РАЗДЕЛ IV. ИЗМЕРЕНИЕ, МОДЕЛИРОВАНИЕ И УПРАВЛЕНИЕ В ЭКОЛОГИИ, НАУКАХ О ЧЕЛОВЕКЕ И ОБЩЕСТВЕ

УДК 004.934:004.912:008:001.8:519.76

МЕТОД ИЗМЕРЕНИЯ АВТОРСКИХ И КУЛЬТУРНО-ЯЗЫКОВЫХ ОСОБЕННОСТЕЙ ТЕКСТОЛОГИЧЕСКОГО МАТЕРИАЛА НА ОСНОВЕ ЧАСТОТНЫХ ХАРАКТЕРИСТИК

П.И. Балабанов, О.В. Головань

В статье описан оригинальный метод измерения характеристик художественных текстов, основанный на установлении количественных параметров функциональной связи между частотой и рангом слов в тексте. Приведено описание программного комплекса для исследования частотных характеристик текстов (программы LangFracDim, C-Analysis и базы данных DB Language Fractal Dimension) и результаты исследований авторских и культурно-языковых особенностей текстов произведений В.М. Гаршина этим методом.

Ключевые слова: математическая лингвистика, частота и ранг слова, закон Ципфа

Количественные методы исследования текстового материала часто используются в прикладных областях лингвистики и основываются на проведении различных подсчетов и косвенных измерениях единиц любого уровня. Поскольку количественные методы в лингвистике часто непосредственно опираются на математическую статистику, то часто их называют также статистическими методами [1, 2].

Основным объектом применения количественных методов является сам текст, как объединенная смысловой связью последовательность знаковых единиц (слов, предлогов, знаков препинания, спец. символов), и его количественные характеристики могут дать определенную информацию, как о самом тексте (объем, количество и доля различных частей речи, количество лексически значимых символов и др.), так и о его характеристиках, связанных с индивидуальными особенностями автора, стиля текста, вкладываемого в него смысла и информации, а также особенностями самого языка, используемого для порождения текста, который, в свою очередь, претерпевает различные трансформации под воздействием объективных культурных и исторических процессов [3].

Сложность и многоаспектность рассмотренных отношений между текстом и влияющими на него факторами сильно усложняет задачу количественного исследования особенностей различного текстового материала, особенно, установления степени влияния отдельных факторов на статистические закономерности текста. Однако использование текстологического подхода, позволяющего решить проблему атрибуции текста на основе контент-анализа и методов психолингвистики, дает возможность исключить из рассмотрения случайные факторы, влияющие на текст и исследовать только те его особенности, которые связаны с объективными причинами трансформации языка, как в пространстве культуры, так и в сфере индивидуального сознания автора [4].

Таким образом, результаты количественных исследований текстов, история возникновения и судьба которых определены однозначно, могут лечь в основу одного из методов измерения их авторских или культурно-языковых особенностей.

В настоящей статье описан один из вариантов количественного метода исследования такого материала, основанный на выявлении в нем определенных частотных закономерностей измерении их характеристик.

Постановка задачи

В основе предлагаемого нами метода лежит исследование обширных массивов текста, предполагающее организацию многократного доступа к их содержанию посредством электронных баз данных (БД) и специализированного программного обеспечения для их обслуживания, поэтому для его реализации ранее нами был разработан комплекс

программ для ЭВМ, позволяющий производить анализ и разбор текстов на отдельные слова (лексемы), занесение их в БД, извлечение из БД по определенным признакам, автоматическое пополнение БД при увеличении объема исследованного текстологического материала, определение авторского и респондентского смысла, вкладываемого в отдельные лексемы, а также проводить определение квантитативных (частотных) характеристик, как отдельных текстов, так и объединенных баз [5].

Программная реализация

В комплекс входят программы для ЭВМ «Фрактальная размерность языка (LangFracDim)» и «Концепт-анализ (C-Analysis)», а также специализированная база данных «БД Фрактальная размерность языка (DB Language Fractal Dimension)» [6 - 8].

Программа «Фрактальная размерность языка» организована по типу анализатора текста, а программа «Концепт-анализ» - опросника. Исходными объектами являются текстовые файлы, слова из которых, после обработки программой, заносятся в базу данных БД (DB Language Fractal Dimension) через используемый нами SQL-сервер Fire Bird v. 1.0 [http://firebird.sourceforge.net/].

Программа LangFracDim предназначена для определения частотных характеристик текста или языка и корреляций между ними. Определение производится путем составления частотного словаря языка либо отдельно взятого текста и определения ранга и частоты слов. Программа позволяет проводить группировку слов в базе по определенным признакам.

