





РАЗДЕЛ I. ПРЕДИКТОРЫ СЛОЖНОСТИ ТЕКСТА: МЕТОДЫ ОЦЕНКИ И ПОДХОДЫ
SECTION I. TEXT COMPLEXITY PREDICTORS:
METHODS AND APPROACHES FOR ASSESSMENT

UDC 811.111'37

DOI: 10.18413/2313-8912-2023-9-1-0-2

Galiya M. Gatiyatullina¹ 
Marina I. Solnyshkina² 
Roman V. Kupriyanov³ 
Chulpan R. Ziganshina⁴ 

Lexical density as a complexity predictor:
the case of Science and Social Studies textbooks

¹ Text Analytics Laboratory, Kazan Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
E-mail: ggaliya-m@mail.ru

² Text Analytics Laboratory, Kazan Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
E-mail: mesoln@yandex.ru

³ Text Analytics Laboratory, Kazan Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
Kazan National Research Technological University
68 Karl Marx St., Kazan, 420015, Russia
E-mail: kroman1@mail.ru

⁴ Naberezhnye Chelny Institute of Kazan Federal University
68/19 Mira Ave., Naberezhnye Chelny, 423812, Russia
E-mail: lane0111@mail.ru

Received 08 February 2023; accepted 18 March 2023; published 30 March 2023

Acknowledgements. This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (“PRIORITY-2030”), Strategic Project №5.

Abstract. An ever-increasing need for quality textbooks and objective linguistic expertise encourages more intensive research into complexity of academic discourse. The current research focuses on lexical density viewed as an effective complexity predictor and defined as the ratio of content words per number of words in a text. Being predominantly quantitative, the study also examines dynamics of Flesh-Kincaid grade levels and ratios of parts of speech across 12 Science and Social Studies textbooks taught in Grades 7 – 12 of American schools. The analysis shows a consistent pattern of strong positive growth of nouns and adjectives across grade levels, while lexical verbal elements slightly decrease across the textbooks. The total adverb count changes slightly, and its movement vector depends on the discourse: it rises in Social Studies textbooks and is stable in Science textbooks. This multidirectional movement of components in Lexical density structure explains its





marginal increase across the grades in Science and Social Studies discourse. The findings indicate discourse sophistication increase realized predominantly in text nominalization. We also discuss challenges which nominalization presents for comprehension of academic texts by readers and suggest that provided with reference values of text complexity features, educators receive a reliable tool to select reading texts and assess their suitability for target learner groups. The findings can be beneficial for textbooks authors, exam material developers and discourse researchers.

Keywords: Lexical density; Readability; Text complexity; Textbooks; Science; Social studies

How to cite: Gatiyatullina, G. M., Solnyshkina, M. I., Kupriyanov, R. V. and Ziganshina, C. R. (2023). Lexical density as a complexity predictor: the case of Science and Social Studies textbooks, *Research Result. Theoretical and Applied Linguistics*, 9 (1), 11-26. DOI: 10.18413/2313-8912-2023-9-1-0-2

УДК 811.111'37

DOI: 10.18413/2313-8912-2023-9-1-0-2

Гатиятуллина Г. М.¹ 
Солнышкина М. И.² 
Куприянов Р. В.³ 
Зиганшина Ч. Р.⁴ 

Лексическая плотность как предиктор сложности
(на материале учебников по естествознанию
и обществознанию)

¹ НИЛ «Текстовая аналитика», Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
E-mail: galiya-m@mail.ru

² НИЛ «Текстовая аналитика», Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
E-mail: mesoln@yandex.ru

³ НИЛ «Текстовая аналитика», Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
Казанский национальный исследовательский технологический университет (КНИТУ)
ул. Карла Маркса, 68, Казань, 420015, Россия
E-mail: kroman1@mail.ru

⁴ Набережночелнинский институт Казанского (Приволжского) федерального университета
пр. Мира, 68/19, Набережные Челны, 423812, Россия
E-mail: iana0111@mail.ru

*Статья поступила 08 февраля 2023 г.; принята 18 марта 2023 г.;
опубликована 30 марта 2023 г.*

Информация об источниках финансирования или грантах: Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета («ПРИОРИТЕТ-2030»), Стратегический проект №5.

Аннотация. Постоянно растущая потребность в качественных учебниках и объективной лингвистической экспертизе побуждает исследователей к более интенсивным исследованиям сложности академического дискурса. Представленное исследование имеет целью изучение лексической плотности, трактуемой авторами как эффективный предиктор сложности текста и рассчитываемый соотношением слов знаменательных частей речи к общему количеству слов в тексте. Исследование также нацелено на изучение динамики и корреляции индекса Флеша-Кинкейда (читабельность) с лексической плотностью в текстах 12 учебников по естественным и общественным наукам, преподаваемым в 7–12 классах американских школ. Анализ подтвердил тенденцию сильного положительного роста существительных и прилагательных и снижение количества лексических глаголов во всех учебниках. Суммарное количество наречий меняется незначительно, а вектор его движения зависит от дискурса: в учебниках по обществознанию оно увеличивается, а в учебниках по естественным наукам остается стабильным. Это разнонаправленное движение компонентов в структуре лексической плотности объясняет ее незначительное увеличение в дискурсе естественных и социальных наук по мере их усложнения от 7 к 12 классу. Полученные данные свидетельствуют о повышении сложности дискурса, реализующемся преимущественно в номинализации текста. В статье также обсуждается проблема значимости номинализации для понимания академических текстов. Предлагаемые читателям референтные значения предикторов сложности текста для изучаемых классов и дискурсов могут явиться надежным инструментом при выборе учебных текстов для целевых групп обучающихся. Результаты исследования также могут быть полезны авторам учебников, разработчикам экзаменационных материалов и исследователям дискурса.

