

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Факультет компьютерных наук

Магистерская программа «Исследования и
предпринимательство в искусственном интеллекте»

**КОРПУСНЫЙ АНАЛИЗ АРХАИЧНОЙ
ЛЕКСИКИ В РУССКОЙ ПОЭЗИИ 18-20
ВЕКОВ**

Исследовательская работа

Выполнил:
студент 2 курса
Пеганов Никита

Москва 2025

Введение

Актуальность. Изучение динамики языковых изменений в художественной литературе является важной задачей исторической лексикологии и корпусной лингвистики. Архаизмы – устаревшие слова и выражения, вытесненные из активного употребления синонимами, – представляют особый интерес как маркеры стилистических и эстетических установок литературных эпох. Процесс архаизации поэтической лексики отражает не только естественную эволюцию языка, но и сознательный выбор авторов, их отношение к традиции и поиск выразительных средств.

Количественный анализ частотности архаизмов в диахронии позволяет выявить закономерности их использования, связанные со сменой литературных направлений, и проверить гипотезы о корреляции между эстетическими парадигмами и лексическим составом текстов.

Цель работы – проанализировать динамику использования архаичной и устаревшей лексики в русской поэзии XVIII–XX веков на материале корпуса из 16 694 стихотворений.

Задачи исследования:

1. Подготовить корпус стихотворений с метаданными (автор, год написания).
2. Составить словарь архаизмов на основе лексикографических источников.
3. Вычислить частотность архаизмов по десятилетиям и литературным течениям.
4. Выявить периоды интенсивной архаизации и интерпретировать полученные результаты.
5. Установить корреляцию между использованием архаизмов и литературными направлениями.

Материал исследования – корпус русской поэзии, включающий 16 694 стихотворения 195 авторов, созданных в период с 1725 по 1996 год. Общий объем корпуса составил 1 726 105 словоупотреблений. Корпус был получен из открытого проекта «Поэтический корпус русского языка» (Гусев 2016), доступного на платформе GitHub. Исходные данные представлены в формате XML и включают метаданные: имя автора, название стихотворения, даты создания, тематические теги.

Методы исследования. В работе применялись методы корпусной лингвистики и квантиitatивного анализа текстов. На первом этапе был выполнен парсинг XML-файла и конвертация данных в табличный формат

(CSV, Parquet). Токенизация текстов проводилась с помощью регулярных выражений для извлечения русских слов.

Словарь архаизмов был составлен на основе «Словаря устаревших слов» с портала «Азбука веры» (1 188 слов), дополнен 18 лексемами, выявленными при визуальном анализе облаков слов, построенных по частотности для XVIII, XIX и XX веков. Для вычисления относительной частотности использовалась метрика «количество архаизмов на 1000 словоупотреблений».

Периодизация материала осуществлялась по десятилетиям и литературным течениям (классицизм, сентиментализм, романтизм, реализм, символизм, акмеизм, футуризм, советская поэзия). Визуализация результатов выполнена с использованием библиотек Python (matplotlib, seaborn, wordcloud).

Основная часть

Подготовка данных

Исходный XML-файл был распарсен с извлечением полей: автор, название, текст стихотворения, годы написания (`date_from`, `date_to`), темы. Для каждого стихотворения был вычислен средний год создания. После фильтрации по наличию даты в анализ вошло 12 857 текстов (77% корпуса), охватывающих период 1725–1996 годов.

Токенизация текстов производилась с выделением последовательностей русских букв длиной не менее 3 символов. Словарь архаизмов был загружен в виде множества лемм для быстрого поиска. Дополнительно был реализован скрипт для визуального анализа наиболее частотных и редких слов по векам, что позволило выявить архаизмы, отсутствовавшие в исходном словаре (например: *ныне*, *взор*, *глас*, *коль*, *старец*).

Результаты количественного анализа

Общая статистика. В корпусе обнаружено 14 922 употребления архаизмов, что составляет среднюю частотность 8,64 архаизма на 1000 словоупотреблений. Максимальная частотность зафиксирована в 1720-х годах – 27,69 на 1000 слов, минимальная – в 1990-х годах (5,92 на 1000). Общее снижение частотности за 270 лет составило 78,6%.

Динамика по десятилетиям. Анализ показал устойчивый тренд снижения частотности архаизмов с XVIII по XX век. Наибольшие значения характерны для периода 1720–1770 годов (от 27,69 до 14,86 на 1000 слов). К концу XIX века частотность снижается до 8–9 на 1000, в XX веке – до 6–7 на 1000 слов.

