КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2017 Т. 9 № 5 С. 837–850
DOI: 10.20537/2076-7633-2017-9-5-837-850

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 51-78, 519.234.3, 519.257, 81-139

Новый метод стилеметрии на основе статистики числительных

А. В. Зенков

Уральский федеральный университет,
Россия, 620002, г. Екатеринбург, ул. Мира, д. 19
Уральский государственный экономический университет,
Россия, 620144, г. Екатеринбург, ул. 8 Марта, д. 62

E-mail: zenkow@mail.ru

Получено 01.07.2017.
Принято к публикации 14.08.2017.

Предложен новый метод статистического анализа текстов. Исследовано распределение частот различных первых значащих цифр в числительных англоязычных текстов. Учитываются количественные и порядковые числительные, выраженные как цифрами, так и словесно. Предварительно из текста удаляются случайно попавшие в него числительные, не отражающие авторский замысел (номера страниц, маркеры списков, идиоматические выражения, устойчивые обороты речи и тому подобное). Обнаружено, что для *сборных* текстов разного авторства частоты первых значащих цифр приближенно соответствуют известному закону Бенфорда, но с резким преобладанием встречаемости единицы. В *связных авторских* текстах возникают характерные отклонения от закона Бенфорда; показано, что эти отклонения являются статистически устойчивыми и значимыми авторскими особенностями, позволяющими при определенных условиях ответить на вопрос об авторстве и различить тексты разных авторов. Требуется, чтобы текст был достаточно длинным (не менее чем порядка 200 кБ). Распределение первых значащих цифр конца ряда $\{1, 2, \ldots, 8, 9\}$ подвержено сильным флуктуациям и не показательно для нашей цели. Цель теоретического обоснования найденной эмпирической закономерности в работе не ставится, но продемонстрировано ее практическое использование для атрибуции текстов. Предлагаемый подход и сделанные выводы подкреплены примерами компьютерного анализа художественных текстов У. М. Теккерея, М. Твена, Р. Л. Стивенсона, Дж. Джойса, сестер Бронте, Дж. Остин. На основе разработанной методологии рассмотрены проблемы авторства текста, ранее приписывавшегося Л. Ф. Бауму (результат согласуется с полученным другими методами), а также известного романа Харпер Ли «Убить пересмешника»; показано, что к написанию первоначального варианта этой книги («Пойди, поставь сторожа») мог быть причастен Трумен Капоте, но финальный текст, вероятно, принадлежит Харпер Ли. Результаты подтверждены на основе параметрического критерия Пирсона, а также непараметрических U-критерия Манна – Уитни и критерия Крускала – Уоллиса.

Ключевые слова: атрибуция текстов, первая значащая цифра числительных

COMPUTER RESEARCH AND MODELING
2017 VOL. 9 NO. 5 P. 837–850
DOI: 10.20537/2076-7633-2017-9-5-837-850

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 51-78, 519.234.3, 519.257, 81-139

A novel method of stylometry based on the statistic of numerals

A. V. Zenkov

Ural Federal University,
Mira st. 19, Ekaterinburg, 620002, Russia
The Ural State University of Economics,
8th of March st. 62, Ekaterinburg, 620144, Russia

E-mail: zenkow@mail.ru

Received 01.07.2017.
Accepted for publication 14.08.2017.

A new method of statistical analysis of texts is suggested. The frequency distribution of the first significant digits in numerals of English-language texts is considered. We have taken into account cardinal as well as ordinal numerals expressed both in figures, and verbally. To identify the author's use of numerals, we previously deleted from the text all idiomatic expressions and set phrases accidentally containing numerals, as well as itemizations and page numbers, etc. Benford's law is found to hold approximately for the frequencies of various first significant digits of *compound* literary texts by different authors; a marked predominance of the digit 1 is observed. In *coherent* authorial texts, characteristic deviations from Benford's law arise which are statistically stable significant author peculiarities that allow, under certain conditions, to consider the problem of authorship and distinguish between texts by different authors. The text should be large enough (at least about 200 kB). At the end of $\{1, 2, \ldots, 9\}$ digits row, the frequency distribution is subject to strong fluctuations and thus unrepresentative for our purpose. The aim of the theoretical explanation of the observed empirical regularity is not intended, which, however, does not preclude the applicability of the proposed methodology for text attribution. The approach suggested and the conclusions are backed by the examples of the computer analysis of works by W. M. Thackeray, M. Twain, R. L. Stevenson, J. Joyce, sisters Brontë, and J. Austen. On the basis of technique suggested, we examined the authorship of a text earlier ascribed to L. F. Baum (the result agrees with that obtained by different means). We have shown that the authorship of Harper Lee's "To Kill a Mockingbird" pertains to her, whereas the primary draft, "Go Set a Watchman", seems to have been written in collaboration with Truman Capote. All results are confirmed on the basis of parametric Pearson's chi-squared test as well as non-parametric Mann – Whitney U test and Kruskal – Wallis test.

Keywords: text attribution, first significant digit of numerals

Citation: *Computer Research and Modeling*, 2017, vol. 9, no. 5, pp. 837–850 (Russian).

1. Введение

Исследования, результатом которых явилась настоящая работа, инициированы известным законом Бенфорда [Benford, 1938]. Он описывает вероятность появления определенной первой значащей цифры в разнообразных распределениях величин, взятых из окружающего мира. Вопреки кажущемуся логичным предположению о том, что появление любой первой значащей цифры должны быть равновероятно, для многих массивов числовых данных первой значащей цифрой заметно чаще других является единица. Согласно закону Бенфорда, при записи числа в десятичной системе счисления вероятность появления цифры d в качестве его первой значащей цифры равна

$$P(d) = \lg\left(1 + \frac{1}{d}\right), \tag{1}$$

так что $d = 1$ должна встречаться с вероятностью $\lg 2 \approx 0.30$, $d = 2$ — с вероятностью 0.18 и т. д.

