ISSN 1814-1196
Научный вестник НГТУ
том 78, № 1, 2020, с. 75–86

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 78, No. 1, 2020, pp. 75–86

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND CONTROL

УДК 519.246.8

DOI: 10.17212/1814-1196-2020-1-75-86

Компьютерный анализ школьных изложений на основании закона Хипса*

Н.С. ЗАКРЕВСКАЯ[1,a], М.А. КОВАЛЕВСКАЯ[2,b], М.Г. ЧЕБУНИН[3,c]

[1] *630073, РФ, г. Новосибирск, пр. К. Маркса, 20, Новосибирский государственный технический университет*
[2] *630090, РФ, г. Новосибирск, ул. Пирогова, 1, Новосибирский государственный университет*
[3] *630090, РФ, г. Новосибирск, пр. Академика Коптюга, 4, Институт математики им. С.Л. Соболева; 630090, РФ, г. Новосибирск, ул. Пирогова, 1, Новосибирский государственный университет*

[a] *natali.erlagol@gmail.com* [b] *maryartkey@gmail.com* [c] *chebuninmikhail@gmail.com*

Идея изложения художественного текста школьниками состоит не в копировании авторского текста, а в пересказе основного содержания другими словами. Компьютерный анализ текста изложения должен учитывать, что и исходный авторский текст, и его изложение школьником подчиняются закону Ципфа распределения частот слов, но с разными словарями и с разными параметрами. Различие словарей и параметров закона Ципфа у автора и учащихся обусловлено различием социальной среды, круга чтения и общения, способности запоминать и использовать слова разных языковых пластов. Математический аппарат для анализа соответствия текста закону Ципфа разработан в ряде теоретических работ по теории вероятностей и математической статистике. Методами теории случайных процессов описывается поведение последовательности количеств разных слов текста. Отметим, что степенной закон роста числа разных слов с увеличением длины текста в прикладной статистике называется законом Хипса. Теоретические результаты на основании элементарной вероятностной модели позволяют анализировать значимость отклонений от закона Хипса.

В настоящей работе разработан алгоритм анализа изложений. Он включает сравнение текста изложения с исходным авторским текстом, выявляет повторяющиеся слова и выражения. Отдельно анализируется появление стандартных триграмм (троек слов) с помощью библиотеки наиболее частых триграмм, составленной по библиотеке русской классики. Затем анализируется однородность текста на основании соответствия закону Хипса. В заключение изучается однородность комбинированных текстов, составленных из исходного текста и текста изложения.

Разработанный подход дает новую информацию о тексте изложения и его соответствии исходному тексту. Он может применяться как для компьютерного анализа изложений, так и для проверки однородности текста.

* *Статья получена 16 декабря 2019 г.*

Исследование выполнено при частичной финансовой поддержке РФФИ и правительства Новосибирской области (грант №19-41-543004) и поддержке программы фундаментальных научных исследований СО РАН № I.1.3. (проект № 0314-2019-0008).

Ключевые слова: закон Ципфа, закон Хипса, слабая сходимость, гауссовский процесс, триграмма, эмпирический мост, статистика омега-квадрат, статистический критерий, реально достигаемый уровень значимости, однородность текста

ВВЕДЕНИЕ

Автоматизация проникла во многие области анализа данных, но сочинения и изложения учащихся по-прежнему проверяются учителями без использования компьютера. В данной работе предлагаются подходы, позволяющие анализировать ряд характеристик текста изложения алгоритмически с использованием специальной программы.

Методами теории случайных процессов на основании элементарной вероятностной модели анализируется процесс появления новых слов в тексте. Основы этого подхода заложены в работах Бахадура [1] и Карлина [2]. Барбур и Гнедин [3] предложили нормальную аппроксимацию, а Барбур [4] – пуассоновскую аппроксимацию для числа слов, встретившихся один раз. Муратов и Зуев [5] изучили свойства числа разных слов как цепи Маркова. Чебунин и Ковалевский [6] доказали функциональную центральную предельную теорему для числа разных слов. Дюреу и Ванг [7] доказали подобную теорему для рандомизированного процесса, в котором каждому слову присваивается случайным образом значение «1» или «–1» независимо от процесса появления слов. Впоследствии Чебунин [8] обобщил функциональную центральную предельную теорему на случай сверхтяжелых хвостов распределения. Современное состояние исследований этой вероятностной модели изложено в работах Бен-Хаму, Бошерона, Оганесяна [9] и Декруеза, Грабчака, Париса [10].

