Ю. Н. Орлов, К. П. Осминин

Определение жанра и автора литературного произведения статистическими методами

В настоящей статье излагается метод классификации текстов на основе анализа статистических закономерностей буквенных распределений, т. е. вероятностей встречаемости букв и буквосочетаний. Подробно рассматривается задача кластеризации литературных произведений по определенным жанрам, а также вопрос определения авторства произведения. При этом решение должно быть найдено без вторжения в область литературы, т. е. без анализа синтаксиса, литературных приемов и схем взаимодействий персонажей.

Введение

По-видимому, впервые статистический анализ был применен к вопросу авторства литературного произведения почти сто лет назад А. А. Марковым [1]. Он предположил, что текст представляет собой случайную цепочку из гласных и согласных букв, связанных между собой определенными вероятностями перехода. Тогда авторство может быть установлено путем сравнения соответствующих вероятностей, которые предполагаются постоянными для каждого автора. Ограниченность метода состоит в том, что эти вероятности существенно зависят от объема текста, по которому они рассчитываются, и эволюционируют на протяжении всего произведения, так что погрешность метода оказывается слишком велика.

Тем не менее, хотя литературное произведение не является реализацией Марковского процесса, существуют разные модификации этого метода, поскольку он легко реализуем на практике. Интересным примером развития рассматриваемой методики стала работа Д. В. Хмелева [2], где уточняющим инструментом служит функция максимального правдоподобия, в качестве которой выбрана информационная энтропия для парных буквосочетаний.

В большинстве существующих методик предполагается некоторая инвариантность авторской манеры письма, что приводит к поиску различных «авторских инвариантов». Это может быть доля гласных или согласных, информационная энтропия, распределение используемых слов по длине, переходные вероятности между парами букв, доля союзных слов (см., например, А. Т. Фоменко [3]) и иные функционалы от распределения текста по буквам и буквосочетаниям. К сожалению, к любому методу, основанному на статистике (в том числе и к применяемому в настоящей работе), можно подобрать контрпример. В частности, таковым является роман «Улисс» Джеймса Джойса, каждая из восемнадцати глав которого написана в разном стиле. Авторами проверено, что ни один из известных методов не дает удовлетворительного ответа (в том смысле, что все главы написаны одним и тем же человеком) более чем по трем парам глав из 153 пар. Разумеется, это не является «противозаконным»: автор имеет право изменить стиль, начинать все слова с буквы «о» и т. д. Кроме того, при увеличении числа сравниваемых произведений возникает неизбежное сближение инвариантов, так что, начиная с какого-то количества авторов, расстояние между инвариантами становится меньше, чем среднеквадратичное отклонение инварианта, которым обычно пренебрегают. Поэтому такая методика имеет принципиальные ограничения.

Отметим также еще один недостаток существующих в данной области работ на-

правлении — это отсутствие собственно статистического анализа для установления уровня достоверности в эмпирическом определении тех или иных закономерностей. Как правило, авторы ограничиваются тем, что говорят о «достаточно больших объемах» текстов, но никаких критериев при этом не предъявляют. Понимая, что расчет точности определения эмпирических вероятностей или проведение проверки определенной статистической гипотезы представляет самостоятельное исследование, по объему сопоставимое с самой статьей, считаем, что желательно иметь хотя бы выводы такого анализа. Например, в [3] приводится такое рассуждение: «При величине выборок в 16000 слов процентное содержание служебных слов для каждого автора из нашего списка оказалось приблизительно постоянным вдоль всех его произведений». Какова величина этого отклонения? Какое отклонение следует ожидать, считая, что данная случайная величина имеет стационарное распределение? Что дают в этом случае критерии близости распределений? Какой объем текста необходим, чтобы отклонение с заданным уровнем достоверности лежало бы в определенном интервале? И, наконец, чему равна площадь перекрывающихся частей под графиками плотностей распределения двух авторских инвариантов, и не превосходит ли вероятность ошибки в первоначальном выборе критерия для инварианта вероятности ошибки в определении авторов? Все эти вопросы остаются без ответа.

Для задачи идентификации автора текста предлагаем искать не авторский инвариант, а изучать близость между распределениями букв или буквосочетаний в подходящей норме. Во-первых, распределение букв — это не одно число, а совокупность, например, 32 чисел, поэтому отличие для разных авторов может быть более четким. Во-вторых, близость распределений можно трактовать как схожесть письма, хотя верно как раз обратное утверждение, но это ситуация, типичная для статистики: например, из близости корреляции к нулю часто делается вывод о независимости сравниваемых величин, что, вообще говоря, неверно.

Статистический анализ частот встречаемости букв в различных текстах проводился весьма интенсивно в середине прошлого века (см. [4–7]) в связи с вопросами кодирования и передачи информации. Однако отсутствие текстов в «электронной» форме не позволяло провести детальный анализ, поскольку тексты обрабатывались вручную. Цель такого анализа состояла в оценке вероятности p_i появления отдельных букв или их сочетаний и определении информационной энтропии текста. Подробный обзор результатов того времени содержится в [8].

Еще одним принципиальным моментом, ограничивающим точность статистических методов, является то, что последовательности букв в произведениях одного автора на практике не образуют стационарный ряд. Распределение меняется на протяжении всего текста, а также при переходе от одного произведения к другому. Поэтому достоверно можно сравнивать тексты с близким уровнем стационарности, для чего введем ниже соответствующие определения.

Изучая буквенные распределения, можно сделать ряд интересных наблюдений о творчестве писателей, выполняющихся, как уже говорилось, «в среднем». Для этого надо ввести удобную норму, определить желаемую точность идентификации и убедиться, что расстояние между текстами, начиная с определенного объема, не зависит от выбора начала отрывка. После этого сравниваемые тексты группируются по близости их попарных расстояний. Этот план действий и реализован в данной работе.

1. Выборочные функции распределения и их плотности

Основным объектом исследования в работе являются выборочные функции распределения текстов по буквам. Дадим соответствующие определения.

