Прикладные аспекты лингвистики / Applied Aspects of Linguistics

Научная статья Original article
DOI: 10.12731/3033-5981-2026-18-1-541 EDN: RWPCAF
УДК 81’42

Эволюция лексического богатства русского языка (корпусное исследование на основе диахронических датасетов национального корпуса русского языка)

Т.А. Рычкова
Мурманский арктический университет, Мурманск, Российская Федерация

Аннотация

Обоснование. Актуальность исследования обусловлена необходимостью изучения лексических изменений в русском языке современными методами. Научная новизна работы заключается в разработке и применении комплексных статистических методов для системного количественного анализа русской лексики на основе частотных словарей 1700–1916, 1918–1991 и 1992–2016 г. общим объемом 250 млн. употреблений, что позволило выявить и количественно описать особенности динамики лексического богатства и структуры словарного состава русского языка в диахронической перспективе.

Цель – определение особенностей динамики лексического богатства русского языка на основе диахронических датасетов 1700–1916, 1918–1991 и 1992–2016 г.

Материалы и методы. Материалы исследования – диахронические датасеты Национального корпуса русского языка 1700–1916, 1918–1991 и 1992–2016 г. Методы – компьютерная обработка корпусов и проверка на соответствие закону Ципфа, расчет индексов Херфиндаля-Хиршмана (HHI), Симпсона, Бергера-Паркера, энтропии Шеннона, коэффициента лексического разнообразия Type-Token Ratio (TTR), статистических значимостей (хи-квадрат) и др.

Результаты. Корпусный анализ диахронических данных за периоды 1700–1916, 1918–1991 и 1992–2016 г. выявил снижение общего лексического разнообразия и богатства русского языка от дореволюционного к постсоветскому периоду. Однако это обеднение лексики происходит преимущественно за счет редких и малочастотных слов, в то время как активный словарь, наоборот, расширяется и становится более продуктивным.

Ключевые слова: динамика лексики; эволюция лексики; лексика русского языка; Национальный корпус русского языка; диахронические датасеты; русский

язык дореволюционного периода; русский язык советского периода; русский язык постсоветского периода

Для цитирования. Рычкова, Т. А. (2026). Эволюция лексического богатства русского языка (корпусное исследование на основе диахронических датасетов национального корпуса русского языка). *Russian Social and Humanitarian Studies / Российские социогуманитарные исследования*, *18*(1), 156–181. https://doi.org/10.12731/3033-5981-2026-18-1-541

The evolution of the lexical richness of the Russian language (corpus research based on diachronic datasets from the national corpus of the Russian language)

T.A. Rychkova

Murmansk Arctic University, Murmansk, Russian Federation

Abstract

Background. The relevance of this study is determined by the need to investigate lexical change in Russian using modern methods. The scientific novelty of the work lies in the development and application of complex statistical models and indices for a systematic quantitative analysis of Russian lexis based on new, previously unexplored material—frequency dictionaries for the periods 1700–1916, 1918–1991, and 1992–2016, with a total size of 250 million tokens. This made it possible to identify and quantitatively describe the dynamics of lexical richness and the structure of the vocabulary in a diachronic perspective.

Purpose. To determine the specific features of the dynamics of lexical richness in Russian on the basis of frequency dictionaries for the periods 1700–1916, 1918–1991, and 1992–2016.

Materials and methods. The material of the study consists of diachronic datasets of the Russian National Corpus for the periods 1700–1916, 1918–1991, and 1992–2016. The methods include computer-based corpus processing and testing for compliance with Zipf's law, calculation of the Herfindahl–Hirschman Index (HHI), Simpson's index, Berger–Parker index, Shannon entropy, the Type–Token Ratio (TTR) as a coefficient of lexical diversity, chi-square tests of statistical significance, and other related measures.

Results. Corpus-based analysis of diachronic data for the periods 1700–1916, 1918–1991, and 1992–2016 revealed a decrease in the overall lexical diversity and richness of Russian from the pre-revolutionary to the post-Soviet period. However, this lexical impoverishment occurs mainly at the expense of rare and low-frequency words, whereas the active vocabulary, by contrast, expands and becomes more productive.

Keywords: lexical dynamics; lexical evolution; Russian lexis; Russian National Corpus; diachronic datasets; pre-revolutionary Russian; Soviet-period Russian; post-Soviet Russian

For citation. Rychkova, T. A. (2026). The evolution of the lexical richness of the Russian language (corpus research based on diachronic datasets from the national corpus of the Russian language). *Russian Social and Humanitarian Studies*, *18*(1), 156–181. https://doi.org/10.12731/3033-5981-2026-18-1-541

Введение

Лексический состав языка подвергается постоянным изменениям, и в современной лингвистике для определения этих изменений все более популярными становятся корпусные исследования [6; 10; 11; 14; 17; 21; 23; 24].

На основе лингвистических корпусов составляются частотные словари, которые позволяют оценить лексику системно. Самые известные частотные словари русского языка - это словари Г. Йосельсона 1953 г., Э.А. Штейнфельд 1963 г., Л.Н. Засориной 1977 г., Л. Лённгрена 1993 г., О. Н. Ляшевской и С. А. Шарова 2009 г. [29-33]. Самые крупные из них – словари Л.Н. Засориной и О.Н. Ляшевской и С.А. Шарова. Словарь Л.Н. Засориной был создан на основе корпуса текстов общим объёмом около миллиона словоупотреблений. Словарь О.Н. Ляшевской и С.А. Шарова включает около 92 млн. словоупотреблений и на сегодняшний день остаётся одним из наиболее авторитетных справочных ресурсов по частотности современных русских слов.

Эти словари синхронические, то есть дают возможность изучить состояние языка в один конкретный период времени. Однако для оценки динамики лексики более информативны диахронические словари, и встречаются они значительно реже.

На данный момент в России есть диахронические словари Казанского федерального университета 1920-2019 гг. и 1992-2019 г. [28] и частотные словари на основе корпусов русских рассказов 1900-1930 г. [3]. Диахронические словари КФУ перекрываются по хронологическому охвату, в связи с чем больше подходят для сравнения с другими языками, чем для изучения изменений собственно

русского языка. Частотные словари на основе русских рассказов 1900-1930 г. наиболее близки идее данной работы. Авторы проекта составили частотные словари на основе корпусов русских рассказов 1900-1930 г., разделив их по трем периодам: дореволюционному, революционному и послереволюционному. В полученных словарях были подсчитаны частотные списки лемм, словоформ и части речи, после чего для каждого частотного списка были рассчитаны статистические переменные. Однако сами статистические переменные все же отличаются от тех, которые используются в нашем исследовании, ввиду небольшого, по сравнению с нашим, охвата исторического периода и наличия контекста. На основе сравнения частотных словарей разных периодов авторы сделали выводы о, например, изменении тональности рассказов (максимальное значение тональности приходится на революционные 1917-1918 гг., в которых отрицательно-окрашенная лексика превышает положительно-окрашенную почти в три раза), о том, какие слова и части речи преобладали в текстах (общая тенденция в распределении частей речи – увеличение доли существительных и глаголов за счет снижения доли наречий, личных местоимений и союзов) и т.д. [1; 3; 6-8; 24].