Исчерпывающее объяснение закона Бенфорда, применимое ко всем случаям его реализации, до сих пор отсутствует, хотя и сформулированы некоторые условия, благоприятствующие его появлению. Один из классических опытов Бенфорда, хорошо согласующийся с (1) (подсчет встречаемости числительных на случайно выбранных новостных страницах прессы), находит логичное объяснение в теореме Хилла [Hill, 1995], согласно которой при неоднократном случайном выборе распределения вероятностей и последующем случайном выборе числа согласно этому распределению получается набор чисел, подчиняющийся закону Бенфорда (1). Заметим, что сам Бенфорд учитывал только числительные, выраженные *цифрами*.

Несмотря на неполноту объяснения, закон Бенфорда успешно применяется для выявления подлогов в бухгалтерской отчетности [Nigrini, 2012] и фальсификаций на выборах [Battersby, 2009]; обсуждается применение в различных областях науки, в качестве иллюстрации укажем работы, связанные с физикой и астрономией [Pain, 2013; Biau, 2015; Hill, Fox, 2016], сейсмологией [Sambridge et al., 2011], стеганографией [Andriotis et al., 2013], наукометрией [Alves et al., 2014].

В работе [Зенков, 2015] показана эффективность подсчета частот различных первых значащих цифр числительных для атрибуции текстов. Оказалось, что не только для случайной комбинации разнородных текстов, но и для *связных* (русскоязычных) текстов, к которым не применимы условия теоремы Хилла, распределение частот приближается к бенфордовскому (1); при этом доля единицы заметно превышает 30 % — хотя бы потому, что, формально являясь числительным, слово «один» фактически может выступать в роли неопределенного артикля. Вероятно, играет роль и известная психологам и социологам склонность к округлению числительных (которая в художественном тексте, конечно, менее важна, чем при опросе общественного мнения).

В отличие от традиционной методологии применения закона Бенфорда, трактующей отклонения от закона как указание на возможное присутствие «фальсификаций» (в широком смысле), в работе [Зенков, 2015] сделан акцент на сопоставлении этих отклонений для текстов разных авторов; оказалось, что эти отклонения являются статистически устойчивыми авторскими особенностями, позволяющими различать тексты разных авторов (при определенных условиях, наиболее важным из которых является достаточно большая длина текста).

Исходя из этих идей, мы представляем здесь результаты дальнейших исследований, касающиеся распределения первых значащих цифр числительных в *англоязычных* текстах. Все результаты обоснованы с помощью статистических критериев согласия: параметрического критерия Пирсона, а также непараметрических U-критерия Манна – Уитни и критерия Крускала – Уоллиса.

Исследование носит экспериментальный характер. Цель теоретического обоснования экспериментальных результатов (если таковое вообще возможно) не ставилась, что, однако, не препятствует применению предложенной методологии для практических задач текстологии.

Для всех англоязычных художественных текстов, подвергнутых компьютерному статистическому анализу, мы подсчитывали частоты появления различных первых значащих цифр количественных и порядковых числительных, выраженных как цифрами, так и (значительно чаще) словесно. В последнем случае число переводилось в цифровую форму (например, “Four hundred and seventy-sixth” заменялось на “476”), а затем учитывалась только первая значащая цифра (4). Для исследования авторского употребления числительных предварительно из текста удалялись идиоматические выражения и устойчивые словосочетания, случайно содержащие числительные (например, “put two and two together”, “at first sight”), а также маркеры списков: 1), 2), 3) и тому подобное. Анализируемые тексты в основном взяты с веб-сайта проекта «Гутенберг» (http://www.gutenberg.org).

2. Распределение первых значащих цифр числительных в сборных литературных текстах

Условия теоремы Хилла наилучшим образом удовлетворяются для составных текстов, содержащих фрагменты разного авторства. В этом случае авторские особенности фрагментов (см. ниже, § 3) усредняются и получается частотная зависимость, напоминающая бенфордовскую (1), но отличающаяся более крутым падением; встречаемость цифры 1 заметно превышает ожидаемую по закону Бенфорда. Начиная с цифры 3, теоретическая вероятность по закону Бенфорда обычно превышает наблюдаемую частоту.

Рис. 1. Распределение первых значащих цифр числительных в восьми сборниках англоязычных художественных текстов. Результаты сопоставляются с ожидаемыми в соответствии с законом Бенфорда (1)

На рис. 1 показаны результаты анализа восьми сборных англоязычных художественных текстов [The Project Gutenberg eBooks]. Для каждого сборника мы наблюдаем монотонное убывание частоты; закономерности для разных сборников в целом схожи, вариации могут быть обусловлены особенностями (например, жанром и временем создания) текстов в каждой коллекции.

3. Распределение первых значащих цифр числительных в связных литературных текстах: авторские особенности

Обычно (достаточно длинные) тексты, принадлежащие перу одного автора, имеют устойчивые особенности в статистике первых значащих цифр числительных и их распределение является стабильной характеристикой автора. В качестве примера приведем здесь распределения первых значащих цифр числительных в текстах У. Теккерея, М. Твена, Р. Л. Стивенсона и Дж. Джойса (рис. 2–5).