Степенной закон убывания вероятностей слов предложен Ципфом [11] (гл. 1). Различные классы оценок параметра Ципфа предложены Ничолсом [12], Закревской и Ковалевским [13], Чебуниным [14], Чебуниным и Ковалевским [15, 16]. Кроме того, в работе [16] предложена процедура проверки гипотезы о соответствии текста закону Ципфа. Эта процедура реализована Закревской и Ковалевским [17] при анализе текстов Шекспира. По процессу числа разных слов строится аналог эмпирического моста, предложенного Гусаровой, Ковалевским, Макаренко [18]. Предельное поведение эмпирического моста было изучено Ковалевским и Шаталиным для однопараметрической регрессии [19], простой парной регрессии [20] и ее обобщений [21]. Вычисление предельного распределения статистики типа омега-квадрат основывается на формуле Смирнова [22]. Подробно подходы к ее применению изложены в книге Мартынова [23] (гл. 3). Также предельное распределение может быть изучено методами Дехеувелса и Мартынова [24].

В лингвистике закон степенного роста словаря известен под именем закона Хипса [25] (параграф 7.5, с. 206–208). Впервые он был предложен Херданом [26] (гл. 4).

1. ИСХОДНЫЕ ДАННЫЕ

Исходными данными для анализа являются авторский текст К.Г. Паустовского (отрывок из повести «Скрипучие половицы») и его изложения школьниками 9-го класса лицея № 126 г. Новосибирска. Изложения выпол-

нялись учениками на компьютерах в компьютерном классе после двух прослушиваний текста оригинала.

Все изложения были сохранены в текстовом формате, орфографические ошибки исправлены. Сравнение исходного текста (оригинала) и изложений показывает, во-первых, что изложения гораздо короче оригинала и, во-вторых, что изложения почти не содержат эпитетов, использование которых столь характерно для оригинала и вообще для сочинений Паустовского.

Так, фраза Паустовского «Он играл. Он добивался ясности мелодии – такой, чтобы она была понятна и мила и Фене, и даже старому Василию, ворчливому леснику из соседней помещичьей усадьбы» передана в изложении ученика словами: «Он очень долго играл». А вместо фразы «Он долго простоял на обрыве Рудого Яра. С зарослей липы и бересклета капала роса. Столько сырого блеска было вокруг, что он невольно прищурил глаза. Но больше всего в этот день Чайковского поразил свет. Он вглядывался в него, видел все новые пласты света, падавшие на знакомые леса. Как только он раньше не замечал этого? С неба свет лился прямыми потоками, и под этим светом особенно выпуклыми и кудрявыми казались вершины леса, видного сверху, с обрыва» в другом изложении лишь короткое предложение: «Вокруг яра было так сыро и много света».

Фрагмент текста: «Он знал, что сегодня, побывав там, вернется – и давно живущая где-то внутри любимая тема о лирической силе этой лесной стороны перельется через край и хлынет потоками звуков. Так и случилось» в изложении выглядит так: «Он знал, что сегодня побывает там. Так и случилось».

Значения параметров текста позволяют интегрально и непредвзято оценить отличия словаря изложения от словаря исходного текста.

Для анализа разладки в текстах были созданы гибридные тексты. К авторскому тексту присоединялся текст изложения.

2. БИБЛИОТЕКА ТРИГРАММ

Для анализа стандартных словосочетаний в изложениях создана библиотека триграмм (троек слов) по произведениям русской классики. С сайта библиотеки Мошкова (lib.ru) для анализа были взяты собрания сочинений четырех авторов: Федора Михайловича Достоевского, Льва Николаевича Толстого, Ивана Сергеевича Тургенева, Антона Павловича Чехова. Собрания сочинений сравнивались попарно (все шесть пар) после исключения всех знаков препинания, и в библиотеку записывались тройки последовательных слов, совпадающие в сочинениях разных авторов. Этот подход к построению библиотеки троек слов обеспечивает независимость от конкретного автора или персонажа. Если отказаться от этого подхода, то в число наиболее частых могут попасть триграммы, не характерные для языка в целом. Например, одной из самых частых в романе «Анна Каренина» Л.Н. Толстого является триграмма «Сергей Сергеевич сказал».

Ниже представлены наиболее частые из библиотеки триграмм русской классики вместе с количеством их вхождений в библиотеку:

в самом деле – 6178,	в том что – 2522,
по крайней мере – 5734,	до сих пор – 2356,
в это время – 4038,	о том что – 2072,

на другой день – 3134, за то что – 1984,
в то время – 3089, с тех пор – 1749.

Всего библиотека содержит 6060 триграмм с частотами вхождений не менее двух, т. е. каждая из триграмм библиотеки встретилась в произведениях каждого из двух авторов, причем в произведениях хотя бы одного из них по крайней мере дважды. Триграммы, встретившиеся только один раз в собрании сочинений каждого из авторов, исключены из рассмотрения, так как их появление можно считать случайным.

Полностью библиотека триграмм русской классики приведена на сайте https://ciu.nstu.ru/WebInput/persons/750/a/file_get/300639?nomenu=1

3. АЛГОРИТМ И РЕЗУЛЬТАТЫ АНАЛИЗА

Алгоритм анализа изложен в статье [16]. По тексту вычисляется оценка параметра Ципфа, строится эмпирический мост числа разных слов и вычисляется статистика омега-квадрат.

