




DOI: 10.22363/2618-8163-2024-22-4-579-597

EDN: AYFCSE

Научная статья

## Лексическое обогащение в учебниках филологического блока: корпусный и статистический подходы

Х.Н. Галимова<sup>1</sup>, Е.В. Мартынова<sup>1</sup>, С.А. Москвичева<sup>2</sup><sup>1</sup>Казанский (Приволжский) федеральный университет, Казань, Российская Федерация<sup>2</sup>Российский университет дружбы народов, Москва, Российская Федерация galikha@mail.ru

**Аннотация.** Актуальность представленного исследования определяется значимостью объективных данных о частоте употребления лексических единиц в учебниках русского языка, а также неизученностью процессов освоения лексики в процессе обучения родному языку в школе. Описан опыт создания частотного словаря учебников филологического блока с опорой на лингвистический корпус учебников русского языка и литературы для 5–7 классов. Учебники филологического предметного блока содержат в себе усредненную модель русского языка и литературы, отражая актуальные для школьника темы и постепенно наращивая объем лексического состава от простого к более сложному. Цель исследования — оценка лексического обогащения в учебных текстах филологического предметного блока для 5–7 классов, а также усовершенствование методики формирования частотных списков. Исследование проведено на материале корпуса, в который вошли 66 учебников по русскому языку и литературе общим объемом 1 553 224 словоформ. Использование методов корпусной и компьютерной лингвистики, а также сравнительно-сопоставительного и статистического методов, в частности программы IKSWEB, среды Google Colab, библиотек Pandas, NLTK и Rmorphy позволило выявить, что объем частотного словаря учебников филологического блока 5 класса составляют 8984 лексемы, 6 класса — 7572 лексемы, 7 класса — 7321 лексемы. «Обогащение» лексики в 6 классе составляют 258 лексем, в 7 классе — 150 лексем. Лексическим ядром трех частотных списков являются слова следующих тематических групп: «Филологические термины», «Глаголы, обозначающие учебные операции», «Природа», «Родственные и дружеские отношения», «Искусство» и «Время». Выявлено, что обогащение словарного запаса у учащихся 6 класса осуществляется за счет архаизмов и историзмов; терминов, характеризующих формы общенационального языка, и терминов словообразования. В 7 классе обогащение частотного словаря осуществляется за счет лингвистических терминов по теме «Наименование глагольных форм», лексико-тематической группы «Религия» и общественно-политической лексики. Частотные списки подтвердили гипотезу о тематической сбалансированности текстов в современных учебниках русского языка и литературы среднего звена и ядерном положении терминологии в текстах рассматриваемых учебников. Перспектива

© Галимова Х.Н., Мартынова Е.В., Москвичева С.А., 2024

This work is licensed under a Creative Commons Attribution 4.0 International License  
<https://creativecommons.org/licenses/by-nc/4.0/legalcode>

исследования видится в осуществлении аналогичного исследования на материале учебных текстов филологического и других предметных блоков старшей школы для выявления внутри- и метапредметных связей.

**Ключевые слова:** лемма, частотный словарь, списки частотности, учебный корпус, термин, покрытие лексики

**Вклад авторов:** *Галимова Х.Н.* — формирование идеи; формулировка или развитие ключевых целей и задач; разработка методологии исследования; *Мартынова Е.В.* — сбор данных, анализ и интерпретация полученных результатов; сбор литературы, анализ и обобщение данных литературы; *Москвичева С.А.* — Критический пересмотр текста статьи; работа с графическим материалом; редактирование статьи.

**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.

**Конфликт интересов:** Авторы заявляют об отсутствии конфликта интересов.

**История статьи:** поступила в редакцию 12.04.2024; принята к печати 23.07.2024.

**Для цитирования:** *Галимова Х.Н., Мартынова Е.В., Москвичева С.А.* Лексическое обогащение в учебниках филологического блока: корпусный и статистический подходы // Русистика. 2024. Т. 22. № 4. С. 579–597. <http://doi.org/10.22363/2618-8163-2024-22-4-579-597>

## Введение

Ни одно слово в языке не существует отдельно от общей номинативной языковой системы, а частотность отдельных слов в дискурсах разных периодов и разных языковых личностей определяется значимостью обозначаемых ими реалий в жизни человека и общества (Коростелева, 2013). Частотность слова в речи всегда является отражением его функционального веса в системе языка, тесно связана с «его парадигматической значимостью, важностью, существенностью для языка» (Гиндин, 1982: 22). Как подчеркивает Л.А. Турыгина, «с каждым элементом можно связать число, которое тем больше, чем более употребителен данный языковой элемент» (Турыгина, 1988: 38).

В современной научной парадигме частотность слова рассчитывается как доля количества употреблений данного слова в тексте или корпусе языка от общего числа слов в тексте или корпусе, выраженная в процентах (Глинкина, 2011). Высокочастотная лексика формирует ядро лексической системы, в ее состав входят лексемы, репрезентирующие базовые, т.е. наиболее значимые, для представителей отдельной речевой культуры понятия и концепты (Чурунина, Солнышкина, Ярмакеев, 2023).

Частотность как достоверный предиктор сложности дискурса рассчитывается при помощи списков частотности (Мартынова и др., 2020), создаваемых на основе репрезентативных и сбалансированных корпусов (Rudell, 1983: 460). Современные частотные словари содержат две числовые харак-

теристики слов: их частотность, т.е. количество вхождений в определенном корпусе, и ранг или порядковый номер слова в частотном списке<sup>1</sup>.

Осмысление эмпирической закономерности распределения частот отдельных слов естественного языка, т.е. лемм, началось с выдающихся работ Дж.К. Ципфа (Гиндин, 1982), доказавшего, что частотность леммы в упорядоченном по частотности списке обратно пропорциональна ее порядковому номеру в списке, т.е. рангу<sup>2</sup>. Разработки компьютерных программ для статистического анализа языковых данных в значительной степени ускорили этапы исследований частотности лексики. Выявлению и валидации структурно-количественных закономерностей построения словаря и текста посвящены работы многих авторов, использующих частотные словари как лингвистические модели, изучение свойств которых способствует выявлению законов функционирования языка (Арапов, 1982; Орлов, 1978; Турыгина, 1988; Чурунина, Солнышкина, Ярмакеев, 2023).

Современная научная парадигма в данной области имеет в качестве основных следующие постулаты: (1) достоверность данных частотных словарей как упорядоченного по частоте встречаемости слов в заданном репрезентативном корпусе зависима от его размера, т.е. объема; (2) частотность лексики — один из наиболее значимых предикторов сложности (Лапошина, Лебедева, 2021), и поэтому имеет высокую степень значимости (Соловьев, Солнышкина, Макнамара, 2022).

Учебники филологического предметного блока, т.е. русского языка и литературы, готовят школьника к грамотному использованию русского языка в разных его контекстах (Solovyev et al., 2021), а также призваны формировать целостный и социально ориентированный взгляд на мир (Solnyshkina, Gafiyatova, 2014). Ожидается, что учебники представляют собой усредненную модель русского языка и литературы, отражая актуальные для школьника темы и постепенно наращивая объем лексического состава от простого к более сложному (Лапошина и др., 2019).

