УДК 51-78, 519.234.3, 519.257, 81-139, 519.248.6

Зенков Андрей Вячеславович,
кандидат физико-математических наук, доцент
ФГАОУ ВО «УрФУ имени первого Президента России Б.Н.Ельцина»
e-mail: zenkow@mail.ru
г. Екатеринбург, Россия

НОВЫЙ СТАТИСТИЧЕСКИЙ МЕТОД АТРИБУЦИИ ТЕКСТОВ

Аннотация:
Предложен новый метод статистического анализа в текстологии. Исследовано распределение частот различных первых значащих цифр в числительных связных авторских русскоязычных текстов. Показано, что эти частоты приближённо соответствуют закону Бенфорда с резким преобладанием доли единицы. Отклонения от закона Бенфорда являются статистически устойчивыми авторскими особенностями, позволяющими при некоторых условиях исследовать вопрос об авторстве, в частности, различать тексты разных авторов. Распределение цифр конца ряда $\{1,2,\ldots,7,8,9\}$ подвержено сильным флуктуациям и непоказательно. Предложенный подход проиллюстрирован и выводы подкреплены примерами компьютерного анализа произведений М. Агеева, В. Набокова, М. Шолохова, Н. Некрасова и др. Результаты обоснованы на основе непараметрического *U*-критерия Манна-Уитни и иерархического кластерного анализа.

Ключевые слова:
Закон Бенфорда, стилеметрия, атрибуция текстов, обработка текстов, критерий Манна-Уитни, иерархический кластерный анализ

Введение

В последние годы заметно расширилась сфера практического использования известного уже больше ста лет закона Бенфорда [Benford 1938]. Закон Бенфорда описывает вероятность появления определённой первой значащей цифры в разнообразных распределениях величин, взятых из реальной жизни. Вопреки здравому предположению о том, что частоты появления любой первой значащей цифры должны быть равными, для многих массивов данных в качестве первой значащей цифры чаще других встречается единица! Согласно закону Бенфорда при записи числа в десятичной

системе счисления вероятность появления цифры d в качестве его первой значащей цифры

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

так что $d = 1$ должна встречаться с вероятностью $\lg 2 \approx 0{,}30$, $d = 2$ – с вероятностью 0,18 и т.д.

Исчерпывающего объяснения закона Бенфорда, охватывающего все случаи реализации, до сих пор не предложено, хотя и сформулированы некоторые условия, благоприятствующие его появлению. Один из классических опытов Бенфорда, хорошо согласующийся с (1) – подсчет встречаемости числительных на произвольных страницах прессы – находит логичное объяснение в теореме Хилла [Hill 1995; Berger, Hill 2015], согласно которой в условиях неоднократного случайного выбора распределения вероятностей с последующим случайным выбором числа согласно этому распределению возникает набор чисел, подчиняющийся закону Бенфорда.

Неполнота понимания не препятствует успешному применению закона Бенфорда для выявления подлогов в бухгалтерской отчётности [Nigrini 2012] и фальсификаций на выборах [Roukemaa 2014]; обсуждаются применения в различных науках; как иллюстрацию 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].

Нами показана перспективность подсчета частот различных первых значащих цифр числительных в лингвистике – для задач текстологии [Зенков 2015]. Оказалось, что не только для случайной комбинации текстов, но и для с в я з н ы х текстов, для которых нарушается условие названной теоремы, распределение частот приближается к (1), но доля единицы заметно превышает 30% – хотя бы потому, что, формально являясь числительным, слово «один» фактически может выступать в роли неопределенного артикля.

В отличие от традиционной методологии применения закона Бенфорда, трактующей отклонения от закона как указание на возможное наличие «фальсификаций» (в широком понимании), нами сделан акцент на сравнении этих отклонений для текстов разных авторов; показано, что эти отклонения являются статистически устойчивыми авторскими особенностями, позволяющими различать тексты разных авторов (при некоторых условиях, важнейшее из которых – достаточно большая длина текста).

