ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2023. Том 16. Выпуск 11 | 2023. Volume 16. Issue 11
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Литературные мистификации и авторское использование числительных

Зенков А. В.

Аннотация. Настоящее исследование относится к стилометрии. Известны случаи, когда писатель, добившийся известности, по разным причинам начинает творить под другим именем, пытается писать в другой манере и порой снова добивается успеха в новом воплощении. Цель исследования – проверка осуществимости намеренного значительного изменения авторского литературного стиля. В качестве маркера стиля используются числительные, присутствующие в текстах того или иного автора. На примерах из англо-, франко- и русскоязычной литературы показано, что использование числительных является авторским инвариантом, который проявляется во всех или большинстве достаточно длинных текстов данного автора. Полученные результаты показали, что, вопреки попыткам автора писать «по-новому», манера использования числительных консервативна и позволяет распознавать фиктивное авторство. Этот вывод сделан на основании анализа произведений Р. Гари и Б. Акунина (Г. Чхартишвили), известных своими литературными мистификациями. Анализ употребления числительных применён также к проблеме авторства романа «Убить пересмешника» Х. Ли. Выводы о схожести/различии литературных стилей сделаны на основе иерархического кластерного анализа и подкреплены критерием Пирсона. Научная новизна работы состоит в новом подходе к поиску авторского инварианта и атрибуции текстов.

EN

Literary mystifications and the authorial use of numerals

Zenkov A. V.

Abstract. This study pertains to stylometry. There are cases when a writer who has achieved fame, for various reasons, begins to create under a different name, attempts to write in a different manner and sometimes achieves success again in a new incarnation. The aim of the study is to test the feasibility of intentionally making significant changes to an author's literary style. Numerals present in the texts by a particular author are used as a style marker. Examples from English, French and Russian literature demonstrate that the use of numerals is a literary 'fingerprint' that manifests in all or most of sufficiently long texts by that author. The obtained results show that, contrary to an author's attempts to write in a 'new' way, the usage of numerals is conservative and allows for the recognition of fictitious authorship. This conclusion is drawn based on the analysis of works by R. Gary and B. Akunin (G. Chkhartishvili), who are known for their literary hoaxes. The analysis of numerals usage is also applied to the issue of authorship regarding Harper Lee's novel 'To Kill a Mockingbird'. Conclusions about the similarity/difference of literary styles are made based on hierarchical cluster analysis and are supported by the Pearson chi-squared test. The scientific originality of the paper lies in taking a new approach to the search for a literary 'fingerprint' and text attribution.

Введение

Актуальность. В стилометрии есть актуальная, до конца не решённая задача поиска авторского инварианта (fingerprint, «*отпечаток пальца*» – англ.) – количественного признака (или совокупности признаков), величина которого примерно постоянна и индивидуальна для всех (или большинства) текстов данного автора. Авторский инвариант был бы полезен, в частности, в задачах определения авторства текстов: написаны ли данные тексты одним автором? Написаны ли они данным конкретным автором? Кто из круга предполагаемых авторов скорее всего является автором данного текста? И т. д. Разумеется, ответы на поставленные вопросы всегда имеют вероятностный характер.

К сожалению, в стилометрии, при всём обилии предложенных количественных методов, до сих пор отсутствует такой, который не дал бы заведомо абсурдного результата на каком-нибудь проверочном примере.

Научная статья (original research article) | https://doi.org/10.30853/phil20230568

К традиционным практикам в стилометрии относятся нахождение средней длины слов и предложений, частоты определённых знаменательных и/или служебных частей речи, частоты *n*-грамм и т. п. (Stamatatos, 2009; Tempestt, Kalaivani, Aneez et al., 2017). Затем полученные численные данные обрабатываются в рамках некоторого вычислительного аппарата от теории вероятностей и математической статистики до кибернетики и теории информации. Казалось бы, совместное использование нескольких методов должно повышать надёжность получаемых результатов, но зачастую они противоречат друг другу.

Большие надежды внушает использование искусственного интеллекта (Brocardo, Traore, Woungang et al., 2017; La Inteligencia Artificial ayuda a descubrir una obra desconocida de Lope de Vega en los fondos de la BNE, Biblioteca Nacional de España. 2023. https://www.bne.es/es/noticias/inteligencia-artificial-ayuda-descubrir-obra-desconocida-lope-vega-fondos-bne), но проблема состоит в непрозрачности работы нейронных сетей и трудности истолкования результатов.

Задачи настоящего исследования:

1. Проверить, имеются ли авторские различия в употреблении числительных в литературных текстах.
2. Выяснить, имеются ли закономерности в распределении частот встречаемости числительных.
3. Установить, отличается ли использование числительных в произведениях, подписанных подлинным именем автора, и произведениях, опубликованных под псевдонимом.

Нами разработан метод решения задач стилометрии, основанный на анализе использования числительных в (литературном) авторском тексте (Zenkov, 2018; 2021; Zenkov, Místecký, 2019; 2022). Этот метод имеет немало преимуществ перед традиционными. Во-первых, в силу самой природы числительных к ним легко применить количественную меру. Во-вторых, получаемые результаты допускают прозрачное филологическое толкование. В-третьих, встречаемость числительных в тексте практически инвариантна относительно перевода текста на другой язык.

Как и всякий статистический метод стилометрии, метод учёта числительных требует достаточно большой длины текста (файлы размером от десятков кБ в кодировке UTF-8).

В истории литературы известны примеры, когда автор творил под разными именами, и иногда эти литературные мистификации оказывались успешными. В тех примерах, которые мы рассмотрим ниже, речь шла не просто о смене имени, но о попытке писать «иначе». Возникает вопрос: может ли сознательное намерение автора изменить свой литературный стиль повлиять на характер использования числительных в тексте?

Рассмотрению этого вопроса и посвящена наша работа. Она построена следующим образом.

После описания методики исследования (Раздел 1) следует сопоставительный анализ литературных текстов разных авторов, демонстрирующий постоянство авторских особенностей встречаемости числительных в текстах. Это показано на примере англо-, франко- и русскоязычных текстов (Раздел 2).

В Разделе 3 наша стилометрическая техника применяется к анализу литературных текстов Ромена Гари и Бориса Акунина, известных своими литературными мистификациями, экспериментировавших со стилем и публиковавшихся под несколькими псевдонимами. Затем мы исследуем проблему авторства литературных текстов Х. Ли, в существенном влиянии на которые подозревается Т. Капоте.

Работа завершается заключением и подведением итогов.

В Приложение вынесены некоторые вопросы вычислительного характера.

Теоретическая база настоящего исследования подробно изложена в опубликованных работах автора (Zenkov, 2018; 2021; Zenkov, Místecký, 2019; 2022). Подобно тому, как в других методах стилометрии учитываются некоторые количественные статистические характеристики (средняя длина слов, предложений и т. п.) авторских текстов (Stamatatos, 2009; Tempestt, Kalaivani, Aneez et al., 2017), независимо от сознательного намерения авторов проявляющиеся в достаточно длинных текстах, метод, применённый в настоящей работе, тоже опирается на статистику, но акцент сделан на частотах встречаемости числительных. Автору известны только две работы предшественников, в которых разрабатывались отчасти сходные идеи (Benford, 1938; Hungerbühler, 2007).

Материалами исследования являются англо-, франко- и русскоязычные художественные тексты, среди критериев отбора которых, в частности, были большой размер произведений и нахождение текстов в свободном доступе в сети Интернет. Сравнительному анализу с точки зрения встречаемости числительных были подвергнуты на языке оригинала следующие произведения (перечисленные в том порядке, как они выстроились на дендрограммах (Рис. 1-4)):

- Англоязычные:

1. Ч. Диккенс: *Наш общий друг* (*Our Mutual Friend*); *Крошка Доррит* (*Little Dorrit*); *Дэвид Копперфильд* (*David Copperfield*); *Домби и сын* (*Dombey and Son*).