Сравнение, частотных списков учебников осуществляется в современной научной парадигме с применением двух мер: покрытие лексики («coverage») и обогащение лексики («enrichment»), предложенных Марко Барони (Baroni et al., 2009) и далее описанных О.В. Блиновой (Блинова, 2019). Меры призваны определить количество совпадающих слов в сравниваемых списках, т.е. в какой степени количество лемм в одном частотном списке «покрывается» количеством лемм в другом списке. «Обогащение» одного частотного списка относительно другого определяет долю новых слов (лемм) в корпусе при переходе из класса в класс (Блинова, 2019).

В рамках исследования проведено сравнение частотных списков лемм, имеющих относительную частоту больше или равно 5, а леммы с частотностью ниже данного порога, в полном соответствии с традицией современной научной парадигмы (Лапошина, Лебедева, 2022) исключили из списка.

<sup>1</sup> Алексеев П.М. Частотные словари : учебное пособие. СПб. : Изд-во С.-Петербург. ун-та, 2001. 156 с.

<sup>2</sup> Закон Ципфа:  $fr = c$ , где  $f$  — частота встречаемости слова в тексте;  $r$  — ранг, порядковый номер;  $c$  — постоянная величина, значение которой различается для разных языков.

В основе указанной традиции лежит положение о том, что появление в тексте редкого слова рассматривается как случайное, поскольку обусловлено исключительно решением автора для представления специфического задания или, например, прецедентного текста (Немова, 2015). Слова с низкой частотностью не предназначены для освоения школьниками и поэтому не включаются в списки лексического «обогащения». Низкочастотные слова образуют множество так называемых «легоменов», в которое входят гапак легомены (от греч. *hapax legomenon* «сказанное однажды»), т.е. слова, использованные в корпусе единожды (Творогов, 1995), дважды (*dis*), трижды (*tris*), и четырежды (*tetrakis*) (Malmkjær, 2002). Например, слова *дурачина* и *простофиля* в Частотном словаре 5 класса встречаются три раза и имеют источником исключительно текст «Сказки о золотой рыбке». Редкая лексика обычно используется в изучаемых учебниках для демонстрации специфических языковых явлений. Так, например, слово *фонарищик*<sup>3</sup> используется в тексте учебника единожды только для отработки и закрепления суффикса *-щик*. Многие устаревшие слова также встречаются в текстах учебников однократно: *Вяземский* (1<sup>4</sup>), *фолиант* (1).

Сложные дефисные слова также составляют отдельный пласт редкой лексики, например, *слово-образец* (4), *ученый-лингвист* (3), *медленно-медленно* (3), *рассуждение-доказательство* (3), *дятел-самец* (3), *город-крепость* (2) и т.д. Знание этих слов необходимо для сохранения национальной культуры и пополнения словарного запаса учеников, однако они в значительной степени повторяют простые слова. При этом важно подчеркнуть, что низкочастотная лексика представляет особый интерес для исследователей, поскольку является «потенциально недооцененной в имеющихся пособиях лексикой» (Лапошина, Лебедева, 2022: 92), однако в рамках нашей работы рассматривается как перспектива исследования.

**Цель исследования** — выявление специфики словарного состава учебников филологического блока 5, 6 и 7 классов российских школ. Планируется определить (1) объем лексики в изучаемых учебниках, (2) частотность использованных в них слов и (3) динамику изменения лексического состава.

## Методы и материалы

**Алгоритм** исследования учебников филологического блока включает:

1. Создание исследовательского корпуса учебников по русскому языку и литературе для 5–7 классов.
2. Усовершенствование методики формирования списков частотности лексики с использованием программ современной компьютерной лингвистики.
3. Формирование списков частотности лексики учебников филологического блока 5–7 классов, включающее следующие этапы: 1) преобразование

<sup>3</sup> Баранов М.Т., Ладыженская Г.А., Тростенцова Л.А. и др. Русский язык. 6 класс: учебник для общеобразоват. организаций : в 2 частях / науч. ред. Н.М. Шанский. 5-е изд. М. : Просвещение, 2015. 191 с. и 175 с.

<sup>4</sup> Здесь и далее в скобках указана частотность слова в частотном словаре соответствующего класса.

текстов в формат ТХТ; 2) токенизация текстов при помощи программы IKSWEB, предполагающая разбиение всех текстов на словоформы (токены); 3) лемматизация, т.е. приведение всех словоформ (токенов) к соответствующей лемме; 4) удаление (А) следующих групп слов: а) имена собственные, номинирующие героев художественных произведений и частных лиц, например, Саша, Леня, Даша и т.д. При этом в списке были сохранены все именованные сущности, представляющие специальные знания, например имена писателей, поэтов, известных деятелей, топонимы и проч.; б) числительные; в) стоп-слова, слова, затрудняющие индексирование страницы поисковыми системами (частицы, союзы, предлоги и т.д.); (Б) специальных символов, включая ударение и диакритические знаки; 5) снятие грамматической омонимии осуществлялось автоматически на основании контекста слова, например, *богатый* имя прилагательное и имя существительное; 6) присвоение каждой лексеме частеречного тэга произведено при помощи доработанной авторами программной библиотеки Rymystem; 7) расчет абсолютной нормализованной частотности слова в тексте учебника  $\text{Freq}(\text{ipt})$  по формуле

$$\text{Freq}(\text{ipt}) = \frac{m}{n} \times 1\,000,$$

где  $m$  — количество употреблений словоформ в корпусе;  $n$  — количество словоформ в корпусе без учета числительных и служебных частей речи; 8) присвоение каждой лемме ранга частотности на основе упорядоченного по частоте списка лемм, соответственно, ранг 1 присваивается самому частотному слову в корпусе, ранг 2 — менее частотному слову и т.д.

4. Выявление динамики изменения лексического состава учебников, т.е. словарного обогащения на каждом из этапов: 5→6, 6→7 на основе сравнительного анализа списков частотности учебников 5, 6 и 7 классов.

5. Тематическая классификация словарного обогащения на каждом из этапов: 5→6, 6→7.

Материалом исследования послужил корпус учебников по русскому языку и литературе для 5–7 классов, входящих в федеральный перечень<sup>5</sup>, т.е. допущенных к использованию организациями, осуществляющими образовательную деятельность на территории РФ, и выпущенных российскими издательствами «Просвещение», «Дрофа», «Русское слово», «Титул».

Исследовательский корпус учебников по русскому языку и литературе для 5–7 классов включает 66 учебников и 1 553 224 словоформ (табл. 1). Корпус содержит генеральную совокупность всех учебников ФГОС 2022 г., поэтому может быть признан сбалансированным и репрезентативным. Метаописание учебников содержит указания на жанр, язык, год издания, год обучения, год изучения дисциплины, облегчая поиск в корпусе.

<sup>5</sup> Приказ Министерства просвещения РФ от 21 сентября 2022 г. № 858 «Об утверждении федерального перечня учебников, допущенных к использованию при реализации имеющих государственную аккредитацию образовательных программ начального общего, основного общего, среднего общего образования организациями, осуществляющими образовательную деятельность и установления предельного срока использования исключенных учебников».