В настоящей работе данный подход развит, и представлены новые результаты исследований.

Работа носит экспериментальный характер. Цель теоретического обоснования результатов (если таковое, вообще, возможно) не ставилась, что, однако, не умаляет применимости предложенной методологии для практических задач текстологии.

Для всех (русскоязычных) текстов, подвергнутых статистическому анализу, с помощью ЭВМ подсчитывались частоты появления различных первых значащих цифр в количественных и порядковых числительных, выраженных как цифрами, так и (значительно чаще) словесно. В последнем случае вначале числительные переводились в цифровую форму записи, так что, например, для числительного «тысяча четыреста» (1400) учитывалась только первая значащая цифра 1.Для выявления авторского употребления числительных предварительно из текста удалялись идиоматические выражения, случайно содержащие числительные («семь пятниц на неделе»).

Распознавание авторства текстов

Авторство «Романа с кокаином»

На протяжении шестидесяти лет в российском литературоведении оставался нерешённым вопрос об авторстве «Романа с кокаином», опубликованного в 1934г. под псевдонимом «М. Агеев». В отсутствие достоверной информации об авторе и каких-либо других значимых публикаций под этим именем получила распространение гипотеза о литературной мистификации. В силу некоторой жанровой и стилистической близости «Романа с кокаином» ранним романам В.В. Набокова перу последнего стали приписывать и роман М. Агеева. Публикация в 1990-х гг. ранее неизвестных архивных материалов [Сорокина, Суперфин 1994] опровергла эту гипотезу. Хотя данный частный филологический вопрос уже снят, покажем, к каким результатам приводит бенфордовская методология.

Ниже приведены результаты статистического исследования «Романа с кокаином» (Рис. 1) и русскоязычных произведений Набокова (на Рис. 2, 3 в качестве примера приведены результаты для двух романов). Отметим резкое различие во встречаемости значащей цифры 1 в романе Агеева, с одной стороны, и в романах Набокова, с другой стороны. С учетом длины проанализированных текстов это различие трудно объяснить случайными флуктуациями (в отличие от последующих значащих цифр, для которых даже в книгах одного автора не усматривается общая закономерность). Это характерные авторские различия стилей. Мы склонны связать их с психологическими особенностями (в частности, склонностью к округлению чисел), которые, независимо от воли и сознания автора, сказываются на его текстах. Для Агеева, по указанной выше причине, материал для сравнения отсутствует, но все произведения первого (русскоязычного) периода творчества Набокова имеют аналогичную встречаемость единицы как первой значащей цифры.

Рис. 1. Распределение первых значащих цифр числительных в «Романе с кокаином» Агеева (1934г.). Результаты здесь и ниже сопоставляются с ожидаемыми согласно закону Бенфорда

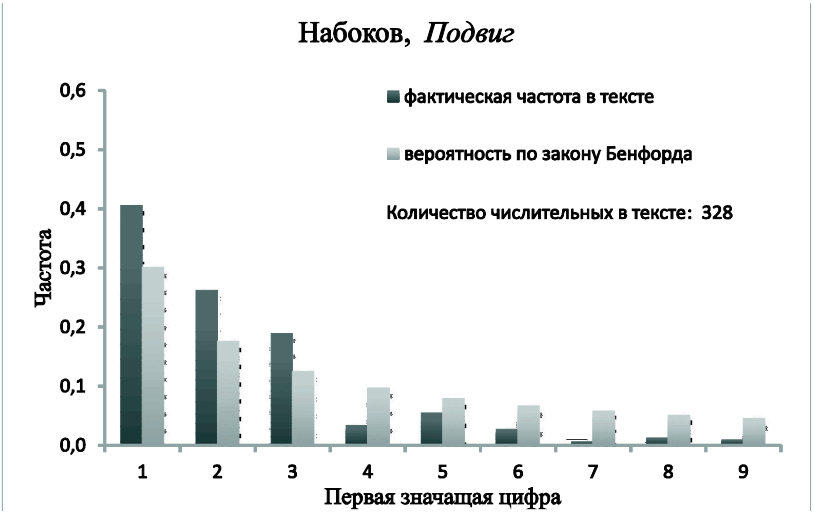

Рис. 2. Распределение первых значащих цифр числительных в романе Набокова «Подвиг» (1931г.)

