introduced the TAMMI method, in which the density of derived and non-derivational words is measured in the text, among other metrics of (inflectional) morphological complexity. To calculate the Derived Word Tokens per Word index, they divide the number of word tokens with any derivational affixes by the number of words in a text. The Derived Word Types per Type index is the number of distinct word types with any derivational affixes divided by the number of types in a text. Similar calculations apply to the measurement of Non-Derivational Word Tokens per Word and Non-Derivational Word Types per Type indices.

Tywoniw and Crossley's approach is a generalisation over statistical measures developed for tokens (N) and types (V) of a given morphological class in corpora-based research (Plag et al. 1999). Other derivational complexity indicators recommended in the literature include:

- the length of the affix;
- the complexity-based rank, constructed on the affix ordering principle: suffixes with a rank greater than the rank of a given affix may follow that suffix in a word, while suffixes with rank lower than this will never follow it;

– the derivative’s junctural phonotactics: the probability of the sequence of sounds spanning the juncture between the morphemes (e.g. “nh”, which is highly unlikely to occur within morphemes and more likely to create morphological boundaries, as in *inhumane*) (see Baayen 2009 for an overview).

Table 1. Numbers of distinct affixes attested in English L1 corpora, dictionaries, and other resources according to (Laws & Ryder 2014: 6, 10)

	Word-initial affixes	Word-final affixes	Totals
Marchand (1969)	65	104	169
Hay and Baayen (2002)	26	54	80
Stein (2007)	547	296	843
<i>Incl. prefixes/suffixes</i>	171	164	
combining forms	125	107	
both	251	25	
Affixes in BNC	268	222	490
<i>Incl. prefixes/suffixes</i>	96	141	
combining forms	41	61	
both	131	20	
Affixes in MorphoQuantics, not in BNC	286	59	345
<i>Incl. prefixes/suffixes</i>	81	22	
combining forms	84	32	
both	121	5	

Bauer and Nation (1993) proposed a scale for categorising English affixes on the following criteria:

- 1) frequency – the number of different words an affix occurs in;
- 2) productivity – the likelihood that the affix will be used to form new words;
- 3) the predictability of the meaning of the affix;
- 4) the regularity of the orthographic form of the base – the predictability of change in the written form when the affix is added;
- 5) the regularity of the spoken form of the base – the amount of phonetic change when the affix is added;
- 6) the regularity of the spelling of the affix – the number of allomorphs attested in different words;
- 7) the regularity of the pronunciation of the affix;
- 8) the regularity of function – the degree to which the affix attaches to a base of a known form-class and produces a word of a known form-class.

Their classification implies that the affixes are acquired at different stages of language mastery. For example, the first level means that a speaker perceives a different form as a different word. Being on the second level, they understand that regularly inflected words are part of the same family. At subsequent levels, different derivational affixes are acquired, see Table 2 for higher level suffixes. One can notice that the same orthographic manifestations of suffixes can be attested at different levels (marked with asterisk in Table 2), cf. the suffix *-able* that appears at both the third and 6th levels. The third level includes cases of attaching this suffix

to transitive verbs, which is a very frequent, productive, and regular word-formation construction. At the 6th level, the base has to be truncated (e.g., the suffix *-ate* removed) when the suffix *-able* is attached, as in *attenuable*, *permeable*, thus leading to more complexity. At the last, seventh, level, the student demonstrates near-native knowledge of every existing morpheme including rare combinations of affixes and roots of Latin and Greek origin, cf. *differentiate*.

Table 2. Levels of English suffixes, according to (Bauer & Nation 1993)

Level	Suffixes
3: the most frequent and regular affixes	<i>-able*</i> (<i>eatable</i>), <i>-er</i> (<i>writer</i>), <i>-ish</i> (<i>selfish</i>), <i>-less</i> (<i>endless</i>), <i>-ly*</i> (<i>fortunately</i>), <i>-ness</i> (<i>kindness</i>), <i>-th*</i> (<i>fourth</i>), <i>-y*</i> (<i>smelly</i>)
4: frequent, orthographically regular affixes	<i>-al*</i> (<i>normal</i>), <i>-ation</i> (<i>preparation</i>), <i>-ess</i> (<i>heiress</i>), <i>-ful</i> (<i>useful</i>), <i>-ism</i> (<i>socialism</i>), <i>-ist*</i> (<i>socialist</i>), <i>-ity</i> (<i>sensitivity</i>), <i>-ize</i> (<i>legalize</i>), <i>-ment</i> (<i>government</i>), <i>-ous</i> (<i>ambitious</i>)
5: regular but infrequent affixes	<i>-age</i> (<i>percentage</i>), <i>-al*</i> (<i>approval</i>), <i>-ally</i> (<i>idiotically</i>), <i>-an</i> (<i>American</i>), <i>-ance</i> (<i>clearance</i>), <i>-ant</i> (<i>consultant</i>), <i>-ary</i> (<i>revolutionary</i>), <i>-atory</i> (<i>confirmatory</i>), <i>-dom</i> (<i>kingdom</i>), <i>-en</i> (<i>wooden</i>), <i>-en</i> (<i>widen</i>), <i>-ence</i> (<i>emergence</i>), <i>-ly*</i> (<i>leisurely</i>) ...
6: frequent but irregular affixes	<i>-able*</i> (<i>permeable</i>), <i>-ee</i> (<i>nominee</i>), <i>-ic</i> (<i>geographic</i>), <i>-ify</i> (<i>quantify</i>), <i>-ion</i> (<i>description</i>), <i>-ist*</i> (<i>tobacconist</i>), <i>-ition</i> (<i>addition</i>), <i>-ive</i> (<i>representative</i>), <i>-th*</i> (<i>length</i>), <i>-y*</i> (<i>diplomacy</i>)

The consistency of Bauer and Nation’s hierarchy was confirmed in Leontjev (2016): the study tested how successfully the learners of English recognised affixes belonging to different levels. With the exception of no difference between levels 5 and 6, the lower the affix level, the easier it is for a student to recognise it. However, other studies suggest that the Bauer and Nation levels “only partly agree with learner knowledge data” (Nation 2021: 969). There are also known challenges in mapping the affix levels to the CEFR Lexical Profile levels and word meanings (Capel 2010).