Определение корреляций между частотными характеристиками основывается на установлении зависимости между частотой и рангом слова по закону Ципфа [9], параметры которой, после линеаризации, определяются методом наименьших квадратов (МНК):

$$Ч = k \cdot P^{\alpha}, \qquad (1)$$

где $Ч$ - частота встречаемости слов в тексте; k - коэффициент пропорциональности; P - ранг слов; α - степень развитости и наполняемости текста различными лексическими единицами (для корпусов большинства современных и естественного языка близка к –1).

Специальная организация работы основного алгоритма программы и ее взаимодействия с БД позволяет ставить в соответствие каждому слову его уникальный номер (**ID**), связанный с темой и словарем, что позволяет избегать повтора одинаковых слов и ускорять извлечение слова из БД SQL-сервером, а связь ID с определенной темой или словарем не только ускоряет работу комплекса ПО - БД, но позволяет осуществлять программе «LangFracDim» оперативное построение таблиц и графиков зависимостей частоты от ранга для слов определенной темы, словаря или общего списка слов, находящихся в БД (рисунок 1).

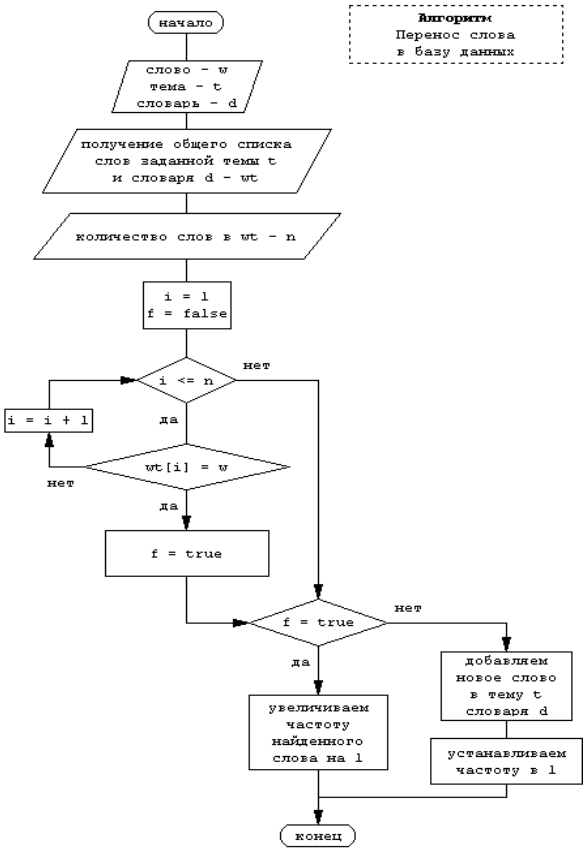

Рисунок 1 - Блок-схема основного алгоритма переноса слова в БД

Результаты и их обсуждение

Для иллюстрации возможностей разработанного метода рассмотрим результаты частотного анализа текстов произведений В.М. Гаршина программой «Фрактальная размерность языка», и их сопоставление с полученными ранее результатами психолингвистического эксперимента [10] для установления главных компонентов культурно-языкового концепта «трагическое».

Так в рассказе «Четыре дня» (1877) среди всех 1538 лексем, использованных автором в тексте, ядерные компоненты «трагического», проявившиеся в психолингвистическом эксперименте [10]: «война», «воспоминание», «судьба», «воображение» встречаются только 6 (!) раз, в то время как перифе-

рийные - «убийство», «кровь», «враг» (в значении описания солдата противоборствующей стороны), «раненый», «жизнь» – 10 (!) раз, то есть составляют 0,39 % и 0,65 % от всех использованных писателем слов, соответственно.

Отдельные же частоты и ранги словрепрезентантов (*Ч; Р*) «трагического» по результатам обработки текста следующие: «война» (1; 1348), «воспоминание» (3; 191), «судьба» (1; 801), «воображение» (1; 1221), «убийство» (1; 1040), «кровь» (2; 334), «враг-турок» (4; 108), «раненый» (2; 303) и «жизнь» (6; 72).

Обнаруженный нами результат на первый взгляд может выглядеть парадоксально, особенно в свете известной гипотезы А. Вежбицкой о частотной значимости культурнозначимых слов языка [11]. Однако здесь мы имеем дело не с общим корпусом текстов русского языка, а с конкретным художественным произведением, автор которого – большой мастер слова, используя различные литературные приемы, погружает читателя в трагические переживания главного героя.