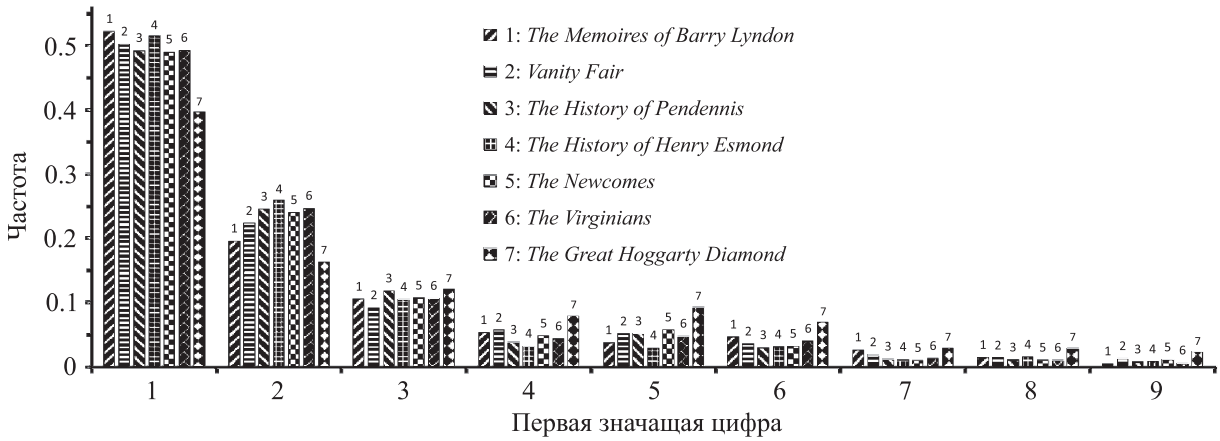

Рис. 2. Распределение первых значащих цифр числительных в текстах Теккерея. Помимо объемных текстов, для сравнения проанализирован и более короткий (“The Great Hoggarty Diamond”). Для текстов № 1–6 асимптотическая значимость $p = 0.999$ (здесь и на нижеследующих рисунках — согласно критерию Крускала – Уоллиса, см. § 4)

Рис. 3. Распределение первых значащих цифр числительных в текстах М. Твена. Асимптотическая значимость $p = 0.998$ (см. § 4)

Рис. 4. Распределение первых значащих цифр числительных в текстах Стивенсона. Асимптотическая значимость (без выбросов, №№ 3 и 6) $p = 0.999$ (см. § 4)

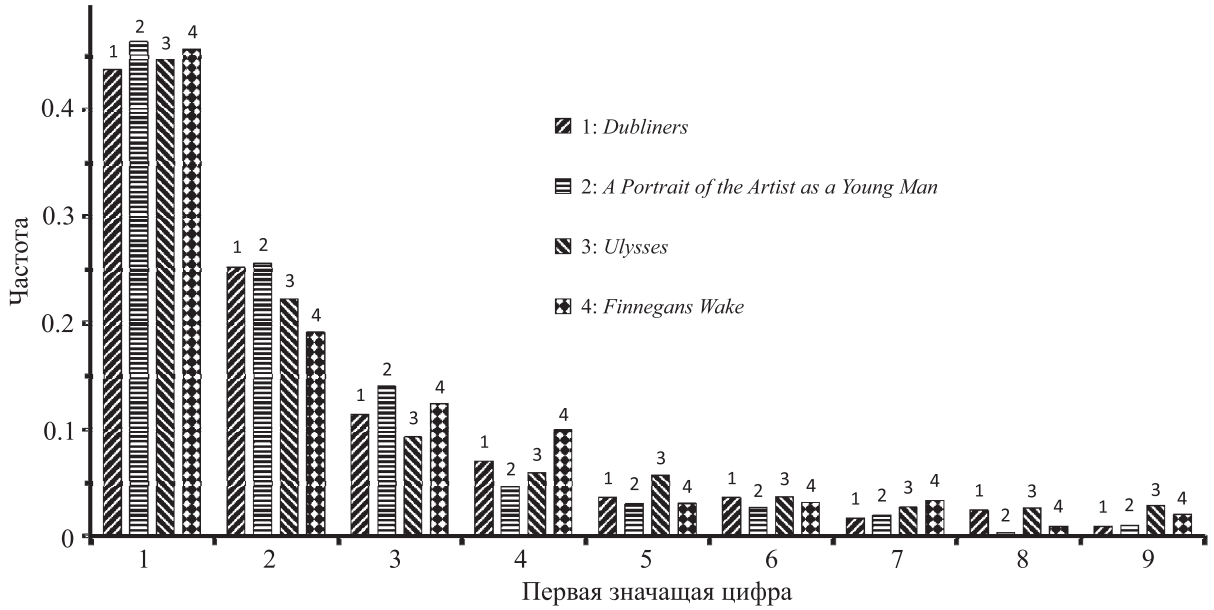

Рис. 5. Распределение первых значащих цифр числительных в текстах Джойса. Асимптотическая значимость $p = 0.998$ (см. § 4)

Рис. 6. Распределение первых значащих цифр числительных в текстах сестер Бронте. Асимптотическая значимость $p = 0.999$ (см. § 4)

Обратим внимание на два выброса, соответствующие значащей цифре 1, для текстов, *не полностью* написанных Стивенсоном (рис. 4).

Для текстов Джойса (рис. 5) частота единицы меньше, чем в других примерах (рис. 2–4, 6).

Различия в статистике первых значащих цифр числительных для текстов разных авторов могут быть не очень заметными, как в случае романов сестер Бронте (рис. 6). Это можно связать с общностью их биографии и воспитания.

Частота первой значащей цифры 1 может достигать значения в два раза большего, чем по закону Бенфорда (рис. 1–6). Именно встречаемость этой цифры, а также цифр 2 и 3 (в меньшей степени) определяет авторскую специфику текстов в нашем подходе. Встречаемость последующих цифр подвержена сильным колебаниям, которые препятствуют получению полезной информации из их распределения.

Для достижения статистической устойчивости интересующих нас частотных характеристик тексты должны быть достаточно длинными: роман, повесть, но, видимо, не рассказ. На рис. 2, кроме больших произведений У. Теккерея, рассмотрен более короткий текст (“The Great Hoggarty Diamond”), который имеет существенно иные характеристики.

Невозможно точно указать универсальную длину текста, при которой частотные характеристики приобретают статистическую устойчивость, поскольку для разных авторов она индивидуальна. В любом случае, по нашим наблюдениям, для текстов размером более 200 кБ (размер файла в формате *txt*) частоты первых значащих цифр начинают стабилизироваться.

Частота первой значащей цифры 1 является характерной авторской особенностью, которая иногда позволяет различать тексты разных авторов, если эта частота сильно отличается для них. Насколько сильным должно быть различие, чтобы считаться значимым? Мы ответим на этот вопрос в следующем параграфе.

