

О применении компьютерных технологий для анализа художественных текстов (на примере произведений В.Г. Короленко)

Р.В. Майер

Аннотация — В статье рассмотрены некоторые возможности использования компьютерных технологий при анализе художественных текстов. Предметом исследования являются следующие произведения В. Г. Короленко: «Сон Макара», «Старый звонарь», «В ночь под Светлый праздник», «За иконой», «По пути», «Птицы небесные», «На Волге», «На затмении», «Слепой музыкант», написанные в период 1883 – 1898 гг. При этом были изучены: 1) зависимость количества словоупотреблений от ранга слова; 2) соответствие текста распределению Ципфа и правилу Парето; 3) особенности концептуализации важнейших концептов ДОБРО и ЗЛО, БОГАТСТВО и БЕДНОСТЬ; 4) проведение сентимент-анализа текстов, получение графика зависимости тональности слова от его позиции в тексте; 5) спектральный анализ зависимости тональности от позиции слов в тексте. Используются компьютерные программы, написанные в ABCPascal, и он-лайн ресурсы Интернета, в частности, нейросеть Qwen, позволяющая решать разнообразные задачи. В результате получены графики зависимостей тональности слов от позиции в тексте, а также спектр нерегулярных колебаний тональности вдоль текста. Показано, что интеллектуальный анализ текстов помогает установить статистические закономерности, связанные с чередованием периодических структур (слов, предложений, микротем и т.д.).

Ключевые понятия: закон Ципфа, Интернет, концептуализация, нейросеть, понятие, правило Парето, программирование, спектральный анализ, текст.

I. ВВЕДЕНИЕ

Проведение научных и учебных литературоведческих исследований часто требует анализа художественных текстов и определения их характеристик. С позиций Общей Теории Систем, любой текст представляет собой систему связанных и взаимодействующих между собой элементов: абзацев, предложений, слов, букв. С другой стороны, это многомерный информационный объект, который может быть описан целым рядом показателей. Художественный текст можно рассматривать как средство моделирования действительности, передачи смысла и эмоционального воздействия на читателя.

Статья получена 11 сентября 2025 г.

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

Развитие Интернета и искусственного интеллекта, в частности нейросетей, привело к появлению новых возможностей и методов анализа художественных текстов. Использование компьютерных технологий для обработки текста математическими методами и его кодирования позволяет поднять лингвистические исследования на более высокий уровень. Методы компьютерной лингвистики и автоматической обработки текстов обеспечивают решение широкого круга задач [1; 9]: извлечение сущностей, разбор предложений, определение частей речи, семантический анализ, машинный перевод, реферирование, анализ тональности, классификация и кластеризация текста, построение баз знаний и др.

Компьютерные технологии помогают извлечь из анализируемого текста количественные данные, которые можно сравнивать и визуализировать с помощью графов и диаграмм. Определенный интерес представляет собой вопрос о соответствии текста: 1) закону Ципфа, согласно которому частота использования слова обратно пропорциональна его рангу [8]; 2) правилу Парето 20/80, утверждающему, что на 20% самых часто употребляемых слов приходится около 80% словоупотреблений. Немалое значение имеет проблема оценки тональности (или сентимент-анализ), различных предложений и фрагментов текста [5]. При этом применяют методы спектрального анализа [3], основанные на изучении спектров взаимодействия букв, их сочетаний, слов, смысловых отрезков, что позволяет выявить особенности авторского стиля писателя, идентифицировать автора, установить жанр. Немаловажен вопрос об особенностях концептуализации ключевых понятий, таких как «добро» и «зло», «богатство» и «бедность» и т.д.

Цель статьи: изучить возможности применения компьютерных технологий для анализа художественных текстов (проверки закона Ципфа, правила Парето, анализа тональности, установления особенностей концептуализации ключевых понятий, выявления слов, ассоциирующихся с ключевыми концептами и т.д.).

Методологической основой являются работы следующих учёных: С. Д. Белов, Д. П. Зрелова, П. В. Зрелов и В. В. Кореньков [1], И. В. Богословская [2], С. Тичер, М. Мейер, Р. Водак и Е. Веттер [10], С. D. Manning, P. Raghavan & H. Schütze [13], K. L. Sumathy & M. Chidambaram [14] (автоматическая обработка текста); В. И. Бортников [4], А. В. Колмогорова, А. А.

Калинин и А. В. Маликова [6], Ю. Н. Марчук [9], М. W. Berry [12] (компьютерная лингвистика); Л. А. Борисов, А. Ю. Ивченко, Н. А. Митин и Ю. Н. Орлов [3] (спектральный метод); А. А. Двойникова и А. А. Карпов [5] (определение тональности текста); Ю. И. Манин [8] (закон Ципфа), Л. Н. Ясницкий [11] (искусственный интеллект).

II. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

1. Проверка выполняемости закона Ципфа и правила Парето. Согласно закону Ципфа, если все слова, встречающиеся в достаточно большом тексте, упорядочить по убыванию частоты (то есть проранжировать по частоте), с которой они встречаются в некотором корпусе текстов, то частота слова будет обратно пропорциональна его рангу [8]. Из закона Ципфа следует принцип Парето (или правило 80/20): пятая часть элементов системы обеспечивает четыре пятых всей суммы (то есть на 20% часто употребляемых слов приходится 80% словоупотреблений). Оба этих утверждения соответствуют неравномерному распределению и описывают закономерности в социальных, экономических и лингвистических системах.

Для проверки выполняемости правила Парето была проанализирована лексика следующих произведений В.Г.Короленко: «Сон Макара», «Старый звонарь», «В ночь под Светлый праздник», «За иконой», «По пути», «Птицы небесные», «На Волге», «На затмении», «Слепой музыкант» (1883 – 1898 гг.) [7]. При этом применялся следующий метод:

1. Объединяют все тексты анализируемых произведений в один текстовый файл text.txt, убирают из него различные комментарии и ссылки. Общий объем текста около 86000 слов.

2. Разделяют текст на 6 фрагментов объемом не более 100000 символов.

3. С помощью Интернет-сервиса advego.com анализируют эти фрагменты и получают списки слов с указанием количества их вхождений в анализируемые тексты. Интернет-ресурс превращает причастия и деепричастия в глаголы.

4. Используя специальную компьютерную программу КП-1, написанную в ABCPascal, объединяют полученные 6 списков слов в один так, чтобы в нем указывалось суммарное количество употреблений каждого слова в исходном файле text.txt.

5. С помощью редактора MS Word упорядочивают слова по алфавиту. Получают словарь В.Г. Короленко (файл slovar.txt) с указанием числа вхождений слов в анализируемые произведения. Изучая его, можно выделить группы слов, относящиеся к одинаковым темам: сельский быт, скитания, природа, мысли и чувства, религия и т.д.

6. Создают программу КП-2, которая, анализируя файл slovar.txt, выводит на экран или записывает в файл все слова, число употреблений которых превышает заданное пороговое значение F, а также их число W и общее количество словоупотреблений N.

7. Многократно запуская программу КП-2 при различных значениях F, определяют количество слов W, встречающихся чаще F, и соответствующее общее число словоупотреблений N. Заполняют таблицу в Excel, содержащую столбцы F, число слов W, количество словоупотреблений N.

8. Строят графики $W=W(F)$ и $N=N(\ln W)$.

Результаты представлены в табл. 1. Из неё видно, что 5 самых часто употребляемых слов (каждое из которых встречается более F=1200 раз) суммарно используются 9783 раза, 10 часто употребляемых слов (каждое – более F=800 раз) используются 14559 раз и т.д.

Таблица 1. Зависимость W и N от порога F.

F	W	N	ln(W)	F	W	N	ln(W)
3700	1	3724	0,000	400	21	20506	3,045
3000	1	3724	0,000	350	25	21977	3,219
2000	2	5809	0,693	300	29	23243	3,367
1500	2	5809	0,693	250	35	24945	3,555
1200	5	9783	1,609	200	48	27865	3,871
1000	6	10974	1,792	150	67	31113	4,205
900	8	12910	2,079	100	106	35849	4,663
800	10	14559	2,303	50	207	42996	5,333
700	12	16036	2,485	25	470	52326	6,153
600	13	16678	2,565	15	771	58206	6,648
500	16	18327	2,773	5	2028	69603	7,615

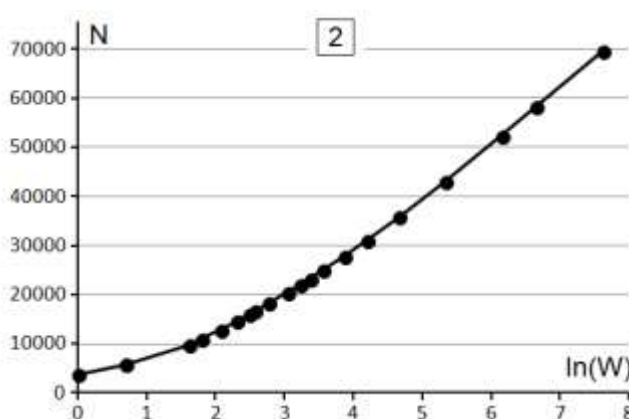
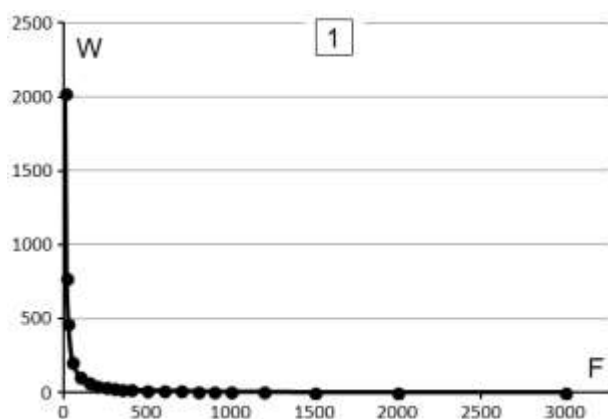


Рис. 1. Графики зависимостей $W=W(F)$ и $N=N(\ln W)$.