2. У. М. Теккерей: *История Генри Эсмонда* (*The History of Henry Esmond*); *История Пенденниса* (*The History of Pendennis*); *Мемуары Барри Линдона* (*The Memoires of Barry Lyndon*); *Ярмарка тщеславия* (*Vanity Fair*).

3. Г. Дж. Уэллс: *Война миров* (*The War of the Worlds*); *Остров доктора Моро* (*The Island of Doctor Moreau*); *Человек-невидимка* (*The Invisible Man*); *Машина времени* (*The Time Machine*).

4. В. В. Набоков (сочинения, написанные по-английски): *Бледный огонь* (*Pale Fire*); *Ада, или радости страсти* (*Ada, or Ardor*); *Смотри на арлекинов!* (*Look at the Harlequins!*); *Подлинная жизнь Себастьяна Найта* (*The Real Life of Sebastian Knight*); *Прозрачные вещи* (*Transparent Things*); *Под знаком незаконнорождённых* (*Bend Sinister*).

- Франкоязычные:

1. М. Пруст: *В поисках утраченного времени* (*À la recherche du temps perdu*) – вся гепталогия: *По направлению к Свану* (*Du côté de chez Swann*); *Под сенью девушек в цвету* (*À l'ombre des jeunes filles en fleurs*); *У Германтов* (*Le côté de Guermantes*); *Содом и Гоморра* (*Sodome et Gomorrhe*); *Пленница* (*La prisonnière*); *Исчезнувшая Альбертина* (*Albertine disparue*); *Обретённое время* (*Le Temps retrouvé*).

2. Э. Золя: *Жерминаль* (*Germinal*); *Разгром* (*La débâcle*); *Чрево Парижа* (*Le Ventre de Paris*); *Западня* (*L'assommoir*); *Проступок аббата Муре* (*La faute de l'abbé Mouret*); *Карьера Ругонов* (*La fortune des Rougon*); *Деньги* (*L'argent*); *Дамское счастье* (*Au bonheur des Dames*); *Земля* (*La terre*); *Мечта* (*Le rêve*); *Доктор Паскаль* (*Le docteur Pascal*); *Радость жизни* (*La Joie de vivre*); *Страница любви* (*Une page d'amour*); *Добыча* (*La curée*); *Его превосходительство Эжен Ругон* (*Son Excellence Eugène Rougon*); *Нана* (*Nana*).

3. Г. де Мопассан: *Милый друг* (*Bel ami*); *Пьер и Жан* (*Pierre et Jean*); *Жизнь* (*Une vie*); *Сильна как смерть* (*Fort comme la mort*); *Наше сердце* (*Notre cœur*).

4. Ф. Мориак: *Тереза Дескейру* (*Thérèse Desqueyroux*); *Клубок змей* (*Le Nœud de vipères*).

5. А. Доде: *Бессмертный* (*L'Immortel*); *Малыш* (*Le petit chose*).

6. А. Жид: *Тесные врата* (*La porte étroite*); *Записные книжки Андре Вальтера* (*Les cahiers d'André Walter*); *Школа жён* (*L'école des femmes*); *Женевьева* (*Geneviève*); *Фальшивомонетчики* (*Les faux-monnayeurs*).

7. Ж. Верн: *Пятнадцатилетний капитан* (*Un capitaine de quinze ans*); *Вокруг света в 80 дней* (*Le tour du monde en quatre-vingts jours*).

- Русскоязычные:

1. Ф. М. Достоевский: *Идиот*; *Преступление и наказание*; *Братья Карамазовы*; *Подросток*; *Униженные и оскорблённые*; *Бесы*; *Записки из мёртвого дома*.

2. И. А. Гончаров: *Обломов*; *Обыкновенная история*.

3. А. И. Герцен: *Кто виноват?*; *Былое и думы*.

4. Н. С. Лесков: *Захудалый род*; *Леди Макбет Мценского уезда*; *Очарованный странник*.

5. И. С. Тургенев: *Накануне*; *Новь*; *Дворянское гнездо*; *Отцы и дети*.

Практическая значимость работы состоит в том, что в ней на конкретных примерах демонстрируется применимость предложенного нового метода стилометрии к решению конкретных задач текстологии. Новый метод можно использовать наряду с традиционными, которые он удачно дополняет.

Обсуждение и результаты

1. Предмет и метод исследования

Нами разработаны компьютерные программы для анализа текстов на русском, английском и французском языках, отыскивающие в тексте количественные и порядковые числительные, выраженные как цифрами (числа), так и словесно в разных словоформах. Не учитывались собирательные числительные (*двое*, *трое*, …), мультипликативные числительные (*одиночный*, *двойной*, *тройной*, …), дистрибутивные числительные (*по одному*, *по двое*, *по трое*, …), неопределённо-количественные числительные (*мало*, *много*, *несколько*, …), которые невозможно квантифицировать. Что касается дробных числительных (*две пятых*, *семь десятых*, …), то мы отдельно учитывали числители и знаменатели, как если бы это были самостоятельные числительные.

Предварительно из текста автоматически удалялись идиоматические выражения и устойчивые фразы, случайно содержащие числительные (*семеро одного не ждут*, *два сапога пара* и т. п.); вручную удалялись числительные, не связанные с авторским художественным замыслом, – номера страниц, глав, перечисления 1), 2), 3), … и т. п. Впрочем, влияние удалённых элементов на результат в любом случае пренебрежимо мало в силу их редкости.

Казалось бы, метод учёта числительных наталкивается на непреодолимое препятствие в языках, в которых числительное *один* формально неотличимо от неопределённого артикля (*ein* в немецком языке, *un* – во французском и т. п.). Но набор числительных, встречающихся в тексте, – это, пожалуй, единственный признак, практически полностью сохраняющийся при переводе на другой язык. Это позволяет в случае необходимости (текст на языке, в котором имеет место подобное совпадение, или недоступность текста на языке-оригинале) анализировать авторский стиль, прибегая к помощи перевода на языке-посреднике.

В прежних работах (Zenkov, 2018; 2021; Zenkov, Místecký, 2019; 2022) нами уже было показано, что использование числительных в текстах специфично для каждого автора, зависит от художественного направления, жанра и стиля. В данной работе мы приведём новые примеры, подтверждающие эти выводы.

Анализ текстов выполнялся следующим образом. С помощью компьютерной программы из текстов извлекались числительные, и для каждого текста формировалась сводка обнаруженных числительных и их абсолютных частот. Поскольку тексты различаются по объёму, для сравнимости абсолютных частот в разных текстах объём одного из них выбирался в качестве эталонного, и исправленные абсолютные частоты получались умножением абсолютных частот на поправочный коэффициент.

Для выявления внутренней структуры в массиве исправленных абсолютных частот был применён иерархический кластерный анализ, объединяющий объекты в кластеры по принципу сходства. Как известно, мерой его является метрика ρ («расстояние»): чем меньше «расстояние» между объектами, тем больше сходство между ними.

В зависимости от характера данных в кластерном анализе применяются разные метрики, как, например, евклидова

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sum_{i}^{n} (x_i - y_i)^2 \quad (1)$$

и манхэттенская (расстояние «городских кварталов»)

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sum_{i}^{n} |x_i - y_i|, \quad (2),$$

где в нашем случае $\boldsymbol{x}$ и $\boldsymbol{y}$ – n-мерные векторы, компонентами которых являются исправленные абсолютные частоты первых n натуральных чисел в двух анализируемых текстах.

Каждое последующее числительное встречается в текстах, вообще говоря, со всё меньшей частотой (см. Раздел 3 и Приложение), поэтому наличие квадрата в формуле (1) означает, что «расстояние» между текстами фактически определяется различиями в частотах лишь числительного *один* – они вносят подавляющий вклад в сумму.

Мы применили манхэттенскую метрику (2), более равномерно учитывающую различия между текстами в частотах не только числительного *один*, но и 2, 3, …, n.

В процессе кластеризации использован метод средней связи (Average Linkage), являющийся золотой серединой между методами ближайшего (Single Linkage) и далёкого (Complete Linkage) соседа, которые, соответственно, преувеличивают и преуменьшают сходство между объектами (Moisl, 2015).