Плотностью функции распределения (ПФР) дискретной случайной величины ξ,

принимающей значения из множества $x_1, x_2, \ldots, x_n$, называется дискретная функция $f(i)$, представляющая вероятность того, что $\xi = x_i, \quad i = 1, 2, \ldots, n$.

Рассмотрим последовательность из N значений величины ξ. Элементы этой последовательности обозначим $b_j, \ j = 1, 2, \ldots, N$. Пусть среди них значение x_i встретилось k_i раз. Тогда выборочной функцией распределения (ВПФР) по заданной выборке объема N называется совокупность $f_N(i)$ величин

$$f_N(i) = \frac{k_i}{N}, \quad i = 1, 2, \ldots, n. \qquad (1)$$

Предположим, что существует некоторый максимальный объем выборки $N_{\max}$ (в нашем случае это число знаков в конкретном литературном произведении). Этот объем генерирует распределение $f_{\max}(i)$ по формуле (1), которое будем называть ПФР данного текста. Рассмотрим выборку объема $N \le N_{\max}$, начинающуюся с любого номера $j \le N_{\max} - N + 1$. Соответствующую ВПФР обозначим $f_N(i; j)$. Номер j играет роль «начального момента времени» в упорядоченной выборке. Очевидно, что с увеличением N наблюдается поточечная сходимость величин $f_N(i; j)$ к $f_{\max}(i)$. Для нашего исследования важно, как быстро осуществляется эта сходимость для всех величин (1), и насколько она равномерна по «времени».

Введем понятие длины ε-стационарности как такого минимального объема выборки $L(\varepsilon)$, что для всех моментов j и всех таких выборок, что $N \ge L(\varepsilon)$, выполняется условие

$$\sum_{i=1}^{n} \left| f_N(i; j) - f_{\max}(i) \right| \le \varepsilon. \qquad (2)$$

Можно показать [9], что при $N \ge N_\varepsilon$, где

$$N_\varepsilon = \left[\left(1 - \frac{\varepsilon}{2} \right) N_{\max} \right] \qquad (3)$$

и квадратные скобки обозначают целую часть числа, условие ε-стационарности (2) заведомо выполнено. Интерес представляют ситуации, когда $L(\varepsilon)$ существенно меньше, чем N_ε при таких значениях ε, которые отвечают точности, достаточной для практических нужд, например, $\varepsilon = 0{,}05$. В таком случае с достоверностью $1 - \varepsilon$ можно считать, что на длине $L(\varepsilon)$ распределение становится неотличимым от распределения всей выборки, которую можно приближенно трактовать как генеральную совокупность. Соответствующий временной ряд b_j будем называть квазистационарным, если его ВПФР ε-стационарна.

Формула (2) использует понятие расстояния в пространстве суммируемых функций. Определим в соответствии с ним расстояние между двумя ВПФР, построенными по выборкам объемов N_1 и N_2:

$$\rho_{12} = \left\| f_{N_1} - f_{N_2} \right\| = \sum_{i=1}^{n} \left| f_{N_1}(i) - f_{N_2}(i) \right|. \qquad (4)$$

Это расстояние, будучи индикатором близости распределений, применятся далее для задачи кластеризации литературных текстов по авторам и жанрам.

Если N_1 и N_2 — полные объемы двух данных текстов, то расстояние (4) корректно отражает различие между текстами только в том случае, если больший объем имеет длину стабилизации $L(\lambda)$ на некотором уровне λ, существенно меньшем, чем само ρ_{12}, причем $L(\lambda)$ не превосходит объема меньшего из текстов. Уточним понятие «существенно меньше». Пусть имеется K текстов одного автора, и L_0 есть длина минимального из них. Для k-го текста этой длине отвечает определенный уровень квазистационарности ε_k:

$$\varepsilon_k = \max_j \sum_{i=1}^{n} \left| f_{L_0}^{(k)}(i; j) - f_{\max}^{(k)}(i) \right|. \qquad (5)$$

Если положить $\lambda = \max_k \varepsilon_k$, то каждый текст на длине L_0 будет λ-стационарным. Рассмотрим K $(K-1)/2$ попарных расстояний ρ_{ij} между текстами. Пусть распределение этих расстояний имеет среднее $\bar{\rho}$ и дисперсию σ^2. Зададим точность δ, с которой предполагается различать тексты. Если $1 - \delta$ — квантиль эмпирического распре-

деления попарных расстояний больше, чем расстояние между данным неизвестным текстом и любым из по крайней мере $[(1-\delta)K]$ базовых текстов, то этот текст с доверительной вероятностью $1-\delta$ будем считать принадлежащим перу того же автора. Этот вывод корректен, если только $\lambda < \delta$. Кроме того, если оказалось, что $\lambda > \sigma / \bar{\rho}$, то сам автор пишет настолько разнообразно, что его нельзя точно идентифицировать. Такие авторы и представляют собой контрпримеры. Следовательно, точность метода определяется долей плохо идентифицируемых авторов в выборке текстов, и потому характеризует не столько метод, сколько саму выборку.

Таким образом, разнообразие существующих статистических методов необходимо, поскольку каждый из них имеет разную мощность по отношению к трудно определяемым авторам. Возможно, что для идентификации имеет смысл применять несколько методов.

2. Квазистационарность однобуквенных ВПФР

Рассмотрим литературное произведение на примитивном с точки зрения читателя уровне — просто как последовательность букв, игнорируя его смысловую составляющую и не обращая внимания на те или иные художественные приемы. Пробелы, знаки препинания и прочие небуквенные символы не учитываем, поскольку они отчасти характеризуют «авторский стиль», требующий не только статистического, но и литературного анализа. Буквы е и ё для удобства не различаем, поскольку во многих печатных текстах обозначение «ё» не используется. Осью времени считаем нумерацию страниц в книге в направлении возрастания номеров, внутри страницы — по строкам сверху вниз, по строке — слева направо. Погрешности, вносимые возможными опечатками, считаем пренебрежимо малыми. Каждой букве ставим в соответствие ее порядковый номер в алфавите. Таким образом, текст в данной работе рассматривается как упорядоченная во «времени» цепочка чисел от 1 до 32. Кроме того, можно нумеровать не только отдельные буквы, но и буквосочетания — пары, тройки и т. д. При таком анализе текстов возникают следующие задачи и вопросы.