Построим график зависимости количества слов W , у которых частота превышает пороговое значение F , от значения F (рис. 1.1). Получается убывающая кривая, так как с ростом F число слов, встречающихся чаще F , уменьшается. На рис. 1.2 представлен график зависимости суммарного числа словоупотреблений N от $\ln(W)$, где W – количество первых слов, расположенных в порядке убывания частоты. При больших значениях N и W точки ложатся на прямую, то есть N прямо пропорционально $\ln(W)$. Если допустить, что частота слов связана с рангом зависимостью $n = 3538 / r^{0,798}$, то на компьютере можно рассчитать суммарное количество словоупотреблений W часто употребляемых слов:

$$N = \sum_{r=1}^W \frac{3538}{r^{0,798}}.$$

Получающийся при этом график N от $\ln(W)$ почти совпадает с графиком на рис. 1.2. Это доказывает, что словарь анализируемых произведений приблизительно соответствует закону Ципфа, который справедлив для «естественных текстов».

В рассматриваемых произведениях общее число словоупотреблений 84203, количество уникальных слов (включая союзы, предлоги, междометия и т.д.) – 9981. Правило Парето 20/80 приблизительно выполняется: на 20% самых часто употребляемых слов (1996) приходится около 69000 словоупотреблений, что составляет 82% от всех словоупотреблений. Следовательно, для изучения статистических свойств текста (например, определения долей существительных, прилагательных, глаголов) достаточно ограничиться учетом 20% всех слов, которые используются наиболее часто. В то же время на 2 % самых часто употребляемых слов (их 190) приходится 42097 словоупотреблений (половина от 84203), то есть столько же, сколько на оставшиеся 98 %.

2. Изучение особенностей концептуализации ключевых понятий. Важной проблемой анализа художественных произведений является выявление ключевых концептов и изучение особенностей их концептуализации. Последнее означает исследование формирования смысла понятия и построение его модели. Это требует выявления существенных признаков концепта, определения его границ, установления связей с другими понятиями в рамках анализируемого произведения. Рассмотрим решение обозначенной проблемы на примере перечисленных выше произведений В. Г. Короленко [7]. В качестве ключевых концептов выберем понятия ДОБРО, ЗЛО, БОГАТСТВО и БЕДНОСТЬ. Нами использовалась программа КП-3, которая считывала файл text.txt, содержащий тексты анализируемых произведений, выявляла слова, содержащие сочетания букв «добр», «зло», «зла», «богат», «бедн», «беден» и т.д. После этого она печатала в файл vyhod.txt строку символов, содержащую 10 букв до и 10 букв после найденного слова. Это позволило установить, с какими словами сочетаются рассматриваемые концепты в данных текстах. Получилось так:

Концепт ДОБРО: 1) хозяйское добро (ценные вещи), краденное добро; взвесить добро (сделанное человеком – 2 раза); 2) добрый человек (3), добрый поп(ик) (3),

добрый Роман, добрый католик, добрый голос, доброе лицо, доброе слово, доброе сердце (2), доброе семя, добрые звезды, добрые соседи, добрая тьма, добрая почва, добрая баба, он добр; 3) добродушный голос, добродушное лицо (2), добродушный послушник, добродушный кандыба, добродушные шутки, добродушная фигура, добродушная мешанка, добродушная улыбка, он добродушен, добродушно заговорил, добродушно сказал; 4) эти одобрения, простодушное одобрение, снисходительное одобрение; неодобрение пассивное; 5) одобрять нас (2); 6) своя добродетель, непоколебимая добродетель; 7) доброжелательство фамильярное; 8) добрейшая душа; 9) добровольный подвиг; 10) неодобрительно покачал.

Концепт ЗЛО: 1) взвесить зло (то есть оценить количество сделанного зла – 2 раза), не желать зла, отчеканить (то есть сказать) зло; 2) злоба дня, искаженное от злобы лицо, вложила злобу; 3) сказать со злостью (3), спросить со злостью, выть от злости, захлебываться от злости; 4) злая тайга, злая дряхлость, злая непогода, злая цель, что-то злое, злое страдание, злая воля, не злое лицо; 5) злобный крик, злобная дума, закончил злобно; 6) озлобление (2), беспредметная озлобленность, озлоблен; 7) злорадное любопытство; злорадно сказал; 8) незлобливое спокойствие; 9) злобно сжимает кулаки, злобно шипит; 10) зловещий голос, зловещий вид (2), зловеще сказал, зловеще крикнул, зловещее что-то; 11) незлобливый человек.

Концепт БЕДНОСТЬ, БЕДНЫЙ: бедный человек (3), бедный Макар (2), бедняга (2), бараксан (бедняга), поддал бедняге сзади, победное знамя, жил бедно, бедные комночи, бедные кости, бедная жизнь, бедная земля, бедный простор, бедная мать (2), бедный мальчик (2), бедная женщина, он беден.

Концепт БОГАТСТВО, БОГАТЫЙ: богатые меха, богатые хоромы (2 раза), богатый ворог, богату быть (2), чем богат, разбогатеть (2), живет богато, богатый монастырь, богатая нервная организация, богатство обширного регистра, богато одетый, богатая красками.