Статистика омега-квадрат (интеграл от квадрата эмпирического моста) для исходного текста близка к 0,01. Для текстов изложений она принимает также небольшие значения (табл. 1), что свидетельствует об однородности изложений. На рис. 1 и 2 изображены процесс накопления слов и эмпирический мост.

Таблица 1

Table 1

Характеристики исходного текста и изложений

Characteristics of the source text and essays

Изложение № essay No.	Слов words	Разных different	Параметр Ципфа Zipf parameter	Сдвиг shift	Омега-квадрат omega-squared
Скрипучие половицы Creaking floorboards	441	328	0.898120386	1.116124	0.011777359
1	199	155	0.95419631	–0.72516	0.006651562
2	142	102	0.904241017	–0.32953	0.013911314
3	199	156	0.876011283	1.971861	0.018892307
4	152	120	0.96437609	–0.81456	0.021083086
5	226	156	0.876011283	0.388949	0.06921515
6	186	141	0.969626351	–0.87568	0.029806418
7	127	96	0.884522783	0.340465	0.006134134
8	104	82	0.818393194	3.696324	0.0127207
9	163	119	0.828728573	2.756179	0.018858037
10	241	173	0.849665727	2.284645	0.019610636
11	184	139	0.929116514	–0.48036	0.004004929
12	136	104	0.854949667	1.746372	0.020132858

Окончание табл. 1

End of Tab. 1

Изложение № essay No.	Слов words	Разных different	Параметр Ципфа Zipf parameter	Сдвиг shift	Омега-квадрат omega-squared
13	48	43	0.871675903	2.041135	0.014516098
14	83	69	0.938599455	–0.4767	0.008009747
15	74	61	0.864648147	1.366616	0.007252226
16	135	106	0.946560741	–0.66408	0.017029592
17	127	105	0.881355504	1.781554	0.01645335
18	88	75	0.888968688	1.288099	0.010209497
19	165	120	0.884523	0.257893	0.01429
20	91	72	0.847996907	1.763144	0.004936503

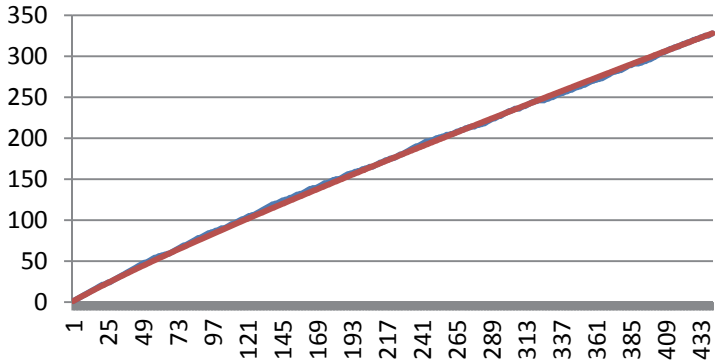

Рис. 1. Процесс числа разных слов для исходного текста

Fig. 1. The process of numbers of different words for the source text

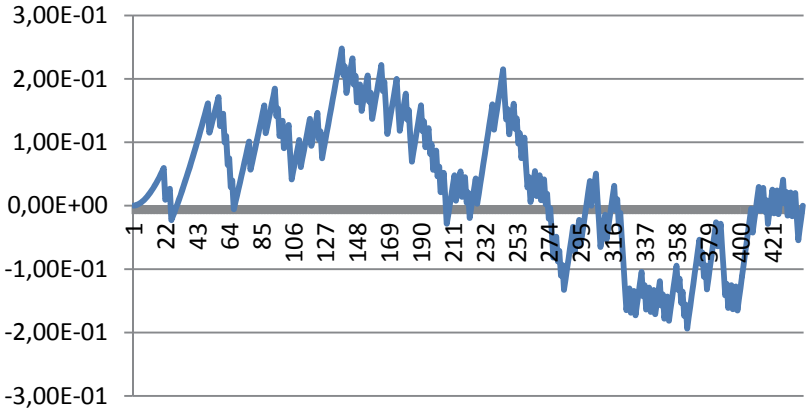

Рис. 2. Эмпирический мост числа разных слов для исходного текста

Fig. 2. An empirical bridge of numbers of different words for the source text

Другая картина получается при изучении текстов, составленных из исходного текста и изложения (табл. 2).

Здесь статистика омега-квадрат гораздо больше, чем для исходного текста и изложений по отдельности. При этом значения омега-квадрат колеблются в пределах от 0,06 до 3,19. Причину этого выясним, изучив изложения 1 и 6 с наибольшим и наименьшим значениями статистики.

Процесс роста числа разных слов для исходного текста с присоединенным к нему изложением 1 (рис. 3) показывает, что после окончания авторского текста (441-го по порядку слова) новые слова почти не появляются, т. е. учащийся почти полностью обходится словами исходного текста.

