УДК 811.161.1
DOI: 10.26907/2782-4756-2023-72-2-33-44

ЛЕКСИЧЕСКИЕ ПРЕДИКТОРЫ СЛОЖНОСТИ УЧЕБНЫХ ТЕКСТОВ ПО РУССКОМУ ЯЗЫКУ КАК ИНОСТРАННОМУ

© Эльзара Гафиятова, Лейсан Галявиева, Марина Солнышкина

LEXICAL PREDICTORS OF TEXT COMPLEXITY: THE CASE OF RUSSIAN AS A FOREIGN LANGUAGE

Elzara Gafiyatova, Leysan Galyavieva, Marina Solnyshkina

The article presents results of a comparative analysis of lexical complexity of educational texts in teaching Russian as a foreign language. The corpus size of the study is about 0.5 million words evenly distributed among six levels of language proficiency (A1-C2, Russian National System of Certification Levels of General Proficiency in Russian as a Foreign Language, further – RNSCL). The analysis algorithm is demonstrated based on B2 level texts, for which we estimated the values of the eight complexity predictors using the automatic analyzers RuLex (rulex.kpfu.ru) and RuLingva (rulingva.kpfu.ru): the number of tokens and types, sentence length, word length, lexical diversity (LD), terminological density, readability (MSIS) and frequency. B2 texts demonstrate significant differences in all the parameters, except for the word length. The validated B2 average word length is 2.26 syllables. The increase of lexical diversity from A1 to C2 is insignificant being within the range of 0.3 - 0.5. The complexity growth in RFL texts is accompanied by an increase of terminological density and the readability index. Since the RFL text is an important source of linguocultural information, the research findings may be useful to researchers, developers of educational resources and test materials, and teachers for text selection processes.

Keywords: academic text, text complexity, complexity predictors, readability, lexical diversity

В статье представлены результаты сравнительного анализа лексической сложности учебных текстов, используемых в практике преподавания русского языка как иностранного. Размер корпуса исследования составляет около 0.5 млн. слов и равномерно распределен между шестью уровнями владения языком (A1-C2, Российская государственная система сертификационных уровней общего владения русским языком как иностранным, далее – РГССУ). Алгоритм анализа был продемонстрирован на материале текстов уровня B2, для которых с помощью автоматических анализаторов RuLex (rulex.kpfu.ru) и RuLingva (rulingva.kpfu.ru) были рассчитаны значения 8 предикторов сложности: количество словоформ и лемм, длина предложения, длина слова, лексическое разнообразие (ЛР), терминологическая плотность, читабельность (MSIS) и частотность. На основании полученных метрик данных параметров было осуществлено сравнение сложности текстов уровня B2, продемонстрировавшее существенные различия всех параметров за исключением длины слова. Типичной для уровня B2 следует признать среднюю длину слова в 2.26 слога. Рост лексического разнообразия от уровня к уровню незначительный и находится в диапазоне 0.3 – 0.5. Рост сложности текстов по РКИ сопровождается увеличением терминологической плотности и ростом индекса читабельности. Поскольку текст по РКИ является важным источником лингвокультурной информации, полученные выводы могут быть полезны исследователям, разработчикам учебных и контрольно-измерительных материалов, а также преподавателям в процессе выбора учебного текста.

Ключевые слова: русский язык как иностранный, лексическое разнообразие, учебные тексты, сложность текста, предикторы сложности, читабельность

Для цитирования: Гафиятова Э. В., Галявиева Л. Ш., Солнышкина М. И. Лексические предикторы сложности учебных текстов по русскому языку как иностранному // Филология и культура. Philology and Culture. 2023. №2 (72). С. 33–44. DOI: 10.26907/2782-4756-2023-72-2-33-44

Введение

Преподавание русского языка как иностранного (далее – РКИ), как и многие сферы общественной жизни, в настоящее время переживает период коренных изменений. В первую очередь это касается не только и не столько целей, задач, методов, которые, безусловно, видоизменяются и постепенно пополняются новыми. Сегодня наи-

большее внимание эксперты уделяют отбору учебных материалов и ресурсов: новое время и новые условия требуют не только новых учебных текстов, но и пересмотра самих принципов и приемов их отбора. Первостепенным становится содержание текстов, их информативность, соответствие как лингводидактическим, так и профессиональным целям обучающихся. Последнее особо справедливо в рамках компетентностного подхода, ядерным компонентом которого является текстоориентированность, предполагающая широкое использование научных, публицистических, художественных и других типов текстов. При этом целый ряд научно-исследовательских вопросов, связанных с параметрами учебных текстов, используемых в практике преподавания русского языка как иностранного, способами их отбора, до настоящего момента остаются малоизученными. В списке нерешенных исследовательских вопросов находятся, в первую очередь, вопросы оценки когнитивной сложности текста и его информативности, а также вопросы выявления лингвистической сложности. Таким образом, актуальность настоящего исследования определена высокой значимостью разработки алгоритма отбора учебного текста заданной сложности и его возможной адаптации для обучающихся с различным уровнем владения языком.

Известно, что восприятие и понимание текста в полной мере зависят от трех взаимосвязанных составляющих: (1) от когнитивной и лингвистической готовности читателя, (2) сложности текста и его соответствия психофизиологическому развитию обучающихся и уровню владения языком, а также (3) прагмалингвистической ситуации самого процесса чтения [1]. Каждая из составляющих требует отдельного рассмотрения как объект научного исследования. В рамках представленной статьи рассмотрена лексическая сложность и предложен авторский алгоритм отбора учебного текста по русскому языку как иностранному.

Современная научная парадигма имеет в своем распоряжении ряд методик оценки сложности текста, валидированных в корпусных и лингвостатистических исследованиях, в том числе для русского языка, и осуществляемых на основе параметризации текста [Там же]. Оценка такого рода традиционно имеет целью прогнозирование читательского адреса. В контексте преподавания РКИ целевая аудитория, то есть «читательский адрес» текста, определяется на основе уровня владения языком РГССУ, то есть уровня развития всех видов речевой деятельности [2].

Одним из наименее изученных вопросов в рамках обозначенной проблемы для текстов по РКИ является вопрос их лексического разнообразия и его влияния на восприятие текста читателем. Каков оптимальный диапазон лексического разнообразия на различных уровнях владения языком? При каких значениях плотность новых слов становится столь высокой или низкой, что читатель утрачивает интерес и мотивацию к чтению? Поставленные исследовательские вопросы решены для ряда европейских языков, а расчет лексического разнообразия является необходимым условием при отборе учебных текстов и осуществлении лингвистической экспертизы учебных изданий во многих странах [3], [4], [5]. Исследования по сложности учебного текста на русском языке успешно осуществляются в ряде отечественных научных школ (см. [1]), однако референтные значения параметра «лексическое разнообразие» для текстов различных типов до сих пор остаются исследовательской нишей. Гипотеза исследования состоит в том, что лексическое разнообразие учебных текстов по РКИ увеличивается по мере усложнения учебного текста. Таким образом, представленное исследование имеет две взаимосвязанные цели: 1) Какова динамика лексического разнообразия в диапазоне текстов по РКИ от А1 до С2? 2) Каковы референтные значения лексического разнообразия текстов по РКИ на каждом из уровней от А1 до С2?