Таблица 1

Объем исследовательского корпуса<sup>6</sup>

Класс	Предмет	Количество учебников	Объем в словоформах
5	Русский язык	12	352332
6	Русский язык	12	323259
7	Русский язык	8	355296
Всего		32	1030887
5	Литература	12	184 936
6	Литература	12	178 619
7	Литература	10	158 782
Всего		34	522 337
<b>ИТОГО</b>		<b>66</b>	<b>1 553 224</b>

Table 1

## Size of the research corpus

Grade	Subject	Textbooks	Volume in wordforms
5	Russian	12	352332
6	Russian	12	323259
7	Russian	8	355296
In total		32	1030887
5	Literature	12	184 936
6	Literature	12	178 619
7	Literature	10	158 782
In total		34	522 337
<b>TOTAL</b>		<b>66</b>	<b>1 553 224</b>

Основу и достоверность результатов исследования обеспечили следующие критерии отбора: (1) общая предметная область — филология; (2) структурированность и сбалансированность по уровням обучения и объему — корпус разделен на три подкорпуса по уровням обучения: 5, 6 и 7 классы; (3) период выпуска учебника — одно десятилетие, с 2012 до 2022 г. Данные критерии обусловили выполнение всех принципов построения репрезентативного и сбалансированного корпуса: системность, жанровое единство, структурное единство, аутентичность и пр. (Нагель, 2008; Солнышкина, Гатиятуллина, 2020; Казачкова, Галимова, 2022) (см. табл. 1).

На основе лингвистического корпуса учебников по русскому языку и литературе для 5–7 классов было создано три частотных словаря: список из учебников филологического блока 5 класса состоит из 8984 лемм, 6 класса — из 7572 лемм, и 7 класс включает в себя 7321 лемм. Каждая лемма в списке снабжена двумя индексами: частотностью и рангом. Аналогично «Частотному словарю русского языка» под редакцией Л.Н. Засориной (1977) слова с одинаковой частотой имеют одинаковый ранг. Словари 5 и 7 классов запатентованы (Свидетельства о государственной регистрации № 2024622527, № 2024623508<sup>7</sup>). Заявления о патентовании словаря 6 класса находится на рассмотрении.

<sup>6</sup> Библиографические данные исследовательского корпуса и список источников размещены на сайте НИЛ «Мультидисциплинарные исследования текста». URL : <http://surl.li/zgmoqu> (дата обращения : 24.06.2024).

<sup>7</sup> Федеральный институт промышленной собственности. URL : <https://www.fips.ru/elektronnye-servisy/informatsionno-poiskovaya-sistema/index.php> (дата обращения : 15.05.2024).

Исследование осуществлялось с использованием **методов** корпусной и компьютерной лингвистики, а также сравнительно-сопоставительного и статистического методов. Токенизация текстов была произведена при помощи программы IKSWEB<sup>8</sup>. Списки частотности разрабатывались в среде Google Colab<sup>9</sup>, предназначенной для разработки и выполнения программного кода в облаке с помощью библиотеки Pandas<sup>10</sup>. Для анализа слов использовались библиотеки NLTK<sup>11</sup> и Rymorphy<sup>12</sup>.

## Результаты

Изучение динамики изменения и лексического состава филологического корпуса учебников 5–7 классов выявило ядро наиболее частотной лексики — 1211 лемм, объединенных в шесть основных тематических групп: «Термины», «Учебные действия», «Родственные и дружеские отношения», «Профессии», «Искусство», «Время».

Лексическое обогащение на этапе 5→6 классы составило 258 лексем, на этапе 6→7 — 150 лексем. В 6 классе словарный запас учащихся обогащается за счет историзмов и архаизмов, терминов, характеризующих формы общенационального языка и терминов словообразования. Обогащение частотных словарей учащихся 7 класса осуществляется за счет лингвистических терминов по теме «Наименования глагольных форм», лексики по теме «Религия» и общественно-политической лексики.

Доля лексического ядра в общем объеме каждого из учебников находится в диапазоне от 13 до 17 % и составляет: 13,4 % в 5 классе, 15,9 % в 6 классе, 16,5 % в 7 классе.

Нормализованная частотность наиболее частотных лемм в изученных подкорпусах 5–7 классов находится в диапазоне от 128 до 5. Данная лексика является подтверждением преемственности и согласованности словарного состава рассмотренных учебников и ядерного положения терминологии в текстах изучаемых учебников.

## Обсуждение

Разработка частотных словарей учебников филологического блока 5–7 классов опиралась на методику отечественной квантитативной лексикографии, применяемую семь десятилетий. Создаваемые первоначально исключительно для прикладных задач, а именно для совершенствования систем стенографии и методики обучения языкам (Несова, Бобрицких, 2018), частотные словари занимают достойное место в прикладной лингвистике. Особо следует указать на два первых частотных словаря русского языка: (1) словарь Г. Йоссельсона, изданный для преподавания русского языка

<sup>8</sup> SEO инструменты. URL : <https://iksweb.ru/> (дата обращения : 15.05.2024).

<sup>9</sup> Добро пожаловать в Colab! URL : [colab.research.google.com/](https://colab.research.google.com/) (дата обращения : 15.05.2024).

<sup>10</sup> PANDAS. URL : <https://blog.skillfactory.ru/glossary/pandas/> (дата обращения : 15.05.2024).

<sup>11</sup> NLTK. URL : <https://www.nltk.org/> (дата обращения : 15.05.2024).

<sup>12</sup> Морфологический анализатор rymorphy2. URL : <https://rymorphy2.readthedocs.io/en/stable/> (дата обращения: 15.05.2024).

в США (Josselson, 1953) и частотный словарь Э.А. Штейнфельд<sup>13</sup>, разработанный и опубликованный в Эстонии с целью определения лексического минимума детей-инофонов в начальной и средней школах (Shteifeldt, 1963). Словарь Э.А. Штейнфельд был составлен на основе статистических подсчетов встречаемости слов в коллекции текстов объемом свыше 400 тысяч слов, в которую входили тексты оригинальной (А. Гайдар, Н. Носов, Э. Успенский) и переводной (Марк Твен, Ханс Кристиан Андерсен, Шарль Перро) художественной литературы, молодежных газет, журналов и материалов радиопередач для молодежи (Shteifeldt, 1963). На основе данного словаря были составлены частотные словари-минимумы для учебных и методических целей. Таков, например, учебный словарь для зарубежных школ под редакцией Н.М. Шанского «4 000 наиболее употребительных слов русского языка»<sup>14</sup>.

Словарь под редакцией Л.Н. Засориной<sup>15</sup> «отражает устойчивую часть лексики, общеупотребительную и нейтральную относительно темы, жанра, автора, составляющую общую основу для всех жанров и разновидностей современной речи» (Засорина, 1977). Словарь содержит около 40 тысяч единиц, охватывает не только язык художественной литературы, но также тексты СМИ. Однако корпус этого словаря значительно устарел: состав корпуса включает большое количество слов из идеологических источников периода 1920–1960 гг., например, работы советских государственных партийных деятелей, материалы съездов КПСС, а также средства массовой информации СССР (Ляшевская, Шаров, 2009). Именно поэтому слова *социалистический*, *советский*, *товарищ*, *пятилетка* и т.п. зафиксированы в данном словаре в первой сотне слов наряду со служебными словами.

