DOI: https://doi.org/10.15688/jvolsu2.2019.2.13

UDC 81'42
LBC 81.055.1

Submitted: 05.02.2019
Accepted: 23.05.2019

NATURAL TEXT: MATHEMATICAL METHODS OF ATTRIBUTION [1]

Vladimir V. Popov
Volgograd State University, Volgograd, Russia

Tatyana V. Shtelmakh
Volgograd State University, Volgograd, Russia

Abstract. The article proposes two algorithms for substandard texts filtering. The first of these is based on the fact that the frequency of *n*-grams occurrence in a quality text obeys the Zipf law, and when the words of the text are rearranged, the law ceases to act. Comparison of the frequency characteristics of the source text with the characteristics of the text resulting from the permutation of words enables researchers to draw conclusions regarding the quality of the source text. The second algorithm is based on calculating and comparing the rate new words appear in good quality and randomly generated texts. In a good text, this rate is, as a rule, uneven whereas in randomly generated texts, this unevenness is smoothed out, which makes it possible to detect low-quality texts.

The methods for solving the problem of substandard texts filtering are statistical and are based on the calculation of various frequency characteristics of the text. As compared to the "bag of words" model, a graph model of the text, in which the vertices are words or word forms, and the edges are pairs of words, as well as models with higher order structures, in which the frequency characteristics of *n*-grams are used with $n > 2$, takes into account the mutual disposition of word pairs, as well as triples of words in a common part of the text, for example, in one sentence or one *n*-gram.

Key words: natural text, pseudo-text, text filtering, Zipf's law, *n*-grams, the rate of appearance of new words, "bag of words" model of the text, graph model of the text.

Citation. Popov V.V., Shtelmakh T.V. Natural Text: Mathematical Methods of Attribution. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2019, vol. 18, no. 2, pp. 147-158. (in Russian). DOI: https://doi.org/10.15688/jvolsu2.2019.2.13

УДК 81'42
ББК 81.055.1

Дата поступления статьи: 05.02.2019
Дата принятия статьи: 23.05.2019

ЕСТЕСТВЕННЫЙ ТЕКСТ: МАТЕМАТИЧЕСКИЕ МЕТОДЫ АТРИБУЦИИ [1]

Владимир Валентинович Попов
Волгоградский государственный университет, г. Волгоград, Россия

Татьяна Владимировна Штельмах
Волгоградский государственный университет, г. Волгоград, Россия

Аннотация. В статье предложено два алгоритма фильтрации некачественных текстов. Первый алгоритм основан на том, что частота появления *n*-грамм в качественном тексте подчиняется закону Зипфа, а в случай-

но генерированных текстах данный закон перестает действовать. Сравнение частотных характеристик двух типов текстов позволяет делать выводы относительно качества исходного текста. Второй алгоритм основан на сравнении скорости появления новых слов в текстах. В качественном тексте эта скорость, как правило, неравномерна, а в случайных текстах неравномерности нивелируются, что дает возможность обнаруживать некачественные тексты.

Основные методы решения задачи фильтрации некачественных текстов – статистические. Они базируются на вычислении различных частотных характеристик текста. В отличие от модели «мешка слов», не учитывающей порядок следования слов в тексте, графовая модель текста (в ней вершинами являются слова или словоформы, а ребрами – пары слов), а также модели со структурами более высокого порядка, в которых используются частотные характеристики n-грамм при $n > 2$, позволяют учитывать взаимное расположение пар и троек слов в какой-либо общей части текста: в одном предложении или одной n-грамме.

Ключевые слова: естественный текст, псевдотекст, фильтрация текстов, закон Зипфа, n-граммы, скорость появления новых слов, «мешок слов», графовая модель текста.

Цитирование. Попов В. В., Штельмах Т. В. Естественный текст: математические методы атрибуции // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2019. – Т. 18, № 2. – С. 147–158. – DOI: https://doi.org/10.15688/jvolsu2.2019.2.13

Введение

Интенсивное развитие информационных технологий способствует качественному изменению процессов порождения текстов, что выводит на первый план вопрос о разграничении естественных текстов и текстов, созданных при помощи различных программных средств. В связи с большим количеством текстового материала, функционирующего в пространстве электронной коммуникации, возникает задача разработки способов автоматической идентификации «случайных» текстов, которая должна осуществляться с опорой на количественный анализ текстовых характеристик.

В настоящее время применение количественных методов в языкознании аналогично их использованию в других естественных и социальных науках. Поскольку количественные показатели несут определенную информацию о самих текстах, одной из важнейших прикладных лингвистических задач становится моделирование топика коллекции текстов для систематизации документов и их использования в образовательных, маркетинговых и иных целях [Bakalov et al., 2012; Wallach, 2006; Yao, Mimno, McCallum, 2009; Zeng et al., 2012].

Квантитативные методы широко применяются для описания и классификации текстов, например, при установлении авторства анонимных текстов. Это связано с тем, что большинство двусторонних единиц и конструкций языка могут служить различению текстов. Доказано, что части речи играют немаловажную роль в формировании функциональных стилей языка. Статистические различия между ними, а также между жанрами составляют основу корпусной стилистики [McIntyre, Walker, 2019].

Количественные методы используются для идентификации индивидуального стиля автора. Учитывая важность данных статистического материала, еще Б.Н. Головин ставил ряд важных вопросов, требующих творческого решения: 1. Связаны ли показываемые статистикой особенности функционирования частей речи, предложений и их членов в речи писателей некоторыми внутренними зависимостями, то есть носят ли они системный характер? 2. Стоят ли за различиями активности частей речи, членов предложений и предложений у разных писателей устойчивые различия художественного содержания их произведений? 3. Следует ли думать, что в необследованных фрагментах текста (не вошедших в выборки) активность изучаемых элементов будет такой же, как и в выборках? [Головин, 1970, с. 9]. Как отмечал исследователь, активность частей речи в произведениях проявляется регулярно, следовательно, в разных местах разных произведений она закономерно характеризует стиль того или иного автора.