3. Выявление понятий, ассоциирующихся с важнейшими концептами. Возможен иной подход, заключающийся в использовании нейросети Qwen, для выявления слов, связанных с изучаемыми концептами ДОБРО и ЗЛО ассоциативно или парадигматически. В качестве примера рассмотрим анализ рассказа В.Г. Короленко «Старый звонарь». Задание для нейросети может выглядеть так: *“Дан текст: «Стемнело. Небольшое селение, приютившееся над дальнею речкой, в бору, тонULO в том особенном сумраке, которым полны весенние звездные ночи, когда тонкий туман, подымаясь с земли, сгущает тени лесов и застилает открытые пространства серебристо лазурною дымкой... Эй, посылайте на смену! Старый звонарь отзвонил...»*

Выяви и перечисли все слова из текста, ассоциирующиеся с концептами ДОБРО и ЗЛО. Используй примеры таких ассоциаций: ДОБРО ассоциируется с: свет, ангел, любовь, милосердие, жертвенность, честность, справедливость, мир, гармония, добродетель, награда, рай, защита слабых, альтруизм, покаяние, благодать, спасение, добро, добродетель, благо, милосердие, щедрость, сострадание, честность, справедливость,

порядочность, бескорыстие. ЗЛО ассоциируется с: тьма, демон, ненависть, жестокость, предательство, ложь, несправедливость, война, хаос, грех, наказание, ад, эгоизм, разрушение, искушение, проклятие, убийство, подлость, тиранья, зло, порок, жестокость, подлость, ненависть, зависть, алчность, ложь, предательство, несправедливость."

В ответ на этот запрос нейросеть отвечает, что «в произведении «Старый звонарь» ярко выражены духовные и моральные контрасты между добром и злом, светом и тьмой, жертвенностью и несправедливостью». Нейросеть также выдает:

1) слова, ассоциирующиеся с концептом ДОБРО (они связаны со «светом, верой, любовью, жертвенностью, честностью, спасением, красотой природы, надеждой, памятью о добрых людях»): «светят» (окна церкви светят — символ духовного света), «свет» (фонарик как "взлетевшая звезда" — метафора света среди тьмы), «звёздное небо» (ассоциируется с божественным, вечным, чистым), «Божьи звезды» (божественное присутствие), «горящее миллионами огней», «лазурь» (небесный свет, чистота, высота), «серебристо лазурная дымка» (красота, нежность, светлая таинственность), «приютившееся» (уют, защищённость), «молодые почки», «ароматный запах», «радостный клич: «Христос воскрес», «слава Те, Господи!», «молиться», «слушать пение», «детские голоса», «истово крестится», «цветущий здоровьем», «надежда на счастье», «радость жизни», «добрая была баба», «Царствие Небесное», «пройдена честно», «сырая земля — ему мать», «перекрестился» «пение», «любовная ласка», «окрыжен сыновьями и внуками», «радостные голоса», «счастье и радость», «восковые свечи», «торжественный звон», «пели и трепетали, смеялись и плакали», «чудная вереница», «звездное небо», «спасение», «покаяние».

2) слова, ассоциирующиеся с концептом ЗЛО (они связаны с «тьмой, смертью, страданием, несправедливостью, разрушением, утратой, физическим и моральным угасанием»): «темнота» (символ тьмы, неизвестности, смерти), «темная» (колокольня, вышка, уголок кладбища — ассоциируется с тьмой, одиночеством), «сумрак» (тьма, полумрак, предчувствие смерти), «убогие хаты» (бедность, нужда, страдание), «поизносился», «плохо видят глаза», «хоронил сыновей, хоронил внуков» «проводил в домовину» «все еще жив», «тяжело», «могилы», «вечного сна», «будет лежать... под крестом», «Бог не шлет смерти», «неисходное бабье горе», «иссушат красивую бабу», «потускнеют глаза», «вечного тупого испуга», «нужда да работа», «людская неправда», «осилила неправда», «богатый враг», «кровавые сиротские слезы», «замаливая», «кипит-разгорается сердце», «сурово глядят», «людское горе», «неправда», «Бог вас суди», «поникает седой головой», «слезы льются», «срам», «неуверенно дрогнул и смолк», «смущенные подголоски», «печально гудящую долгую ноту», «дрожит и льется, и плачет», «изнеможенно опустил», «бледные щеки», «старый пень, разбитый непогодой», «угрюмая и тесная вышка», «один», «вековой тупой испуг», «алчность, жестокость, предательство». Выявленные словосочетания помогают охарактеризовать авторский стиль В. Г. Короленко.

4. Анализ тональности художественных текстов.

Для проведения сентимент-анализа текста может быть использована нейросеть Qwen. Сущность метода заключается в следующем:

1. Текст загружают в окно запроса нейросети и требуют: 1) удалить все предлоги, союзы и междометия (в результате получается очищенный список значимых слов, расположенных по порядку); 2) выявить список эмоционально окрашенных слов или выражений и их координаты (порядковые номера слов, отсчитанные от начала текста).