Материалы и методы

В данной работе представлен анализ текстовых корпусов, а именно диахронических датасетов Национального корпуса русского языка (НКРЯ). Корпуса НКРЯ характеризуются репрезентативностью, сбалансированностью корпусов по жанрам и типам текстов, большим объемом, что делает их идеальным материалом для корпусного исследования.

В нашем исследовании использованы скачиваемые диахронические датасеты НКРЯ, которые представляют собой наборы текстов трех разных периодов (без метаразметки): 1700–1916, 1918–1991 и 1992–2016 гг., соответствующие досоветскому, советскому и постсоветскому периодам нашей истории. Общий объем датасетов, как указано на сайте НКРЯ, – 250 млн. словоупотреблений [27].

Полученные от НКРЯ датасеты были обработаны с помощью программы Python, лемматизированы и преобразованы в частотные словари в виде списков лемм с указанием частотности словоупотребления и перечислением всех словоформ.

Таким образом было получено три частотных словаря, сохраненных в виде файлов Excel:

1) частотный словарь досоветского периода 1700–1916 г. (досоветский);

2) частотный словарь 1918–1991 г. (советский);

3) частотный словарь 1992–2016 г. (постсоветский).

Далее понятия досоветские, советские, постсоветский «словари», «корпуса» используются как синонимичные.

После компьютерной обработки частотные словари были очищены от цифр и иноязычных вкраплений и подсчитаны с точки зрения следующих параметров: общее количество лемм (лексем, то есть исходных форм слов, в данном случае эти понятия используются как синонимы), их форм; был сделан перерасчет абсолютной частоты употребления каждой леммы в ipm (instances per million), то есть было рассчитано количество вхождений (употреблений) слова на миллион слов для того, чтобы избежать искажения результатов из-за разного объема словарей.

В очищенном от цифр и иноязычных вкраплений досоветском файле оказалось 427 963 русских лексем, общая сумма словоупотреблений которых 72 180 232, ipm частота всех словоупотреблений – 1 128 065.

В очищенном советском файле – 478 211 русских лексем, общая сумма словоупотреблений – 93 152 300, ipm – 983 236.

В очищенном постсоветском словаре – 412 823 лексемы, общее количество словоупотреблений - 81 363 980, ipm – 969 646.

Для определения лексического богатства и разнообразия в получившихся словарях были использованы проверка соответствия распределения частот закону Ципфа, расчет индексов концентрации Хиршмана (HHI), индексов разнообразия Симпсона и Бергера–Паркера, определение энтропии Шеннона, а также вычисление

коэффициента лексического разнообразия (Type-Token Ratio, TTR) и критерия хи-квадрат (χ^2).

Результаты

1. Закон Ципфа

Закон Ципфа утверждает, что частота употребления слова в языке обратно пропорциональна его рангу в списке, упорядоченном по убыванию частоты: первое по частоте слово встречается примерно вдвое чаще второго, второе — вдвое чаще третьего и так далее [22; 26].

Для оценки словаря корпуса с помощью закона Ципфа слова ранжируют по частоте от самого частотного слова к наименее частотным, каждому слову в зависимости от частоты присваивают определенный ранг (самое частотное слово имеет ранг 1, второе по частотности 2 и т.д.), затем строят график зависимости частоты слова от его ранга и проверяют, насколько данные соответствуют гипотезе обратной пропорциональности, характерной для закона Ципфа. В данном случае данные были рассчитаны для 10.000 самых частотных слов в каждом корпусе.

В описании закона Ципфа используются понятия наклона и коэффициента детерминации R^2. Наклон – это параметр, определяющий крутизну уменьшения частоты слова с ростом его ранга при построении графика частотности на логарифмической шкале. Параметр наклона у естественного языка в классической формулировке равен -1. Коэффициент детерминации R^2 показывает качество аппроксимации эмпирических данных моделью, соответствующей закону Ципфа [22; 26].

Если распределение частот в исследуемом корпусе близко к закону Ципфа, это означает, что словарь типичен для естественного языка. Наоборот, значительные отклонения от закона могут указывать на искусственность текста, его неприменимость для изучения естественного языка. Закон Ципфа помогает также проанализировать принципы функционирования лексики: как часто используются основные слова и как быстро насыщается корпус новыми слова-

ми, сравнить разные корпуса по интенсивности использования слов и лексической насыщенности [22].

В результате проверки на соответствие закону Ципфа были получены следующие результаты (см. таблицу 1 и график 1).

Таблица 1.

Корпус	Наклон	R^2
Дореволюционный	-1.087332	0.992138
Советский	-1.053374	0.992757
Постсоветский	-1.067211	0.988439

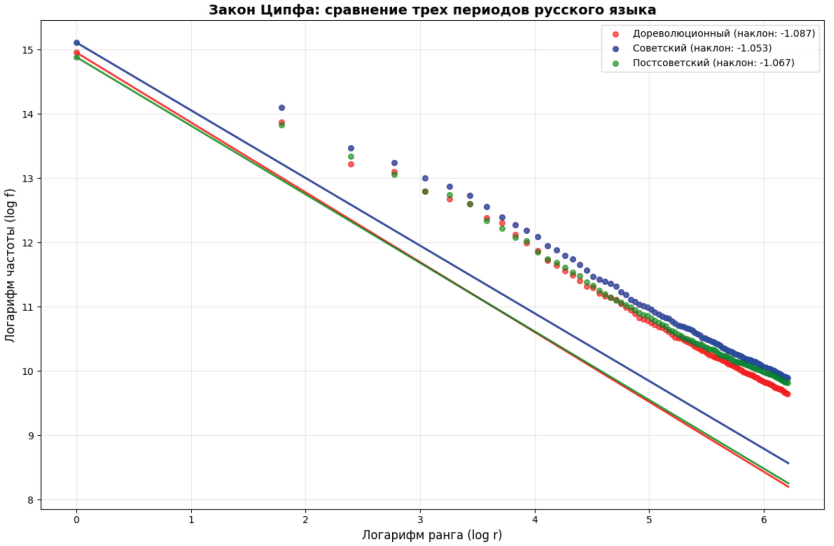

График 1

Все три периода демонстрируют соответствие закону Ципфа: линии практически идеально прямые; точки плотно прилегают к линиям тренда; наклон ≈ -1.0 во трех корпусах. $R^2 > 0.98$ свидетельствует об очень высокой линейной аппроксимации. $R^2 > 0.98$ для всех периодов также указывает на то, что ядро языка остаётся статистически стабильным, несмотря на социальные потрясения и перемены рассматриваемых исторических периодов.