Ключевые слова: Лексическая плотность; Читабельность; Сложность текста; Учебники; Естествознание; Обществознание

Информация для цитирования: Гиниятуллина Г. М., Солнышкина М. И., Куприянов Р. В., Зиганшина Ч. Р. Лексическая плотность как предиктор сложности (на материале учебников по естествознанию и обществознанию) // Научный результат. Вопросы теоретической и прикладной лингвистики. 2023. Т. 9. № 1. С. 11-26. DOI: 10.18413/2313-8912-2023-9-1-0-2

Introduction

The problems of text complexity and text comprehension have been in the focus of numerous researchers for a number of decades. Modern transdisciplinary research into text comprehension views rigorous text leveling based on its complexity as the core of successful reading and learning (Solnyshkina, Harkova, Kazachkova, 2020). Benefits and advantages of leveled reading as a strategy of assigning children to books that match their reading skills have been substantiated by hundreds of studies. Popular text leveling systems, including Lexile, Guided Reading Level, Developmental Reading Assessment,

Accelerated Reader and Scholastic Reading Levels (de-la-Peña, Luque-Rojas, 2021) rely on two main ideas: (1) a text presents challenges related to its informative/ cognitive and linguistic features; (2) a reader employs his reading skills at best when the reading stimuli lie within his zone of proximal development (Vygotsky, 1978). Therefore, reader/text matching algorithms imply assessment of text complexity and readers' fluency, accuracy and comprehension abilities. While the existing approaches to testees' reading abilities estimates are primarily based on cloze or open tests and assess how well a testee comprehends levels

of meaning, idea, composition and language conventionality (Fox, 2009), concepts on text complexity assessment are plentiful and vary depending on the range of the text features considered. Since manual procedures of text complexity features measurement present numerous difficulties, researchers encourage development of machine analysis and NLP tools, which are expected to provide accurate text features estimates and compete with analysis conducted by humans (Khurana, Koli, Khatter, 2023).

Validated text complexity predictors and their significant correlations build a theoretical foundation for designing sophisticated text profilers which are capable of defining vocabulary levels of reading stimuli, producing customizable word frequency lists¹, identifying text complexity and aligning it with a category of readers². The idea behind a text profiler is the dialogic nature of a text since a text is always addressed to a specific reader and as such can and should be aligned to a category of readers, i.e. 'profiled'. Once a text is automatically 'profiled', software users receive access to textual analyses and, which is more important, its alignment to a category of readers. As for categories of readers, they are typically identified and presented either based on the number of years of formal schooling/ grades (readability formulas) or vocabulary acquired (Lexile.com).

The growing number of studies on text complexity published worldwide, has not exhausted the topic even for the English language (Solnyshkina, Solovyev, Gafiyatova, Martynova, 2022). There are still numerous research niches emerging, one of which we outline as the impact of lexical density on text complexity. In this article we explore to what

extent lexical density affects complexity of English academic texts thus testing the well-known belief that lexical density predicts text complexity (Daller, Van Hout and Treffers-Daller, 2003). For this purpose, we outline the following research questions:

- RQ 1: What is the range of lexical density metrics in school Science and Social Studies textbooks across Grades 7-12?
- RQ 2: How do shares of different parts of speech vary in school Science and Social Studies textbooks across Grades 7-12?
- RQ 3: How does lexical density correlate with readability in school Science and Social Studies textbooks across Grades 7-12?

The research is conducted to confirm the *hypothesis* that lexical density dynamics in school Science and Social Studies textbooks slightly differ, and nominalization index is higher in Social Studies textbooks.

Literature review

Text complexity

As a concept and a notion 'text complexity' can be defined as a complex of factors affecting and contributing to text comprehension, in other words, they are a set of features which make understanding of a text challenging for a group of people or a particular person. Quantitative dimensions of text complexity which are typically measured by computer software include multiple features clustered into codependent complexity predictors. (cf. McNamara, Graesser, McCarthy and Cai, 2014). Over decades of intensive studies researchers have proposed numerous text complexity predictors including word and sentence length, word frequency, abstractness, syntactic complexity, (Solovyev, Solnyshkina, McNamara, 2022) lexical diversity or TTR and a number of TTR modifications (Templin, 1957) including Guiraud's Index (Giroud, 1954), Corrected TTR (Carroll, 1964), Index of Herdan (Herdan, 1960, 1964), Maas (Maas, 1972, Tweedie and Baayen, 1998, Treffers-Daller, 2013); U Notational variant of Maas (Dugast, 1978; 1979), D score (Malvern, Richards, 1997) and MLTD original (McCarthy, Jarvis, 2010).

¹ Vocab Profilers, available at: <https://www.lex Tutor.ca/vp/> (Accessed 20 February 2023).
 VocabKitchen, available at: <https://www.vocabkitchen.com/home> (Accessed 20 February 2023).

² MultilingProfiler, available at: <https://www.multilingprofiler.net/> (Accessed 20 February 2023).

Flesh-Kincaid readability formula

The first text indices to be selected and derived into a formula able to predict text challenges for readers of different educational backgrounds were word and sentence length (Flesch, 1948: 233). The formula based on these features, the Flesch Reading Ease, became the first readability formula designed to select reading material for people of a certain educational status. Later it was redesigned to convert readability indices into school grade levels. At present, there are more than 50 readability formulae developed to predict English text complexity (Crossley et al., 2008). The most common and robust formula to measure text readability for mainstream readers is Flesh-Kincaid grade level score. The input parameters in the formula are word length and sentence length:

Flesch-Kincaid Grade Level (FKGL)=
 $0.39 \times (\text{Total Words} / \text{Total Sentences}) + 11.8 \times (\text{Total Syllables} / \text{Total Words}) - 15.59$
(Kincaid, Fishburne, Rogers, and Chissom, 1975).

Scholars of text complexity argue that there are a variety of factors contributing to its dynamics across grades/readability levels and types of discourse (Halliday, 2008). As for specifics of academic discourse, according to Hyland (2006b: 13-4), its typical features can be summed up by high lexical density among two more features, which are, high nominal style, i.e. nominatization, and impersonal constructions.

Lexical density

Lexical density was initially studied to compare spoken and written language (Ure, 1971), interviews and conversations (Zora and John-Lewis, 1989) and oral exam answers (O'Loughlin, 1995) to show the difference in mode or between spontaneous and planned speech. Later, lexical diversity was used to define complexity levels in IELTS writing test papers (To et al., 2013), newspapers, conversation, academic register (Biber, 2021), English textbooks (To, Fan, Thomas, 2013), high school English textbooks (Putra, Lukmana, 2017), textbooks for junior high schools (Mulyanti and Soeharto, 2019),

eighth-grade California history textbooks (Schleppegrell et al., 2004).

