Частотный анализ текста с помощью компьютера

Р.В. Майер

Аннотация **— Рассмотрен метод частотного анализа текста, приводящий к получения спектральных распределений букв, слов, смысловых отрезков. Цель статьи заключается: 1) в создании компьютерных программ, позволяющих получить спектры распределения слов и отдельных символов в текстах большого объема; 2) в их апробации при анализе повести В.Г. Короленко «Дети подземелья»; 3) в построении вероятностной модели писателя. Представлены три программы на языке ABCPascal, позволяющие получить: 1) частотное распределение букв и их сочетаний; 2) спектральное распределение слов и смысловых отрезков; 3) количество переходов от смысловых отрезков длиной n к смысловым отрезкам длиной m. В статье приводятся: 1) спектральные распределения гласных «о», «а», «е», «и», «у», «я», «ю» в анализируемом тексте; 2) частотные распределение слов по длине; 3) спектр смысловых отрезков текста, ограниченных знаками препинания; 4) матрица переходов от смысловых отрезков длиной n к смысловым отрезкам длиной m; 5) таблица вероятностей этих переходов; 6) граф вероятностного автомата, моделирующего порождение автором текста. Его вершины соответствуют количеству слов в смысловых отрезках текста, отделенных знаками препинания, а ребра – наиболее вероятным переходам. Все это характеризует индивидуальные особенности стиля писателя и может быть использовано для установления авторства.**

Ключевые слова **— вероятностный автомат, текст, В.Г. Короленко, программирование, стохастическая матрица, частотный анализ.**

I. Введение

Один из методов изучения художественных, публицистических, научных и других текстов заключается в его статистическом анализе, изучении частотного распределения букв, слогов, слов и т.д. Выявление ключевых слов, часто используемых словосочетаний и определение их частот позволяет классифицировать тексты по темам. Известные семантические методы классификации текстов [1, 3], основанные на вычислении косинусоидальной меры близости, меры Дайса и других мер, могут быть улучшены путем учета спектрального распределения букв и их сочетаний.

В общем случае текст имеет периодическую структуру [10], в нем наблюдаются упорядоченное и неупорядоченное повторения отдельных элементов (символов и их сочетаний). Наряду с периодическими конструкциями, регулярными колебаниями и ритмами, происходящими с невысокими «частотами», имеют место «высокочастотные» хаотические колебания, характеризующиеся непрерывным спектром. При этом чередуются диалогические и монологические сцены, сюжетно-повествовательные и описательные эпизоды, начала и окончания предложений, существительные и глаголы, слова и промежутки между ними, гласные и согласные буквы, знаки препинания, отдельные буквы, их сочетания (биграммы, триграммы) и т.д.

У каждого писателя свои стилистические, художественные и языковые особенности. Любое произведение может быть охарактеризовано целой совокупностью статистических характеристик авторского стиля, среди которых спектральное распределение букв, их сочетаний, слов, смысловых отрезков предложений. Статистическое распределение различных элементов может быть использовано для оценки качества текстового материала, выявления особенностей его структуры и идиостиля писателя, для распознавания авторства текстов [7].

Изучение зависимости числа букв, биграмм, триграмм от периода повторения, а также количества слов от их длины позволяет: 1) определить среднюю длину слова в тексте; 2) оценить разнообразие слов и их словоформ; 3) выявить повторяющиеся буквенные шаблоны; 4) изучить стилистические особенности прозаических и поэтических произведений; 5) установить жанр текста; 6) идентифицировать автора, установить особенности его идиостиля [8]; 7) выявить ошибки при сканировании и оптическом распознавании текста.

Изучение спектральных характеристик текста предусматривает комплексное применение методов компьютерного анализа, статистики и визуализации данных [2, 5]. Для этого проводят токенизацию текста, то есть выделяют слова и знаки препинания, а затем подсчитывают количество употреблений в тексте токенов каждого типа и получают соответствующую гистограмму или график, характеризующие их частотное распределение.

