



АКТУАЛЬНЫЕ ПРОБЛЕМЫ ИССЛЕДОВАНИЙ РУССКОГО ЯЗЫКА

KEY ISSUES OF RUSSIAN LANGUAGE RESEARCH

DOI: 10.22363/2618-8163-2024-22-4-518-539

EDN: APRGCY

Научная статья

Предикативная сила лексических параметров: оценка сложности текста в учебниках по русскому языку для 5–7 классов

М.И. Андреева^{1, 2}  , Р.Р. Замалетдинов² , А.С. Борисова³ 

¹Казанский государственный медицинский университет, *Казань, Российская Федерация*

²Казанский (Приволжский) федеральный университет, *Казань, Российская Федерация*

³Российский университет дружбы народов, *Москва, Российская Федерация*

 mariia99andreeva@yandex.ru

Аннотация. Рассмотрена оценка влияния лексических параметров на сложность текста и реализована на материале специализированного лингвистического корпуса, в состав которого вошли тексты 15 действующих учебников по русскому языку для 5–7 классов общим объемом 811911 слов. Исследование нацелено на определение объема и динамики изменения лексического состава учебников для 5–7 классов современных линеек учебников по русскому языку. Алгоритм исследования: (а) выявление объема и состава лексики в текстах учебников; (б) оценка доли лингвистических терминов в их составе; (в) идентификация предикторов сложности, т.е. параметров, демонстрирующих статистически значимую корреляцию с читабельностью. Осуществлению аналитической части исследования предшествовали метаописание корпуса, его токенизация, лемматизация, сегментирование на отрывки приблизительно по 1000 слов. Расчеты параметров текста производились при помощи текстового профайлера RuLingva, оценка силы корреляционных зависимостей выполнялась при помощи программы STATISTIKA. Для достоверности результатов исследования анализ влияния лексиче-



ских параметров на читабельность текста проведен на двух уровнях: учебники (рассчитывались средние показатели по 15 учебникам 5–7 классов), 1000-словные сегменты текста. Выявлен несколько сниженный (в среднем на 1.0–1.5. уровней) от ожидаемого индекс читабельности изученных текстов, который может служить характеристикой текста учебника по русскому языку как жанра и свидетельствует об эклектичности учебного текста, включающего фрагменты научного стиля (правила), художественного (упражнения) и делового (инструкции и задания к упражнениям). Установлено, что доля терминов не превышает 2 % в словарном составе учебников, но их доля в тексте поднимается до 13 %. Доказано, что предиктивной силой роста сложности текста обладают индексы «лексической плотности», связности (локальный и глобальный повторы существительного и аргумента), «дескриптивность» (отношение прилагательных к существительным), «нарративность» (отношение глаголов к существительным), а также доля имен существительных в родительном падеже. Перспектива исследования видится в изучении роли глаголов и местоимений в текстах учебников по русскому языку как предикторов сложности.

Ключевые слова: учебный текст, корпус языка, сложность учебного текста, частотная лексика, лингвистическая терминология, лексическая плотность, связность

Вклад авторов: *Андреева М.И.* — сбор материала, концепция и дизайн исследования, осуществление исследования; *Замалетдинов Р.Р.* — руководство исследованием, анализ полученных данных, написание статьи; *Борисова А.С.* — обработка материала, исследование, верификация данных.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Финансирование. Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ–2030). Работа выполнена в рамках проекта № 050738-0-000 системы грантовой поддержки научных проектов РУДН.

История статьи: поступила в редакцию 05.03.2024; принята к печати 18.06.2024.

Для цитирования: *Андреева М.И., Замалетдинов Р.Р., Борисова А.С.* Предикативная сила лексических параметров: оценка сложности текста в учебниках по русскому языку для 5–7 классов // Русистика. 2024. Т. 22. № 4. С. 518–539. <http://doi.org/10.22363/2618-8163-2024-22-4-518-539>

Введение

Сложность текста как его важнейшая характеристика во многом определяет успешность понимания текста читателем и снижает риск потери интереса к чтению. Одна из наиболее значимых аксиом дискурсивной комплексологии и теории сложности текста включает понимание непрерывности сложности, т.е. постепенного роста когнитивной и лингвистической сложности от начальной к средней, а затем — старшей школе (Goldman, Lee, 2014). Предполагается, что сложность текста учебника, в т.ч. по русскому

языку, должна повышаться в течение учебного года, а учебный материал при этом меняется от простого к сложному. Аналогичным образом ожидается, что сложность линеек учебников должна возрастать от класса к классу.

Сложность учебного текста все еще остается исследовательской нишей, поскольку нерешенными остаются вопросы о характере представления учебного материала и его характеристиках. Исследования сложности текста, начиная с классической формулы читабельности Флеша — Кинкейда, преимущественно сосредоточены на разноуровневых параметрах текста, в большей или меньшей степени влияющих на совокупную, аддитивную его сложность (Dubay, 2004). Исходное предположение, используемое в качестве рабочей гипотезы в многочисленных исследованиях и повсеместно признанное, заключается в том, что аддитивная сложность как количественная мера может быть рассчитана на основе значений (индексов) предикторов сложности. К последним относятся, например, морфологические, лексические, синтаксические и дискурсивные (Solnyshkina et al., 2018). Однако открытым остается вопрос о взаимовлиянии указанных параметров и о неравномерном росте значений каждого из них. Например, использование абстрактных или низкочастотных, т.е. сложных для восприятия и понимания, слов, может быть причиной перехода автора к более простому синтаксису (Vahrusheva et al., 2023). Этим объясняется интерес к проблеме взаимосвязи, например, синтаксической и лексической сложности учебного текста по русскому языку. Снижают ли авторы учебников по русскому языку синтаксическую сложность текста, стремясь избавить учащихся от излишней когнитивной перегрузки, в то время как лексическая сложность растет? Последнее происходит преимущественно либо за счет увеличения лексического разнообразия, либо за счет лингвистической, например, терминологической, плотности.

Индекс лексического разнообразия (type-token ratio, TTR) — одна из основных характеристик языка писателя и индивидуально-авторского своеобразия, рассчитывается как отношение лемм ко всем словоформам текста (Templin, 1957; Graesser et al., 2004: 1). Если индекс TTR равен 1.0, это означает, что все слова в тексте — разные, ни одно слово не повторяется. Очевидно, что такого рода искусственно созданный текст будет сложен для восприятия даже читателям с высоким объемом оперативной памяти. В отношении учебных текстов логично предположить, что их лексическое разнообразие во много зависит от количества используемых автором(-ами) терминов. Известно и другое: научное исследование традиционно имеет более высокий уровень TTR. Но означает ли это, что тексты по русскому языку старших классов имеют более высокие индексы TTR или авторы, в стремле-

нии не перегружать школьника, будут сохранять определенный невысокий уровень лексического разнообразия?

В отличие от лексического разнообразия индекс лексической плотности рассчитывается как отношение числа только знаменательных частей речи к числу словоупотреблений (Ure, 1971). Отражая информационную структуру текста, лексическая плотность текста верифицирована как один из предикторов сложности (Halliday, 1985), в т.ч. учебных текстов по естественнонаучным и гуманитарным дисциплинам (Gatiyatullina et al., 2023).

Что касается читабельности текста, то она трактуется не просто как свойство текстового материала, характеризующее легкость восприятия его человеком (Окладникова, 2010), т.е. во многом индивидуальная характеристика, но и как объективный параметр оценки сложности текста в целом (Solovyev, Dascalu, Solnyshkina, 2023), рассчитываемый при помощи формулы читабельности с учетом средней длины предложения и средней длины слова (Solnyshkina et al., 2018). Поскольку читабельность текста, т.е. зависимость восприятия и понимания текста читателем от указанных параметров, доказана для многих языков мира, то подтверждение предиктивной силы новых параметров сложности осуществляется на основе доказанных корреляций избранного параметра именно с читабельностью текста (Paraschiv, Dascalu, Solnyshkina, 2023).