Таблица 2

Table 2

Характеристики составных текстов (сначала следует исходный текст, за ним изложение)

Characteristics of composite texts (first the source text, followed by the essay)

Изложение № essay No.	Слов words	Разных different	Параметр Ципфа Zipf parameter	Сдвиг shift	Омега-квадрат omega-squared
1	690	344	0.401125	159.6238	3.19319
2	605	390	0.736966	22.83753	0.24811
3	653	378	0.608046	56.24678	0.951314
4	605	386	0.722092	25.91739	0.296835
5	679	412	0.680876	36.46613	0.45806
6	636	424	0.79423	12.78176	0.065608
7	583	375	0.736966	21.76538	0.373437
8	555	359	0.739647	20.56639	0.416502
9	619	364	0.612977	54.02718	0.988777
10	694	419	0.677383	37.8343	0.511081
11	645	379	0.62354	51.53493	0.882135
12	597	378	0.710493	28.10251	0.427383
13	540	346	0.710114	26.8219	0.584571
14	559	371	0.780388	13.57649	0.240804
15	569	367	0.744849	19.51301	0.41823
16	594	359	0.64549	43.51465	0.856879
17	577	386	0.791557	12.3093	0.182099
18	553	349	0.705616	27.00296	0.670425
19	623	386	0.69159	32.64327	0.441373
20	617	364	0.619051	51.84014	1.108292

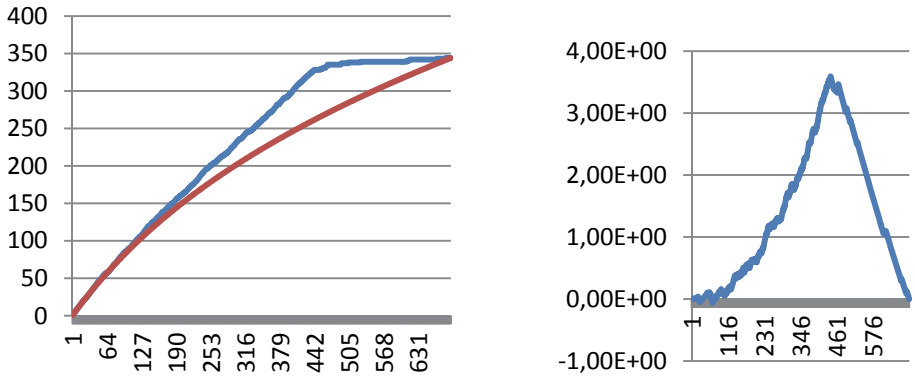

Рис. 3. Процесс числа разных слов, его приближение в соответствии с законом Ципфа (слева) и эмпирический мост числа разных слов (справа) для текста, составленного из исходного текста и изложения 1

Fig. 3. The process of numbers of different words, its approximation according to Zipf's law (left) and the empirical bridge of numbers of different words (right) for a text composed of the source text and essay 1

В результате эмпирический мост имеет острый и высокий пик, точка максимума достигается на 441-м слове, т. е. в точности указывает на окончание авторского текста. Для компенсации столь резкого уменьшения скорости роста словаря требуется низкое значение параметра Ципфа 0,4 и громадный сдвиг 160, а статистика омега-квардат принимает значение 3,19.

В случае присоединения к исходному тексту изложения 6 процесс роста числа разных слов (рис. 4) показывает продолжение роста, хорошо согласующееся с ростом числа слов в исходном тексте, т. е. учащийся пересказывает текст своими словами. В результате эмпирический мост имеет значительно менее высокий пик, но точка максимума также достигается на 441-м слове и служит хорошей оценкой момента разладки, т. е. окончания авторского текста и начала изложения. Оценка параметра Ципфа 0,79, а сдвига 12,8 говорит об отличии от авторского текста, но гораздо менее значительном. Статистика омега-квадрат принимает значение 0,00656.

Итоговый анализ включает выделение слов, не встречавшихся у автора, и стандартных триграмм. Также отыскивается предложение, в котором максимально число слов, отсутствующих у автора. Пример анализа (знаки препинания исключаются):

В это утро Чайковский проснулся рано Встал с {кровати} И не глядя в окно вспоминал {свое} любимое место Его дом {находился} на пригорке {Зеленый} {лес} {уходил} {зарослями} вниз Его {любимым} {местом} был {Рудный} {яр} Бывало и среди ночи он {начнет} вспоминать дорогу в это место Она {проходила} и по {небольшому} {мосту} и по {корабельному} {порту} {находящемуся} {вверху} В это утро он {понял} что ему {нужно} как {можно} {скорее} {попасть} в это {удивительное} место Долго он стоял на {опушке}Его {взор} {падал} на заросли {деревьев} {окутанных} {теплым} {солнечным} светом И в этот {раз} его поразил свет Чайковский долго вглядывался в него Он лился даже с неба {прямым} {потоком} на {полянку} А на