Задача 1. Определить минимальный объем выборки, в данном случае длину $L(\varepsilon)$ цепочки символов, идущих подряд, но начинающихся с любого места произведения, для которого такую длину можно определить в направлении возрастания времени, ВПФР которого ε-стационарна в рамках данного произведения. Сравнить $L(\varepsilon)$ для разных произведений одного и того же автора, когда объем текста допускает такое сравнение. Насколько велик разброс этой величины в зависимости от ε по совокупности произведений одного автора? Зависит ли $L(\varepsilon)$ от жанра произведения? Может ли средняя по произведениям функция $L(\varepsilon)$ служить опознавательным знаком отдельного писателя?

Задача 2. Следует выяснить, можно ли по виду ПФР определить, к какому жанру относится данное произведение — триллер, ужасы, любовный роман, детектив, комедия, технический текст и т. п. Разумеется, мы далеки от мысли, что автор при написании текста сознательно стремится к созданию некоторой ПФР. Но возможно, характерная ПФР возникает непроизвольно в силу самой тематической направленности текста, и тогда она представляет собой некоторый инструмент измерения или сравнения в такой тонкой области, как литературное творчество.

Задача 3. Если окажется, что произведения, написанные в одном жанре, кластеризуются указанным выше образом — по ПФР или по длинам ε-стационарности, то интересно сравнить между собой различия в ПФР для произведений разных авторов и жанров. Кроме того, можно проверить, отличаются ли по этому показателю «признанные шедевры» и, как говорится, «обычная литература».

Поскольку последовательность букв в тексте образует нестационарный временной ряд, необходимо понять, какой смысл имеет ВПФР. Ведь эмпирическая вероятность есть предел отношения (1) при $N \to \infty$, если таковой существует, поэтому значений k_i для каждого i должно быть достаточно много. Тогда ВПФР представляет собой набор вероятностей использования букв в тексте, объем которого должен быть достаточно большим, чтобы эти вероятности определялись с заданным уровнем точности в предположении стационарности выборки. Ошибка δ в определении вероятностей отличается от уровня квазистационарности ε, более того, она должна быть существенно меньше, иначе само понятие длины стационарности не будет иметь практического смысла. Оценим соответствующий минимальный объем текста.

Как известно (см., например, [10]) отклонение выборочного среднего значения $\bar{x}(N)$ стационарной случайной величины, определяемое по выборке объема N, от генерального среднего μ распределено асимптотически нормально с нулевым средним и стремящейся к нулю дисперсией σ^2 / N, где σ^2 есть дисперсия этой величины по гипотетической генеральной совокупности $f(i)$. Рассмотрим в качестве такой случайной величины количество n_i буквы «i» в тексте объема N. Тогда среднее значение этого количества n_i / N даст выборочную эмпирическую вероятность использования данной буквы. Значение σ_i представляет собой среднеквадратичное отклонение этой вероятности, а $\sigma_i / \sqrt{N}$ — отклонение среднего значения этой вероятности от значения по генеральной совокупности. Однако в условиях, когда генеральная дисперсия не известна, а оценивается только по выборочной дисперсии $s^2(N)$, следует рассматривать статистику

$$t = \sqrt{N-1}\frac{\bar{x}(N)-\mu}{s(N)}. \qquad (6)$$

Предположим, что выборочные отклонения частот использования букв с увеличением объемов выборки асимптотически нормальны. Тогда для каждой из n букв статистика (6) имеет распределение Стьюдента с $N - 1$ степенями свободы. Пренебрегая отличием N от $N - 1$, с доверительной вероятностью α получаем, что $|f_N(i) - f(i)|$ не превосходит $t_\alpha s / \sqrt{N}$, где t_α оценим сверху как α-квантиль предельного распределения Стьюдента с бесконечным числом степеней свободы. В частности, для α = 0,95; 0,97; 0,99 соответствующие значения t_α приближенно равны 1,96; 2,20; 2,58 [10]. В качестве оценки выборочной дисперсии также возьмем максимальную по 32 буквам: $s = \max s_i$. Тогда из (6) получаем следующую оценку для минимального объема текста:

$$\sum_{i=1}^{n} |f_N(i) - f_{\max}(i)| \le \frac{t_\alpha}{\sqrt{N}} \sum_{i=1}^{n} s_i \le \frac{t_\alpha n s}{\sqrt{N}}. \qquad (7)$$

Зададим число λ как величину интегральной близости $f_N(i)$ к некоторой гипотетической $f(i)$: $\sum_{i=1}^{n} |f_N(i) - f(i)| \le \lambda$. Тогда из (7) получаем, что если объем текста превосходит величину $N_{\min}$, приближенно являющуюся решением уравнения

$$N = \left(\frac{t_\alpha n s(N)}{\lambda} \right)^2, \qquad (8)$$

то с вероятностью α его распределение на этом объеме близко к стационарному с точностью λ. Эмпирическая зависимость $s(N)$ была проанализирована для 100 произведений различных авторов и жанров (см. далее п. 3). Полагая n = 32 и λ = 0,01, получаем в результате численного решения уравнения (8), что для вышеуказанных значений α величины $N_{\min}$ соответственно равны примерно 8 тыс., 10 тыс. и 15 тыс. знаков. Для корректного сравнения текстов между собой их уровень стационарности на этих длинах должен во всяком случае превосходить уровень ошибки, с которой были определены эмпирические частоты, т. е. $\varepsilon > \lambda$.

Анализ текстов показал, что чем меньше ε, т. е. выше задаваемый уровень стационарности, тем больший разброс наблюдается в длинах $L(\varepsilon)$. Для ε = 0,05 всевозможные $L(0{,}05)$ заключены между 10 тыс.