The notion and the term of 'lexical density' was introduced by Jean Ure in 1971 (Ure, 1971) as the ratio of the number of content words per number of running words (tokens). Content carrying words include nouns, verbs, adverbs and adjectives. Prepositions, conjunctions, auxiliary verbs and pronouns are viewed as non-content words.

M. Halliday (1985) advocates lexical density as a text complexity predictor based on the fact that it relates to the text information structure and as such contributes to its complexity. For example, a conversation has lower lexical density compared to the written texts. M. Halliday argues that written language is "more dense" or "semantically loaded", i.e. lexical density shows "how closely packed the information is" (Halliday, 1985: 62, 66) or "information package" (Johansson, 2008). In his seminal work "The language of science" M. Halliday concludes that "higher lexical density results in higher textual complexity" (2004: 83). D. Biber et al. (2021) claim that linguistic features perform textual tasks of the two major types: marking information structure and cohesion specifying that "text information structure refers to the way in which referential information is packaged or presented within clauses, as well as the way in which clauses are packaged or presented within texts" (Biber et al., 2021: 42). S. Eggins (2004: 94-95) argues that conversation is more dynamic with higher distribution of verbs, linked sequences of clauses while written language tends to have higher distribution of abstract ideas linked by verbs of being in condensed sentences. M. Halliday (1985; 1993), D. Biber and B. Gray (2016) distinguish morphological and syntactic differences affecting complexity of spoken and formal discourse arguing that complexity in conversation is clausal, while academic texts complexity is phrasal and is primarily presented by nominal phrases.

As lexical density refers to statistical indices, researchers suggest different ways of

measuring it. J. Ure (1971) calculates lexical density as the ratio of the number of content words per number of running words:

$$LD = \frac{N_{lex}}{N} * 100$$

M. Halliday (1989: 67) suggested measuring lexical density by calculating lexical items per total number of clauses. S. Eggins (2004: 97) estimates lexical density as ratio of content-carrying words per total number of words in the text. D. Biber (2021) also calculates lexical density, i.e. the sum of content words, per text and further normalizes it per thousand words. The notion of content or lexical words shared by many linguists comprises nouns, adjectives, verbs, and adverbs. M. Halliday (1989: 63), O'Loughlin (1995) also proposed to take into account items consisting of more than one word such as phrasal verbs (to catch up on), idioms (to kick the bucket) or contractions (they're, isn't).

To the best of our knowledge a comprehensive study of Flesh Kincaid readability and lexical density of school Science and Social Studies textbooks has not been performed. Fang et al. (2006) examined indices of lexical density in the 3rd, 5th, and 10th grades textbooks of different subjects: it was registered as 51% in the fable studied in the 3rd grade, 53% – in the 5th grade Science texts, and 59% – in the 10th grade History texts. Two studies on lexical density progress were performed on Indonesian English textbooks used in junior high school (7th, 8th, and 9th grades) (Mulyanti and Soeharto, 2020) and senior high school (10th, 11th, and 12th grades) (Putra and Lukmana, 2017). The results indicate increase of lexical density across the grades. To, Fan, Thomas (2013) conducted research on four short extracts from reading passages in four English textbooks for elementary, pre-intermediate, intermediate and upper-intermediate levels to challenge the correlation between lexical

density, readability (Flesch's Reading Ease Scale), and text levels. The findings confirmed that the lexical density level may increase steadily with the text levels and their readability, however there was no strong relation identified between them in the reading comprehension texts. To and Mahboob (2019) explored lexical density in 24 texts reading passages from four English textbooks for elementary, pre-intermediate, intermediate and upper-intermediate levels and found increase of lexical density level from elementary to intermediate level and slight decrease in upper-intermediate level. As for the Russian language, the morphological patterns of academic texts reported in (Gatiyatullina et al., 2020; Solnyshkina et al., 2017) demonstrate a significant nominal over verbal dominance with nouns making 40-46%, verbs – 12-17%, adjectives – 14%, adverbs – 4-5% of tokens in the text (Gatiyatullina et al., Solnyshkina et al., 2017: 398). The research also confirmed a stable growth of nouns and decrease of verbs in Biology and Social science textbooks across grades 5-11.

Material and Methods

Material

The research corpus with the total size of 2.715.682 tokens comprises two subcorpora: Science (Biology) and Social studies (Civics, Government) (see Corpus Material). Each consists of textbooks for grades 7 – 12 of American secondary and high schools and published between 2008 and 2020 (cf. Table 1). As the textbooks under study are predominantly written for schoolchildren of more than one grade and as such are studied for two or three school years we divided all the texts into three levels based on the age of the target audience of readers: Level I – aged 12-14, Level II – aged 15-17, Level III- aged 18-19.

Table 1. Corpus Size
Таблица 1. Размер корпуса

Level #	Textbook	Number of chapters	Number of tokens	Textbook	Number of chapters	Number of tokens	Total tokens
Science							
Level I	Science in Focus 7 th grade	13	106285	Science Green level 7 th grade	24	100631	206916
Level II	Biology Georgia 10 th grade	34	311207	Biology On level 10 th grade	35	173949	485156
Level III	Biology AP Campbell	38	276541	Biology AP	56	554009	830550
						Total	1522622
Social Studies							
Level I	Civics in Practice 7 th grade	23	118459	Civics Today 7 th grade	28	145100	263559
Level II	Government California 10 th grade	26	248907	Government Roots and Reform 10 th grade	22	260814	509721
Level III	Government Enhanced	18	280882	Government AP	20	138898	419780
						Total	1193060

For the purposes of the study, we combined texts of the same level and discourse into 6 groups of readability levels: Science I – III and Social Science I – III. Calculations of nouns, verbs, adjectives and adverbs were made both per sentence or per 1000 words (cf. Table 2).

Method

The algorithm of the analysis included 4 stages.

On Stage 1, with the help of TextInspector (<https://textinspector.com/>), we measured values of the following features in each group of texts: Flesh-Kincaid, verbal elements per sentence, noun elements per sentence, nouns, adjectives, verbs, adverbs, verbs in present tense, verbs in past tense and later normalized each part of speech, i.e. nouns, verbs, adjectives and adverbs, as well as the sum, i.e. lexical density, to 1000 tokens.