Цель статьи состоит: 1) в создании компьютерных программ, позволяющих получить частотные распределения слов и отдельных символов в текстах большого объема; 2) их апробации при анализе повести В.Г. Короленко «Дети подземелья»; 3) в построении вероятностной модели писателя. В качестве методологической основы использовались работы следующих ученых: Г.Г. Москальчук и Н.А. Манаков [5], Д.В. Руднев и С.В. Друговейко-Должанская [8], А. И. Трубкина [9] (лингвистика); Г.Г. Белоногов [2], Д.А. Куусела [4], Р.Г. Пиотровский, К.Б. Бектаев и А.А. Пиотровская [6], Т.Б. Радбиль и М.В. Маркина [7]

Статья получена 12 мая 2024 г.
Майер Роберт Валерьевич, доктор педагогических наук, профессор кафедры физики Глазовского инженерно-педагогического университета имени В.Г. Короленко, (e-mail: robert_maier@mail.ru).

(компьютерная лингвистика). Основными методами исследования являются метод выявления и подсчета сочетаний символов в текстовом файле с помощью компьютерных программ, написанных в ABCPascal.

II. Результаты исследования

1. Частотное распределение букв. Для анализа была выбрана повесть В.Г. Короленко «Дети подземелья» [4] объемом около 82 тыс. знаков. Текст помещают в текстовый файл Text.txt, из него удаляют все символы (пробелы, знаки препинания) за исключением букв русского алфавита. Все буквы преобразуют в строчные; получается так: «речкачерезкоторуюперекинутупомяну тыймоствытекалаизпрудаивпадалавдругой». При этом текст рассматривается как линейное пространство, элементами которого являются буквы. Каждый символ имеет координату x, отсчитываемую от начала текста; она пропорциональна времени, требуемому чтецу для достижения символа.

Допустим, символ «а» встречается в точках с координатами $x_1 = 4$, $x_2 = 9$, $x_3 = 12$, $x_4 = 18$ и т.д. Появлению буквы «а» соответствуют периоды $x_2 - x_1 = 5$, $x_3 - x_2 = 3$, $x_4 - x_3 = 6$; это соответствует трем «спектральным линиям». Если бы буква «а» встречалась строго через 5 символов, то ее спектр представлял бы собой вертикальный отрезок при $\Delta x = 5$, длина которого пропорциональна «яркости» N, то есть количеству букв «а» в тексте. Обычно получается непрерывный спектр в некотором диапазоне изменения Δx (от 1 до 100).

Алгоритм получения частотного распределения букв (или их сочетаний) в тексте состоит в следующем:

1. Подготовить текстовый файл к обработке: удалить все пробелы, знаки препинания, дефисы, заменить прописные буквы на строчные. В результате получается файл 1.txt, содержащий строки из букв.
2. Используя программу 1 (Приложение), получить частотное распределение букв (или их сочетаний) и сохранить его в текстовом файле 2.txt.
3. Полученные значения загрузить в Excel и на их основе построить график распределения «яркости» N от периода появления Δx.