4. Проверка статистической значимости результатов

Разумеется, нельзя полностью полагаться на непосредственно наблюдаемые близость/различия в гистограммах. Сравнение двух и более эмпирических распределений (в нашем случае — распределений первых значащих цифр числительных в текстах тех или иных авторов) связано с проверкой соответствующих статистических гипотез о значимости/незначимости различий между распределениями [Поллард, 1982; Сидоренко, 2001].

Сформулируем гипотезы. Нулевая гипотеза H_0 утверждает, что все проверяемые совокупности распределены одинаково. Альтернативная гипотеза H_1: распределения отличаются друг от друга. Проверку H_0 удобно проводить[1], используя непараметрические ранговые U-критерий Манна – Уитни и критерий Крускала – Уоллиса, первый из которых рассматривает два распределения, а второй является обобщением на случай большего числа распределений.

Например, для романов “Wuthering Heights” («Грозовой перевал») Эмили Бронте и “The Tenant of Wildfell Hall” («Незнакомка из Уайлдфелл-Холла») Энн Бронте (№ 5 и 7 на рис. 6) U-критерий Манна – Уитни дает асимптотическую значимость (вероятность H_0) $p = 0.566$. Такое значение означает статистическую *незначимость* различий[2]. Для всех текстов Бронте (рис. 6) критерий Крускала – Уоллиса дает асимптотическую значимость $p = 0.999$. Снова подтверждается нулевая гипотеза[3].

К сожалению, ранговые критерии согласия, заменяющие исходные статистические данные их рангами (при этом часть информации неизбежно теряется), имеют при прочих равных условиях меньшую мощность (разрешающую способность) по сравнению с параметрическими критериями. Критерии Манна – Уитни и Крускала – Уоллиса обычно не способны подтвердить визуально наблюдаемые очевидные различия между распределениями.

Для подтверждения *различий* между распределениями первых значащих цифр мы использовали параметрический критерий согласия χ^2 Пирсона, который, помимо прочих применений, используется и для сопоставления эмпирических распределений одного и того же признака (для проверки *однородности* распределений). В нужной нам форме соответствующая процедура отсутствует в стандартных статистических пакетах, поэтому опишем ее подробно на примере текстов Теккерея “Virginians” и “The Great Hoggarty Diamond”, статистические характеристики которых визуально существенно различны (рис. 2). Наши исходные статистические данные, касающиеся встречаемости разных первых значащих цифр в двух текстах, приведены в таблице 1.

[1] Тем более что для них есть эффективная реализация в стандартных статистических пакетах (мы применяли SPSS).
[2] Применение параметрического критерия Пирсона (см. ниже) дает тот же результат.
[3] Во всех вычислениях (если не указано иное) уровень значимости $\alpha = 0.05$.

Таблица 1. Эмпирические результаты статистической обработки текстов

Первая значащая цифра	"The Virginians"			"The Great Hoggarty Diamond"			Сумма частот по строке
	Частота	Относительная частота, %	Метка ячейки	Частота	Относительная частота, %	Метка ячейки	
1	1194	49.20	I	261	39.67	II	1194 + 261 = **1455**
2	597	24.60	III	107	16.26	IV	597 + 107 = **704**
3	257	10.59	V	79	12.01	VI	257 + 79 = **336**
4	104	4.29	VII	52	7.90	VIII	104 + 52 = **156**
5	114	4.70	IX	61	9.27	X	114 + 61 = **175**
6	95	3.91	XI	45	6.84	XII	95 + 45 = **140**
7	30	1.24	XIII	19	2.89	XIV	30 + 19 = **49**
8	23	0.95	XV	19	2.89	XVI	23 + 19 = **42**
9	13	0.54	XVII	15	2.28	XVIII	13 + 15 = **28**
	Σ = 2427	**Σ = 100 %**		**Σ = 658**	**Σ = 100 %**		**ΣΣ = 3085**

Сопоставим эмпирическим частотам теоретические, получаемые с учетом того, что количество числительных в текстах различно: 2427 в "The Virginians" и 658 в "The Great Hoggarty Diamond". Таким образом, из общей суммы 2427 + 658 = 3085 числительных в двух текстах на долю первого приходится 2427/3085 = 0.79, а на долю второго — 658/3085 = 0.21 всех числительных. Во всех строках *теоретические* частоты, относящиеся к первому и второму текстам, должны составлять соответственно 0.79 и 0.21 от суммарной частоты соответствующей строки. Если эмпирические распределения, подлежащие сравнению, не отличаются друг от друга, то эмпирические частоты не должны существенно отличаться от теоретических, полученных из пропорции.

Перекомпонуем (см. таблицу 2) данные таблицы 1, помещая относительные частоты для обоих текстов в порядке, указанном метками, в одном столбце (это будут эмпирические частоты f_{emp}), а в другом столбце мы поместим теоретические частоты f_{theor} рассчитанные, согласно предыдущему, следующим образом:

$$f_{\text{theor}} = \frac{(\sum \text{частот по строке}) \cdot (\sum \text{частот по столбцу})}{\sum\sum = 3085}.$$

При сопоставлении эмпирических распределений по критерию Пирсона учитывается количество степеней свободы $df = (r-1)(c-1)$, где r — количество строк в таблице эмпирических частот, c — количество сопоставляемых распределений. В нашем случае $r = 9$, $c = 2$. Итак, количество степеней свободы $df = (9-1)(2-1) = 8$.

При таком df табличные критические значения распределения χ^2 для двух уровней значимости следующие:

$$\chi^2_{\text{cr}} = \begin{cases} 15.507 & (\alpha = 0.05), \\ 20.090 & (\alpha = 0.01). \end{cases}$$

Поскольку эмпирическое χ^2_{emp} превышает каждое из этих критических, гипотеза H_0 отвергается даже при уровне значимости $\alpha = 0.01$; эмпирические распределения различаются *значимо*.