Thus, the finalized list of the metrics compared and contrasted included the following: Flesh-Kincaid, Verbal elements per sentence, Noun elements per sentence, Nouns per 1000 words, Adjectives per 1000 words, Verbs per 1000 words, Adverbs per 1000 words, Lexical density per 1000 words, Verbs in present tense per 1000 words, Verbs in past tense per 1000 words (see Table 2). Following the universally accepted classification installed in TextInspector (<https://textinspector.com/>) which we use as a tool to measure text features, we also distinguish between and measured separately grammatical verbs or auxiliaries, and lexical verbs. Phrasal verbs, e.g., account for, were treated as one lexical item, i.e. *account*, and one grammatical item, i.e. ‘for’.

On Stage 2, we pursued an intra-discourse analysis: compared and contrasted

the values across grades separately in Science and in Social Studies subcorpora.

On Stage 3, we contrasted the metrics across discourses.

On Stage 4, we identified the role of each part of speech in lexical density values

across grades and discourses.

Research results

On Stage 1, we measured and normalized to 1000 words those text features which are expected to be confirmed as complexity predictors (cf. Table 2).

Table 2. Linguistic features of texts of three grade levels (I – III)

Таблица 2. Лингвистические параметры текстов трех образовательных уровней (I – III)

Feature	Level I			Level II			Level III		
	Mean SocS (N = 51)	Mean Sci (N = 43)	p-value, Mann-Whitney U	Mean SocS (N = 63)	Mean Sci (N = 97)	p-value, Mann-Whitney U	Mean SocS (N = 56)	Mean Sci (N = 101)	p-value, Mann-Whitney U
1. Flesh-Kincaid	9.79	8.49	< .01*	13.35	8.51	< .01*	13.31	13.04	0.04*
2. Verbal elements/sentence	1.32	0.99	< .01*	1.60	0.81	< .01*	1.66	1.32	< .01*
3. Noun elements/sentence	1.92	1.41	< .01*	2.56	1.19	< .01*	2.69	2.00	< .01*
4. Nouns per 1000 words	319.66	313.64	0.34	325.31	327.31	0.29	331.63	325.86	0.03*
5. Adjectives per 1000 words	81.41	82.88	0.57	90.17	89.63	0.30	92.48	102.07	< .01*
6. Verbs per 1000 words	45.79	45.59	0.99	36.28	44.80	< .01*	37.86	36.77	0.43
7. Adverbs per 1000 words	34.77	32.77	0.06	39.33	36.77	0.01*	34.60	36.52	0.01*
8. Lexical density per 1000 words	481.63	474.89	0.22	491.09	498.52	0.05*	496.57	501.22	0.12
9. Verbs in present tense per 1000 words	59.09	73.47	< .01*	37.84	76.96	< .01*	38.81	62.66	< .01*
10. Verbs in past tense per 1000 words	23.28	13.35	< .01*	31.18	13.19	< .01*	29.53	10.01	< .01*

* $p < .05$ – statistically significant differences

The intra-discourse analysis on Stage 2 revealed that readability indices of Social science textbooks are higher than those of Science (line 1) which means that they are

more difficult to comprehend as their word and/or sentence lengths are longer.

Noun and verbal ratios per sentence (lines 2, 3) which reflect lexical density per

sentence are also higher across the grades in texts of Social science books. We can also see that while indices of nouns and adjectives per 1000 (lines 4, 5) grow slightly across grades, the number of verbs (line 6) declines in texts of both discourses. Metrics of adverbs and lexical density raise marginally in the texts of Social science and Science (lines 7, 8). The lexical density dynamics in Science and Social Studies textbooks slightly differs, and nominalization index is marginally higher in Social Studies textbooks (line 8).

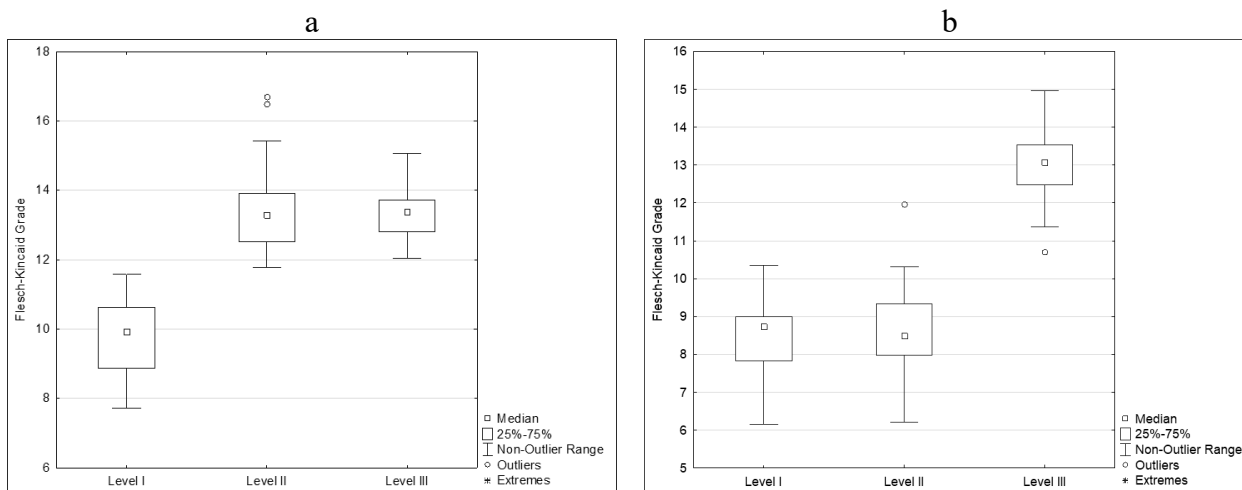
Striking distinctions are observed in the number of verbs in present and past tenses (lines 9, 10): the share of verbs in the present tense is much higher than that of the past which may be viewed as a discourse dissimilarity. The index of “Verbs in past tense per 1000 words” is more than two times higher in the past tense in Social Studies texts than in Science. The opposite trend is observed in the parameter “Verbs in present tense per 1000 words”: verbs in the present tense are approximately 1.5 times less common in Social Studies texts than in

Science. The obvious reason is disciplines specifics. Science texts present and describe animal species, their habitats, as well as the work of organs and physiological systems of the body. All the above are areas of functioning present tense verbs. The past tense is used less frequently than in the social sciences and is largely related to fewer topics: history and development of biology, theory of animal evolution, and description of extinct species. In Social studies, ratio of past tense verbs is higher, because practically every social problem has its own background, distant past and in some cases even antiquity.