Важно отметить **циклический паттерн**: наблюдается локальный рост частотности архаизмов к концу каждой литературной эпохи. Пики зафиксированы в конце классицизма (1800-е), сентиментализма (1810-е), романтизма (1830-е) и реализма (1890-е). Этот паттерн отсутствует у модернистских течений (символизм, футуризм, советская поэзия), где наблюдается стабильное снижение без циклических колебаний.

Анализ по литературным течениям. Для каждого течения была вычислена средняя частотность архаизмов (таблица 1).

Таблица 1: Частотность архаизмов по литературным течениям

Литературное течение	Период	Частотность (на 1000 слов)
Классицизм	1730–1800	16,93
Сентиментализм	1770–1820	13,17
Романтизм	1800–1840	11,68
Реализм	1840–1890	8,70
Символизм	1890–1910	8,38
Акмеизм	1910–1920	7,70
Футуризм	1910–1930	7,53
Советская поэзия	1920–1990	6,62

Стандартное отклонение между течениями составило $\sigma = 3,55$, что свидетельствует о статистически значимом влиянии литературного направления на частотность архаизмов. Максимальная разница между классицизмом и советской поэзией достигает 10,31 единицы (разница в 2,6 раза).

Анализ по авторам. Среди топ-10 авторов по количеству стихотворений наибольшую частотность архаизмов демонстрируют представители романтизма: Михаил Лермонтов (12,81 на 1000 слов), Федор Тютчев (12,44), Александр Пушкин (11,42). Минимальные значения – у поэтов XX века: Анна Ахматова (5,48) и Владимир Высоцкий (5,51).

Визуализация результатов

На рисунке 1 представлена динамика частотности архаизмов по десятилетиям с 1720 по 2000 год. График демонстрирует общий нисходящий тренд с локальными пиками в конце каждой литературной эпохи до конца XIX века.

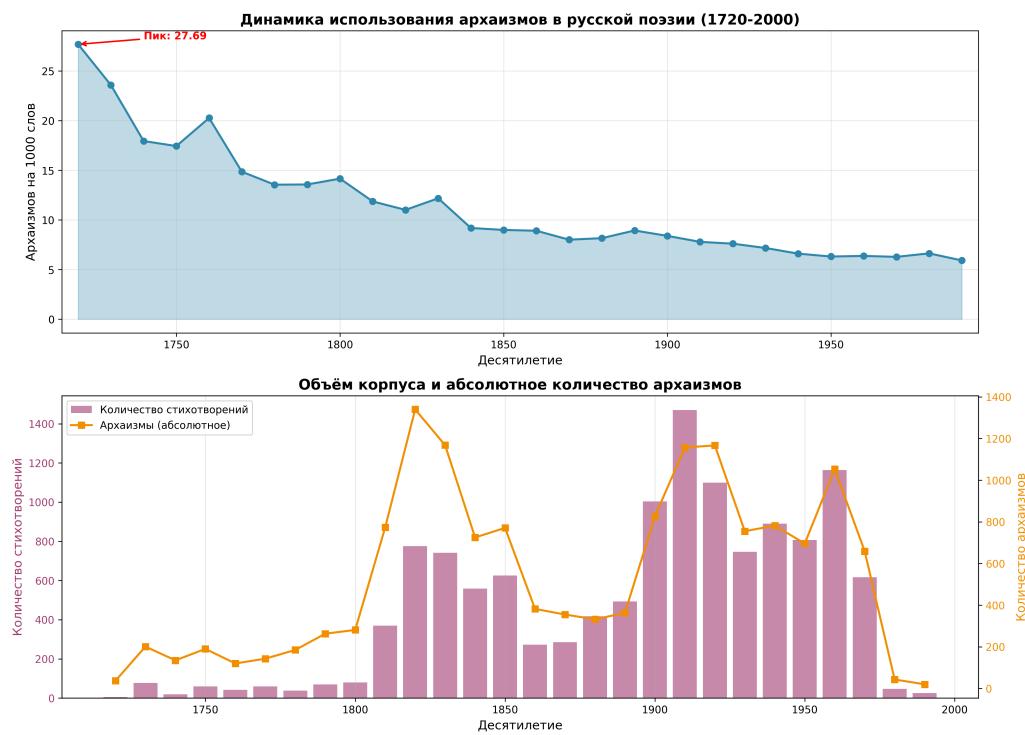


Рис. 1: Динамика частотности архаизмов по десятилетиям (1720–2000)

Рисунок 2 показывает сравнение средней частотности архаизмов по литературным течениям. Чётко видна градация от классицизма (максимум) к советской поэзии (минимум).