Принимая во внимание значимость для науки и эффективность анализа взаимозависимости сложности / читабельности с лексическими параметрами учебного текста, наше **исследование нацелено** на изучение объема и динамики изменения лексического состава учебников по русскому языку для 5–7 классов. Задача исследования — определить: (1) степень различия объема и словарного состава и роль лингвистических терминов в составе лексики текстов учебников 5–7 классов в современных линейках учебников по русскому языку; (2) статистически значимые и влияющие на читабельность, т.е. являющиеся или способные являться предикторами сложности, лексические параметры рассматриваемых текстов, включая лексическое разнообразие и лексическую плотность.

Методы и материалы

Исследование основано на материале авторского корпуса учебников по русскому языку, в который вошли 15 учебников 5–7 классов общим объемом 811911 словоформ (табл. 1). Отбор учебников базировался на перечне ФГОС, актуальном на период 2018–2024 гг.

Таблица 1

Размер и состав корпуса учебников 5–7 классов по русскому языку

Учебник	Код учебника	Класс	Количество словоформ
VoMa	05 VoMa ¹	5	28664
	06 VoMa ²	6	28363
	07 VoMa ³	7	21273
RyAl	05 RyAl ⁴	5	63076
	06 RyAl ⁵	6	68564
	07 RyAl ⁶	7	49923
LaBa	05 LaBa ⁷	5	40050
	06 BaLa ⁸	6	50348
	07 BaLa ⁹	7	56003
ByKi	05 ByKi ¹⁰	5	80486
	06 ByKi ¹¹	6	78437
	07 ByGo ¹²	7	63144
ShFI	05 ShFI ¹³	5	106689
	07 ShFI ¹⁴	7	76963
LiPi	06 LiPi ¹⁵	6	50276
	ИТОГО	15	811911

И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

¹ Здесь и далее в работе используется авторская маркировка учебников корпуса буквенно-числовым кодом, соответствующим году обучения и начальным буквам фамилий двух первых авторов. *Воителева Т.М., Марченко О.Н., Смирнова Л.Г., Текучёва И.В.* Русский родной язык: учебник для 5 класса общеобразовательных организаций. 2-е изд. М. : Русское слово — учебник, 2022. 168 с. (ФГОС. Инновационная школа).

² *Воителева Т.М., Марченко О.Н., Смирнова Л.Г., Шагин И.В.* Русский родной язык : учебник для 6 класса общеобразовательных организаций. 2-е изд. М. : Русское слово — учебник, 2022. 160 с.

³ *Воителева Т.М., Марченко О.Н., Смирнова Л.Г., Шагин И.В.* Русский родной язык : учебник для 7 класса общеобразовательных организаций. 2-е изд. М. : Русское слово — учебник, 2022. 112 с. (ФГОС. Инновационная школа).

⁴ *Рыбченкова Л.М., Александрова О.М., Глазков А.В., Лисицын А.Г.* Русский язык. 5 класс : учебник для общеобразоват. учреждений : в 2 частях. 2-е изд., испр. М. : Просвещение, 2012. 127 с.

⁵ *Рыбченкова Л.М., Александрова О.М., Загоровская О.В., Нарушевич А.Г.* Русский язык. 6 класс : учебник для общеобразоват. организаций : в 2 частях. 3-е изд. М. : Просвещение, 2014. 159 с.

⁶ *Рыбченкова Л.М., Александрова О.М., Загоровская О.В., Нарушевич А.Г.* Русский язык. 7 класс : учебник для общеобразоват. организаций. 2-е изд. М. : Просвещение, 2014. 207 с.

⁷ *Ладыженская Т.А., Баранов М.Т., Тростенцова Л.А., Григорян Л.Т., Кулибаба И.И., Ладыженская Н.В.* Русский язык. 5 класс : учебник для общеобразоват. учреждений : в 2 частях / науч. ред. Н.М. Шанский. М. : Просвещение, 2012. Ч. 1 — 192 с. Ч. 2 — 176 с.

⁸ *Баранов М.Т., Ладыженская Т.А., Тростенцова Л.А., Ладыженская Н.В., Григорян Л.Т., Кулибаба И.И.* Русский язык. 6 класс : учебник для общеобразоват. организаций : в 2 частях / науч. ред. Н.М. Шанский. 5-е изд. М. : Просвещение, 2015. Ч. 1 — 191 с. Ч. 2 — 175 с.

⁹ *Баранов М.Т., Ладыженская Т.А., Тростенцова Л.А., Ладыженская Н.В., Григорян Л.Т., Александрова О.М., Дейкина А.Д., Кулибаба И.И.* Русский язык. 7 класс : учебник для общеобразоват. организаций : в 2 частях. 3-е изд. М. : Просвещение, 2021. Ч. 1 — 176 с. Ч. 2 — 149 с.

¹⁰ *Быстрова Е.А., Кибирева Л.В., Гостева Ю.Н., Калмыкова И.Р., Юрьева Е.С.* Русский язык : учебник для 5 класса общеобразовательных учреждений : в 2 частях / под ред. Е.А. Быстровой. 4-е изд. М. : Русское слово — учебник, 2015. Ч. 1 — 280 с. Ч. 2 — 288 с.

¹¹ *Быстрова Е.А., Кибирева Л.В., Гостева Ю.Н., Антонова Е.С.* Русский язык : учебник для 6 класса общеобразовательных учреждений : в 2 частях / под ред. Е.А. Быстровой. 3-е изд. М. : Русское слово — учебник, 2014. Ч. 1 — 160 с. Ч. 2 — 157 с.

¹² *Быстрова Е.А., Гостева Ю.Н., Кибирева Л.В., Воителева Т.М.* Русский язык : учебник для 7 класса общеобразовательных организаций / под ред. Е.А. Быстровой. 5-е изд. М. : Русское слово — учебник, 2017. 304 с. (Инновационная школа).

¹³ *Шмелёв А.Д., Флоренская Э.А., Габович Ф.Е.* Русский язык : учебник для 5 кл. общеобразоват. учреждений : в 2 частях. М. : Вентана-Граф, 2015. Ч. 1 — 297 с. Ч. 2 — 320 с. + вкл. 3,0.

¹⁴ *Шмелёв А.Д., Флоренская Э.А., Савчук Л.О., Шмелёва Е.Я.* Русский язык: 7 класс : учебник для учащихся общеобразоват. организаций / под ред. А.Д. Шмелёва. М. : Вентана-Граф, 2016. 400 с.

¹⁵ *Лидман-Орлова Г.К., Пименова С.Н., Еремеева А.П., Купалова А.Ю. и др.* Русский язык. Практика. 6 кл. : учебник / под ред. Г.К. Лидман-Орловой. 2-е изд., стереотип. М. : Дрофа, 2014. 318 с.

Table 1

Size and contents of Corpus of Russian language textbooks for Grades 5–7

Textbook	Textbook code ¹⁶	Grade	Tokens
VoMa	05_VoMa	5	28664
	06_VoMa	6	28363
	07_VoMa	7	21273
RyAl	05_RyAl	5	63076
	06_RyAl	6	68564
	07_RyAl	7	49923
LaBa	05_LaBa	5	40050
	06_BaLa	6	50348
	07_BaLa	7	56003
ByKi	05_ByKi	5	80486
	06_ByKi	6	78437
	07_ByGo	7	63144
ShFI	05_ShFI	5	106689
	07_ShFI	7	76963
LiPi	06_LiPi	6	50276
	TOTAL	15	811911

S o u r c e : Calculated by the M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

В состав корпуса вошли четыре линейки учебников, каждая из которых маркирована (см. табл. 1) первыми буквами фамилий авторов, а класс — соответствующими цифрами: 1) под редакцией Воителевой Т.М. (2022 г.) — VoMa; 2) под редакцией Рыбченковой Л.М. (5 класс 2012 г., 6, 7 классы 2024 г.) — RyAl; 3) под редакцией Ладыженской Т.А. (5 класс 2012 г., 6 класс 2015 г., 7 класс 2021 г.) — LaBa; 4) под редакцией Быстровой Е.А. (5 класс 2015 г., 6 класс 2014 г., 7 класс 2017 г.) — ByKi. Учебники разных линеек, предназначенные для школьников одного класса, отличаются объемом, однако в целом корпус исследования следует признать сбалансированным как по количеству учебников (5 в каждом из 5–7 классов), так и по объему (около 300000 словоформ в каждом классе). Поскольку изучаемые тексты представляют генеральную совокупность учебников, то и требование дифференциальной полноты корпуса достигнуто (Ure, 1971, Gatiyatullina et al., 2023): в корпусе представлены тексты всех классов, для которых предполагается оценить долю лингвистических терминов в языке и тексте учебника, а также выявить лексическое разнообразие и лексическую плотность.