In Lyashevskaya et al. (2021), Bauer & Nation’s classification of suffixes was operationalised through four quantitative indices for levels 3 to 6. To calculate the index of each level, they divide the number of tokens with suffixes of level N by the number of tokens that have any inflectional or derivational suffix. Although a wider variety of indices can be designed that take into account all types of word-formation devices (e.g. prefixes and morphemes within composite words), the relation between the suffix-based derivational indices and inflectional complexity index is straightforward for English morphology since both calculation methods rely on the number of suffixes that have to be correctly supplied in writing.

3. Related research

While many studies have addressed the relationship between morphological complexity and the level of proficiency in a foreign language, unsurprisingly, most research is based on inflectional complexity only under the notion of morphological complexity. Over the years, a wealth of methods, such as MCI, and tools, such as

LancsBox (Brezina et al. 2020), have been developed to facilitate the empirical study of inflectional data. One of the illustrative examples is Brezina and Palloti (2019), whose aim was to reveal the relationship between the realised inflectional complexity in texts of Italian learners and their level of language proficiency. They show significant correlation between measures of inflectional complexity and other indicators of complexity such as a standardised type-token ratio and sentence length. However, the effects are not observed in groups of advanced learners in the study based on written argumentative essays in English produced by Italian university students, taken from the International corpus of learner English (ICLE). In the case of English, which is less complex than Italian in terms of verbal inflection, the authors at CEFR levels B1 to C1 demonstrate a native-like ability, thus reaching a threshold “after which inflectional diversity remains constant” (Brezina & Palloti 2019: 99).

Ehret and Szmrecsanyi (2019) studied the relationship between the number of years of L2 instruction in English and morphological and syntactic complexity as well as holistic, information-theory-based complexity of text (Kolmogorov complexity). The authors found that writers with higher levels of language mastery wrote more complex texts, although the relationship between complexity and proficiency was not always linear. While holistic text complexity and morphological complexity of the text increased, the syntactic complexity might decrease as more advanced students were less likely to adhere to the correct word order.

The effect of the derivational complexity of the native language in L2 acquisition was investigated in van der Slik et al. (2019). In particular, the authors found a link between the derivational complexity of a student's native language and their success in learning Dutch. Students whose native language was morphologically less difficult than Dutch found it more difficult to acquire the morphological system of Dutch.

In Kimppa et al. (2019), the acquisition of word formation in adult students of Finnish was studied using psycholinguistic methods. It was found that more advanced students showed performance comparable to that of native Finnish speakers when processing derivational morphology. This suggests that with an increase in the level of proficiency in a foreign language, the processing of derivational morphology also develops, and some learners can reach the level of a native speaker.

Our study aims at narrowing the gap in the studies of word-formation complexity by focusing attention on its relationship to accuracy. It addresses the following research questions:

1. Is there an association between word-formation complexity and error-free use of the word-formation models in L2 English production?

2. Is there a task effect? Does the word-formation complexity affect the frequencies of the word-formation errors in Task 1 texts the same way as in Task 2 texts? Which specific levels of word-formation complexity contribute to the accuracy of word formation the most?

4. Data and method

This study is based on REALEC, Russian Error-Annotated English Learner Corpus. This corpus currently includes about 6,000 texts (roughly 1.5 million words) of the examination essays written by Russian-speaking learners of English during their second-year examination at university. The writing tasks are similar to those used in the IELTS examination. The first task tests the ability to describe graphic material in the task (graph description essay), and in the second one the participant is expected to express their opinion about a certain problem given in a short written text prompt (argumentative essay). The required size of the first type of essay is at least 150 words; the second one, at least 250. A substantial part of the corpus was manually processed by EFL experts: they annotated and corrected errors at different linguistic levels from orthographic and morphological to discursive.

We have selected from the corpus 1307 examination texts containing errors in suffixal word formation. In our sample, there are no texts without derivational errors due to the well-known problem of recall in spotting such errors by experts. The errors that we selected were labelled with the following tags: “Formational suffix”, “Word formation” and “Confusion of categories”. The “Formational suffix” tag marks inappropriate use of a suffix or its absence where a suffix is needed – see examples (1) and (2). The “Word formation” tag combines three types of errors: incorrect use of both a suffix and a prefix, or the absence of both where they are needed, or the combination of the first two types – see (3) and (4). The corrections suggested by experts for the errors in focus in this research are given in square brackets, while corrections suggested for all other errors are not presented, so the authors’ spelling, grammar and vocabulary are intact in the examples.

- (1) *Business books empahsize, that society and all its members must feel confident in every step taking by **politics [politicians]** and what is more, to have an essence of non-restricted protection.*
- (2) *First of all, if we speak of equality of men and women we should make a **notice [note]** that this also mean that women could not do some work which is not suit them (take heavy things).*
- (3) *One of researches showed, that the **borned [inborn]** characteristics more important for our personality and development.*
- (4) *Today a lot of international organization move their businesses to **undeveloped [developing]** countries.*

The “Confusion of categories” label is used to mark cases of incorrect choice of a part of speech from a word family, see (5). This tag, unlike the previous two, describes the nature of the error rather than the formal type of what needs to be corrected in the error span. However, among the errors labelled with this tag, there are also cases of incorrect use of word-formation suffixes.