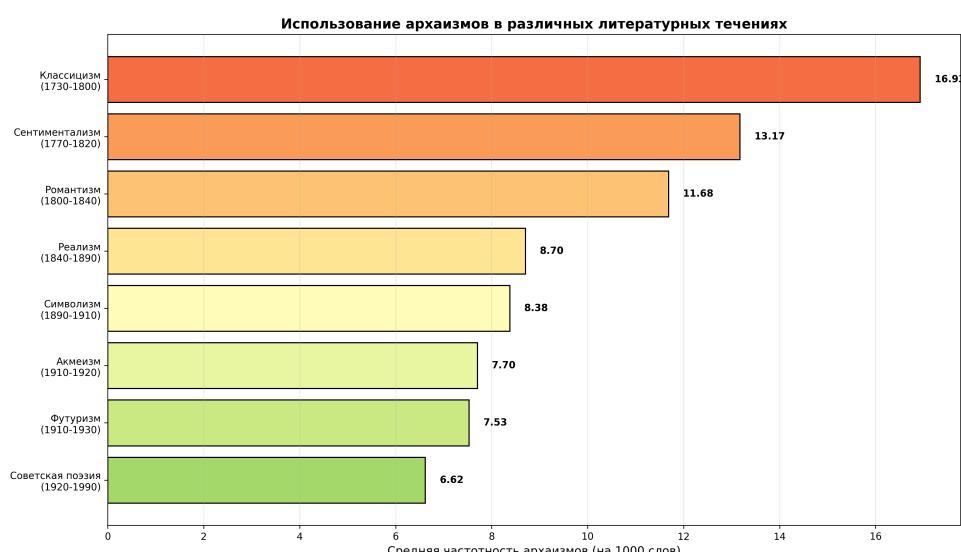


Рис. 2: Сравнение литературных течений по частотности архаизмов

На рисунке 3 представлена временная шкала, совмещающая динамику частотности с периодами литературных течений. Цветные полосы иллюстрируют хронологические рамки каждого направления.



Рис. 3: Временная шкала частотности архаизмов с литературными течениями

Рисунок 4 демонстрирует сравнение частотности архаизмов по векам (XVIII, XIX, XX). Заметно прогрессивное снижение показателя от 18,4 (XVIII век) до 7,0 (XX век).

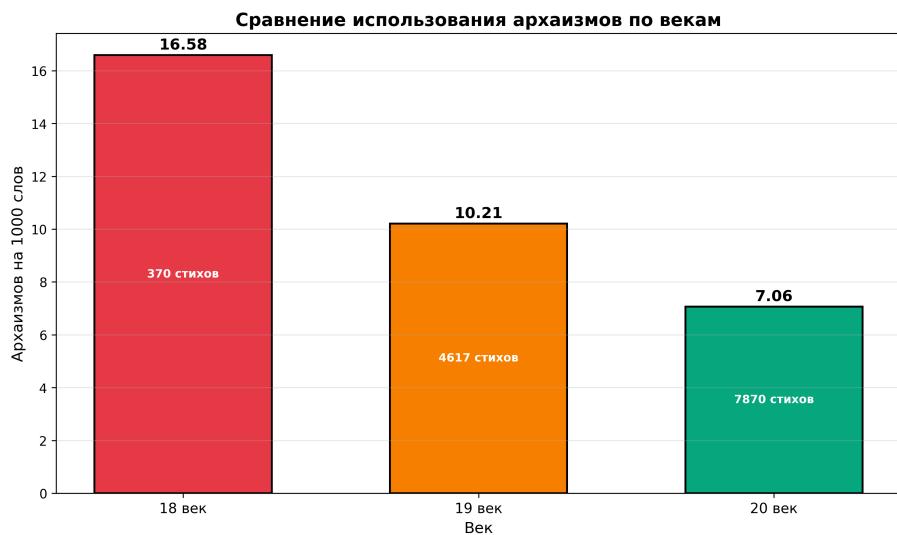


Рис. 4: Сравнение частотности архаизмов по векам

Интерпретация результатов

Полученные данные подтверждают гипотезу о снижении частотности архаизмов в русской поэзии с XVIII по XX век. Этот процесс обусловлен несколькими факторами.