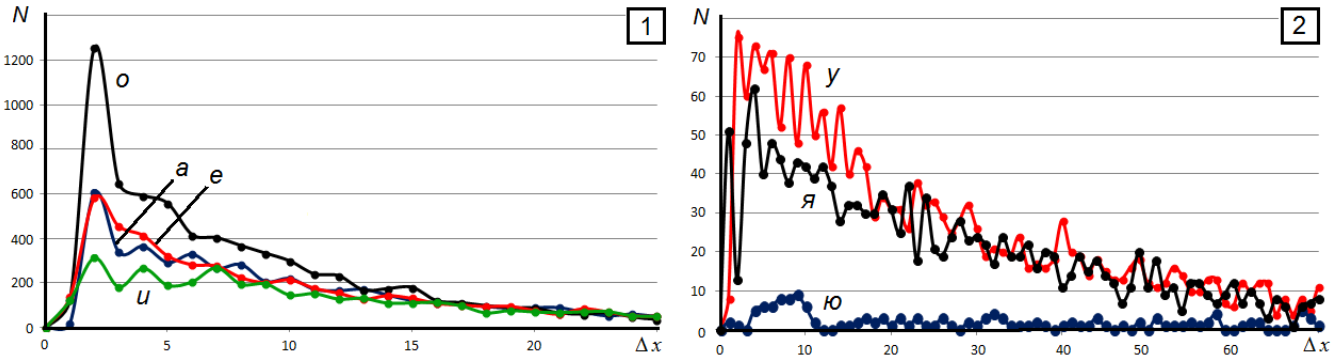

Рис. 1. Спектры гласных «о», «а», «е», «и», «у», «я», «ю» в тексте [4].

Частотные распределения некоторых гласных, полученные в результате анализа текста В.Г. Короленко [4], представлены на рис. 1.1 и 1.2. По горизонтали откладывается периодичность Δx появления той или иной буквы, по вертикали – их количество N. Из графиков видно, что спектры различных букв сильно отличаются друг от друга.

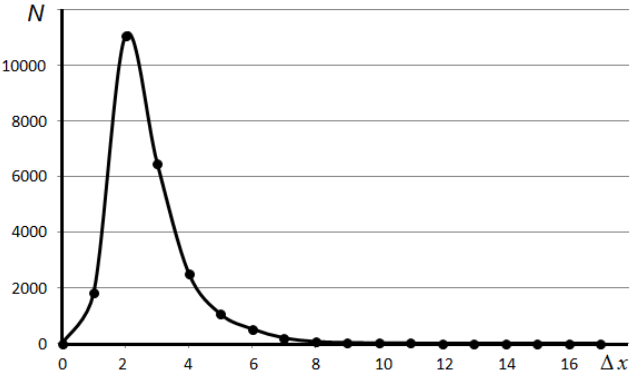

Рис. 2. Общий спектр гласных «а», «о», «е», «у».

Глобальный максимум для букв «а», «е» и «о» приходится на $\Delta x = 2$; это означает, что очень часто расстояние между буквами «а» («е» или «о») равно 2 (как для «о» в слове «городок»). В спектрах букв «а» и «и» присутствуют чередующиеся минимумы и максимумы, соответствующие различным периодам Δx. Для буквы «а» максимумы соответствуют $\Delta x = 2$, 4, 6, а минимумы $\Delta x = 3$, 5, 8 (рис. 1.1). Спектры гласных «у» и «я» содержат большое количество максимумов и минимумов (рис. 1.2). Общий спектр гласных «а», «о», «е», «у» похож на гладкую кривую с одним максимумом, соответствующим $\Delta x = 2$ (рис. 2).

2. Спектральное распределение слов и смысловых отрезков. Если из текста удалить знаки препинания, оставив пробелы, то промежутки Δx между ними будут характеризовать длины слов и периодичность их появлений. Кроме того, можно исключить стоп-слова, к которым относят предлоги, союзы, междометия, частицы и другие части речи, не несущие смысловой нагрузки: а, ан, бы, в, до, же, за, и, из, к, ко, ли, на, над, не, ни, но, о, об, от, по, под, с, со, то, у, уж. Если такой текст содержит только слова длиной 5 и 10 букв, то его спектр состоит из двух вертикальных отрезков, соответствующих $d = \Delta x - 1 = 5$ и 10.

На рис. 3.1 представлены: 1) спектр слов (зависимость числа слов от длины) в тексте [4] – черный график 1; 2) спектры слов в тексте, из которого исключены стоп-слова – красный график 2. Так как стоп-слова обычно состоят из одной или двух букв, то для слов из трех и более букв оба распределения слов совпадают.