- (5) *What is about global warming, recent studies say that it is a result of **climatic [climate]** changes which are essential for the earth.*

Spelling mistakes in affixes and on morpheme boundaries such as *useage* (cf. *usage*), *privelage* (cf. *privilege*), *begining* (cf. *beginning*) are tagged as

“Spelling” and not as errors in word formation, according to the REALEC annotation scheme. Such cases are excluded from consideration. We also removed texts in which only the incorrect uses of prefixes and composites, and not suffixes, were annotated.

The resulting dataset consists of 595 graph description essays and 712 argumentative essays. Table 3 reports the number of errors labelled in the texts of each examination task type. Most of the texts contain only one derivational error. Texts in which no more than four derivational errors were reported account for 96% of documents in each type.

Table 3. Distribution of the number of errors in two types of examination writing

# errors	graph description	argumentative essay	total
1	376	391	767
2	118	170	288
3	44	95	139
4	33	25	58
5	12	17	29
6	8	5	13
7	2	5	7
8	1	1	2
9		2	2
10		1	1
11	1		1
total	595	712	1307

There is no available information regarding the proficiency level of the authors, however, it can be estimated within the range from B1 to C1 on the CEFR scale.

To measure derivational complexity, we used the application Inspector (Lyashevskaya et al. 2021). In each text, it calculates (token-wise) the number of word-formation affixes on each of the levels 3 (most frequent and regular), 4 (frequent and orthographically regular), 5 (infrequent and regular), and 6 (frequent and irregular) of Bauer & Nation’s classification of word affixes (see Section 2). The simple base forms are identified using the nltk package PorterStemmer, after which the total number of suffixed words in the text is calculated, thus taking into account both inflectional and derivational morphemes at the end of the word. The relative metrics of each level are calculated according to the formula:

$$\frac{\text{number of suffixes on } n\text{th level}}{\text{number of suffixed words}}$$

The basic statistics of the mean, standard deviation (SD), median values of the suffix level indices, and accuracy index (number of errors) are summarised in Table 4. The average length of texts is: 182.8 words ($SD=37.4$) for the graph description essays and 275.2 words ($SD=62.3$) for argumentative essays.

Table 4. Descriptive statistics of word-formation complexity measures and word-formation errors

task	stats	level 3	level 4	level 5	level 6	# errors
graph description	Mean	0.033	0.083	0.046	0.042	1.709
	SD	0.032	0.071	0.041	0.042	1.234
	Median	0.026	0.069	0.038	0.029	1.000
argumentative essay	Mean	0.049	0.080	0.067	0.052	1.829
	SD	0.034	0.042	0.039	0.034	1.256
	Median	0.045	0.075	0.065	0.045	1.000

Several types of statistical techniques were used to investigate the research questions. The Pearson's analysis, which presupposes the continuous distribution of variables, was conducted to detect possible correlation among complexity indices, while the non-parametric rank-based Kendall correlation analysis was applied to measure pairwise the ordinal association between continuous measures of complexity and a discrete (paucal integer) measure of accuracy.

Analysis of variance of the complexity and accuracy measures was conducted to compare the groups of texts such as graph description and opinion essays, or essays with one vs. more than one word-formation error. Since the measures are distributed non-normally we performed a non-parametric Kruskal-Wallis rank-sum test (*kruskal.test* function for two groups in the R package *stats*, R Core Team 2019, Hollander & Wolfe 1973: 115–120, 185–194).

In addition, we applied two regression algorithms. We assume that the word-formation errors follow a Poisson distribution, since it describes the likelihood of events that occur over a fixed period of time, and the events are independent of each other. When a student writes an essay, an error may or may not occur at any given moment. We used a Poisson regression (*vglm* function in the R package VGAM, Yee 2015) to model the number of errors (count dependent variable that ranges from 1 to 11) by the indices of derivational complexity (four independent non-normally distributed variables). The zero-truncated model, based on a positive Poisson distribution, is better suited for data in which no zeros in the response variable is attested, as in our case.

In order to determine whether we need to apply a one-inflated Poisson regression model due to the excess of ones in our response variable (# errors=1, see Table 3), we fitted a binary logistic regression model (*glm* function in the R package *stats*, R Core Team 2019, Dobson 1990) which estimated the probability of the response falling into one of two groups: texts having one error vs. texts having two and more errors. The same four complexity indices were used in this model as independent variables.

5. Results

We ran Pearson's correlation test to evaluate possible correlation among four levels of the complexity scores, considering values ($0.5 \leq r < 1$), with $p \leq 0.05$ to be

a strong correlation. Only a weak correlation was observed between levels 3 and 6 ($r=0.23$), levels 3 and 4 ($r=-0.09$), and levels 4 and 5 ($r=0.08$) in graph description essays. As for argumentative essays, there was a medium positive correlation between levels 4 and 5 ($r=0.38$, $p<0.05$) and a weak correlation between some other levels ($r=0.24$ for levels 3 and 6, -0.19 for levels 3 and 4, -0.09 for levels 4 and 6, -0.08 for levels 5 and 6).

In what follows, we assess the effect of the examination task using a non-parametric analysis of variance, and argue for the need for a separate analysis of data in two task types. After that, we show the results of a Poisson regression analysis that estimates the effect of complexity on the number of errors in essays of each examination type. In each group, we further split the data into two subgroups by the number of word-formation errors and present the results of non-parametric analysis of variance and regression analysis performed on these subgroups.