Тот факт, что в советский период наклон смягчается по сравнению с досоветским (-1.087 → -1.053) свидетельствует о снижении лексического разнообразия языка в советский период. Вероятно, это было связано с советской цензурой и общим стремлением к унификации в языке в этот период. В постсоветский период наклон немного увеличился (-1.053→-1.067), что говорит о расширении лексического запаса. Однако если оценивать общую тенденцию, то все же по сравнению с дореволюционным периодом разнообразие общей лексической системы языка снижается.

2. Топ-100 самых частотных слов по периодам

В данном разделе было подсчитано соотношение между частотностью и рангом слова для 100 самых частотных слов в каждом корпусе (см. график 2).

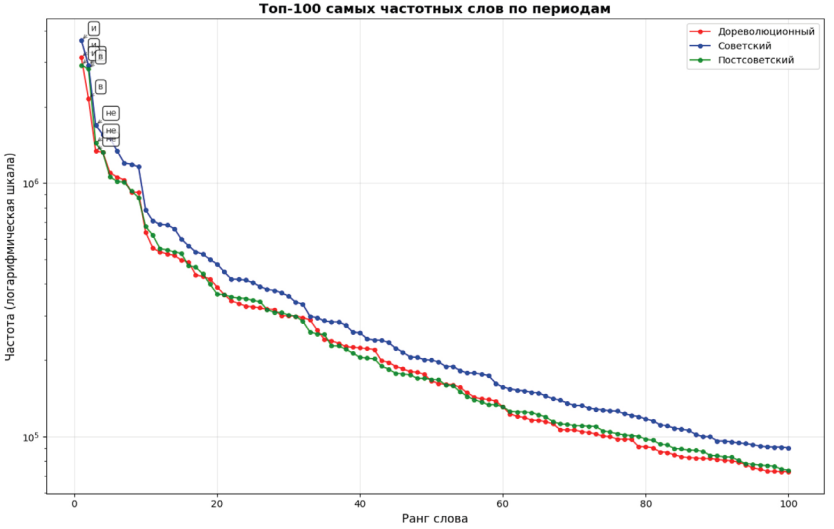

График 2

Как видно на графике, все три линии имеют одинаковую форму с резким спадом частоты употребления в начале и плавным снижением дальше. Самые резкие падения находятся между 1-10 рангами,

после 50-го ранга линии почти параллельны. Это типичная картина для распределения закона Ципфа.

Для советского корпуса (синяя линия) характерны самые высокие абсолютные частоты вхождения для топ-100 слов, у дореволюционного (красная линия) - средние позиции, у постсоветского (зеленая линия) - немного ниже дореволюционного.

Наибольшая частота у самых употребительных слов в советский период (пик частотности) свидетельствует о более низком лексическом разнообразии, о том, что авторы чаще обходились одним и тем же набором слов, что подтверждает выводы, сделанные в предыдущем пункте. Вероятными причинами этого является, как уже говорилось, стандартизация языка, цензура, требования идеологического однообразия текстов, характерные для советского периода.

Более низкие частоты в дореволюционном корпусе, напротив, говорят о большем лексическом разнообразии и менее идеологически и стилистически унифицированном корпусе письменных текстов. Постсоветский корпус занимает среднее положение между советским и дореволюционным, что также, как и в предыдущим пункте, говорит об увеличении лексического многообразия по сравнению с советским периодом, но все же это многообразие не достигает уровня дореволюционного периода.

Таким образом, график частотности топ-100 самых частотных слов подтверждает выводы, сделанные в предыдущем разделе, и также показывает снижение общего лексического многообразия в корпусах.

3. Коэффициент лексического разнообразия (TTR)

Коэффициент лексического разнообразия (Type-Token Ratio) – еще один коэффициент, измеряющий лексическое богатство словаря. Этот индекс рассчитывается довольно просто: количество уникальных лемм в корпусе делится на общий объем слов и умножается на 100%. [19, с. 16-20].

Высокий TTR говорит о том, что в корпусе много уникальных слов и мало повторений, и, соответственно, о более сложном или богатом языке текста. Низкий TTR свидетельствует о частом повторении слов и меньшем лексическом разнообразии. TTR 0.3-0.6% го-

ворит о среднем разнообразии, более низкие или высокие значения оцениваются соответственно как более низкое или более высокое многообразие.

Подсчеты TTR в исследуемых корпусах показали следующие результаты (см. график 3).

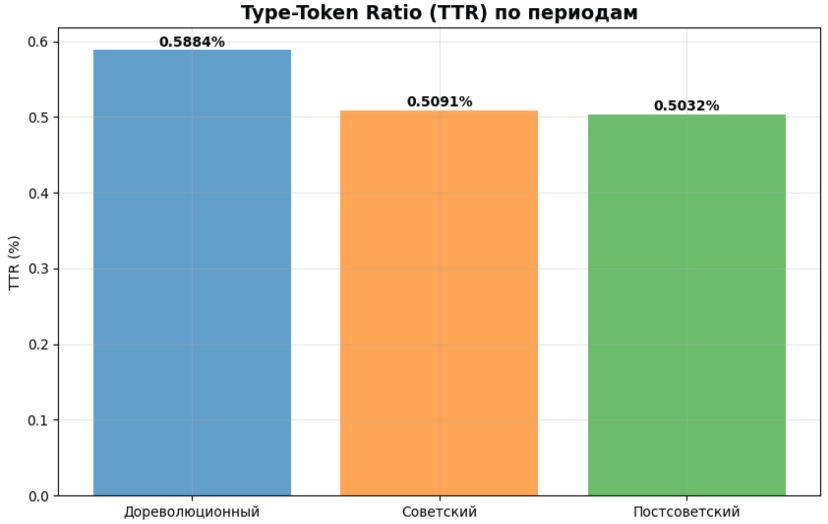

График 3

Все периоды попадают в категорию «среднее разнообразие» (0.3-0.6%), но дореволюционный период ближе к «высокому разнообразию» (>0.6%)

TTR последовательно снижается от дореволюционного к постсоветскому периоду (дореволюционный 0.5884% → советский 0.5091% → постсоветский 0.5032%), что свидетельствует о снижении разнообразия и общем обеднении словаря.

Следует отметить, что TTR не самый надежный показатель и должен учитываться только вместе с другими подсчетами. Проблема TTR в том, что он зависим от размера корпуса: чем больше корпус, тем больше повторений слов в нем будет, поэтому самый большой корпус (в данном случае советский) автоматически по-

лучает более низкий TTR (то есть снижение TTR не обязательно означает обеднение языка, а может отражать рост объёма текстов).

Однако согласованность динамики TTR с результатами, полученными по объемно-независимым метрикам, указанным в предыдущем разделе, позволяют предположить, что выявленное снижение TTR отражает не только эффект масштабирования выборки, но и общую тенденцию к снижению разнообразия за счет сокращения пласта редкой лексики.