Рис. 3. Распределение первых значащих цифр числительных в романе Набокова «Дар» (1937г.)

Разумеется, сравнение распределений не может основываться только на выявлении субъективных визуальных сходства и различий между ними. Нами применен непараметрический *U*-критерий Манна-Уитни. Нулевая гипотеза H_0, утверждающая отсутствие значимых различий в рассмотренных распределениях, оказалась отвергнутой и принятой именно в тех случаях, как описано выше. Различие между романами Набокова оказалось незначимым, а «Роман с кокаином» Агеева значимо отличается от каждого из них.

Эти выводы подтверждаются и дендрограммой, визуализирующей результаты иерархического кластерного анализа распределений частот первых значащих цифр числительных в текстах с точки зрения сходства/различия этих распределений. Для кластеризации здесь и ниже использован метод межгрупповых связей (average linkage between groups) [Gan et al. 2007] (как сбалансированный метод, избегающий крайностей методов ближайшего и дальнего соседей) с чебышевской метрикой, определяющей расстояние ρ между n-мерными числовыми векторами $\boldsymbol{x}$ и $\boldsymbol{y}$ как максимум модуля разности их компонент: $\rho(\boldsymbol{x},\boldsymbol{y}) = \max_{1\le i\le n} |x_i - y_i|$. В нашем случае компонентами векторов являются частоты той или иной первой значащей цифры в каждом из двух анализируемых текстов. Очевидно, что максимум модуля разности может достигаться на том значении i, $i = 1,2,\ldots,9$, для которого частота изначально не мала, а это, как правило, значащие цифры 1, 2, 3. Но именно частоты этих цифр (особенно, единицы) и определяют специфику

текста в нашей методологии, чем и обусловлен выбор чебышевской метрики.

Мы произвели кластеризацию для «Романа с кокаином» и почти всех романов Набокова, написанных по-русски или имеющих авторский перевод на русский язык (Рис. 4). Расстояние ρ отсчитывается по горизонтальной шкале, чем оно больше, тем менее похожи анализируемые объекты (тексты). Роман Агеева стоит особняком среди всех учтенных текстов, присоединяясь к ним на финальной стадии кластеризации.

Итак, статистический метод, основанный на подсчете первых значащих цифр числительных, способен ответить на вопрос об авторстве текста.

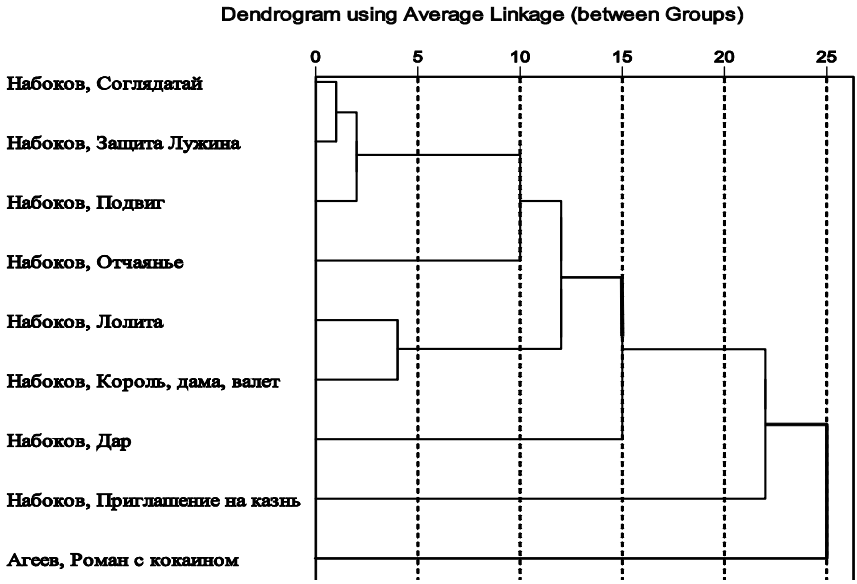

Рис. 4. Дендрограмма кластеризации распределения частот первых значащих цифр числительных в текстах Набокова и Агеева

