УДК 519.23

ПРИМЕНЕНИЕ ЧАСТОТНЫХ ХАРАКТЕРИСТИК ДЛЯ ОПРЕДЕЛЕНИЯ АВТОРСТВА ЛИТЕРАТУРНЫХ ТЕКСТОВ

Суетин В.Ю.
РГУ имени А.Н. Косыгина, г. Москва

Поступила в редакцию 29.04.2022, после переработки 15.06.2022.

В настоящей работе гипотеза И. Амлински о том, что автором романа «12 стульев» является М.А. Булгаков, опровергнута методами математической статистики. В качестве авторского инварианта использовано относительное количество служебных слов (частиц, союзов, предлогов) на объёмах 6000 слов. Проведено сравнение указанных частотных характеристик романов «12 стульев», «Мастер и Маргарита», «Белая Гвардия». Использован критерий Манна-Уитни.

Ключевые слова: авторский инвариант, критерий Манна-Уитни, критерий Шермана равномерного распределения.

Вестник ТвГУ. Серия: Прикладная математика. 2022. № 2. С. 84–89.
https://doi.org/10.26456/vtpmk637

Введение

В 2013 году искусствовед Ирина Амлински опубликовала книгу [1], в которой на основании сравнительного лингвистического анализа текстов романов «12 стульев» и «Мастера и Маргариты» сделала вывод о том, что автором «12 стульев» вполне мог быть М.А. Булгаков. Проверка этой гипотезы методами математической статистики впервые была проведена В.Ю. Суетиным и Е.С. Анненковой в работе [2]. В этой работе в качестве авторского инварианта, то есть числовой характеристики, присущей конкретному автору, использовано отношение числа служебных слов (предлогов, союзов, частиц) к общему числу слов выборки. Правомерность использования указанного числового параметра текстов показана в работе [3], где отмечено, что такой параметр не является устойчивым на малых объёмах слов и рекомендован объём в 16000 слов, на котором параметр уже является стабильным. При этом авторы исследования [3] работали с короткими текстами, например, рассказами А.П.Чехова. Мы же исследуем «длинные» тексты, в результате чего параметр становится устойчивым на гораздо меньших объёмах: при использовании выборок по 6000 слов мы показали, что имеет место равномерность распределения относительных значений доли служебных слов, а для произведений одного автора выборки отличаются не существенно. В различных работах в качестве авторских инвариантов использовались, например, длина предложения,

длина слов, доля употребления существительных, доля употребления глаголов, частота употребления частицы «не» или предлога «в» и т.д. Использование доли служебных слов выгодно отличается от перечисленных признаков некой «интегральностью»: как указано в работе [3], "большое число служебных слов, используемых в русском языке, делает этот параметр невероятно трудно контролируемым на сознательном уровне". В соответствии с предложенной в [3] методологией, в работе [2] и в настоящей работе мы использовали следующие служебные слова:

предлоги — в, на, с, за, к, по, из, у, от, для, во, без, до, о, через, со, при, про, об, ко, над, из-за, из-под, под;

союзы — и, что, но, а, да, хотя, когда, чтобы, если, тоже, или, то есть, зато, будто;

частицы — не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, то есть, нибудь, уже, либо.

1. Частотный анализ

Мы выделили из текстов романов «12 стульев», «Мастер и Маргарита» и «Белая гвардия» по 9 равновеликих (по 6000 слов) выборок через интервалы в 10 страниц текста. Проверка равномерности распределений проведена по критерию Шермана. Затем мы попарно сравнили через критерий Манна-Уитни данные для романов «Мастер и Маргарита» и «Белая гвардия», «12 стульев» и «Мастер и Маргарита», «12 стульев» и «Белая гвардия».

Таблица 1: Ранжированные наборы частот служебных слов

Мастер и Маргарита	Белая гвардия	12 стульев
0,2313	0,2077	0,1955
0,2332	0,2093	0,2003
0,2380	0,2177	0,2012
0,2388	0,2312	0, 2052
0,2413	0,2323	0,2068
0,2418	0,2328	0,2080
0,2497	0,2393	0,2140
0,2508	0,2408	0,2201
0,2568	0,2460	0,2211

Равномерность распределений проверяем по критерию Шермана [4]: для вариационного ряда $x_1 < x_2 < \cdots < x_n$ проводим проверку простой гипотезы о равномерности распределения выборки меньшего объёма $n-2$, соответствующей ряду $x_2 < x_3 < \cdots < x_{n-1}$, на отрезке $[x_1; x_n]$. Введём величины

$$U_{i-1} = \frac{x_i - x_1}{x_n - x_1}, \quad i = 2, \ldots, n-1;\ U_0 = 0;\ U_{n-1} = 1.$$