4. Индекс энтропии Шеннона

Индекс энтропии Шеннона в корпусной лингвистике измеряет степень неопределённости и разнообразия в распределении слов по частотам: чем больше энтропия, тем менее предсказуем выбор слова, тем ближе распределение к случайному. Если распределение слов очень равномерное – много разных слов с примерно одинаковыми частотами – энтропия будет высокой, что отражает высокую степень неопределённости (вариативности) при выборе следующего слова в тексте и, соответственно, более богатый и менее клишированный лексический состав [20].

Абсолютная энтропия показывает фактический уровень неопределённости или разнообразия в данных – это «объём информации» в корпусе. Чем она выше, тем больше разнообразие и неопределенность. Относительная энтропия (нормированная) показывает, насколько фактическая энтропия близка к максимальной, насколько близко реальное распределение к полной случайности и равномерности. Она отражает, насколько распределение слов в корпусе приближено к равномерному [20] (см. график 4).

Мы видим рост от дореволюционного периода к современному абсолютной энтропии (11.19 → 11.44 → 11.56), что демонстрирует увеличение количества информации и непредсказуемости в распределении слов. Относительная энтропия тоже растет (0.598 → 0.606 → 0.620), указывая на растущую равномерность распределения частот слов.

Таким образом, несмотря на то что предыдущие расчеты показывали снижение лексического разнообразия (TTR 0.588% → 0.503%), энтропия показывает рост информационной насыщенности (11.19 → 11.56).

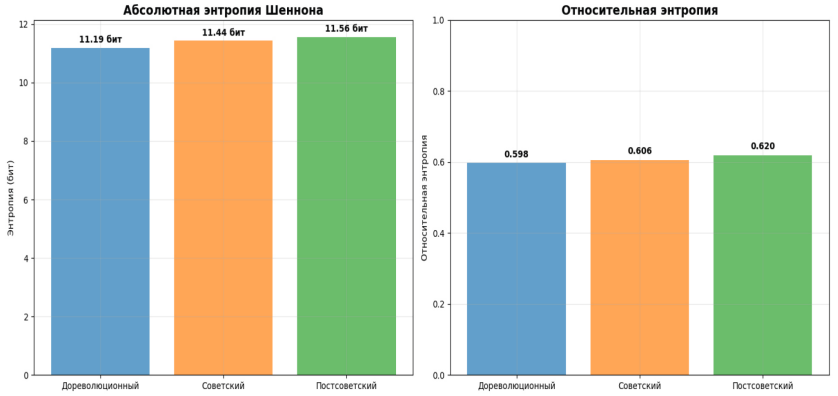

График 4

Это кажущееся противоречие объясняется тем, что TTR отражает лексическое богатство словаря, а энтропия – эффективность его использования, т.е. словарь становится менее разнообразным, но зато используется более эффективно, становится более функциональным, информационно насыщенным. Можно сказать, что в современный период мы передаем больше информации меньшими лексическими средствами.

5. Индекс Херфиндаля-Хиршмана

Индекс Херфиндаля-Хиршмана (HHI) в корпусной лингвистике используется для оценки концентрации активного словарного запаса, то есть для измерения того, насколько частотное распределение слов сосредоточено вокруг небольшого ядра наиболее употребляемых лексем или, напротив, равномерно распределено между множеством слов. Чем выше значение HHI, тем сильнее концентрация частотности у небольшого числа слов и, соответственно, меньше общее лексическое разнообразие. Более низкие значения индекса свидетельствуют о более равномерном распределении и большем разнообразии лексики [13, с. 1-3]. Для дополнительной интерпретации результатов HHI и наглядности также были рассчитаны совокупные доли, приходящиеся на 10 и 100 самых частотных слов каждого корпуса (см. график 5).

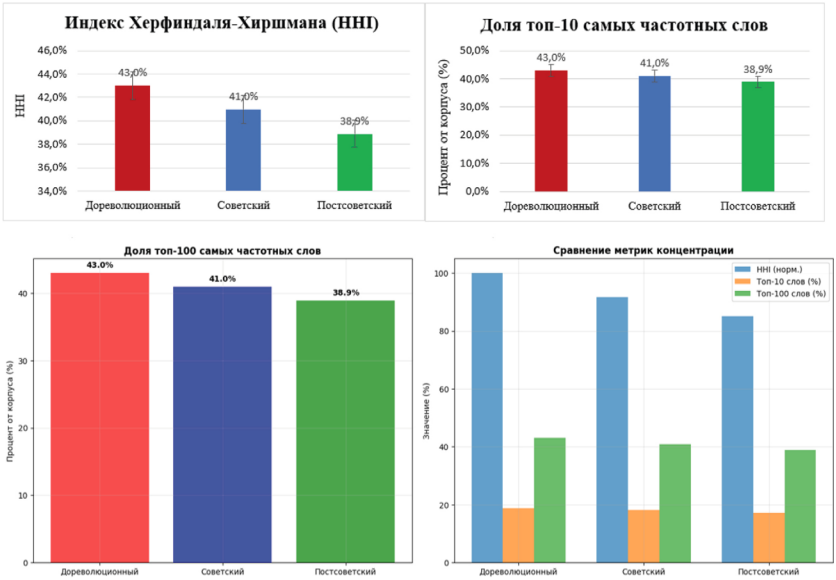

График 5

С точки зрения концентрации словаря все корпуса попадают в категорию «Разнообразный словарь» (HHI 50-100), при этом постсоветский период ближе всего к «Очень разнообразному» (<50).

С точки зрения эволюции лексического разнообразия прослеживается четкая тенденция падения индекса HHI от дореволюционного (54.70) к советскому (50.16) и постсоветскому корпусу (46.54). Это свидетельствует о том, что активный словарь последовательно становится более разнообразным со временем. Концентрация активного словаря вокруг самых частотных слов снижается на 15%, что является довольно значимым изменением и также говорит о более свободном распределении активного словарного запаса.

С точки зрения распределения частотности среди топ-10 слов (дореволюционный 18.70% → советский18.09% → постсоветский 17.14%) мы видим уменьшение доли топ-10 в корпусе на 1.56%, что говорит о снижении зависимости от самых частотных слов и расширении активного словаря.

Распределение частотности среди топ-100 слов (дореволюционный 43.00% → советский 40.96% → постсоветский 38.91%) показывает ту же самую тенденцию: уменьшение доли топ-100 на 4.09%, что также свидетельствует о значимом расширении активного словаря.

Таким образом, от дореволюционного к постсоветскому периоду «ядро» самых частотных слов становится менее сконцентрированным, более разнообразным, что говорит об увеличении лексического разнообразия в пределах активного словаря.

