Компьютерный анализ текстов

Применение вычислительных методов корпусного анализа к исследованию текстов литературных произведений

Н.Л. Аванесян[I], О.В. Губина[II], А.М. Чеповский[I,II]

[I] Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия
[II] Российский университет дружбы народов им. Патриса Лумумбы, г. Москва, Россия

Аннотация. Статья посвящена применению математических методов корпусного анализа для исследований литературных текстов русских писателей. Для исследований создан корпус прозаических текстов художественной литературы России XIX века, состоящий из пяти подкорпусов. Каждый подкорпус содержит тексты одного из авторов. На примере созданного корпуса продемонстрированы возможности применения метода анализа соответствий, интегрированного в корпусную платформу TXM в качестве одного из средств статистического метода исследований. В качестве другого метода рассматривается анализ коэффициентов попарной ранговой корреляции для сравнения частотных характеристик текстов различных подкорпусов. Описанные методики дают коррелированные результаты и позволяют констатировать возможность выделения дифференцирующих признаков. Они могут использоваться как для лингвистических исследований, так и создания корректных обучающих текстовых наборов для задач искусственного интеллекта.

Ключевые слова: *корпусная лингвистика, платформа TXM, анализ соответствий, корреляционный анализ.*

DOI: 10.14357/20790279240204 **EDN: IKHGUO**

Введение

С развитием тенденции на машинное обучение для систем искусственного интеллекта возрастает необходимость в создании большего числа подходящих для данной задачи выборок данных. Корпусная лингвистика может служить средством создания специализированных коллекций текстов для машинного обучения в области обработки естественного языка, решения задач, относящихся к проблемам классификации текстов (распознавания образов). Корпусы литературных текстов предоставляют обширный объем информации и помогают в построении моделей для задач искусственного интеллекта.

Методы статистических исследований применялись для анализа корпусов текстов террористической и антиправовой направленности, описанные в [1–3], для анализа корпусов текстов пользователей социальных сетей [4,5].

В работах в [1,2 и 4] для корпусных исследований применялась платформа TXM ([6,7]), которая представляет собой программное обеспечение с открытым кодом, предназначенное для подготовки, обработки, анализа и публика-

ции корпусов среднего размера (примерно до 10 000 000 словоупотреблений). Платформа TXM развивается в лаборатории IHRIM Высшей нормальной школы Лиона и Национального центра научных исследований Франции (CNRS & ENS de Lyon, France) [7].

Корпусная платформа TXM дает возможность быстро и достоверно решать задачи, связанные с анализом текста. Морфологическая разметка осуществляется автоматически при загрузке текстов на платформу TXM. Используемые процедуры морфологического анализа, описанные в [8], включают снятие омонимии в корпусе. Корпусная платформа TXM позволяет добавлять нестандартные для корпусных средств методы морфологического анализа [9]. В платформу TXM интегрированы математические процедуры для статистических исследований, такие как анализ специфичности и анализ соответствий, описанный в [10,11].

В работах в [3 и 5] для анализа корпусов текстов применялся корреляционный ранговый анализ словарей, составленных для объединенных наборов текстов, которые можно рассматривать как специальные текстовые корпуса. Составление словарей для данного исследования основано на оригинальных процедурах морфологического анализа, описанных в [12].

В работе [13] процедуры платформы TXM были применены для анализа корпуса текстов русских рассказов 1901-1930 годов, в котором тексты большого числа авторов разделялись на подкорпуса по времени написания рассказов.

В данной работе на примере созданного корпуса текстов литературных произведений русских классиков, разбитого на подкорпуса по авторам произведений, мы изучаем одновременное использование двух технологий: средств корпусного анализа платформы TXM и методов корреляционного рангового анализа словарей текстов.

1. Исследуемый корпус текстов

В данной работе мы анализируем корпус текстов художественной литературы, созданный авторами данной работы из прозаических текстов художественной литературы России 1824–1913 гг. В качестве литературных текстов, которые формируют корпус, выбраны следующие литературные произведения и сборники произведений пяти авторов-прозаиков России: А.С. Пушкин – «Повести Белкина», «Дубровский», «Капитанская дочка», «Моцарт и Сальери», «Пиковая дама»; Н.В. Гоголь – «Невский проспект», «Игроки», «Нос», «Повесть о том, как поссорился Иван Иванович с Иваном Никифоровичем», «Шинель», «Старосветские помещики», «Записки сумасшедшего», «Женитьба»; А. П. Чехов – «Моя жизнь», «Степь», «Учитель словесности»; И. С. Тургенев – «Холостяк», «Му-му», «Первая любовь», «После смерти», «Три встречи»; Максим Горький – «Детство», «Мой спутник», «Супруги Орловы».