С помощью специальной программы, подобной программам 1 и 2, удалось получить частотное распределение смысловых отрезков (СО) текста, – слов и словосочетаний, ограниченных знаками препинания. Учитывались все знаки препинания, кроме кавычек: запятая, точка, точка с запятой, двоеточие, многоточие, восклицательный, вопросительный знаки, тире. Знаки препинания выделяют элементарные мысли автора. Используя программу «Блокнот», их заменяют символом « = », причем расположенные рядом два знака препинания (например, запятая и тире) рассматриваются как один. Пробелы из текста удаляют. Расстояние Δx между знаками препинания на единицу больше суммарного количества букв (не считая пробелов), содержащихся в соответствующей части предложения. Для анализируемого текста спектр распределения СО $N(d)$ содержит большое количество чередующихся максимумов и минимумов (рис. 3.2).

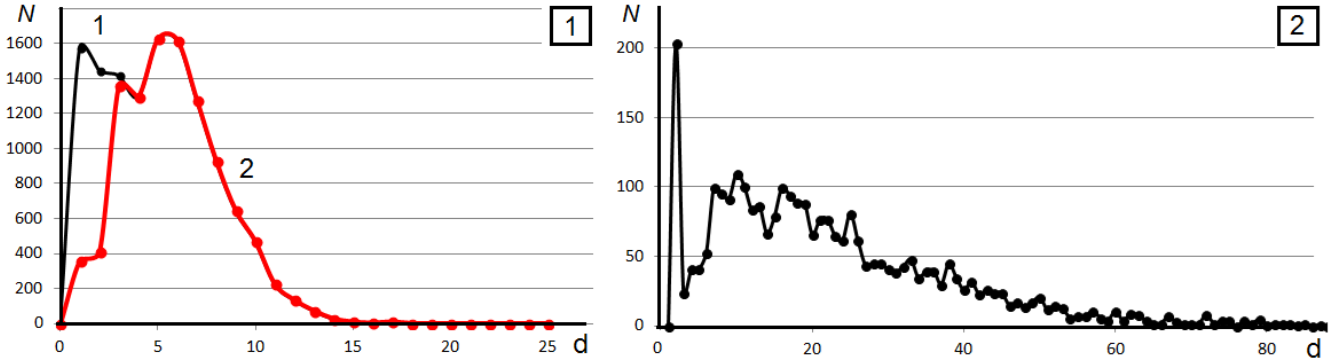

Рис. 3. Распределение слов по длине. Спектр смысловых отрезков текста [4].

Таблица 1. Матрица переходов писателя от СО из n слов к СО из m слов

	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	10	11	12	13	14	…	16	…	21	
m=1	127	89	76	51	45	27	15	15	8	6	6	0	0	0		0		0	465
m=2	115	109	88	73	62	25	19	17	6	3	2	2	1	0		0		0	522
m=3	85	108	99	65	63	30	22	20	14	3	2	2	4	0		1		0	518
m=4	53	73	71	63	54	35	17	15	7	2	2	2	0	1		0		0	395
m=5	25	58	72	64	47	28	18	11	7	4	2	1	1	1		0		0	339
m=6	30	31	33	26	29	15	6	15	5	2	1	2	0	0		0		1	195
m=7	9	18	24	18	15	15	7	6	3	2	1	1	0	0		1		0	120
m=8	13	9	27	14	15	10	9	6	2	1	2	0	0	0		0		0	108
m=9	4	14	8	11	5	3	4	2	4	1	1	0	0	0		0		0	57
10	1	4	6	2	0	6	2	1	1	0	0	0	0	0		0		1	23
11	1	6	6	5	1	0	0	0	0	0	0	0	0	0		0		0	19
12	0	0	4	2	1	2	1	0	0	0	0	0	0	0		0		0	10
13	1	1	3	0	1	0	0	0	0	0	0	0	0	0		0		0	6
14	0	1	1	0	0	0	0	0	0	0	0	0	0	0		0		0	2
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0		0	0
16	1	0	1	0	0	0	0	0	0	0	0	0	0	0		0		0	2
…	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0		0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0		0	0
21	0	1	0	1	0	0	0	0	0	0	0	0	0	0		0		0	2
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0		0	0
S_n	465	522	519	395	338	196	120	108	57	24	19	10	6	2	0	2	0	2	2783

3. Вероятностная модель писателя. Любой текст может быть охарактеризован вероятностями $p(n,m)$ следования СО из m слов после СО из n слов. Для их определения из исходного текста были удалены стоп-слова, а знаки препинания автоматически заменены символом « = ». Ниже приводится фрагмент получившегося текста: «= эта фраза стала выражением крайней степени нищеты = старый замок радушно принимал покрывал временно обнищавшего писца = сиротливых старушек = безродных бродяг = все эти бедняки терзали внутренности дряхлого здания =».