В квантитативной лингвистике принято считать, что даже при сопоставлении текстов двух авторов доказательную силу об их сходстве будет иметь лишь анализ, в котором к изучаемым текстам добавлены посторонние тексты других авторов (т. н. impostors) (Koppel, Winter, 2014). Мы учитывали это требование в нашем анализе.

2. Манера употребления числительных – авторская стилистическая особенность

Мы подтвердим этот тезис на примере англо-, франко- и русскоязычных текстов.

А. Англоязычные тексты

На Рис. 1 представлены результаты кластеризации данных о встречаемости числительных в текстах англоязычных авторов, перечисленных выше. Применены подходы, сформулированные в конце Раздела 1; в формуле (2) взято $n = 7$, т. к. во всех исследованных текстах встречались числительные от одного до семи.

Произведения распределились на дендрограмме в полном соответствии с авторством. Литературному стилю Ч. Диккенса свойственно наиболее единообразное употребление числительных (высота слияния невелика). Наибольшие различия – между произведениями Г. Уэллса. Интересно отметить, что два суперкластера (Диккенс – Теккерей и Набоков – Уэллс) в общем соответствуют хронологическому делению на литературу XIX и XX вв.

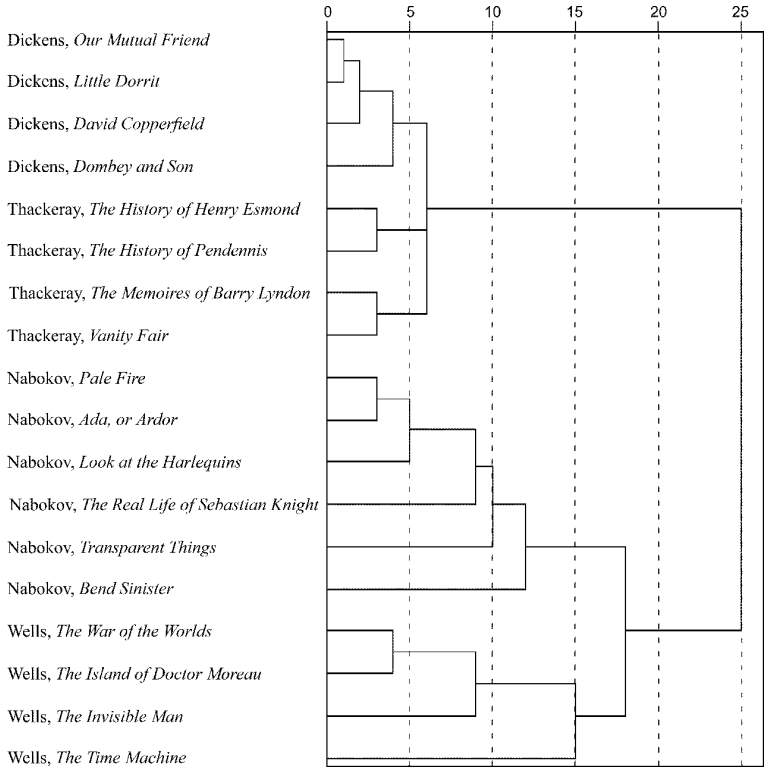

Рисунок 1. *Результат применения иерархического кластерного анализа к литературным текстам Ч. Диккенса, У. М. Теккерея, В. В. Набокова и Г. Дж. Уэллса. По горизонтальной оси отложено «расстояние» между текстами в произвольных единицах*

B. Франкоязычные тексты

На Рис. 2 представлена дендрограмма для франкоязычных текстов в рамках вышеприведённых подходов; в формуле (2) взято $n = 8$, т. к. во всех исследованных текстах встречались числительные от одного до восьми.

Снова можно констатировать распределение произведений по дендрограмме в соответствии с авторством. Единственные два автора, произведения которых не полностью локализованы на дендрограмме, – Э. Золя и Г. де Мопассан. В критической литературе неоднократно высказывались замечания о близости их стилей (Artinian, 1941; Dugan, 1973; Lloyd, 2020).

Проверим, могла ли удачная дендрограмма получиться случайно. Для этого добавим ещё авторов (impostors) – Ф. Мориака и Ж. Верна (Рис. 3). Дендрограмма почти не изменилась, что говорит о разумности нашей идеи о числительных в текстах как устойчивой особенности авторского стиля.

C. Русскоязычные тексты

На Рис. 4 представлена дендрограмма для произведений русскоязычных авторов.

Снова произведения распределились на дендрограмме в соответствии с авторством. Литературному стилю Ф. М. Достоевского свойственно очень единообразное употребление числительных (высота слияния невелика). Естественным исключением являются «Записки из мёртвого дома», которые носят полудокументальный характер.

Приведённые примеры показывают, что манера употребления числительных в текстах является авторским инвариантом и может использоваться в задачах стилометрии. Разумеется, кластерный анализ сам по себе не имеет доказательной силы, а представляет собой, скорее, средство визуализации данных. Но при необходимости можно продемонстрировать сходство/различие данных по числительным для разных авторов средствами математической статистики (см. Приложение), которая подтверждает полученные результаты.

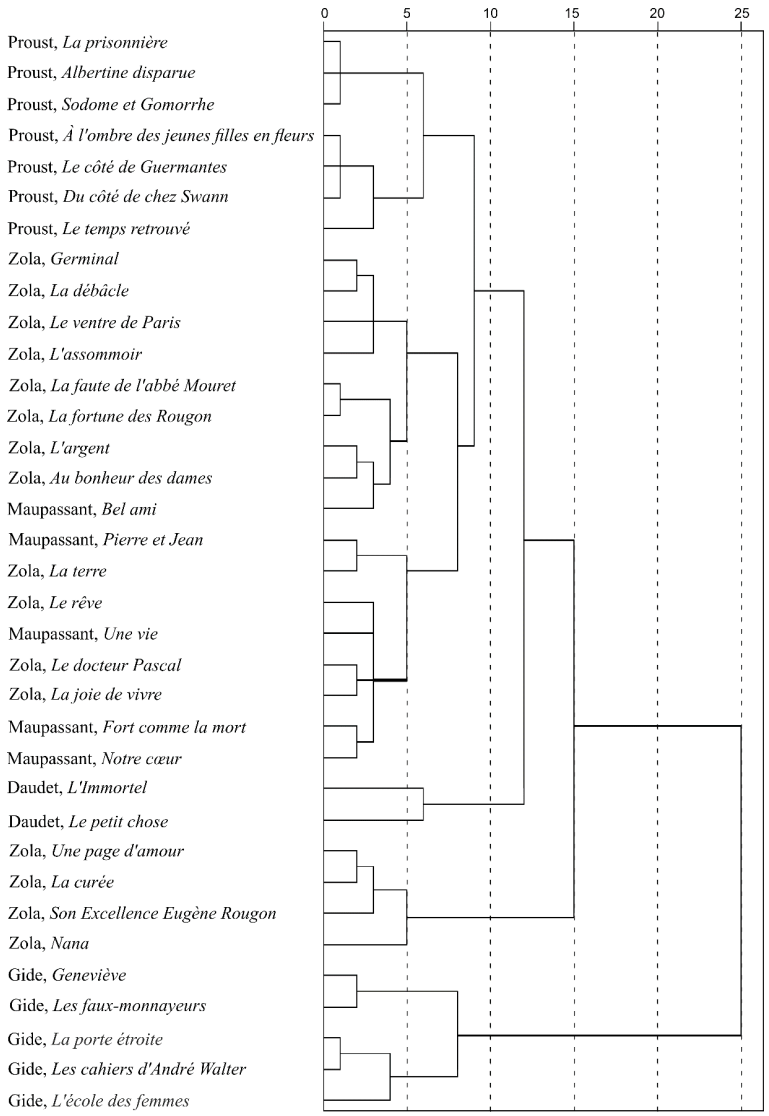

Рисунок 2. *Результат применения иерархического кластерного анализа к франкоязычным литературным текстам. По горизонтальной оси отложено «расстояние» между текстами в произвольных единицах*