Итак, статистика предоставляет большие возможности для систематического изучения языкового функционирования и развития.

Ключевым моментом, который объединяет все квантитативные методики анализа текста, является то, что в их основе лежат

представления о некоторой единице анализа. Ее определение крайне важно, поскольку она выступает своего рода аналогом исследуемых (но неконтролируемых) переменных в эксперименте. Под единицей, вслед за Л.С. Выготским, мы подразумеваем «такой продукт анализа, который в отличие от элементов обладает всеми основными свойствами, присущими целому, и который является далее не разложимыми живыми частями этого единства» [Выготский, 1999, с. 14].

Однако их трактовка при конкретном методе анализа текста может быть принципиально различной. Так, под единицей лингвистического анализа текста понимаются инварианты различных лингвистических моделей описания языка, соотносящиеся с языком или языковым стандартом (морфема, фонема, предложение, словосочетание, высказывание и др., понимание которых в разных лингвистических направлениях может различаться). Единицами анализа выступают коллокации, которые в отечественной лингвистике понимаются как несвободные сочетания, не относящиеся к идиомам: с одной стороны, ключевое слово этих сочетаний может появляться в контексте с разными языковыми единицами, с другой стороны, эти единицы (то есть контекст ключевого слова) можно перечислить в виде закрытого («полузакрытого») списка (ср., напр., работы Л.Н. Иорданской, И.А. Мельчука и их последователей по изучению лексических функций и моделей управления [Иорданская, Мельчук, 2007]). В последние годы применение компьютерных программ обработки большого текстового материала позволяет использовать другие сущности в качестве единиц текстового анализа, к которым относятся, например, текстовые n-граммы. Они необходимы для решения разнообразных прикладных лингвистических задач, в частности автоматической категоризации и классификации текстов [Cavnar, Trenkle, 2001].

Целью нашей работы является поиск количественных критериев, позволяющих автоматически распознать естественный текст и отличить данный текст от созданного при помощи определенных алгоритмов порождения текста. Единицей анализа избрана *n*-грамма.

Прежде чем приступить к анализу языкового материала, обратимся к дефиниции понятия «текст». Вслед за И.Р. Гальпериным мы считаем, что текст – это «произведение речетворческого процесса, обладающее завершенностью, объективированное в виде письменного документа, литературно обработанное в соответствии с типом этого документа, произведение, состоящее из названия (заголовка) и ряда особых единиц (сверхфразовых единств), объединенных разными типами лексической, грамматической, логической, стилистической связи, имеющее определенную целенаправленность и прагматическую установку» [Гальперин, 2006, с. 18]. В качестве объекта исследования мы рассматриваем текст в целом, а не его составляющие, разделяя мнение А.А. Леонтьева о том, что сложные закономерности формирования целостности и завершенности текста участвуют в формировании его смыслового аспекта: «В противоположность связности цельность есть характеристика текста как смыслового единства, как единой структуры, и определяется на всем тексте» [Леонтьев, 1979, с. 12]. Кроме того, мы изучаем тексты без ограничения на объем, обладающие характеристиками отдельнооформленности, связности и цельности (о признаках текста см.: [Мурзин, Штерн, 1991]). Такие тексты в данной работе мы будем называть *естественными*.

Если на естественном языке сформирована некоторая последовательность слов, то возникает вопрос о реализации указанных выше фундаментальных свойств текста. Назовем *псевдотекстом* любую последовательность слов естественного языка, полученную на основе некоторой вероятностной модели сочетания слов в тексте. При этом считаем, что псевдотекст не получен многократным дублированием одной из своих частей. В качестве лексической единицы текста будем рассматривать слово в тексте (словоформу), а список слов будет образовывать *словарь* текста. Псевдотексты получим как результат случайной перестановки слов естественного текста, тогда распределение вероятностей слов в псевдотексте останется неизменным. Статистические модели текста в связи с необходимостью его автоматической обработки, создания машинных алгоритмов вводятся

и обосновываются, в частности, в [Пиотровский, 1975].

В части I данной статьи текст рассмотрен в статическом состоянии (оно, по мнению Л.А. Новикова, соответствует тексту как «результату речемыслительной деятельности» [Новиков, 1983]), в части II – в динамическом состоянии, то есть в процессе его порождения, при этом не затрагиваются вопросы его психологического восприятия и понимания.

Исследования естественных текстов и псевдотекстов проводим на основе графовой модели, описанной в [Григорьева и др., 2017]. Граф образуется как совокупность двух множеств: множество вершин – слова текста, множество ребер – пары слов, находящихся в одной *n*-грамме. Каждому ребру приписывается вес – количество *n*-грамм, содержащих эту пару слов.

I. Структура *n*-грамм текста

Пусть имеется некоторый текст D. Выберем целое число $n \geq 2$. Под *n-граммой* будем понимать последовательность из n подряд идущих слов одного текста (о таком толковании *n*-граммы см.: [Бузикашвили, Самойлов, Крылова, 2000]). Составим список всех *n*-грамм текста D и для каждой *n*-граммы подсчитаем ее частоту. Расположим *n*-граммы в порядке убывания их частот, полученный ряд значений назовем *вектором частот n*-грамм. Компьютерные эксперименты показывают, что *n*-граммы естественного (связного) текста подчиняются аналогу закона Зипфа, то есть имеются *n*-граммы с достаточно высокой частотой, а затем их частота быстро убывает. Отметим, что, как правило, частоты псевдотекста небольшие и убывают сравнительно медленно. Например, вектор частот 3-грамм естественного текста: (15, 14, 13 13, 12, 12, 10, 9, 9, 8, 8, 8, 7, 7, 7, 7, 7, 7, 6, 6...); псевдотекста: (4, 3, 3, 2, 2, 2, 2, 2,...).

