Распознавание автора текста с использованием цепей А.А. Маркова

Д.В. Хмелёв

май 1999

Статья опубликована в журнале Вестник МГУ, сер.9: Филология, №2, 2000, с.115-126

Аннотация

В статье посредством формального анализа текста решается задача определения авторства текста. Новый метод основывается на формальной математической модели последовательности букв текста как реализации цепи А.А. Маркова. Оказывается, частоты употребления пар букв очень хорошо характеризуют автора. Последнее утверждение проверено в объемном статистическом эксперименте на произведениях 82 писателей.

Содержание

1 Введение **1**

2 Математические основания **3**

2.1 Марковская модель 3

2.2 Схема эксперимента 4

2.3 Анализ частот употреблений букв (схема Бернулли) 5

3 Модельный эксперимент **5**

4 Результаты более объемного вычислительного эксперимента **9**

5 Заключение **13**

1 Введение

В последние десятилетия наметилась тенденция поиска и выявления характерных структур авторского языка путем применения формально-количественных, статистических методов. Первые пробные шаги на этом пути предпринял еще в начале XIX века Н.А. Морозов [1]. Интересно, что тогда же известный математик А.А. Марков выступил с критикой Н.А. Морозова [2]. А.А. Марков критиковал Н.А. Морозова за то, что он не произвел тщательной статистической проверки утверждений относительно устойчивости некоторых элементов авторского стиля (например, частицы "не"). Примером правильного статистического подхода А.А. Марков считал свое исследование в статье [3], где он изучал распределение доли гласных и согласных среди первых 20000 букв "Евгения Онегина". Отметим, что работа [3] посвящена первому применению "испытаний, связанных в цепь", получивших впоследствии название цепей А.А. Маркова. Работа [3] удивительным образом предоставляет историческую основу методу определения авторства, изложенному в настоящей статье.

Истории вопроса определения авторства текста посвящена первая глава книги [4]. Несмотря на то, что среди перечисленных в [4] присутствуют весьма изощренные методики определения структуры авторского стиля, все они страдают, на взгляд автора настоящей статьи, одним общим недостатком. Ни одна из этих методик не проходила проверку на большом числе писателей. Дело в том, что многие из методик имеют трудно формализуемый этап сведения естественно-литературного произведения к предлагаемой математической модели. Например, для некоторых из них необходима классификация всех слов текста по грамматическим классам, что требует участия человека. При таком подходе большой вычислительный эксперимент с целью проверки методики на большом числе авторов практически неосуществим. Поэтому все такие методики пытались использовать на небольшом числе авторов.

Другого подхода придерживались авторы статьи [5]. Они исследовали несколько простых параметров авторского стиля и на огромном числе произведений писателей XVIII-XX веков статистически доказали, что доля всех служебных слов в длинном прозаическом произведении является т.н. авторским инвариантом. В настоящей статье предложен новый, независимый от [5], а также всех методик, перечисленных в [4], формальный метод установления автора текста. Наша постановка задачи отличается от [5]. Мы предполагаем, что в нашем распоряжении имеются достаточно длинные фрагменты прозаических произведений ряда авторов на русском языке. Про некоторый анонимный фрагмент текста известно, что он принадлежит одному из этих авторов, но какому — неизвестно. Требуется узнать — кому именно.

Новый метод основывается на формальной математической модели последовательности букв текста как реализации цепи А.А. Маркова. По тем произведениям автора, которые достоверно им созданы, вычисляется матрица переходных частот употреблений пар букв. Она служит оценкой матрицы вероятностей перехода из буквы в букву. Матрица переходных частот строится для каждого из авторов. Для каждого автора оценивается вероятность того, что именно он написал анонимный фрагмент текста. Автором анонимного текста полагается тот, у которого вычисленная оценка вероятности больше.