Программа 3 (приложение) распознает отдельные слова и считает их количество между разделительными символами « = », заполняя одномерную матрицу Sw[n], где n – номер СО. Получается так: … 1-8-2-4-7 … Затем она определяет общее число переходов z[n, m] от СО длиной n к СО длиной m (n, m=1, 2, …). Значения z[1, 2], z[1, 3], …, z[2, 1], …, z[20, 20], …, соответствующие переходам 1 → 2, 1 → 3, …, 2 → 1, …, 20 → 20, … сохраняются в файле Vihod.txt в виде столбца. Это позволяет легко перенести их в Excel и получить матрицу переходов (табл. 1).

Таблица 2. Таблица вероятностей переходов n → m.

	n=1	n=2	n=3	4 и 5	6 и 7	8 и 9	10, 11, …
m=1	0,27	0,17	0,15	0,13	0,13	0,14	0,18
m=2	0,25	0,21	0,17	0,18	0,14	0,14	0,12
m=3	0,18	0,21	0,19	0,17	0,16	0,21	0,18
4 и 5	0,16	0,25	0,28	0,31	0,31	0,24	0,25
6 и 7	0,09	0,09	0,11	0,12	0,14	0,18	0,17
8 и 9	0,04	0,04	0,07	0,06	0,08	0,08	0,08
10 и т.д.	0,01	0,03	0,04	0,02	0,04	0,01	0,02

Для повышения статистической значимости некоторые столбцы графа объединяют (например, 6 и 7, или 8 и 9, или 10, 11 и т.д.), а указанные в них частоты суммируют. Вероятность перехода от словосочетания длиной n к словосочетанию длиной m равна:

$$p(n,m) = z[n,m]/S_n,$$

где S_n – сумма чисел в столбце n (табл. 1). Получается стохастическая матрица вероятностей (табл. 2); сумма ее чисел в каждом столбце равна 1. Не смотря на достаточно большой объем текста, некоторые вероятности не являются статистически значимыми, так как количество соответствующих переходов очень мало (например, всего 2 перехода 10 → 7, табл. 1). Этой матрице соответствует вероятностный автомат, моделирующий порождение писателем текста.

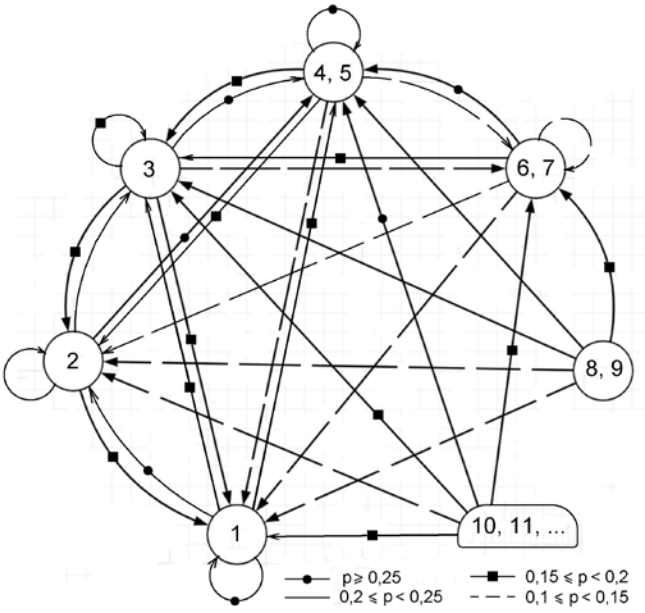

Рис. 4. Граф переходов вероятностного автомата.