Российские ученые активно создают и используют специализированные частотные словари общенаучной лексики<sup>16</sup> и словари языка поэтов и писателей<sup>17</sup>. В современной отечественной лингвистике особую значимость имеет частотный словарь русского языка под редакцией О.Н. Ляшевской и С.А. Шарова (2009)<sup>18</sup>, созданный на коллекции текстов Национального корпуса русского языка (НКРЯ)<sup>19</sup> 1950–2007 гг. Словарь представлен

<sup>13</sup> Штейнфельдт Э.А. Частотный словарь современного русского литературного языка : 2 500 наиболее употребительных слов / под ред. В.А. Ицковича. Таллинн : НИИ педагогики СССР, 1968. 316 с

<sup>14</sup> Шанский Н.М., Даунене З.П., Бакеева Н.З., Гайдарова М.П., Караева Н.Б., Судавичене Л.В. 4 000 наиболее употребительных слов русского языка / под ред. действ. члена АПН СССР Н.М. Шанского. М. : Рус. яз., 1979. 712 с.

<sup>15</sup> Частотный словарь русского языка / под ред. Л.Н. Засорина. М. : Рус. яз, 1977. 936 с.

<sup>16</sup> Частотный словарь общенаучной лексики / под общ. ред. Е.М. Степановой. М. : Изд-во Моск. ун-та, 1970.

<sup>17</sup> Словарь языка Пушкина : в 4 томах / отв. ред. акад. АН СССР В.В. Виноградов. 2-е изд., доп. / Российская академия наук. Ин-т рус. яз. им. В.В. Виноградова. М. : Азбуковник, 2000; Словарь языка Достоевского / гл. редактор Ю.Н. Караулов. М., Азбуковник, вып. 1, 2001. 442 с., вып. 2, 2003, 510 с.; Словарь поэтического языка Марины Цветаевой : в 4 томах. Т. 1 : А-Г / отв. ред. М.Ю. Белякова. М. : Дом-музей Марины Цветаевой, 1996. 320 с.

<sup>18</sup> Ляшевская О.Н., Шаров С.А. Новый частотный словарь русской лексики. URL : <http://dict.ruslang.ru/freq.php> (дата обращения : 20.05.2024).

<sup>19</sup> Национальный корпус русского языка. URL : <http://www.ruscorpora.ru> (дата обращения : 24.06.2024).



и имеет высокую степень сбалансированности жанрового многообразия материала, включает коллекцию текстов разных типов, жанров и стилей, в т.ч. и тексты русской литературы зарубежья.

Создаваемые в рамках нашего исследования частотные словари характеризуют, с одной стороны, язык учебного текста соответствующего класса, а с другой стороны, ядро и периферию его словаря. Лексическое ядро учебников, т.е. список лемм, частотность которых больше или равна 5 словоупотреблений на 1000, в учебниках 5 класса составляют 1211 лемм, 6 класса — 1794 леммы и 7 класса — 1947 лемм. Наиболее частотные лексемы приведены в табл. 2.

Таблица 2

Частотная лексика филологического предметного блока 5–7 классов

5 класс		6 класс		7 класс	
Лемма	Freq (ipt)	Лемма	Freq (ipt)	Лемма	Freq (ipt)
правильно	128	звук	52	деепричастие	67
сегодня	125	Россия	44	писать	39
будущий	101	рассказ	39	страдательный	31
существительное	98	читать	38	деепричастный	29
начала	98	категория	37	наречие	26
инфинитив	92	утро	37	отглагольный	26
наклонение	69	профессионализм	34	писать	35
фрагмент	59	старославянизм	32	нарекать	21
фольклор	39	едва	24	блудный	18
качественный	16	печенег	22	обстоятельственный	17

Table 2

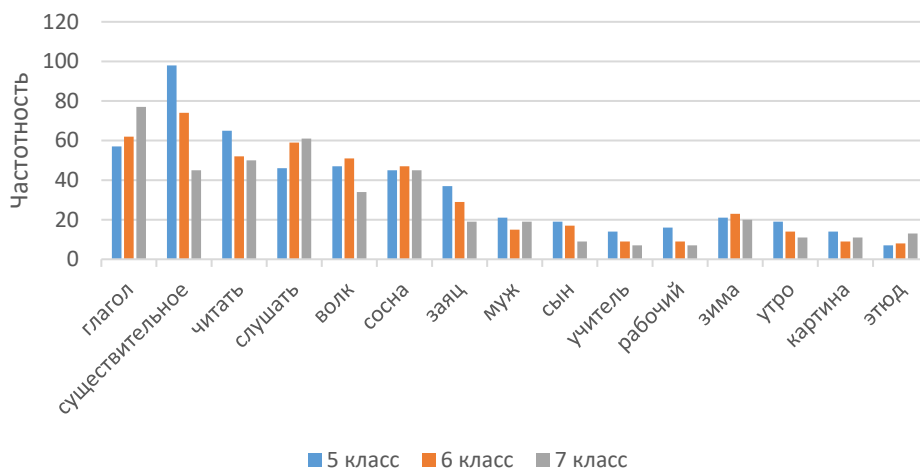
Frequency of philological vocabulary in textbooks of grades 5–7

Grade 5		Grade 6		Grade 7	
Lemma	Freq (ipt)	Lemma	Freq (ipt)	Lemma	Freq (ipt)
right	128	sound	52	Participle	67
today	125	Russia	44	to write	39
future	101	story	39	passive	31
noun	98	to read	38	verbal participle	29
at first	98	category	37	adverb	26
infinitive	92	morning	37	verbal	26
mood	69	professionalism	34	to write	35
fragment	59	Old Slavonism	32	to name	21
folklore	39	barely	24	prodigal	18
qualitative	16	pecheneg	22	circumstantial	17

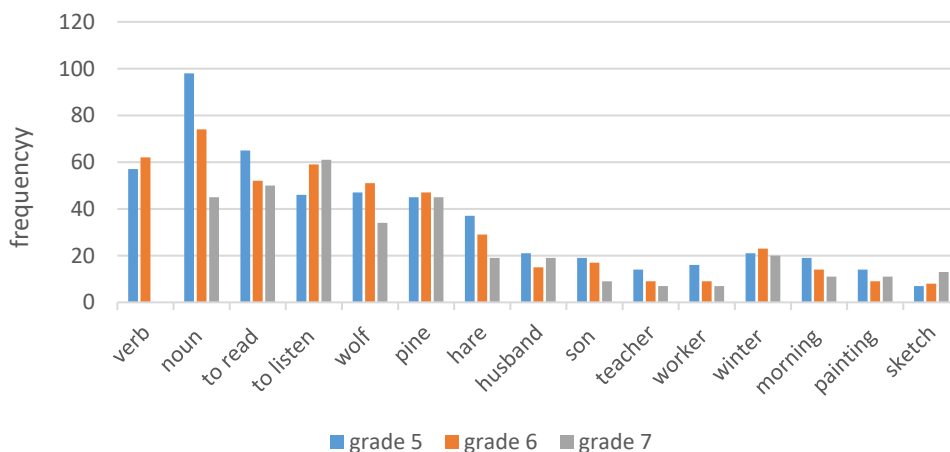
Общее лексическое ядро всех изучаемых учебников или «покрытие» составляют 1211 лемм.