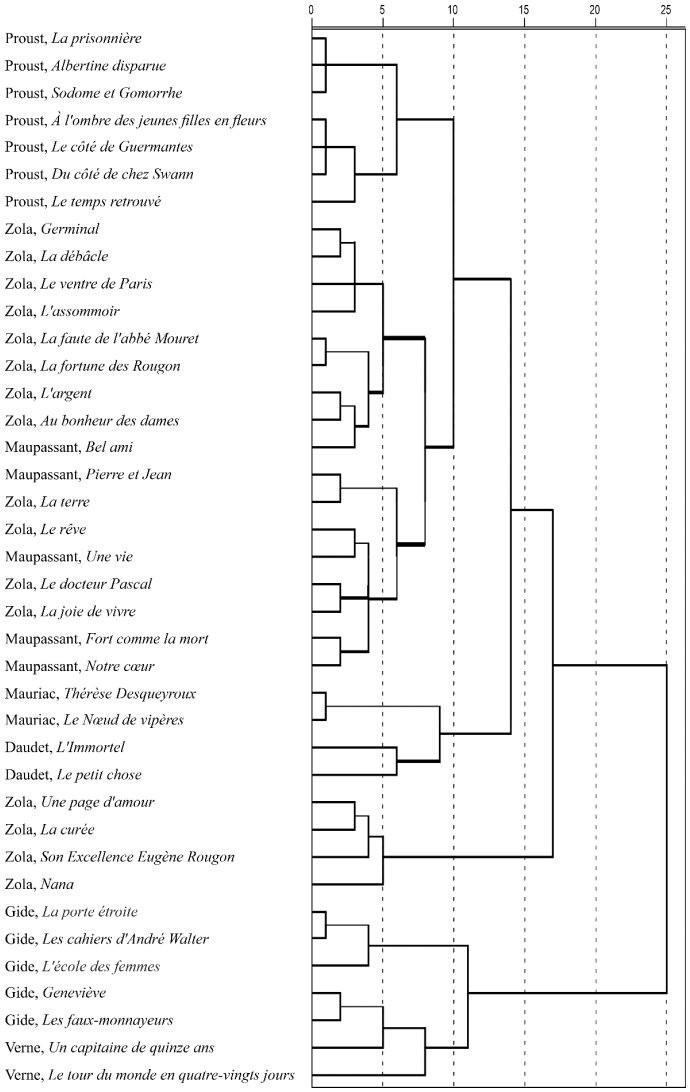

Рисунок 3. *Результат применения иерархического кластерного анализа к франкоязычным текстам, с добавлением текстов д р у г и х авторов (impostors) – Ф. Мориака и Ж. Верна*

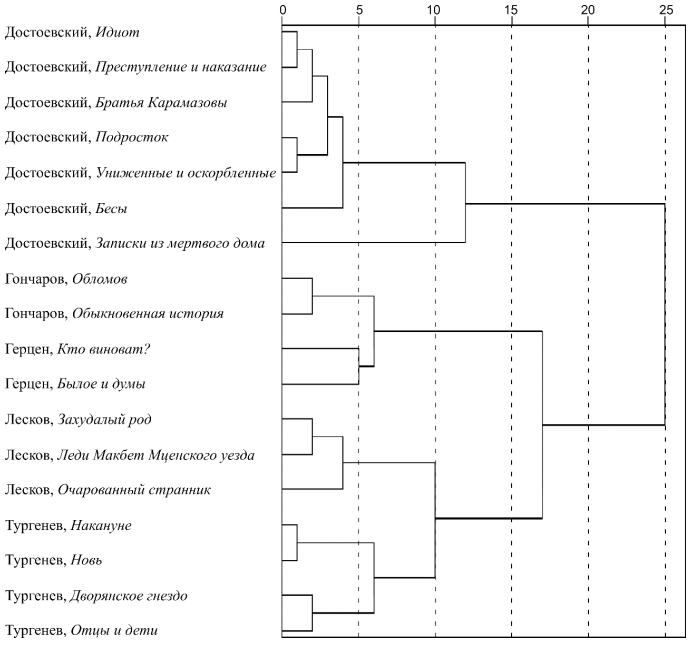

Рисунок 4. *Результаты иерархической кластеризации художественных произведений Ф. М. Достоевского, И. А. Гончарова, А. И. Герцена, Н. С. Лескова, И. С. Тургенева*

3. Авторство, скрытое псевдонимом, и манера употребления числительных

Приступим к главной задаче настоящей работы.

А. Литературное наследие Р. Гари

Французский писатель Ромен Гари (1914-1980) был склонен к литературным мистификациям. Кроме произведений, опубликованных под именем «Ромен Гари» (которое само является псевдонимом), он публиковался также под именами «Эмиль Ажар», «Фоско Синибальди» и «Шатан Бога». Наконец, его первый роман *Вино мертвецов* (*Le vin des morts*, опубл. 1937) вышел под его подлинным именем «Роман Кацев». Единственный писатель, дважды получивший Гонкуровскую премию (впервые как Гари и повторно как Ажар), Р. Гари, по его собственным словам, оставил в текстах произведений «Ажара» много намёков, которые позволяли установить истинного автора, но критика в большинстве своём оказалась слепа и намёков не распознала (Boisen, 1996; Poier-Bernhard, 1996; Hocus Bogus…, 2010).

Мы проверили, сильно ли различаются литературные стили Р. Гари и вымышленных авторов с точки зрения нашей методологии.

Для этого были проанализированы:

- произведения, выпущенные под именем «Ромен Гари»:

Европейское воспитание (*Education européenne*, 1945),
Тюльпан (*Tulipe*, 1946),
Большая барахолка (*Le grand vestiaire*, 1949),
Корни неба (*Les racines du ciel*, 1956, Гонкуровская премия),
Обещание на рассвете (*La promesse de l'aube*, 1960),
Слава нашим выдающимся пионерам (*Gloire à nos illustres pionniers / Les oiseaux vont mourir au Pérou*, 1962),
Леди Л. (*Lady L.*, 1963),
Пожиратели звёзд (*Les mangeurs d'étoiles*, 1966),
Пляска Чингиз-Хаима (*La danse de Gengis Cohn*, 1967),
Повинная голова (*La tête coupable*, 1968),
Прощай, Гари Купер! (*Adieu Gary Cooper*, 1969),
Белая собака (*Chien blanc*, 1970),
Европа (*Europa*, 1972),
Чародеи (*Les enchanteurs*, 1973),
Дальше ваш билет недействителен (*Au-delà de cette limite votre ticket n'est plus valable*, 1975),
Свет женщины (*Clair de femme*, 1977),
Спасите наши души (*Charge d'âme*, 1977),
Грустные клоуны (*Les clowns lyriques*, 1979),
Воздушные змеи (*Les cerfs-volants*, 1980),

- произведения, выпущенные под псевдонимом «Эмиль Ажар»:

Голубчик (*Gros calin*, 1974),
Вся жизнь впереди (*La vie devant soi*, 1975, повторная Гонкуровская премия),
Псевдо (*Pseudo*, 1976),
Страхи царя Соломона (*L'Angoisse du roi Salomon*, 1979),

- произведение, выпущенное под псевдонимом «Фоско Синибальди»:

Человек с голубем (*L'homme à la colombe*, 1958),

- произведение, выпущенное под псевдонимом «Шатан Бога»:

Головы Стефании (*Les têtes de Stéphanie*, 1974),

- произведение, выпущенное под настоящим именем «Роман Кацев»:

Вино мертвецов (*Le vin des morts*, 1937).

На Рис. 5 (левый график) представлена дендрограмма данных, касающихся встречаемости числительных (принципы построения описаны во введении; в формуле (2) взято $n = 8$).