Эту процедуру мы использовали всюду в данной работе для обоснования значимости различий между распределениями.

Таблица 2. Вычисления для критерия согласия χ^2 Пирсона

Ячейка	Эмпирическая частота f_{emp}	Теоретическая частота f_{theor}	$\left(f_{\text{emp}} - f_{\text{theor}}\right)^2 \Big/ f_{\text{theor}}$
I	1194	$1455 \cdot 2427/3085 = 1144.66$	2.13
II	261	$1455 \cdot 658/3085 = 310.34$	7.84
III	597	$704 \cdot 2427/3085 = 553.84$	3.36
IV	107	$704 \cdot 658/3085 = 150.16$	12.40
V	257	264.33	0.20
VI	79	71.67	0.75
VII	104	122.73	2.86
VIII	52	33.27	10.54
IX	114	137.67	4.07
X	61	37.33	15.02
XI	95	110.14	2.08
XII	45	29.86	7.68
XIII	30	38.55	1.90
XIV	19	10.45	6.99
XV	23	33.04	3.05
XVI	19	8.96	11.26
XVII	13	22.03	3.70
XVIII	15	5.97	13.65
	Σ = 3085	**Σ = 3085**	**Σ = 109.48** = χ^2_{emp}

5. Распознавание авторства текстов

Джейн Остин и ее эпигоны

Романы нравов Джейн Остин (1775–1817) вызвали многочисленные подражания. Близкие темы и даже осознанное намерение писать в похожей манере не уберегли имитаторов от резких отличий в статистике встречаемости первых значащих цифр (рис. 7). Для текстов № 1–7 тест Крускала – Уоллиса дает асимптотическую значимость $p = 0.998$ (нулевая гипотеза сохраняется — разница между текстами Остин не значима). Согласно критерию Пирсона, любой текст другого авторства (№ 8–11) по частотному распределению первых значащих цифр *значимо* отличается от текстов Остин (при уровне значимости $\alpha = 0.05$).

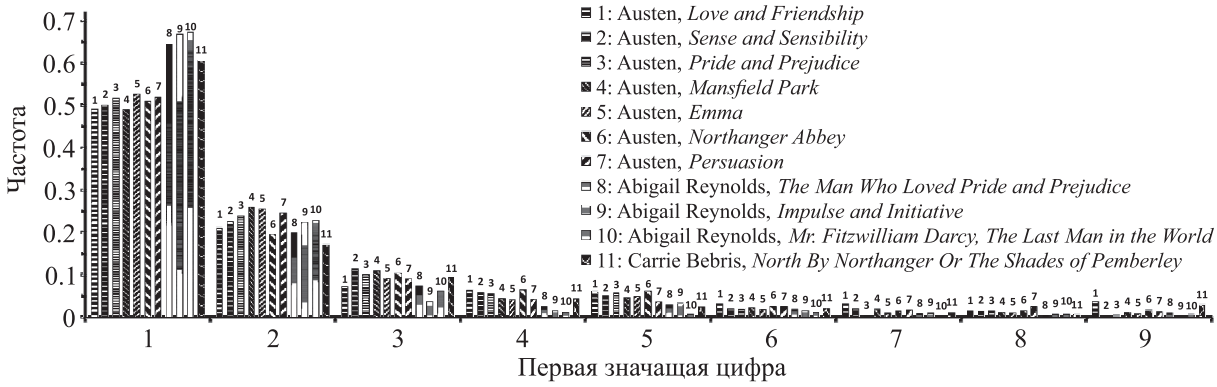

Рис. 7. Распределение первых значащих цифр числительных в текстах Джейн Остин и ее эпигонов

Итак, авторство текста может влиять на статистику первых значащих цифр числительных в нем.

Авторство 15-й книги о стране Оз

Л. Ф. Баум, плодовитый писатель, чей «Удивительный волшебник из страны Оз» имел большой успех, написал до своей смерти 13 продолжений этой книги. Серия была настолько популярна, что издатели решили не прекращать ее. «Королевская книга страны Оз», 15-я часть, опубликованная после смерти Баума, была написана «Л. Фрэнком Баумом, . . . , расширена и отредактирована Рут Пламли Томпсон», как отмечено на титульной странице первого издания (1921 г.).

Впоследствии распространилась (и сейчас является общепринятой) точка зрения, подтверждаемая лингвистически и статистически, что Томпсон не пользовалась какими-либо черновиками, оставшимися от Баума, поэтому «Королевская книга страны Оз» является полностью ее собственной работой [Binongo, 2003]. Хотя этот филологический вопрос уже решен, покажем результаты применения нашей методологии. Ниже приведены результаты статистического изучения книг Баума, а также их продолжений, принадлежащих Томпсон и другим авторам (рис. 8–10).

Обратим внимание на существенную разницу во встречаемости первой значащей цифры 1 в текстах Баума, с одной стороны, и текстах Томпсон (в частности, в «Королевской книге страны Оз» — № 6 на рис. 9), с другой стороны. Ввиду объемности анализируемых текстов это разительное отличие вряд ли может быть случайным (в отличие от последующих значащих цифр, которые даже в книгах одного и того же автора встречаются с разной частотой); оно демонстрирует авторство Томпсон.

Для текстов Баума (рис. 8) тест Крускала – Уоллиса дает асимптотическую значимость $p = 0.999$ (различие между текстами *не значимо*). Такова же ситуация и для текстов Томпсон ($p = 0.998$). Согласно критерию Пирсона частотные распределения первых значащих цифр для любого текста Баума и любого текста Томпсон *существенно* различаются (при уровне значимости $\alpha = 0.05$).

Помимо Томпсон, многие другие авторы создавали произведения по мотивам «Волшебника из страны Оз». Общность темы не обусловила схожести распределений (рис. 10). Мы рассматриваем различия распределений как особенности авторских стилей, независимо от воли и намерения автора влияющих на его тексты.