Обзор литературы

В современной научной парадигме сложность объекта трактуется как «количество информации, необходимой для его описания» [6, с. 54], следовательно, сам объект детерминирует как тип, так и способ оценки его сложности. При этом из всего многообразия характеристик объекта наиболее значимым признается (не)сводимость целого к элементам, то есть «идиоматичность» или «эмерджентность» [7]. Если объект признается идиоматичным, то его сложность трактуется как неаддитивная категория. Аддитивная (дескриптивная) сложность, трактуемая как сумма сложностей составляющих, свойственна объектам с простой структурой. Сложность идиоматичных объектов рассчитывается на основе «объема информации, необходимого для разрешения (уменьшения) неопределенности (неясности) в системе» [6, с. 55]. Таким образом, опираясь на данную основополагающую трактовку, можно рассматривать текст как аддитивную категорию только с высокой долей условности. Например, очевидно, что художественный текст не является только суммой входящих в него элементов, а его целое не сводимо к частям. Оценка неаддитивной сложности

текста весьма нетривиальна, она может осуществляться только на основе критериального подхода, то есть на основе оценки понимания текста читателем [8], [9].

Что касается текстов нехудожественных, то их сложность принято трактовать как структурную и рассчитывать на основе оценки количества элементов и многообразия связей между ними [1]. В качестве предикторов сложности текста как категории ученые называют до 200 разноуровневых параметров: морфологических, лексических, синтаксических и дискурсивных [10]. Преимущественное большинство экспертов, работающих в данной области, признают, что при всей значимости других параметров именно сложность используемой в тексте лексики составляет ядро его лингвистической сложности [11]. Оценка последней осуществляется на основе метрик лексических параметров: частотности, разнообразия, абстрактности и некоторых других [12]. К лексическому разнообразию как параметру текста и предиктору сложности – особое внимание ученых, поскольку именно это свойство текста может существенно изменить трудность его восприятия [13]. Определяя лексическое разнообразие (далее – ЛР) как «словарный запас, отражающий способность говорящего находить необходимые единицы лексикона» [14, с.1415], или «как диапазон, вариативность словарного запаса, реализуемого в тексте заданной длины» [15, с. 387], ученые справедливо указывают на его различия у каждой языковой личности. Автор генерируемого текста может сказать только о тех референтах, обозначения и наименования которых есть в его словарном запасе, а описание одного и того же объекта разными авторами всегда отличается выбором слов и их составом. Лексическое разнообразие является одним из критериев оценки письма, а также мастерства художников слова. «Языку писателя», на основании которого делаются выводы не только о самом писателе, но и литературно-художественной традиции и даже исторической эпохе, посвящены сотни исследований [16].

В дискурсивной комплексологии ЛР рассчитывается как отношение количества лемм к количеству словоформ в тексте [17]. Если ни одно из слов в тексте не повторяется, то ЛР = 1, а если лексических повторов в тексте очень много, то есть ЛР ниже 0,2, то текст такого рода, как правило, не только легок для понимания, но и весьма скучен. При многочисленных повторах утрачивается новизна, а сам текст несет минимальную информативность. Текст с низким ЛР способен демотивировать компетентного читателя, поскольку информативность такого текста весьма невысока [3].

Современные исследователи указывают на необходимость рассчитывать ЛР исключительно на текстовых отрывках не более чем в 1000 словоформ, поскольку высокая повторяемость функциональных и служебных слов будет давать более низкие показатели ЛР на текстах, длина которых превышает 1000 словоформ [18], [19]. Онлайн-анализаторы текста нового поколения, RuLingva (rulingva.kpfu.ru) и Rulex (rulex.kpfu.ru), осуществляют данные операции – деление на отрывки по 1000 словоупотреблений и вычисление среднего показателя ЛР – в автоматическом режиме.

Методы и материалы

Материал исследования составил корпус текстов по РКИ, созданный НИЛ «Текстовая аналитика» Казанского (Приволжского) федерального университета объемом 0.5 млн. словоупотреблений (Свидетельство 2020622254). В корпус вошли тексты 168 учебников по русскому языку как иностранному, используемых в практике преподавания в вузах РФ, а также тексты типовых тестов по РКИ. Заявленные авторами учебников уровни сложности текстов РКИ (A1- C2) верифицировались двумя способами: при помощи экспертного мнения, а далее – и при помощи автоматического анализатора текстов Rulex (rulex.kpfu.ru). RuLex – профайлер учебных текстов на русском языке (см. рис.1).

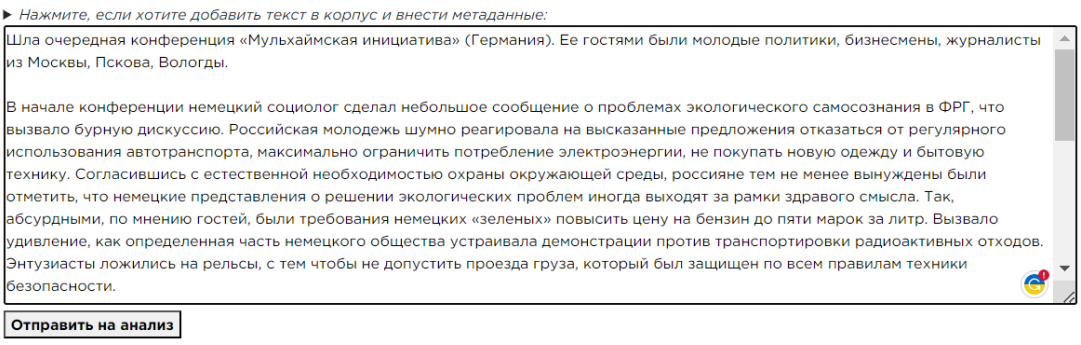

Рис. 1. Интерфейс RuLex

RuLex позволяет загружать текст объемом до 60000 слов и осуществлять автоматический анализ текста по 12 параметрам (см.рис 2), включая количество предложений, словоформ и слов (лемм), количество одно-, дву-, трех- и четырехсложных слов, среднюю длину слов в слогах и буквах, среднюю длину предложений в словах, индексы лексического разнообразия и читабельности (PSIS). Формула читабельности (сложности) Флеша-Кинкейда адаптирована для русского языка и валидирована в ряде отечественных и зарубежных исследований (см. [10], [12], [19], [20]).