Такой метод оказывается удивительно точным для естественно-языковых текстов. Мы обсуждаем здесь особенности применения метода и сравниваем его с методом определения автора на основе частот употребления каждой из букв в отдельности. Проверка метода проводилась на произведениях 82 писателей, среди которых есть русские писатели как XIX, так и XX века.

2 Математические основания

2.1 Марковская модель

Обозначим через A некоторое множество букв. Через A^k обозначим множество слов длины k над алфавитом A. Пусть $A^* = \cup_{k>0} A_k$. Обозначим длину слова $f \in A^*$ через $|f|$.

Задачу определения автора текста можно сформулировать следующим образом. Пусть заданы n классов C_i, где $i = 0, \ldots, n-1$. В каждом классе C_i находятся последовательности $f_{i,j} \in A^*$, где $j = 1, \ldots, m_i$, т.е. $C_i = \{f_{i,j} | j = 1, \ldots, m_i\}$. Наша задача состоит в том, чтобы отнести $x \in A^*$ к одному из классов C_i.

Предположим, что последовательности букв $f_{i,j}$ являются реализациями цепи Маркова с переходной матрицей Π^i. Построим оценку P^i. Обозначим через $h_{i,j,kl}$ число переходов букв $k \mapsto l$ в фрагменте $f_{i,j}$, положим $h_{i,kl} = \sum_j h_{i,j,kl}$, а $h_{i,k} = \sum_l h_{i,kl}$. Положим $P^i_{kl} = h_{i,kl}/h_{i,k}$. Возможно, некоторые P^i_{kl} равны нулю. Обозначим через Z_i множество таких упорядоченных пар (k, l), что $P^i_{kl} > 0$.

Предположим, что x также является реализацией цепи Маркова с матрицей переходных вероятностей P^θ , где θ неизвестный параметр, который принимает одно из значений 1, …, n.

Обозначим через $\nu_{k,l}$ число переходов $k \mapsto l$ в x. Пусть также $\nu_k = \sum_l \nu_{k,l}$. Обозначим через

$$L_i(x) = -\sum_{(k,l)} \nu_{k,l} \times \ln(\nu_{k,l}/(P^i_{kl} \times \nu_k)),$$

где сумма берется по парам $(k, l) \in Z_i$. Грубо говоря, $L_i(x)$ равно минус логарифму вероятности x при условии, что x — реализация цепи Маркова с матрицей переходных вероятностей P^i. Назовем $t(x)$ оценкой максимального правдоподобия для неизвестного параметра θ

$$t(x) = \mathrm{argmin}_{i=0,\dots,n-1} L_i(x). \tag{2.1}$$

Мы не будем обсуждать и доказывать какие-либо математические свойства оценки (2.1), хотя это и представляет интересную задачу математической статистики (более подробно см. [6, с. 224]). Зато мы покажем удивительную эффективность оценки (2.1) при установлении автора текста.

2.2 Схема эксперимента

Возьмем $A = \{$маленькие буквы кириллицы$\} \cup \{$символ пробела$\}$. Предположим, что у нас имеются в распоряжении достаточно длинные фрагменты произведений n авторов на русском языке. Обозначим j-тый фрагмент i-того автора через $g_{i,j}$. Можно считать, что фрагмент $g_{i,j}$ является последовательностью символов некоторого расширенного алфавита B, который включает, например, знаки пунктуации, большие буквы, латинские буквы и т.д. (на персональном компьютере B обычно совпадает с расширенным множеством символов ASCII).

Каждый фрагмент $g_{i,j} \in B^*$ можно отобразить в A^* посредством некоторой функции $F : B^* \to A^*$. Пусть, например, F превращает все заглавные буквы в маленькие, склеивает перенесенные слова, выбрасывает все знаки пунктуации и излишние знаки пробела, оставляя их по одному между словами, а также вставляет один пробел в начале и один пробел в конце фрагмента в случае отсутствия таковых.

