УДК 004:81-139

*Амиева А.М., Филимонов В.В., Живодёров А.А.,*
*Горбич Л.Г., Крамаренко А.А.*

# МАШИННАЯ АТРИБУЦИЯ РУССКОЯЗЫЧНЫХ ТЕКСТОВ И ХРОНОЛОГИЧЕСКИЙ ФАКТОР

*Аннотация.* В работе описано исследование русскоязычных текстов, проведённое с применением статистики $\chi^2$. Обнаружены закономерности изменения средних значений $\chi^2$ в зависимости от времени написания за последние 200 лет. Работа выполнена на кафедре полиграфии и веб-дизайна ИРИТ-РтФ УрФУ.
*Ключевые слова:* статистика $\chi^2$, текст, классификация.

*Abstract.* The paper describes the study of Russian texts, carried out using the statistics $\chi^2$. Regularities in the variation of the mean values of $\chi^2$ as a function of writing time over the past 200 years are found. The work was completed at the department of printing art and web-design, Ural Federal University.
*Keywords:* statistics $\chi^2$, text, classification.

## Введение

Задача классификации текстов по жанрам кажется простой, но при попытке формализации оказывается весьма нетривиальной. «Так, если «объяснение» компьютеру понятий рифмы и ритмической размерности еще можно себе представить, то анализ «музыкальности», «образности» и «эстетического воздействия» кажется задачей, превосходящей по сложности проблему компьютерного анализа смысла текстов» [1].

На данном этапе наша работа связана с разработкой методики машинной атрибуции текстов. Мы предполагаем, что полученные нами результаты помогут в решении следующих задач:

1) построение математической модели текста;
2) исследовательские задачи (установление авторства и пр.);
3) оценка удобочитаемости (юзабилити);
4) учёт особенностей текста при редактировании и вёрстке.

## Ход работы

Исследования проводились на материале «Корпуса текстов русского языка» (далее Корпус) и с помощью программ «*Coder*» и «*QLines*». Корпус на сегодняшний день состоит из около 1 500 текстов различных направлений и жанров. Каждое направление представлено отдельным подкорпусом. Тексты были взяты из открытых интернет источников [2].

С помощью программы «*Coder*» для проведения исследования все гласные буквы были переведены в цифровой код. В кодированных текстах не представлены согласные буквы, пробелы и знаки препинания. При подсчётах буква «ё» учитывалась как «е», т.к. в большей части опубликованных в сети интернет текстов буква «ё» не использована.

На следующем этапе исследовалась количество букворазмещений по три. Оно было получено двумя способами.

Сначала как произведение частоты появлений букворазмещений ($\omega_i^{theor}$) на количество гласных букв в тексте ($N$):

$$n_i^{theor} = \omega_i^{theor} \cdot N \quad (1)$$

$$\omega_i^{theor} = \omega_{i_1} \cdot \omega_{i_2} \cdot \omega_{i_3} \quad (2)$$

где $\omega_{i_1}, \omega_{i_2}, \omega_{i_3}$ – частоты появления первой, второй, третьей букв в букворазмещении, которые в свою очередь рассчитываются по следующей формуле:

$$\omega_{i_j} = \frac{n_j}{N} \quad (3)$$

где $n_j$ – количество появлений соответствующей отдельной гласной буквы в тексте.

Получаемое таким образом количество букворазмещений было названо «теоретическим количеством».

Затем при помощи программы «*QLines*» были пересчитаны все тройки гласных во всех текстах Корпуса. Полученные значения были названы «наблюдаемым количеством» ($n_i^{emp}$).

Для оценки отличия предполагаемого распределения троек гласных от их действительного распределения была применена статистика $\chi^2$:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i^{theor} - n_i^{emp}\right)^2}{n_i^{theor}} \cdot \frac{50000}{N} \quad (4)$$

Нормировка на длину текста в 50 000 гласных применяется для того, чтобы значения $\chi^2$ для текстов разной длины можно было сравнивать между собой.