2. Полученный список эмоционально окрашенных слов загружают в окно запроса и предлагают нейросети оценить тональность каждого слова или выражения по шкале от –1 (максимально отрицательная эмоция: горе, боль, смерть, отчаяние) до 1 (максимально положительная эмоция: радость, свет, любовь, надежда, святость).

3. Результаты переносят в электронные таблицы Excel и на их основе строят график зависимости тональности Т слов от их координаты х. Вычисляют среднее значение тональности по всему тексту.

Рассмотрим результаты сентимент-анализа рассказа В.Г. Короленко «Старый звонарь». Получившийся список слов с указанием их координат х и тональности Т частично представлен в табл. 2. В полном списке 173 слова, координата х изменяется от 4 до 1163. Нейросеть также обнаружила сложную эмоциональную палитру текста, в котором преобладают отрицательные и нейтральные эмоции (старость, утрата, грусть), но в кульминации наблюдается всплеск положительных эмоций (вера, радость, духовное просветление).

Таблица 2. Список слов с указанием координаты х и тональности Т.

	Слово или выражение	х	Т
1	приютившееся	4	0,6
2	тонуло	9	-0,3
3	особенном сумраке	11	-0,2
4	тонкий туман	18	-0,1
5	сгущает тени	22	-0,3
6	серебристо лазурною дымкой	29	0,8
7	тихо	32	0
8	задумчиво	33	-0,2
9	грустно	34	-0,8
10	дремлет	37	-0,1
11	убогие	38	-0,7
12	мерцают	45	0,4
13	чуткая	51	0,5
.....			
195	смущенные	1126	-0,7
196	печально гудящую	1135	-0,8
197	дрожит и льется, и плачет	1141	-0,9
198	изнеможенно	1148	-0,9
199	слезы катятся	1154	-0,8
200	бледным щекам	1156	-0,7
201	отзвонил	1163	0,7

На основе полученных данных был построен график зависимости тональности слов от их положения в тексте Т(х) (рис. 2). Из него видно, что тональность Т

совершает нерегулярные колебания в диапазоне от -1 до $+1$, в них присутствуют высокие и низкие частоты.

Нейросеть выделила: 1) Самые положительные выражения: «Христос воскрес» ($+1.0$), «надежды на счастье», «радости жизни», «цветущий здоровьем», «слава Те, Господи», «Божьи звезды» ($+0.9$); 2) Самые отрицательные выражения: «хоронил», «проводил в домовину», «муки», «непосильный труд», «вечный тупой испуг», «кровавые сиротские слезы», «неправда», «дрожит и льется, и плачет» (-0.9); 3) Смешанные / нейтральные: «смеялись и плакали» (0.0), «иллюзия счастья» ($+0.5$ – положительное чувство, но в

иллюзорной форме). При этом делается вывод, что «текст обладает глубоко трагическим фоном (старость, потеря, физическое угасание, социальная несправедливость), но в кульминации наблюдается мощный всплеск положительной духовной энергии через веру, звон и воскресение Христа. Это придаёт произведению «Старый звонарь» высокую эмоциональную напряжённость и духовное разрешение. Средняя тональность всего текста близка к нулю; в финале тональность изменяется в положительную сторону, что соответствует теме надежды на фоне страдания».

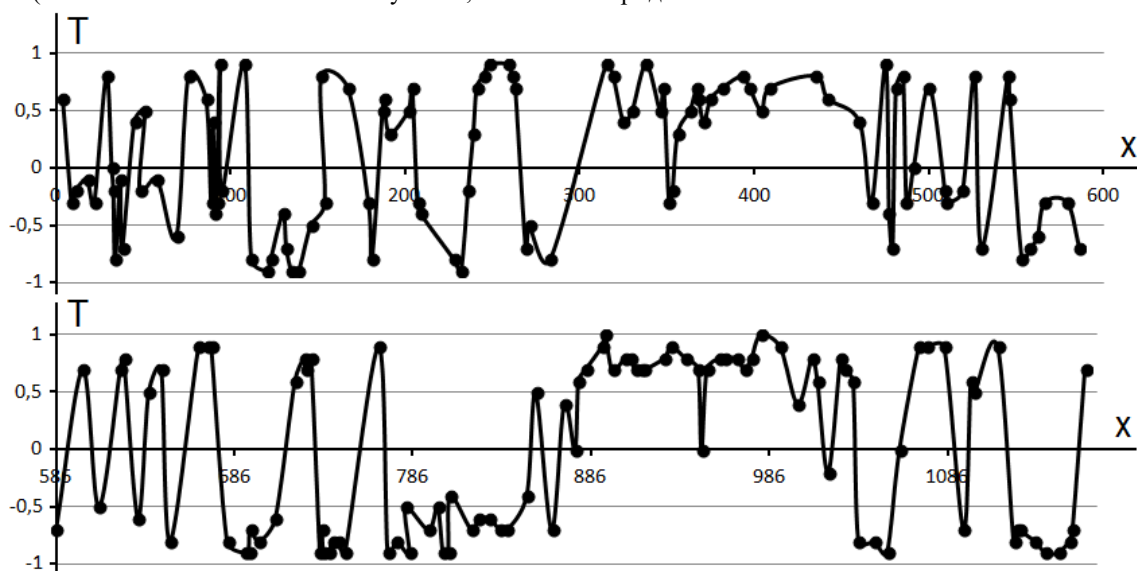


Рис. 2. Зависимость тональности от координаты слов и выражений.