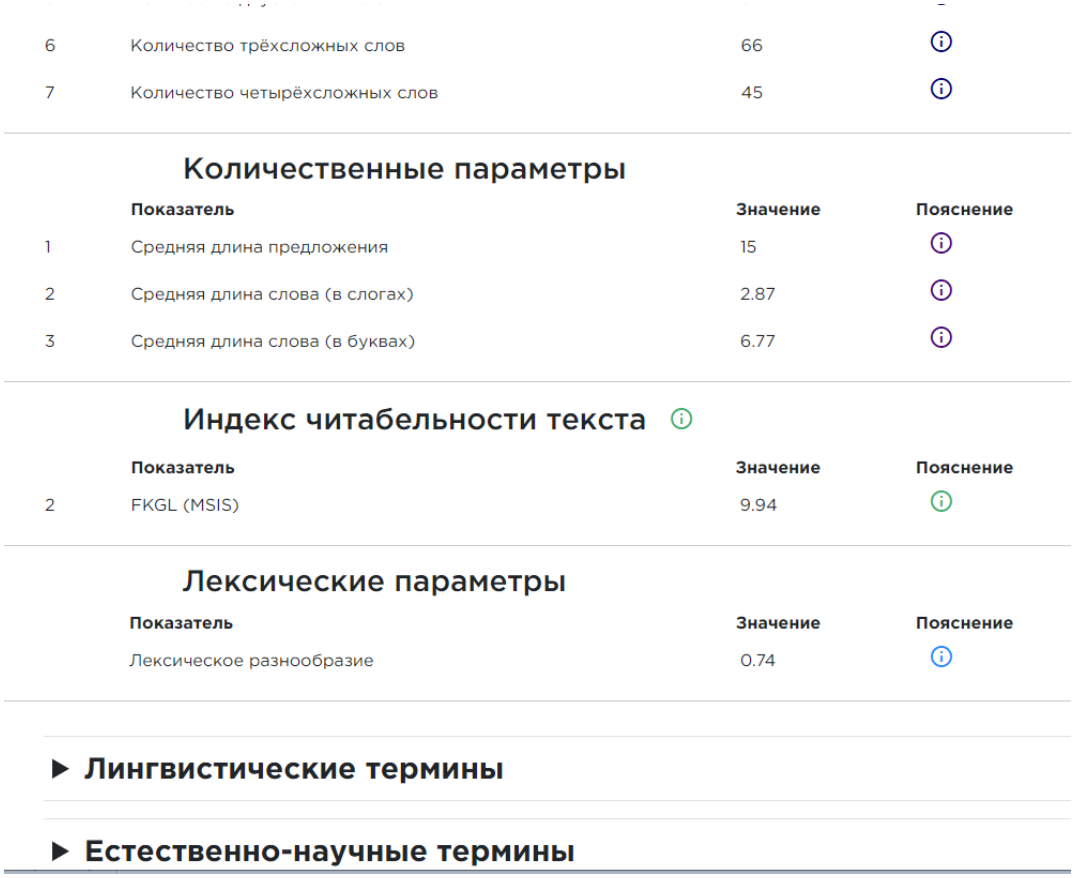

Рис. 2. Список параметров RuLex

Отдельная функция сервера Rulex – извлечение терминов семи предметных областей (см. рис.3): лингвистической, естественно-научной, математической и др. При этом следует отдельно указать на то, что в ситуации отсутствия снятия омонимии [21] одно и то же слово может извлекаться в составе двух и более групп терминов. Таким образом, данная функция RuLex дополнительно демонстрирует внутриязыковую неоднозначность. Например, слово *тема* является термином лингвистики и термином изобразительного искусства (рис.3). Общеупотребительные слова и общенаучные термины также могут быть извлечены в составе терминологических групп отдельных предметных областей: предлог *до* и термин музыки *до*; *система* как общенаучный термин и термин информатики. Такого рода межпредметная омонимия и лексическая полисемия рассматриваются современными учеными как валидированные предикторы когнитивной сложности, поскольку они отражают «неопределенность информации», заключенную в форме языкового знака, имеющего два и более значения [5], [13].

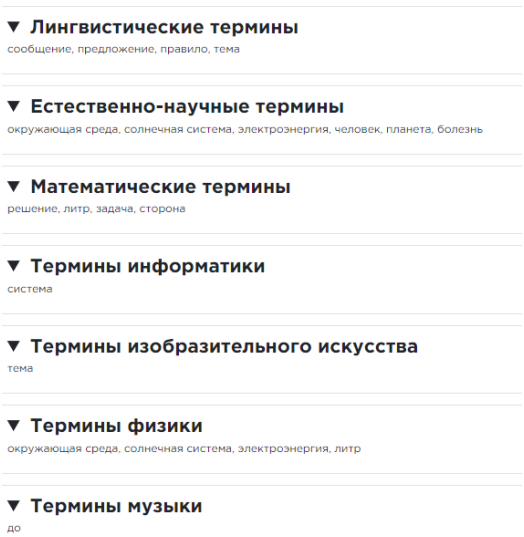

Рис. 3. Термины, извлекаемые Rulex

Немаловажной функцией RuLingva является оценка частотности лексики в диапазоне от первой тысячи наиболее частотных слов русского языка (0К – 1К[1]) до 50000 по словарю О. Н. Лященской и С. И. Шарова [22] и визуализация данных (см. рис.4).

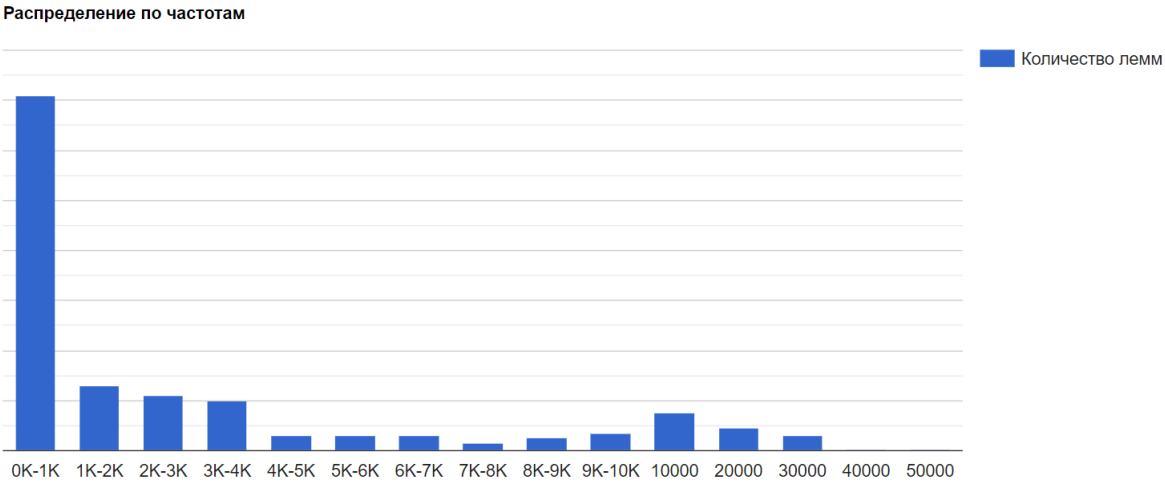

Рис. 4. Частотность лексики, рассчитываемая при помощи Rulingva

Аналогичной является и функция распределения слов по уровням владения A1-C2, предлагаемая современными лингвистическими онлайн-сервисами: для английского языка – TextInspector (textinspector.com), для русского языка: RuLingva (см. рис.5) и Текстометр (textometr.ru).