Проблема «Тихого Дона»

Другой известной проблемой атрибуции текстов является вопрос об авторстве романа «Тихий Дон» и, шире, всего литературного наследия М.А. Шолохова. Имеются веские аргументы в пользу версии о плагиате и некоторые доводы против нее. Роман содержит восемь частей, объединенных в четыре книги. Лингвистическое и статистическое изучение романа привело многих исследователей к выводу о том, что текст крайне неоднороден; авторство первых частей (или, по крайней мере, их литературной первоосновы, использованной Шолоховым) многими специалистами приписывается

писателю Ф.Д. Крюкову, хотя есть и другой кандидат – В.А. Краснушкин, а в тексте последующих частей усматривают стиль А.С. Серафимовича, Б.А. Пильняка, А.А. Фадеева (неисчерпывающий список). Высказывалось мнение и о том, что сомнительно авторство не только «Тихого Дона»; что «Поднятая целина» и «Они сражались за Родину» также написаны не Шолоховым, а другими авторами (в частности, называлась фамилия А.П. Платонова) [Кузнецов 2003].

Не вдаваясь подробно в филологический обзор состояния проблемы, приведем результаты нашего статистического исследования в рамках бенфордовской методологии.

Во-первых, нами проведен статистический анализ трех романов Шолохова (Рис. 5). Распределение первых значащих цифр числительных в них очень различно, при том, что обычно это распределение характерно для автора.

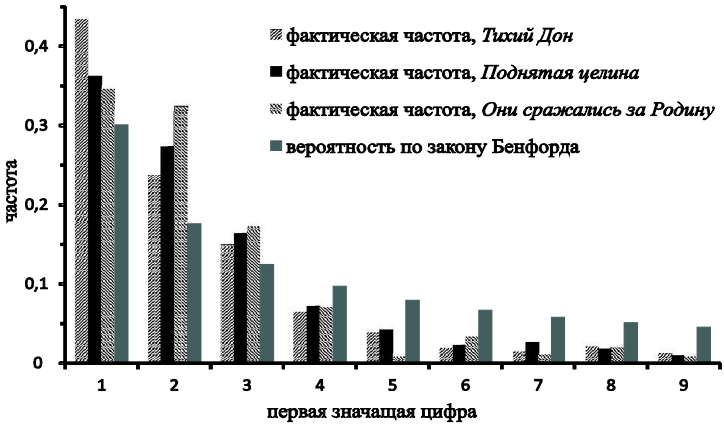

Рис. 5. Распределение первых значащих цифр числительных в романах Шолохова «Тихий Дон», «Поднятая целина», «Они сражались за Родину»

Данный результат сделал необходимым более детальный сопоставительный анализ основных произведений, приписываемых Шолохову, и текстов некоторых авторов, в которых видят истинных создателей этих произведений. Дендрограмма кластеризации распределений первых значащих цифр числительных представлена на Рис. 6. Некоторые выводы (подтверждаемые критерием Манна-Уитни):

1) Разные части «Тихого Дона» и «Поднятой целины» распределяются по разным кластерам, что говорит о внутренней статистической неоднородности текстов с точки зрения распределения первых значащих цифр числительных (ср. со статистически близкими «Разгромом» и «Молодой гвардией» Фадеева);

2) Предположения о том, что Платонов, Пильняк, Серафимович могли участвовать в создании текста «Тихого Дона» и первой книги «Поднятой целины», не лишены основания;
3) Авторство Краснушкина в отношении «Тихого Дона» более сомнительно;
4) «Они сражались за Родину» и вторая книга «Поднятой целины», хронологически создававшиеся в одну эпоху, могут принадлежать одному автору;
5) Тексты Крюкова статистически близки началу «Тихого Дона».
6) В высшей степени сомнительно, что «Донские рассказы», с одной стороны, и «Тихий Дон», «Поднятая целина», «Они сражались за Родину» принадлежат одному автору.