Алгоритм исследования включал этапы: (1) создание и предобработка корпуса исследования — отдельных учебников и линеек учебников по русскому языку 5–7 классов; (2) создание списков словарного состава каждого учебника; (3) извлечение лингвистических терминов и выявление их доли в словарном составе текста учебника; (4) сегментирование текстов корпуса и расчеты значений лингвистических параметров текстов учебников для (5) идентификации статистически значимых корреляций лингвистических параметров с читабельностью/сложностью текста.

На этапе 1 тексты учебников в формате pdf были конвертированы в txt формат. Далее, каждый текстовый документ загружался в оффлайн инстру-

¹⁶ Hereinafter in the research the authors M.I. Andreeva, A.S. Borisova, R.R. Zamaletdinov developed a code for the textbooks comprising a letter and a number to mark the family names of the textbook authors and the grades.

мент Notepad ++ для предобработки, которая предполагала удаление лишних символов, пробелов, нечитаемых знаков, проверку полноты предложений по исходному pdf файлу двумя экспертами. Проверенный текстовый файл получал метаописание и размещался на удаленном сервере в архиве. Тексты учебников используются исключительно для исследовательских целей. Авторский код метаописания каждого учебника, например, Tbook(M)_R_R_06_06_VoMa_0_A_0_2022_28363, обеспечивает быстрый поиск в корпусе и содержит 11 зон, последовательно отражающих язык текста (R — русский), предметную область (R — русский язык), класс (06), год изучения дисциплины (06), автора/ов (VoMa — Воителева Т.М., Марченко Л.Г.), сферу функционирования (A — Academic, учебный), год издания (2022), количество словоформ (28363).

На *этапе 2* проводились следующие операции: (1) токенизация, т.е. разбиения исходного текста на отдельные, разделенные пробелами, единицы, в качестве которых выступают слова, числа, знаки пунктуации, позволяющая работать с текстом на уровне составляющих его элементов; (2) лемматизация словарного состава учебников, нацеленная на объединение словоформ одного слова и осуществлявшаяся при помощи специальных алгоритмов морфологического анализатора `rumorphy3.MorphAnalyzer`. В практических целях из списков лемм также были удалены так называемые стоп-слова, т.е. функциональные слова высокой частотности (Гадасин и др., 2022). Лемматизация считается завершенной по окончании подсчета словоформ каждой леммы, использованной в тексте. Список лемм каждого учебника с указанием количества их словоупотреблений фактически является частотным списком словарного состава, однако данные частоты абсолютны, т.е. отражают количество употреблений слова в изучаемом тексте.

Этап 3 включал составление списков лингвистических терминов из каждого учебника при помощи функции «Извлечение терминов» текстового профайлера RuLex¹⁷ для последующего сравнения их долей в лексическом составе изучаемых учебников и линеек учебников (рис. 1).

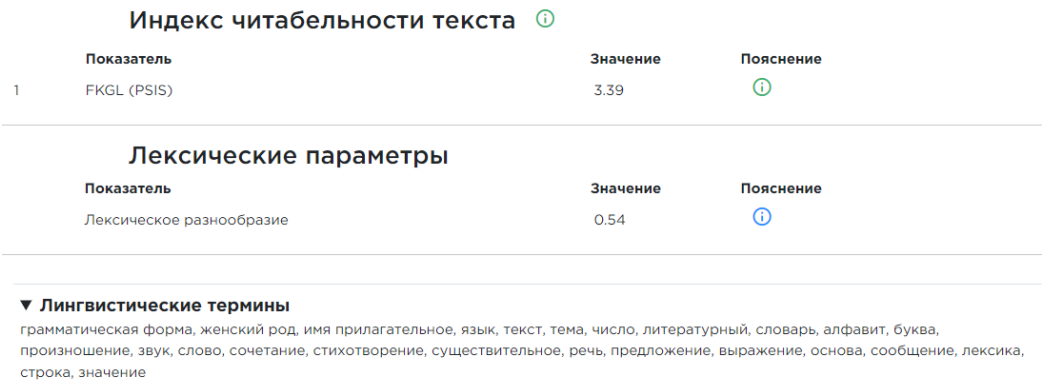


Рис. 1. Интерфейс текстового профайлера RuLex: функция извлечения терминов
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

¹⁷ RuLex text profiler. URL : <https://rulex.kpfu.ru/nlp> (дата обращения: 24.02.2024).

Text readability index ⓘ			
	Index	Meaning	Explanation
1	FKGL (PSIS)	3.39	ⓘ

Lexical parameters			
	Index	Meaning	Explanation
	Lexical diversity	0.54	ⓘ

▼ Linguistic terms
грамматическая форма, женский род, имя прилагательное, язык, текст, тема, число, литературный, словарь, алфавит, буква, произношение, звук, слово, сочетание, стихотворение, существительное, речь, предложение, выражение, основа, сообщение, лексика, строка, значение

Fig. 1. RuLex text profiler interface: linguistic terms extraction option
Source : Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Расчеты лингвистических параметров текстов осуществлялись при помощи профайлера RuLingva¹⁸, функционал которого позволяет рассчитывать 49 параметров текстов на русском языке (Andreeva et al., 2020).

Отметим, что достоверность значений ряда параметров, включая TTR, достигается только в тексте, длина которого не превышает 1000 слов (Biber, 2006, Вахрушева и др., 2021), именно поэтому расчеты метрик лингвистических параметров на *этапе 4* осуществлялись после сегментирования каждого текста на фрагменты по 1000 словоформ¹⁹. При этом важно, что число 1000 является условным, поскольку сегментированные отрывки содержат исключительно законченные предложения, т.е. их длина может быть несколько меньше или больше 1000 слов. В соответствии с современной парадигмой текстовой аналитики сегментирование производилось для последующего выявления значений (метрик) каждого параметра в текстовых отрывках одинаковой длины (Ranzato, 2018; Sardinha, 2002).

Выявление корреляций лингвистических параметров текстов учебников с читабельностью текста на *этапе 5* было осуществлено при помощи программы STATISTIKA²⁰. Читабельность рассчитывалась на основе формулы

$$FK(mod) = 0,36 \times \text{СДП} + 5,76 \times \text{СДС} - 11,97,$$

где СДП — средняя длина предложения в словах; СДС — средняя длина слова в слогах (Solnyshkina et al., 2018). Особенностью формулы читабельности FK(mod) является ее корреляция с годом обучения или классом школы.

¹⁸ RuLingva text profiler. URL: <https://rulingva.kpfu.ru/> (дата обращения: 24.02.2024).

¹⁹ RuLingva. Segmentatorus. URL: <https://rulingva.kpfu.ru/chunkizer> (дата обращения : 24.02.2024).

²⁰ StatSoft. Statistika. URL: <https://statsoftstatistica.ru/> (дата обращения : 24.02.2024).

Результаты

Линейки учебников имеют значительные различия в объеме и словарном составе: линейки учебников RyAl 5–7 и VyKi 5–7 не только объемнее, но и списки их словарного состава в среднем в 1,5 раза больше, чем в учебниках VoMa 5–7 и LaBa 5–7 (табл. 2).

Таблица 2

Словарный состав учебников 5–7 классов по русскому языку

Параметр	VoMa			RyAl			VyKi			LaBa		
Класс	5	6	7	5	6	7	5	6	7	5	6	7
Количество лемм	4981	5746	4690	7237	8096	7350	7134	8795	8538	4075	6586	8472

И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой, на материале авторского корпуса исследования.

Table 2

Vocabulary in Russian language textbooks for Grades 5–7

Paremeter	VoMa			RyAl			VyKi			LaBa		
Grade	5	6	7	5	6	7	5	6	7	5	6	7
Number of lemmas	4981	5746	4690	7237	8096	7350	7134	8795	8538	4075	6586	8472

S o u r c e : Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Среднее количество терминов в учебниках 5 класса составляет 116, в 6 и 7 классах — по 109. Наиболее «терминологичны» учебник ShFl 5, содержащий 129 терминов и учебники RyAl 6 и RyAl 7, в которых выявлено 120 и 121 термин соответственно. Однако доля терминов в текстах учебников весьма различна (табл. 3).