Уровень	Type	Token
A1	83 (33.6%)	115 (40.35%)
A2	12 (4.86%)	14 (4.91%)
B1	51 (20.65%)	53 (18.6%)
B2	49 (19.84%)	50 (17.54%)
C1	15 (6.07%)	16 (5.61%)
C2	37 (14.98%)	37 (12.98%)

Рис. 5. Доли лексики A1-C2, рассчитываемые при помощи Rulingva

В рамках данного исследования для каждого из заявленных текстов были рассчитаны метрики следующих предикторов сложности: (1) количе-

[1] К маркирует 1000.

ство словоформ, (2) количество лемм, (3) средняя длина слова (в слогах), (4) средняя длина предложения (в словах), (5) индекс лексического разнообразия и (6) индекс читабельности (сложности) Флеша-Кинкейда (см. Табл.1-3), (7) количество терминов (терминологическая плотность текста). Выбор указанных количественных параметров обусловлен тем, что именно они являются базовым набором показателей, для которых рассчитаны и верифицированы референтные индексы, позволяющие осуществлять текстовую аналитику, то есть интерпретировать полученные при анализе текстов числовые показатели [23]. Отдельно следует указать, что индекс читабельности Флеша-Кинкейда традиционно используется для сравнительной оценки сложности различных текстов, включая тексты, генерируемые обучающимися, и тексты, предлагаемые при обучении РКИ. Во всех случаях единой валидированной шкалой при ранжировании текста по категориям коммуникантов является диапазон периода обучения в школе [14], [15], [19].

Результаты

В Таблице 1 предложены полученные в результате автоматического анализа типовых тестов по РКИ Второго сертификационного уровня из пособия Аверьяновой Г. Н. и др. [24] (тексты 1 – 4) и пособия Чепковой Т. П. [25] (тексты 5 – 7). Данные пособия были выбраны случайным образом для демонстрации алгоритма анализа исследования. При этом a priori предполагается, что каждый из текстов, включенных в пособие, независимо от типа формируемых / контролируемых навыков и вида речевой деятельности, должен находиться в диапазоне одного уровня лексической сложности, то есть полностью соответствовать лексическому минимуму B2.

Таблица 1

Предикторы лексической сложности текстов РКИ

№ текста	Уровень CEFR	Предикторы сложности					
		Кол-во словоформ	Кол-во лемм	Лексическое разнообразие	Ср. длина слова	Ср. длина предложения	Читабельность
Т.1 [24, с.12–13]	B2	374	214	0,57	2,44	13,32	6,89
Т. 2 [24, с.14–15]	B2	285	210	0,74	2,87	15	9,94
Т. 3 [24, с.16–18]	B2	559	340	0,61	2,2	11,41	4,79
Т. 4 [24, с.90]	B2	244	135	0,55	2,23	10,17	4,51
Т. 5 [24, с.22–23]	B2	461	230	0,5	2,22	7,33	3,77
Т. 6 [25, с.27–31]	B2	1196	471	0, 39	2,22	12,08	5,19
Т. 7 [25, с.65–67]	B2	648	291	0,45	2,17	12,96	5,22

Анализ показывает (см. табл.1), что изученные тексты демонстрируют широкий диапазон метрик каждого из параметров за исключением длины слова, которая находится в выбранных для демонстрации методики исследования текстах в пределах 2.17 до 2.87 слогов. Средние показатели *индекса лексического разнообразия* варьируются от 0,39 до 0.74, что свидетельствует о значительных различиях в уровнях когнитивной сложности, задаваемых данными текстами. Очевидными причинами столь разительных отличий индексов ЛР могут быть, например, различия в жанровой и стилистической характеристиках текстов: Текст 1 – фрагмент автобиографической повести В. Шаламова «Четвертая Вологда», Текст 2 – журнальная статья, а Текст 4 – предлагаемый в качестве образца пересказ сценария кинофильма «Ирония судьбы». Средние показатели *индекса читабельности,* как видно из табл.1, находятся в пределах от 3,77 (текст 5, индекс читабельности 3,77, то есть предназначен для носителей языка, имеющих начальное (4 класса) образование) до 9,94 (текст 2, индекс читабельности 9,94, то есть предназначен для носителей языка, имеющих незаконченное среднее образование (9 классов). Учитывая, что тексты 1–4 помещены в пособие типовых тестов второго сертификационного уровня (B2, РГССУ), естественно ожидать, что тексты будут демонстрировать приблизительно одинаковую читабельность и лексическое разнообразие. Однако расчеты по-

казывают, что различия между, например, текстами 2 и 4 весьма многочисленны (см. табл. 1). Обращение к самим текстам подтверждает количественные и качественные различия двух текстов даже на примере их фрагментов: текст 2, как указывалось, – фрагмент публицистического текста [24, с.14–15], текст 4 – пересказ сценария кинофильма «Ирония судьбы» [24, с. 90], который содержит значительно более простой синтаксис и преимущественно общеупотребительную лексику.

Сравнение долей лексики A1-C2, рассчитываемых при помощи Rulingva, подтверждает предположение о значительных различиях в их уровнях сложности.

Таблица 2

Доли лексики A1-C2 в тексте 2 и тексте 4

	Текст 2		Текст 4	
Уровень	Леммы	Словоформы	Леммы	Словоформы
A1	83 (33.6%)	115 (40.35%)	81 (48.5%)	141 (57.79%)
A2	12 (4.86%)	14 (4.91%)	16 (9.58%)	22 (9.02%)
B1	51	53	31	36
	(20.65%)	(18.6%)	(18.56%)	(14.75%)
B2	49 (19.84%)	50 (17.54%)	16 (9.58%)	16 (6.56%)
	15 (6.07%)	16 (5.61%)	8 (4.79%)	9 (3.69%)
C2	37 (14.98%)	37 (12.98%)	15 (8.98%)	20 (8.2%)

Доля лексики уровня A1 в тексте 4 значительно больше, чем в тексте 2, (48.5% vs 33.6%), а доля лексики уровня B2 (заявленный авторами пособия уровень) – значительно меньше: 9.58% vs 19.84%. Обе метрики также являются предикторами сложности и дифференцируют тексты по составу лексики различных уровней (A1 – C2) в каждом из них.

Сравнение объемов разночастотных групп лексики так же однозначно показывает различия каждого из текстов: доля высокочастотной лексики (1 – 4К) отличается незначительно, но в тексте 2 присутствует низкочастотная лексика (7 – 8К), которая отсутствует в тексте 4, и доля лексики уровня 10000 выше в тексте 2, что также свидетельствует о его более высокой сложности (см. рис. 6).