В результате было получено распределение текстов Корпуса по кластерам, связанным с величиной $\chi^2$ и прагматикой самих текстов:

– в интервале от 0 до 2 000 имеются 3 текста М.В. Ломоносова;
– в интервале от 2 000 до 2 400 встречаются только стихотворные произведения;
– художественные тексты (стихи и проза) встречаются во всём диапазоне значений $\chi^2$, но большинство расположены в интервале от 2 000 до 6 000;
– научные тексты начинают появляться с 4 000, большинство их расположено в интервале от 6 000 до 12 000;
– в интервале от 4 000 до 12 000 расположены религиозные, социально-политические и публицистические тексты;
– в следующем интервале от 10 000 до 30 000 расположены административные тексты.

Тексты оказались объединёнными в кластеры, которые в основном совпали с подкорпусами, выделенными экспертно, хоть и машина не ориентировалась на смысл текста и его название, проанализировав только последовательность знаков.

В предыдущей работе [1] в качестве дополнительного параметра для более полного разделения групп был выбран год рождения автора произведения, который «отражает историческое время деятельности создателей текстов».

В результате две группы текстов – поэзия и проза – оказались полностью разделены. Распределение 102 произвольных текстов на две группы по двум параметрам может свидетельствовать о том, что в основе такого разделения лежит некоторый закон. Тогда было выдвинуто предположение, что «нормативы поэзии изначально существенно отличались от прозаических, затем, с течением времени, постепенно сближались, почти смыкаясь, и в настоящее время снова расходятся» [1].

В настоящей работе в качестве дополнительного параметра было выбрано время написания текста. Для русскоязычных текстов указывалось время написания автором, а для переводных – год перевода текста на русский язык. Исследование было проведено на материале трёх подкорпусов – поэзия, художественная проза и научные тексты.

Тексты в подкорпусе поэзии (Рис. 1) представлены от 1726 по 2006 год. На Рисунке прослеживается волнообразный характер изменения величины $\chi^2$ в зависимости от времени. Отчётливо видны минимумы в интервалах между 1800 и 1822 гг., между 1830 и 1880 гг., между 1918 и 1954 гг. и между 1965 и 1988 гг. В это время поэтические тексты, по-видимому, максимально

обособляются, поэтические правила соблюдаются более строго, чем в другие временные промежутки, когда поэтические тексты по своей структуре приближаются к прозаическим.

Среднее значение $\chi^2$ для подкорпуса художественной прозы (Рис. 1) с течением времени изменяется мало. Значения $\chi^2$ концентрируются вокруг величины равной 5000 с разбросом порядка 1000. Устойчивых тенденций к повышению или понижению значений $\chi^2$ не наблюдается. Можно сказать, что структура текстов художественной прозы в течение последних двухсот лет существенно не меняется.