Спектр тематического многообразия наиболее частотных слов «покрытия» весьма узок и включает небольшое количество основных групп (рис. 1). Тематический анализ, осуществленный на основе классификации

Л.Г. Бабенко<sup>20</sup> выявил следующие группы: «Филологические термины» (1/3<sup>21</sup>), «Глаголы, обозначающие учебные операции» (1/6), «Природа» (1/7), «Родственные и дружеские отношения» (1/8), «Профессии» (1/9), «Искусство» (1/10), «Время» (1/10). Оставшийся пласт лексики составляют лексические единицы, принадлежащие различным тематическим группам.



**Рис. 1.** Нормализованная частотность лексем ядра «покрытия»  
И с т о ч н и к : составлено Х.Н. Галимовой, Е.В. Мартыновой, С.А. Москвичевой с использованием программы Microsoft Excel.



**Figure 1.** Normalized frequency of the “coverage”

S o u r c e : Compiled by Kh.N. Galimova, E.V. Martynova, S.A. Moskvitcheva using the Microsoft Excel program.

Наибольшую частотность демонстрируют терминологические единицы (*наречие* (52<sup>22</sup>), *наклонение* (32), *роман* (20) и др.), составляющие примерно

<sup>20</sup> Большой толковый словарь русских существительных : свыше 15000 имен существительных, идеографическое описание, синонимы, антонимы / ред. Л.Г. Бабенко. 2-е изд., стереотип. М. : АСТ-ПРЕСС, 2008. 864 с.

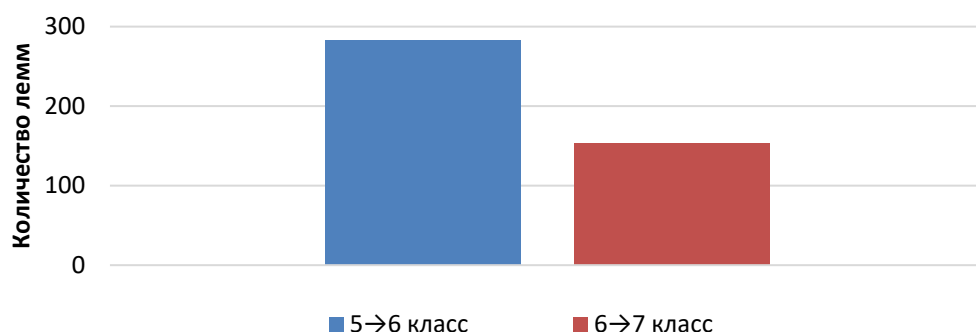
<sup>21</sup> В скобках указаны совокупные доли лексики соответствующей тематической группы.

<sup>22</sup> В скобках указана нормализованная частотность слова в корпусе — Freq (ipt).

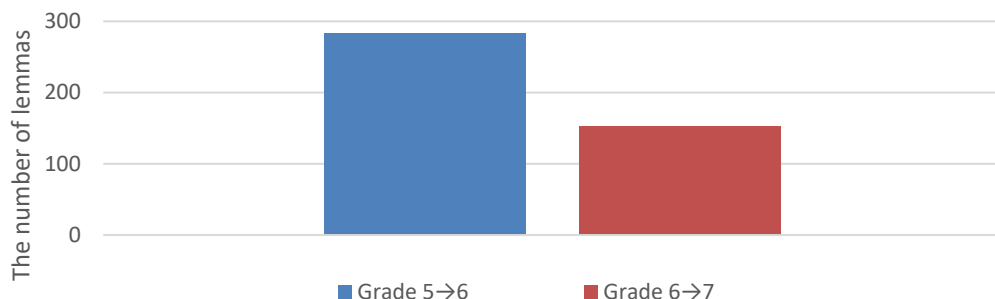
одну треть общего лексического ядра и демонстрирующие значительный рост по сравнению с аналогичным блоком в текстах учебников начальной школы. Предшествующие исследования показали, что «набор» терминологической лексики в текстах учебников начальной школы «весьма ограничен: несмотря на то, что формулировки заданий и справочная информация занимают более 60 % объема учебника, лексика этих блоков составляет 11 % от всех уникальных лемм учебника» (Лапошина и др., 2019: 6). Высокая частотность и наполняемость тематической группы «Термины» в изучаемых учебниках свидетельствуют о росте абстрактности текстов учебников филологического блока средней школы.

1/6 часть лексического «покрытия» принадлежит глаголам, обозначающим учебные операции (*повторять (26), списать (19), просмотреть (39)*). Они занимают второе место по количеству входящих в них единиц. На третьем месте — тематическая группа «Природа», в составе которой высокую частотность имеют лексические единицы подгруппы «Растения» (*каштан (13), пальма (12), бессмертник (6)*) и «Животные» (*медведь (11), лев (11), сокол (6)*). Далее следует группа «Родственные и дружеские отношения» (*сын (29), падчерица (22), товарищеский (5)*) и «Профессии» (*актер (7), певица (7), плотник (6)*). Лексические единицы тематической группы «Искусство» также имеют высокую нормализованную частотность: *оркестр (7), опера (6), хор (8)*. В эти списки вошли и различные слова с семантикой «Время» (*навсегда (7), редко (6), наспех (5)*).

Сопоставление частотных словарей позволило обнаружить и изменения в их составе: при переходе от класса к классу совершенствуется навык чтения учащихся, вместе с тем увеличиваются словарный запас и лексическое разнообразие. На рис. 2 представлено «обогащение» лексики, т.е. увеличение объема лексического состава при переходе из класса в класс, на каждом из этапов. Список «обогащения» этапа 5→6, т.е. лексем, впервые появляющихся в учебниках 6 класса и имеющих частотность больше или равно 5, включает 258 лексических единиц. Аналогичный список этапа 6→7 содержит 150 лексических единиц (рис. 2).



**Рис. 2.** Размер списков обогащения лексики филологического блока на этапах 5→6 и 6→7  
Источники: составлено Х.Н. Галимовой, Е.В. Мартыновой, С.А. Москвичевой с использованием программы Microsoft Excel.



**Figure 2.** Vocabulary enrichment lists at stages 5→6 and 6→7

Source: Compiled by Kh.N. Galimova, E.V. Martynova, S.A. Moskvitcheva using the Microsoft Excel program.

Как мы видим, список обогащения на этапе 5→6 значительно превосходит список 6→7.

В табл. 3 приведены 10 наиболее частотных слов, которые пополнили частотные словари русского языка и литературы в 6 и 7 классах соответственно.

**Обогащение лексики в учебниках 6 и 7 классов**

Таблица 3

Ранг	5→6 классы		6→7 классы	
	Лемма	Частотность	Лемма	Частотность
1	категория	37	деепричастие	67
2	определительный	36	страдательный	31
3	профессионализм	34	деепричастный	29
4	старославянизм	32	наречие	26
5	суффиксальный	28	отглагольный	26
6	архаизм	26	праведник	19
7	аршин	26	культурный	19
8	жаргонизм	24	завет	17
9	печенег	22	обстоятельный	17
10	историзм	22	оппонент	11

**Vocabulary enrichment in textbooks of Grades 6 and 7**

Table 3

Rank	Grades 5→6		Grades 6→7	
	Lemma	Frequency	Lemma	Frequency
1	category	37	Participle	67
2	definitive	36	passive	31
3	professionalism	34	verbal participle	29
4	Old Slavonic	32	adverb	26
5	suffix	28	verbal	26
6	archaism	26	saint	19
7	Arshin	26	cultural	19
8	jargon	24	covenant	17
9	Pecheneg	22	circumstantial	17
10	istorizm	22	opponent	11

При переходе из класса в класс объем материала расширяется, добавляются новые темы. Обогащение словарного запаса учащихся 6 класса осуществляется за счет лексики следующих тематических блоков (табл. 4): (1) устаревшие слова (историзмы и архаизмы), составляющие около 25 %

от всей лексики: *аршин* (26<sup>23</sup>), *опричник* (13), *губерния* (11), *атаман* (9), *объездчик* (6), *милостивый* (6), *сажень* (10), *быличка* (9) и др.; (2) термины, характеризующие формы общенационального языка, около 25 %: *архаизм* (26), *жаргонизм* (24), *историзм* (22) и др.; (3) термины словообразования, составляющие около 35 %: *вопросительно-относительный* (19), *суффиксальный* (28), *бессуффиксный* (21).