и 40 тыс. знаков. Из оценки (8) следует, что соответствующие вероятности определены с точностью λ от 0,005 до 0,01. Ошибка в определении ε для каждой длины из диапазона 10÷40 тыс. знаков, обусловленная неточностью определения вероятностей, имеет величину порядка $\lambda^2 / (2\varepsilon)$, что не превосходит 0,001 (относительная ошибка менее 2% по сравнению с $\varepsilon = 0{,}05$). Это означает, что 0,05-стационарность определена достаточно корректно. Такой же вывод можно сделать и для 0,03-стационарности. В то же время разброс для $L(0{,}01)$ оказался очень велик, от 40 тыс. до почти 400 тыс. знаков. Поэтому, чтобы иметь относительную ошибку на уровне 2%, необходимо рассматривать тексты с длинами, большими, чем 250 тыс. знаков. В противном случае ошибка, вносимая неточностью в эмпирических вероятностях, может повлиять на статистические выводы о длине стационарности текста, и, в конечном счете, на критерий группировки текста.

Кроме того, анализ показал, что функции $L(\varepsilon)$ для разных произведений одного и того же автора могут существенно различаться, а для разных авторов, напротив, быть весьма близки. Поэтому $L(\varepsilon)$ не может служить опознавательным знаком отдельного писателя. В то же время стабилизация ПФР самих произведений позволяет сделать предположение, что ПФР различных авторов могут быть статистически различимы. Основанием для корректного сравнения авторских ПФР является 0,03-стабилизация всех произведений с объемом более 100 тыс. знаков на этом минимальном объеме независимо от объема самого произведения. Важно также и то, что установление достаточно высокого уровня стационарности происходит на объемах, существенно меньших тех, которые следуют из формулы (3).

3. Кластеризация ПФР по жанрам

Применим анализ ПФР к вопросу об объединении (кластеризации) произведений с близкими распределениями. Заметим, что возможность такой кластеризации заранее не очевидна. Хотя отличия в ПФР двух произведений могут быть близки (например, расстояние между ними меньше 0,05), это не означает, что если два текста близки третьему, то они близки и между собой.

Рассмотрим вопрос о том, существует ли статистическая связь между произведениями, написанными в одном тематическом жанре. Поскольку предполагается сравнивать жанры, а не авторов, то в отдельных примерах будем использовать не только русскоязычную, но и переводную прозу.

Для анализа возьмем 100 произведений объемом более ста тысяч знаков каждый, распределенных по десяти жанрам: классический детектив, «дамский» детектив, киберпанк, ужасы и мистика, классическая фантастика, фэнтэзи, боевик, любовный роман, русская классическая проза, советская проза второй половины XX века. Это деление достаточно условно, и может оказаться, что роман является многожанровым, но в целом имеет смысл попытаться отобрать типичные произведения. В каждом жанре возьмем десять романов, по одному на каждого автора. Для анализа были выбраны следующие произведения:

1. Боевик. Ч. Абдуллаев — Обретение ада, А. Белов — Битва за масть, А. Бушков — Стервятник, В. Горшков — Тюрьма особого назначения, В. Доценко — Месть Бешеного, А. Ильин — Полковник, Д. Корецкий — Пешка в большой игре, Ф. Незнанский — Смертельные игры, А. Таманцев — Гонки на выживание, Д. Черкасов — Невидимки.

2. Дамский детектив. И. Арбенина — Черное солнце, Е. Арсеньева — Моя подруга месть, А. Данилова — Волчья ягода, Д. Донцова — Дама с коготками, О. Играева — Две дамы и король, Л. Ильина — Вредность не порок, С. Климова — Подражание королю, Н. Левитина — Дилетант, А. Маринина — Украденный сон, Т. Полякова — Мой любимый киллер.

3. Классический детектив. Б. Акунин — Статский советник, Ф. Буало, Т. Нарсежак —

Последний трюк каскадера, А. и Г. Вайнеры — Визит к Минотавру, П. Вале, М. Шеваль — Запертая комната, С. Жапризо — Убийственное лето, А. Конан-Дойл — Этюд в багровых тонах, А. Кристи — Смерть в облаках, Ж. Сименон — Гнев Мегрэ, Р. Стаут — Через мой труп, Г. Честертон — рассказы о патере Брауне.

4. Киберпанк. П. Амнуэль — Люди Кода, А. Белаш — Война кукол, У. Гибсон — Нейромантик, А. Лазаревич — Сеть Нанотех, С. Лукьяненко — Лабиринт отражений, Д. Нун — Вирт, О. Палек — Реальная виртуальность, Р. Рукер — Программа, М. Суэнвик — Вакуумные цветы, А. Тюрин — Танцы с Виртуэллой.

5. Ужасы и мистика. У. Блэтти — Изгоняющий дьявола, А. Дашков — Войны некромантов, М. и С. Дяченко — Хозяин колодцев, С. Кинг — Оно, Д. Кунц — Кукольник, Г. Лавкрафт — Хребты безумия, Г. Майринк — Голем, Э. По — рассказы, Б. Стокер — Дракула, А. Толстой — Упырь \ Семья вурдалака.

6. Классическая научная фантастика. А. Азимов — Конец вечности, А. Беляев — Ариэль, Р. Брэдбери — 451 градус по Фаренгейту, П. Буль — Планета обезьян, Р. Желязны, Д. Линдскольд — Хрономастер, А. Казанцев — Купол надежды, А. Кларк — Город и звезды, В. Михайлов — Тогда придите и рассудим, К. Саймак — Магистраль вечности, А. и Б. Стругацкие — Понедельник начинается в субботу.

7. Фэнтэзи. Т. Гудкайнд — Первое правило волшебника, К. Еськов — Последний кольценосец, Л. Кудрявцев — Охотник на магов, Г. Кук — Черный отряд, С. Неграш — Сказочное королевство, И. Новак — Книга дракона, Н. Романецкий — Чародей Свет, М. Семенова — Волкодав, Дж. Толкиен — Сильмариллион, И. Эльтеррус — Бремя императора.