На соответствующем логарифмическом графике зависимости частоты, как функции ранга слова (1), полученном для рассказа «Четыре дня», выявленные при когнитивном эксперименте репрезентанты «трагического» мы также обнаружим в разных областях частотограммы (рисунок 2).

Высокие частоты и низкие ранги в тексте рассмотренного рассказа имеют различные служебные части речи, от частоты 86 (частица «не») вплоть до частоты 12 смыслозначимых лексем в тексте не обнаруживается. Первые смыслозначимые лексемы «быть» и «здесь» имеют частоту употребления в тексте рассказа 12.

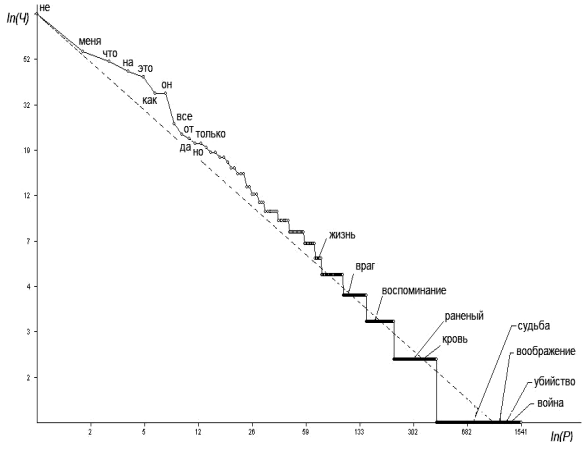

Рисунок 2 - Частотограмма рассказа В.М. Гаршина «Четыре дня»

На частотограмме можно обнаружить и некоторые характеристические элементы. Так, начиная с частотности 12, как это следует из приведенного графика, на зависимости Ч=f(Р) проявляются «площадки» - группы слов имеющих одинаковую частоту использования в тексте, но различный ранг.

Например, в группе слов с одинаковой частотой 12 в рассказе «Четыре дня» находятся лексемы: «что-то», «него» и «быть». На различных частотах эти площадки имеют различную протяженность, которая достигает максимального значения в самой многочисленной группе слов с частотой использования 1, однако количество одночастотных слов в группах изменяется неравномерно. Так, в приведенном примере хорошо видно, что в диапазоне частот от 12 до 10, количество таких слов, и соответственно протяженность площадок, растет, затем, в интервале от 10 до 7 уменьшается, проходит через минимум на частоте 6, потом снова увеличивается и растет вплоть до частоты 1. То есть функция количества одночастотных слов текста от частоты убывающая, негладкая и имеет экстремумы на своих участках.

Рассмотрим поведение этой функции в нашем примере. При частотах от 86 до 22 каждому значению частоты соответствует одно слово. В интервале от 21 до 12 – 1,2 слова. С частотой 11 употребляется 3 слова, 10 и 9 – 7 слов, 8 – 12 слов, с частотой 7 – 10 слов. В рассмотренном примере координаты экстремума функции Ч=f(Р) соответствуют частоте 6. В самой малочисленной группе слов с этой частотой использования и на соответствующей площадке находятся 7 лексем: «вдруг», «мой», «больше», «равно», «жизнь», «ты» и «чем». Как видим, в этой экстремальной группе находится и репрезентант «трагического» - «жизнь», имеющий самую высокую частоту использования среди всех остальных, обнаруженных нами ранее в психолингвистическом эксперименте [10].

В группе слов с частотой 5 находятся уже 29 лексем, в группе с частотой 4 – 43 лексемы, среди которых репрезентант «трагического» - «враг», в группе с частотой 3 – 76 лексем, среди которых еще один репрезентант – «воспоминание», с частотой 2 - 203, тут находятся репрезентанты «раненый» и «кровь», и, наконец, в группе с частотой 1 находится уже 1111 лексем – 72,2 % всех слов текста, среди которых остальные репрезентанты «трагического», обнаруженные нами и в психолингвистическом эксперименте: «судьба», «воображение», «убийство» и «война».

При пересчете доли (%) словрепрезентантов «трагического» в группах остальных лексем, имеющих с ними одинаковую частотность получим следующие значения: «жизнь» – 14,28; «враг» - 2,32; «воспоминание» - 1,31; «раненый» и «кровь» - 0,98; «судьба», «воображение», «убийство» и «война» - 0,36 %. При такой стратегии частотной интерпретации значимости словрепрезентантов «трагического» в тексте она возрастает многократно. Например, «жизнь» в 36 (!) раз (с 0,39 до 14,28 %), а «война» в 6 раз (с 0,06 до 0,36 %).