Кроме того, мы будем рассматривать функцию G, которая устроена так же, как и функция F, с тем дополнением, что все слова, которые в фрагменте $g_{i,j}$ начинались с заглавной буквы, отбрасываются. Например, если

$y =$ "Кроме␣того,␣мы␣будем␣рассматривать␣функцию␣G,", то

$F(y) =$ "␣кроме␣того␣мы␣будем␣рассматривать␣функцию␣", а

$G(y) =$ "␣того␣мы␣будем␣рассматривать␣функцию␣".

Теперь предположим, что некий фрагмент текста $y \in B^*$ принадлежит одному из n авторов, и нам неизвестно, кому именно. Наша задача: определить автора фрагмента y. Мы можем найти автора, применяя оценку (2.1) к последовательности $x = F(y)$ или к $x = G(y)$. Следовательно, мы получаем два способа определения автора:

1) истинный автор — $t(F(y))$,

2) истинный автор — $t(G(y))$.

Важно отметить, что оценки $t(F(y))$ и $t(G(y))$ вычисляются на основе информации о частотах употребления пар букв. Поскольку между словами вставлены пробелы, оценки $t(F(y))$ и $t(G(y))$ никак не зависят от порядка самих слов. По-видимому, $t(F(y))$ и $t(G(y))$ характеризуют последовательности морфем в словоформах русского языка, но, конечно, совсем не учитывают синтаксисическую информацию (на основе последней пытались устанавливать авторство в [4]).

Обычно ни для одного из естественно-языковых текстов гипотеза о том, что он является реализацией соответствующей цепи А.А. Маркова, не выдерживает статистической проверки. Между тем, мы можем формально произвести все вычисления и найти оценку (2.1). Статистический эксперимент показывает, что авторы определяются очень уверенно.

2.3 Анализ частот употреблений букв (схема Бернулли)

Схемой Бернулли в теории вероятностей называется последовательность независимых одинаково распределенных случайных величин. Формально мы можем предположить, что последовательности $f_{i,j}$ и x являются реализациями последовательности независимых одинаково распределенных случайных величин, принимающих значения в A, а x распределен как величины класса η, где η — неизвестный параметр. Тогда оценка (2.1) принимает вид

$$e(x) = \mathrm{argmin}_i G_i(x), \qquad (2.2)$$

где

$$G_i(x) = -\sum_k \nu_k \ln((\nu_k \times h_i)/(hi,k \times \nu)),$$

где сумма вычисляется по таким k, что $\nu_k > 0$, а $\nu = \sum_k \nu_k$, $h_i = \sum_k h_{i,k}$ и. Грубо говоря, производя оценку $\eta(x)$ мы производим частотный анализ текста. Статистический эксперимент показывает, что оценка $e(x)$ существенно хуже оценки $t(x)$.

3 Модельный эксперимент

Сначала проведем проверку нашей методики на следующем примере. Рассмотрим следующие произведения К. Булычева, А. Волкова, Н.В. Гоголя и В. Набокова.

Мы хотим проверить эффективность оценки $t(F(y))$. Предлагается следующий способ: выбрать каждого автора i $(i = 0, 1, 2, 3)$ по одному контрольному произведению у i, оценить матрицы Π^i по другим произведениям $f_{i,j}$, а затем найти $t(F(y^i))$. Если оценка работает хорошо, то для каждого автора i должно быть $t(F(y^i)) = i$.

0) К. Булычев: Умение кидать мяч (y^0); Белое платье золушки ($g_{0,1}$); Великий дух и беглецы ($g_{0,2}$); Глубокоуважаемый микроб ($g_{0,3}$); Закон для дракона ($g_{0,4}$); Любимец [Спонсоры] ($g_{0,5}$); Марсианское зелье ($g_{0,6}$); Миниатюры ($g_{0,7}$); "Можно попросить Нину?"($g_{0,8}$); На днях землетрясение в Лигоне ($g_{0,9}$); Перевал ($g_{0,10}$); Показания Оли Н. ($g_{0,11}$); Поминальник XX века ($g_{0,12}$); Раскопки курганов в долине Репеделкинок ($g_{0,13}$); Тринадцать лет пути ($g_{0,14}$); Смерть этажом ниже ($g_{0,15}$);