Казалось бы, график не подтверждает нашу идею о специфичности использования числительных авторами. Но при добавлении в анализ двух произведений Ж. Верна (правая панель на Рис. 5) становится понятно, что всё дело в масштабе (по горизонтальной оси): высота слияния impostors и текстов Гари почти в два раза превышает высоту внутреннего слияния текстов Гари. Заметим, что максимальная высота всегда нормирована на 25.

Какая-либо хронологическая последовательность в распределении произведений по дендрограмме не просматривается, однако отметим, что в одном из двух суперкластеров явным особняком стоит ранний роман *Le vin des morts* (1937) – очевидно, что Р. Гари только ещё вырабатывал свой стиль в литературе.

Произведения, подписанные собственным именем Р. Гари, без какой-либо системы перемежаются произведениями, опубликованными под псевдонимами. Итак, в манере использования числительных при попытке Р. Гари поменять свой литературный стиль существенных изменений не произошло.

В. Литературные мистификации Бориса Акунина

Русскоязычный писатель, учёный-японист, литературовед, переводчик, либеральный общественный деятель Григорий Чхартишвили (род. 1956) публикует небеллетристические тексты под своим настоящим именем, но в художественной литературе с 1998 г. несравненно более известен под псевдонимом «Б. Акунин». С 2007 г. начали публиковаться произведения под псевдонимами «Анатолий Брусникин» (*Девятный спас*, *Герой иного времени*, *Беллона*) и «Анна Борисова» (*Там…*, *Креативщик*, *Vremena goda*). Впоследствии Г. Чхартишвили признал авторство и этих произведений.

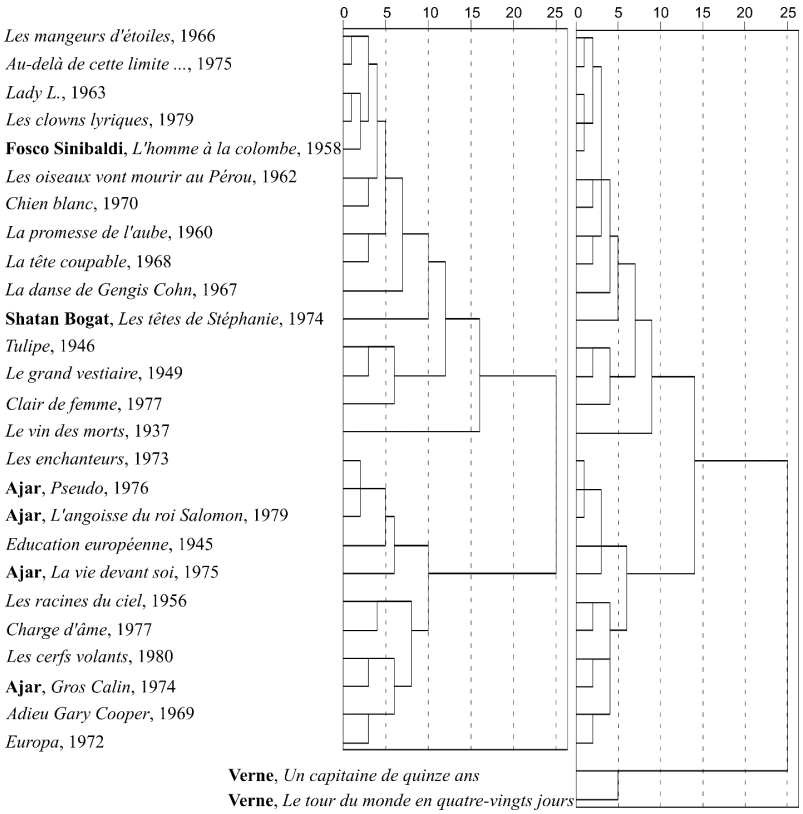

Рисунок 5. *Левый график: результаты иерархической кластеризации художественных произведений Р. Гари, опубликованных под собственным именем и под псевдонимами (в последнем случае имя явно указано). Правый график: то же, но с добавлением impostors: произведений Ж. Верна Пятнадцатилетний капитан (Un capitaine de quinze ans) и Вокруг света в 80 дней (Le tour du monde en quatre-vingts jours)*

Проверим, изменился ли его литературный стиль (в том, что касается числительных) при письме под псевдонимом.

На Рис. 6 представлены результаты кластеризации данных по использованию числительных в произведениях «Б. Акунина», «Анатолия Брусникина» и «Анны Борисовой» (принципы кластеризации описаны во введении; в формуле (2) взято $n = 10$).

Рисунок 6. *Результаты иерархической кластеризации художественных произведений Г. Чхартишвили, опубликованных под наиболее известным псевдонимом «Б. Акунин» и под менее известными именами «Анатолия Брусникина» и «Анны Борисовой»*

Закономерности в распределении имён по дендрограмме недостаточно отчётливы, чтобы с определённостью утверждать, что Г. Чхартишвили существенно по-разному использует числительные в произведениях, написанных под разными именами.

Итак, примеры Р. Гари и Г. Чхартишвили приводят нас к (предварительному) выводу о том, что манера использования числительных инвариантна для каждого писателя и изменить её практически невозможно. Конечно, этот вывод, для полной обоснованности, нуждается в подкреплении и другими примерами.

С. Проблема авторства произведения Х. Ли

В предыдущих примерах (Р. Гари и Г. Чхартишвили) наш метод учёта числительных служил целям констатации, а сейчас мы приступаем к проблеме, ещё не имеющей окончательного решения.

Харпер Ли (1926–2016) – американская писательница, автор прославленного романа *Убить пересмешника* (*To Kill a Mockingbird*, 1960), который является её единственным крупным литературным произведением. В 2015 году была издана книга Х. Ли *Пойди, поставь сторожа* (*Go Set a Watchman*), которая была написана ранее романа *Убить пересмешника*, но не была в своё время опубликована. По мнению критиков, это не отдельный роман, а лишь первоначальная версия романа *Убить пересмешника*, которую теперь попытались, исходя из коммерческого интереса, представить как самостоятельное произведение (Shields, 2016).

Х. Ли была в многолетних дружеских отношениях с Т. Капоте (1924–1984) вплоть до его смерти. Его многочисленные литературные и документальные произведения считаются литературной классикой. В свете сказанного понятно появление неоднократно высказывавшихся подозрений, что и роман *Убить пересмешника* написал тоже Капоте.

На Рис. 7 представлены результаты кластеризации данных по использованию числительных в двух романах Х. Ли, а также в основных произведениях Т. Капоте *Голоса травы* (*Луговая арфа*, *The Grass Harp*, 1951), *Рождественские воспоминания* (*A Christmas Memory*, 1956), *Завтрак у Тиффани* (*Breakfast at Tiffany's*, 1958), *Услышанные молитвы* (*Answered Prayers*, 1966-1980), *Летний круиз* (*Summer Crossing*, 1943), *Другие голоса, другие комнаты* (*Other Voices, Other Rooms*, 1948) (принципы кластеризации описаны во введении; в формуле (2) взято $n = 10$).

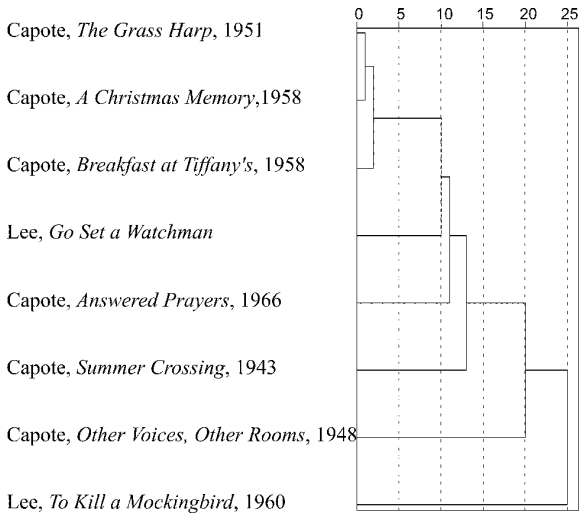

Рисунок 7. *Результаты иерархической кластеризации художественных произведений Х. Ли и Т. Капоте*

Дендрограмма показывает, что первоначальный вариант романа Х. Ли близок с точки зрения использования числительных романам Капоте, и, следовательно, он *мог* влиять на текст Х. Ли. В окончательном варианте – романе *Убить пересмешника* – влияние Т. Капоте, если таковое и было, менее существенно.