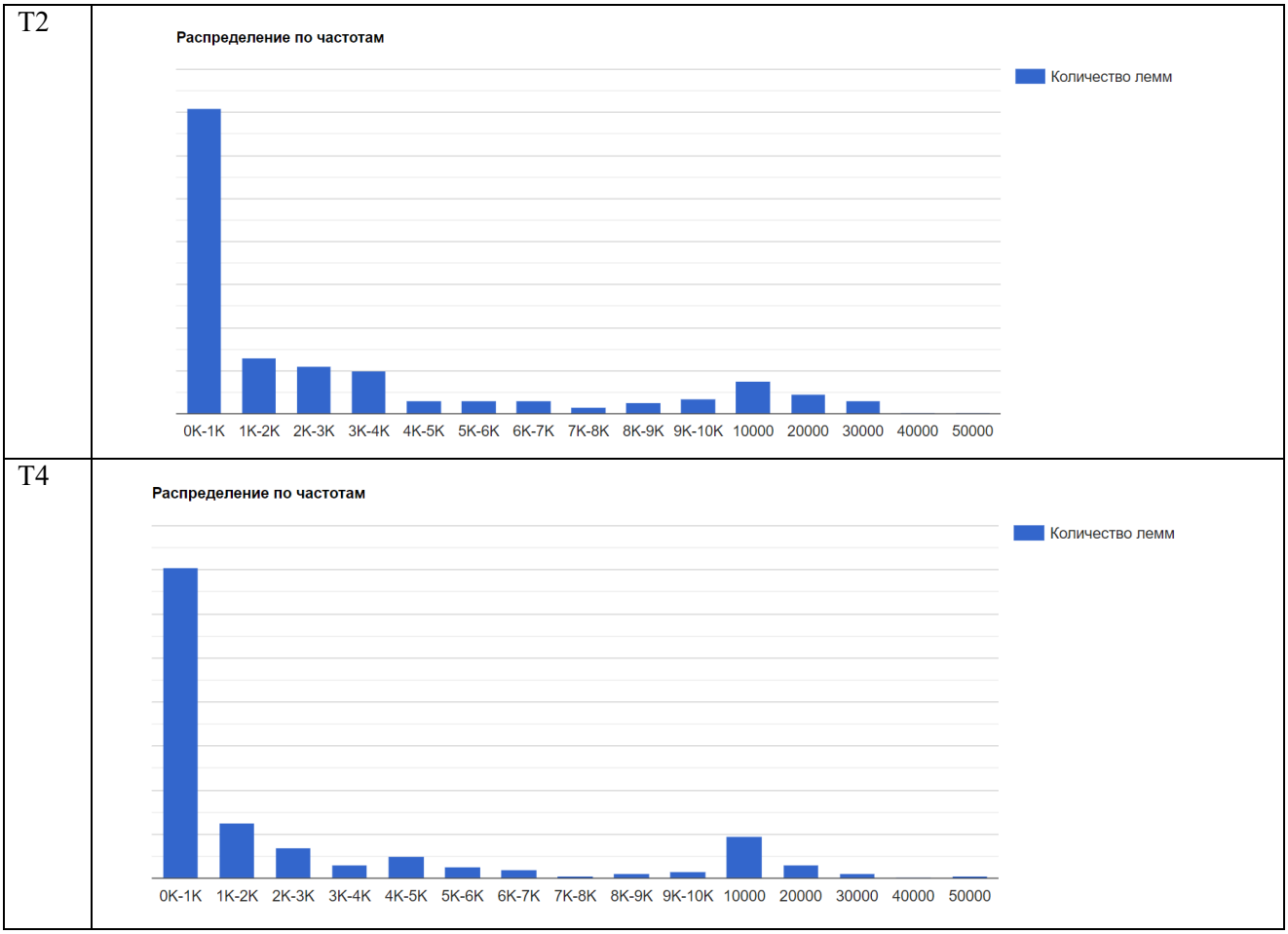

Рис.6. Сравнение частот лексики текстов 2 и 4 (по Ляшевской и Шарову)

Что касается терминологической плотности изучаемых текстов, то в тексте 2 было выявлено 20 терминов, среди которых 4 лингвистических термина (*сообщение*, *предложение*, *правило*, *тема*), 6 естественно-научных терминов (*окружающая среда*, *солнечная система*, *электроэнергия*, *человек*, *планета*, *болезнь*), 4 математических термина (*решение*, *литр*, *задача*, *сторона*), один термин информатики (*система*), один термин изобразительного искусства (*тема*), 4 термина физики (*окружающая среда*, *солнечная система*, *электроэнергия*, *литр*) и один термин музыки (*до*). Для текста 4 количество терминов значительно ниже – 7 единиц, из них: лингвистический термин (*слово*), естественно-научные термины (*мужчина*, *рука*, *вода*, *сон*), термин изобразительного искусства (*свет*) и термин музыки (*до*).

Аналогичный алгоритм анализа был применен для всех текстов в корпусе, среднестатистические величины изучаемых 8 параметров всех текстов каждого из уровней представлены в табл. 3.

Таблица 3

Диапазон метрик предикторов сложности текстах A1-C2

Уровень	Кол-во текстов	Кол-во словоформ	Кол-во лемм	Лексическое разнообразие	Ср. длина слова	Ср. длина предложения	Читабельность	Кол-во терминов
A1	234	50-274	25 - 172	0,39 - 0,45	2,1 - 2,31	4,03 - 6,85	2,83 - 3,82	2 - 8
A2	169	327-611	169 - 251	0,41 - 0,5	2,14 - 2,35	9,7 - 14,3	3,4 - 6,7	2 - 12
B1	137	345-508	180 - 249	0,42- 0,5	2,30 - 2,38	13,1 - 13,8	5,8 - 6,72	4 -14
B2	132	212-291	137 - 177	0,64 - 0,65	2,29 - 2,56	12,65 - 19,27	5,75-9,69	6 - 25
C1	135	500-719	284 - 367	0,51 - 0,57	2,21 - 2,7	11,9 - 17,61	5,29-6,83	12 - 37
C2	121	365-823	206 - 382	0,52 - 0,61	2,45 - 2,79	13,04 - 22,25	6,82- 12,12	12 - 42

Как видим, в целом, от уровня A1 до уровня C2 наблюдается незначительный рост длины слова, однако данный показатель не превышает трех слогов. Положительная динамика наблюдается в следующих мерах: длина предложения, читабельность, терминологическая плотность. Длина текста и, соответственно, количество лемм авторы учебных текстов выбирают довольно произвольно: тексты уровня B2 в среднем оказались короче текстов B1, а некоторые из текстов C2 значительно уступают текстам B1 и C1 по длине. Динамика метрик параметра лексическое разнообразие растет незначительно: от 0.39 (нижний показатель уровня A1) до 0.65 (наиболее высокий показатель, выявленный для уровня B2).