С помощью нейросети текст «Старый звонарь» был разбит на смысловые блоки или логические микротемы, отражающие определенное настроение, образ, развитие повествования и т.д. Общее количество фрагментов, позволяющих точно передать структуру текста, – 21. Для каждого фрагмента были выписаны эмоционально окрашенные слова и определена их алгебраическая сумма $T(f)$, где f – номер фрагмента. При этом используется шкала: « $+1$ – максимальное добро, свет, радость, святость; 0 – нейтрально-эмоциональные или описательные выражения; -1 – глубокое зло, страдание, смерть, отчаяние.»

Ответ нейросети выглядит так: «1) Стемнело. Небольшое селение, приютившееся над дальнею речкой,

в бору, тонуло в том особенном сумраке, которым полны весенние звездные ночи, когда тонкий туман, подымаясь с земли, сгущает тени лесов и застилает открытые пространства серебристо лазурною дымкой... Все тихо, задумчиво, грустно. (Эмоционально окрашенные слова: приютившееся, $T = +0.6$; тонуло, $T = -0.3$; особенном сумраке, $T = -0.2$; тонкий туман, $T = -0.1$; сгущает тени, $T = -0.3$; серебристо лазурною дымкой, $T = +0.8$; тихо, $T = 0.0$; задумчиво, $T = -0.2$; грустно, $T = -0.8$). 2) Село тихо ...» Зависимость тональности T фрагмента от его номера F представлена на рис. 3. Хорошо видно, что тональность текста совершает колебания с периодом 11 – 13 фрагментов.

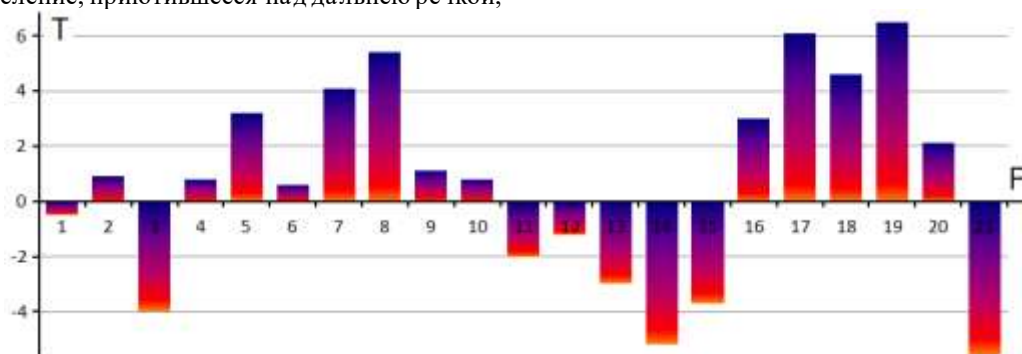


Рис. 3. Зависимость тональности от номера фрагмента $T(f)$.

Нейросеть отмечает, что текст выстроен как «духовное восхождение старика к последнему звону», при этом каждая микротема символизирует «шаг к завершению

жизненного пути», наполненный «памятью, болью, верой и торжеством души над телом». Финал звонаря не трагичен: «отзвонил» – значит, «жизнь прожита, служение завершено, покой наступил».

5. Спектральный анализ текста. Для анализа периодических функций (сигналов и т.д.) часто применяют метод спектрального анализа. Он заключается в определении характерных частот колебаний и соответствующих им амплитуд. Текст также может быть рассмотрен как периодический сигнал $y(x)$, которому соответствует последовательность значений некоторой величины y , изменяющихся вдоль оси «позиция в тексте» x . В качестве переменной y может использоваться эмоциональная тональность, частота употребления ключевых слов, сложность предложений (длина, количество придаточных), тематическая насыщенность, интенсивность образов (свет/тьма, добро/зло и т.д.).

Метод спектрального анализа текста заключается в следующем: 1) текст разбивают на элементы или фрагменты (слова, предложения, абзацы); 2) у каждого фрагмента определяют числовое значение признака y ; 3) получают функцию $y = y(x)$, где x – порядковый номер элемента, $y(x)$ – значение признака; 4) применяют преобразование Фурье, вычисляют амплитудный спектр, определяют преобладающие частоты (периоды, циклы); 5) интерпретируют полученные результаты, исходя из того, что «спектральным линиям» соответствуют повторяющиеся структуры в тексте. Например, если измерять длины предложений L , то спектральные линии будут соответствовать ритмическим повторениям этой характеристики. В некоторых случаях можно сделать вывод, что автор использует волновой ритм, чередуя короткие и длинные предложения с периодом 4-8 предложений. Если длины предложений одинаковые, то спектр текста будет состоять из одной спектральной линии.

Применим метод спектрального анализа к функции $T = T(x)$, график которой приведен на рис. 2. Для этого в окно запроса нейросети Qwen загрузим табл. 2 (201 пар значений) и сформулируем задание: «Выполни спектральный анализ заданной функции $y = y(x)$ и определи амплитуды A первых 15 гармоник».