Для текста D зафиксируем целые числа h и m, где $h, m \geq 2$. Пусть $D_1, D_2, ..., D_m$ – тексты, полученные на шаге 1, 2, ..., m перемешивания D. Обозначим через $Sum(D)$, $Sum(D_i)$, $1 \leq i \leq m$ суммы частот h первых n-грамм (то есть n-грамм с наибольшими частотами) соответствующих текстов. Обозначим через MS среднее арифметическое величин $Sum(D_1)$, $Sum(D_2)$, ..., $Sum(D_m)$:

$$MS = \frac{Sum(D_1) + Sum(D_2) + ... + Sum(D_m)}{m},$$

и положим:

$$\theta(D) = \frac{Sum(D)}{MS}.$$

Отметим, что величина $\theta(D)$ зависит не только от документа D и параметров n, h и m, но и от того, какие перестановки списка слов документа D применялись при формировании документов $D_1, D_2, ..., D_m$. Если эти перестановки выбираются случайно, то при повторном применении алгоритма будут получаться различные, хотя и достаточно близкие между собой значения. Так, при трех запусках компьютерной программы с параметрами $n = 3$, $h = 20$ и $m = 6$ могут быть получены значения величины θ, равные 1.775, 1.738 и 1.753 соответственно. Для реализации алгоритма исследований нам необходимо зафиксировать набор возможных параметров: вектор $\alpha = (n,i,j,k,m,h)$, влияющих на частоту n-грамм и величины $Sum(D)$, MS, и $\theta(D)$:

n – число слов в одной n-грамме;

i – следует ли рассматривать n-граммы лексем (при $i = 1$) или словоформ ($i = 0$);

j – надо ли сортировать слова в каждой n-грамме по алфавиту ($j = 1$) или не надо ($j = 0$);

k – минимальная длина учитываемых слов;

m – число формируемых псевдотекстов;

h – число учитываемых n-грамм с наибольшей частотой.

Исследования проводились на коллекции из 60 естественных текстов, число слов в которых насчитывается от 485 до 460326.

Анализ коллекции естественных текстов

Упорядоченные по алфавиту, то есть без учета порядка слов, наиболее частотные 3-граммы из одного текста получаем, например, такие:

– словоформные: *как после того, для того чтобы, такое что это, дело том что, знает черт что, все что это, было как слышно, несколько секунд через, кроме того что, том убедился что, время некоторое через, было видно что*;

– лексемные: *как после тот, для тот чтобы, дело тот что, тот убедиться что,*

такой что это, быть мочь это, весь что это, она тот что, быть как слышный, несколько секунда через, мысль тот что, кроме тот что.

Неупорядоченные по алфавиту словоформы этого же текста: *после того, для того чтобы, дело том что, черт знает что, слышно было как, через несколько секунд, что это такое, через некоторое время, кроме того что, через несколько минут, убедился том что, позвольте вас спросить*.

В результате расчетов θ сформированы таблицы определенного вида, которые продемонстрируем на таблицах 1 и 2 для одного из текстов коллекции при α = (3, 0, 1, 3, 12, 200).

В ходе исследования всей коллекции получены следующие свойства θ при α = (*n*,*i*,*j*,*k*,*m*,*h*), *m* = 3, *h* = 200:

1. Для *n* = 2 количество естественных текстов, близких по расчетным показателям θ к псевдотекстам, значительно выше, чем для *n* = 3 с такими же значениями остальных параметров в α. Поэтому далее рассматриваем *n* = 3.

2. Значения θ не превышали значения 2 для всех псевдотекстов (расчеты в таблицах типа 2) при различных вариантах *i*, *j* и *k* = 3 вектора параметров α, кроме α = (3, 1, 1, 3, 12, 200) (не выше 2.7).

3. Значение минимума θ не ниже значения 1 для всех естественных текстов (расчеты в таблицах типа 1) при различных вариантах *i*, *j* и *k* = 3 вектора параметров α.

4. Значение максимума θ не ниже значения 2 для всех естественных текстов (расчеты в таблицах типа 1) при α = (3, 0, 0, 3, 12, 200).

5. Длина диапазона значений θ (размах θ) при α = (3, 0, 1, 3, 12, 200) и α = (3, 0, 0, 3, 12, 200) для естественного текста во всех 60 случаях больше длины диапазона значений его же псевдотекста. При этом минимальное значение θ (при указанных параметрах) для большинства из рассмотренных текстов строго больше, чем максимальное значение этой же величины для соответствующего псевдотекста (кроме 5 из 60 для первого вектора параметров и 1 из 60 для второго вектора параметров).

6. При α = (3, 0, 0, 3, 12, 200) наименьший размах θ естественных текстов равен 0.927 и только 2 псевдотекста из 60 имеют размах θ больше этого минимума (1.059 и 0.996), что продемонстрировано на рисунке 1. У 56 из 60 естественных текстов размах θ больше 1.059 (от 1.24 до 55.797), и они могут идентифицироваться представленным методом как естественные. А значит, размах θ также может служить отличительным признаком псевдотекстов.

При рассмотрении лексемных *n*-грамм условия 4 и 5 нарушаются для 30 из 60 текстов коллекции, что продемонстрировано на рисунке 2.