8. Любовный роман. Э. Арсан — Эммануэль, С. Хоум — 69 мест, которые надо посетить с мертвой принцессой, К. Холландер — Парижское Танго, Э. Шаукат — Пламя страсти, Э. Макнейл — Девять с половиной недель; а также тексты: Греческая смоковница, Частные уроки, Дневник Бетти, Каникулы в Калифорнии, История О.

9. Русская классика. Н. Гоголь — Мертвые души, И. Гончаров — Обломов, Ф. Достоевский — Идиот, А. Куприн — Поединок, Н. Лесков — Обойденные, Д. Мамин-Сибиряк — Приваловские миллионы, А. Пушкин — художественная проза (Повести Белкина, Капитанская дочка, Дубровский), М. Салтыков-Щедрин — Господа Головлевы, Л. Толстой — Воскресение, И. Тургенев — Дым.

10. Советская литература. Ф. Абрамов — Дом, Ч. Айтматов — И дольше века длится день, В. Астафьев — Царь-рыба, Ю. Герман — Дорогой мой человек, Д. Гранин — Зубр, Ф. Искандер — Сандро из Чегема, В. Орлов — Альтист Данилов, А. Приставкин — Ночевала тучка золотая, В. Распутин — Прощание с Матерой, В. Солоухин — Трава.

Три крупных жанра (остросюжетный роман, фантастика, социально-психологический роман) были специально разбиты на большее количество категорий с целью выяснить, можно ли более детально идентифицировать жанр произведения. Кроме того, в некоторых группах содержатся произведения, которые следовало бы отнести к другому жанру. Многие книги в стиле «фэнтэзи» являются по сути боевиками, а фантастические романы или романы ужасов часто захватывают читателя своим детективным сюжетом. В этой связи представляется интересным, в какую категорию будут отнесены такие экземпляры.

Построим однобуквенные распределения частот в каждом романе и сравним попарные расстояния между ними. Попробуем сопоставить жанру как таковому некоторую характерную именно для него однобуквенную ПФР. Такая «жанровая ПФР» получается путем смешения всех произведений данного жанра в одно среднее распределение жанра.

Расчеты показали, что расстояние от произведения до жанровой ПФР не может служить индикатором кластеризации про-

изведения по принадлежности к тому или иному жанру, поскольку расстояния между жанровыми ПФР для разных жанров в основном оказались меньше, чем между отдельным произведением и средней ПФР того же жанра.

Тогда вместо неэффективного инструмента «средней жанровой ПФР» рассмотрим сами расстояния между распределениями отдельных текстов. Поскольку все рассматриваемые произведения являются квазистационарными, с объемами, превосходящими максимальную из длин 0,03-стационарности, то будем определять средние расстояния по группе текстов как средние арифметические. Результаты расчетов представлены в табл. 1. Поскольку она симметрична, заполнена только верхняя треугольная часть.

На главной диагонали табл. 1 находятся средние попарные расстояния между романами, написанными в одном жанре. В остальных ячейках приведены средние значения от попарных расстояний между произведениями двух жанров, отвечающих номерам строки и столбца. Отметим, что среднеквадратичное отклонение для этих величин, как и для расстояний между произведениями, написанными в разных жанрах, варьируется по жанрам от 1 до 2 (в единицах, использованных в этой таблице, т. е. в процентах). Это достаточно большая величина, которая не позволяет четко идентифицировать жанр по распределению букв текста, поскольку в интервале ширины 2σ с центром в некотором диагональном элементе оказывается больше половины расстояний между произведениями. Тем не менее, для 68 пар из 90 (т. е. в 75% случаев) эта таблица позволяет правильно сгруппировать романы по их тематике в соответствии с исходным перечнем жанров. Также оказалось возможным указать на более подходящую, чем предполагалось в начале, группировку некоторых произведений. Так, три первых детективных и две последних классических категории характеризуются тем, что для них расстояния между романами внутри каждой из групп значительно меньше межгрупповых расстояний, расположенных в тех же столбцах и строках, что и указанные диагональные элементы. Кроме того, классическая научная фантастика (6-ой жанр) и любовные романы (8-ой жанр) отделяются от остальных жанров, кроме детективных, хотя и не так значимо.

Два жанра — «киберпанк» и «фэнтэзи» (соответственно 4-ый и 7-ой), как и предпо-

Таблица 1

Средние попарные расстояния между однобуквенными распределениями, %

Жанр	1	2	3	4	5	6	7	8	9	10
1	**5,5**	6,2	5,8	6,4	6,5	6,4	6,8	7,1	6,4	6,7
2		**6,2**	6,3	7,3	7,8	7,1	7,4	7,1	6,7	7,0
3			**5,7**	6,6	7,0	6,4	6,9	7,4	6,5	7,0
4				**6,8**	7,4	6,8	7,7	8,0	7,5	7,4
5					**8,0**	7,1	7,8	8,2	7,6	7,9
6						**6,5**	7,3	7,6	7,0	8,1
7							**7,6**	8,1	7,4	7,8
8								**7,4**	7,7	7,6
9									**6,2**	7,1
10										**6,4**

лагалось, не выделяются в самостоятельные группы, а должны быть объединены с боевиками, которыми они по сути и являются. Разумеется, этот вывод сделан не обо всех произведениях указанных жанров, а только о тех, которые были рассмотрены в нашем анализе. Следует также отметить, что «Сильмариллион», будучи скорее мифологическим произведением, не вписывается ни в один из рассмотренных жанров, т. к. его распределение отличается от любого из произведений более чем на 11% (а чаще всего на 13%).

Ужасы и мистика (5-ый жанр) оказались сильно разнородной группой, единственной, для которой внутригрупповые расстояния больше межгрупповых. Это указывает на то естественное обстоятельство, что разные люди «пугаются» различных вещей, необъяснимость которых проявляется в детективной, фантастической и иных формах. Тем самым ужас как таковой не является самостоятельным жанром: он выступает, например, как страшный детектив, фантастика с монстрами, жуткая любовная история и т. п.