6. Индексы Симпсона и Бергера-Паркера

Индекс Симпсона рассчитывается как сумма квадратов относительных частот каждого слова (типа) в тексте и используется для оценки степени лексического богатства и выявления доминирования отдельных слов в корпусе. Чем выше индекс Симпсона, тем больше доминирование нескольких слов и меньше разнообразие.

Интерпретация индекса Симпсона:

• D = 0.0 - максимальное разнообразие (все слова с одинаковой частотой);

• D = 1.0 - максимальная концентрация (весь корпус состоит из повторений одного слова);

• 1-D = разнообразие (чем ближе к 1, тем более разнообразен словарь).

Индекс Бергера-Паркера измеряет доминирование наиболее частотного слова и высчитывается как доля самого частотного слова к общему числу слов в корпусе [15].

Для дореволюционного, советского и постсоветского корпусов были получены следующие результаты (см. график 6).

Таким образом, индекс Симпсона (D) показывает снижение концентрации слов вокруг частотного «ядра» от дореволюционного корпуса к постсоветскому: 0.005470 → 0.005016 → 0.004654.

Индекс Симпсона (1-D), наоборот, показывает повышение разнообразия слов от дореволюционного к постсоветскому корпусу: 0.994530 → 0.994984 → 0.995346.

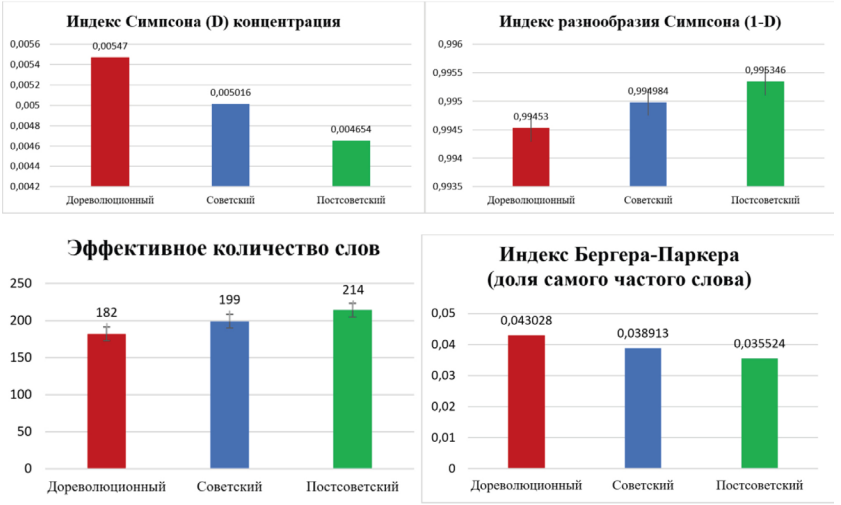

График 6

Дополнительно к индексу Симпсона был рассчитан показатель эффективного количества слов [16, с. 363-375]. Рост этого показателя от 183 до 215 означает, что современный русский язык стал более сбалансированным – редкие слова стали использоваться чаще, а частотные – реже.

Индекс Бергера-Паркера показывает, что доля самого частотного слова снижается от корпуса к корпусу: 0, 0430 → 0,0389 → 0,0355.

Можно сделать вывод, что в дореволюционный период наблюдается наибольшая концентрация вокруг частотных слов. Это говорит о более иерархической структуре словаря и наличии классического «ядра» частотной лексики.

В советский период начинается выравнивание распределения частот, расширение активного словаря и уменьшение зависимости от самых частых слов.

В постсоветский период представлена наименьшая концентрация частотности в высокоранговой лексике и наибольшее разнообразие в ее использовании.

Лексическая система эволюционирует от модели с выраженным доминированием ограниченного ядра высокочастотных слов к мо-

дели с более равномерным и сбалансированным распределением частотности по всему словарному составу.

7. Статистические значимости (хи-квадрат)

Для выявления слов, частота употребления которых претерпела статистически значимые изменения между историческими периодами, был применен критерий согласия хи-квадрат (χ^2) [15] (см. график 7).

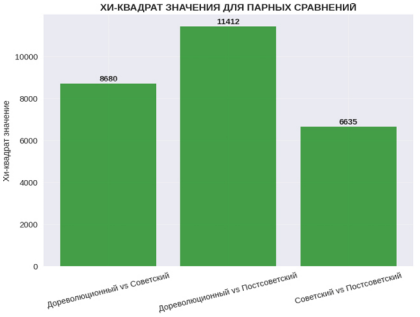

График 7

Наибольшие различия демонстрируют дореволюционный и постсоветский корпуса (11,411). Этот результат ожидаем, потому что между дореволюционным и постсоветским периодами прошло больше всего времени. Пара «дореволюционный и советский корпуса» показывают средние значения (8,680). Наименьшие различия обнаруживают «советский и постсоветский корпуса» (6,635).

Таким образом, если сравнивать пары «дореволюционный и советский» и «советский и постсоветский» корпуса, мы можем увидеть, что наибольшие изменения в лексической системе языка наблюдаются в первой паре, то есть в тот период происходил более значительный эволюционный сдвиг в лексике русского языка.

Обсуждение

Полученные в настоящем исследовании результаты о снижении общего лексического разнообразия при одновременном росте функциональной насыщенности активного словаря требуют осмысления

в контексте существующих корпусных исследований как русского, так и других языков. Несмотря на распространённость количественных методов в современной лингвистике, комплексный анализ эволюции лексического богатства на материале масштабных диахронических корпусов русского языка до сих пор не проводился.

Существующие исследования на материале русского языка демонстрируют методологическое сходство с настоящей работой, но существенно отличаются по целям, охвату и набору анализируемых показателей.

Так, есть две работы, которые используют те же диахронические корпусы НКРЯ [14; 17], однако в них фокус смещён на выявление изменений в значениях отдельных слов с помощью векторных моделей (word embeddings), а не на выявление глобальных изменений словарного состава. Хотя эти исследования в целом подтверждают факт значимой лексической перестройки на рубежах эпох, что согласуется с нашими выводами, полученными по критерию χ^2, однако они не дают количественной оценки общего лексического богатства словаря.

Те исследования, которые нацелены именно на измерение лексического разнообразия в русском языке, ограничиваются только определенными областями. Например, работа, посвященная анализу лексической сложности документов Конституционного Суда РФ [10], использует широкий набор индексов лексического разнообразия (коэффициент лексического разнообразия (TTR), его скорректированную версию (CTTR), характеристику Юла K (Yule's K) и др.), но сосредоточена на одном жанре и коротком временном отрезке (1992–2018 гг.). Исследование А. Пиперского [21], посвящённое лексическому разнообразию русской поэзии, включает схожие с нашими индексы (стандартизированный TTR, индекс Симпсона) и фиксирует рост разнообразия и выравнивание активного словаря с XVIII века, что можно рассматривать как раннее проявление в художественном регистре общей тенденции, которую мы наблюдаем для языка в целом. Исследования узких лексических подсистем (например, цветообозначений в поэзии, [18]) также подтверждают

ценность диахронического подхода с нормализацией частот и использованием статистических критериев, однако они не ставят целью описание эволюции всей лексической системы.