Таблица 1

Значения величины θ(*D*) при различных *m* и *h* (*D* – естественный текст)

	m = 1	*m* = 2	*m* = 3	*m* = 4	*m* = 5	*m* = 6	*m* = 7	…	*m* = 12
h = 1	3.75	5.0	5.0	5.0	4.411	4.285	4.375	…	4.285
h = 2	4.142	5.272	5.437	5.272	4.833	4.702	4.720	…	4.578
h = 3	4.2	5.25	5.478	5.419	5.0	4.941	4.9	…	4.893
…	…	…	…	…	…	…	…	…	…
h = 100	2.558	2.584	2.588	2.587	2.573	2.573	2.571	…	2.573
…	…	…	…	…	…	…	…	…	…
h = 200	2.283	2.341	2.313	2.328	2.328	2.357	2.364	…	2.389

Таблица 2

Значения величины θ(*D*) при различных *m* и *h* (*D* – псевдотекст)

	m = 1	*m* = 2	*m* = 3	*m* = 4	*m* = 5	*m* = 6	*m* = 7	…	*m* = 12
h = 1	1.5	1.2	1.125	1.2	1.153	1.125	1.05	…	1.058
h = 2	1.5	1.2	1.125	1.2	1.153	1.125	1.076	…	1.107
h = 3	1.333	1.066	1.0	1.066	1.052	1.021	0.982	…	1.010
…	…	…	…	…	…	…	…	…	…
h = 100	1.01	1.0	0.996	1.0	1.0	0.997	0.996	…	0.998
…	…	…	…	…	…	…	…	…	…
h = 200	0.985	0.954	0.968	0.982	0.988	0.976	0.977	…	0.971

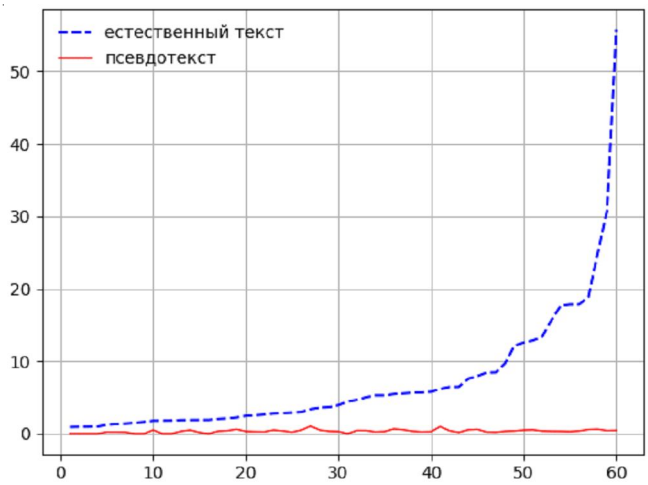

Рис. 1. Значения размаха θ, упорядоченные по его возрастанию, для 60 естественных текстов и псевдотекстов при α = (3, 0, 0, 3, 12, 200)

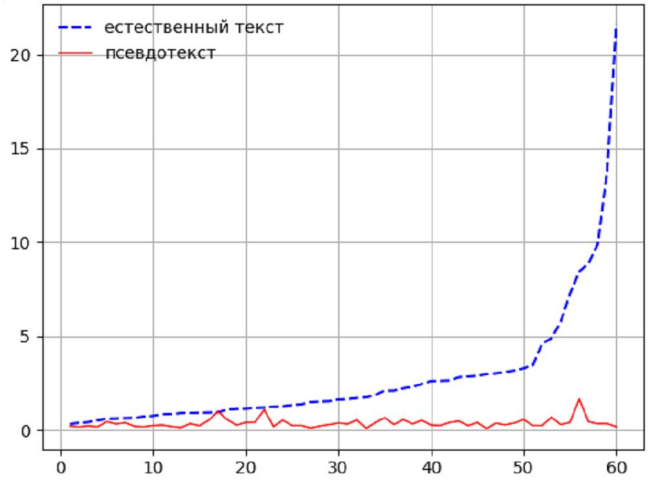

Рис. 2. Значения размаха θ, упорядоченные по его возрастанию, для 60 естественных текстов и псевдотекстов при α = (3, 1, 1, 3, 12, 200)

Не идентифицированные как естественные по условию 6 тексты (4 из 60, то есть точность 93 %) имеют max θ = 2, то есть не проходят так же условие 2. Такого вида тексты требуют дополнительных исследований. Отметим, что в рассматриваемой коллекции они имеют наименьший объем, не более 6000 слов. С ростом объема текста размах θ, как правило, увеличивается. Но этот факт не является однозначной закономерностью, что видно по рисунку 3.

Приведенные результаты позволяют предложить следующий алгоритм, основанный на вычислении величины $u(D)$, который позволяет разделять тексты на «хорошие» (близкие по свойствам к естественным текстам) и «подозрительные» (близкие по свойствам к псевдотекстам). Пусть D – некоторый текст, подлежащий исследованию.

Алгоритм 1

1. Полагаем $n = 3$, $i = 0$, $j = 0$, $k = 3$.

2. Для каждого m от 1 до 10 и каждого h от 1 до 100 вычисляем величину $\theta_{m,h} = \theta(D)$ для вектора параметров $\alpha = (m, i, j, k, m, h)$.

3. Если все величины $\theta_{m,h}$, найденные в пункте 2, удовлетворяют неравенствам: $\max \theta_{m,h} > 2$, то считаем, что D – «хороший» текст.

4. Если $\max \theta_{m,h} < 2$, то считаем, что это «подозрительный» текст.