Рассматривая расстояния между отдельными произведениями, а не только между жанрами в целом, можно более точно классифицировать их тематическую направленность. Так, «дамский детектив» С. Климовой более близок к боевикам, расстояние до которых в среднем составило 4,8, а роман А. Марининой — к классическому детективу (расстояние 4,7), что, в общем-то, справедливо. «Частные уроки», формально отнесенные к любовным романам, по расстоянию гораздо ближе к классическому детективу, что также соответствует фактическому содержанию этого произведения. Мистические романы «Дракула» и «Голем» также оказались наиболее близки к детективам, а «Хребты безумия» — к фантастике.

Таким образом, проведенный анализ показал определенную жанровую кластеризацию литературных текстов, написанных разными авторами. Однако разделить произведения по жанрам, используя только однобуквенное распределение, с достаточно высокой достоверностью (например, не хуже 0,9) оказалось невозможным. Далее рассмотрим распределения, отвечающие текстам, написанным одним автором.

4. Кластеризация ПФР по авторам

Исследуем вопрос кластеризации произведений, написанных одним автором, по той же схеме, что и анализ различных жанров. Задача состоит в построении среднего «авторского» распределения и в определении средних расстояний между произведениями, написанными одним и, соответственно, разными авторами. Можно ли по расстояниям между текстами сгруппировать те из них, которые написаны одним писателем?

Для анализа были взяты десять произведений одного автора, написанные приблизительно в одинаковом жанре, определены попарные расстояния между ними, а также среднее расстояние между произведениями данного автора. Такая процедура была проделана с произведениями десяти авторов: Ч. Айтматов, Б. Акунин, М. Булгаков, Н. Гоголь, Д. Донцова, Ф. Достоевский, О. Маркеев, В. Набоков (однако, разные переводчики), Л. Толстой, И. Тургенев. Это не вполне репрезентативная выборка писателей, но она хорошо иллюстрирует методику. Заметим, что существующий численный алгоритм позволяет за короткое время создать расчетный файл выходных данных по 100 тысячам произведений объемом более 500 тысяч знаков каждый (в том числе и на иностранных языках), т. е. при необходимости можно провести сравнительный анализ распределений для практически всех достаточно плодовитых писателей.

Рассмотрим величину $L(\varepsilon)$ для выбранных десяти авторов. Для сравнения их ПФР необходимо, чтобы все сравниваемые произведения имели бы установившиеся распределения на длине минимального из текстов. Средние авторские величины длин $L(\varepsilon)$, вычисленные для 100 книг десяти авторов при $\varepsilon = 0{,}02$; 0,03, приведены в табл. 2. Из этих

Таблица 2

Средние авторские длины квазистационарности

Автор	*L* (0,03)	*L* (0,03), % от длины книги	*L* (0,02)	*L* (0,02), % от длины книги
Ч. Айтматов	59 589	33	100 763	50
Б. Акунин	51 277	13	103 005	25
М. Булгаков	51 685	30	97 637	51
Н. Гоголь	42 008	42	66 446	61
Д. Донцова	38 578	10	100 446	26
Ф. Достоевский	86 066	31	149 588	46
О. Маркеев	87 889	24	170 736	39
В. Набоков	42 773	17	72 196	30
Л. Толстой	76 797	22	176 870	45
И. Тургенев	59 025	27	99 695	43
Среднее	**59 568**	**25**	**113 738**	**41**

данных видно, что в рассмотренных произведениях при допустимом отклонении в 3% квазистационарность достигается в среднем на четверти произведения, или на 60 тыс. знаках. Эта величина может быть существенно меньше для конкретного автора. Например, тексты Донцовой 0,03-стационарны в среднем уже на 10% текста. Аналогичные наблюдения и выводы можно сделать в отношении Акунина и, по-видимому, многих других «сериальных» писателей.

В первом приближении все распределения достаточно больших текстов весьма похожи, так что почти все локальные максимумы и минимумы распределений приходятся на одни и те же буквы. Однако средние авторские распределения отражают предпочтения авторов в использовании тех или иных букв в большей степени, чем средние жанровые ПФР, что позволяет провести кластеризацию большинства произведений вокруг средних авторских ПФР.

Будем проводить кластеризацию отдельных произведений путем сравнения близости их ПФР. Рассмотрим, насколько близки ПФР отдельных произведений к средним авторским ПФР. Результаты расчетов приведены в табл. 3. Эта таблица несимметрична, поскольку, например, среднее расстояние от отдельных произведений Тургенева до «среднего Толстого» не должно быть равно среднему расстоянию от отдельных произведений Толстого до «среднего Тургенева».

Из таблицы 3 видим, что все расстояния, стоящие в некоторых строке и столбце, не меньше (а за исключением двух значений из ста — строго больше), чем расстояния, находящиеся на их пересечении на главной диагонали. Это свидетельствует о четкой отделимости писателей по их авторской ПФР. В данной выборке неоднозначно идентифицируются только 15 произведений из 100. Кроме того, все авторы, кроме Акунина, которого можно спутать с Гоголем или Толстым, идентифицируются однозначно. Характерно, что при этом Гоголь четко отделяется от Толстого.