5.1. Effect of examination task

The results of comparisons based on Kruskal-Wallis rank sums are given in Table 5. The analysis reveals that at all levels of derivational complexity and with respect to the number of word-formation errors, there is a significant difference ($p<0.05$) between the texts of graph description and argumentative essays. This is in line with the conclusion of (Lyashevskaya et al. *forthc.*) that the texts of the two examination tasks invoke different patterns of complexity and accuracy. In Sections 5.2, 5.3, and 5.4 the analysis is conducted separately for the two task types.

Table 5. *Non-parametric analysis of variance in the groups of graph description and opinion essays*

	<i>H</i> chi-squared	<i>p</i> -value
der_level3	96.732 .	2.2e-16 ***
der_level4	3.9733 .	0.04623 ** .
der_level5	107.96 .	2.2e-16 ***
der_level6	45.258 .	1.727e-11 ***
errors	7.5932 .	0.005859** .

5.2. Derivational complexity and accuracy in graph description

Table 6 shows the estimated coefficients and statistical significance of Poisson's zero-truncated regression model fitted for graph description essays. The number of errors in the model is conditioned by four word-formation complexity metrics. The complexity of suffixes at level 5 (orthographically regular but infrequent affixes) and level 6 (frequent but orthographically irregular models) was found to be significant predictors ($p < 0.05$). The model suggests that the number of errors decreases by 30% with each additional 0.1-point increase in the level 5 complexity, and increases by 29.5% for 0.1-increase in the level 6 complexity.

Table 6. Summary of Poisson's zero-truncated regression model for graph description data

	Estimate	p-value
const	0.2884 .	0.00728 ***
der_level3	-1.1467 .	0.43706 ..
der_level4	-0.5626 .	0.40022 ..
der_level5	-3.4913 .	0.00495 ***
der_level6	2.5884 .	0.00981 ***

5.3. Derivational complexity and accuracy in argumentative essays

The analysis was repeated for the texts of argumentative writing. These data also show a very weak rank correlation between each of the suffix level indices and the number of errors (Kendall's $\tau=0.006$ in the case of level 3, level 4 – 0.02, level 5 – 0.013, and level 6 – 0.029, all $p < 0.001$).

Table 7 reports the output of Poisson's zero-truncated regression model that predicts the count of word-formation errors conditioned by four suffix level complexity measures. Only the suffix complexity at level 6 (frequent but orthographically irregular models) was found to be a significant predictor ($p < 0.05$). The model suggests that with each additional 0.1-point increase in the level 6 complexity, the average number of word-formation errors increases by 30.8% while holding all other variables in the model constant.

Table 7. Summary of Poisson's zero-truncated regression model for opinion essays data

	Estimate	p-value
const	0.1818 .	0.1500 ..
der_level3	-1.3744 .	0.2635 ..
der_level4	0.7608 .	0.4453 ..
der_level5	-0.1944 .	0.8548 ..
der_level6	2.6819 .	0.0145 *

It should be mentioned that no interaction was observed between the complexity measures in the models for both task types, suggesting that these measures are independent and combine additively such that the outcome is better predicted by a simple weighted sum of the indices.

5.4. Texts with one error vs. more than one error

56% of essays have only one word-formation error in our sample. So it is possible that the regression models we presented above underestimate the probability of ones in the response – an effect known as one-inflation (Hassanzadeh & Kazemi 2017).

We ran a rank-sum one-way analysis of variance, dividing the graph description essays into two groups: texts with one error (376 documents) and texts with two and more errors (219 documents). The results suggest that there is no difference between these two groups in regard to their complexity indices, except for level 6, see Table 8. Furthermore, the effect of the suffix complexity was not found in the binary logistic regression models for these groups conditioned by complexity ($p\text{-value}>0.05$ for all four coefficients in various combinations, also with backward elimination of predictors from a full model).

We repeated the same experiments with the argumentative essays (391 documents with one error, 321 documents with more than one error). No effect of complexity was found in both non-parametric analysis of variance and logistic regression (all p -values > 0.05). Therefore we conclude that there is not enough evidence to support the need for selective modelling one-inflation in our datasets.

*Table 8. Non-parametric analysis of variance
in the subgroups with one error vs. more than one error*

graph description		argumentative	
	p -value		p -value
der_level3	0.6891	der_level3	0.9654
der_level4	0.3823	der_level4	0.2786
der_level5	0.1477	der_level5	0.7961
der_level6	0.01621**	der_level6	0.3318

6. Analysis

According to our analysis, the use of level 5 and level 6 word-formation suffixes affects the number of derivational errors in graph description, and the use of level 6 suffixes affects the number of errors in argumentative writing. With the increase in the frequency of level 5 suffixes, the number of errors decreases, and with the increase in the frequency of level 6 suffixes, the number of errors increases.

We have to bear in mind that the expected CEFR level of learner proficiency in the examination is stated as B2, in other words, its range is from low intermediate to high intermediate level. At the intermediate level of English, the learners are expected to have acquired frequent and regular suffixal models such as *-er* in *writer* (level 3) and *-ity* in *sensitivity* (level 4). Regular but infrequent suffixes, such as *-ence* in *emergence* (level 5), had most likely been encountered by students during training and had most likely been practised sufficiently. If so, their performance in using such suffixal constructions might be at the top-right end of the U-shaped curve (Abrahamsson 2013). Much the same can be said of the level 6 suffixes (e. g. *-th* in *length*), with the refinement that irregular word-formation models are likely to be more prone to errors. Admittedly, the lexical unit encoded at level 6 is not only an idiosyncratic, non-prototypical form-function pairing (due to the complexity and irreproducibility of the form), but also belongs to a word family having a non-prototypical and non-transparent structure.