5. Если $\theta_{m,h} = 2$, $\min \theta_{m,h} \geq 1$, то полагаем $n = 3$, $i = 0$, $j = 1$, $k = 3$ и выполняем пункт 2. Если $\max(\theta) < 2$, $\min(\theta) < 1$, то считаем, что D – «подозрительный» текст, иначе считаем, что требуется другой способ исследования.

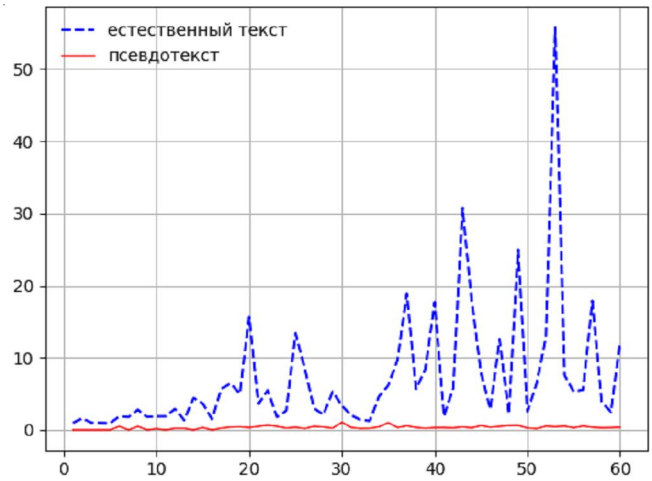

Рис. 3. Значения размаха θ, упорядоченные по возрастанию количества слов в тексте, для 60 естественных текстов и псевдотекстов при α = (3, 0, 0, 3, 12, 200)

Рассмотрим использование метода *n*-грамм на примере. На 28-м международном фестивале рекламы, проходившем 15–16 ноября 2018 года (ТВ3.ру, 2018), был представлен совместный проект канала ТВ-3, «Яндекс» и С. Лукьяненко, в котором использовали рассказ «Дурной договор», созданный нейросетью, обученной на массиве русской прозы и произведениях Н.В. Гоголя, по сюжету С. Лукьяненко. Результаты исследований этого текста:

– α = (3, 0, 0, 3, 12, 200). Исходный текст: вектор частот (2, 1, 1, 1, 1, 1 ...), max(θ) = 2.0, min(θ) = 1.004, размах θ равен 0.996; его псевдотекст: max(θ) = 1.0, min(θ) = 1.0, размах θ равен 0. По свойствам θ текст похож на один из четырех естественных текстов коллекции, не идентифицированных методом *n*-грамм;

– α = (3, 0, 1, 3, 12, 200). Исходный текст: вектор частот (2, 2, 2, 2, 2, 2, 2, 2, 2, 1, ...), max(θ) = 1.384 < 2, min(θ) = 0.8 < 1, размах θ равен 0.584; его псевдотекст: max(θ) = 1.147, min(θ) = 1.0, размах θ равен 0.147.

Таким образом, в связи с нарушением условий 3 и 4 и достаточно малым размахом θ рассказ «Дурной договор» можно идентифицировать как текст, близкий по свойствам к псевдотекстам.

II. Структура словарей фрагментов текста

Пусть имеется некоторый текст D. Составим список $L(D)$ всех его слов в том порядке, в котором они появляются в тексте. Обозначим через $L(D,t)$ начальный отрезок (срез) списка $L(D)$, состоящий из первых t слов, где t не превосходит длины $len(D)$ всего списка $L(D)$.

Приводя каждое из слов $w \in L(D,t)$ к лемме и учитывая каждую лемму один раз, получим словарь $V(D,t)$ начального отрезка $L(D,t)$ текста D. Обозначим через $f(D,t)$ длину этого словаря, то есть число слов в нем.

Если $t = len(D)$, то $V(D,t)$ будет словарем, который соответствует всему исходному тексту. Обозначим через $V(D)$ этот словарь, а через $len(V(D))$ – его длину, то есть число слов в нем.

В работе [Baker, 1988] введено понятие скорости (*pace*), с которой появляются новые слова в авторском тексте. Используя введенные только что обозначения и учитывая тот факт, что *pace* выражают в процентах, можно записать формулу для вычисления величины *pace*:

$$pace = \frac{len(V(D))}{len(D)} \cdot 100\ \%.$$

Компьютерные эксперименты показывают, что в естественном тексте скорость появления новых слов неравномерна. Например, эта скорость увеличивается при смене объекта, который описывается в тексте. Далее эта скорость постепенно уменьшается (см. рис. 4). При перемешивании слов текста скорость появления новых слов усред-

няется, поэтому естественные тексты можно обнаружить, если сравнивать описанные скорости. Проследить указанную закономерность для одного из текстов коллекции (рассматриваемый ниже в расчетах Текст1) можно на рисунке 4.

В данном разделе будет получена более детальная характеристика поведения функций $L(D,t)$, $V(D,t)$ и $f(D,t)$.

Пусть исходный текст D – естественный текст. Формируем псевдотексты D_1, D_2, ..., D_m, перемешивая слова D случайным образом m раз, $m \geq 2$. При каждом целом t, $1 \leq t \leq len(D)$, обозначим через $\overline{u} = \overline{u(t)}$ среднее арифметическое величин $u_i = f(D_i,t)$, $i = 1, 2, ..., m$, а через $s = s(t)$ их среднее квадратическое отклонение:

$$\overline{u} = \frac{u_1 + u_2 + ... + u_m}{m}$$

$$s = \sqrt{\frac{1}{m}((u_1 - \overline{u})^2 + (u_2 - \overline{u})^2 + ... + (u_m - \overline{u})^2)}.$$

Обозначим через $Gt2$ количество тех значений t, $1 \leq t \leq len(D)$, для которых выполнено двойное неравенство

$$\overline{u} - 2 \cdot s < f(D,t) < \overline{u} + 2 \cdot s,$$

а через $Lt2$ – количество тех значений t, $1 \leq t \leq len(D)$, для которых оно не выполнено. Если $s = 0$ при некотором t, то это t не относим ни к $Gt2$, ни к $Lt2$.