Если состав авторов неизвестен, т.е. авторские ПФР отсутствуют, то задача кластеризации произведений решается так же, как и в случае с жанрами, т.е. сравнением близости всех возможных пар. В таблице 4 при-

Таблица 3

Средние попарные расстояния между отдельными произведениями и авторскими ПФР, %

Автор	Айтм. (ср.)	Акун. (ср.)	Булг. (ср.)	Гог. (ср.)	Донц. (ср.)	Дост. (ср.)	Марк. (ср.)	Наб. (ср.)	Толст. (ср.)	Тург. (ср.)
Айтм.	4,4	7,3	7,6	7,2	7,5	8,3	7,8	7,0	6,8	6,8
Акун.	6,0	2,4	4,3	6,0	6,3	6,7	4,5	4,2	4,8	4,9
Булг.	7,3	5,4	4,0	6,9	6,6	8,9	4,9	6,2	6,7	6,3
Гог.	7,5	7,1	7,5	6,0	9,2	7,2	7,9	8,4	6,2	7,0
Донц.	6,6	6,1	6,0	7,9	2,1	9,0	5,6	7,1	7,0	6,6
Дост.	7,5	7,6	9,1	7,0	9,6	3,9	9,8	7,8	6,4	6,5
Марк.	7,1	5,1	4,2	7,2	5,9	9,9	2,8	6,3	7,1	6,8
Наб.	6,5	5,0	5,6	6,3	7,3	7,6	6,2	3,8	5,7	5,5
Толст.	7,1	6,5	6,9	6,6	8,4	7,2	7,4	6,7	4,8	6,0
Тург.	6,1	5,5	5,9	6,7	7,1	6,0	6,8	5,5	4,9	3,4

ведены результаты расчетов средних меж-авторских расстояний, определяемых как средние арифметические по всем соответствующим парам произведений. На диагонали этой таблицы расположены средние расстояния между произведениями данного автора, а в остальных клетках — средние расстояния между произведениями двух данных авторов.

Из таблицы 4 следует, что в отличие от жанровой кластеризации, точность идентификации авторов по средним попарным расстояниям несколько хуже — около 80%, хотя большинство авторов по-прежнему весьма четко отделяются друг от друга.

Однако следует подчеркнуть, что большие расстояния между ПФР не обязательно свидетельствуют о разных авторах этих текстов, т. е. близость расстояний между произведениями одного и того же писателя в значительной мере обусловлена жанром. Так, расстояние между повестями Н. Гоголя из «Вечеров на хуторе близ Диканьки» равно в среднем 0,027, а между частями «Мертвых душ» 0,034, что также невелико. Однако расстояние между «вечерами» и «душами» значительно больше — оно равно 0,062 — хотя и не такое, как в среднем для разных писателей, но все же существенное.

Наиболее ярко тематическая и одновременно авторская кластеризация выражена у современных «сериальных» писателей. Так, у Б. Акунина, Д. Донцовой, и О. Маркеева весьма небольшое расстояние между текстами — в среднем 0,024, причем разброс расстояний также очень мал, их отклонения в среднем квадратичном равны 0,012.

Задача определения авторства 100 вышеуказанных произведений по расстоянию до среднего авторского распределения, образованного остальными девятью произведениями автора, показала, что в 85 случаях из 100 это удается успешно осуществить: минимальное из десяти расстояний от какого-либо произведения до средней авторской ПФР отвечает правильному автору этого произведения. Из оставшихся 15 несовпадений в половине случаев правильный ответ отделяет от них расстояние меньше полпроцента.

Рассмотрим еще один пример применения буквенных распределений к одной проблеме авторства, ставшей уже «классической». Речь идет о спорах вокруг авторства романа «Тихий Дон». Не обсуждая литературные аргументы «за» и «против», приведем результат сравнительного статистического анализа четырех частей этого рома-

Таблица 4

Средние попарные расстояния между отдельными произведениями, %

Автор	Айтм.	Акун.	Булг.	Гог.	Донц.	Дост.	Марк.	Наб.	Толст.	Тург.
Айтм.	**6,2**	7,9	8,7	8,7	7,8	9,0	8,4	7,9	8,4	7,6
Акун.		**3,6**	6,0	7,3	6,7	8,0	5,7	5,7	7,1	6,0
Булг.			**6,0**	8,3	7,0	9,9	5,8	7,0	8,2	7,2
Гог.				**6,1**	9,4	8,2	8,5	7,6	7,8	9,5
Донц.					**3,3**	9,8	6,4	7,7	8,7	7,4
Дост.						**5,8**	10,8	8,7	8,5	7,4
Марк.							**4,2**	7,1	8,3	8,4
Наб.								**5,6**	7,7	6,6
Толст.									**7,6**	7,1
Тург.										**4,9**

на с другими произведениями М. Шолохова: романами «Поднятая целина», «Они сражались за Родину», повестями (в целом) и рассказами (в целом). Все части «Тихого Дона» имеют близкое распределение, как и должно быть для произведения, написанного одним автором в одном стиле: отличие составляет 0,030. «Поднятая целина», имеющая объем, приблизительно равный одной части «Тихого Дона», отличается от каждой его части на 0,056. Приблизительно такое же расстояние и между другими произведениями Шолохова: среднее расстояние между ними (без учета анализируемой эпопеи) составило 0,058, причем среднеквадратичное отклонение этих расстояний очень мало и равно 0,007. Таким образом, «Тихий Дон» явно группируется с остальными произведениями Шолохова. Поэтому, скорее всего, мы имеем дело с разными произведениями, написанными одним автором.

Разумеется, при использовании статистического метода всегда имеется вероятность ошибки. Но, поскольку количество анализируемых произведений ограничено и не очень велико, априори оценить эту вероятность не представляется возможным, поэтому получаемые результаты могут носить только рекомендательный характер.

5. Двухбуквенные ПФР

Построим теперь двухбуквенные ПФР (2-ПФР) для отдельных литературных произведений. Как и выше, попробуем воспользоваться расстоянием между распределениями для задачи кластеризации произведений по жанрам и авторам. Действуя так же, как и для 1-ПФР в п. 2, можно показать, что на текстах длиной более 100 тыс. знаков вероятности определяются с точностью, не влияющей на расстояния между ПФР.

Определив расстояния между 2-ПФР для произведений, перечисленных в п. 3, можно провести кластеризацию по этому признаку, объединив в одну группу тексты, попарные расстояния между которыми приблизительно одинаковы и значительно меньше, чем с другими произведениями. В указанной выборке удалось правильно сгруппировать 85 текстов из 100. Этот результат выше, чем для однобуквенных распределений, где аналогичный метод дал точность в 75%.