When analysing the complexity and accuracy of the university examination writing in English, several further considerations are to be taken into account. First, equivalents of English words with level 5 and 6 suffixes should have been acquired by undergraduate students in their native language. More often than not, in the case of Russian L2, such words are loanwords and/or the product of word-formation, and one can argue for the existence of near-equivalent suffixes and near-equivalent word-formation models in two languages. If a word has not been acquired and/or sufficiently trained in L2, the learner can still be successful resting on the mechanisms of generalisation, or overgeneralise and thus come to failure.

Second, a word-formation construction can be acquired in terms of morphology but not in terms of syntactic behaviour and co-occurrence. This is the case of partial lexical equivalence – when an existing L2 word-formation construction is inappropriate in a given context, for example, when a learner uses a correctly formed gerund wrongly applying a pattern of the noun to it (the decreasing in the number instead of either decreasing the number or the decrease in the number).

Third, examination writing can be considered as a product of a trade-off between complexity and accuracy according to Skehan's (1998, 2009) Trade-off Hypothesis. In principle, the learner should be interested in maximising the text complexity, including word-formation, but not at the expense of accuracy, and can therefore adopt various strategies to increase the success rate.

Fourth, task complexity and available cognitive resources can either facilitate or inhibit interactions between complexity and the quality of the output according to the later version of the Trade-off Hypothesis and the Cognition Hypothesis by Robinson (2001, 2011), see also overview in (Vasylets et al. 2017). The examination tasks in question differ in many ways and essentially diverge in that, in the case of graph description, the author apparently adheres to specific academic-like stylistically stringent register and can rely on the task prompt as a source of lexical material, whereas in the the case of argumentative writing, the text is expected to be longer, can be less formal and objective, and has to involve argumentation. When sharing his/her opinion, the student has to demonstrate advanced vocabulary knowledge and a rich supply of diverse constructions (Vinogradova et al. 2017). It is usually agreed that the learner might experience greater cognitive load in the latter task, for example, because of the need for adjusting the discourse strategy, perspective taking, choosing appropriate time and space reference, and more complex task planning in general. This can hypothetically result in a beneficial effect of increasing complexity on attention and control processes. At the same time, less advanced students may strive to avoid underdeveloped derivational patterns, but downgrading the complexity does not necessarily interfere with the number of errors.

Level 3 and 4 suffixes are not significant predictors of accuracy in either task type group, which means that lower complexity indices do not account for variance in accuracy in the essays of intermediate learners on their way towards advanced proficiency, even of the learners with a mature vocabulary in their L1. This confirms the intuition that “if we are examining text coverage for high-proficiency learners, Level 6 of Bauer and Nation is likely to be suitable” (Nation 2021).

Qualitative analysis of errors reveals a few noticeable patterns. Expectedly, non-existent derivational constructions are observed in place of the models with irregular suffixes, cf. a wrong choice of the suffix *-ion* in the word **tendention*:

(6) *There we can see an upward **tendention** [tendency] throughout the years.*

Such an error can be explained as L1 interference, cf. Russian *tendencija*.

Nevertheless, the incorrect use of existing words is more common. In most cases, it is accompanied by the incorrect choice of the part of speech:

(7) The **long [length]** of railways a more than 100 kilometres.

(8) Moreover, it has to be noticed, that population of aged people has tendency to **growth [grow]**.

Note that in (8), the error presumably appeared due to interference, since there is an expression *tendentsija k rostu*, lit. ‘tendency to growth’ in Russian.

Only the most frequent level 3 model triggers the errors in word usage, as illustrated by examples (6) and (7).

Examples (9) and (10) show the error in use of the very frequent level 3 model.

(9) But in projection for 2050 in Yemen of population the number of **workable [working]** people will increase and will be 57,3%.

(10) Overall, Instagram is more **usable [used more]** by people of the age of 18–29 (approximately 50%).

Even though the forms *workable* and *usable* are attested in L1 English, they are inappropriate in a given context. Such errors show that learning the syntactic properties of derived forms and understanding the relationship between the functions of the base and derived forms should be part of word-formation instruction.

7. Conclusion

Our study of the association between derivational complexity and the number of errors in English examination writing was motivated by the hypothesis that with increasing complexity, the number of errors will decrease: the more complex suffixes a student uses, the higher his/her level of language proficiency is. To support this hypothesis, we used the classification of derivational suffixes by Bauer and Nation (1993) and a number of statistical methods, such as non-parametric analysis of variance and regression models. Our analysis shows that the two examination tasks applied in the end-of-course examination exhibit partly different patterns.

For shorter and more formal texts, which contain descriptions of the graphical materials, only the use of advanced word-formation structures have a significant effect on the number of errors in word formation. Moreover, the effect is twofold: with an increase in the frequency of level 5 suffixes, the number of errors in word formation decreases, and with the increase in the frequency of level 6 suffixes, it increases. This may indicate that the acquisition of morphology is not linear, but wave-like: the level 5 suffixes are learned well and used confidently, whereas the level 6 suffixes may be familiar to the student (that is, he/she most probably has come across them before), yet they can still be used incorrectly. But one could make an alternative argument: it is the irregularity of word-formation models attested at level 6 that accounts for the decrease in derivational accuracy. The latter approach

indirectly supports both parallel and (semi-)sequential acquisition of the level 5 and 6 suffixes.

As for the texts written in answer to the second examination task, argumentative essays, the word-formation complexity effect narrows down to the suffixes at level 6. We can stipulate that the very format of opinion essays (rather, absence of one specific format) allows authors to choose words more at will and, accordingly, adjust the level of morphological complexity to their level of L2 acquisition in word formation in order to avoid inappropriate usage of words with certain infrequent suffixes.