Эти выводы хорошо согласуются с кратко описанными выше результатами, полученными другими (в основном, филологическими) методами.

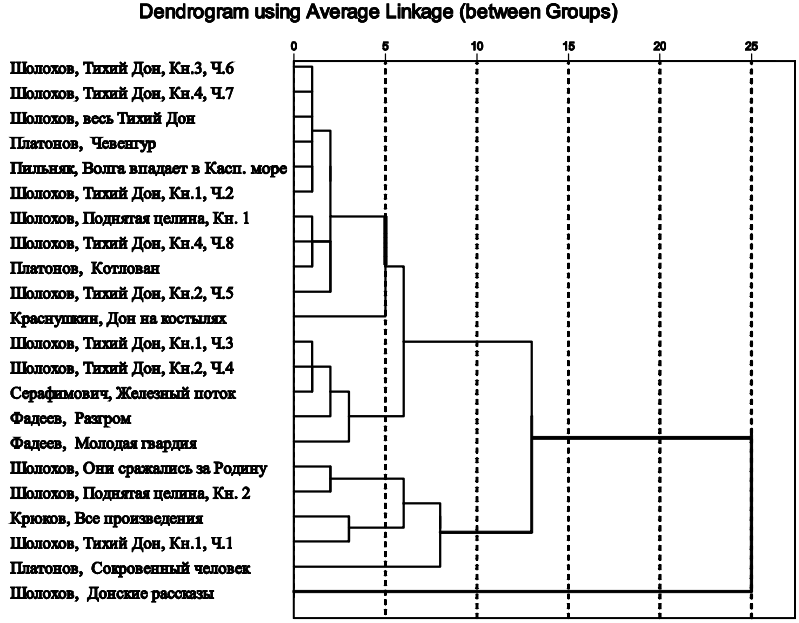

Рис. 6. Дендрограмма кластеризации распределения частот первых значащих цифр числительных в текстах Шолохова и предполагаемых авторов приписываемых ему книг

Заметим, что (все доступные для анализа) тексты Крюкова сравнительно невелики по объему, и для статистического анализа пришлось объединить их в один файл. То же – для «Донских рассказов» Шолохова. Доступным для анализа оказался только один текст Краснушкина («Дон на косты-

лях»). Рассказ «Сокровенный человек» Платонова сравнительно невелик, что могло сказаться на статистической значимости результатов для него.

Итак, бенфордовский анализ может быть полезен при исследовании вопроса об а в т о р с т в е текстов.

Проверка методологии: ранняя проза Н.А. Некрасова

Интересную возможность проверки нашей идеи о связи авторства текста с его статистическими характеристиками предоставляют романы "Три страны света" и «Мертвое озеро» написанные Н.А. Некрасовым, несравненно более известным как поэт, в начале его литературной карьеры совместно с А.Я. Панаевой и впервые опубликованные в 1848–1849 и 1851гг., соответственно.

Рукописи романов не сохранились, поэтому в вопросе о распределении труда между соавторами значимы их собственные свидетельства. В "Воспоминаниях" Панаевой сообщается, что в написании романа "Три страны света" принимали участие оба – и Некрасов, и она; что же касается «Мертвого озера», то участие Некрасова ограничилось разработкой сюжета и написанием незначительной части текста. Руководствуясь филологическими соображениями, литературоведы, вопреки свидетельству Панаевой, усматривают в о б о и х романах существенную часть текста, написанную Некрасовым (с указанием конкретных глав) [Некрасов 1965; Некрасов 1985].