Stage 3. The research shows, that on each level, there are statistically significant differences between the linguistic parameters of educational texts in two disciplines. 5 features on Level I have statistically significant differences: Flesh-Kincaid, Verbal elements/sentence, Noun elements/sentence. On Levels II and III, the number of differences between academic texts increases dramatically.

Figure 1. a) Flesh-Kincaid (Social Studies); b) Flesh-Kincaid (Science)

Рисунок 1. а) Читательность по Флешу-Кинкейду (Обществознание); б) Читательность по Флешу-Кинкейду (Естествознание)



As we can be seen from Figure 1, text complexity increases from Level I to Level III. However, the dynamics of text complexity rise in two discipline discourses differs. In

social studies texts, there is a sharp increase in complexity from Level I to Level II. In Science, the complexity increase is observed on the final level.

Figure 2. a) Lexical density per 1000 words (Social Studies); b) Lexical density per 1000 words (Science)

Рисунок 2. а) Лексическая плотность на 1000 слов (Обществознание); б) Лексическая плотность на 1000 слов (Естествознание)

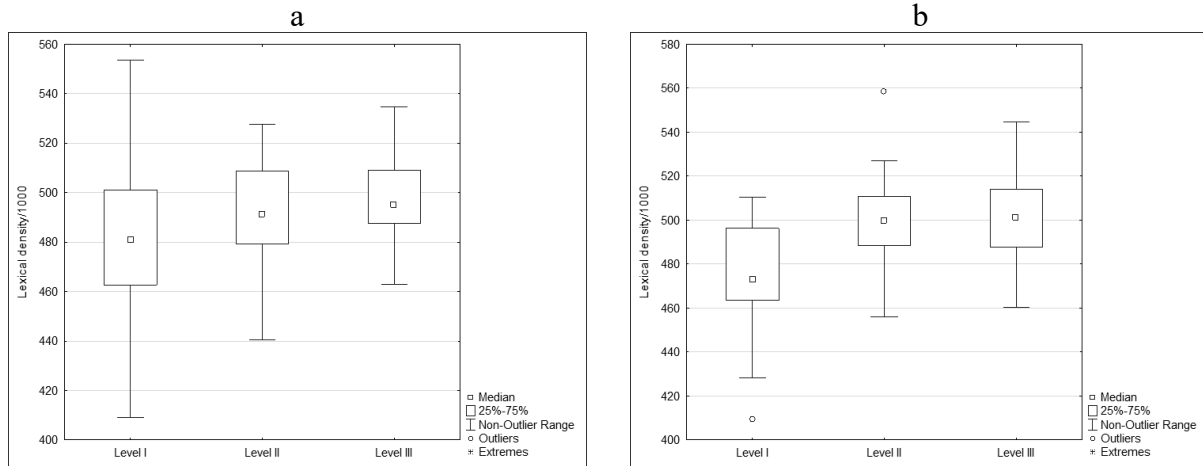


Figure 2 demonstrates that readability growth is accompanied with the lexical density increase: this is the pattern characteristic of both, i.e. Social Studies and Science discourses. However, this increase is far from being significant. For example, Lexical density in Science texts increases from 474.89 (on Level 1) to 501.22 (on Level 3), which is only a 6% increase. In Social Studies texts, these changes are even less visible and amount to 3%. Whereas the parameter ‘Adjectives per 1000 words as a part of Lexical density, increased by 23% in Science texts, and by 14% in Social Studies. In this case, this indicates that with an increase in the texts grade level, they begin to use more adjectives, thereby becoming more

descriptive and allowing a more complete disclosure of a particular concept typically expressed by a noun.

It can also be assumed that a grade level increase is accompanied by a change in the ratio of parts of speech in the text, while the lexical density increases slightly, since its components dynamics are divergent.

On *Stage 4*, we focused on the role of each part of speech in lexical density values across grades and discourses. To identify the relationship between the grade level and the number of parts of speech in the text, we employed Spearman's rank correlation analysis. Table 3 below shows Spearman Rank Order Correlations of the parameters and text grade levels investigated.

Table 3. Correlation Coefficients (Spearman Rank Order Correlations) of Linguistic Parameters of Texts and Level of Texts

Таблица 3. Коэффициенты корреляции (Spearman Rank Order Correlations) лингвистических параметров текстов и уровня текста

I	Feature	Level	
		SocS	Sci
	II	III	IV
1.	Flesch-Kincaid Grade	0,68*	0,79*
2.	Verbal Elements per Sentence	0,56*	0,65*
3.	Noun Elements per Sentence	0,62*	0,66*

	Feature	Level	
		SocS	Sci
4.	Total noun count /1000	0,30*	0,15*
5.	Total adjective count/1000	0,40*	0,55*
6.	Total lexical verb count/1000	-0,48*	-0,44*
7.	Total adverb count/1000	-0,05	0,18*
8.	Lexical density/1000	0,24*	0,34*
9.	Total verbs in present tense count/1000	-0,57*	-0,49*
10.	Total verbs in past tense count/1000	0,28*	-0,15*

The correlations marked * are significant at $p < .05$

As it can be seen in Table 3, most of the text parameters have a statistically significant correlation with the grade level. Texts on Social Studies and Science have differences in correlation coefficients, but the direction and strength of the correlation are identical in most cases. The only exception is Total verbs in past tense /1000, which tends to increase in Social Studies textbooks together with increasing grade levels, while in Science textbooks it is opposite directed. The latter is related to the discipline specifics.

We also revealed that the grade level increase is accompanied with increase in Verbal Elements per Sentence, Noun Elements per Sentence, Total noun count /1000, Total adjective count/1000 and Lexical density/1000. A negative correlation with the grade level is observed with such features as Total lexical verb count/1000 and Total verbs in present tense count/1000.

Thus, we can conclude that Lexical density as a construct contains components with opposite movement vectors: Total noun count and Total adjective count increase, while Total lexical verb count decreases. As for the Total adverb count, it changes very slightly, and its movement vector depends on the discourse: it rises in Social Studies textbooks and is stable in Science textbooks. Such a multidirectional movement of components in Lexical density structure apparently explains its marginal increase across the grades.