опушку падали косые {лучики} света {Сосны} {от} него {приобрели} {тот} {самый} {золотистый} {оттенок} Чайковский заметил что в этот день стволы сосен {отбрасывали} свет на опушку {Мягкий} {едва} {заметный} свет Его {внимание} {привлекли} {ивы} и {ольха} {которые} были освещены {голубым} светом {отбрасываемым} {от} воды Все это свет {опушка} {деревья} {вызвало} у него {чувство} {чуда}

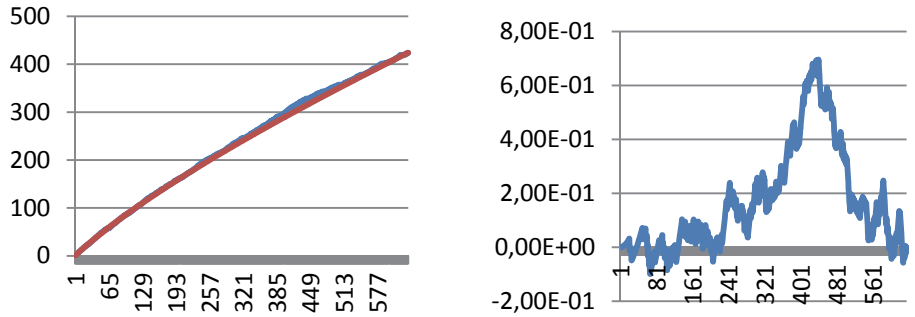

Рис. 4. Процесс числа разных слов, его приближение в соответствии с законом Ципфа (слева) и эмпирический мост числа разных слов (справа) для текста, составленного из исходного текста и изложения 6

Fig. 4. The process of numbers of different words, its approximation according to Zipf's law (left) and the empirical bridge of numbers of different words (right) for a text composed of the source text and essay 6

Триграммы: в этот день – 1, как можно скорее – 1, что ему нужно – 1, в этот раз – 1, он понял что – 1, и в этот – 1.

Цитата: «Она (*дорога*) проходила и по небольшому мосту, и по корабельному порту, находящемуся вверху».

ЗАКЛЮЧЕНИЕ

В статье разработан новый метод компьютерного анализа изложений. Анализ текста, образованного прибавлением текста изложения к авторскому тексту, проводится на основании процесса роста словаря текста. Эмпирический мост текста – это нормированный процесс разности между числом разных слов и его прогнозом на основании закона Ципфа. Оценка качества самостоятельного изложения текста строится по статистике омега-квадрат интеграла от квадрата эмпирического моста. На основании исследования можно предложить следующий критерий оценки самостоятельности изложения (чем статистика меньше, тем лучше): «отлично» – меньше 0,3, «хорошо» – от 0,3 до 0,8, «удовлетворительно» – от 0,8 до 2, «неудовлетворительно» – больше 2.

Анализ новых слов, введенных автором изложения, и стандартных триграмм открывает новые возможности для оценивания и исследования изложений.

Авторы благодарят А.П. Ковалевского за предложенное направление исследований, Д.З. Алдагарова – за помощь в программировании, Е.Н. Лободюк – за помощь в составлении библиотеки триграмм.

СПИСОК ЛИТЕРАТУРЫ

1. *Bahadur R.R.* On the number of distinct values in a large sample from an infinite discrete distribution // Proceedings of the National Institute of Sciences of India. – 1960. – Vol. 26A, Suppl. 2. – P. 67–75.

2. *Karlin S.* Central limit theorems for certain infinite urn schemes // Journal of Mathematics and Mechanics. – 1967. – Vol. 17, N 4. – P. 373–401.

3. *Barbour A.D., Gnedin A.V.* Small counts in the infinite occupancy scheme // Electronic Journal of Probability. – 2009. – Vol. 14. – P. 365–384.

4. *Barbour A.D.* Univariate approximations in the infinite occupancy scheme // Latin American Journal of Probability and Mathematical Statistics. – 2009. – Vol. 6. – P. 415–433.

5. *Muratov A., Zuyev S.* Bit flipping and time to recover // Journal of Applied Probability. – 2016. – Vol. 53, N 3. – P. 650–666.

6. *Chebunin M., Kovalevskii A.* Functional central limit theorems for certain statistics in an infinite urn scheme // Statistics and Probability Letters. – 2016. – Vol. 119. – P. 344–348.

7. *Durieu O., Wang Y.* From infinite urn schemes to decompositions of self-similar Gaussian processes // Electronic Journal of Probability. – 2016. – Vol. 21. – P. 43. – DOI: 10.1214/16-EJP4492.

8. *Чебунин М.Г.* Функциональная центральная предельная теорема в бесконечной урновой схеме для распределений со сверхтяжелыми хвостами // Сибирские электронные математические известия. – 2017. – Т. 14. – С. 1289–1298. – DOI: 10.17377/semi.2017.14.109.