Обсуждение

Примечательно, что изученные тексты не демонстрируют стабильного роста рассчитываемых метрик. Особое значение в аспекте представленного исследования имеет индекс лексического разнообразия, применяемый, как указывалось ранее, для оценки повторяемости лексики текста. Ученые утверждают, что текст, имеющий более высокий индекс лексического разнообразия, является более сложным [26], то есть индекс ЛР способен выступать индикатором сложности текста [27]. А. Н. Лапошина, М. Ю. Лебедева представляют референтные диапазоны для коэффициента лексического разнообразия публицистического текста (0,8) и учебного текста уровня B1 – 0,5 [28], однако наше исследование продемонстрировало, что тексты уровня B2 могут иметь и более низкие метрики ЛР, а положительная динамика величин данного индекса от A1 до C2, хотя и наблюдается, но весьма незначительна. Можно было бы ожидать, что тексты уровня A1 имеют заметно более низкий показатель лексического разнообразия, поскольку обучающиеся владеют меньшим словарным запасом, однако, справедливости ради, следует сказать, что диапазон лексического разнообразия, будучи ниже на уровне A1, чем на уровне C2, все-таки остается в пределах 0,3-0,5. Показательно, что данные выводы в целом совпадают с выводами ряда российских и зарубежных ученых [19], [28], [29].

Таким образом, гипотеза исследования оказывается частично подтвержденной: мы вправе

говорить о диапазоне лексического разнообразия учебных текстов по РКИ уровней A1-C2 в пределах 0,3- 0,5. Тексты с более низким ЛР не выявлены, а тексты, ЛР которых превышает индекс 0.5, как правило, являются первичными аутентичными текстами, предназначенными для носителей языка с высоким уровнем когнитивной и лингвистической готовности. Что касается читабельности, то она демонстрирует незначительные флуктуации и рост в диапазоне от A1 к C2. Корреляция между читабельностью (MSIS) и лексическим разнообразием (ЛР) не выявлена: независимо от уровня читабельности (от 2.83 до 12,12) тексты по РКИ обладают относительно средним уровнем лексического разнообразия.

Заключение

Отбор текстов для обучения РКИ на различных уровнях владения языком – краеугольный камень успеха образовательного процесса, поскольку уровень лингвистической сложности учебного материала играет в нем весьма важную роль. Предлагаемый алгоритм отбора текстов на основе 8 параметров – длина текста (в словоформах и леммах), предложения и слова, читабельность, лексическое разнообразие, терминологическая плотность, частотность лексики – суть параметры, способные объективно, на основе верифицированных референтных индексов дифференцировать тексты различных уровней. Актуальными для отбора текстов следует признать все указанные индексы, однако индексы «лексическое разнообразие» и «читабельность», имея установленные диапазоны референтных значений, обладают более высокой демонстрационной силой.

Таким образом, исследование лингвистических характеристик текстов по РКИ позволило выявить некую диффузность значений параметров изучаемых текстов, что может свидетельствовать об отсутствии строгих границ между текстами смежных уровней. В первую очередь это касается текстов, используемых для обучения различным видам речевой деятельности – чтению (Текст 2) и говорению (Текст 4). Следовательно, причиной «наложения» и некоторой диффузности параметров следует, очевидно, признать многожанровость используемых в практике РКИ текстов, отсутствие строгих границ между текстами смежных уровней. В первую очередь это касается текстов, используемых для обучения различным видам речевой деятельности.

Результаты осуществленного исследования предоставляют экспертам и исследователям алгоритм отбора текстов по РКИ для каждого из шести уровней сложности, а данные о качественных и количественных различиях текстов уровней A1-C2 могут быть использованы учеными и практиками при разработке учебных материалов и в лингвистической экспертизе. Нормативы по лексическому разнообразию учебных текстов могут стать основой автоматического определения уровня текста и быть использованы, например, в онлайн-серверах и при отборе текстов с заявленным уровнем для учебных пособий и КИМов.

В качестве дополнительного вывода осуществленного исследования следует указать на подтвержденную валидность базового комплекса из восьми выбранных параметров, достаточных для установления различий в жанровой специфике текста. Для установления референтных значений каждого из параметров для текстов различных жанров, и в этом видится перспектива исследования, планируется провести дополнительный анализ с привлечением разножанровых текстов одного уровня сложности.

Список источников

1. *Солнышкина М. И., Соловьев В. Д., Гафиятова Э. В., Мартынова Е. В.* Сложность текста как междисциплинарная проблема: отечественная и зарубежная парадигмы // Вопросы когнитивной лингвистики. 2022. № 1. С.18– 39.

2. Российская государственная система сертификационных уровней общего владения русским языком как иностранным (ТРКИ) (РГССУ) URL: https://gct.msu.ru/testirovanie-TRKI/ (дата обращения: 27.12.2022)

3. *Duran, P., Malvern, D., Richards, B., Chipere, N.* “Developmental Trends in Lexical Diversity” Applied Linguistics OUP 25/2, 2004. Pp. 220–242.

4. *McCarthy, P. M., & Jarvis, S.* ‘vocd: A theoretical and empirical evaluation’. Language Testing, 24, 2007. Pp. 459–488

5. *McCarthy, P.M.*, & *Jarvis, S.* MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, Behavior Research Methods, 42(2), 2010. Pp. 381–392

6. *Романов В. Н.* Техника анализа сложных систем. СПб: СЗТУ, 2011. 287 с.

7. *Горлушкина Н. Н.* Системный анализ и моделирование информационных процессов и систем. Санкт-Петербург: Университет ИТМО, 2016. 120 с.

8. *Виноградова Е. М.* Пропозициональный анализ художественного текста как основа его интерпретации // Известия Уральского государственного университета. 2006. № 41. С. 145–152.

9. *Безруких М. М., Адамовская О. Н., Иванов В. В., Филиппова Т. А.* Особенности зрительного восприятия и окуломоторной активности у второклассников при чтении текстов различной сложности // Новые исследования. 2017. №, 4 (53). С. 46 – 63.

10. *Ivanov V. V., Solnyshkina M. I., Solovyev V. D.* Efficiency of text readability features in Russian academic texts // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2018. Pp. 267–283.

11. *McCarthy, K. S., McNamara, D. S., Solnyshkina, M. I., Tarasova, F. Kh., Kupriyanov, R. V.* The Russian language test: towards assessing text comprehension // Science Journal of Volgograd State University. Linguistics, 2019. №4. Pp. 231–247

12. *Solnyshkina M., Solovyev V., Ivanov V., Danilov A.* Studying text complexity in Russian academic corpus with multi-level annotation // CEUR Workshop Proceedings. Proceedings of Computational Models in Language and Speech Workshop, co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics, TEL 2018. 2018. Pp. 1–11.