Таким образом, хотя в научных работах по корпусному анализу современного русского языка имеются работы, применяющие отдельные элементы нашей методологии, настоящее исследование является первым, которое на едином материале трёх крупных исторических периодов (1700–2016 гг., 250 млн словоупотреблений) реализует комплексный подход, включающий одновременный расчёт и интерпретацию закона Ципфа (наклон, R^2), индексов концентрации (Херфиндаля-Хиршмана, Симпсона, Бергера-Паркера), энтропии Шеннона, соотношения типов и токенов (TTR) и статистической значимости хи-квадрат (χ^2).

В свою очередь, межъязыковые параллели демонстрируют, что основные выявленные нами тренды не являются уникальными для русского языка, а вписываются в общую картину эволюции лексических систем. Исследование корпусов английского, испанского и турецкого языков [23], выявляет схожее фундаментальное соотношение: при увеличении объёма текстов и усложнении коммуникации энтропия Шеннона растёт, в то время как соотношение TTR снижается. Этот вывод полностью соответствует нашему ключевому наблюдению о парадоксальном сочетании снижения TTR с ростом энтропии от дореволюционного к постсоветскому периоду, что свидетельствует о росте информационной насыщенности единицы текста.

Анализ динамики энтропии «ядра» словаря по данным Google Books Ngram для семи языков [12], показывает специфичную для русского языка траекторию с резким падением показателей около 1920-х годов и подъёмом в конце XX века. Эта картина качественно подтверждает нашу интерпретацию советского периода как эпохи унификации и снижения лексического разнообразия (смягчение наклона Ципфа, пик частотности топ-слов), за которой последовало выравнивание распределения в постсоветское время.

Похожие результаты выявляются в исследованиях классического китайского языка, где фиксируется устойчивый рост лексической

сложности при одновременном упрощении синтаксических структур, что трактуется как повышение информационной эффективности коммуникации [25].

Таким образом, проведённое исследование методологически соотносится с международными исследованиями крупных корпусов различных языков. При этом для русского языка оно является новым по своему масштабу (охват более трёх столетий) и комплексности используемого аналитического аппарата.

Заключение

1. Корпусный анализ диахронических датасетов 1700-1916, 1918-1991 и 1992-2016 г. показал, что общее разнообразие и богатство лексического состава русского языка снижаются от дореволюционного к постсоветскому периоду.

2. Динамика лексической системы носит дифференцированный характер. Общее снижение разнообразия (TTR) происходит на периферии словаря — за счёт выхода из употребления редких слов. При этом центральная часть (активное ядро), наоборот, становится более разнообразной и равномерно используемой, что подтверждается ростом эффективного числа слов и снижением доли самых частотных лексем (топ-10, топ-100).

3. Для дореволюционного периода характерна наиболее высокая концентрация частотности вокруг узкого ядра высокоранговых лексем (максимальные значения индексов HHI, Симпсона и доли самого частотного слова – индекс Бергера-Паркера). Это указывает на более иерархическую структуру словарного состава с чётким разделением на малое сверхчастотное ядро и обширный низкочастотный периферийный слой.

4. Эволюция лексической системы русского языка характеризуется повышением её информационной эффективности: несмотря на сокращение общего числа лемм и доли редкой лексики, активное ядро словаря расширяется и становится более сбалансированным (снижение индексов HHI, Симпсона и Бергера-Паркера), а его способность кодировать информацию растет (увеличение энтропии

Шеннона). Это указывает на тенденцию к передаче большего объема информации менее разнообразным, но более эффективно используемым набором слов.

Благодарности. Выражаем благодарность программисту Евгению Сергеевичу Столетову за компьютерную обработку диахронических датасетов и лингвисту-аналитику Валерии Михайловне Запорожцевой за помощь с расчетами статистических данных.

Список литературы

1. Завьялова, И. С., & Шерстинова, Т. Ю. (2022). О морфологических различиях в текстах русской малой прозы 1900–1930 гг. *Человек: Образ и сущность. Гуманитарные аспекты*, (2), 176–204. https://doi.org/10.31249/chel/2022.02.12. EDN: https://elibrary.ru/OEIGOJ
2. Комарькова, М. А. (2021). Тенденции лингвистических изменений в современном английском языке. *Современное педагогическое образование*, (6), 153–155. EDN: https://elibrary.ru/DNEYUT
3. Мартыненко, Г. Я., Шерстинова, Т. Ю., Попова, Т. И., Мельник, А. Г., & Замирайлова, Е. В. (2018). О принципах создания корпуса русского рассказа первой трети XX века. В кн.: *Труды международной конференции по компьютерной и когнитивной лингвистике* (с. 180–197). EDN: https://elibrary.ru/YFFGSO
4. Соловьёв, В. Д. (2012). Статистические методы анализа диахронических корпусов текстов как инструмент исследования языковой динамики. В кн.: *Материалы международной конференции «Русский язык: функционирование и развитие»* (с. 47). Казань: Казанский университет.
5. Черкасова, Г. А. (2015). Сопоставительные исследования коэффициентов «Лексического разнообразия» и «Лексического богатства» Ю. Н. Караулова на материале русских ассоциативных словарей. *Вопросы психолингвистики*, (25), 93–104. EDN: https://elibrary.ru/UDLHEJ
6. Шерстинова, Т. Ю., & Завьялова, И. С. (2022). Динамика дистрибуции частеречных и грамматических категорий в русском рассказе 1900–1930. В кн.: *Русская грамматика в диалоге научных школ, направлений, методов* (с. 324). EDN: https://elibrary.ru/LLVVYK
7. Шерстинова, Т. Ю. (2021). Русская литература 1900–1930: что изменилось в языке и стиле после Октябрьской революции? В кн.: *Второй российский эстетический конгресс* (с. 622–624). EDN: https://elibrary.ru/PZGGQT
8. Шерстинова, Т. Ю., Колпащикова, Е. О., Сейнова, А. Р., Максименко, П. И., & Родионов, Р. А. (2023). Русский рассказ 1900–1930-х и его восприятие чи-

тателем: опыт квантитативного анализа оценки художественного текста. *Человек: Образ и сущность. Гуманитарные аспекты*, (2), 164–184. https://doi.org/10.31249/chel/2023.02.09. EDN: https://elibrary.ru/GZYNIO