Отметим, что ранний вариант нашего стилометрического метода, основанный на учёте первых значащих цифр числительных в тексте (Zenkov, 2018), привёл к аналогичному выводу об авторстве романа Х. Ли.

Отметим также, что в данном анализе введение добавочных авторов (impostors) неуместно, т. к. круг возможных авторов изначально ограничен фигурами Ли и Капоте. В работе (Choiński, Eder, Rybicki, 2017/2018) авторы приходят к аналогичному выводу относительно влияния Т. Капоте на текст романа Х. Ли. Правда, они рассматривают ещё и тексты Терезы Хохофф (Therese von Hohoff Torrey, “Tay Hohoff”, 1898–1974), редактора, посвятившего много сил совершенствованию первоначальной рукописи Х. Ли. Но среди четырёх известных собственных произведений Хохофф два являются не художественными, а документальными (*A Ministry to Man: The Life of John Lovejoy Elliott, a Biography*; *The Author and his Audience: With a Chronology of Major Events in the Publishing History of J. B. Lippincott Company*), одно рассчитано на детскую аудиторию (*The Cat Who Wanted Out*) и одно мемуарное (*Cats and Other People*). Это очень специальные тексты, вряд ли пригодные для анализа на предмет использования числительных; к сожалению, эти книги были нам недоступны.

Рис. 8 представляет частотную зависимость числительных, встречающихся в вышеперечисленных произведениях Х. Ли и Т. Капоте. Абсолютные частоты пересчитаны с учётом разного размера текстов. Для удобства восприятия числительные ограничены диапазоном [1; 40].

Непосредственно из графика видно следующее:

1. Общими свойствами для всех текстов является уменьшение частоты с увеличением числительного (т. е. числа, которое им обозначено); наличие на круглых числах (10, 20, 30, …) локальных максимумов, высота которых тоже постепенно уменьшается; постепенное разрежение ряда чисел (появление пропусков по оси числительных). Эти выводы являются универсальными и справедливы для всех текстов, анализом которых мы занимались.
2. В текстах Х. Ли, по сравнению с текстами Т. Капоте, особенно велика частота числительного *один*, но уже числительное *два* (и последующие) имеет сравнительно меньшую частоту, т. е. Ли прибегает к числительным реже.
3. В текстах Т. Капоте более велико разнообразие числительных.
4. Частотная зависимость числительных в тексте *Пойди, поставь сторожа* ближе к таковым для текстов Т. Капоте, чем в тексте *Убить пересмешника*. Это согласуется с выводом, полученным выше из дендрограммы, о большем родстве раннего варианта романа Х. Ли с произведениями Т. Капоте.

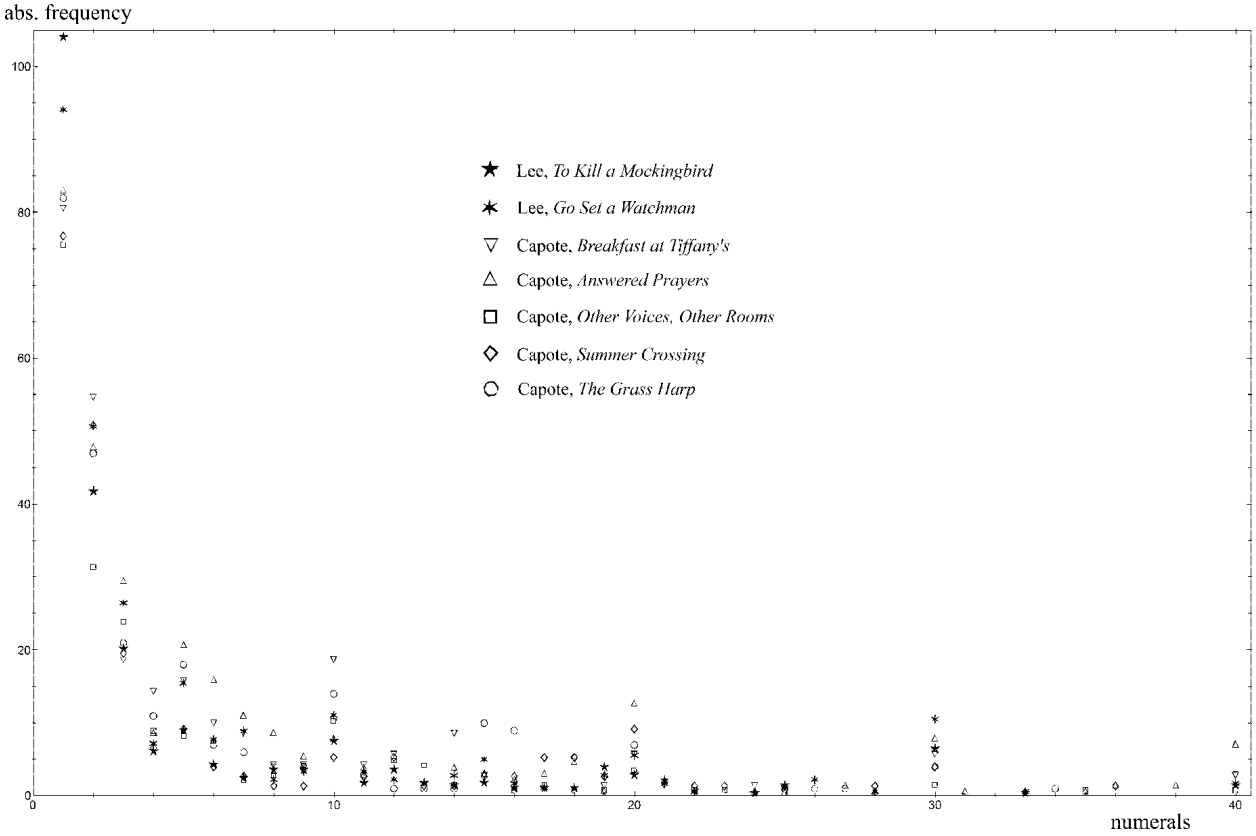

Рисунок 8. *Частотная зависимость числительных, встречающихся в художественных произведениях Х. Ли и Т. Капоте*

4. Приложение

Изложим здесь более подробно некоторые вычислительные аспекты нашей работы.

Согласно Рис. 7, окончательный вариант произведения Х. Ли *Убить пересмешника* с точки зрения использования числительных далёк от романа Т. Капоте *Завтрак у Тиффани*, а первоначальный вариант Х. Ли *Пойди, поставь сторожа*, наоборот, близок этому роману. Визуальные сходство/различие можно подкрепить статистическим критерием согласия Пирсона.

Сопоставление эмпирических распределений (в нашем случае – распределений абсолютных частот числительных в текстах тех или иных авторов) связано с проверкой соответствующих статистических гипотез о значимости/незначимости различий между распределениями.

Сформулируем гипотезы. Нулевая гипотеза H_0 предполагает, что проверяемые совокупности распределены одинаково. Альтернативная гипотеза H_1: распределения отличаются одно от другого.

Параметрический критерий согласия χ^2 Пирсона, помимо прочего, применяется для сопоставления эмпирических распределений одного и того же признака (для проверки однородности распределений). В нужном нам виде соответствующая процедура отсутствует в стандартных статистических пакетах, поэтому опишем её подробно.

Наши исходные статистические данные, касающиеся встречаемости числительных *один*, *два*, …, *десять* в трёх текстах, приведены в Табл. 1. Конечно, в текстах встречаются и бóльшие числительные, но со всё меньшей частотой.