Discussion

Our results showed that lexical density is the lowest in the 7th grade (Level I) and the highest in the 12th grade (Level III) in both Science and Social Studies textbooks. The obtained results are consistent with that of earlier research: specifically, D. Biber (2021: 68) showed that lexical density in academic texts is about 500 content words per thousand.

Nouns are the most frequent part of speech across the levels and the studied disciplines. High nominal style was also noted by a number of researchers. Our findings are consistent with D. Biber et al. (1999: 64) who report that nouns being the most frequent word class in academic register have the ratio of about three to four nouns per lexical verb.

Both Science and Social Studies have the lowest distribution of nouns in secondary school level, namely in the 7th grade textbook. Nouns show consistent growth up to level III both in Science and Social Studies books. However, Social Studies textbooks demonstrate a higher distribution of nouns than in Science both in Level II and Level III. Noun frequency growth across the grades suggests higher nominalization which was also identified by a number of scholars in science discourse (Halliday 1993; Halliday, 2004; Eggins, 2004). Being the most common feature of scientific texts, nominalization is the expression of meanings in a form of a noun or noun phrase that might more be expressed in a verb, adjective, or whole clause (Martin, 1991, 1997). "Nominalizations allow us to pack in more lexical content per sentence" (Eggins, 2004: 96). In much

scientific writing, almost all the lexical items in any clause occur inside just one or two nominal groups (noun phrases) (Halliday and Martin, 1993: 76). D. Biber et al. (2011: 10) claim that “alternative grammatical characteristics (associated with complex noun phrases rather than embedded clauses) are much more appropriate measures of grammatical complexity in academic writing”. For this reason, we suppose that complexity of textbooks increases with the growth of nominal phrases across the grades regardless of discipline.

Adjectives are the second most frequent content parts of speech both in Science and Social Studies. Like with that of nouns the frequency of adjectives grows across the levels. However, the growth rate of adjectives is higher in Science than in Social Studies. As such frequency of adjectives in Level III Science textbooks is 10.5, while in Social Studies it is 9.4. The findings are consistent with those of Fang and Cao (2015: 131) where the difference between density of adjectives in natural and social sciences was similar (9.2 in Natural Sciences and 8.1 in Social Sciences). The results suggest that (1) it is common for both science and social studies textbooks to have densely distributed adjectives in phrasal structure of nominalizations which are likely to function as parts of terminological word combinations; (2) Science texts are known to have densely nominalized adjectives as collective nouns. E.g. *the vertebrae, carnivore, Euglenozoans, Carbohydrates, trans fats, Steroids, a membrane potential, enzyme-substrate complex, induced fit, facultative anaerobes*, etc. (Urry et al., 2016).

Frequency of lexical verbs is quite low in both disciplines. In Science its highest distribution is observed on Level I while the lowest is in Level III texts. Unlike Science textbooks, texts in Social studies demonstrate minor fluctuation of verbs on Level II. The increase in the number of nouns and the decrease in distribution of lexical verbs may suggest the tendency to de-verbalization or higher nominalization as mentioned in

D. Biber et al. (2013; 2021b). However, the metrics of verbs per sentence have a strong correlation with Flesh Kincaid grade level (0.98 in Level I Science textbooks and 1.3 in Level III textbooks). This suggest that (1) textbooks syntax complicates as a sentence acquires more clauses and that (2) Level III textbooks tend to use verbal forms rather than lexical verbs, which is in line with D. Biber et al. (2013; 2021b). Adverbs are the least frequent content part of speech both in Science and Social Studies. 7th grade textbooks in Science have the lowest distribution of adverbs across the studied disciplines and levels. The distribution of adverbs is stable and is 3.9 in textbooks in Science both in Level II and Level III.

Conclusion

Text complexity defined as a complex phenomenon affected by numerous text features still attracts a lot of research aimed at identifying the best ways to assess it and align texts and readers. One of the areas of text complexity studies application are text leveling systems developed to mitigate challenges of complex cognitive and linguistic content. Modern text leveling systems are viewed as instruments of prime importance for all types of readers including mainstream and especially readers with speech impairments.

The current study showed a high lexical density of both science and social studies textbooks. Lexical density increases on the account of the growth of nouns and adjectives which is a peculiar feature of academic register. Nominalization, as the process of converting verbs and adjectives into nouns observed in academic texts, creates additional difficulties for understanding because phrases that have undergone the nominalization process lose some of their original semantics. The latter leads to higher ambiguity, difficulty in decoding the text message and mental reconstructing its structure. In addition to highlighting how various text features affect text complexity, our findings specifically support the idea of science and social studies discourses

differences exemplified mostly by ratios of lexical verbs and specifically by verbs in the present and past tenses.

The range of lexical density metrics in school Science and Social Studies textbooks across Grades 7-12 varies between 482 – 496 in Social Studies and 474 – 501 in Science discourse. While the number of nouns and adjectives increase over grades, the number of lexical verbs decrease in both discourses. The share of lexical verbs in the present tense is much higher than that of the past which may be viewed as a discourse dissimilarity. The research confirmed the *hypothesis* that lexical density dynamics in school Science and Social Studies textbooks slightly differ, and based on the metrics of nouns, adjectives and lexical verbs per 1000 words we may argue that degree of nominalization is higher in Social Studies textbooks.