Таблица 3

Терминологическая плотность учебников 5–7 классов по русскому языку

Параметр	VoMa			RyAl			VyKi			LaBa		
Класс	5	6	7	5	6	7	5	6	7	5	6	7
Количество лемм терминов	105	97	90	118	120	104	120	110	113	111	113	116
	2,1	1,6	1,9	1,6	1,4	1,4	1,6	1,2	1,3	2,7	1,7	1,3
Количество словоформ терминов	1987	2211	1376	6568	7421	4658	9584	7286	8217	2441	5252	5397
	6,9	7,7	6,2	10,4	10,8	9,3	11,9	9,2	13	6	10,4	9,6

Примечание. В графе с заливкой указана доля терминов в словарном составе и тексте.
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Table 3

Terminological density of the Russian language textbooks for Grades 5–7

Parameter	VoMa			RyAl			ByKi			LaBa		
Grade	5	6	7	5	6	7	5	6	7	5	6	7
Count of terms' lemma	105	97	90	118	120	104	120	110	113	111	113	116
	2,1	1,6	1,9	1,6	1,4	1,4	1,6	1,2	1,3	2,7	1,7	1,3
Count of terms' tokens	1987	2211	1376	6568	7421	4658	9584	7286	8217	2441	5252	5397
	6,9	7,7	6,2	10,4	10,8	9,3	11,9	9,2	13	6	10,4	9,6

Note. Highlighted lines mark the distribution of the terms among the tokens and in the text.
Source: Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Лингвистические термины, извлеченные из текста каждого учебника при помощи платформы RuLingva, демонстрируют высокую степень сходства во всех линейках учебников (рис. 2).

Значения *индекса читабельности* всех изученных учебников находятся в пределах нормы, т.е. могут отклоняться от цифры, обозначающей год обучения на 1.0–1.5 единицы (Solnyshkina, Shoeva, 2024). Учебники по русскому языку демонстрируют положительную динамику *индекса читабельности*, при этом в учебниках VoMa читабельность с 5 по 7 класс растет минимально.

Показатели *лексического разнообразия* сохраняют значения в диапазоне от 0,4 до 0,5 и имеют незначительный рост с 5 по 7 класс только в линейках учебников ByKi (от 0,48 до 0,51) и LaBa (от 0,46 до 0,52). Полученные результаты подтвердили выводы предыдущих исследований: зависимость лексического разнообразия и читабельности не выявлены.

Индекс лексической плотности имеет приблизительно одинаковый диапазон для учебников всех классов и линеек: от минимального 0,54 до максимального 0,8, т.е. корреляция с классом не выявлена. Однако, как уже указывалось, читабельность сегментов в изучаемых учебниках имеет довольно значительные флуктуации от минимального 2,78 до максимального 9,13, и анализ на уровне сегментов текстов выявил статистическую зависимость лексической плотности и читабельности текстов по русскому языку. Данная зависимость характеризуется как средняя, но позволяет утверждать, что индекс лексической плотности может являться предиктором читабельности учебного текста по русскому языку.

Обсуждение

Терминологическое ядро учебников 5 класса составляют 87 лингвистических терминов, 6 класса — 79, 7 класса — 81 термин. Полные списки терминов размещены на сайте научных проектов НИЛ «Междисциплинарные исследования текста»²¹. Общими для частотных списков учебников 5–7 классов являются 63 лингвистических термина со средним показателем частотности от 1372,3 (**слово**) до 178,5 (**звук**) (рис. 2).

²¹ URL: <http://surl.li/zgmoqu> (дата обращения: 24.02.2024).

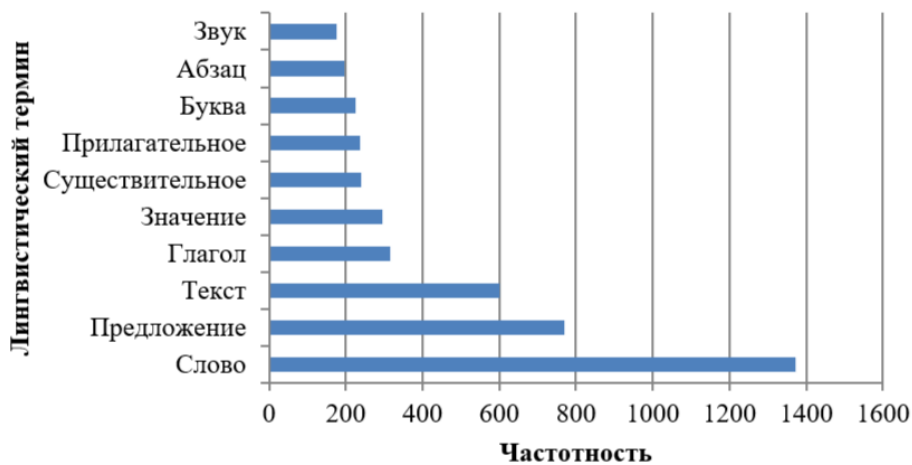


Рис. 2. Наиболее частотные лингвистические термины в текстах учебников 5–7 классов
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой, на материале авторского корпуса исследования.

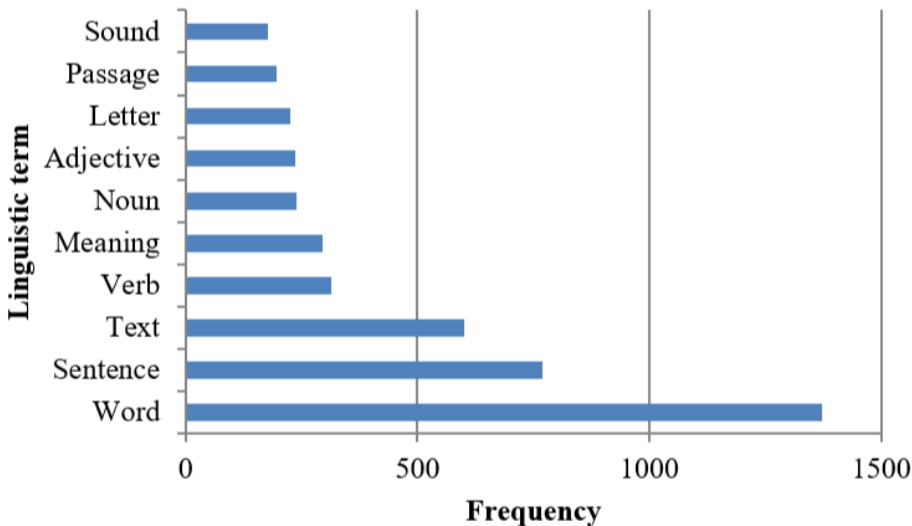


Fig. 2. Most frequent linguistic terms in textbooks for Grades 5–7
S o u r c e : Calculated by the M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Средние показатели индекса читабельности учебников находятся в диапазоне от 4,36 (05_RyAl) до 6,82 (07_ShFl), т.е. в целом они несколько ниже, чем в учебниках других предметных областей на русском языке (Gatiatullina et al., 2023). Очевидной причиной такого явления следует признать высокую долю иллюстраций из художественных произведений, индексы читабельности которых традиционно ниже (Solnyshkina et al., 2024). Во всех учебниках для пятого класса значения индекса читабельности как базового параметра сложности находятся в пределах нормы, т.е. от 4,5 до 5,5 по FK(mod) (табл. 4). Менее сложным, чем другие, является учебник RyAl_05, его читабельность — 4,36. Средняя читабельность всех рассмотренных учебников 7 класса несколько ниже уровня седьмого года обучения и находится в диапазоне 5,53–6,64.

Таблица 4

Индекс читабельности учебников по русскому языку 5–7 классов

Параметр	VoMa			RyAl			ByKi			LaBa		
Класс	5	6	7	5	6	7	5	6	7	5	6	7
Количество лемм	5,02	5,56	5,53	4,36	6,05	6,64	5,31	5,07	6,14	5,09	5,00	5,91

И с т о ч н и к : Расчеты выполнены М.И. Андреевой, А.С. Борисовой, Р.Р. Замалетдиновым на материале авторского корпуса исследования.