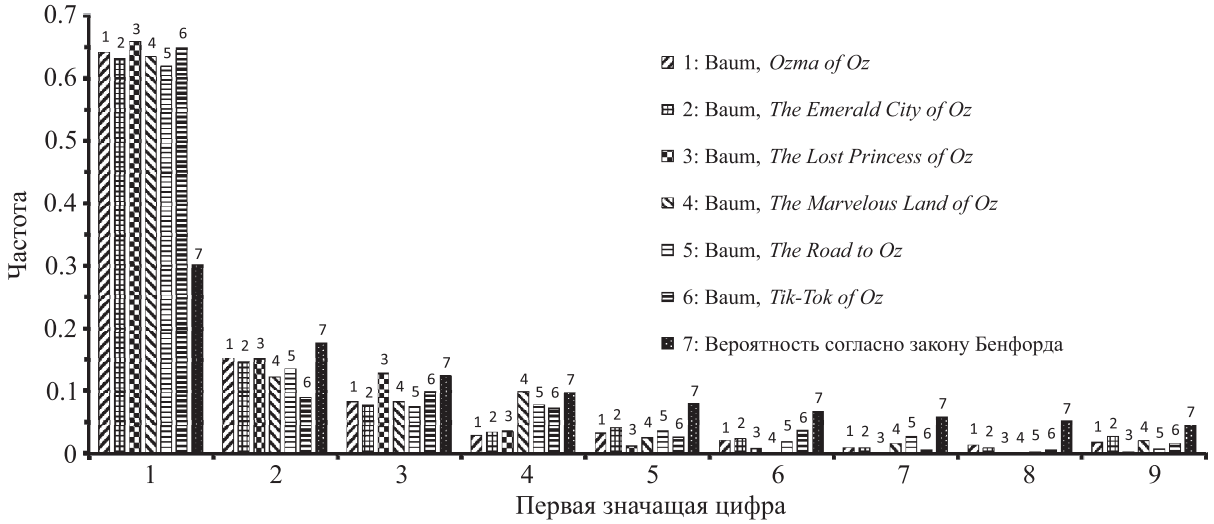

Рис. 8. Распределение первых значащих цифр числительных в текстах Баума

Рис. 9. Распределение первых значащих цифр числительных в текстах Р. П. Томпсон, продолжательницы Баума

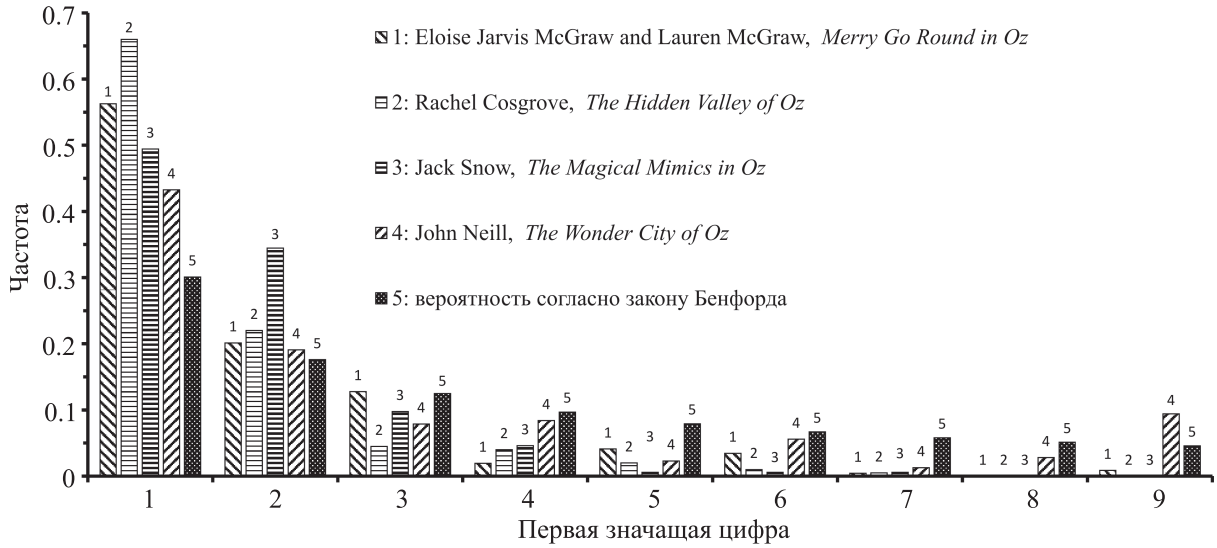

Рис. 10. Распределение первых значащих цифр числительных в продолжениях «Удивительного волшебника из страны Оз», созданных прочими авторами

Таким образом, статистический метод, основанный на подсчете частот первых значащих цифр числительных, может быть способен ответить на вопрос об авторстве текста.

6. Проверка методологии: Харпер Ли и Трумэн Капоте

Роман Харпер Ли “To Kill a Mockingbird” («Убить пересмешника»), опубликованный в 1960 г., считается одним из величайших произведений американской литературы. В 2015 г., незадолго до смерти писательницы, был опубликован ее второй роман — “Go Set a Watchman” («Пойди, поставь сторожа»). Изначально поданный издателем как продолжение «Пересмешника», сейчас он, напротив, считается первоначальным вариантом знаменитого романа.