Нейросеть выдаёт следующий результат: $f_1 = 0,012295$, $A_1 = 1,2842$; $f_2 = 0,024591$, $A_2 = 1,2588$; $f_3 = 0,036887$, $A_3 = 1,1698$; $f_4 = 0,006148$, $A_4 = 1,0746$; $f_5 = 0,018443$, $A_5 = 1,0333$; $f_6 = 0,030739$, $A_6 = 0,9176$; $f_7 = 0,043035$, $A_7 = 0,8684$; $f_8 = 0,009838$, $A_8 = 0,8121$; $f_9 = 0,049180$, $A_9 = 0,7544$; $f_{10} = 0,021885$, $A_{10} = 0,7196$; $f_{11} = 0,034181$, $A_{11} = 0,6837$; $f_{12} = 0,003074$, $A_{12} = 0,6403$; $f_{13} = 0,037481$, $A_{13} = 0,6321$; $f_{14} = 0,01537$, $A_{14} = 0,6278$; $f_{15} = 0,027666$, $A_{15} = 0,6186$. Частота f измеряется в 1/слов. Получающийся спектр представлен на рис. 4.

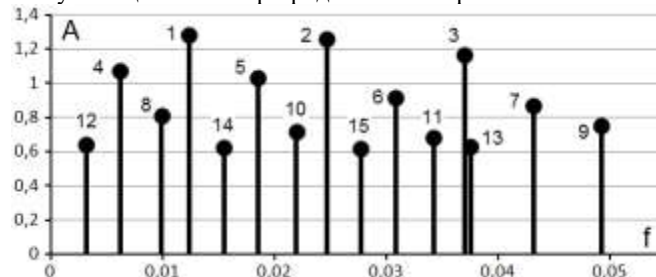


Рис. 4. Спектр функции $T=T(x)$.

Основная частота $f_1 = 0,012295$ 1/слов, период 81,3 слова, что соответствует временному циклу появления

слов той или иной тональности. Чем больше амплитуда, тем в большей степени колебания соответствующей частоты присутствуют в тексте.

III. ЗАКЛЮЧЕНИЕ

В статье продемонстрированы некоторые возможности использования информационных технологий для анализа художественных произведений. В частности изучены: 1) зависимость количества словоупотреблений от ранга слова; 2) соответствие текста распределению Ципфа и принципу Парето; 3) особенности концептуализации важнейших концептов; 4) анализ тональности художественных произведений при делении текста на значимые слова и микротемы; 5) получение спектра зависимости тональности слова от его позиции в тексте. Показано, что интеллектуальный анализ текстов (с помощью компьютерных программ, он-лайн ресурсов и нейросетей) позволяет решить разнообразные задачи, требующие выявления слов и словосочетаний с заданными признаками, оценки их характеристик, установления статистических закономерностей чередования периодических структур (слов, предложений, эмоционально окрашенных выражений, микротем и т.д.).

БИБЛИОГРАФИЯ

- [1] Белов С. Д. и др. Обзор методов автоматической обработки текстов на естественном языке / С. Д. Белов, Д. П. Зрелова, П. В. Зрелов, В. В. Кореньков // Системный анализ в науке и образовании: сетевое научное издание. – 2020. № 3. – С. 8–22.
- [2] Богословская И. В. Количественный анализ результатов автоматизированной обработки текстов // Теория языка и межкультурная коммуникация. Курский гос. университет. – 2019. №4. – С. 32–38.
- [3] Борисов Л. А., Ивченко А. Ю., Митин Н. А., Орлов Ю. Н. Тематическая классификация текстов с помощью спектральных портретов // Препринты ИПМ им. М.В.Келдыша. 2017. № 106. – 22 с.
- [4] Бортников В. И. Лингвистический анализ текста: учебно-методическое пособие. – Екатеринбург: Изд-во Урал. ун-та, 2020. – 112 с.
- [5] Двойникова А. А., Карпов А. А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных. Информационно-управляющие системы, 2020, № 4, с. 20–30.
- [6] Колмогорова А. В., Калинин А. А., Маликова А. В. Лингвистические принципы и методы компьютерной лингвистики для решения задач sentiment-анализа русскоязычных текстов // Актуальные проблемы филологии и педагогической лингвистики. 2018. №1 (29). С. 139–148.
- [7] Короленко В. Г. Повести, рассказы, очерки: сборник. – Москва: Московский рабочий, 1977. – 260 с.
- [8] Манин Ю. И. Закон Ципфа и вероятностные распределения Левина // Функциональный анализ и его приложения. 2014, том 48, выпуск 2. С. 51–66.
- [9] Марчук Ю. Н. Компьютерная лингвистика: учеб. пособие. – М.: АСТ: Восток-Запад, 2007. – 317 с.
- [10] Тичер С., Мейер М., Водак Р., Веттер Е. Методы анализа текста и дискурса. Харьков: Гуманитарный центр, 2009. – 356 с.
- [11] Ясницкий Л. Н. Введение в искусственный интеллект. Учеб. пособие для ВУЗов. – М.: Изд. центр «Академия», 2005. – 176 с.
- [12] Berry M. W. Survey of Text Mining: Clustering, Classification and Retrieval. – Springer-Verlag, New York, Inc., 2004. p. 244.
- [13] Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval. – Cambridge University Press, 2008. – 520 p.
- [14] Sumathy K. L., Chidambaram M. Text Mining: Concepts, Applications, Tools and Issues – An Overview International Journal of Computer Applications, Vol. 80. No. 4. 2013. pp. 29–32.