Table 4

Readability index of the Russian language textbooks for Grades 5–7

Paremeter	VoMa			RyAl			ByKi			LaBa		
Grade	5	6	7	5	6	7	5	6	7	5	6	7
Number of lemmas	5.02	5.56	5.53	4.36	6.05	6.64	5.31	5.07	6.14	5.9	5.00	5.91

S o u r c e : Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Важно указать, что индексы читабельности отдельных сегментов текста имеют значительно более широкий диапазон различий: от минимального 2,78 (ByKi 05) до максимального 9,13 (RyAl 06), что вполне естественно и, очевидно, связано с присутствием в тексте учебников по русскому языку отрывков разных функционально-стилистических типов речи — от рассуждения (правила и теория) до повествования (иллюстративные примеры и тексты упражнений).

Динамика читабельности во всех линейках учебников положительная (рис. 3), однако рост читабельности в учебниках VoMa крайне низок — от 5,02 в 5 классе до 5,53 в 7 классе (см. табл. 4), что может служить косвенным показателем роста других предикторов сложности, например, лексической плотности или роста абстрактности.

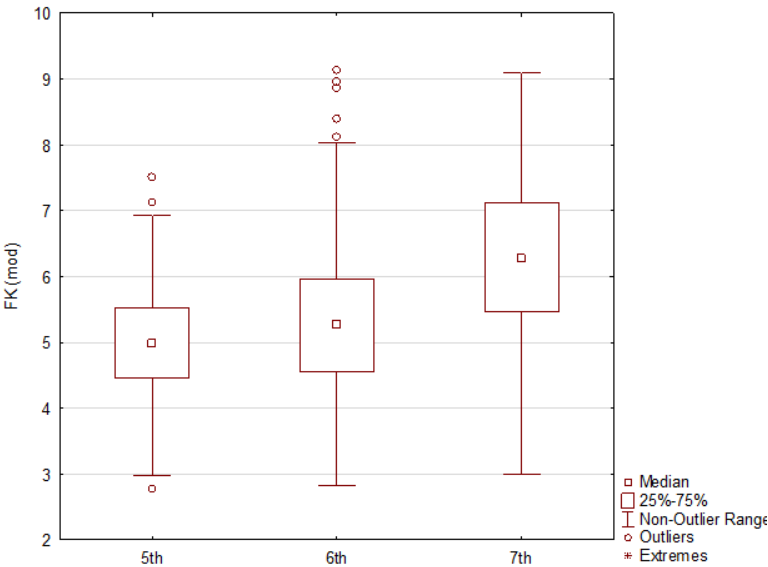


Рис. 3. Читабельность FK(mod) учебников по русскому языку 5–7 классов
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Fig. 3. Readability FK(mod) of Russian language textbooks for Grades 5–7
S o u r c e : Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Корреляционный анализ продемонстрировал различную степень зависимости лингвистических параметров изученных текстов с читабельностью (табл. 5).

Таблица 5

Параметры, имеющие статистически значимую корреляцию с читабельностью

Группы параметров	Параметр		r (сила корреляции по Спирману)
TTR	1	Лексическое разнообразие	0,08
Связность	2	Глобальный повтор существительного	0,53*
	3	Глобальный повтор аргумента	0,48*
	4	Локальный повтор существительного	0,46*
	5	Локальный повтор аргумента	0,42*
Категории существительных	6	Родительный падеж (сущ.)	0,47*
Знаменательные части речи	7	Лексическая плотность	0,46*
	8	Существительные	0,37*
	9	Прилагательные	0,33*
	10	Отношение прилаг. / сущ.	0,14
	11	Отношение глаг. / сущ.	–0,32*

Примечание. *— маркирует зависимости, имеющие статистическую значимость.
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Table 5

Parameters exhibiting correlation with text readability

Clusters of parameters	Parameter		r (Spearman correlation)
TTR	1	Type / token ratio	0.08
Cohesion	2	Global noun overlap	0.53*
	3	Global argument overlap	0.48*
	4	Local noun overlap	0.46*
	5	Local argument overlap	0.42*
Noun	6	Genitive case (noun)	0.47*
Content words	7	Lexical density	0.46*
	8	Nouns	0.37*
	9	Adjectives	0.33*
	10	Adjective / Noun ratio	0.14
	11	Verb / Noun ratio	–0.32*

Note. *— marks statistically significant correlations.
S o u r c e : Calculated by the M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Из 49 параметров, рассчитываемых RuLingva, в табл. 5 были включены преимущественно статистически значимые, которые можно рассматривать в качестве дополнительных к длине предложения и длине слова предикторов сложности текста. Предиктивная сила указанных параметров подтверждена в ряде исследований (см. обзор Solovyev et al., 2019) и не требует дополнительных доказательств. Последовательность размещения значений параметров в табл. 5 определялась на основе силы корреляции по убыванию. Исключение составил параметр TTR, т.е. лексическое разнообразие, описание которого необходимо для подтверждения отсутствия его предиктивной силы при расчете сложности (читабельности) текста, поэтому он включен в список первым. Из табл. 5 также были исключены параметры, зависящие от длины предложения, а именно, количество существительных / прилагательных / глаголов на предложение, поскольку их значения будут расти из класса в класс вместе с увеличением длины предложения (Gatiyatullina et al., 2020). Таким образом, табл. 5 содержит объединенные в три группы метрики параметров, имеющих преимущественно среднюю ($0,5 < r < 0,69$) и умеренную ($0,3 < r < 0,49$) корреляцию (по Спирману) с читабельностью (по FK(mod)). Дополнительно, как указывалось, мы рассматривали влияние параметра «лексическое разнообразие», чтобы подтвердить или опровергнуть вывод об отсутствии его влияния на сложность текста (Чурунина и др., 2023).

Обратимся к полученным данным. Как видим, корреляция TTR с читабельностью весьма незначительна — 0,08 (см. табл. 5), что в целом подтверждает выводы, сделанные ранее на материале учебников по русскому языку начальной школы (Чурунина и др., 2023). Показательно, что некоторый минимальный рост лексического разнообразия все же имеет место (рис. 4), а его диапазон, установленный на материале большого корпуса, скорее может характеризовать учебник как жанр, но не служить предиктором сложности.

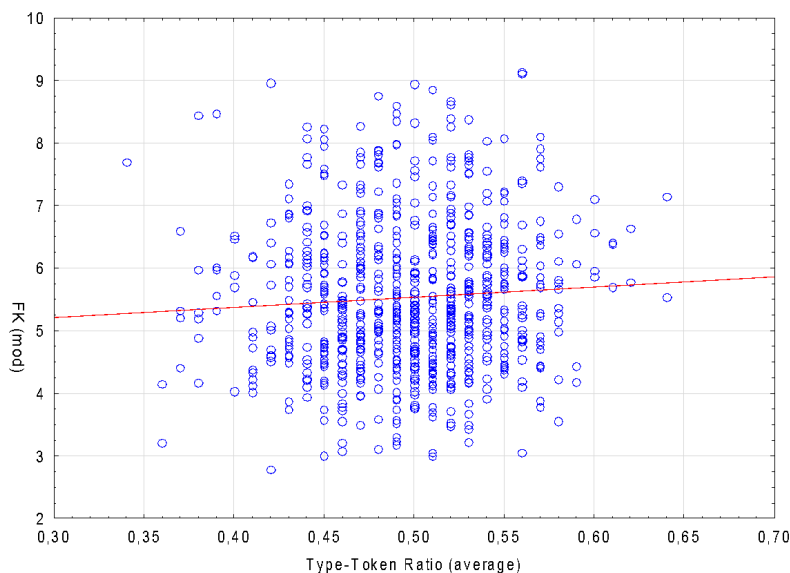


Рис. 4. Зависимость лексического разнообразия и читабельности FK(mod) учебников по русскому языку 5–7 классов

И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Fig. 4. Type / token ratio and FK (mod) correlation in Russian textbooks for Grades 5–7
S o u r c e : Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Корреляция средней и умеренной степеней зависимости наблюдается в трех группах параметров: (1) связность, (2) знаменательные части речи и (3) категории знаменательных частей речи.