Приведем результаты компьютерных экспериментов (проведенных при $n = 3$ и $m = 5$) для трех текстов коллекции и текста нейросети:

Текст	*Gt*2	*Lt*2	*Gt*2/*Lt*2
Текст1	19162	6057	3.1636
Текст2	76996	27391	2.8109
Текст3	95194	21717	4.3833
Текст1, псевдотекст	10377	14848	0.6988
Текст1, (первые 15000 слов)	927	2334	0.3971
«Дурной договор»	249	2125	0.1171

Проведенные исследования позволяют сделать вывод, что если D – естественный текст, то $Gt2$, как правило, превосходит $Lt2$, поэтому отношение $Gt2/Lt2$ существенно больше 1. Если рассматривать текст D в его исходном виде и $L(D)$ – список словоформ, то получаем следующие результаты (см. рис. 5):

1. У 11 из 60 текстов коллекции (18 %) $Gt2 / Lt2 < 1$ и у 22 из 60 текстов коллекции (37 %) $Gt2 / Lt2 < 2.5$.

2. Для 5 псевдотекстов из 60 выполнено $Gt2 / Lt2 > 1$, для одного $Gt2 / Lt2 > 2$. Для всех псевдотекстов выполнено $Gt2 / Lt2 < 2.5$.

3. Для двух из четырех естественных текстов, указанных в I разделе статьи, как близкие по свойствам к псевдотекстам, выполнено соотношение $1 < Gt2 / Lt2 < 2$, для остальных двух $Gt2 / Lt2 < 1$, и все четыре текста остаются в группе «подозрительных» текстов.

Если преобразовать текст D и рассматривать $L(D)$ как список лексем, то получаем следующие результаты (см. рис. 6):

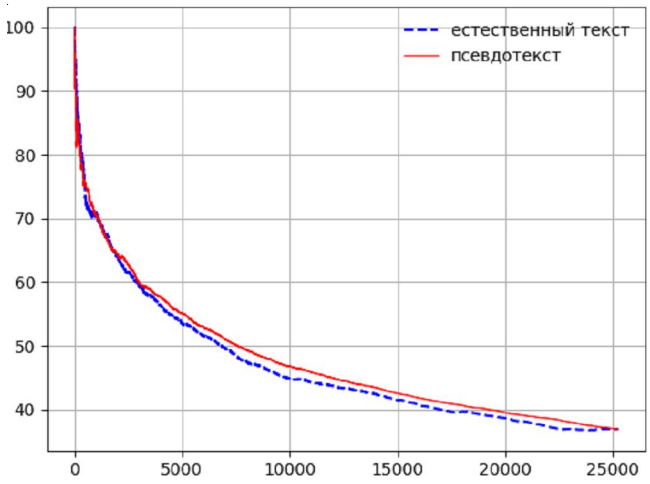

Рис. 4. Изменение скорости появления новых слов (*pace*) в естественном тексте и псевдотексте

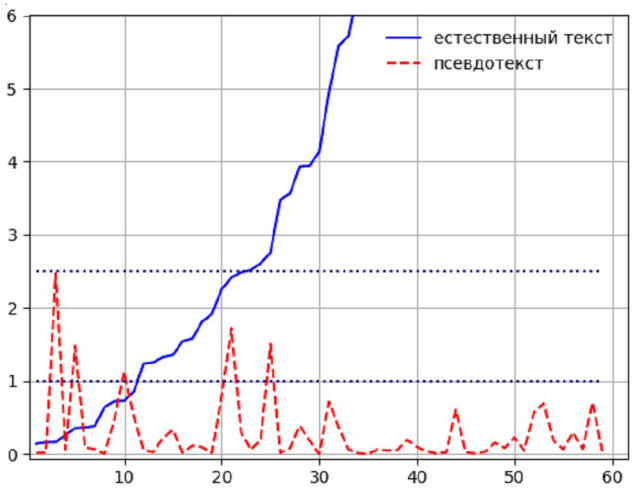

Рис. 5. Значения величины *Gt*2/*Lt*2 для 60 естественных текстов (часть графика в окрестности значений 2.5 и 1, значения упорядочены по возрастанию) и псевдотекстов в виде списка словоформ

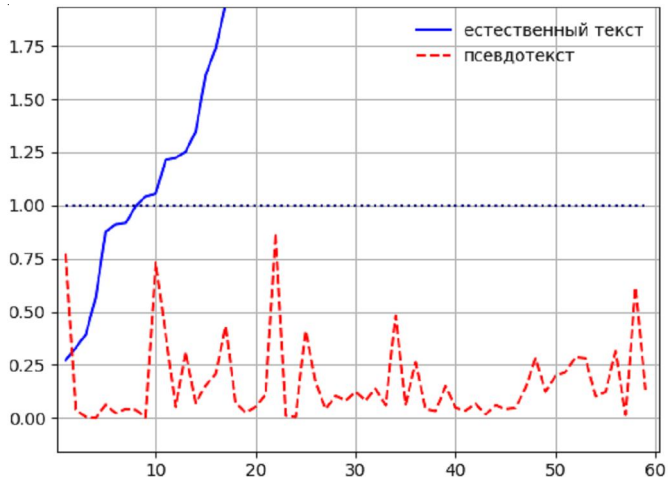

Рис. 6. Значения величины *Gt*2 / *Lt*2 для 60 естественных текстов (часть графика в окрестности значения 1, значения упорядочены по возрастанию) и псевдотекстов в виде списка лексем

4. Для 8 из 60 текстов коллекции (14 %) выполнено *Gt*2 / *Lt*2 < 1.