On the use of computer technology for the analysis of literary texts (on the example of V.G. Korolenko's works)

R.V. Mayer

Abstract – The article discusses some of the possibilities of using computer technology in the study of literary texts. The subject of the research are the following works by V.G. Korolenko: "Makar's Dream", "The Old Bell Ringer", "On the night of a Bright Holiday", "Behind the Icon", "Along the Way", "Birds of the Sky", "On Volga", "On the Eclipse", "The Blind Musician", written by in the period 1883-1898. At the same time, the following are studied: 1) the dependence of the number of word uses on the rank of the word; 2) the correspondence of the text to the Zipf distribution and the Pareto rule; 3) the features of conceptualization of the most important concepts of GOOD and EVIL, WEALTH and POVERTY; 4) conducting a sentimental analysis of texts, obtaining a graph of the dependence of the word tonality on its position in the text; 5) the spectral analysis of the dependence of tonality on the words position in the text. Computer programs written in ABCPascal and online Internet resources (in particular, the Qwen neural network) are used which allows solving various tasks. As a result, graphs of the dependence of the word tonality on the position in the text, as well as a spectrum of irregular fluctuations of tonality along the text, are obtained. It is shown that the intellectual analysis of texts helps to establish statistical patterns associated with the alternation of periodic structures (words, sentences, microthemes, etc.).

Keywords: Zipf's law, Internet, conceptualization, neural network, concept, Pareto rule, programming, spectral analysis, text.

REFERENCES

- [1] Belov S. D. i dr. Obzor metodov avtomaticheskoy obrabotki tekstov na estestvennom jazyke / S. D. Belov, D. P. Zrel'ova, P. V. Zrel'ov, V. V. Koren'kov // Sistemyj analiz v nauke i obrazovanii: setevoe nauchnoe izdanie. – 2020. # 3. – С. 8–22.
- [2] Bogoslovskaja I. V. Kolichestvennyj analiz rezul'tatov avtomatizirovannoj obrabotki tekstov // Teorija jazyka i mezhkul'turnaja kommunikacija. Kurskij gos. universitet. – 2019. #4. – S. 32-38.
- [3] Borisov L. A., Ivchenko A. Ju., Mitin N. A., Orlov Ju. N. Tematicheskaja klassifikacija tekstov s pomoshh'ju spektral'nyh portretov // Preprinty IPM im. M.V.Keldysa. 2017. # 106. – 22 s.
- [4] Bortnikov V. I. Lingvisticheskij analiz teksta: uchebno-metodicheskoe posobie. – Ekaterinburg: Izd-vo Ural. un-ta, 2020. – 112 s.
- [5] Dvojn'kova A. A., Karpov A. A. Analiticheskij obzor podhodov k raspoznavaniju tonal'nosti russkojazychnyh tekstovych dannyh. Informacionno- upravljajushhie sistemy, 2020, # 4, s. 20–30.
- [6] Kolmogorova A. V., Kalinin A. A., Malikova A. V. Lingvisticheskie principy i metody komp'yuternoj lingvistiki dlja reshenija zadach sentiment-analizarusskojazychnyh tekstov // Aktual'nye problemy filologii i pedagogicheskoy lingvistiki. 2018. #1(29). S. 139-148.
- [7] Korolenko V. G. Povesti, rasskazy, ocherki: sbornik. – Moskva: Moskovskij rabochij, 1977. – 260 s.
- [8] Manin Ju. I. Zakon Cipfa i verojatnostnye raspredelenija Le vina // Funkcional'nyj analiz i ego prilozhenija. 2014, tom 48, vypusk 2. S. 51–66.
- [9] Marchuk Ju. N. Komp'yuternaja lingvistika: ucheb. posobie. – M.: AST; Vostok-Zapad, 2007. – 317 s.
- [10] Ticher S., Mejer M., Vodak R., Vetter E. Metody analiza teksta i diskursa. Har'kov: Gumanitamyj centr, 2009. — 356 c.
- [11] Jasn'ckij L. N. Vvedenie v iskusstvennyj intellekt. Ucheb. posobie dlja VUZov. – M.: Izd. centr «Akademija», 2005. – 176 s.
- [12] Berry M. W. Survey of Text Mining: Clustering, Classification and Retrieval. – Springer-Verlag, New York, Inc., 2004. p. 244.
- [13] Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval. – Cambridge University Press. 2008. – 520 p.
- [14] Sumathy K. L., Chidambaram M. Text Mining: Concepts, Applications, Tools and Issues – An Overview International Journal of Computer Applications, Vol. 80. No.4. 2013. pp. 29-32.