1) А. Волков: Семь подземных королей (y^1); Волшебник изумрудного города ($g_{1,1}$); Урфин Джюс и его деревянные солдаты ($g_{1,2}$); Огненный бог Марранов ($g_{1,3}$); Гениальный пень ($g_{1,4}$); На войне, как на войне ($g_{1,5}$); О чем молчали газеты... ($g_{1,6}$); Преступление и наказание ($g_{1,7}$); Эпилог ($g_{1,8}$); Желтый Туман ($g_{1,9}$); Тайна заброшенного замка ($g_{1,10}$);

2) Н.В. Гоголь: Рассказы и повести (y^2, названия повестей: "Повесть о том, как поссорился Иван Иванович с Иваном Никифоровичем", "Старосветские помещики", "Вий", "Записки сумасшедшего"); Ревизор ($g_{2,1}$); Тарас Бульба ($g_{2,2}$); Вечера на хуторе близ Диканьки ($g_{2,3}$);

3) В. Набоков: Другие берега (у3); Король, дама, валет ($g_{3,1}$); Лолита ($g_{3,2}$); Машенька ($g_{3,3}$); Рассказы ($g_{3,4}$); Незавершенный роман ($g_{3,5}$).

Например, у А. Волкова контрольным произведением является y^1, т.е. "Семь подземных королей."Все остальные произведения используются для вычисления Π^i. Результаты вычислений представляются следующей таблицей.

Таблица 1

N	Автор	c_1	c_2	c_3	c_4
0	К. Булычев	0	15	2345689	75161
1	А. Волков	0	8	1733165	233418
2	Н.В. Гоголь	0	3	723812	243767
3	В. Набоков	0	5	1658626	367179

Столбец c_2 содержит общее число файлов, в которых хранятся произведения автора. Заметим, что число файлов может не совпадать с числом произведений по двум причинам: во-первых, несколько произведений одного автора могут находится в одном файле (здесь такое произошло с А. Волковым — три повести "Желтый Туман", "Тайна заброшенного замка"и "Огненный бог Марранов"были в одном файле); во-вторых, одно большое произведение может разбиваться на несколько частей (последнее необходимо учитывать при изучении таблицы 2).

В колонке c_3 содержится суммарное число символов (букв и пробелов) в $F(g_{i,j})$: $c_3 = \sum_j |F(g_{i,j})|$. В колонке c_4 содержится число символов в $F(y^i)$, т.е. $c4 = |F(y^i)|$. Например, для К. Булычева общий объем текстов $\sum_j F(g_{0,j})$ составляет 2'345'689. Общий объем $F(y^1)$, т.е. число символов A в повести "Умение кидать мяч", выбранной в качестве контрольного текста, равно 75'161.

В столбце c_1 в строке j находится ранг числа $L_j(F(y^j))$ среди чисел $\{L_i(F(y^j))|i = 0, 1, 2, 3\}$. Под рангом мы подразумеваем номер $L_j(F(y^j))$ среди чисел $\{L_i(F(y^j))|i = 0, 1, 2, 3\}$, расположенных в порядке невозрастания. Например, если $j = 1$ и L_i расположились в порядке $L_0 \leq L_3 \leq L_2 \leq L_1$, то рангом L_1 будет 3. А если $j = 0$ и L_i расположились в том же порядке $L_0 \leq L_3 \leq L_2 \leq L_1$, то рангом L_0 будет 0. Ранг $L_j(F(y^j))$, среди чисел $\{L_i(F(y^j)|i = 0, 1, 2, 3\}$ совпадает с рангом $L_j(F(y^j))/|F(y^j)|$, среди чисел $\{L_i(F(y^j))/|F(y^j)| \mid i = 0, 1, 2, 3\}$. Расположим в строках $j = 0, 1, 2, 3$ следующей матрицы по 4 числа $L_i(F(y^j))/|F(y^j)|, i = 0, 1, 2, 3$:

$$L = \begin{pmatrix} 2.484569 & 2.508425 & 2.504301 & 2.49377 \\ 2.501061 & 2.473907 & 2.516797 & 2.492874 \\ 2.499033 & 2.504508 & 2.480202 & 2.483829 \\ 2.541367 & 2.538101 & 2.548842 & 2.520018 \end{pmatrix}.$$

В каждой строке найдем ранги чисел L_i:

$$R = \begin{pmatrix} 0 & 3 & 2 & 1 \\ 2 & 0 & 3 & 1 \\ 2 & 3 & 0 & 1 \\ 2 & 1 & 3 & 0 \end{pmatrix}.$$

Искомые числа столбца c_1 стоят на диагонали. Вспоминая формулу (2.1), мы заключаем, что $t(F(y^j)) = j$ тогда и только тогда, когда ранг $L_j(F(y^j))/|F(y^j)|$ среди чисел $\{L_i(F(y^j))/|F(y^j)||i = 0, 1, 2, 3\}$ просто равен 0. Следовательно, если в какой-либо строке в столбце $c1$ таблицы 1 стоит 0, то авторство контрольного текста определено правильно. Из таблицы 1 мы видим, что у всех писателей авторство определено верно.

Прежде, чем обсудить этот результат, поясним, почему столбец $c1$ задан таким образом. Дело в том, что если авторство определено неверно (т.е., оказалось $t(F(y^j)) \neq j$), то нас может интересовать, насколько мы были близки к правильному ответу. Если ранг $L_j(F(y^j))/|F(y^j)|$ среди чисел $\{L_i(F(y^j))/|F(y^j)||i = 0, 1, 2, 3\}$ равен 1, то мы ошиблись всего на одного писателя. Такой случай существенно лучше случая ранга $L_j(F(y^j))/|F(y^j)|$ равного 3, поскольку тут правильный писатель оказывается в списке претендентов на его собственное произведение последним, что свидетельствует о большей ошибке.

Кроме того, матрица R сама по себе допускает интересные интерпретации. Например, из первой строки мы видим, что контрольное произведение К. Булычева "Умение кидать мяч"после самого К. Булычева больше походит на В. Набокова, затем на Н. Гоголя, и в последнюю очередь на произведения А. Волкова. Из последующих двух строк можно сделать вывод, что контрольные произведения А. Волкова и Н. Гоголя также в первую очередь походят на произведения В. Набокова. Может быть, это вызвано тем, что сам Набоков исторически находится между Н. Гоголем и парой писателей: А. Волковым и К. Булычевым? Если эта гипотеза верна, то наша метод чувствителен к исторической эпохе, в которую создано произведение. Некоторое подтверждение тому мы находим в последней строке матрицы R: контрольное произведение В. Набокова похоже в первую очередь на пару А. Волкова и К. Булычева, и лишь затем — на Н. Гоголя. Если бы пара А. Волкова и К. Булычева разбивалась Н. Гоголем. то мы имели бы аргумент против нашей гипотезы. Впрочем, возможны другие интерпретации матрицы R, и автор нисколько не настаивает на выше приведенной.

Можно интересоваться зависимостью матрицы R от

а) числа и объема текстов обучающих выборок;

б) однородности по жанру;

в) однородности по тематике;

г) длины контрольного текста;

д) единицы анализа (на уровне букв, слов и предложений)

и многих других параметров. Ниже мы приводим информацию относительно пункта а). Вкратце вывод таков: методика работает удовлетворительно (то есть, на диагонали матрицы R в основном стоят 0) при объеме обучающей выборки свыше 100 тысяч символов ASCII, и объеме контрольного текста свыше 100 тысяч символов ASCII.