Once university students have mastered how to deal with the range of word-formation means, their accuracy at this level seems to be getting more dependent on other factors, such as syntactic and discursive complexity of their writing, the range of their lexis, and individual psychological and neurophysiological reaction to the complex cognitive task.

It is important to note that our empirical findings and analysis call for a future research agenda in the area of EFL word formation, such that will examine more direct interactions at particular levels of acquisition, involves analysis of other morphological, syntactic, and discourse structure parameters as well as empirical measurements of extralinguistic factors and, as a result, will develop the methods of exploratory data analysis and computational modelling to reveal distinct learner group profiles and register-sensitive text clusters (Crossley 2020).

Acknowledgements

The research was carried out within the HSE University project ADWISER – Automated detection of writing inaccuracies for students of English in Russia, 2021. The work of Olga Lyashevskaya was partly supported by the Korean National Research Foundation (2021 General Joint Research Support Project TROIKA, 2021–2023).

The authors are grateful to Irina Panteleeva and Anna Scherbakova, who developed Inspector, a tool that has been used to calculate the complexity parameters. We also thank Lilia Rodionova for her assistance in the design and analysis of the study.

REFERENCES

- Abrahamsson, Niclas. 2013. U-shaped learning and overgeneralization. In Peter Robinson (ed.), *The routledge encyclopedia of second language acquisition*, 663–664. London: Routledge. <https://doi.org/10.4324/9780203135945>
- Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 899–919. Berlin, New York: De Gruyter Mouton. <https://doi.org/10.1515/9783110213881.2.899>
- Baerman, Matthew, Dunstan Brown & Greville G. Corbett (eds.). 2015. *Understanding and Measuring Morphological Complexity*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198723769.001.0001>
- Bardovi-Harlig, Kathleen & Theodora Bofman. 1989. Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition* 11(1). 17–34.
- Bauer, Laurie & Paul Nation 1993. Word families. *International Journal of Lexicography* 6(4). 253–279.

- Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.
- Brezina, Vaclav & Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second Language Research* 35(1). 99–119. <https://doi.org/10.1177/0267658316643125>
- Brezina, Vaclav, Pierre Weill-Tessier & Antony McEnery. 2020. #LancsBox v. 5.x. URL: <http://corpora.lancs.ac.uk/lancsbox> (accessed 25.05.2021).
- Brown, Dale, Tim Stoeckel, Stuart Mclean & Jeff Stewart. 2020. The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, amaa061. <https://doi.org/10.1093/applin/amaa061>
- Bulté, Bram & Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 21–46. Amsterdam: John Benjamins. <https://doi.org/10.1075/llt.32.02bul>
- Bulté, Bram & Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26. 42–65. <https://doi.org/10.1016/j.jslw.2014.09.005>
- Capel, Annette. 2010. A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal* 1(1). 2–7. <https://doi.org/10.1017/S2041536210000048>
- Crossley, Scott. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research* 11(3). 415–443. <https://doi.org/10.17239/jowr-2020.11.03.01>
- de la Torre García, Nuria, María Cecilia Ainciburu & Kris Buyse. 2021. Morphological complexity and rated writing proficiency: The case of verbal inflectional diversity in L2 Spanish. *ITL – International Journal of Applied Linguistics* 172(2). 290–318. <https://doi.org/10.1075/itl.20009.del>
- Dobson, Annette J. 1990. *An Introduction to Generalized Linear Models*. London: Chapman and Hall.
- Ehret, Katharina & Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research* 35(1). 23–45. <https://doi.org/10.1177/0267658316669559>
- Hassanzadeh, Fatemeh & Iraj Kazemi. 2017. Regression modeling of one-inflated positive count data. *Statistical Papers* 58(3). 791–809. <https://doi.org/10.1007/s00362-015-0726-7>
- Hay, Jennifer & R. Harald Baayen. 2002. Parsing and productivity. In Geert E. Booij & Jaap Van Marle (eds.), *Yearbook of morphology 2001*, 203–235. Dordrecht: Kluwer Academic. https://doi.org/10.1007/978-94-017-3726-5_8.
- Hollander, Myles & Douglas A. Wolfe. 1973. *Nonparametric Statistical Methods*. New York: John Wiley & Sons.
- Horst, Marlise & Laura Collins. 2006. From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review* 63(1). 83–106. <https://doi.org/10.1353/cml.2006.0046>
- Kimppa, Lilli, Yury Shtyrov, Suzanne C.A. Hut, Laura Hedlund, Miika Leminen & Alina Leminen. 2019. Acquisition of L2 morphology by adult language learners. *Cortex* 116. 74–90. <https://doi.org/10.1016/j.cortex.2019.01.012>
- Lahuerta, Ana Cristina. 2018. Study of accuracy and grammatical complexity in EFL writing. *International Journal of English Studies* 18(1). 71–89. <https://doi.org/10.6018/ijes/2018/1/258971>
- Laufer, Batia, Stuart Webb, Su Kyung Kim & Beverley Yohanan. 2021. How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *ITL – International Journal of Applied Linguistics* 172(2). 229–258. <https://doi.org/10.1075/itl.20020.lau>