В кластер «связность» объединены параметры «глобальный повтор существительных», «локальный повтор существительных», «глобальный повтор аргумента», «локальный повтор аргумента». Очевидно, что количество повторов частотных слов языка в учебном тексте также немаловажно для запоминания и понимания текста. Повторы слов напрямую влияют на лексическую связность текста. Значимыми при анализе связности являются повторы существительных и «аргументов», т.е. существительных, местоимений и именных сочетаний, в смежных предложениях. Экспериментальным путем доказано влияние лексических повторов на восприятие текста (Crossley et al., 2007; Kupriyanov et al., 2022). Каждый из параметров связности текста (табл. 5) имеет корреляцию с читабельностью в диапазоне 0,53–0,42, что указывает на среднюю и умеренную степень корреляции. Наглядно данная зависимость представлена на рис. 5.

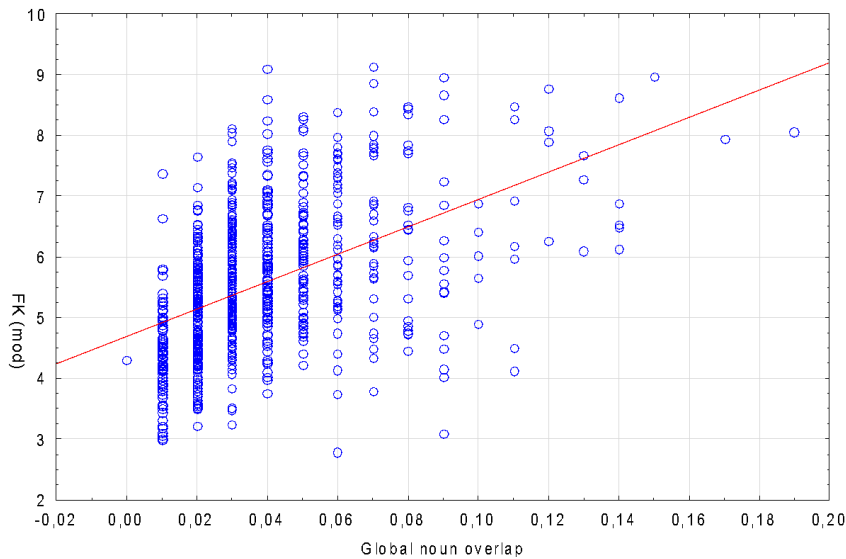


Рис. 5. Корреляция глобального повтора существительного и читабельности FK (mod) учебников по русскому языку 5–7 классов
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Fig. 5. Global noun overlap and FK (mod) correlation in Russian textbooks for Grades 5–7
S o u r c e : Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Статистическая значимость параметров локальных повторов существительных и глобальных повторов аргумента была доказана в ряде экспериментов НИЛ «Текстовая аналитика» (Kupriyanov et al., 2022). При сравнении параметров связности учебных текстов разных тематических областей ученые выявили высокие значения повторов существительных и аргументов

в филологических текстах, по сравнению с математическим и естественно-научными, что указывает на номинативный и дескриптивный характер учебников по русскому языку (Kurpiyanov et al., 2022).

Лексическая плотность изучаемых текстов растет преимущественно за счет увеличения долей имен существительных и прилагательных. Показательно, что в отличие от текстов учебников других предметных областей, в которых доля глаголов уменьшается от 5 к 7 классам (в гуманитарной и естественнонаучной областях) (Gatiyatullina et al., 2023), в учебниках по русскому языку доля глаголов меняется незначительно за счет их высокой плотности в текстах заданий, т.е. инструкций²², присутствующих абсолютно во всех учебниках русского языка независимо от уровня обучения.

Показатель *лексической плотности*, т.е. отношение количества знаменательных частей речи к общему количеству слов, растет вместе с читабельностью, поэтому его корреляция так хорошо заметна на рис. 6.

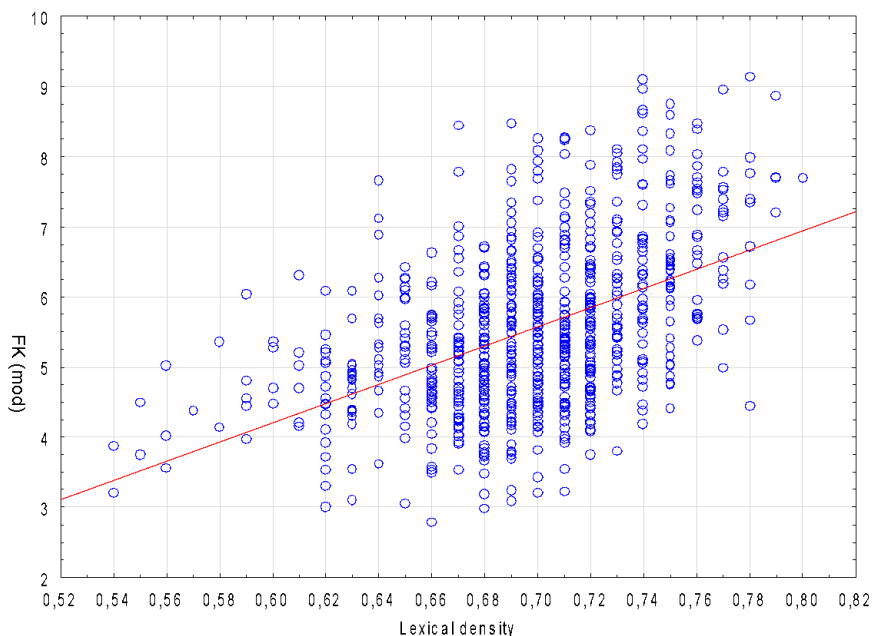


Рис. 6. Корреляция лексической плотности и читабельности FK(mod) в учебниках по русскому языку 5–7 классов

И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Fig. 6. Lexical density and FK(mod) correlation in textbooks for grades 5–7

S o u r c e : Calculated by the M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Результаты работы выявили взаимозависимость читабельности, «повествовательности», определяемой как отношение глаголов к существительным, и «дескриптивности», т.е. отношение прилагательных к существительным (рис. 7, 8).

²² Например, «Прочитайте текст и определите, к какому стилю он относится. Своё мнение обоснуйте. Для чего автор рассказал эту историю?», «Спишите текст, подчёркивая в нём слова, которые помогают связать предложения» (06_RyAl).

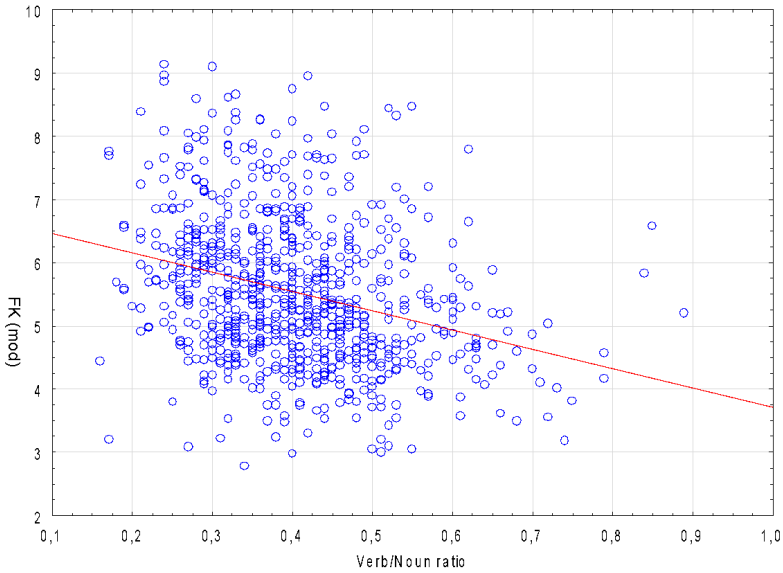


Рис. 7. Корреляция отношения глаголов к прилагательным и читабельности FK(mod) учебников по русскому языку 5–7 классов
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Fig. 7. Verb / noun ratio and FK (mod) correlation in Russian textbooks for grades 5–7
S o u r c e : Calculated by M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