Таблица 4

Тематическое обогащение лексики в учебниках 6 и 7 классов

5→6 классы		6→7 классы	
Лексика	Доля от общего количества слов, %	Лексика	Доля от общего количества слов, %
Устаревшие слова (архаизмы-историзмы)	25	Наименования глагольных форм	25
Термины, характеризующие формы общенационального языка	25	Религия	30
Термины словообразования	35	Общественно-политическая	25
Другие	15	Другие	20

Table 4

Thematic vocabulary enrichment in Textbooks of Grades 6–7

Grades 5–6		Grades 6–7	
Vocabulary	Number of words, %	Vocabulary	Number of words, %
Obsolete words (archaisms-historicisms)	25	Names of verb forms	25
Terms characterizing forms of the national language	25	Religion	30
Word-formation terms	35	Socio-political	25
Other	15	Other	20

Изучение устаревшей лексики способствует сохранению и трансляции культурного кода, популяризации традиционных устоев народов России. Устаревшие слова составляют «традиционный исторический пласт» (Генералова, 2019) и входят в состав национально- и культурно-учебной предмаркированной лексики русского языка.

В 7 классе обогащение частотных словарей осуществляется за счет (1) лингвистических терминов по теме «Наименования глагольных форм», составляющих около 25 % лексики (*деепричастие* (67<sup>24</sup>), *отглагольный* (26), *страдательный* (31), *инфинитив* (10) и др.), (2) лексики по теме «Религия» — около 30 % (*завет* (17), *псалом* (7), *псалтирь* (6), *мусульманский* (9), *благочестивый* (6), *святыцы* (5), *праведность* (5), *монашество* (5), *христианка* (5), *христианский* (5) и др.). Общественно-политическая лексика составляет примерно 10 %. Например, *парламент* (6), *оппонент* (11), *стратегия* (9)

<sup>23</sup> В скобках указана нормализованная на 1000 словоупотреблений частотность в учебнике 6 класса.

<sup>24</sup> В скобках указана нормализованная на 1000 словоупотреблений частотность в учебнике 7 класса.

и др. Увеличение доли филологических терминов вполне закономерно, поскольку Рабочая программа учебной дисциплины «Русский язык» 7 класса предполагает обобщение и систематизирование знания учащихся о глаголе, причастии и деепричастии, причастном и деепричастном оборотах<sup>25</sup>. Знание религиозной лексики способствует не только формированию представления о материальной и духовной культуре своего народа, его прошлом, но также и формированию духовности, высокой нравственности, культуры и толерантности.

В целом лексический состав учебников «Русский язык» и «Литература» дает благодатный материал для воспитания важнейших качеств современного человека и формирования культурного кода.

### **Заключение**

Результаты представленного исследования, имея высокую значимость для русистики, могут быть использованы для проведения ряда научных изысканий. Перспектива нашего исследования видится в следующих направлениях:

во-первых, корпус текстов учебников филологического предметного блока может быть использован для получения достоверных данных о жанровой специфике учебного текста. Верификацию на материале данного корпуса может получить, например, фиксируемая учеными гетерогенность лексической системы различных типов специализированных дискурсов. И хотя учебный текст включает преимущественно нейтральную и кодифицированную лексику, большой интерес в современных условиях «демократизации» учебного дискурса представляет вопрос о многообразии регистров, представленных в текстах учебников и тематическом наполнении их лексического состава;

во-вторых, логично предположить, и это может быть использовано как гипотеза будущего исследования, что текст учебника должен иметь преимущественно положительную или нейтрально окрашенную лексику. Автоматизированный контент- и сентимент-анализ иллюстративных текстов учебников русского языка позволит выявить отношение автора(-ов) к объектам, явлениям и событиям, речь о которых идет в тексте. Особый интерес в этой связи представляет позиционирование наименований культурно-значимых для страны объектов;

в-третьих, весьма перспективным представляется проведение аналогичного исследования на материале учебных текстов филологического и других предметных блоков старшей школы для выявления внутри- и мета-предметных связей.

---

<sup>25</sup> Рабочая программа (ID 4220440) учебного предмета «Русский язык. Базовый уровень» для обучающихся 7 классов. URL : [https://1school-lobnya.ru/assets/files/program/2024-2025/2024\\_7\\_Русский%20язык.pdf](https://1school-lobnya.ru/assets/files/program/2024-2025/2024_7_Русский%20язык.pdf) (дата обращения : 12.06.2024).

### Список литературы

- Арапов М.В.* Текст и язык — целостность и организменность // Учен. зап. тартуского ун-та. Тарту, 1982. Вып. 628.
- Блинова О.В.* Низкочастотные слова в русском языке и подходы к моделированию общезыковой частотности // Социо- и психолингвистические исследования. 2019. № 7. С. 7–13.
- Генералова Е.В.* Устаревшая лексика русского языка: вопросы преподавания и лексикографической интерпретации // Journal of applied linguistics and lexicography. 2019. № 2. С. 370–380. <https://doi.org/10.33910/2687-0215-2019-1-2-371-380>
- Гиндин С.И.* Частота слова и его значимость в системе языка // Tartu ülikooli toimetised. 1982. Вып. 658. С. 22–54.
- Глинкина Л.А.* Частотность как значимый регистр лексикографии и фразеологии // Проблемы истории, филологии, культуры. 2011. № 3 (33). С. 7–11.
- Казачкова М.Б., Галимова Х.Н.* Создание лингвистического корпуса учебников английского языка // Иностранные языки в школе. 2022. № 2. С. 32–38.
- Коростелева Л.В.* Высокочастотные имена существительные, прилагательные и числительные в современном русском языке (по материалам лексикографии) : монография. Нижневартовск : Изд-во Нижневарт. гос. ун-та, 2013. 115 с.
- Лапошина А.Н., Веселовская Т.С., Лебедева М.Ю., Курпрещенко О.Ф.* Лексический состав текстов учебников русского языка для младшей школы: корпусное исследование // Компьютерная лингвистика и интеллектуальные технологии : по материалам международной конференции «Диалог 2019». 2019. Т. 18 (25). С. 351–363.
- 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. <https://doi.org/10.22363/2618-8163-2021-19-3-331-345>
- Лапошина А.Н., Лебедева М.Ю.* Формирование частотного словаря-минимума русского языка для детей-инофонов на основе корпусных данных // МИРС. 2022. № 3. С. 90–99. <https://doi.org/10.24412/1811-1629-2022-3-90-99>
- Мартынова Е.В., Солнышкина М.И., Мерзлякова А.Р.* Лексические параметры учебного текста (на материале текстов учебного корпуса русского языка) // Филология и культура. 2020. № 3 (61). С. 72–80. <https://doi.org/10.26907/2074-0239-2020-61-3-72-80>
- Нагель О.В.* Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура. 2008. № 4. С. 53–59.
- Немова А.Н.* Прецедентные тексты как культурный код в процессе изучения литературы // Нижегородское образование. 2015. № 1. С. 22–26.
- Несова Н.М., Бобрицких Л.Я.* Представление словаря в теоретической и учебной лексикографии // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2018. Т. 9. № 2. С. 439–450. <https://doi.org/10.22363/2313-2299-2018-9-2-439-450>
- Орлов Ю.К.* Модель частотной структуры лексики // Исследования в области вычислительной лингвистики и лингвостатистики. М., 1978. С. 59–118.
- Солнышкина М.И., Гатиятуллина Г.М.* История развития корпусной лингвистики (на примере англоязычных корпусов) // Вестник Томского государственного университета. Филология. 2020. № 63. С. 133–157. <https://doi.org/10.17223/19986645/63/8>
- Творогов О.В.* Гапаксы «Слова» // Энциклопедия «Слова о полку Игореве» : в 5 томах. СПб. : Дмитрий Буланин, 1995. Т. 2. С. 12–15.
- 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. <https://doi.org/10.22363/2687-0088-30161>