- Laws, Jacqueline & Chris Ryder. 2014. Getting the measure of derivational morphology in adult speech a corpus analysis using MorphoQuantics. *University of Reading Language Studies Working Papers* 6. 3–17. [http://morphoquantics.co.uk/Resources/Laws%20&%20Ryder%20\(2014\).pdf](http://morphoquantics.co.uk/Resources/Laws%20&%20Ryder%20(2014).pdf) (accessed 25.06.2021)
- Leontjev, Dmitri. 2016. L2 English derivational knowledge: Which affixes are learners more likely to recognise? *Studies in Second Language Learning and Teaching* 6(2). 225–248. <https://doi.org/10.14746/ssl.2016.6.2.3>
- Lyashevskaya, Olga, Irina Pantelev & Olga Vinogradova. 2021. Automated assessment of learner text complexity. *Assessing Writing* 49, article 100529. <https://doi.org/10.1016/j.asw.2021.100529>
- Lyashevskaya, Olga, Olga Vinogradova & Anna Scherbakova. (forthc.) Accuracy, syntactic complexity, and task type at play in examination writing: A corpus-based study. In Agnieszka Leńko-Szymańska & Sandra Götz (eds.), *Complexity, accuracy, and fluency in learner corpus research*.
- Marchand, Hans. 1969. *The Categories and Types of Present-Day English Word-Formation*. 2nd ed. Munich: C. H. Beck.
- Nation, Paul. 2021. Thoughts on word families. *Studies in Second Language Acquisition* 43(5). 969–972. <https://doi.org/10.1017/S027226312100067X>
- Norris, John & Lourdes Ortega. 2009. Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics* 30(4). 555–578. <https://doi.org/10.1093/applin/amp044>
- Plakans, Lia, Atta Gebril & Zeynep Bilki. 2019. Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing* 36(2). 161–179. <https://doi.org/10.1177/0265532216669537>
- Plag, Ingo, Christiane Dalton-Puffer & Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language & Linguistics* 3(2). 209–228.
- R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.r-project.org> (accessed 25.06.2021).
- Robinson, Peter. 2001. Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22(1). 27–57. <https://doi-org.proxylibrary.hse.ru/10.1093/applin/22.1.27>
- Robinson, Peter. 2011. Second language task complexity, the Cognition Hypothesis, language learning, and performance. In Peter Robinson (ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*, 3–39. Amsterdam: John Benjamins. <https://doi.org/10.1075/tblt.2.05ch1>
- Skehan, Peter. 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.
- Skehan, Peter. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30(4). 510–532. <https://doi.org/10.1093/applin/amp047>
- Stein, Gabriele. 2007. *A Dictionary of English Affixes: Their Function and Meaning*. Munich: Lincom Europa.
- Tywoniw, Rurik & Scott Crossley. 2020. Morphological complexity of L2 discourse. In Eric Friginal & Jack A. Hardy (eds.), *The Routledge handbook of corpus approaches to discourse analysis*, 269–297. London: Routledge. <https://doi.org/10.4324/9780429259982-17>
- van der Slik, Frans, Roeland van Hout & Job Schepens. 2019. The role of morphological complexity in predicting the learnability of an additional language: The case of La (additional language) Dutch. *Second Language Research* 35(1). 47–70. <https://doi.org/10.1177/0267658317691322>

- Vasylets, Olena, Roger Gilabert & Rosa M. Manchón. 2017. The effects of mode and task complexity on second language production. *Language Learning* 67(2). 394–430 <https://doi.org/10.1111/lang.1222>
- Vinogradova, Olga, Olga Lyashevskaya & Irina Panteleeva. 2017. Multi-level student essay feedback in a learner corpus. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2017*. 373–387. Moscow.
- Yee, Thomas W. 2015. *Vector Generalized Linear and Additive Models*. Springer. <https://doi.org/10.1007/978-1-4939-2818-7>
- Yoon, Hyung-Jo. 2017. Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66. 130–141. <https://doi.org/10.1016/j.system.2017.03.007>

Supplementary materials

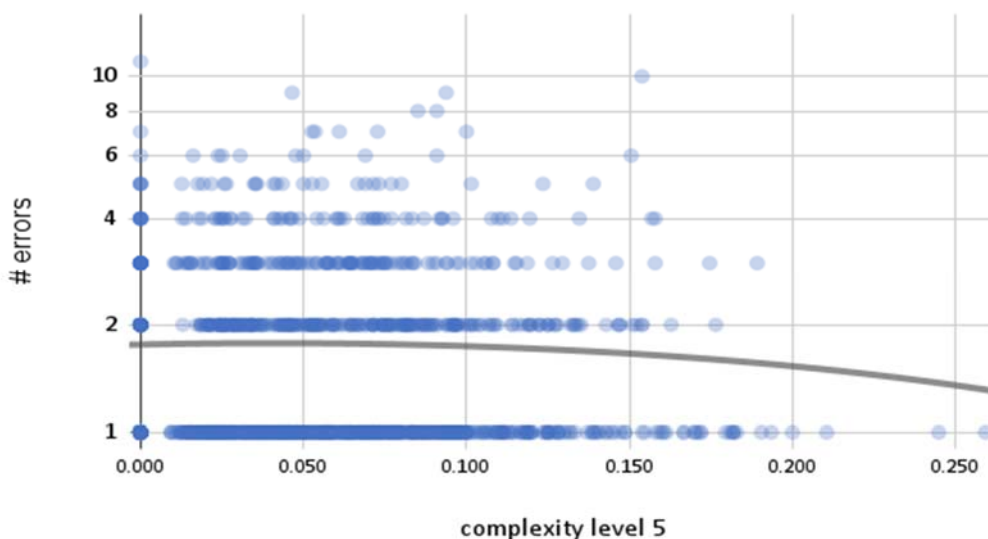


Figure B. Complexity at level 6 and number of errors (both task types)