Во-первых, естественное устаревание лексики. Часть слов, использовавшихся поэтами XVIII–XIX веков как нейтральная или современная лексика, к концу XX века перешла в разряд архаизмов в результате естественной эволюции языка. Таким образом, рост доли архаизмов в более ранних текстах отчасти является артефактом исторической дистанции: чем старше текст, тем больше вероятность, что его лексика устарела к моменту анализа (конец XX – начало XXI века).

Во-вторых, эволюция литературного языка в сторону сближения с разговорной нормой и отказ от церковнославянской книжной традиции. Классицизм, ориентированный на античные образцы и высокий стиль, активно использовал старославянизмы как маркеры торжественности. Романтизм сохранил интерес к архаике через обращение к исторической и фольклорной тематике.

В-третьих, смена эстетических парадигм в конце XIX – начале XX века. Модернистские течения (символизм, футуризм) провозгласили разрыв с традицией, что привело к минимизации использования устаревшей лексики. Советская поэзия продолжила этот тренд, тяготея к современному урбанизированному языку.

В-четвертых, циклический паттерн роста архаизмов к концу каждой эпохи может быть интерпретирован как «прощание с традицией» перед сменой литературной парадигмы. Поэты переходного периода обращались к архаичной лексике, чтобы подчеркнуть связь с уходящей эпохой и её эстетическими идеалами. Отсутствие этого паттерна у модернистов указывает на более радикальный характер разрыва с прошлым.

Корреляция между частотностью архаизмов и литературным течением ($\sigma = 3,55$) свидетельствует о том, что архаизация была не пассивным следом традиции, а сознательным **эстетическим выбором**, соответствующим идеологии и стилистике направления.

Заключение

В ходе исследования был проведен корпусный анализ частотности архаизмов в русской поэзии XVIII–XX веков на материале 16 694 стихотворений (1 726 105 словоупотреблений). Основные результаты работы:

1. Установлено устойчивое снижение частотности архаизмов на 78,6% за период 1720–1990 годов (с 27,69 до 5,92 на 1000 слов).

2. Выявлена статистически значимая корреляция между литературным течением и уровнем архаизации ($\sigma = 3,55$). Классицизм демонстрирует максимальную частотность (16,93 на 1000), советская поэзия – минимальную (6,62 на 1000).
3. Обнаружен циклический паттерн: локальный рост частотности архаизов к концу каждой литературной эпохи (классицизм, сентиментализм, романтизм, реализм). Этот паттерн отсутствует у модернистских течений, что указывает на принципиально иной характер отношения к языковой традиции.
4. Среди авторов наибольшую частотность архаизов демонстрируют представители романтизма: М. Лермонтов (12,81), Ф. Тютчев (12,44), А. Пушкин (11,42).

Полученные результаты позволяют утверждать, что использование архаизов в поэзии было не механическим наследованием традиции, а осознанным стилистическим приёмом, коррелирующим с эстетическими установками литературных направлений. Циклический паттерн архаизации свидетельствует о рефлексивном отношении поэтов к смене эпох.

Перспективы исследования включают расширение словаря архаизов с учётом морфологических форм, анализ контекстов употребления устаревшей лексики и сопоставление с прозаическими текстами.

Список литературы

1. Виноградов В. В. Очерки по истории русского литературного языка XVII–XIX веков. М.: Высшая школа, 1982.
2. Грановская Л. М. Русский литературный язык в конце XIX и XX вв. М.: Элпис, 2005.
3. Гусев И. О. Поэтический корпус русского языка // GitHub repository. 2016. URL: <https://github.com/IlyaGusev/PoetryCorpus> (дата обращения: 20.12.2025).
4. Словарь устаревших слов // Азбука веры. URL: <https://azbyka.ru/otechnik/Spravochniki/slovar-ustarevshih-slov/> (дата обращения: 20.12.2025).
5. Пеганов Н. С. Корпусный анализ архаизации русской поэтической лексики: исходный код исследования // GitHub repository. 2025. URL: <https://github.com/NikPeg/poetry-archaization-analysis> (дата обращения: 24.12.2025).

6. Успенский Б. А. Краткий очерк истории русского литературного языка (XI–XIX вв.). М.: Гнозис, 1994.
7. Lyashevskaya O., Sharov S. Frequency dictionary of modern Russian: the Russian National Corpus // Online: <http://dict.ruslang.ru/freq.php>, 2009.