Трумэн Капоте был пожизненным другом Харпер Ли и даже прототипом одного из персонажей «Пересмешника». В отличие от Харпер Ли, которая фактически является автором единственной книги, он был гораздо более плодовитым, и многие его произведения признаны лите-

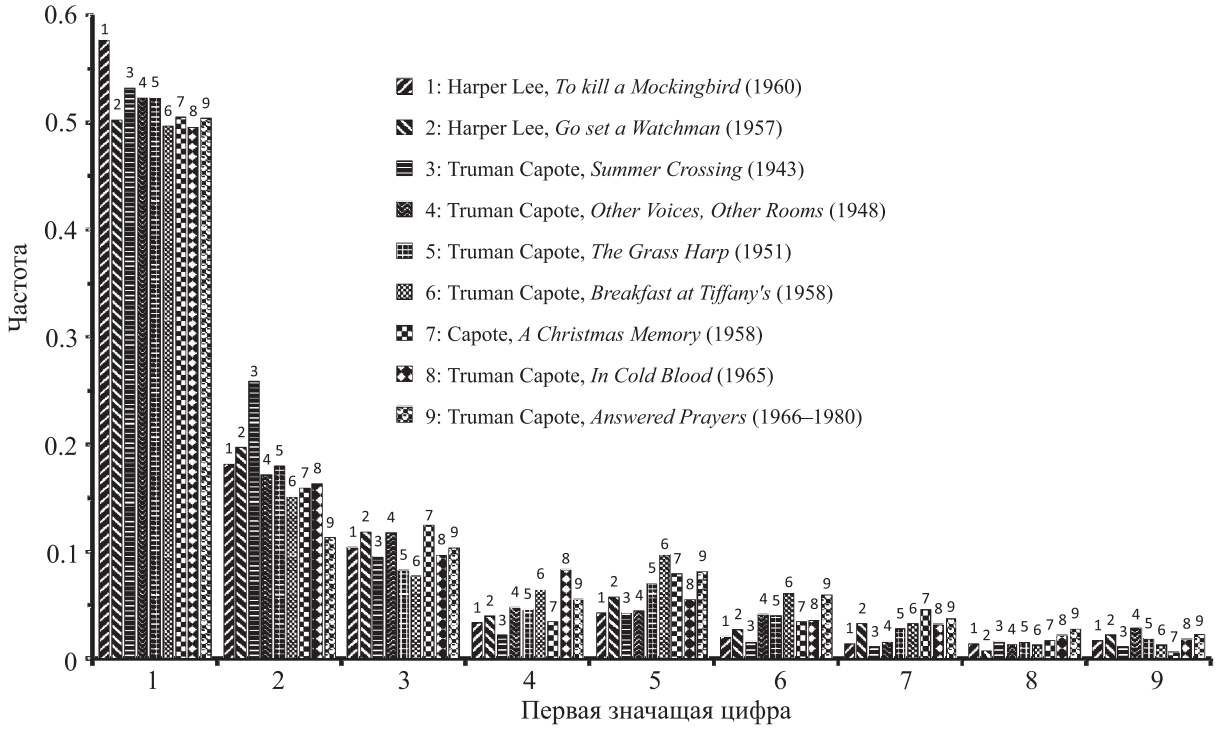

Рис. 11. Распределение первых значащих цифр числительных в текстах Харпер Ли и Трумэна Капоте

ратурной классикой. Неоднократно высказывалось подозрение, что Капоте является истинным автором книг Харпер Ли.

Проверка этой гипотезы является интересным применением нашей идеи о связи между авторством текста и статистикой числительных в нем. Мы подсчитали частоты различных первых значащих цифр числительных в романах Харпер Ли и Трумэна Капоте (рис. 11). Результаты оказались неожиданными: свойства романа «Убить пересмешника» далеки от свойств текстов Капоте, но первоначальная версия («Пойди, поставь сторожа») близка к ним[1]. Картина такова, что Капоте как будто мог помочь Харпер Ли в написании первоначального текста (обратим внимание на сходство частотных характеристик текста № 2 и текстов № 6, 7, написанных в те же годы; см. рис. 11). Приобретя писательский опыт, Харпер Ли, вероятно, *сама* создала роман «Убить пересмешника».

7. Заключение

Мы полагаем, что наша методология может быть полезным дополнением к традиционным стилеметрическим практикам учета длины предложений, длины слов, частот использования служебных слов и определенных знаменательных частей речи и т. д. [Mitkov, 2003; Aronoff, Rees-Miller, 2004; Ryabko et al., 2016]. Разумеется, для применимости нашей методологии нужно, чтобы числительные не совпадали по форме с неопределенными артиклями (как, например, *ein* в немецком или *un* во французском языке).

Результаты, полученные в рамках современных методов, основанных на нейронных сетях [Brocardo et al., 2017], могут быть впечатляющими, но сама техника, к сожалению, является «черным ящиком»: четкая интерпретация результатов обычно затруднена. В этом смысле наш подход представляется лингвистически более содержательным.

[1] Согласно критерию Пирсона частотное распределение текста № 1 (Харпер Ли) значимо отличается от частотного распределения текста № 3 (Капоте) при $\alpha = 0.01$. Если вместо № 1 взять № 2, то различие будет значимым только при $\alpha = 0.1$.

Итак, подведем итоги.

1. Закон Бенфорда приближенно выполняется для связных текстов.
2. Отклонения от закона Бенфорда являются статистически значимыми устойчивыми авторскими особенностями, позволяющими при некоторых условиях (главное из которых — достаточная длина) различить тексты разных авторов. Разумеется, сходство этих отклонений для нескольких текстов еще не означает тождественности их авторства.
3. Фактическая частота появления обычно превышает вероятность по закону Бенфорда для значащих цифр 1, 2 и иногда 3; для последующих цифр ситуация обратна. Распределение цифр конца ряда $\{1, 2, \ldots, 9\}$ подвержено сильным флуктуациям и не показательно.

Список литературы (References)

Зенков А. В. Отклонения от закона Бенфорда и распознавание авторских особенностей в текстах // Компьютерные исследования и моделирование. 2015. — Т. 7, № 1. — С. 197–201.
Zenkov A. V. Otklonenia ot zakona Benforda i raspoznavanie avtorskikh osobennostei v tekstakh [Deviation from Benford's law and identification of author peculiarities in texts] // Computer Research and Modeling. — 2015. — Vol. 7, no. 1. — P. 197–201 (in Russian).

Поллард Дж. Справочник по вычислительным методам статистики. — М.: Финансы и статистика, 1982. — 344 с.
Pollard J. H. A Handbook of Numerical and Statistical Techniques. — Cambridge: Cambridge University Press, 1977. — 344 p. (Russ. ed.: *Pollard Dzh.* Spravochnik po vychislitelnym metodam statistiki. — Moscow: Finansy i statistika, 1982. — 344 p.)