Статистика критерия Шермана имеет вид

$$\omega_n = \frac{1}{2} \sum_{i=1}^{n+1} \left| U_i - U_{i-1} - \frac{1}{n+1} \right|,$$

то есть в нашем случае

$$\omega_7 = \frac{1}{2}\sum_{i=1}^{8}\left|U_i - U_{i-1} - \frac{1}{8}\right|.$$

Для «Мастера и Маргариты» величина ω_7 составила 0,358, для «Белой гвардии» – 0,377 , для «12 стульев» – 0,316 при критическом значении, соответствующим уровню значимости $\alpha = 0,05$, равном 0,488 [4]. Таким образом, все рассмотренные распределения можно считать равномерными.

2. Сравнительный анализ

Для проверки устойчивости авторского инварианта на выборках в 6000 слов для текстов большой длины и для сравнения произведений (возможно) разных авторов применяем ранговый критерий Манна-Уитни. Сначала сравним данные по произведениям М.А.Булгакова «Белая гвардия» и «Мастер и Маргарита». Нулевая гипотеза H_0: выборки получены из одной генеральной совокупности, различия в значениях не являются существенными и носят случайный характер. Альтернативная гипотеза H_1: выборки получены из разных генеральных совокупностей, различия данных носят существенный характер.

Таблица 2: Ранговое сравнение частот по романам «Белая гвардия» и «Мастер и Маргарита»

Значение	Роман	Ранг
0,2077	БГ	1
0,2093	БГ	2
0,2177	БГ	3
0,2312	БГ	4
0,2313	ММ	5
0,2323	БГ	6
0,2328	БГ	7
0,2332	ММ	8
0,2380	ММ	9
0,2388	ММ	10
0,2393	БГ	11
0,2408	БГ	12
0,2413	ММ	13
0,2418	ММ	14
0,2460	БГ	15
0,2497	ММ	16
0,2508	ММ	17
0,2568	ММ	18

Сумма рангов для «Мастера и Маргариты» $n_x = 110$, для «Белой гвардии» $n_y = 61$, большая из двух ранговых сумм $T = 110$, статистика Манна-Уитни

$$U = n_x \cdot n_y + \frac{n_x(n_x+1)}{2} - T = 16.$$

Критическое значение статистики Манна-Уитни для уровня значимости $\alpha = 0,01$ равно 11, критерий левосторонний, так что эмпирическая статистика, большая, чем критическое значение, не попадает в критическую область и, следовательно, нет оснований отвергать нулевую гипотезу. Как видим, для романов одного автора использование доли служебных слов даёт ожидаемый результат.

Сравним теперь попарно данные по романам «Белая гвардия» и «12 стульев» (таблица 3) и по романам «Мастер и Маргарита» и «12 стульев» (таблица 4) с такими же нулевой и альтернативной гипотезами.

Таблица 3: Сравнительный анализ романов «Белая гвардия» и «12 стульев»

Значение	Роман	Ранг
0,1955	12С	1
0,2003	12С	2
0,2012	12С	3
0,2052	12С	4
0,2068	12С	5
0,2077	БГ	6
0,2080	12С	7
0,2093	БГ	8
0,2140	12С	9
0,2177	БГ	10
0,2201	12С	11
0,2211	12С	12
0,2312	БГ	13
0,2323	БГ	14
0,2328	БГ	15
0,2393	БГ	16
0,2408	БГ	17
0,2560	БГ	18

Таблица 4: Сравнительный анализ романов «Мастер и Маргарита» и «12 стульев»

Значение	Роман	Ранг
0,1955	12С	1
0,2003	12С	2
0,2012	12С	3
0,2052	12С	4
0,2068	12С	5
0,2080	12С	6
0,2140	12С	7
0,2201	12С	8
0,2211	12С	9
0,2388	ММ	10
0,2393	ММ	11
0,2408	ММ	12
0,2413	ММ	13
0,2418	ММ	14
0,2460	ММ	15
0,2497	ММ	16
0,2508	ММ	17
0,2568	ММ	18

Сумма рангов в таблице 3 для «Белой гвардии» $n_x = 117$, для «12 стульев» $n_y = 54$, большая из двух ранговых сумм $T = 117$, статистика Манна-Уитни

$$U = n_x \cdot n_y + \frac{n_x(n_x+1)}{2} - T = 9.$$

Здесь эмпирическая статистика меньше критического значения для $\alpha = 0,01$, то есть попадает в критическую область, а значит есть основание отвергнуть нулевую гипотезу. Таким образом, отличие значений доли служебных слов в выборках из романов «Белая гвардия» и «12 стульев» существенно.