Нами выполнен подсчет частот различных первых значащих цифр числительных в частях каждого из романов, приписываемых литературоведами конкретным авторам (Некрасов, Панаева), и, для сравнения, аналогичный анализ для «Воспоминаний» Панаевой и ранних прозаических произведений, единоличным автором которых является Некрасов (Рис. 7).

Некоторые выводы:

1) Распределение первых значащих цифр числительных в частях «Мертвого озера», приписываемых Некрасову и Панаевой, в целом схоже и сопоставимо с результатами для части «Трех стран света», приписываемой Панаевой (за исключением цифры 3, в которой график обнаруживает выброс). Для «Воспоминаний» Панаевой получены похожие результаты.
2) Распределение первых значащих цифр числительных в части «Трех стран света», приписываемой Некрасову, существенно отличается от трех указанных выше распределений, но схоже с результатами для ранней художественной прозы Некрасова. Не исключено участие Панаевой в написании и этой части романа.
3) Отсюда следует, что разные части «Мертвого озера», вероятно, написаны одним автором, а именно – Панаевой, а разные части «Трех стран света», действительно, имеют разное авторство.

4) Итак, нет оснований не доверять Панаевой в ее свидетельстве о процессе написания двух ее совместных с Некрасовым романов.

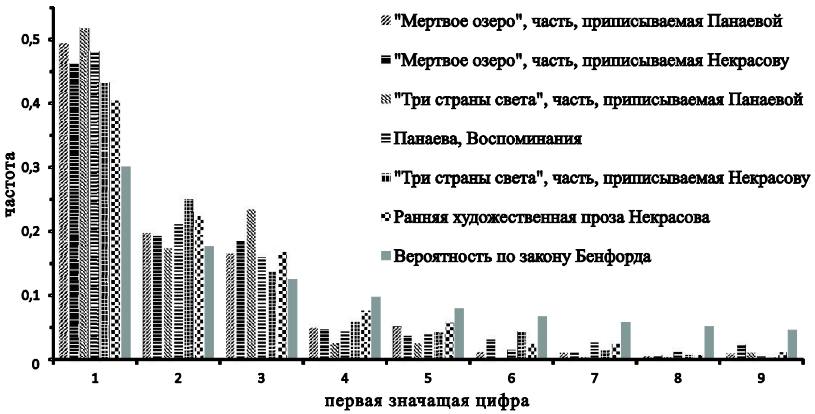

Рис. 7. Распределение первых значащих цифр числительных в текстах Некрасова и Панаевой [1]

Мы полагаем, что разработанная нами методология может быть полезным дополнением к традиционным текстологическим практикам учета длины предложений, длины слов, частот употребления служебных слов и определенных знаменательных частей речи и т.д. [Ryabko 2016].

Заключение

1) Закон Бенфорда приближенно выполняется для связных текстов.

2) Отклонения от закона Бенфорда являются статистически значимыми устойчивыми авторскими особенностями. Существенное различие этих отклонений позволяет 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 превышает вероятность согласно закону Бенфорда для значащих цифр 1, 2, 3; для последующих цифр ситуация обратна. Распределение цифр конца ряда $\{1,2,\ldots,7,8,9\}$ подвержено сильным флуктуациям и непоказательно.

[1] В текст, обозначенный на рисунке 7 как ранняя художественная проза Некрасова, включены «Повесть о бедном Климе», «Жизнь и похождения Тихона Тростникова», «Сургучов», «Тонкий человек, его приключения и наблюдения», "В тот же день часов в одиннадцать утра..." [Некрасов 1984].