С целью исследования различий в употребляемой лексике авторов одного периода времени корпус составлялся из наборов текстов, составленных из произведений пяти авторов. Корпус состоит из пяти подкорпусов текстов, каждый из которых составлен из литературных текстов одного из известных авторов. Каждый из пяти подкорпусов текстов составлен из суммарного набора текстов сравнимого размера, что необходимо для корректного сравнительного анализа. Подкорпуса составлены так, чтобы наблюдались равноценные распределения частотностей словоупотреблений между образованными подкорпусами. Корпус состоит из подкорпусов текстов прозы разных авторов, указанных в табл. 1.

Сформированный корпус был загружен на платформу TXM. При загрузке на корпусную платформу TXM морфологическая разметка осуществлялась автоматически на базе программной процедуры TreeTagger [8]. Формирование словарей для корреляционного анализа осуществлялось по итогам морфологического анализа, методами описанными в [12]. Размеры основных составленных словарей указаны в табл. 1.

2. Анализ соответствий (*correspondence analysis*) для корпусов текстов

Для анализа созданного корпуса текстов применялся множественный анализ соответствий [10,11,14], который можно трактовать как аналогичный факторному анализу метод снижения размерности анализируемого пространства признаков. Данный метод позволяет сравнить подкорпус с другим подкорпусом и с корпусом в целом. Корпус должен быть относительно сбалансирован по объемам подкорпусов. В нашем случае все подкорпуса специально создавались приблизительно одного размера (порядка 100 000 словоупотреблений).

Множественный анализ соответствий – это статистический метод анализа данных, основная цель которого состоит в формировании факторов из некоторого выделенного набора значений параметра (значений из лексической таблицы). Найденные факторы формируются таким образом, чтобы

Табл. 1

Состав корпуса текстов и размеры словарей

№	Подкорпус, тексты	Число словоупотреблений	Размеры словарей в количестве записей		
			Существительных	Глаголов	Прилагательных
1	А. П. Чехова	103534	6270	5806	3467
2	Н. В. Гоголя	102561	5579	5194	2972
3	М. Горького	104581	5592	6382	3129
4	А. С. Пушкина	103203	5770	5607	3030
5	И. С. Тургенева	97975	4422	5261	2639

первый фактор объяснял наибольшую долю различий между подкорпусами, второй – меньшую, третий – еще меньшую и т.д.

Объектом анализа соответствий являются таблицы сопряженности. В корпусном анализе – это таблицы встречаемости лексических характеристик для различных подкорпусов. В наших исследованиях в таблицы включаются 200 наиболее частотных в текстах корпуса лексических единиц и рассматриваются отдельно для частотных наборов всех лемм и разных частей речи, исключающих имена собственные.

Таблица сопряженности представляется в бинарной форме, которая представляет собой матрицу, значениями элементов которой могут быть только 0 и 1, а число столбцов равно суммарному количеству категорий для всех признаков. Формируется матрица Берта [14] – симметричная матрица блочного вида. Вычленение факторов есть следствие решения задачи на собственное значение для сформированной матрицы.

Анализ соответствий позволяет представить полученные результаты графически, что существенно облегчает интерпретацию (рис. 1, 2). Процент, предоставленный для каждой оси, указывает на ее вес в сформированном факторе разделяющем подкорпуса. На рис. 1, например, две первые оси захватывают почти две трети (34,86% + 29,70% = 64,56%) от общей информации о корпусе, а на рис. 2 представлены результаты описывающие более двух третей (47,09% + 24,27% = 71,36%) характеристик (признаков соответствующего пространства). Визуальное представление данных дает возможность увидеть и описать определенные закономерности, взаимосвязи подкорпусов.