9. *Ben-Hamou A., Boucheron S., Ohannessian M.I.* Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications // Bernoulli. – 2017. – Vol. 23, N 1. – P. 249–287.

10. *Decrouez G., Grabchak M., Paris Q.* Finite sample properties of the mean occupancy counts and probabilities // Bernoulli. – 2018. – Vol. 24, N 3. – P. 1910–1941.

11. *Zipf G.K.* The psycho-biology of language. – London: Routledge, 1936.

12. *Nicholls P.T.* Estimation of Zipf parameters // Journal of the American Society for Information Science. – 1987. – Vol. 38 (6). – P. 443–445.

13. *Закревская Н.С., Ковалевский А.П.* Однопараметрические вероятностные модели статистик текста // Сибирский журнал индустриальной математики. – 2001. – Т. 4, № 2. – С. 142–153.

14. *Чебунин М.Г. Оценивание параметров* вероятностных моделей по числу различных элементов выборки // Сибирский журнал индустриальной математики. – 2014. – Т. 17, № 3. – С. 135–147.

15. *Chebunin M., Kovalevskii A.* Asymptotically normal estimators for Zipf's law // Sankhya A. – 2019. – Vol. 81, iss. 2. – P. 482–492. – DOI: 10.1007/s13171-018-0135-9.

16. *Chebunin M.G., Kovalevskii A.P.* A statistical test for the Zipf's law by deviations from the Heaps' law // Сибирские электронные математические известия. – 2019. – Т. 16. – С. 1822–1832.

17. *Zakrevskaya N., Kovalevskii A.* An omega-square statistics for analysis of correspondence of small texts to the Zipf-Mandelbrot law // Applied methods of statistical analysis. Statistical computation and simulation – AMSA'2019, 18–20 September 2019, Novosibirsk: Proceedings of the International Workshop. – Novosibirsk: NSTU, 2019. – P. 488–494.

18. *Гусарова Г.В., Ковалевский А.П., Макаренко А.Г.* Критерии наличия разладки // Сибирский журнал индустриальной математики. – 2005.– Т. 8, № 4. – С. 18–33.

19. *Kovalevskii A.P., Shatalin E.V.* Asymptotics of sums of residuals of one-parameter linear regression on order statistics // Theory of Probability and Its Applications. – 2015. – Vol. 59, N 3. – P. 375–387.

20. *Ковалевский А.П., Шаталин Е.В.* Выбор регрессионной модели зависимости массы тела от роста с помощью эмпирического моста // Вестник Томского государственного университета. Математика и механика. – 2015. – № 5 (37). – С. 35–47. – DOI: 10.17223/19988621/37/3.

21. *Kovalevskii A., Shatalin E.* A limit process for a sequence of partial sums of residuals of a simple regression on order statistics // Probability and Mathematical Statistics. – 2016. – Vol. 36, Fasc. 1. – P. 113–120.

22. *Смирнов Н.В.* О распределении ω2-критерия Мизеса // Математический сборник. – 1937. – Т. 2, № 5. – С. 973–993.

23. *Мартынов Г.В.* Критерии омега-квадрат. – М.: Наука, 1978. – 79 с.

24. *Deheuvels P., Martynov G.V.* Cramer-von mises-type tests with applications to tests of independence for multivariate extreme-value distributions // Communications in Statistics – Theory and Methods. – 1996. – Vol. 25, N 4. – P. 871–908.

25. *Heaps H.S.* Information retrieval: computational and theoretical aspects. – New York: Academic Press, 1978.

26. *Herdan G.* Type-token mathematics. – The Hague: Mouton, 1960.

Закревская Наталья Станиславовна, аспирант факультета прикладной математики и информатики, преподаватель кафедры высшей математики Новосибирского государственного технического университета. Основное направление научных исследований – математическая статистика. Имеет более 10 публикаций. E-mail: natali.erlagol@gmail.com

Ковалевская Мария Артемовна, магистрант философского факультета Новосибирского государственного университета. Основное направление научных исследований – математическая лингвистика. Не имеет публикаций. E-mail: maryartkey@gmail.com

Чебунин Михаил Георгиевич, научный сотрудник Института математики им. С.Л. Соболева, и.о. зав. лабораторией прикладной вероятности Новосибирского государственного университета. Основное направление научных исследований – математическая статистика. Имеет более 15 публикаций. E-mail:chebuninmikhail@gmail.com

Zakrevskaya Natalia S., post-graduate student of the faculty of applied mathematics and computer science, teacher of the department of higher mathematics in Novosibirsk State Technical University. The main area of research is mathematical statistics.She is the author of more than 10 publications. E-mail: natli.erlagol@gmail.com

Kovalevskaya Maria A., master's student of the faculty of philosophy of Novosibirsk State University. The main area of her research is mathematical linguistics. She has no publications. E-mail: maryartkey@gmail.com