Функциональная зависимость Ч=f(Р) для рассказа «Четыре дня» в логарифмических координатах апроксимируется прямой (пунктирная линия тренда на рисунок 2), описываемой следующим уравнением:

$$ln(Ч) = -0{,}6238 \times ln(P) + 4{,}2906, \qquad (2)$$

с $\sigma(x)$ экспериментальных точек от теоретической кривой 0,0832 и коэффициентом линейной корреляции r = -0,9024, свидетельствующими о существовании сильной обратной связи между частотой и рангом слов в произведении.

Нами были исследованы и другие произведения этого писателя, так или иначе связанные с трагическими событиями конца XIX в. в России: очерк «Очень коротенький роман» (1878); рассказ «Трус» (1879); рассказ «Из воспоминаний рядового Иванова» (1883) и автобиографический очерк «Аяслярское дело» [12].

Если расположить найденные нами коэффициенты пропорциональности и показатели степени в уравнении закона Ципфа (1) для четырех произведений в порядке возрастания общего количества лексем (N) в соответствующих текстах (813, 1538, 2018, 2525, 5225), получим следующие ряды: для k 19,4531, 73,0102, 54,4132, 105,4045, 191,8485; для α -0,4812, -0,6238, -0,5604, -0,6335, -0,6495 (рисунок 3).

Из рисунка 3 виден в целом согласованный характер изменения первой величины, и асимптотическое изменение показателя степени α в законе Ципфа, который для рассмотренных нами произведений одного автора находится в интервале от -0,4812 до -0,6495, приближаясь к последнему значению.

Заключение

Таким образом, авторские особенности текстов количественно проявляются в показателе степени в уравнении закона Ципфа для тематической выборки произведений, что подтверждается предельным характером зависимости $\alpha = f(N)$.

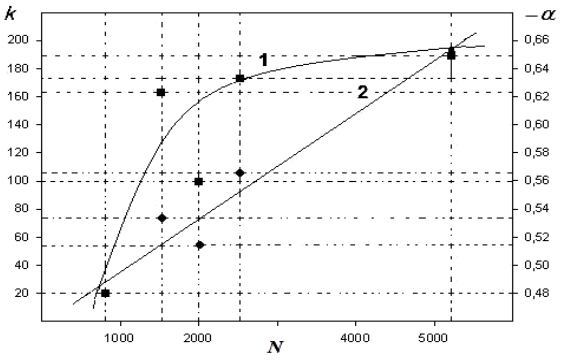

Рисунок 3 - Зависимость коэффициентов α (1) и k (2) в уравнении закона Ципфа от количества лексем для произведений В.М. Гаршина, репрезентирующих «трагическое»

Обнаруженные и зафиксированные в художественных текстах В.М. Гаршина составляющие культурно-языкового концепта «трагическое» определяют его функционирование в русской языковой культуре конца XIX в. и дают возможность исследования культурного образа мира этого времени, отражают особенности содержание культуры и языка.

СПИСОК ЛИТЕРАТУРЫ

1. Андреев, Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении/ Н.Д. Андреев. – СПб.: 1997.
2. Гладкий, А.В. Элементы математической лингвистики./ А.В.Гладкий, И.А.Мельчук –М.: 1989.
3. Лотман Ю.М. Структура художественного текста./ Ю.М. Лотман. – М., 1970.
4. Гришунин, А. Л. Исследовательские аспекты текстологии./ А. Л. Гришунин. -М., 1998.
5. Головань, О.В. // О.В. Головань.- Ползуновский альманах. 2008. № 2. С. 153-156.
6. Св-во № 2005610982 (RU) от 22.04.2005 / Головань О.В. // Опубл. Бюлл. № 2. 2005.
7. Св-во № 2005611226 (RU) от 25.05.2005 / Головань О.В., Барсуков А.А. // Опубл. Бюлл. № 2. 2005.
8. Св-во № 2005620308 (RU) от 28.11.2005 / Головань О.В. // Опубл. Бюлл. № 4. 2005.
9. Zipf G.K. The psycho-biology of language. / G.K. Zipf . -Boston, 1935.
10. Головань, О.В. Семантико-ассоциативная структура концепта «война»./ О.В. Головань – Барнаул, 2001
11. Вежбицкая, А. Язык. Культура. Познание. / А. Вежбицкая– М.: 1997.
12. Гаршин, В.М. Сочинения./ В.М. Гаршин - Москва, 1955.

*Д.ф.н.,профессор **Балабанов П.И.**, тел. (3842) 35-95-03, Кемеровский государственный университет культуры и искусств, НИИ прикладной культурологии; к.ф.н., доцент, докторант **Головань О.В.**, e-mail: oleg6888@rambler.ru (г. Кемерово).*