13. *Graesser A. C., McNamara D. S., Louwerse M. M.,* et al. Coh-Metrix: Analysis of text on cohesion and language // Behavior research methods, instruments, & computers. 2004. Vol. 36, Is. 2. Pp. 193–202.

14. *Fergadiotis, G., & Wright, H.* Lexical diversity for adults with and without aphasia across discourse elicitation task. Aphasiology, 2011. Pp. 1414–1430.

15. *Казачкова М. Б., Галимова Х. Н.* Лексическое разнообразие текста как параметр сложности текстов // Вестник Марийского государственного университета. 2021. №3 (43). Т. 15. С. 384–390.

16. *Кравченко А. В.* «Язык писателя» как семиотический конструктор // Актуальные проблемы филологии и педагогической лингвистики. 2014. №16. С. 21–29

17. *Соловьев В. Д., Солнышкина М. И., Макнамара Д. С.* Компьютерная лингвистика и дискурсивная комплексология: парадигмы и методы исследований // Russian Journal of Linguistics. 2022. Т. 26. № 2. С. 275–316.

18. Douglas Biber University Language: A corpus-based study of spoken and written registers // Studies in Corpus Linguistics, 2007. V.23. Pp. 624– 627.

19. *Вахрушева А. Я., Солнышкина М. И., Куприянов Р. В., Гафиятова Э. В., Климагина И. О.* Лингвистическая сложность учебных текстов. // Вопросы журналистики, педагогики, языкознания, БелГУ. 2021. №40 (1). С. 88–99

20. *Мартынова Е. В., Солнышкина М. И., Мерзлякова А. Ф., Гизатулина Д. Ю.* Лексические параметры учебного текста (на материале текстов учебного корпуса русского языка) // Филология и культура. Philology and Culture. 2020. № 3 (61). С. 72–80.

21. *Гомон Д.Н.* Проблема снятия омонимии // Карповские научные чтения: сб. науч. ст. Вып. 5: в 2 ч. Ч. 1. Минск: «Белорусский Дом печати», 2011. С.166–170.

22. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник. 2009. URL: http://dict.ruslang.ru/freq.php (дата обращения: 26.12.2022)

23. *Kupriyanov, R. V., Solnyshkina, M. I., Dascalu, M.* and *Soldatkina, T. A.* Lexical and syntactic features of academic Russian texts: a discriminant analysis, Research Result. Theoretical and Applied Linguistics, 2022, № 8 (4). Pp. 105–122

24. *Аверьянова Г. Н.* Типовые тесты по русскому языку как иностранному. Второй сертификационный уровень. Общее владение / Г. Н. Аверьянова и др. М-СПб: «Златоуст». 1999. 112 с.

25. *Чепкова Т. П.* Русские фразеологизмы. Узнаем и учим: учеб. пособие / Т. П. Чепкова, Ю. Б. Мартыненко, Е. В. Степанян. М: ФЛИНТА. 2013. 107 с.

26. *Richards, B.* Type/Token Ratios: what do they really tell us? Journal of Child Language, 1987, № 14. Pp. 201–209.

27. *To VT, Le T.* Lexical density and readability: a case study of English textbooks // Proceedings of the Australian Systemic Functional Linguistics Association Conference. Melbourne, 2013. Pp. 61–71.

28. *Лапошина А. Н., Лебедева М. Ю.* Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. 2021. Т. 19. № 3. С. 331–345.

29. *Чурунина А. А., Солнышкина М. И. Ярмакеев И. Э.* Лексическое разноообразие как предиктор сложности учебников по русскому языку // Русистика. 2023. № 2. (в печати).

References

1. Solnyshkina, M. I., Solov'ev, V. D., Gafiyatova, E. V., Martynova, E. V. (2022). *Slozhnost' teksta kak mezhdistsiplinarnaya problema: otechestvennaya i zarubezhnaya paradigmy* [Text Complexity as an Interdisciplinary Problem: Domestic and Foreign Paradigms]. Voprosy kognitivnoi lingvistiki. No. 1, pp. 18–39. (In Russian)

2. *The Russian National System of Certification Levels of General Proficiency in Russian as a Foreign Language* (TORFL) (RGSSU) URL: https://gct.msu.ru/testirovanie-TRKI/ (accessed: 12.27. 2022). (In Russian)

3. Duran, P., Malvern, D., Richards, B., Chipere, N. (2004). *"Developmental Trends in Lexical Diversity".* Applied Linguistics OUP 25/2, pp. 220–242. (In English)

4. McCarthy, P. M., & Jarvis, S. (2007). *'vocd: A Theoretical and Empirical Evaluation'*. Language Testing, 24, pp. 459–488 (In English)

5. McCarthy, P. M., & Jarvis, S. (2010). *MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment*. Behavior Research Methods. No. 42(2), pp. 381–392 (In English)

6. Romanov, V. N. (2011). *Tehnika analiza slozhnyh system* [Technique of Analysis of Complex Systems]. 287 p. St. Petersburg, SZTU. (In Russian)

7. Gorlushkina, N. N. (2016). *Sistemnyi analiz i modelirovaniye informatsionnykh protsessov i sistem* [System Analysis and Modeling of Information Processes and Systems]. 120 p. St. Petersburg, ITMO University. (In Russian)

8. Vinogradova, E. M. (2006). *Propozitsional'nyi analiz hudozhestvennogo teksta kak osnova ego interpretatsii* [Propositional Analysis of a Literary Text as the Basis of Its Interpretation]. Izvestiya Ural'skogo

gosudarstvennogo universiteta. No. 41, pp. 145– 152. (In Russian)

9. Bezrukikh, M. M., Adamovskaya, O. N., Ivanov, V. V. & Filippova, T. A. (2017). *Osobennosti zritel'nogo vospriyatiya i okulomotornoi aktivnosti u vtoroklassnikov pri chtenii tekstov razlichnoi slozhnosti.* [Specifics of Visual Perception and Oculomotor Activity in Second-Graders When Reading Texts of Varying Complexity]. Al'manakh "Novye issledovaniya", 4 (53), pp. 46–63. (In Russian)

10. Ivanov, V. V., Solnyshkina, M. I., Solovyev, V. D. (2018). *Efficiency of Text Readability Features in Russian Academic Texts.* Komp'yuternaya Lingvistika i Intellektual'nye Tehnologii. Pp. 267–283. (In English)

11. McCarthy, K. S., McNamara, D. S., Solnyshkina, M. I., Tarasova, F. Kh., Kupriyanov, R. V. (2019). *The Russian Language Test: Towards Assessing Text Comprehension.* Science Journal of Volgograd State University. Linguistics. No. 4, pp. 231–247. (In English)

12. Solnyshkina, M., Solovyev, V., Ivanov, V., Danilov, A. (2018). *Studying Text Complexity in Russian Academic Corpus with Multi-Level Annotation*. CEUR Workshop Proceedings. Proceedings of Computational Models in Language and Speech Workshop, co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics, TEL 2018. Pp. 1–11. (In English)