На основе таблицы 2 построен граф переходов вероятностного автомата (рис. 4), который моделирует писателя. На нем показаны переходы с вероятностью $p \geq 0{,}1$. Из него, в частности, следует, что после СО длиной n=3 автор делает одно из следующих действий: 1) с вероятностью больше 0,25 пишет СО с m=4 или 5; 2) с вероятностью 0,15 – 0,2 пишет СО с m=3; 3) с вероятностью 0,15 – 0,2 пишет СО с m=2; 4) с вероятностью 0,15 – 0,2 пишет СО с m=1; 5) с вероятностью 0,1 – 0,15 пишет СО с m=6 или 7; 6) с вероятностью меньше 0,1 пишет СО с m=8, 9 или с m ⩾ 10 (эти переходы на графе не показаны).

III. Заключение

В статье рассмотрены методы получения частотных распределений различных букв, знаков препинания, слов и их сочетаний. Предложены три компьютерные программы на ABCPascal. На примере повести В.Г. Короленко «Дети подземелья» апробированы предлагаемые программы и проанализированы результаты. Получены частотные распределения букв, знаков препинания как зависимости их количества («яркости») от расстояния между ними (периода повторения). На основе вычисленной матрицы переходов построен вероятностный автомат, моделирующий порождение автором текста. Он представляет собой граф, вершины которого соответствуют количеству слов в смысловых отрезках текста, отделенных знаками препинания, а ребра – наиболее вероятным переходам. Все это характеризует идиостиль автора текста, его индивидуальные особенности и может быть использовано для установления авторства.

Приложение

Программа 1. Частотное распределение букв

```
uses crt; const chislo_strok=700;
var x,k,i,j,r,u,N,pp: integer;
a: array[0..5200]of string;
stroka: string; Vh,F: text;
w: array[0..350]of integer;
dx: array[0..10000]of integer;
BEGIN stroka:='а'; r:=0;
Assign(Vh,'c:\T1.txt'); Reset(Vh);
Assign(F,'c:\Vihod.txt'); Rewrite(F);
For k:=1 to chislo_strok do begin
Readln(Vh,a[k]); end;
For k:=1 to chislo_strok do begin
writeln(a[k],' ',k);  end;
For k:=1 to chislo_strok do begin if k>1
then pp:=pp+length(a[k-1]);
For j:=1 to length(a[k]) do If
stroka=copy(a[k],j,length(stroka))then
begin x:=j+pp;
inc(r); dx[r]:=x-u; u:=x; end; end;
For i:=1 to r do begin
  For j:=0 to 330 do If dx[i]=j then
inc(w[j]); end;
For i:=1 to 330 do writeln(F,w[i]);
writeln(F,r); Close(Vh); Close(F); END.
```

Программа 2. Распределение слов по длине

```
uses crt; const chislo_strok=2500;
alf='абвгдеёжзийклмнопрстуфхцчшщъыьэюя';
var  k,i,j,r,dl,pp,N,sovpalo: integer;
x:longint; Vh,F: text;
a: array[0..5200]of string; w:
array[0..250]of integer;
dlina: array[0..29200]of integer;
BEGIN Assign(Vh,'c:\T2.txt'); Reset(Vh);
Assign(F,'c:\Vihod.txt'); Rewrite(F);
For k:=1 to chislo_strok do begin
Readln(Vh,a[k]); end;
For k:=1 to chislo_strok do begin
a[k]:=' '+a[k]+' '; end;
For k:=1 to chislo_strok do For j:=1 to
length(a[k]) do begin sovpalo:=0;
For i:=1 to 33 do If copy(a[k],j,1)=
copy(alf,i,1) then sovpalo:=1;
If sovpalo=1 then inc(dl) else begin If
dl>0 then begin inc(r); dlina[r]:=dl;
end; dl:=0; end; end;
For i:=1 to r do begin
For j:=0 to 130 do If dlina[i]=j then
inc(w[j]); end;
For i:=0 to 130 do writeln(F,w[i]);
writeln(F,r);
Close(Vh); Close(F); END.
```