Corpus Materials

Ford, L. E., Bardes, B. A., Schmidt, S. W. and Shelley, M. C. (2020). *American Government and Politics Today*, Enhanced Eighteenth Edition. (In English)

Fisher, A. G. (2008). *Focus on Life Science*, Grade 7, California Edition. (In English)

Massing, G. I. (2009). *Civics in Practice: Principles of Government and Economics*, Holt McDougal, Austin, TX. (In English)

McGraw, H. (2008). *Glencoe Integrated Science*, Level Green, Grade 7, Student Edition. (In English)

Miller, K. and Levine, J. (2010). *Biology*, On-Level Student Edition, Prentice Hall. (In English)

Nowicki, S. (2008). *Biology*, Student Edition McDougal Littell, Georgia. (In English)

O'Connor, K., Sabato, L. and Yanus, A. (2011). *American Government: Roots and Reform*, Pearson, 11th Edition. (In English)

Remy, R. C., Patrick, J. J., Shaffel, D. C. and Clayton, G. E. (2010). *Civics Today: Citizenship, Economics, and You*, Glencoe/McGraw-Hill, Columbus, OH. (In English)

Sidlow, E., Henschen, B., Gerston, L. and Christensen, T. (2011). *Govt.*, Student edition, Wadsworth, Cengage Learning, California. (In English)

Urry, L., Cain, M., Wasserman, S., Minorsky, P. and Reece, J. (2016). *Campbell Biology*, (Campbell Biology Series), 11th Edition. (In English)

Wolfford, D. (2020). *Advanced Placement: United States Government and Politics*, 3rd Edition, Perfection Learning. (In English)

Zedalis, J. and Eggebrecht, J. (2018). *Biology for AP® Courses*, OpenStax College, Rice University, Houston, TX. (In English)

References

Biber, D. and Gray, B. (2013). Nominalising the verb phrase in academic scientific writing, in Aarts, B., Close, J., Leech, G. and Wallis, S. (eds.), *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, Cambridge University Press, Cambridge, 99-132. <https://doi.org/10.1017/CBO9781139060998.006> (In English)

Biber, D. and Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*, Cambridge University Press, Cambridge, UK. <https://doi.org/10.1017/CBO9780511920776> (In English)

Biber, D., Gray, B. and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45 (1), 5-35. (In English)

Biber, D., Johansson S., Leech, G., Conrad, S. and Finegan, E. (2021). *Grammar of spoken and written English*, John Benjamins, Amsterdam. <https://doi.org/10.1075/z232> (in English)

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *The Longman grammar of spoken and written English*, Longman, London, England. (In English)

Carroll, J. B. (1964). *Language and Thought*, Prentice Hall, Englewood Cliffs, N. J., 118. (In English)

Crossley, S. A., Cobb, T. and McNamara, D. S. (2008). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications, *System*, Elsevier, 41 (4), 965-981.

<http://dx.doi.org/10.1016/j.system.2013.08.002> (In English)

- Daller, H., Van Hout, R. and Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals, *Applied Linguistics*, 24 (2), 197-222. <http://dx.doi.org/10.1093/applin/24.2.197> (In English)
- de-la-Peña, C. and Luque-Rojas, M. J. (2021) Levels of Reading Comprehension in Higher Education: Systematic Review and Meta-Analysis, *Frontiers in Psychology*, 12. <http://doi.org/10.3389/fpsyg.2021.712901> (In English)
- Dugast, D. (1978). Sur quoi se fonde la notion d'étendue théorique du vocabulaire? *Le français moderne*, 46, 25-32. (In French)
- Dugast, D. (1979). Vocabulaire et Stylistique. I - Théâtre et Dialogue. *Travaux de Linguistique Quantitative*, Slatkine-Champion, Geneva. (In French)
- Eggs, S. (2004). *An introduction to systemic functional linguistics* (2nd ed.), Pinter, London, UK. (In English)
- Fang, A. C. and Cao, J. (2015). *Text Genres and Registers: The Computation of Linguistic Features*, Springer Berlin Heidelberg, Berlin, Heidelberg, Germany. <http://dx.doi.org/10.1007/978-3-662-45100-7> (In English)
- Fang, Z., Schleppegrell, M. J. and Cox, B. E. (2006). Understanding the Language Demands of Schooling: Nouns in Academic Registers, *Journal of Literacy Research*, 38 (3), 247-273. https://doi.org/10.1207/s15548430jlr3803_1 (In English)
- Flesch, R. (1948). A new readability yardstick, *Journal of Applied Psychology*, 32, 221-233. (In English)
- Fox, E. (2009). The Role of Reader Characteristics in Processing and Learning from Informational Text, *Review of Educational Research*, 79 (1), 197-261. <http://www.jstor.org/stable/40071165> (In English)
- Gatiyatullina, G., Solnyshkina, M., Solovyev, V., Danilov, A., Martynova, E. and Yarmakeev, I. (2020). Computing Russian Morphological distribution patterns using RusAC Online Server, *Proceedings of the 13th International Conference on Developments in eSystems Engineering (DeSE)*, Liverpool, UK, 393-398. <https://doi.org/10.1109/DeSE51703.2020.9450753> (In English)
- Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire: Essai de méthodologie*, Presses Universitaires de France, Paris, France. (In French)
- Halliday, M. A. K. (1989). *Spoken and written language*, Oxford University Press, Oxford, UK. (In English)
- Halliday, M. A. K. (2004). *The language of science*, Continuum, New York, USA. (In English)
- Halliday, M. A. K. and Greaves, W. S. (2008). *Intonation in the grammar of English*, Equinox, London, UK. <http://dx.doi.org/10.1017/S136067430999044X> (In English)
- Halliday, M. A. K. and Martin, J. R. (1993). *Writing science: Literacy and discursive power*, University of Pittsburgh Press, Pittsburgh, PA. (In English)
- Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*, Hodder Arnold, London, UK. (In English)
- Herdan, G. (1960). *Type-Token Mathematics: A Textbook of Mathematical Linguistics*, Mouton, The Hague. (In English)
- Herdan, G. (1964). *Quantitative Linguistics*, Butterworth, London, UK. (In English)
- Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective, *Working papers*, Lund University, Department of Linguistics and Phonetics, 53, 61-79. (In English)
- Khurana, D., Koli, A., Khatter, K. and Singh, S. (2022). Natural language processing: state of the art, current trends and challenges, *Multimedia Tools and Applications*, 82, 3713-3744. <https://doi.org/10.1007/s11042-022-13428-4> (In English)
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L. and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count, and Fesch Reading Ease Formula) for Navy enlisted personnel, *Research Branch Report*, 8-75. (In English)
- Maas, H. D. (1972.) Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes, *Zeitschrift für Literaturwissenschaft und Linguistik*, 2 (8), 73-96. (In German)
- Malvern, D. D. and Richards, B. J. (1997). A new measure of lexical diversity, in Ryan, A.

and Wray, A. (eds.), *Evolving models of language*, Multilingual Matters, Clevedon, 58–71. (In English)