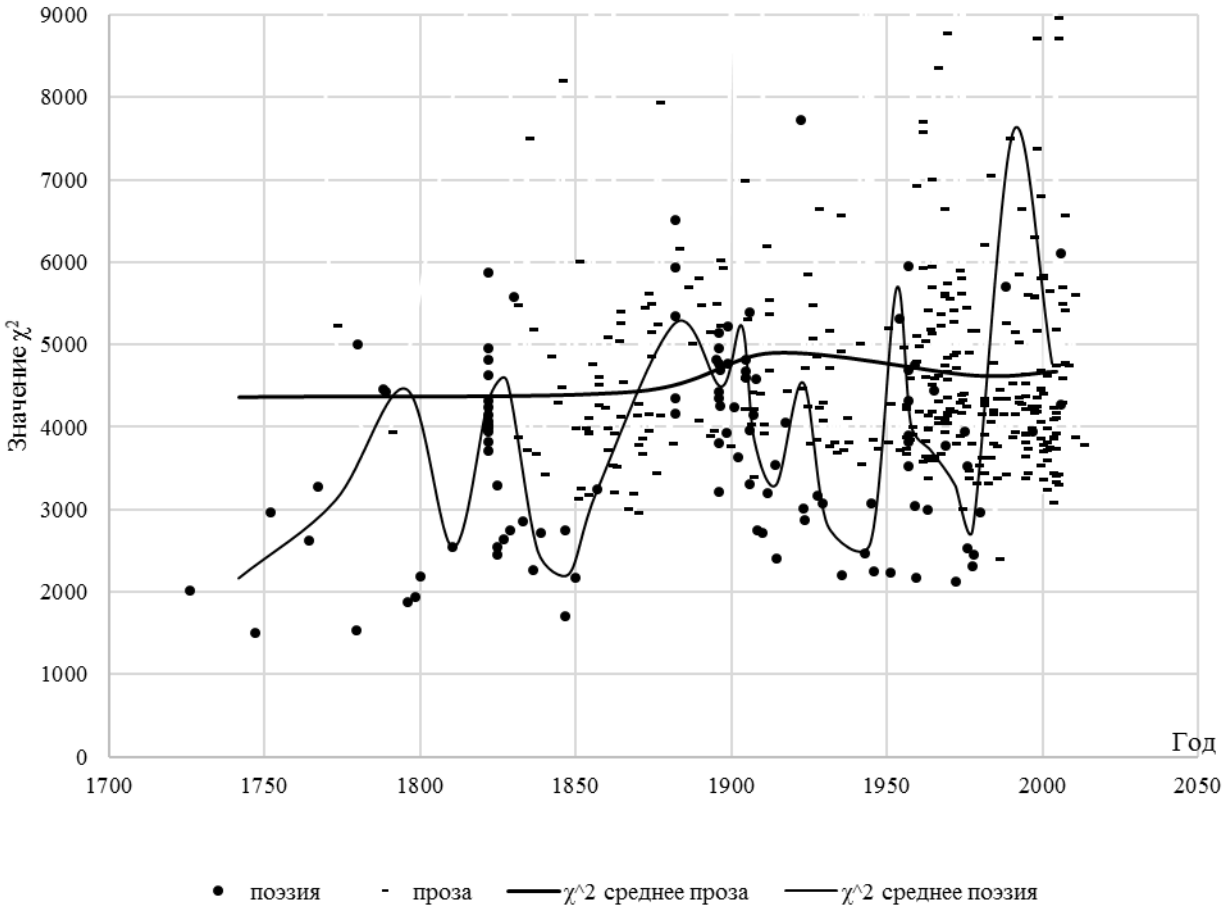


Рисунок 1 – Диаграмма значений $\chi^2$ для подкорпусов поэзии и художественной прозы в зависимости от времени

Первые по времени написания тексты научного подкорпуса принадлежат М.В. Ломоносову и имеют значения $\chi^2$ ниже 2 000 (1239, 1626, 1650). Возникает вопрос, можно ли считать произведения Ломоносова в полном смысле слова научными, не смотря на темы, которые он в них затрагивает (история государства и риторика). По своему строю и по языку они больше напоминают поэтические произведения.

С 1860-х гг. по сегодняшний день разброс значений $\chi^2$ устойчиво увеличивается (Рис. 2). Формируется три диапазона. В первый диапазон (4000–8000) входят литературоведческие, искусствоведческие и языковедческие тексты, во втором диапазоне (8000–12000) к ним добавляются тексты по естественной истории и истории техники, медицине, экономике, социологии, психиатрии, в третьем (больше 12000) расположены тексты по математике, естественным и техническим наукам. Тексты по философии и психологии встречаются во всех диапазонах.

Очевидно, можно говорить о постепенном формировании специфических научных языков, характерных для различных дисциплин. Специфика отдельных научных языков находит своё отражение в структуре текста и оказывается связанной с величиной статистики $\chi^2$.