5. Для всех псевдотекстов выполнено *Gt*2 / *Lt*2 < 1.

6. Для двух из четырех естественных текстов, указанных в I разделе статьи как близкие по свойствам к псевдотекстам, выполнено соотношение *Gt*2 / *Lt*2 > 2, и эти тексты можно отнести к группе «хороших».

Опишем алгоритм (точность 86 %), основанный на вычислении величин *Gt*2 и *Lt*2, который позволяет разделять тексты на «хорошие» и «подозрительные».

Пусть задан некоторый текст *D*.

Алгоритм 2

1. Полагаем $n = 3$ и $m = 10$.

2. Приводим каждое слово текста *D* к лемме.

3. Вычисляем для текста *D* величины *Gt*2 и *Lt*2.

4. Если выполнено неравенство *Gt*2 / *Lt*2 > 1, то считаем, что *D* – «хороший» текст. В противном случае считаем, что это «подозрительный» текст.

В результате применения алгоритмов 1 и 2 из 60 исследуемых естественных текстов в группе «подозрительных» остались два текста (4 %).

Выводы

В работе предложены два алгоритма, которые позволяют разделять тексты на качественные и «случайные». Первый их них

основан на том, что для качественных текстов закон Зипфа выполняется как для отдельных слов, так и для *n*-грамм, а в случайных текстах закон Зипфа продолжает действовать для отдельных слов, но перестает действовать для *n*-грамм.

Второй алгоритм основан на подсчете скорости появления новых слов. В «хорошем» тексте эта скорость неравномерна. В случайных текстах она усредняется, поэтому оригинальные тексты можно обнаружить, если сравнивать описанные скорости.

Возможности применения обоих алгоритмов зависят от ряда параметров, поэтому в дальнейшем предполагается проведение более объемных компьютерных экспериментов для выбора оптимальных параметров.

ПРИМЕЧАНИЕ

[1] Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований и Администрации Волгоградской области, проект № 18-412-340007 «Лингвистические исследования проблем автоматизированной обработки корпусов текстов для региональных организационно-управленческих целей и разработка соответствующего программного комплекса».

The study was supported by the Russian Foundation for Basic Research and the Volgograd Region Administration, project no. 18-412-340007 “Linguistic Foundation for Information Retrieval from Corpus and Computer Software Development Relevant to Document-Government”.

СПИСОК ЛИТЕРАТУРЫ

Бузикашвили Н. Е., Самойлов Д. В., Крылова Г. А., 2000. N-граммы в лингвистике // Методы и средства работы с документами : сб. ст. М. : Едиториал УРРС. С. 91–130.

Выготский Л. С., 1999. Мышление и речь. Изд. 5-е, испр. М. : Лабиринт. 352 с.

Гальперин И. Р., 2006. Текст как объект лингвистического исследования. 4-е изд. стер. М. : КомКнига. 144 с.

Головин Б. Н., 1970. Язык и статистика. М. : Просвещение. 190 с.

Григорьева Е. Г., Клячин В. А., Помельников Ю. В., Попов В. В., 2017. Алгоритм выделения ключевых слов на основе графовой модели лингвистического корпуса // Вестник Волгоградского государственного университета. Серия 2, Языкознание. Т. 16, № 2. С. 58–67. DOI: https://doi.org/10.15688/jvolsu2.2017.2.6.

Иорданская Л. Н., Мельчук И. А., 2007. Смысл и сочетаемость в словаре. М. : Языки славянских культур. 672 с.

Леонтьев А. А., 1979. Понятие текста в современной лингвистике и психолингвистике // Психолингвистическая и лингвистическая природа текста и особенности его восприятия / под ред. Ю. А. Жлуктенко, А. А. Леонтьева. Киев : Вища школа. С. 7–17.

Мурзин Л. Н., Штерн А. С., 1991. Текст и его восприятие. Свердловск : Изд-во УГУ. 172 с.

Новиков А. И., 1983. Семантика текста и ее формализация. М. : Наука. 215 с.

Пиотровский Р. Г., 1975. Текст, машина, человек. Л. : Наука. 327 с.

Bakalov A., McCallum A., Wallach H., Mimno D., 2012. Topic models for taxonomies // Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries (Washington, DC, USA, June 10–14, 2012). P. 237–240. DOI: https://doi.org/10.1145/2232817.2232861.

Baker J. C., 1988. Pace: A Test of Authorship Based on the Rate at which New Words Enter an Author’s Text // Literary and Linguistic Computing. Vol. 3, no. 1. P. 36–39.

Cavnar W., Trenkle J., 2001. N-Gram-Based Text Categorization. URL: https://www.researchgate.net/publication/2375544_N-Gram-Based_Text_Categorization.

McIntyre D., Walker B., 2019. Corpus Stylistics: Theory and Practice. Edinburgh University Press. 376 p.

Wallach H. M., 2006. Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning (Pittsburgh, Pennsylvania, USA, June 25–29, 2006). P. 977–984. DOI: https://doi.org/10.1145/1143844.1143967.

Yao L., Mimno D., McCallum A., 2009. Efficient methods for topic model inference on streaming document collections // Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (Paris, France, June 28 – July 01, 2009). P. 937–946. DOI: https://doi.org/10.1145/1557019.1557121.