13. Graesser, A. C., McNamara, D. S., Louwerse, M. M., et al. (2004). *Coh-Metrix: Analysis of Text on Cohesion and Language*. Behavior research methods, instruments, & computers. Vol. 36, Is. 2, pp. 193–202. (In English)

14. Fergadiotis, G., & Wright, H. (2011). *Lexical Diversity for Adults with and without Aphasia across Discourse Elicitation Task.* Pp. 1414–1430. Aphasiology. (In English)

15. Kazachkova, M. B., Galimova, H. N. (2021). *Leksicheskoe raznoobrazie teksta kak parametr slozhnosti tekstov* [Lexical Diversity of the Text as a Parameter of the Complexity of Texts]. Vestnik Marijskogo gosudarstvennogo universiteta. V. 15. No. 3 (43), pp. 384–390. (In Russian)

16. Kravchenko, A. V. (2014). *"Yazyk pisatelya" kak semioticheskii konstruktor* ["The Writer's Language" as a Semiotic Constructor]. Aktual'nye problemy filologii i pedagogicheskoi lingvistiki. No. 16, pp. 21–29. (In Russian)

17. Solov'ev, V. D., Solnyshkina, M. I., Maknamara, D. S. (2022). *Komp'yuternaya Lingvistika i Diskursivnaya Kompleksologiya: Paradigmy i Metody Issledovanii* [Computational Linguistics and Discursive Complexology: Paradigms and Research Methods]. Russian Journal of Linguistics. V. 26. No. 2, pp. 275–316. (In Russian)

18. *Douglas Biber University Language: A Corpus-Based Study of Spoken and Written Registers* (2007). Studies in Corpus Linguistics. V. 23, pp. 624–627. (In English)

19. Vahrusheva, A. Ya., Solnyshkina, M. I., Kupriyanov, R. V., Gafiyatova, E. V., Klimagina, I. O. (2021). *Lingvisticheskaya slozhnost' uchebnyh tekstov* [Linguistic Complexity of Educational Texts]. Voprosy zhurnalistiki, pedagogiki, yazykoznaniya. Belarusian State University. No. 40 (1), pp. 88–99. (In Russian)

20. Martynova, E. V., Solnyshkina, M. I., Merzlyakova, A. F., Gizatulina, D. Yu. (2020). *Leksicheskie parametry uchebnogo teksta (na materiale tekstov uchebnogo korpusa russkogo yazyka)* [Lexical Parameters of the Educational Text (based on the texts of the educational corpus of the Russian language)]. Filologiya i kul'tura. No. 3 (61), pp. 72–80. (In Russian)

21. Gomon, D. N. (2011). *Prblema snyatiya omonimii* [The Problem of Disambiguation]. Karpovskie nauchnye chteniya: Sb.nauchnykh statei. No. 5: v 2-kh chastyakh. Chast' 1. Pp. 166–170. Minsk, "Belarusskii Dom pechati". (In Russian)

22. Lyashevskaya, O. N., Sharov, S. A. (2009). *Chastotnyi slovar' sovremennogo russkogo yazyka (na materialah Nacional'nogo korpusa russkogo yazyka)* [Russian Dictionary of Modern Frequency (based on the materials of the National Corpus of the Russian Language)]. Moscow, Azbukovnik. URL: http://dict.ruslang.ru/freq.php (accessed: 26.12.2022). (In Russian)

23. Kupriyanov, R. V., Solnyshkina, M. I., Dascalu, M. and Soldatkina, T. A. (2022). *Lexical and Syntactic Features of Academic Russian Texts: A Discriminant Analysis, Research Results.* Theoretical and Applied Linguistics. No. 8 (4), pp. 105–122. (In Russian)

24. Aver'yanova, G. N. et al. (1999). *Tipovye testy po russkomu yazyku kak inostrannomu. Vtoroi sertifiktscionnyi uroven'. Obshhee vladenie* [Standard Tests in Russian as a Foreign Language. The Second Certification Level. General Proficiency]. 112 p. Moscow- St. Petersburg. Zlatoust. (In Russian)

25. Chepkova, T. P. et al. (2013). *Russkie frazeologizmy. Uznaem i uchim: ucheb. Posobie* [Russian Phraseological Units. Studying and Learning: A Study Guide]. Moscow, FLINTA 107 p. Moscow, FLINTA. (In Russian)

26. Richards, B. (1987). *Type/Token Ratios: What Do They Really Tell Us?* Journal of Child Language. No 14, pp. 201–209. (In English)

27. To, V., Le, T. (2013). *Lexical Density and Readability: A Case Study of English Textbooks*. Proceedings of the Australian Systemic Functional Linguistics Association Conference. Melbourne. Pp. 61–71. (In English)

28. Laposhina, A. N., Lebedeva, M. Yu. (2021). *Tekstometr: onlain-instrument opredeleniya urovnya slozhnosti teksta po russkomu yazyku kak inostrannomu* [Textometer: An Online Tool for Determining the Level of Complexity of a Text in Russian as a Foreign Language]. Rusistika. V. 19. No. 3, pp. 331–345. (In Russian)

29. Churunina, A. A., Solnyshkina, M. I. Yarmakeev, I. E. (2023). *Leksicheskoe raznooobrazie kak prediktor slozhnosti uchebnikov po russkomu yazyku* [Lexical Diversity as a Predictor of the Complexity of Textbooks on the Russian Language]. Rusistika. No. 2. (In print). (In Russian).

The article was submitted on 21.06.2023
Поступила в редакцию 21.06.2023

Гафиятова Эльзара Василовна,
доктор филологических наук,
доцент,
Казанский федеральный университет,
420008, Россия, Казань,
Кремлевская, 18.
rg-777@yandex.ru

Галявиева Лейсан Шагиахматовна,
кандидат филологических наук,
доцент,
Казанская государственная академия ветеринарной медицины,
420029, Россия, Казань,
Сибирский тракт, 35.
g-leysan@mail.ru

Солнышкина Марина Ивановна,
доктор филологических наук,
профессор,
Казанский федеральный университет,
420008, Россия, Казань,
Кремлевская, 18.
mesoln@yandex.ru

Gafiyatova Elzara Vasilovna,
Doctor of Philology,
Associate Professor,
Kazan Federal University,
18 Kremlyovskaya Str.,
Kazan, 420008, Russian Federation.
rg-777@yandex.ru

Galyavieva Leysan Shagiakhmatovna,
Ph.D. in Philology,
Associate Professor,
Kazan State Academy of Veterinary Medicine,
35 Sibirskii Trakt,
Kazan, 420029, Russian Federation.
g-leysan@mail.ru

Solnyshkina Marina Ivanovna,
Doctor of Philology,
Professor,
Kazan Federal University,
18 Kremlyovskaya Str.,
Kazan, 420008, Russian Federation.
mesoln@yandex.ru