Вернемся к обсуждению таблицы 1. Поскольку в столбце $c1$ все числа равны 0, авторство всех контрольных произведений определено верно. Результат тем более неожиданный, что мы использовали столь примитивную информацию о тексте, как частоты употребления пар букв. На самом деле простейший компьютерный эксперимент (результаты которого здесь не приведены) показал, что при небольшом числе подозреваемых писателей (меньше шести) даже оценка (2.2), основанная всего лишь на подсчете частот употребления букв, дает очень хорошие результаты. В следующем разделе описан значительно более объемный статистический эксперимент. Из него становится ясно, что методика устойчиво работает на очень большом числе авторов.

4 Результаты более объемного вычислительного эксперимента

В электронной библиотеке "Самые любимые книжки"нашлось $n = 82$ различных автора, которые творили в XIX-XX веках. Количество произведений разных авторов колебалось от 1 до 30 (например, у Аркадия и Бориса Стругацких). У немногих авторов, у которых нашлось лишь одно произведение (например, у Бориса Стругацкого), оно было поделено на две части, одна из которых использовалась в качестве контрольного текста. При отборе произведений учитывался объем: выбирались авторы, суммарный объем произведений которых превышал 100000 символов ASCII. Общее число произведений (романов, повестей, рассказов и т.п.) превысило 1000. Они были представлены в 386 файлах. Общий объем данных составил 128 $\times$

10^6 символов ASCII.

Для каждого автора мы составили список $g_{i,j}$ текстов, из которых были получены оценки Π^i, и оставили один текст y^i, подлежащий распознаванию и не используемый при оценке Π^i. Следуя схеме, описанной в предыдущем разделе, мы провели эксперименты для проверки качества оценок $t(F(\cdot))$, $t(G(\cdot))$, $e(F(\cdot))$, $e(G(\cdot))$ на этих 82 писателях. Для экономии места мы приведем лишь таблицу, отображающую информацию об эффективности оценки $t(G(\cdot))$. Эта таблица составлялась подобно таблице 1. Ради экономии места соответствующие таблицы L и R не приведены.

Таблица 2

N	Автор	c_1	c_2	c_3	c_4
0	К. Булычев	0	15	2007724	64741
1	О. Авраменко	0	6	1733113	223718
2	А. Больных	0	6	1294721	373611
3	А. Волков	0	8	1478932	202495
4	Г. Глазов	0	5	1398323	184593
5	М. и С. Дяченко	0	5	1754213	197039
6	А. Етоев	0	5	267096	80358
7	А. Кабаков	0	4	905502	222278
8	В. Каплан	0	6	515029	129608
9	С. Казменко	3	4	1846161	156768
10	В. Климов	0	3	250231	179903
11	И. Крашевский	0	2	1183722	481795
12	И. Кублицкая	0	1	282377	170469
13	Л. Кудрявцев	1	3	583239	179093
14	А. Курков	0	6	628041	218726
15	Ю. Латынина	10	2	2628781	283565
16	А. Лазаревич	46	3	310553	94629
17	А. Лазарчук	0	5	2395669	210151
18	С. Лем	0	7	1568013	343519
19	Н. Леонов	0	2	568854	279377
20	С. Логинов	14	13	1998543	159247
21	Е. Лукин	0	4	602216	125694
22	В. Черняк	0	2	920056	201636
23	А.П. Чехов	0	2	662801	343694
24	И. Хмелевская	0	4	1524905	203684