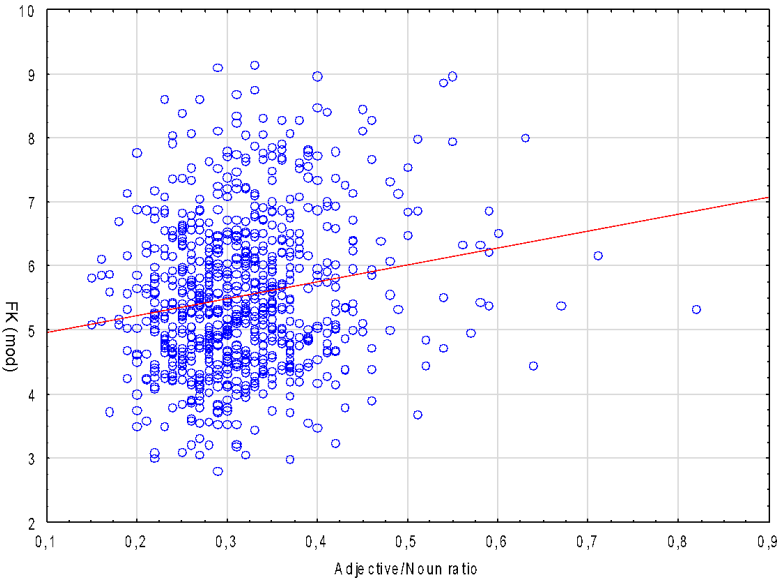


Рис. 8. Корреляция отношения параметра «описательность» (прил./сущ.) и читабельности (FK(mod)) учебников 5–7 классов
И с т о ч н и к : Расчеты выполнены М.И. Андреевой, Р.Р. Замалетдиновым, А.С. Борисовой на материале авторского корпуса исследования.

Fig. 8. The descriptiveness (adjective / noun) ratio and FK (mod) correlation in textbooks for grades 5–7
S o u r c e : Calculated by the M.I. Andreeva, R.R. Zamaletdinov, A.S. Borisova based on the research corpus.

Подчеркнем, что проведенный анализ более полно раскрыл словарный состав учебников по русскому языку и выявил корреляцию читабельности не с лексическим разнообразием текста, а с лексической плотностью.

Заключение

Полученные результаты, с одной стороны, подтверждают значимость лексической плотности и связности текста как предикторов сложности учебных текстов по русскому языку, с другой стороны, раскрывают перспективы дальнейших исследований.

Весьма перспективным считаем изучение влияния глаголов на читабельность текста. В отличие от других типов учебных текстов, в которых читабельность обратно зависима от доли глаголов в тексте, в учебных текстах по русскому языку этот предиктор не значим, что обусловлено спецификой текстов заданий, в большей степени представленных глаголами, что несет высокую плотность на всех образовательных уровнях. Поскольку текст учебника по русскому языку объединяет три типа формантов, которые в определенной степени связаны с различными функционально-стилистическими типами речи (теория (или правила) — рассуждение, задания — инструкция, упражнения — повествование), логично предположить, что роль глагола в каждом из них различна.

Отдельный интерес представляет изучение количества и роли местоимений как фактора, влияющего на лексические параметры, в частности, на лексическое разнообразие (TTR), повтор существительного и повтор аргумента. Логично предположить, что выявленный нами рост «терминологичности» учебного текста влечет за собой рост повторов имен существительных, который будет снижать неоднозначность воспринимаемого текста. В стремлении к однозначности авторы учебников снижают количество местоимений, увеличивая таким образом лексическую плотность текста.

Список литературы / References

- Andreeva, M., Solnyshkina M., Bukach, O., Zaikin, A., & Zamaletdinov, R. (2020). Assessment of comparative abstractness: Quantitative approach. In *CEUR Workshop Proceedings* (pp. 132–144). Kazan.
- Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. John Benjamins Publ. <https://doi.org/10.1075/scl.23>
- Churunina, A.A., Solnyshkina, M.I., & Yarmakeev, I.E. (2023). Lexical diversity as a predictor of the complexity of textbooks on the Russian language. *Russian Studies*, 21(2), 212–227. (In Russ.). <https://doi.org/10.22363/2618-8163-2023-21-2-212-227>.
- Чурунина А.А., Солнышкина М.И., Ярмеев И.Э. Лексическое разнообразие как предиктор сложности учебников по русскому языку // Русистика. 2023. Т. 21. № 2. С. 212–227. <https://doi.org/10.22363/2618-8163-2023-21-2-212-227>

- Crossley, S.A., Louwerse, M.M., McCarthy, P.M., & McNamara, D.S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1), 15–30.
- Dubay, W. (2004). *The Principles of Readability*. CA.
- Gadasin, D.V., Pak, E.V., Korovushkina, V.M., & Melkova, E.K. (2022). Preprocessing of textual information based on natural language terms. *REDS: Telecommunication Devices and Systems*, 12(1), 4–11. (In Russ.).
Гадасин Д.В., Пак Е.В., Коровушкина В.М., Мелькова Е.К. Предобработка текстовой информации на основе термов естественного языка // REDS : Телекоммуникационные устройства и системы. 2022. Т. 12. № 1. С. 4–11.
- Gatiyatullina, G., Solnyshkina, M., Solovyev, V., Danilov, A., Martynova, E., & Yarmakeev, I. (2020). Computing Russian morphological distribution patterns using RusAC online server. In *2020 13th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 393–398). IEEE Publ. <https://doi.org/10.1109/DeSE51703.2020.9450753>
- Gatiyatullina, G.M., Solnyshkina, M.I., Kupriyanov, R.V., & 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. <https://doi.org/10.18413/2313-8912-2023-9-1-0-2>
- Goldman, S.R., & Lee, C.D. (2014). Text complexity: State of the art and the conundrums it raises. *The Elementary School Journal*, 115(2), 290–300. <https://doi.org/10.1086/678298>
- Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Zh. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2), 193–202. <http://doi.org/10.3758/BF03195564>
- Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Hodder Arnold Publ.
- Kupriyanov, R.V., Solnyshkina, M.I., Dascalu, M., & Soldatkina, T.A. (2022). Lexical and syntactic features of academic Russian texts: A discriminant analysis. *Research Result. Theoretical and Applied Linguistics*, 8(4), 105–122. <http://doi.org/10.18413/2313-8912-2022-8-4-0-8>
- Okladnikova, S.V. (2010). A model for a comprehensive assessment of the readability of test materials. *Caspian Journal: Management and High Technologies*, (3), 63–71. (In Russ.).
Окладникова С.В. Модель комплексной оценки читабельности тестовых материалов // Прикаспийский журнал : управление и высокие технологии. 2010. № 3. С. 63–71.
- Paraschiv, A., Dascalu, M., & Solnyshkina, M.I. (2023). Classification of Russian textbooks by grade level and topic using Readerbench. *Research result. Theoretical and applied linguistics*, 9(1), 73–86. <https://doi.org/10.18413/2313-8912-2023-9-1-0-4>
- Ranzato, P.L.R. (2018). *A text segmentation technique based on language models*. Master thesis, Milano.
- Sardinha, T.B. (2002). Segmenting corpora of texts. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 18(2), 273–286. <https://doi.org/10.1590/S0102-44502002000200004>
- Solnyshkina, M., Guryanov, I., Gafiyatova, E., & Varlamova, E. (2018). Readability Metrics: the Case of Russian Educational Texts. In *Abstracts & Proceedings of ADVED 2018 — 4th International Conference on Advances in Education and Social Sciences* (pp. 676–681). Istanbul.
- Solnyshkina, M.I., Kupriyanov, R.V., & Shoeva, G.N. (2024). Linguistic profiling of text genres: adventure stories vs. textbooks. *Research Result. Theoretical and Applied Linguistics*, 10(1), 115–132. <https://doi.org/10.18413/2313-8912-2024-10-1-0-7>
- Solnyshkina, M.I., & Shoeva, G.N. (2024). Towards a taxonomy of textbooks as a genre: The case of Russian textbooks. *RUDN Journal of Language Studies, Semiotics and Semantics*, 15(3), 313–328. <https://doi.org/10.22363/2313-2299-2024-15-2-313-328>