Программа 3. Распределение СО по длине

```
uses crt; const chislo_strok=2500;
alf='абвгдеёжзийклмнопрстуфхцчшщъыьэюя';
var k,i,j,r,r_max,bukva,chisl_slov,tt:
integer; x:longint; Vh,F: text;
z: array[0..150,0..150]of integer;
a,slovo: array[0..5200]of string;
chislo: array[0..30000]of integer;
w: array[0..235]of integer;
BEGIN Assign(Vh,'c:\T3.txt'); Reset(Vh);
Assign(F,'c:\Vihod.txt'); Rewrite(F);
For k:=1 to chislo_strok do begin
Readln(Vh,a[k]); end;
For k:=1 to chislo_strok do begin
a[k]:=' '+a[k]+' '; end; chisl_slov:=1;
For k:=1 to chislo_strok do For j:=1 to
length(a[k]) do begin
For i:=1 to 33 do If copy(a[k],j,1)=
copy(alf,i,1) then bukva:=1;
If (bukva=1)and(copy(a[k],j,1)=' ') then
inc(chisl_slov);
If (bukva=1)and(copy(a[k],j,1)='=') then
begin inc(r); chislo[r]:=chisl_slov;
bukva:=0; chisl_slov:=1; end; end;
r_max:=r;
For i:=1 to r do begin
For j:=0 to 130 do
If chislo[i]=j then inc(w[j]); end;
For r:=1 to r_max do begin
inc(z[chislo[r],chislo[r+1]]); inc(tt);
end;
For i:=1 to 22 do begin writeln(F,i);
For j:=1 to 22 do writeln(F,z[i,j]);
writeln(F,' ',tt);
end; Close(Vh); Close(F);
END.
```

БИБЛИОГРАФИЯ

[1] *Андриевская, Н.К.* Гибридная интеллектуальная мера оценки семантической близости // Проблемы искусственного интеллекта. – 2021. № 1. – С. 4-17.

[2] Белоногов, Г. Г. Компьютерная лингвистика и перспективные информационные технологии: теория и практика построения систем автомат. обраб. текстовой информ. / Г.Г. Белоногов, Ю.П. Калинин, А.А. Хорошилов. – Москва: Рус. мир, 2004. – 246 с.

[3] *Бородащенко, А.Ю.* Анализ текстов на семантическое сходство на основе аппарата теории графов // Известия ОрелГТУ. Серия "Информационные системы и технологии". 2008. № 1-2. С. 46-52.

[4] Короленко, В. Г. Дети подземелья. Махаон, 2022. 96 с.

[5] Куусела, Д. А. Анализ лексических спектров текстов с помощью математических методов // StudArctic Forum. 2023. Т. 8, № 2. С. 30 – 35.

[6] Москальчук, Г.Г., Манаков, Н.А. Форма текста как многоуровневый конструкт // Знание. Понимание. Умение. 2014 – №4. С. 291 – 302.

[7] Пиотровский, Р.Г., Бектаев, К.Б., Пиотровская, А.А. Математическая лингвистика: учеб. пособие для пед. ин-тов. М.: Высш. шк., 1977. 383 с.

[8] Радбиль, Т.Б., Маркина, М.В. Вероятностно-статистические модели в производстве автороведческой экспертизы русскоязычных текстов // Политическая лингвистика. 2019. № 2 (74). С. 156 – 166.

[9] Руднев Д.В., Друговейко-Должанская С.В. Распределение знаков препинания в современной деловой письменности // Язык и метод. Русский язык в лингвистических исследованиях XXI века. Вып. 7. Русская пунктуация в коммуникативном аспекте. Краков: Изд-во Ягеллонского ун-та, 2021. С. 75–86.