Chebunin, Mikhail G., research associate at the S.L. Sobolev Institute of mathematics, acting head. Of the laboratory of applied probability in Novosibirsk State University. The main area of his research is mathematical statistics. He has published more than 15 publications. E-mail: chebuninmikhail@gmail.com

DOI: 10.17212/1814-1196-2020-1-75-86

Computer analysis of school paraphrase essays based on Heaps' law[*]

N.S. ZAKREVSKAYA[1,a], M.A. KOVALEVSKAYA[2,b], M.G. CHEBUNIN[3,c]

[1] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation*

[2] *Novosibirsk State University, 1 Pirogov Street, Novosibirsk, 630090, Russian Federation*

[3] *Sobolev Institute of Mathematics, 4 Koptjug Prospekt, Novosibirsk, 630090, Russian Federation; Novosibirsk State University, 1, Pirogov Street, Novosibirsk, 630090, Russian Federation*

[a] *natali.erlagol@gmail.com* [b] *maryartkey@gmail.com* [c] *chebuninmikhail@gmail.com*

Abstract

The purpose of writing a paraphrase essay is not to copy the author's text, but to re-tell the main content in other words. A computer analysis of the text should take into account that both the original author's text and its paraphrase made by a student obey the Zipf's law about the frequency distribution of words, but with different dictionaries and with different parameters. The difference

[*] *Received 16 December 2019.*

The research was carried out with partial financial support from the RFBR and the government of the Novosibirsk region (grant no. 19-41-543004) and support from the basic research program of the SB RAS no. I. 1. 3. (project no. 0314-2019-0008)

between dictionaries and parameters of the Zipf's law for the author's text and for students' texts is due to different social environment, the circle of reading and communication, the ability to remember and use words of different language layers. The mathematical apparatus for analyzing correspondence of the text to Zipf's law was developed in a number of theoretical works on probability theory and mathematical statistics. Methods of the stochastic processes theory describe the behavior of a sequence of quantities of different words in a text. Note that the power law of the growth of the number of different words with increasing text length is called the Heaps' law in applied statistics. Theoretical results based on the elementary probabilistic model allow us to analyze the significance of deviations from the Heaps' law. In this work, an algorithm for analyzing paraphrase essays is developed. It includes comparing the text of the paraphrase with the original author's text, identifying duplicating words and phrases. A separate analysis is made for standard trigrams (triples of words) using a library of the most frequent trigrams compiled from the library of Russian classics. Then, the uniformity of the text is analyzed based on its compliance with the Heaps' law. In conclusion, the homogeneity of combined texts composed of the source text and the text of the essay is studied. The developed approach gives new information about the text of the essay and its compliance with the source text. It can be used both for computer analysis of statements, and for checking the uniformity of the text.

Keywords: Zipf law, Heaps' law, weak convergence, Gaussian process, trigram, empirical bridge, omega-square statistics, statistical test, p-value, text uniformity

REFERENCES

1. Bahadur R.R. On the number of distinct values in a large sample from an infinite discrete distribution. *Proceedings of the National Institute of Sciences of India*, 1960, vol. 26A, Suppl. 2, pp. 67–75.

2. Karlin S. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 1967, vol. 17, no. 4, pp. 373–401.

3. Barbour A.D., Gnedin A.V. Small counts in the infinite occupancy scheme. *Electronic Journal of Probability*, 2009, vol. 14, pp. 365–384.

4. Barbour A.D. Univariate approximations in the infinite occupancy scheme. *Latin American Journal of Probability and Mathematical Statistics*, 2009, vol. 6, pp. 415–433.

5. Muratov A., Zuyev S. Bit flipping and time to recover. *Journal of Applied Probability*, 2016, vol. 53, no. 3, pp. 650–666.

6. Chebunin M., Kovalevskii A. Functional central limit theorems for certain statistics in an infinite urn scheme. *Statistics and Probability Letters*, 2016, vol. 119, pp. 344–348.

7. Durieu O., Wang Y. From infinite urn schemes to decompositions of self-similar Gaussian processes. *Electronic Journal of Probability*, 2016, vol. 21, p. 43. DOI: 10.1214/16-EJP4492.

8. Chebunin M.G. Funktsional'naya tsentral'naya predel'naya teorema v beskonechnoi urnovoi skheme dlya raspredelenii so sverkhtyazhelymi khvostami [Functional central limit theorem in an infinite urn scheme for distributions with superheavy tails]. *Sibirskie elektronnye matematicheskie izvestiya = Siberian Electronic Mathematical Reports*, 2017, vol. 14, pp. 1289–1298. DOI: 10.17377/semi.2017.14.109. (In Russian).

9. Ben-Hamou A., Boucheron S., Ohannessian M.I. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 2017, vol. 23, no. 1, pp. 249–287.