Zeng Q. T., Redd D., Rindflesch T. C., Nebeker J. R., 2012. Synonym, topic model and predicate-based query expansion for retrieving clinical documents // AMIA. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540443.

ИСТОЧНИК

ТВ3.ру, 2018 – «Нейро-Гоголь» вошел в шорт-лист Red Apple // ТВ3.ру. URL: https://tv3.ru/post/luchshaya-innovatsiya-v-reklame (дата обращения: 25.01.2019).

REFERENCES

Buzikashvili N.E., Samoylov D.V., Krylova G.A., 2000. N-grammy v lingvistike [N-Grams in Linguistics]. *Metody i sredstva raboty s dokumentami* [Methods and Means of Working with Documents]. Moscow, Editorial URRS Publ., pp. 91-130.

Vygotskiy L.S., 1999. *Myshlenie i rech* [Thinking and Speech]. Moscow, Labirint Publ. 352 p.

Galperin I.R., 2006. Tekst kak obyekt lingvisticheskogo issledovaniya [Text as an Object of Linguistic Research]. Moscow, KomKniga Publ. 144 p.

Golovin B.N., 1970. Yazyk i statistika [Language and Statistics]. Moscow, Prosveshchenie Publ. 190 p.

Grigoryeva E.G., Klyachin V.A., Pomelnikov Yu.V., Popov V.V., 2017. Algoritm vydeleniya klyuchevykh slov na osnove grafovoy modeli lingvisticheskogo korpusa [Algorithm of Key Words Search Based on Graph Model of Linguistic Corpus]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2, Yazykoznanie* [Science Journal of VolSU. Linguistics], vol. 16, no. 2, pp. 58-67. DOI: https://doi.org/10.15688/jvolsu2.2017.2.6.

Iordanskaya L.N., Melchuk I.A., 2007. *Smysl i sochetaemost v slovare* [Meaning and Compatibility in the Dictionary]. Moscow, Yazyki slavyanskikh kultur Publ. 672 p.

Leontev A.A., 1979. Ponyatie teksta v sovremennoy lingvistike i psikholingvistike [The Concept of Text in Modern Linguistics and Psycholinguistics]. *Psikholingvisticheskaya i lingvisticheskaya priroda teksta i osobennosti ego vospriyatiya* [Psycholinguistic and Linguistic Nature of Text and Features of Its Perception]. Kiev, Vishcha shkola Publ., pp. 7-17.

Murzin L.N., Shtern A.S., 1991. *Tekst i ego vospriyatie* [The Text and Its Perception]. Sverdlovsk, Izd-vo UGU. 172 p.

Novikov A.I., 1983. *Semantika teksta i ee formalizatsiya* [Text Semantics and Its Formalization]. Moscow, Nauka Publ. 215 p.

Piotrovskiy R.G., 1975. *Tekst, mashina, chelovek* [Text, Machine, Person]. Leningrad, Nauka Publ. 327 p.

Bakalov A, McCallum A, Wallach H, Mimno D., 2012. Topic Models for Taxonomies. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 237-240.

Baker J.C., 1988. Pace: A Test of Authorship Based on the Rate at Which New Words Enter an Author's Text. *Literary and Linguistic Computing*, vol 3, no. 1, pp. 36-39.

Cavnar W., Trenkle J., 2001. *N-Gram-Based Text Categorization*. URL: https://www.researchgate.net/publication/2375544_N-Gram-Based_Text_Categorization.

Mclntyre D., Walker B. 2019. *Corpus Stylistics: Theory and Practice*. Edinburgh University Press. 376 p.

Wallach H.M. 2006. Topic Modeling: Beyond Bag-of-Words. *Proceedings of the 23rd International Conference on Machine Learning (Pittsburgh, Pennsylvania, USA, June 25–29, 2006)*, pp. 977-984.

Yao L., Mimno D., Mccallum A., 2009. Efficient Methods for Topic Model Inference on Streaming Document Collections. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Paris, France, June 28 – July 1, 2009)*, pp. 937-946. DOI: https://doi.org/10.1145/1557019.1557121.

Zeng Q.T., Redd D., Rindflesch T.C., Nebeker J.R., 2012. Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents. *AMIA*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540443.

SOURCE

«Neyro-Gogol» voshel v short-list Red Apple ["Neuro-Gogol" Was Included in Red Apple Shortlist], 2018. *TV3.ru*. URL: https://tv3.ru/post/luchshaya-innovatsiya-v-reklame (accessed 25 January 2019).

Information about the Authors

Vladimir V. Popov, Candidate of Sciences (Physics and Mathematics), Associate Professor, Department of Computer Science and Experimental Mathematics, Volgograd State University, Prosp. Universitetsky, 100, 400062 Volgograd, Russia, popov.vlaval@volsu.ru, kiem@volsu.ru, https://orcid.org/0000-0003-0419-2874

Tatyana V. Shtelmakh, Senior Lecturer, Department of Computer Science and Experimental Mathematics, Volgograd State University, Prosp. Universitetsky, 100, 400062 Volgograd, Russia, shtelmakh_tv@mail.ru, https://orcid.org/0000-0002-5320-7406

Информация об авторах

Владимир Валентинович Попов, кандидат физико-математических наук, доцент кафедры компьютерных наук и экспериментальной математики, Волгоградский государственный университет, просп. Университетский, 100, 400062 г. Волгоград, Россия, popov_v_v@rambler.ru, https://orcid.org/0000-0003-0419-2874

Татьяна Владимировна Штельмах, старший преподаватель кафедры компьютерных наук и экспериментальной математики, Волгоградский государственный университет, просп. Университетский, 100, 400062 г. Волгоград, Россия, shtelmakh_tv@mail.ru, https://orcid.org/0000-0002-5320-7406