25	Л. и Е. Лукины	0	3	837198	122999
26	С. Лукьяненко	0	14	3682298	483503
27	Н. Маркина	0	1	266297	93647
28	М. Наумова	0	3	306514	337821
29	С. Павлов	0	2	751836	453448
30	Б. Райнов	0	4	1405994	420256
31	Н. Рерих	0	3	1011285	211047
32	Н. Романецкий	2	6	1305096	117147
33	А. Ромашов	0	1	88434	87744
34	В. Рыбаков	0	6	715406	121497
35	К. Серафимов	0	1	186424	75276
36	И. Сергиевская	0	1	109118	50786
37	С. Щеглов	10	2	253732	55188
38	А. Щеголев	0	2	848730	105577
39	В. Шинкарев	29	2	156667	80405
40	К. Ситников	0	7	419872	109116
41	С. Снегов	0	2	824423	408984
42	А. Степанов	0	5	1223980	93707
43	А. Столяров	11	1	350053	137135
44	Р. Светлов	0	2	454638	268472
45	А. Свиридов	63	3	660413	235439
46	Е. Тильман	0	2	705352	464685
47	Д. Трускиновская	0	8	2005238	118351
48	А. Тюрин	0	18	4109050	110237
49	В. Югов	0	5	829209	66657
50	А. Молчанов	0	1	398487	206541
51	Ф.М. Достоевский	1	3	613825	88582
52	Н.В. Гоголь	0	3	638339	215540
53	Д. Хармс	0	2	199449	114889
54	А. Житинский	0	2	2137325	543037
55	Е. Хаецкая	2	2	723167	204091
56	В. Хлумов	0	3	788562	183358
57	В. Кунин	0	3	1335918	296463
58	А. Мелихов	0	1	615548	458086
59	В. Набоков	0	5	1522633	342774
60	Ю. Никитин	0	2	1342176	702383
61	В. Сегаль	0	2	320218	75917
62	В. Ян	0	1	507502	600636

63	А. Толстой	0	1	129664	97842
64	И. Ефремов	0	1	536604	256521
65	Е. Федоров	0	1	1120665	221388
66	О. Гриневский	0	1	158762	96085
67	Н. Гумилев	0	1	70181	71042
68	Л.Н. Толстой	0	1	1225242	199903
69	В. Михайлов	0	1	254464	84135
70	Ю. Нестеренко	0	1	352988	71075
71	А.С. Пушкин	0	1	170380	57143
72	Л. Резник	0	1	115925	79628
73	М.Е. Салтыков-Щедрин	0	1	239289	101845
74	В. Шукшин	0	1	309524	66756
75	С. М. Соловьев	0	1	2345807	160002
76	А. Кац	0	1	841898	81830
77	Е. Козловский	1	1	849038	889560
78	С. Есенин	0	1	219208	44855
79	А. Стругацкий	0	1	151246	51930
80	А. и Б. Стругацкие	0	29	6571689	345582
81	Б. Стругацкий	0	1	298832	261206

Первый вывод из данных этой таблицы состоит в том, что количество правильных ответов (нулей в колонке c_1) очень велико — 69. Истинный автор произведения оказывается на втором месте в списке претендентов всего в трех случаях (в колонке c_1 стоит 1): Л. Кудрявцев, Ф.М. Достоевский и Е. Козловский. На третьем месте ($c_1 = 2$) — в двух случаях: Н. Романецкий и Е. Хаецкая. На четвертом месте оказывается лишь один автор ($c_1 = 3$) — С. Казменко. Для остальных 7 авторов ошибка очень велика (Ю. Латынина, А. Лазаревич, С. Логинов, С. Щеглов, В. Шинкарев, А. Столяров, А. Свиридов). Они не оказываются даже в десятке претендентов на их собственные произведения.

Мерой неточности оценки $t(G(\cdot))$ может служить средний ранг, равный сумме чисел в колонке c_1, деленной на общее число писателей 82. Здесь средний ранг равен

$$2.35 \approx (3\times1+2\times2+1\times3+2\times10+1\times11+1\times14+1\times29+1\times46+1\times63)/82$$

Все эти числа приведены в таблице 3 в колонке $t(G(\cdot))$. Если выбросить семерых плохо определяемых авторов, средний ранг окажется равным $0.13 \approx 2/15 = (3 \times 1 + 2 \times 2 + 1 \times 3)/75$.