- Турьгина Л.А. Моделирование языковых структур средствами вычислительной техники. М., 1988. 175 с.
- Чурунина А.А., Солнышкина М.И., Ярмакеев 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>
- Baroni M., Bernardini S., Ferraresi A., Zanchetta E. The WaCky Wide Web : A collection of very large linguistically processed webcrawled corpora // Language resources and evaluation. 2009. Vol. 43. Pp. 209–226.
- Malmkjær K. The linguistics encyclopedia. 2nd ed. London ; New York : Routledge, 2002. 87 p.
- Josselson H. The Russian word count and frequency analysis of grammatical categories of standard literary russian. detroit : Wayne University Press, 1953.
- Rudell A. Frequency of word usage and perceived word difficulty : Ratings of Kucera and Francis words // Behaviour research methods, instruments, & computers. 1993. No. 25 (4). Pp. 455–463.
- Shteinfeldt E. Frequency dictionary of a modern Russian literary language : 2500 Most common words. Tallin, 1963. 316 p.
- Solnyshkina M., Gafiyatova E. Modern forestry English : Macro- and microstructure of low register dictionary // Journal of language and literature. 2014. Vol. 5. № 4. Pp. 220–224. <https://doi.org/10.7813/jll.2014/5-4/47>
- Solovyev V., Islamov M., Solnyshkina M., Kupriyanov R., Gafiyatova E. Sentiment analysis for Russian academic texts : A lexicon-based approach // CEUR workshop proceedings. 2021. 3090. Pp. 89–97.

#### **Сведения об авторах:**

Галимова Халида Нурисламовна, кандидат филологических наук, старший научный сотрудник НИЛ «Мультидисциплинарные исследования текста» института филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, г. Казань, ул. Кремлевская, д. 18. *Сфера научных интересов*: сложность текста, сравнительно-историческое, типологическое и сопоставительное языкознание. ORCID: 0000-0003-1817-5004. SPIN-код: 7931-3389. E-mail: galikha@mail.ru

Мартынова Екатерина Владимировна, старший преподаватель кафедры теории и практики преподавания иностранных языков, младший научный сотрудник НИЛ «Мультидисциплинарные исследования текста» института филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, г. Казань, ул. Кремлевская, д. 18. *Сфера научных интересов*: сложность текста, семантические роли, теория языка. ORCID: 0000-0001-5883-0718. SPIN-код: 9431-7981. E-mail: katerinamarty@yandex.ru

Москвичева Светлана Алексеевна, кандидат филологических наук, доцент кафедры общего и русского языкознания, филологический факультет, Российский университет дружбы народов, Российская Федерация, 117198, г. Москва, ул. Миклухо-Маклая, д. 10/2. *Сфера научных интересов*: социолнгвистика, дискурс-анализ. ORCID: 0000-0002-8047-7030. SPIN-код: 9596-7692. E-mail: moskvicheva-sa@rudn.ru.



DOI: 10.22363/2618-8163-2024-22-4-579-597

EDN: AYFCCE


Research article

## Lexical enrichment of philological textbooks: corpus and statistical approaches

Khalida N. Galimova<sup>1</sup>, Ekaterina V. Martynova<sup>1</sup>,  
Svetlana A. Moskvitcheva<sup>2</sup>

<sup>1</sup>Kazan (Volga Region) Federal University, *Kazan, Russian Federation*

<sup>2</sup>RUDN University, *Moscow, Russian Federation*

 galikha@mail.ru

**Abstract.** The relevance of the study is determined by the need to study objective data on vocabulary frequency in Russian language textbooks and mastering vocabulary in teaching Russian as the native language at school. The article describes the experience of creating a frequency dictionary of philological textbooks based on the linguistic corpus of textbooks on the Russian language and literature for 5–7 grades. Philological textbooks present an average model of the Russian language and literature, reflecting topics relevant to the student and gradually increasing the volume of lexical complexity. The aim of the article is to assess lexical enrichment in philological textbooks for 5–7 grades and to improve the methodology for compiling frequency lists. The study was carried out on the material of a corpus including 66 textbooks on the Russian language and Literature with the total size of 1,553,224 tokens. Methods of corpus and computational linguistics methods, comparative-contrastive, and statistical methods (IKSWEB program, the Google Colab environment, the Pandas, NLTK and Pymorphy libraries) revealed that the frequency list of the 5th grade comprises 8984 lemmas; the 6th grade, 7572 lemmas; the 7th grade, 7321 lemmas. Vocabulary “enrichment” in the 6th grade consists of 258 lexemes, and in the 7th grade, 150 lexemes. The lexical core of the three frequency lists are words of the thematic groups “Philological terms”, “Verbs denoting educational actions”, “Nature”, “Family and friendly relations”, “Art”, and “Time”. The 6th grade vocabulary “enrichment” includes archaisms and historicisms, terms denoting forms of the national language, and word-formation terms. The 7th grade “enrichment” comprises of linguistic terms on the themes “Names of verb forms”, “Religion”, and socio-political vocabulary. The frequency lists confirmed the hypothesis about the thematic balance of texts in modern textbooks on the Russian language and Literature and linguistics terminology being the core in the textbooks. The prospects of the study are seen in conducting a similar research of educational texts in Philology and other subjects form the textbooks for senior school in order to define intra- and meta-subject links.

**Keywords:** lemma, frequency dictionary, frequency lists, Academic corpus of the Russian language, term, Philology, lexical coverage, lexical enrichment

**Contribution:** *Galimova Kh.N.* — Formulation of the idea; formulation or development of key goals and objectives; development of the research methodology; *Martynova E.V.* — Data collection, analysis and interpretation of the obtained results; collection of literature, analysis and generalization of literature data; *Moskvicheva S.A.* — Critical revision of the text of the article; work with graphic material; the editing of the article.