[10] Трубкина, А.И. Система периодических конструкций в языке и дискурсе: проблема статуса // Известия Сочинского государственного университета. 2013. № 3 (26). С. 251-254.

Frequency analysis of text using a computer

R.V. Mayer

Abstract **– The frequency analysis method of the text is considered, which leads to obtaining spectral distributions of letters, words, and semantic segments. The purpose of the article: 1) to create computer programs that allow you to obtain the spectra of the distribution of words and individual characters in large texts; 2) to test them in the analysis of V.G. Korolenko's novel "Children of the Dungeon"; 3) to build a probabilistic model of the writer. There are three programs in the ABCPascal language that allow to get: 1) the frequency distribution of letters and their combinations; 2) the spectral distribution of words and semantic segments; 3) the number of transitions from semantic segments of length n to semantic segments of length m. The article provides: 1) the spectral distributions of the vowels "o", "a", "e", "i", "u", "ya", "yu" in the analyzed text; 2) the frequency distribution of words along the length; 3) the spectrum of semantic segments of the text limited by punctuation marks; 4) the matrix of transitions from semantic segments of length n to semantic segments of length m; 5) the probabilities table of these transitions; 6) the graph of probabilistic automaton simulating the generation of text by the author. Its vertices correspond to the number of words in semantic text segments separated by punctuation marks, and the edges correspond to the most likely transitions. All this characterizes the individual characteristics of the style and can be used to establish authorship**.

Keywords **— probabilistic automaton, text, V.G. Korolenko, programming, stochastic matrix, frequency analysis.**

REFERENCES

[1] Andrievskaja, N.K. Gibridnaja intellektual'naja mera ocenki semanticheskoj blizosti // Problemy iskusstvennogo intellekta. – 2021. # 1. – S. 4-17.

[2] Belonogov, G. G. Komp'juternaja lingvistika i perspektivnye informacionnye tehnologii: teorija i praktika postroenija sistem avtomat. obrab. tekstovoj inform. / G.G. Belonogov, Ju.P. Kalinin, A.A. Horoshilov. – Moskva: Rus. mir, 2004. – 246 s.

[3] Borodashhenko, A.Ju. Analiz tekstov na semanticheskoe shodstvo na osnove apparata teorii grafov // Izvestija OrelGTU. Serija "Informacionnye sistemy i tehnologii". 2008. # 1-2. S. 46-52.

[4] Korolenko, V. G. Deti podzemel'ja. Mahaon, 2022. 96 s.

[5] Kuusela, D. A. Analiz leksicheskih spektrov tekstov s pomoshh'ju matematicheskih metodov // StudArctic Forum. 2023. T. 8, # 2. S. 30 – 35.

[6] Moskal'chuk, G.G., Manakov, N.A. Forma teksta kak mnogourovnevyj konstrukt // Znanie. Ponimanie. Umenie. 2014 – #4. S. 291 – 302.

[7] Piotrovskij, R.G., Bektaev, K.B., Piotrovskaja, A.A. Matematicheskaja lingvistika: ucheb. posobie dlja ped. in-tov. M.: Vyssh. shk., 1977. 383 s.

[8] Radbil', T.B., Markina, M.V. Verojatnostno-statisticheskie modeli v proizvodstve avtorovedcheskoj jekspertizy russkojazychnyh tekstov // Politicheskaja lingvistika. 2019. # 2 (74). S. 156 – 166.

[9] Rudnev D.V., Drugovejko-Dolzhanskaja S.V. Raspredelenie znakov prepinanija v sovremennoj delovoj pis'mennosti // Jazyk i metod. Russkij jazyk v lingvisticheskih issledovanijah XXI veka. Vyp. 7. Russkaja punktuacija v kommunikativnom aspekte. Krakov: Izd-vo Jagellonskogo un-ta, 2021. S. 75–86.

[10] Trubkina, A.I. Sistema periodicheskih konstrukcij v jazyke i diskurse: problema statusa // Izvestija Sochinskogo gosudarstvennogo universiteta. 2013. # 3 (26). S. 251-254.