Условием применимости критерия Пирсона является ограничение, чтобы абсолютная частота для каждой ячейки таблицы была не меньше 5, поэтому строки **8** и **9** придётся объединить (Табл. 2).

Таблица 1. *Эмпирические абсолютные частоты числительных в анализируемых текстах*

Числительное	Абсолютная частота числительных		
	Lee, *To Kill a Mockingbird*	Lee, *Go Set a Watchman*	Capote, *Breakfast at Tiffany's*
1	289	171	56
2	116	92	38
3	56	48	13
4	17	13	10
5	25	28	11
6	12	14	7
7	7	16	6
8	10	4	3
9	10	6	3
10	21	20	13

Таблица 2. *Эмпирические абсолютные частоты числительных после укрупнения строк*

Числительное	Абсолютная частота числительных		
	Lee, *To Kill a Mockingbird*	Lee, *Go Set a Watchman*	Capote, *Breakfast at Tiffany's*
1	289	171	56
2	116	92	38
3	56	48	13
4	17	13	10
5	25	28	11
6	12	14	7
7	7	16	6
8 и **9**	20	10	6
10	21	20	13

Каждый из текстов Х. Ли будем по отдельности сравнивать с текстом Капоте (Табл. 3).

Таблица 3. *Эмпирические абсолютные частоты числительных в текстах Убить пересмешника и Завтрак у Тиффани после укрупнения строк*

Числительное	Lee, *To Kill a Mockingbird*		Capote, *Breakfast at Tiffany's*		Сумма частот в строке
	Эмпирическая абсолютная частота	Метка ячейки	Эмпирическая абсолютная частота	Метка ячейки	
1	289	I	56	II	289 + 56 = 345
2	116	III	38	IV	116 + 38 = 154
3	56	V	13	VI	56 + 13 = 69
4	17	VII	10	VIII	17 + 10 = 27
5	25	IX	11	X	25 + 11 = 36
6	12	XI	7	XII	12 + 7 = 19
7	7	XIII	6	XIV	7 + 6 = 13
8 и **9**	20	XV	6	XVI	20 + 6 = 26
10	21	XVII	13	XVIII	21 + 13 = 34
	Σ = **563**		Σ = **160**		ΣΣ = **723**

Сопоставим эмпирическим частотам теоретические, получаемые с учётом того, что количество числительных (не превышающих десяти), найденных в текстах, различно: 563 в тексте Х. Ли и 160 – в тексте Т. Капоте. Итак, из полного количества 563 + 160 = 723 числительных в двух текстах на первый из них приходится доля 563 / 723 = 0,78, а на второй – 160 / 723 = 0,22. Во всех строках теоретические частоты, относящиеся к первому и второму текстам, должны составлять соответственно 0,78 и 0,22 от суммарной частоты по соответствующей строке. Если проверяемые эмпирические распределения не отличаются одно от другого, эмпирические частоты не должны существенно отклоняться от теоретических, получаемых из пропорции.

Перегруппируем данные Табл. 3, располагая относительные частоты для обоих текстов в порядке, указываемом метками, в одном столбце (это будут эмпирические частоты f_{emp}), а в другом столбце поместим теоретические частоты f_{theor}, вычисляемые согласно предыдущему как

$$f_{\text{theor}} = \frac{(\Sigma \text{ частот в строке})(\Sigma \text{ частот в столбце})}{\Sigma\Sigma}.$$

В этой формуле ΣΣ = 723 (Табл. 4).

Таблица 4. *Вычисления для применения критерия согласия Пирсона*

ячейка	эмпирическая частота f_{emp}	теоретическая частота f_{theor}	$\frac{(f_{emp} - f_{theor})^2}{f_{theor}}$
I	289	345 · 563 / 723 = 268,65	1,54
II	56	345 · 160 / 723 = 76,35	5,42
III	116	154 · 563 / 723 = 119,92	0,13
IV	38	154 · 160 / 723 = 34,08	0,45
V	56	69 · 563 / 723 = 53,73	0,10
VI	13	69 · 160 / 723 = 15,27	0,34
VII	17	27 · 563 / 723 = 21,02	0,77
VIII	10	27 · 160 / 723 = 5,98	2,70
IX	25	36 · 563 / 723 = 28,03	0,33
X	11	36 · 160 / 723 = 7,97	1,15
XI	12	19 · 563 / 723 = 14,80	0,53
XII	7	19 · 160 / 723 = 4,20	1,87
XIII	7	13 · 563 / 723 = 10,12	0,96
XIV	6	13 · 160 / 723 = 2,88	3,38
XV	20	26 · 563 / 723 = 20,25	0,00
XVI	6	26 · 160 / 723 = 5,75	0,01
XVII	21	34 · 563 / 723 = 26,48	1,13
XVIII	13	34 · 160 / 723 = 7,52	3,99
	Σ = 723	**Σ = 723**	**Σ = 24,81**

При сопоставлении эмпирических распределений по критерию Пирсона учитывается количество степеней свободы $df = (r - 1)(c - 1)$, где r – количество строк в таблице эмпирических частот (после объединения ячеек получилось $r = 9$ – см. Табл. 2), c – количество сопоставляемых распределений ($c = 2$). Итак, количество степеней свободы $df = 8$.

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

$$\chi^2_{cr} = \begin{cases} 15{,}5 & (\alpha = 0{,}05), \\ 20{,}1 & (\alpha = 0{,}01). \end{cases} \quad (3).$$

Поскольку эмпирическое $\chi^2_{emp} = 24{,}81$ превышает каждое из этих критических, гипотеза H_0 отвергается; иными словами, распределения числительных в романах *Убить пересмешника* Х. Ли и *Завтрак у Тиффани* Т. Капоте различаются з н а ч и м о.

Теперь сравним первичный извод *Пойди, поставь сторожа* романа Х. Ли с тем же романом Т. Капоте (Табл. 5).

Таблица 5. *Эмпирические абсолютные частоты числительных в текстах Пойди, поставь сторожа и Завтрак у Тиффани после укрупнения строк*

Числительное	Lee, *Go Set a Watchman*		Capote, *Breakfast at Tiffany's*		Сумма частот в строке
	Эмпирическая абсолютная частота	Метка ячейки	Эмпирическая абсолютная частота	Метка ячейки	
1	171	I	56	II	171 + 56 = 227
2	92	III	38	IV	92 + 38 = 130
3	48	V	13	VI	48 + 13 = 61
4	13	VII	10	VIII	13 + 10 = 23
5	28	IX	11	X	28 + 11 = 39
6	14	XI	7	XII	14 + 7 = 21
7	16	XIII	6	XIV	16 + 6 = 22
8 и **9**	10	XV	6	XVI	10 + 6 = 16
10	20	XVII	13	XVIII	20 + 13 = 33
	Σ = **412**		Σ = **160**		ΣΣ = **572**

Из полного количества 412 + 160 = 572 числительных в двух текстах на первый приходится доля 412 / 572 = 0,72, а на второй – 160 / 572 = 0,28. Во всех строках теоретические частоты, относящиеся к первому и второму текстам, должны составлять соответственно 0,72 и 0,28 от суммарной частоты по соответствующей строке.

Выполняя расчёты, аналогичные проделанным выше, получим Табл. 6.

Критические значения распределения χ^2 остаются теми же (3). В данном случае эмпирическое $\chi^2_{emp} = 8{,}59$ существенно м е н ь ш е критических значений при обоих уровнях значимости, поэтому нет оснований отвергнуть гипотезу H_0: закономерности использования числительных в текстах *Пойди, поставь сторожа* Х. Ли и *Завтрак у Тиффани* Т. Капоте неразличимы (при данных уровнях значимости).