Martin, J. R. (1991). Nominalization in science and humanities: Distilling knowledge and scaffolding, *Functional and systemic linguistics: Approaches and Uses*, De Gruyter Mouton, Berlin, New York, 307–338. <https://doi.org/10.1515/9783110883527.307> (In English)

Martin, J. R. (1997). ‘Analysing genre: functional parameters’, in Christie, F. and Martin, J. (eds.), *Genres and Institutions: Social Processes in the Workplace and School*, Cassell, London, 3–39. (In English)

McCarthy, P. M. and Jarvis, S. (2010). MTL, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment, *Behavior Research Methods*, 42 (2), 381–392. <https://doi.org/10.3758/brm.42.2.381> (In English)

McNamara, D. S., Graesser, A. C., McCarthy, P. M. and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*, Cambridge University Press, Cambridge, UK. <https://doi.org/10.1017/CBO9780511894664> (In English)

Mulyanti, W. and Soeharto, P. (2020). Text Complexity in English Textbooks for Junior High School: A Systemic Functional Perspective, *Advances in Social Science, Education and Humanities Research*, 217–222. <https://doi.org/10.2991/assehr.k.200406.044> (In English)

O’Loughlin, K. (1995). Lexical density in Candidate output on two versions of an oral Proficiency Test, *Melbourne Papers in Language Teaching*, 26–48. (In English)

Putra, D. A. and Lukmana, I. (2017). Text complexity in senior high school English textbooks: A systemic functional perspective, *Indonesian Journal of Applied Linguistics*, 7 (2), 436–444. <https://doi.org/10.17509/ijal.v7i2.8352> (In English)

Schleppegrell, M. J., Achugar, M. and Ote’iza, T. (2004). The grammar of history: Enhancing content-based instruction through a functional focus on language, *TESOL Quarterly*, 38 (1), 67–93. (In English)

Solnyshkina, M. I., Harkova, E. V. and Kazachkova, M. B. (2020). The structure of cross-linguistic differences: Meaning and context of ‘readability’ and its Russian equivalent

‘chitabelnost’, *Journal of Language and Education*, 6 (1), 103–119. <http://doi.org/10.17323/jle.2020.7176> (In English)

Solnyshkina, M. I., Solovyev, V. D., Gafiyatova, E. V. and Martynova, E. V. (2022). Slozhnost' teksta kak mezhdisciplinarnaya problema [Text complexity as interdisciplinary problem], *Issues of Cognitive Linguistics*, 1, 18–40. (In Russian)

Solnyshkina, M. I., Zamaletdinov, R. R., Gorodetskaya, L. A. and Gabitov, A. I. (2017). Evaluating text complexity and Flesch-Kincaid grade level, *Journal of Social Studies Education Research*, 8 (3), 238–248. (In English)

Solovyev, V., Solnyshkina, M. and McNamara, D. (2022). Komputernaya lingvistika i diskursivnaya kompleksologiya: paradigm i metody issledovaniy [Computational linguistics and discourse complexology: Paradigms and research methods], *Russian Journal of Linguistics*, 26 (2), 275–316. <http://doi.org/10.22363/2687-0088-31326> (In Russian)

Templin, M. (1957). *Certain Language Skills in Children: Their Development and Inter-Relationships*, University of Minnesota Press, Minneapolis, MN. (In English)

To, V. and Mahboob, A. (2019). Complexity of English textbook language: A systemic functional analysis, *Linguistics and the Human Sciences*, 13 (3), 264–293. <http://dx.doi.org/10.1558/lhs.31905> (In English)

To, V., Fan, S. and Thomas, D. P. (2013). Lexical density and readability: A case study of English textbooks, *The International Journal of Language, Society and Culture*, 37 (7), 61–71. (In English)

Tweedie, F. J. and Baayen, F. J. (1998). How variable may a constant be? Measures of lexical richness in perspective, *Computers and the Humanities*, 32 (5), 323–352. (In English)

Ure, J. (1971). Lexical density and register differentiation, in Perren, J. E. and Trim, J. L. M. (eds.), *Applications of linguistics*, Cambridge University Press, London, 443–452. (In English)

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*, Harvard University Press, Cambridge, MA. (In English)

Zora, S. and John-Lewis, C. (1989). Lexical Density in interview and conversation, *York Papers in Linguistics*, 14, 89–100. (In English)

Все авторы прочитали и одобрили окончательный вариант рукописи.

All authors have read and approved the final manuscript.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Conflicts of interests: the authors have no conflicts of interest to declare.

Galiya M. Gatiyatullina, Research Assistant, Text Analytics Research Laboratory, Senior Lecturer, Department of Theory and Practice of Teaching Foreign Languages, Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia.

Галия Маратовна Гатиятуллина, младший научный сотрудник, НИЛ «Текстовая аналитика», старший преподаватель кафедры теории и практики преподавания иностранных языков, Институт филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет, Казань, Россия.

Marina I. Solnyshkina, Doctor of Philology, Head and Chief Researcher, Text Analytics Research Laboratory, Professor of the Department of Theory and Practice of Teaching Foreign Languages, Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia.

Марина Ивановна Солнышкина, доктор филологических наук, профессор, профессор кафедры теории и практики преподавания иностранных языков, руководитель и главный

научный сотрудник, НИЛ «Текстовая аналитика», Институт филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет, Казань, Россия.

Roman V. Kupriyanov, Candidate of Psychology, Chief Researcher, Text Analytics Research Laboratory, Institute of Philology and Intercultural Communication, Kazan Federal University; Associate Professor, Department of Social Work, Pedagogy and Psychology, Kazan National Research Technological University, Kazan, Russia.

Роман Владимирович Куприянов, кандидат психологических наук, доцент старший научный сотрудник, НИЛ «Текстовая аналитика», Институт филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет; доцент кафедры социальной работы, педагогики и психологии, Казанский национальный исследовательский технологический университет (КНИТУ), Казань, Россия.

Chulpan R. Ziganshina, Candidate of Philology, Associate Professor, Department of Philology, Naberezhnye Chelny Institute of Kazan Federal University, Naberezhnye Chelny, Russia.

Чулпан Рифовна Зиганшина, кандидат филологических наук, доцент кафедры филологии Набережночелнинского института Казанского (Приволжского) Федерального Университета, Набережные Челны, Россия.