10. Decrouez G., Grabchak M., Paris Q. Finite sample properties of the mean occupancy counts and probabilities. *Bernoulli*, 2018, vol. 24, no. 3, pp. 1910–1941.

11. Zipf G.K. *The psycho-biology of language*. London, Routledge, 1936.

12. Nicholls P.T. Estimation of Zipf parameters. *Journal of the American Society for Information Science*, 1987, vol. 38 (6), pp. 443–445.

13. Zakrevskaya N.S., Kovalevskii A.P. Odnoparametricheskie verotnostnye modeli statistik teksta [One-parameter probabilistic models of text statistics]. *Sibirskii zhurnal industrial'noi matematiki = Journal of Applied and Industrial Mathematics*, 2001, vol. 4, no. 2, pp. 142–153. (In Russian).

14. Chebunin M.G. Otsenivanie parametrov veroyatnostnykh modelei po chislu razlichnykh elementov vyborki [Estimation of parameters of probabilistic models which is based on the number of

different elements in a sample]. *Sibirskii zhurnal industrial'noi matematiki = Journal of Applied and Industrial Mathematics*, 2014, vol. 17, no. 3, pp. 135–147. (In Russian).

15. Chebunin M., Kovalevskii A. Asymptotically normal estimators for Zipf's law. *Sankhya A*, 2019, vol. 81, iss. 2, pp. 482–492. DOI: 10.1007/s13171-018-0135-9.

16. Chebunin M.G., Kovalevskii A.P. A statistical test for the Zipf's law by deviations from the Heaps' law. *Sibirskie elektronnye matematicheskie izvestiya = Siberian Electronic Mathematical Reports*, 2019, vol. 16, pp. 1822–1832.

17. Zakrevskaya N., Kovalevskii A. An omega-square statistics for analysis of correspondence of small texts to the Zipf-Mandelbrot law. *Applied methods of statistical analysis. Statistical computation and simulation – AMSA'2019*, 18–20 September 2019, Novosibirsk: Proceedings of the International Workshop. Novosibirsk, NSTU Publ., 2019, pp. 488–494.

18. Gusarova G.V., Kovalevskii A.P., Makarenko A.G. Kriterii nalichiya razladki [Criteria for the existence of a change point]. *Sibirskii zhurnal industrial'noi matematiki = Journal of Applied and Industrial Mathematics*, 2005, vol. 8, no. 4, pp. 18–33. (In Russian).

19. Kovalevskii A.P., Shatalin E.V. Asymptotics of sums of residuals of one-parameter linear regression on order statistics. *Theory of Probability and Its Applications*, 2015, vol. 59, no. 3, pp. 375–387.

20. Kovalevskii A.P., Shatalin E.V. Vybor regressionnoi modeli zavisimosti massy tela ot rosta s pomoshch'yu empiricheskogo mosta [The choice of a regression model of the body weight on the height via an empirical bridge]. *Vestnik Tomskogo gosudarstvennogo universiteta. Matematika i mekhanika = Tomsk State University Journal of Mathematics and Mechanics*, 2015, vol. 5 (37), pp. 35–47. DOI: 10.17223/19988621/37/3.

21. Kovalevskii A., Shatalin E. A limit process for a sequence of partial sums of residuals of a simple regression on order statistics. *Probability and Mathematical Statistics*, 2016, vol. 36, Fasc. 1, pp. 113–120.

22. Smirnoff N.V. O raspredelenii ω2-kriteriya Mizesa [Sur la distribution de ω2 (critérium de M. v. Mises)]. *Matematicheskii sbornik = Recueil Mathematique*, 1937, vol. 2, no. 5, pp. 973–993. (In Russian).

23. Martynov G.V. *Kriterii omega-kvadrat* [Omega-square tests]. Moscow, Nauka Publ., 1978. 79 p.

24. Deheuvels P., Martynov G.V. Cramer-von mises-type tests with applications to tests of independence for multivariate extreme-value distributions. *Communications in Statistics – Theory and Methods*, 1996, vol. 25, no. 4, pp. 871–908.

25. Heaps H.S. *Information retrieval: computational and theoretical aspects*. New York, Academic Press, 1978.

26. Herdan G. *Type-token mathematics*. The Hague, Mouton, 1960.

Для цитирования:

Закревская Н.С., Ковалевская М.А., Чебунин М.Г. Компьютерный анализ школьных изложений на основании закона Хипса // Научный вестник НГТУ. – 2020. – № 1 (78). – С. 75–86. – DOI: 10.17212/1814-1196-2020-1-75-86.

For citation:

Zakrevskaya N.S., Kovalevskaya M.A., Chebunin M.G. Komp'yuternyi analiz shkol'nykh izlozhenii na osnovanii zakona Khipsa [Computer analysis of school paraphrase essays based on Heaps' law]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta = Science bulletin of the Novosibirsk state technical university*, 2020, no. 1 (78), pp. 75–86. DOI: 10.17212/1814-1196-2020-1-75-86.