Список использованных источников

1. Alves et al. 2014 – 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. Pp. 173–184.
2. Andriotis et al. 2013 – Andriotis P., Oikonomou G., Tryfonas T. JPEG steganography detection with Benford's Law. *Digital Investigation*. 2013. Vol. 9. No. 3–4. Pp. 246–257.
3. Benford 1938 – Benford F. The law of anomalous numbers. *Proceedings of American Philosophical Society*. 1938. Vol. 78. No. 4. Pp. 551–572.
4. Berger, Hill 2015 – Berger A., Hill T. P. *An Introduction to Benford's Law*. Princeton: Princeton University Press, 2015.
5. Biau 2015 – Biau D., The first-digit frequencies in data of turbulent flows. *Physica* A. 2015. Vol. 440, Pp. 147–154.
6. Gan et al. 2007 – Gan G., Ma C., Wu J. *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia: SIAM, 2007.
7. Hill 1995 – Hill T. P. A Statistical Derivation of the Significant-Digit Law. *Statistical Science*. 1995. Vol. 10. Pp. 354–363.
8. Hill, Fox 2016 – 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 pages.
9. Nigrini 2012 – Nigrini M. J. *Benford's Law: applications for forensic accounting, auditing, and fraud detection*. Hoboken: John Wiley & Sons, 2012.
10. Pain 2013 – Pain J.-C. Regularities and symmetries in atomic structure and spectra. *High Energy Density Physics*. 2013. Vol. 9. No. 3. Pp. 392–401.
11. Roukemaa 2014 – Roukemaa B. F. A first-digit anomaly in the 2009 Iranian presidential election. *Journal of Applied Statistics*. 2014. Vol. 41. No. 1. Pp. 164–199.
12. Ryabko 2016 – Ryabko B., Astola J., Malyutov M. *Compression-Based Methods of Statistical Analysis and Prediction of Time Series*. Springer International Publishing Switzerland, 2016.
13. Sambridge et al. 2011 – 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. Pp. 1–6.
14. Зенков 2015 – Зенков А. В. Отклонения от закона Бенфорда и распознавание авторских особенностей в текстах // Компьютерные исследования и моделирование. 2015. Т. 7, вып. 1. С. 197–201.
15. Кузнецов 2003 – Новое о Михаиле Шолохове: Исследования и материалы / Ф.Ф. Кузнецов и др. (ред.). М.: ИМЛИ РАН, 2003.
16. Некрасов 1965 – Некрасов Н.А. Три страны света. Ярославль: Верхне-Волжское книжное издательство, 1965.

17. Некрасов 1984 –Некрасов Н.А. Художественная проза. Незаконченные романы и повести 1841–1856 гг. Полное собрание сочинений и писем в пятнадцати томах, Том 8. Л.: Наука, 1984.
18. Некрасов 1985 – Некрасов Н.А. Мертвое озеро. Полное собрание сочинений и писем в пятнадцати томах, Том 10 книга I, Л.: Наука, 1985.
19. Сорокина, Суперфин 1994 – Сорокина М. Ю., Суперфин Г. Г. «Был такой писатель Агеев…»: версия судьбы или о пользе наивного биографизма // Минувшее: Исторический альманах. Вып. 16. М., СПб.: Феникс-Атенеум, 1994. С. 265–289.

Andrei Zenkov,
PhD, Associate professor
Ural Federal University named after the first President of Russia Boris Yeltsin
e-mail: zenkow@mail.ru
Ekaterinburg, Russia

A NEW STATISTICAL METHOD OF TEXT ATTRIBUTION

Abstract:
A new method of statistical analysis of texts is suggested. The frequency distribution of the first significant digits in numerals of connected authorial Russian-language texts is considered. Benford's law is found to hold approximately for these frequencies with a marked predominance of the digit 1. Deviations from Benford's law are statistically significant author peculiarities that allow, under certain conditions, to consider the problem of authorship and distinguish between texts by different authors. At the end of $\{1,2,\ldots,7,8,9\}$ row, the digits distribution is subject to strong fluctuations and thus unrepresentative for our purpose. The approach proposed and the conclusions are backed by the examples of the computer analysis of works by M. Ageev, V. Nabokov, M. Sholokhov, N. Nekrasov et al. The results are confirmed on the basis of non-parametric Mann-Whitney U test and hierarchical cluster analysis.
Keywords:
Benford's law, stylometry, text attribution, text processing, Mann-Whitney U test, hierarchical cluster analysis