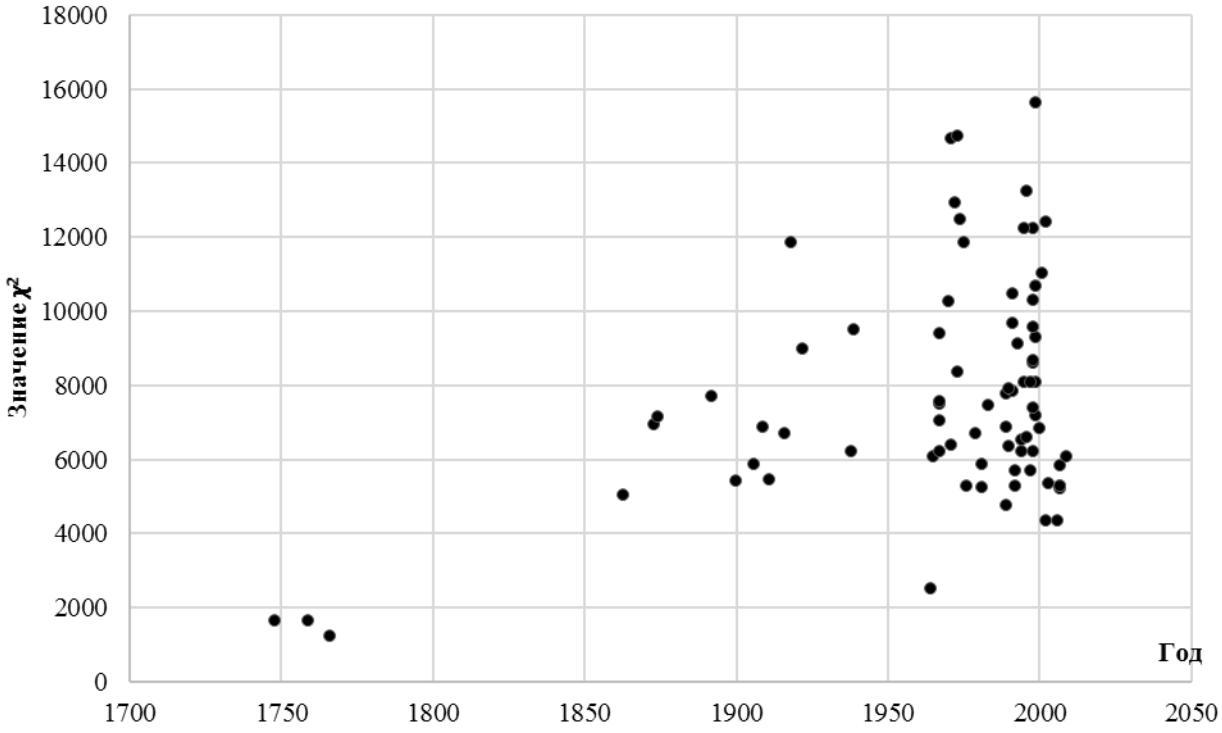


Рисунок 2 – Диаграмма значений $\chi^2$ для подкорпуса научных текстов в зависимости от времени

### *Библиографический список*

1. Горбич Л. Г. Опыт различения поэтических и прозаических текстов на основе сравнения распределений биграмм гласных букв / Л. Г. Горбич, В. В. Филимонов, А. А. Живодёров // Количественные методы в искусствознании : материалы Междунар. науч.-практ. конф. (Екатеринбург, 20–22 сент. 2012 г.). – Екатеринбург, 2013. – С. 163–166.

2. Филимонов В. В. Кластеризация русскоязычных текстов с применением статистики $\chi^2$ / В. В. Филимонов, А. М. Амиева, А. П. Сергеев // Информация: передача, обработка, восприятие материалы : междунар. науч.-практ. конф. (Екатеринбург, 12–13 янв. 2016 г.). – Екатеринбург : УрФУ, 2016. – С. 164–174.
3. Филимонов В. В. Экспрессия и упорядоченность в письменной речи / В. В. Филимонов, А. А. Живодёров, Л. Г. Горбич // Известия Уральского Федерального Университета. Серия 1. Проблемы образования, науки и культуры. – 2012. – № 3 (104). – С. 313–319.
4. Машинная атрибуция русскоязычных текстов: обзор методов / А. М. Амиева [и др.] // Новые информационные технологии в образовании и науке : материалы X междунар. науч.-практ. конференции (Екатеринбург, 27 февр. –3 марта 2017 г.). – Екатеринбург : РГППУ, 2017. – С. 371–375.