Ещё сильнее эти различия при сравнении романов «Мастер и Маргарита» и «12 стульев».

Сумма рангов в таблице 4 для «Мастера и Маргариты» $n_x = 126$, для «12 стульев» $n_y = 45$, большая из двух ранговых сумм $T = 126$, статистика Манна-Уитни

$$U = n_x \cdot n_y + \frac{n_x(n_x+1)}{2} - T = 0.$$

Итак, мы вновь принимаем альтернативную гипотезу о том, что различия данных существенны.

Заключение

В работе рассмотрено применение относительного числа служебных слов (союзов, предлогов, частиц) в качестве авторского инварианта для литературных текстов большой длины. С помощью критерия Манна-Уитни с уровнем значимости $\alpha = 0,01$ опровергнута гипотеза И.Амлински о том, что автором романа «12 стульев» является М.А. Булгаков.

Список литературы

[1] Амлински И. 12 стульев от Михаила Булгакова. Numbrecht: Kirschner Verlag, 2013. 328 с.

[2] Проверка гипотезы об авторстве «12 стульев» методами математической статистики // Материалы III Всероссийской научно-практической конференции "Перспективы развития математического образования в эпоху цифровой трансформации". Тверь: ТвГУ, 2022. С. 22–24.

[3] Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Методы количественного анализа текстов нарративных источников. М.: АН СССР. Институт Истории СССР, 1983.

[4] Лемешко Б.Ю., Блинов П.Ю. Критерии проверки отклонения распределения от равномерного закона (Руководство по применению). Новосибирск: НГТУ, 2015. 182 с.

Образец цитирования

Суетин В.Ю. Применение частотных характеристик для определения авторства литературных текстов // Вестник ТвГУ. Серия: Прикладная математика. 2022. № 2. С. 84–89. https://doi.org/10.26456/vtpmk637

Сведения об авторах

1. **Суетин Валерий Юрьевич**
доцент кафедры высшей математики института экономики и менеджмента Российского государственного университета им. А.Н. Косыгина.

Россия, 119071, г. Москва, улица Малая Калужская, д. 1, РГУ им А.Н. Косыгина. E-mail: suetin-vu@rguk.ru

APPLICATION OF FREQUENCY CHARACTERISTICS TO DETERMINE THE AUTHORSHIP OF LITERARY TEXTS

Suetin Valeriy Yur'evich
Assistant professor at High Mathematics department,
Kosygin Russian State University
Russia, 119071, Moscow, M.Kaluzhskaya str, 1. Kosygin RSU.
E-mail: suetin-vu@rguk.ru

Received 29.04.2022, revised 15.06.2022.

In this paper, the hypothesis of I. Amlinski that the author of the novel "12 chairs" is M.A. Bulgakov is refuted by the methods of mathematical statistics.

Keywords: author's invariant, Mann-Whitney test, Sherman uniform distribution test.

Citation

Suetin V.Yu., "Application of frequency characteristics to determine the authorship of literary texts", *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2022, № 2, 84–89 (in Russian). https://doi.org/10.26456/vtpmk637

References

[1] Amlinski I., *12 stulev ot Mikhaila Bulgakova [12 chairs by Mikhail Bulgakov]*, Kirschner Verlag, Numbrecht, 2013 (in Russian), 328 pp.

[2] "Verification of the hypothesis about the authorship of "12 chairs" by methods of mathematical statistics", *Materialy III Vserossijskoj nauchno-prakticheskoj konferentsii "Perspektivy razvitiya matematicheskogo obrazovaniya v epokhu tsifrovoj transformatsii" [Materials of the III All-Russian Scientific and Practical Conference "Prospects for the development of mathematical education in the era of digital transformation"]*, TvGU, Tver, 2022, 22–24 (in Russian).

[3] Fomenko V.P., Fomenko T.G., *Avtorskij invariant russkikh literaturnykh tekstov. Metody kolichestvennogo analiza tekstov narrativnykh istochnikov [The author's invariant of Russian literary texts. Methods of quantitative analysis of texts of narrative sources]*, USSR Academy OF Sciences. Institute of History of the USSR, Moscow, 1983 (in Russian).

[4] Lemeshko B.Yu., Blinov P.Yu., *Kriterii proverki otkloneniya raspredeleniya ot ravnomernogo zakona (Rukovodstvo po primeneniyu) [Criteria for checking the deviation of the distribution from the uniform law (Application Guide)]*, NGTU, Novosibirsk, 2015 (in Russian), 182 pp.