Второй вывод из данных таблицы 2 состоит в том, что метод работает и на стихотворных произведениях (А.С. Пушкина, С. Есенина и Н. Гумилева). В-третьих, правильно определяются писатели, чьи произведения переводились с польского языка (С. Лем и И. Хмелевская). В-четвертых, среди плохо распознаваемых авторов нет общепризнанных классиков русской литературы.

Для сравнения, в следующей таблице приведены результаты аналогичного исследования с оценками $t(F(x))$, $e(F(x))$, $e(G(x))$ на тех же текстах.

Таблица 3

c_1	$t(F(\cdot))$	$t(G(\cdot))$	$e(F(\cdot))$	$e(G(\cdot))$
0	57	69	1	2
1	4	3	8	8
2	4	2	7	13
3	4	1	2	2
4	0	0	3	7
≥ 5	13	7	61	50
Среднее	3.50	2.35	13.95	12.37

Сделаем два вывода на основании данных последней таблицы. Во-первых, частотный анализ (метод, основанный на схеме Бернулли) работает плохо (имеется максимум два точных ответа). Однако, он все-таки дает какую-то информацию об авторе, ибо в случае совершенно случайного выбора истинного автора средний результат в последних двух столбцах был бы около 40. Во-вторых, выбрасывание слов, начинающихся с заглавной буквы, заметно улучшает результаты (даже при частотном анализе). Действительно, столбцы с функцией $G(\cdot)$ заметно лучше столбцов с функцией $F(\cdot)$.

5 Заключение

Из данных таблицы 3 хорошо видно, что оценка (2.1), основанная на анализе числа употреблений диад (двухбуквенных сочетаний), значительно эффективней оценки (2.2), основанной на частотном анализе одиночных букв, и правильно указывает автора с большой долей уверенности (84% против 3%). Можно было бы ожидать превосходство оценки (2.1), поскольку она использует больше информации об исходном тексте. Следует подчеркнуть удивительную точность (2.1) при распознавании истинного автора произведения (например, метод авторского инварианта [5] принципиально не может различить более 10 писателей, а здесь рассмотрено свыше 80). Такая точность несомненно должна привлечь внимание к изложенному методу.

Отмечается существенное улучшение качества распознавания автора текста при выбрасывании слов, начинающихся с заглавной буквы. Этот феномен еще требует своего объяснения.

Как уже говорилось, А.А. Марков весьма интересовался задачей определения авторства текста (об этом свидетельствует его статья [2]). Знаменательно, что его собственная идея о "связи испытаний в цепь", опробованная им же на литературном материале [3], привносит прогресс в решение этой задачи.

Автор благодарен М.И. Гринчуку за плодотворные дисскуссии по этой тематике. Автор также признателен А.Т. Фоменко и Г.В. Носовскому за оживленное внимание к работе и обсуждения результатов. Кроме них, автор благодарит А.А. Поликарпова, совместные дискуссии с которым существенно повлияли на окончательное оформление работы.

Список литературы

[1] Морозов Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд. // Известия отд. русского языка и словестности Имп.Акад.наук, Т.XX, кн.4, 1915.

[2] Марков А.А. Об одном применении статистического метода. // Известия Имп.Акад.наук, серия VI, Т.X, N4, 1916, с.239.

[3] Марков А.А. Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь. // Известия Имп.Акад.наук, серия VI, Т.X, N3, 1913, с.153.

[4] От Нестора до Фонвизина. Новые методы определения авторства. М.: Издат. группа ?Прогресс?, 1994.

[5] Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Предисловие А.Т. Фоменко. // Фоменко А.Т. Новая хронология Греции: Античность в средневековье. Т. 2. М.: Изд-во МГУ, 1996, с.768-820.

[6] Ивченко Г.И., Медведев Ю.И. Математическая статистика. 2-е изд. М.: Высшая школа, 1992.