**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: 12.04.2024; accepted: 23.07.2024.

**For citation:** Galimova, Kh.N., Martynova, E.V., & Moskvitcheva, S.A. (2024). Lexical enrichment of philology textbooks: corpus and statistical approaches. *Russian Language Studies*, 22(4), 579–597. (In Russ.). <http://doi.org/10.22363/2618-8163-2024-22-4-579-597>

## References

- Arapov, M.V. (1982). Text and language — integrity and organization. *Scientific Journal of the Tartu University*. Tartu. 628. (In Russ.).
- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed webcrawled corpora. *Language Resources and Evaluation*, 43, 209–226. <https://doi.org/10.1007/s10579-009-9081-4>
- Blinova, O.V. (2019). Russian low-frequency words and approaches to modeling general language frequency. *Socio- and Psycholinguistic Studies*, (7), 7–13. (In Russ.).
- 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 Language Studies*, 21(2), 212–227. (In Russ.). <https://doi.org/10.22363/2618-8163-2023-21-2-212-227>
- Generalova, E.V. (2019). Obsolescent vocabulary of the Russian language: educational and lexicographic interpretation issues. *Journal of Applied Linguistics and Lexicography*, (2), 371–380. (In Russ.). <https://doi.org/10.33910/2687-0215-2019-1-2-371-380>
- Gindin, S.I. (1982). The frequency of the word and its significance in the language system. *Tartu Ülikooli Toimetised*, (658), 22–54. (In Russ.).
- Glinkina, L.A. (2011). Frequency as an important characteristic of lexicography and phraseography. *Journal of Historical, Philological and Cultural Studies*, (3), 7–11.
- Josselson, H. (1953). *The Russian word count and frequency analysis of grammatical categories of standard literary Russian*. Detroit: Wayne University Press.
- Kazachkova, M.B., & Galimova, H.N. (2022). A linguistic corpus of English textbooks creation. *Foreign Languages at School*, 2, 32–38. (In Russ.).
- Korosteleva, L.V. (2013). *High-frequency nouns, adjectives and numerals in modern Russian (based on the materials of lexicography)*: monograph. Nizhnevartovsk: Publishing House of Nizhnevartovsk State University. (In Russ.).
- Laposhina, A.N., Veselovskaya, T.S., Lebedeva, M.Yu., & Kupreshchenko, O.F. Lexical composition of the Russian language textbooks for primary school: corpus study. In *Computational linguistics and intellectual technologies: based on the materials of the international conference “Dialogue 2019”*. Vol. 18 (pp. 351–363). (In Russ.).
- Laposhina, A.N., & Lebedeva, M.Yu. (2022). Developing a Russian frequency core vocabulary list for foreign children based on corpus data. *Mir Russkogo Slova*, (3), 90–99. (In Russ.). <https://doi.org/10.24412/1811-1629-2022-3-90-99>
- Laposhina, A.N., & Lebedeva, M.Yu. (2021). Textometr: an online tool for automated complexity level assessment of texts for Russian language learners. *Russian Language Studies*, (3), 331–345. (In Russ.). <https://doi.org/10.22363/2618-8163-2021-19-3-331-345>
- Malmkjær, K. (2002). *The linguistics encyclopedia*. 2nd ed. London; New York: Routledge.
- Martynova, E.V., Solnyshkina, M.I., & Merzlyakova, A.R. (2020). Lexical parameters of the academic text (based on the texts of the academic corpus of the Russian language). *Philology and Culture*, (3), 72–80. <https://doi.org/10.26907/2074-0239-2020-61-3-72-80>

- Nagel, O.V. (2008). Corpus linguistics and its use in computer-based language teaching. *Language and Culture*, 4, 53–59. (In Russ.).
- Nemova, A.N. (2015). Case texts as a cultural code in the process of studying the literature. *Nizhny Novgorod Education*, (1), 22–26. (In Russ.).
- Nesova, N.M., & Bobritskikh, L.Ya. (2018). Representation of the dictionary in theoretical and educational lexicography. *RUDN Journal of Language Studies, Semiotics and Semantics*, 9(2), 439–450. (In Russ.). <https://doi.org/10.22363/2313-2299-2018-9-2-439-450>
- Orlov, Yu.K. (1978). *A model of the frequency structure of vocabulary. Research in computational linguistics and linguostatistics*. Moscow State University, 59–118. (In Russ.).
- Rudell, A. (1993). Frequency of word usage and perceived word difficulty: Ratings of Kucera and Francis words. *Behaviour Research Methods, Instruments, & Computers*, (25), 455–463.
- Shteifeldt, E. (1963). *Frequency dictionary of a modern Russian literary language: 2500 most common words*. Tallin.
- Solnyshkina, M., & Gafiyatova, E. (2014). Modern forestry English: Macro- and microstructure of low register dictionary. *Journal of Language and Literature*, 5(4), 220–224. <https://doi.org/10.7813/jll.2014/5-4/47>
- Solnyshkina, M.I., & Gatiyatullina, G.M. (2020). The history of corpus linguistics (on the example of the English language corpora). *Tomsk State University Journal of Philology*, 63, 133–157. (In Russ.). <https://doi.org/10.17223/19986645/63/8>
- Soloviev, V.D., Solnyshkina, M.I., & McNamara, D.S. (2022). Computational linguistics and discursive complexity: paradigms and research methods. *Russian Journal of Linguistics*, 26(2), 275–316. (In Russ.). <https://doi.org/10.22363/2687-0088-30161>
- Solovyev, V., Islamov, M., Solnyshkina, M., Kupriyanov, R., & Gafiyatova, E. (2021). Sentiment Analysis for Russian Academic Texts: A Lexicon-Based Approach. In *CEUR Workshop Proceedings*, 3090 (pp. 89–97).
- Turygina, L.A. (1988). *Modeling of language structures by means of computer technology*. Moscow. (In Russ.).
- Tvorogov, O.V. (1995). Gapaks “Words”. In *Encyclopedia “Words on Igor's Regiment”*. In 5 vol. Vol. 2 (pp.12–15). St. Petersburg: Dmitry Bulanin. (In Russ.).

#### Bio notes:

*Khalida N. Galimova*, PhD in Philology, Senior Researcher at the Multidisciplinary Text Investigation Research Institute of Philology and Intercultural Communication, Kazan Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests*: the complexity of the text, comparative historical, typological and comparative linguistics. ORCID: 0000-0003-1817-5004. SPIN-code: 7931-3389. E-mail: galikha@mail.ru

*Ekaterina V. Martynova*, Senior Lecturer at the Department of Theory and Practice of Teaching Foreign Languages, Junior Researcher at the Multidisciplinary Text Investigation Research Institute of Philology and Intercultural Communication, Kazan Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests*: text complexity, semantic roles, language theory. ORCID: 0000-0001-5883-0718. SPIN-code: 9431-7981. E-mail: katerinamarty@yandex.ru

*Svetlana A. Moskvicheva*, PhD in Philology, Associate Professor of the General and Russian Linguistics Department, Faculty of Philology, RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation. *Research interests*: sociolinguistics, discourse analysis. ORCID: 0000-0002-8047-7030. SPIN-code: 9596-7692. E-mail: moskvicheva-sa@rudn.ru