Таблица 6. *Вычисления для применения критерия согласия Пирсона*

ячейка	эмпирическая частота f_{emp}	теоретическая частота f_{theor}	$\frac{(f_{emp} - f_{theor})^2}{f_{theor}}$
I	171	227 · 412 / 572 = 163,50	0,34
II	56	227 · 160 / 572 = 63,50	0,89
III	92	130 · 412 / 572 = 93,64	0,03
IV	38	130 · 160 / 572 = 36,36	0,07
V	48	61 · 412 / 572 = 43,94	0,38
VI	13	61 · 160 / 572 = 17,06	0,97
VII	13	23 · 412 / 572 = 16,57	0,77
VIII	10	23 · 160 / 572 = 6,43	1,98
IX	28	39 · 412 / 572 = 28,09	0,00
X	11	39 · 160 / 572 = 10,91	0,00
XI	14	21 · 412 / 572 = 15,13	0,08
XII	7	21 · 160 / 572 = 5,87	0,22
XIII	16	22 · 412 / 572 = 15,85	0,00
XIV	6	22 · 160 / 572 = 6,15	0,00
XV	10	16 · 412 / 572 = 11,52	0,20
XVI	6	16 · 160 / 572 = 4,48	0,52
XVII	20	33 · 412 / 572 = 23,77	0,60
XVIII	13	33 · 160 / 572 = 9,23	1,54
	Σ = 572	**Σ = 572**	**Σ = 8,59**

Подобные (громоздкие!) расчёты нами проделаны для всех приведённых выше дендрограмм, и подтвердилось, что визуальным сходству/различиям использования числительных авторами можно доверять.

Заключение

Исследована встречаемость числительных в оригинальных литературных текстах, написанных на английском (18 произведений четырёх авторов), французском (39 произведений семи авторов), русском (18 произведений пяти авторов) языках. В результате были получены следующие выводы:

1. Обнаружилось, что манера использования числительных у разных авторов может существенно различаться, и для каждого автора она устойчиво воспроизводится в разных текстах. Это может объясняться психологией автора, независимо от его осознанного намерения влияющей на творческий результат. Авторские различия в употреблении числительных не только наблюдаются визуально посредством кластерного анализа на дендрограммах, но и подтверждаются критерием согласия Пирсона.
2. В ряду числительных *один*, *два*, *три*, … каждый следующий элемент встречается в текстах, вообще говоря, всё реже (с естественными всплесками на круглых числах *десять*, *двадцать*, …, *сто*, …). Причины уменьшения частоты не вполне нам понятны; некоторое объяснение даётся экспериментально обнаруженным явлением «Бенфордовского предпочтения» (Benford bias) (Burns, 2020) в человеческой психологии: когда испытуемым было предложено придумывать числа, частотное распределение первых значащих цифр напоминало распределение Бенфорда (Benford, 1938): меньшие цифры встречались чаще.
3. Исследование особенностей использования числительных в авторских литературных текстах, опубликованных под разными псевдонимами, показало, что попытки авторов изменить свой творческий стиль, сочинять «по-другому» практически не сказываются на встречаемости числительных в текстах.
4. Таким образом, манера употребления числительных является авторским инвариантом, и это можно использовать при решении задач об авторстве текстов. Вывод, сделанный нами относительно атрибуции двух романов американской писательницы Х. Ли, согласуется с выводами, полученными другими исследователями в рамках иных методик.

Итак, предложенный нами метод, основанный на анализе встречаемости числительных в текстах, является ещё одним действенным методом стилометрии, который, конечно, не отменяет существующие, а дополняет их.

Перспективы дальнейшего исследования состоят в расширении доказательной базы подхода к стилометрическим задачам, основанного на статистике использования числительных в текстах, а также в решении конкретных задач, связанных с количественным изучением авторского стиля и атрибуцией текстов.

Источники | References

1. Artinian A. Maupassant Criticism in France, 1880-1940, with an Inquiry into His Present Fame and a Bibliography. N. Y.: Kings Crown Press, 1941.
2. Benford F. The Law of Anomalous Numbers // Proceedings of the American Philosophical Society. 1938. Vol. 78. No. 4.

3. Boisen J. Un Picaro métaphysique: Romain Gary et l'art du roman. Odense: Odense University Press, 1996.
4. Brocardo M. L., Traore I., Woungang I., Obaidat M. S. Authorship Verification Using Deep Belief Network Systems // International Journal of Communication Systems. 2017. Vol. 30. Iss. 12. https://doi.org/10.1002/dac.3259
5. Burns B. D. Do People Fit to Benford's Law, or Do They Have a Benford Bias? 2020. https://cognitivesciencesociety.org/cogsci20/papers/0379/index.html
6. Choiński M., Eder M., Rybicki J. Harper Lee and Other People: A Stylometric Diagnosis // Mississippi Quarterly. 2017/2018. Vol. 70/71. No. 3.
7. Dugan J. R. Illusion and Reality, A Study of Descriptive Techniques in the Works of Guy de Maupassant. Berlin – Boston: Mouton, 1973.
8. Hocus Bogus. Romain Gary Writing as Émile Ajar / transl. by D. Bellos. New Haven – L.: Yale University Press, 2010.
9. Hungerbühler N. Benfords Gesetz über führende Ziffern: wie die Mathematik Steuersündern das Fürchten lehrt. 2007. https://ethz.ch/content/dam/ethz/special-interest/dual/educeth-dam/documents/Unterrichtsmaterialien/mathematik/Benfords%20Gesetz%20über%20führende%20Ziffern%20(Artikel)/benford.pdf
10. Koppel M., Winter Y. Determining if Two Documents Are Written by the Same Author // Journal of the Association for Information Science and Technology. 2014. Vol. 65. No. 1.
11. Lloyd C. Guy de Maupassant. L.: Reaktion Books, 2020.
12. Moisl H. Cluster Analysis for Corpus Linguistics. Berlin – München – Boston: De Gruyter Mouton, 2015.
13. Poier-Bernhard A. Romain Gary – das brennende Ich: literaturtheoretische Implikationen eines Pseudonymenspiels. Tübingen: Niemeyer, 1996.
14. Shields C. J. Mockingbird: A Portrait of Harper Lee: From Scout to Go Set a Watchman. 2nd ed. N. Y.: Henry Holt and Co., 2016.
15. Stamatatos E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. 2009. Vol. 60. No. 3.
16. Tempestt N., Kalaivani S., Aneez F., Yiming Y., Yingfei X., Damon W. Surveying Stylometry Techniques and Applications // ACM Computing Surveys. 2017. Vol. 50. No. 6.
17. Zenkov A. V. A Method of Text Attribution Based on the Statistics of Numerals // Journal of Quantitative Linguistics. 2018. Vol. 25. No. 3.
18. Zenkov A. V. Stylometry and Numerals Usage: Benford's Law and Beyond // Stats. 2021. Vol. 4.
19. Zenkov A., Místecký M. Young Vladimír Vašek? – A Numerals Analysis Contribution to the Bezruč-Hrzánský Identity Issue // Naše řeč. 2022. Vol. 105. No. 3.
20. Zenkov A. V., Místecký M. The Romantic Clash: Influence of Karel Sabina over Macha's Cikani from the Perspective of the Numerals Usage Statistics // Glottometrics. 2019. Vol. 46.

Финансирование | Funding

RU Исследование выполнено за счет средств гранта Российского научного фонда № 23-28-00750, https://rscf.ru/project/23-28-00750/, проект «Разработка нового метода стилометрии на основе статистики использования числительных в авторских текстах».

EN The reported study was funded by the Russian Science Foundation, grant No. 23-28-00750, https://rscf.ru/project/23-28-00750/, the project "Development of a new method of stylometry based on statistics of the use of numerals in authorial texts".

Информация об авторах | Author information

RU **Зенков Андрей Вячеславович**[1], к. физ.-мат. н., доц.
[1] Уральский федеральный университет, г. Екатеринбург

EN **Zenkov Andrei Viacheslavovich**[1], PhD
[1] Ural Federal University, Ekaterinburg

[1] *zenkow@mail.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 21.09.2023; опубликовано online (published online): 31.10.2023.

Ключевые слова (keywords): стилометрия; стилеметрия; квантитативная лингвистика; атрибуция текстов; числительные в тексте; stylometry; quantitative linguistics; text attribution; numerals in the text.