9. Юлдашева, Л. У. (2023). Исследование лексического массива русского языка: вопросы сохранения и потери слов в современной эпохе. *Journal of Multidisciplinary Bulletin, 6*(5), 458–466.
10. Blinova, O. V., Belov, S., & Revazov, M. A. (2021). Decisions of Russian constitutional court: lexical complexity analysis in shallow diachrony. В кн.: *CEUR Workshop Proceedings* (с. 61–74).
11. Bochkarev, V. V., Solovyev, V. D., Nestik, T. A., & Shevlyakova, A. V. (2024). Variations in average word valence of Russian books over a century and social change. *Journal of Mathematical Sciences, 285*(1), 14–27. https://doi.org/10.1007/s10958-024-07419-z. EDN: https://elibrary.ru/QYDSPS
12. Buntinx, V., & Kaplan, F. (2018). Negentropic linguistic evolution: A comparison of seven languages. В кн.: *Digital Humanities 2018: Book of Abstracts / Libro de resúmenes*.
13. Dunn, J., Coupe, T., & Adams, B. (2020, November). Measuring linguistic diversity during COVID-19. В кн.: *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science* (с. 1–10). https://doi.org/10.18653/v1/2020.nlpcss-1.1
14. Fomin, V., Bakshandaeva, D., Rodina, Ju., & Kutuzov, A. (2019). Tracing cultural diachronic semantic shifts in Russian using word embeddings. В кн.: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"* (Moscow, May 29 – June 1, 2019). Получено из: https://arxiv.org/pdf/1905.06837
15. Gries, S. T. (2021). *Statistics for linguistics with R: A practical introduction* (3rd ed.). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110718256
16. Jost, L. (2006). Entropy and diversity. *Oikos, 113*(2), 363–375. https://doi.org/10.1111/j.2006.0030-1299
17. Kutuzov, A., & Kuzmenko, E. (2018). Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. В кн.: *Quantitative approaches to the Russian language* (с. 95–112). Routledge. https://doi.org/10.4324/9781315105048-5
18. Lyashevskaya, O., Vlasova, E., & Litvintseva, K. (2019). Lexical diversity and colour hues in Russian poetry: A corpus-based study of adjectives. В кн.: P. Plecháč, M. Skulacheva, & R. Piłś (Eds.), *Quantitative approaches to versification* (с. 131–141). Institute of Czech Literature of the Czech Academy of Sciences.
19. Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development* (с. 16–30). Palgrave Macmillan UK.

20. MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.
21. Piperski, A. (2023). Lexical diversity of Russian poets. В кн.: *Literature, language and computing: Russian contribution* (с. 113–120). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3604-5_10
22. Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6. EDN: https://elibrary.ru/SFDFMF
23. Rosillo-Rodes, P., San Miguel, M., & Sánchez, D. (2025). Entropy and type-token ratio in gigaword corpora. *Physical Review Research, 7*(3), 033054. https://doi.org/10.48550/arXiv.2411.10227. EDN: https://elibrary.ru/XQTDHY
24. Sherstinova, T., & Martynenko, G. (2019, November). Linguistic and stylistic parameters for the study of literary language in the corpus of Russian short stories of the first third of the 20th century. В кн.: *R. Piotrowski's Readings in Language Engineering and Applied Linguistics: Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019)* (Saint Petersburg, Russia, с. 105–120).
25. Song, J., & Lei, L. (2025). Lexical bloom, syntactic retreat: Examining complexity trade-offs within Classical Chinese evolution across two millennia. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2024-0125. EDN: https://elibrary.ru/ZMAJJX
26. Zipf, G. K. (1972). *Human behavior and the principle of least effort: An introduction to human ecology*. Hafner.

Список источников и словарей

27. *Национальный корпус русского языка (НКРЯ)* [Электронный ресурс]. (2003–2025). Скачиваемые корпуса. Получено 19.10.2025, из: https://ruscorpora.ru/page/corpora-datasets/
28. *Диахронический словарь русской лексики* [Электронный ресурс] / Казанский (Приволжский) федеральный университет, Институт филологии и межкультурной коммуникации. Получено 20.10.2025, из: https://kpfu.ru/philology-culture/struktura-instituta/nauchno-obrazovatelnye-centry-noc/noc-po-lingvistike-im-ia-boduena-de-kurtene/nil-39kvantitativnaya-lingvistika39/diahronicheskij-slovar.html
29. Засорина, Л. Н. (Ред.). (1977). *Частотный словарь русского языка: около 40 000 слов*. Москва: Русский язык.
30. Ляшевская, О. Н., & Шаров, С. А. (2009). *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)*. Москва: Азбуковник. Получено 20.10.2025, из: http://dict.ruslang.ru/freq.php

31. Штейфельдт, Э. А. (1963). *Частотный словарь современного русского литературного языка: 2500 наиболее употребительных слов: пособие для преподавателей русского языка*. Таллин: Издательство «Юхисэлу».
32. Lönngren, L. (1993). *The frequency dictionary of modern Russian*. Acta Univ. Ups., Studia Slavica Upsaliensia. Uppsala.
33. Josselson, H. (1953). *The Russian word count and frequency analysis of grammatical categories of standard literary Russian*.

References

1. Zav'yalova, I. S., & Sherstinova, T. Yu. (2022). On morphological differences in the texts of Russian short prose from 1900 to 1930. *Man: Image and Essence. Humanitarian Aspects*, *(2)*, 176–204. https://doi.org/10.31249/chel/2022.02.12. EDN: https://elibrary.ru/OEIGOJ
2. Komarkova, M. A. (2021). Trends in linguistic changes in modern English. *Modern Pedagogical Education*, *(6)*, 153–155. EDN: https://elibrary.ru/DNEYUT
3. Martynenko, G. Ya., Sherstinova, T. Yu., Popova, T. I., Melnik, A. G., & Zamirailova, E. V. (2018). On the principles of creating a corpus of Russian short stories from the first third of the 20th century. In: *Proceedings of the International Conference on Computational and Cognitive Linguistics* (pp. 180–197). EDN: https://elibrary.ru/YFFGSO
4. Solovyev, V. D. (2012). Statistical methods for analyzing diachronic text corpora as a tool for studying language dynamics. In: *Materials of the International Conference "Russian Language: Functioning and Development"* (p. 47). Kazan: Kazan University.
5. Cherkasova, G. A. (2015). Comparative studies of the coefficients of "Lexical Diversity" and "Lexical Richness" by Yu. N. Karaulov based on Russian associative dictionaries. *Journal of Psycholinguistics*, *(25)*, 93–104. EDN: https://elibrary.ru/UDLHEJ
6. Sherstinova, T. Yu., & Zav'yalova, I. S. (2022). Dynamics of distribution of part-of-speech and grammatical categories in Russian short stories of 1900–1930. In: *Russian Grammar in the Dialogue of Scientific Schools, Directions, and Methods* (p. 324). EDN: https://elibrary.ru/LLVVYK
7. Sherstinova, T. Yu. (2021). Russian literature of 1900–1930: what changed in language and style after the October Revolution? In: *Second Russian Aesthetic Congress* (pp. 622–624). EDN: https://elibrary.ru/PZGGQT
8. Sherstinova, T. Yu., Kolpashchikova, E. O., Seinova, A. R., Maksimenko, P. I., & Rodionov, R. A. (2023). Russian short story of 1900–1930 and its reader perception: an experience of quantitative analysis of literary text evaluation. *Man: Image and Essence. Humanitarian Aspects*, *(2)*, 164–184. https://doi.org/10.31249/chel/2023.02.09. EDN: https://elibrary.ru/GZYNIO