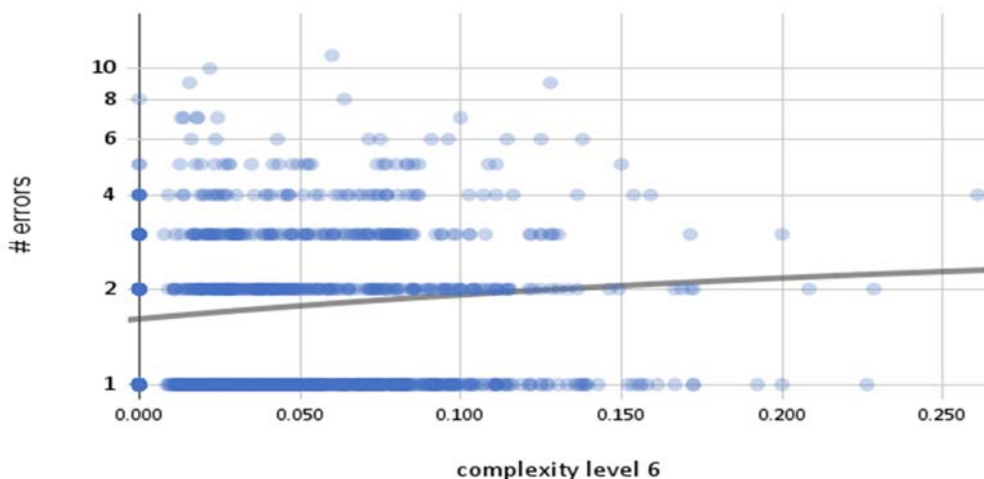


Figure A. Complexity at level 5 and number of errors (both task types)

Article history:

Received: 20 October 2021

Accepted: 08 February 2022

Bionotes:

Olga N. LYASHEVSKAYA is Professor at the School of Linguistics, National Research University “Higher School of Economics”, Moscow, and a Senior Research Fellow at the Vinogradov Russian Language Institute, RAS, Moscow. Her research interests embrace semantics of grammar, lexical semantics, construction grammar, paleo-Slavic grammar, cognitive linguistics, corpus linguistics, lexicography, quantitative data analysis and computational linguistics. She coordinates the natural language processing and database projects of the Russian National Corpus team and PI in a number of lexical resource projects including Russian Framebank and Russian Constructicon.

Contact information:

National Research University “Higher School of Economics”,
room 519, building A, 21/4, Staraya Basmannaya ul., Moscow, Russia

e-mail: olesar@yandex.ru

ORCID: 0000-0001-8374-423X

Julia V. PYZHAK is a student at the Department of Humanities, National Research University “Higher School of Economics”, Moscow. Her research interests include corpus linguistics, machine learning and data analysis, as well as English as a foreign language.

Contact information:

National Research University “Higher School of Economics”,
room 519, building A, 21/4, Staraya Basmannaya ul., Moscow, 117218, Russia

e-mail: jeneavas41@yandex.ru

ORCID: 0000-0003-3439-9788

Olga I. VINOGRADOVA is Associate Professor at the School of Linguistics, National Research University “Higher School of Economics”, Moscow, and a Research Fellow at the Laboratory of Learner Corpora at the Department of Humanities, National Research University “Higher School of Economics”, Moscow. Her areas of research are learner corpus linguistics, computer aided language learning, and corpus-based assessing writing, L2 acquisition and lexical typology. She leads the development of the learner corpus REALEC and a number of tools for automated assessment of learner text complexity and accuracy.

Contact information:

National Research University “Higher School of Economics”.

room 519, building A, 21/4, Staraya Basmannaya ul., Moscow, 117218, Russia

e-mail: olgavinogr@gmail.com

ORCID: 0000-0001-5928-1482

Сведения об авторах:

Ольга Николаевна ЛЯШЕВСКАЯ – профессор Школы лингвистики Национального исследовательского университета «Высшая школа экономики», старший научный сотрудник Института русского языка имени В. В. Виноградова РАН (Москва). Ее исследовательские интересы охватывают широкий круг проблем, включая

грамматическую и лексическую семантику, грамматику конструкций, палеославянскую грамматику, когнитивную лингвистику, корпусную лингвистику, лексикографию, количественный анализ данных и компьютерную лингвистику. Она координирует проекты по разработке Национального корпуса русского языка, русского ФреймБанка, русского Конструктикона и т. д., а также является автором публикаций по когнитивной и функциональной лингвистике, анализу лингвистических данных и языковым технологиям.

Контактная информация:

Национальный исследовательский университет «Высшая школа экономики»
Россия, 117218, Москва, Старая Басманная ул., 21/4, корпус А, комн. 519
e-mail: olesar@yandex.ru
ORCID: 0000-0001-8374-423X

Юлия Вячеславовна ПЫЖАК – студентка факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики», стажер-исследователь научно-учебной лаборатории учебных корпусов факультета гуманитарных наук. Ее исследовательские интересы включают корпусную лингвистику, машинное обучение и анализ данных, английский язык как иностранный.

Контактная информация:

Национальный исследовательский университет «Высшая школа экономики»
Россия, 117218, Москва, Старая Басманная ул., 21/4, корпус А, комн. 519
e-mail: jeneavas41@yandex.ru
ORCID: 0000-0003-3439-9788

Ольга Ильинична ВИНОГРАДОВА – доцент Школы лингвистики Национального исследовательского университета «Высшая школа экономики», научный сотрудник научно-учебной лаборатории учебных корпусов факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (Москва). Область ее научных интересов включает корпусную лингвистику, компьютерные технологии в обучении языку, оценку письменных навыков корпусными методами, преподавание и освоение иностранных языков и лексическую типологию. Она руководит разработкой учебного корпуса REALEC и ряда компьютерных инструментов оценивания сложности и структуры ошибок в текстах, написанных изучающими язык.

Контактная информация:

Национальный исследовательский университет «Высшая школа экономики»
Россия, 117218, Москва, Старая Басманная ул., 21/4, корпус А, комн. 519
e-mail: olgavinogr@gmail.com
ORCID: 0000-0001-5928-1482