Сидоренко Е. В. Методы математической обработки в психологии. — СПб.: Речь, 2001. — 350 с.
Sidorenko E. V. Metody matematicheskoi obrabotki v psikhologii [Methods of mathematical processing in psychology]. — Saint Petersburg: Rech Publishing House, 2001. — 350 p. (in Russian).

Alves A. D., Yanasse H. H., Soma N. Y. Benford's Law and articles of scientific journals: comparison of JCR and Scopus data // Scientometrics. — 2014. — Vol. 98. — P. 173–184.

Andriotis P., Oikonomou G., Tryfonas T. JPEG steganography detection with Benford's Law // Digital Investigation. — 2013. — Vol. 9, no. 3–4. — P. 246–257.

Aronoff M., Rees-Miller J. (eds.) The Handbook of Linguistics. — Oxford (a.o.): Blackwell Publishing, 2004. — 824 p.

Battersby S. Statistics hint at fraud in Iranian election // New Scientist. — 24 June 2009.

Benford F. The law of anomalous numbers // Proceedings of American Philosophical Society. — 1938. — Vol. 78, no. 4. — P. 551–572.

Biau D. The first-digit frequencies in data of turbulent flows // Physica A. — 2015. — Vol. 440. — P. 147–154.

Binongo J. N. Who wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution // Chance. — 2003. — Vol. 16. — P. 9–17.

Brocardo M. L., Traore I., Woungang I., Obaidat M. S. Authorship verification using deep belief network systems // International Journal of Communication Systems. — 2017. — DOI: 10.1002/dac.3259

Hill T. P. A Statistical Derivation of the Significant-Digit Law // Statistical Science. — 1995. — Vol. 10. — P. 354–363.

Hill T. P., Fox R. F. Hubble's Law Implies Benford's Law for Distances to Galaxies // Journal of Astrophysics and Astronomy. — 2016. — Vol. 37, no. 4. — 8 p.

Mitkov R. (ed.) The Oxford Handbook of Computational Linguistics. — Oxford (a.o.): Oxford University Press, Inc., 2003. — 786 p.

Nigrini M. Benford's Law: applications for forensic accounting, auditing, and fraud detection. — Hoboken: John Wiley & Sons, Inc., 2012. — XX + 330 p.

Pain J.-C. Regularities and symmetries in atomic structure and spectra // High Energy Density Physics. — 2013. — Vol. 9, no. 3. — P. 392–401.

Ryabko B., Astola J., Malyutov M. Compression-Based Methods of Statistical Analysis and Prediction of Time Series. — Switzerland: Springer International Publishing, 2016. — 144 p.

Sambridge M., Tkalčić H., Arroucau P. Benford's Law of First Digits: from Mathematical Curiosity to Change Detector // Asia Pacific Mathematics Newsletter. — 2011. — Vol. 1, no. 4. — P. 1–6.

The Project Gutenberg eBooks, http://www.gutenberg.org

The Best American Humorous Short Stories, by George P. Morris, Edgar A. Poe, Caroline M. S. Kirkland, Eliza Leslie, George W. Curtis, Edward E. Hale, Oliver W. Holmes, Mark Twain, Harry S. Edwards, Richard M. Johnston, Henry C. Bunner, Frank R. Stockton, Francis Bret Harte, O. Henry, George R. Chester, Grace MacGowan Cooke, William J. Lampton, and Wells Hastings. The Project Gutenberg eBook, eBook #10947.

The Short-story, by Washington Irving, Edgar A. Poe, Nathaniel Hawthorne, Francis Bret Harte, Robert L. Stevenson, Rudyard Kipling. The Project Gutenberg eBook, transcribed from the 1916 Allyn and Bacon edition, eBook # 21964.

The Lock And Key Library, Classic Mystery And Detective Stories, by Rudyard Kipling, A. Conan Doyle, Egerton Castle, Stanley J. Weyman, Wilkie Collins, and Robert L. Stevenson. The Project Gutenberg eBook, transcribed from the 1909 Review of Reviews Co. edition, eBook # 2038.

Shorter Novels, Eighteenth Century. The History of Rasselas, The Castle of Otranto, Vathek, by Samuel Johnson, Horace Walpole, and William Beckford. The Project Gutenberg eBook, transcribed from the 1903 Aldine House edition, eBook # 34766.

The Best of the World's Classics, Vol. V — Great Britain and Ireland, by James Boswell, William Wordsworth, Walter Scott, Samuel T. Coleridge, Robert Southey, Walter Savage Landor, Charles Lamb, William Hazlitt, Thomas De Quincey, Lord Byron, Percy Bysshe Shelley, George Grote, Thomas Carlyle, Lord Macaulay. The Project Gutenberg eBook, transcribed from the 1909 Funk & Wagnalls Co. edition, eBook # 22182.

The Great English Short-Story Writers, Vol. 1, by Daniel Defoe, James Hogg, Washington Irving, Nathaniel Hawthorne, Edgar A. Poe, John Brown, Charles Dickens, Frank R. Stockton, Mark Twain, Francis Bret Harte, Thomas Hardy, Henry James, and Robert L. Stevenson. The Project Gutenberg eBook, transcribed from the 1910 Readers's Library edition, eBook # 10135.

A House to Let, by Charles Dickens, Wilkie Collins, Elizabeth Gaskell, and Adelaide A. Procter. The Project Gutenberg eBook, transcribed from the 1903 Chapman and Hall edition, eBook #2324.

Masterpieces of Mystery, Vol. 1, Ghost Stories, by Algernon Blackwood, Montague Rhodes James, Katherine Rickford, William F. Harvey, Ralph Adams Cram, Robert L. Stevenson, and Wilbur D. Steele. The Project Gutenberg eBook, transcribed from the 1920 Doubleday, Page & Co. edition, eBook # 27722.