9. Yuldasheva, L. U. (2023). Studying the lexical array of the Russian language: issues of preserving and losing words in the modern era. *Journal of Multidisciplinary Bulletin*, *6*(5), 458–466.
10. Blinova, O. V., Belov, S., & Revazov, M. A. (2021). Decisions of Russian Constitutional Court: lexical complexity analysis in shallow diachrony. In: *CEUR Workshop Proceedings* (pp. 61–74).
11. Bochkarev, V. V., Solovyev, V. D., Nestik, T. A., & Shevlyakova, A. V. (2024). Variations in average word valence of Russian books over a century and social change. *Journal of Mathematical Sciences*, *285*(1), 14–27. https://doi.org/10.1007/s10958-024-07419-z. EDN: https://elibrary.ru/QYDSPS
12. Buntinx, V., & Kaplan, F. (2018). Negentropic linguistic evolution: a comparison of seven languages. In: *Digital Humanities 2018: Book of Abstracts / Libro de resúmenes*.
13. Dunn, J., Coupe, T., & Adams, B. (2020, November). Measuring linguistic diversity during COVID-19. In: *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science* (pp. 1–10). https://doi.org/10.18653/v1/2020.nlpcss-1.1
14. Fomin, V., Bakshandaeva, D., Rodina, Ju., & Kutuzov, A. (2019). Tracing cultural diachronic semantic shifts in Russian using word embeddings. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"* (Moscow, May 29 – June 1, 2019). Получено из: https://arxiv.org/pdf/1905.06837
15. Gries, S. T. (2021). *Statistics for linguistics with R: A practical introduction* (3rd ed.). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110718256
16. Jost, L. (2006). Entropy and diversity. *Oikos, 113*(2), 363–375. https://doi.org/10.1111/j.2006.0030-1299
17. Kutuzov, A., & Kuzmenko, E. (2018). Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. In: *Quantitative approaches to the Russian language* (pp. 95–112). Routledge. https://doi.org/10.4324/9781315105048-5
18. Lyashevskaya, O., Vlasova, E., & Litvintseva, K. (2019). Lexical diversity and colour hues in Russian poetry: A corpus-based study of adjectives. In: P. Plecháč, M. Skulacheva, & R. Piłś (Eds.), *Quantitative approaches to versification* (pp. 131–141). Institute of Czech Literature of the Czech Academy of Sciences.
19. Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development* (pp. 16–30). Palgrave Macmillan UK.
20. MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

21. Piperski, A. (2023). Lexical diversity of Russian poets. In: *Literature, language and computing: Russian contribution* (pp. 113–120). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3604-5_10
22. Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6. EDN: https://elibrary.ru/SFDFMF
23. Rosillo-Rodes, P., San Miguel, M., & Sánchez, D. (2025). Entropy and type-token ratio in gigaword corpora. *Physical Review Research, 7*(3), 033054. https://doi.org/10.48550/arXiv.2411.10227. EDN: https://elibrary.ru/XQTDHY
24. Sherstinova, T., & Martynenko, G. (2019, November). Linguistic and stylistic parameters for the study of literary language in the corpus of Russian short stories of the first third of the 20th century. In: *R. Piotrowski's Readings in Language Engineering and Applied Linguistics: Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019)* (Saint Petersburg, Russia, pp. 105–120).
25. Song, J., & Lei, L. (2025). Lexical bloom, syntactic retreat: Examining complexity trade-offs within Classical Chinese evolution across two millennia. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2024-0125. EDN: https://elibrary.ru/ZMAJJX
26. Zipf, G. K. (1972). *Human behavior and the principle of least effort: An introduction to human ecology*. Hafner.

Sources and dictionaries

27. National Corpus of the Russian Language (NCRL) [Electronic resource]. (2003–2025). Downloadable corpora. Retrieved on October 19, 2025, from: https://ruscorpora.ru/page/corpora-datasets/
28. Diachronic dictionary of Russian vocabulary [Electronic resource] / Kazan (Volga Region) Federal University, Institute of Philology and Intercultural Communication. Retrieved on October 20, 2025, from: https://kpfu.ru/philology-culture/struktura-instituta/nauchno-obrazovatelnye-centry-noc/noc-po-lingvistike-im-ia-boduena-dekurtene/nil-39kvantitativnaya-lingvistika39/diahronicheskij-slovar.html
29. Zasorina, L. N. (Ed.). (1977). *Frequency dictionary of the Russian language: about 40 000 words*. Moscow: Russkiy Yazyk.
30. Lyashevskaya, O. N., & Sharov, S. A. (2009). *Frequency dictionary of modern Russian (based on materials from the National Corpus of the Russian Language)*. Moscow: Azbukovnik. Retrieved on October 20, 2025, from: http://dict.ruslang.ru/freq.php
31. Shteyfeldt, E. A. (1963). *Frequency dictionary of modern standard Russian literature: 2500 most common words: a guide for Russian language teachers*. Tallinn: Yuhiselu Publishing House.

32. Lönngren, L. (1993). *The frequency dictionary of modern Russian*. Acta Univ. Ups., Studia Slavica Upsaliensia. Uppsala.
33. Josselson, H. (1953). *The Russian word count and frequency analysis of grammatical categories of standard literary Russian*.

ДАННЫЕ ОБ АВТОРЕ

Рычкова Татьяна Александровна, кандидат филологических наук, доцент, доцент кафедры филологии, межкультурных коммуникаций и журналистики
Мурманский арктический университет
ул. Коммуны, 9, г. Мурманск, 183038, Российская Федерация
rychkovata@yandex.ru

DATA ABOUT THE AUTHOR

Tatiana A. Rychkova, PhD in Philology, Associate Professor, Associate Professor of the Department of Philology, Intercultural Communications and Journalism
Murmansk Arctic University
9, Kommuny Str., Murmansk, 183038, Russian Federation
rychkovata@yandex.ru
SPIN-code: 8157-7942
ORCID: https://orcid.org/0000-0002-0342-1308
Researcher ID: ABA-4880-2021
Scopus AuthorID: 57 215 913 998
ResearchGate:https://www.researchgate.net/profile/Tatiana-Rychkova?ev=hdr_xprf

Поступила 09.12.2025
После рецензирования 25.01.2026
Принята 12.02.2026

Received 09.12.2025
Revised 25.01.2026
Accepted 12.02.2026