Более существенное увеличение точности группировки было обнаружено в задаче кластеризации текстов по авторам, рассмотренным в п. 4. Если ПФР автора неизвестна,

Таблица 5

Средние по авторам попарные расстояния между 2-ПФР, %

Автор	Айтм.	Акун.	Булг.	Гог.	Донц.	Дост.	Марк.	Наб.	Толст.	Тург.
Айтм.	**19,9**	24,1	26,0	25,4	23,7	25,4	25,0	23,8	25,2	23,2
Акун.		**12,5**	18,4	22,8	18,1	23,5	17,5	17,8	23,1	19,9
Булг.			**19,5**	25,4	21,1	27,0	20,3	20,5	25,5	22,9
Гог.				**21,4**	26,2	23,7	26,0	23,6	23,8	23,0
Донц.					**11,0**	25,2	18,4	21,1	25,8	21,0
Дост.						**18,3**	28,0	25,0	24,3	21,4
Марк.							**14,4**	20,2	25,7	23,0
Наб.								**16,8**	22,8	21,8
Толст.									**20,6**	22,7
Тург.										**16,3**

то точность группировки текстов составила почти 90%. Если же есть возможность сформировать средние авторские ПФР, то автор успешно определяется в 95 случаях из 100. В таблице 5 приведены средние по авторам попарные расстояния между 2-ПФР ста произведений десяти авторов.

Приведем также наиболее часто употребляемые пары $i - j$ (первые 40) в рассматриваемой выборке текстов 10 авторов (табл. 6), где f — частота встречаемости пар. Эти тексты можно рассматривать как представительную выборку русского литературного языка.

Вопрос об использовании ПФР более высоких размерностей связан с естественным ограничением на достоверность оценки частот использования буквенных комплексов. Одной из задач, требующих решения, является определение максимальной величины комплексов букв (тройки, четверки и т. п.), распределение которых позволяет уточнять статистические выводы. Эта величина зависит от объема текста.

Вариантом решения задачи повышения точности классификации произведений может быть использование многомерного фазового пространства и выделение его основных компонент. Таковыми могут служить: авторство, жанровая принадлежность, размер произведения, время написания, принадлежность автора той или иной школе или кругу, его социальный слой и т. п. В этом случае интересной задачей является изучение зависимости расстояния между ВПФР произведений и их расстоянием в таком расширенном фазовом пространстве.

Заключение

В работе показано, что выборочные функции распределения текстов по буквам и парам букв могут служить инструментом для группировки произведений по авторам и жанрам. Несмотря на нестационарность распределений, при объемах более 100 тыс. знаков тексты можно считать квазистационарными с ошибкой не более 0,03, рассчитанной в норме суммируемых функций для однобуквенных распределений.

Распределения букв в произведениях, написанных одним автором в одном жанре, отличаются, как правило, менее чем на 0,055, тогда как для разных авторов отличие в сопоставимых по объему текстах одного жанра не ниже 0,07, чаще же оно порядка 0,1.

Таблица 6

Наиболее употребительные пары букв в совокупности текстов

№	ij	f	№	ij	f	№	ij	f	№	ij	f
1	то	0,01761	11	ро	0,01080	21	ер	0,00903	31	ор	0,00791
2	ст	0,01514	12	ка	0,01062	22	ос	0,00896	32	ом	0,00770
3	на	0,01497	13	го	0,01062	23	ол	0,00891	33	ил	0,00769
4	но	0,01430	14	ни	0,01018	24	ло	0,00872	34	те	0,00747
5	по	0,01376	15	ла	0,01009	25	та	0,00845	35	за	0,00743
6	не	0,01355	16	ен	0,01001	26	ва	0,00839	36	ет	0,00729
7	ал	0,01226	17	пр	0,00942	27	ре	0,00835	37	ве	0,00710
8	ко	0,01217	18	ли	0,00920	28	ел	0,00826	38	ри	0,00698
9	ов	0,01094	19	во	0,00911	29	он	0,00806	39	од	0,00691
10	ра	0,01089	20	от	0,00908	30	ть	0,00805	40	ак	0,00682

Это можно трактовать как различие «фирменных подписей» писателей, которые они непроизвольно оставляют в виде квазистационарных распределений букв.

Критерий кластеризации, основанный на близости между двухбуквенными распределениями текстов, позволил правильно идентифицировать автора с ошибкой не более 5%, а жанр — с ошибкой не более 15%. Однобуквенные распределения дали ошибку соответственно 15 и 25%.

Список литературы

1. *Марков А. А.* Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь. // Известия Имп. Акад. наук, 1913, серия VI, Т. X, № 3, с. 153.
2. *Хмелев Д. В.* Распознавание автора текста с использованием цепей А. А. Маркова. // Вестник МГУ, 2000, сер. 9: филология, № 2, с. 115–126.
3. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов. // Фоменко А. Т. Новая хронология Греции: Античность в средневековье. Т. 2. — М.: Изд-во МГУ, 1996, с. 768–820.
4. *Добрушин Р. Л.* Математические методы в лингвистике. // Математическое просвещение, 1961, вып. 6, с. 37–60.
5. *Лебедев Д. С., Гармаш В. А.* О возможности увеличения скорости передачи телеграфных сообщений. // Электросвязь, 1958, № 1, с. 68–69.
6. *Урбах В. Ю.* К учету корреляций между буквами алфавита при вычислении количества информации в сообщении. // Проблемы кибернетики, вып. 10, 1963, с. 111–117.
7. *Невельский П. Б., Розенбаум М. Д.* Угадывание профессионального текста специалистами и неспециалистами. / В сб. Статистика речи и автоматический анализ текста. — Л.: Наука, 1971, с. 134–148.
8. *Яглом А. М., Яглом И. М.* Вероятность и информация. — М.: КомКнига, 2007. — 512 с.
9. *Орлов Ю. Н., Осминин К. П.* Построение выборочной функции распределения для прогнозирования нестационарного временного ряда. // Математическое моделирование, 2008, № 9, с. 23–33.
10. *Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф.* Справочник по теории вероятностей и математической статистике. — М.: Наука, 1985. — 640 с.