- Solovyev, V., Andreeva, M., Solnyshkina, M., Zamaletdinov, R., Danilov, A., & Gaynutdinova, D. (2019). Computing concreteness ratings of Russian and English most frequent words: Contrastive approach. In *2019 12th International Conference on Developments in eSystems Engineering (DeSE)*, 403–408. <https://doi.org/10.1109/DeSE.2019.00081>
- Solovyev, V.D., Dascalu, M., & Solnyshkina, M.I. (2023). Discourse complexity: driving forces of the new paradigm. *Research Result. Theoretical and Applied Linguistics*, 9(1), 4–10. <https://doi.org/10.18413/2313-8912-2023-9-1-0-1>
- Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.
- Vakhrusheva, A.Y., Solnyshkina, M.I., Kupriyanov, R.V., Gafiyatova, E.V., & Klimagina, I.O. (2021). Linguistic complexity of academic texts. *Issues in Journalism, Education, Linguistics*, 40(1), 89–99. <https://doi.org/10.18413/2712-7451-2021-40-1-89-99>
- Вахрушева А.Я., Солнышкина М.И., Куприянов Р.В., Гафиятова Э.В., Климагина И.О. Лингвистическая сложность учебных текстов // Вопросы журналистики, педагогики, языкознания. 2021. Т. 40. № 1. С. 88–99. <https://doi.org/10.18413/2712-7451-2021-40-1-89-99>
- Vakhrusheva, A., Solovyev, V., Solnyshkina, M., Gafiyatova, E., & Akhtyamova, S. (2023). Revisiting Assessment of Text Complexity: Lexical and Syntactic Parameters Fluctuations. In *International Conference on Speech and Computer*, 430–441. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48309-7_35
- Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics*, 23(7), 443–452.

Сведения об авторах:

Андреева Мария Игоревна, кандидат филологических наук, доцент кафедры иностранных языков, Казанский государственный медицинский университет, Российская Федерация, 420012, Республика Татарстан, г. Казань, ул. Бутлерова, д. 49; старший научный сотрудник НИЛ «Мультидисциплинарные исследования текста», Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, Республика Татарстан, г. Казань, ул. Кремлевская, д. 18, корп. 1. *Сфера научных интересов*: корпусная лингвистика, сложность текстов, семантика, лексикология. ORCID: 0000-0002-5760-0934. SPIN-код: 9243-6995. Researcher ID: ABF-7003-2020. Scopus ID: 57195974758. E-mail: mariia99andreeva@yandex.ru

Замалетдинов Радиф Рифкатович, доктор филологических наук, профессор, директор Института филологии и межкультурной коммуникации, заведующий кафедрой общего языкознания и тюркологии, Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, Республика Татарстан, г. Казань, ул. Кремлевская, д. 18, корп. 1. *Сфера научных интересов*: когнитивная лингвистика, лингвокультурология, сопоставительное языкознание, история и закономерности функционирования татарского и русского языков, билингвизм. ORCID: 0000-0002-2692-1698. SPIN-код: 4027-8784. Researcher ID: M-2174-2013. Scopus ID: 56027359900. E-mail: director.ifmk@gmail.com

Борисова Анна Степановна, кандидат филологических наук, доцент кафедры иностранных языков, филологический факультет, Российский университет дружбы народов, Российская Федерация, 117198, г. Москва, ул. Миклухо-Маклая, д. 6. *Сфера научных интересов*: теория и практика перевода, когнитивная лингвистика, дискурс-анализ. ORCID: 0000-0002-7395-7028. SPIN-код: 2332-6093. Researcher ID: AАН-9347-2019. Scopus ID: 57194527178. E-mail: borisova-as@rudn.ru

DOI: 10.22363/2618-8163-2024-22-4-518-539

EDN: APRGCY

Research article

Predicative potential of lexical parameters: text complexity assessment in Russian language textbooks for 5–7 grades

Mariia I. Andreeva^{1,2}  , Radif R. Zamaletdinov² , Anna S. Borisova³ 

¹Kazan State Medical University, Kazan, Russian Federation

²Kazan (Volga Region) Federal University, Kazan, Russian Federation

³RUDN University, Moscow, Russian Federation

✉ mariia99andreeva@yandex.ru

Abstract. This study addresses the urgent issue of assessing the influence of lexical parameters on text complexity. The research has been conducted on the material of a specialized linguistic corpus, which includes texts of 15 modern Russian language textbooks for 5–7 grades, with a total size of 811911 words. The study is aimed at identifying the scale and dynamics of changes in vocabulary of Russian textbooks for 5–7 grades. The research algorithm included the following stages: (a) identifying the size and content of vocabulary in modern Russian textbooks for 5–7 grades, (b) assessing the share of linguistic terms in their vocabulary, and (c) identifying complexity predictors, i.e. parameters demonstrating a statistically significant correlation with readability. The analytical part of the study was preceded by a meta-description of the corpus, its tokenization, lemmatization, segmentation into fragments of approximately 1000 words. Text parameters were calculated using the text profiler RuLingva, and the correlation strength was assessed with STATISTIKA. To ensure the research results reliability, co-dependencies of lexical parameters and text readability were analyzed at two levels: at the textbook level (with average indicators for 15 textbooks for 5–7 grades) and at the level of 1000-word fragments. We revealed a slightly lower readability index, which was expected to be 1.0–1.5 levels higher. The latter may be a characteristic of Russian language textbook as a genre and indicate eclecticism of academic texts, including fragments of research discourse (rules and theory), fiction (exercises), and instructional discourse (texts of tasks). The research demonstrated that the share of linguistic terms does not exceed 2 % in the textbook vocabulary, but their share in the texts rises to 13 %. The statistical analysis indicates that the indices of ‘lexical density’, cohesion (global and local overlaps of nouns and arguments), ‘descriptiveness’ (ratio between adjectives and nouns), ‘narrativity’ (ratio between verbs and nouns), and the share of nouns in the genitive case are text complexity predictors. The prospects for the research include studying verbs and pronouns as complexity predictors in Russian language textbooks.

Keywords: educational text, corpus of language, educational text readability, word frequency, linguistic terminology, lexical density, cohesion

Contribution: *Andreeva M.I.* — data collection, research idea and design, writing the article; *Zamaletdinov R.R.* — research management, analysis of the obtained data, writing the article; *Borisova A.S.* — conducting the research, data verification, data processing.

Funding. This article has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030). This publication has been supported by the RUDN University Scientific Projects Grant System, project no. 050738-0-000.

Conflict of interests. The authors declare that they have no conflict of interests.

Article history: received 05.03.2024; accepted 18.06.2024.

For citation: Andreeva, M.I., Zamaletdinov, R.R. & Borisova, A.S. (2024). Predicative potential of lexical parameters: text complexity assessment in Russian language text-books for 5–7 grades. *Russian Language Studies*, 22(4), 518–539. (In Russ.). <http://doi.org/10.22363/2618-8163-2024-22-4-518-539>

Bio notes:

Mariia I. Andreeva, PhD in Philology, Associate Professor of the Department of Foreign Languages, Kazan State Medical University, 49 Butlerov St, Kazan, the Republic of Tatarstan, 420012, Russian Federation; Senior researcher of the research laboratory ‘Multidisciplinary Text Investigation’, Kazan (Volga region) Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests:* corpus linguistics, text complexity, semantics, lexicology. ORCID: 0000-0002-5760-0934. SPIN-code: 9243-6995. Researcher ID: ABF-7003-2020. Scopus ID: 57195974758. E-mail: mariia99andreeva@yandex.ru

Radif R. Zamaletdinov, Doctor Habil. (Philology), Professor, Director of the Institute of Philology and Intercultural Communication, Head of the Department of General Linguistics and Turkology, Kazan (Volga region) Federal University; 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests:* cognitive linguistics, linguoculturology, comparative linguistics, history and patterns of functioning of the Tatar and Russian languages. ORCID: 0000-0002-2692-1698. SPIN-code: 4027-8784. Researcher ID: M-2174-2013. Scopus ID: 56027359900. E-mail: director.ifmk@gmail.com

Anna S. Borisova, PhD in Philology, Associate Professor of the Department of Foreign Languages, Faculty of Philology, RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation. *Research interests:* translation studies, cognitive linguistics, discourse analysis. ORCID: 0000-0002-7395-7028. SPIN-code: 2332-6093. Researcher ID: AAH-9347-2019. Scopus ID: 57194527178. E-mail: borisova-as@rudn.ru