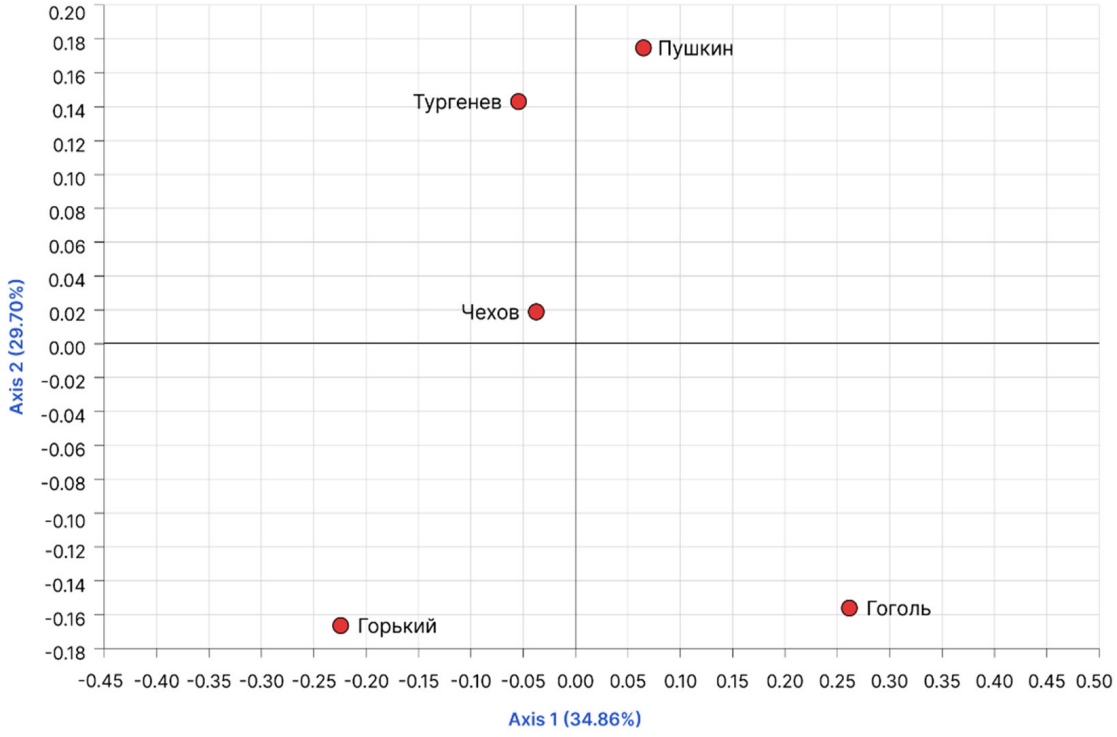

Рис. 1. Результат анализа соответствий для существительных

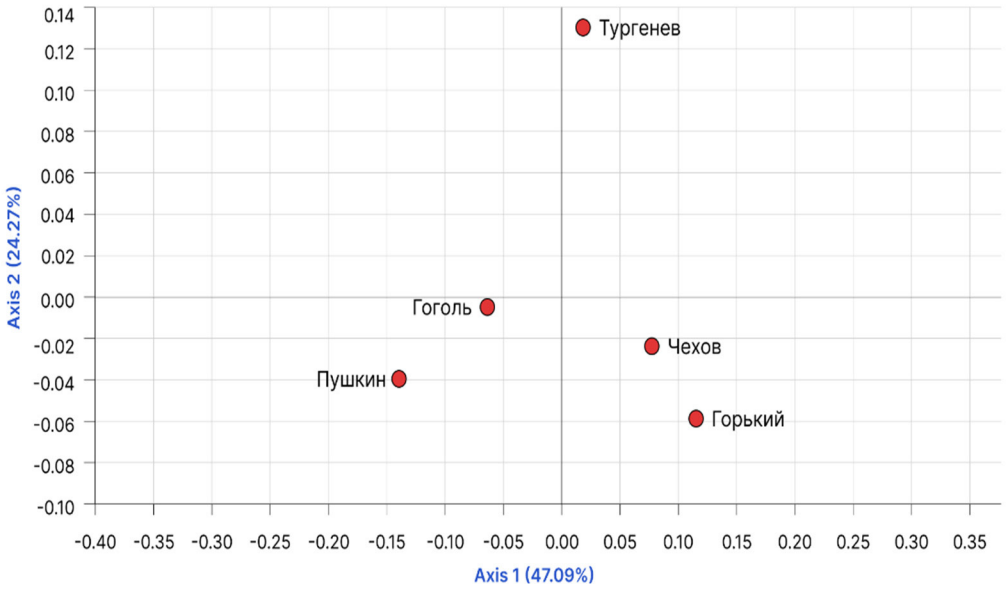

Рис. 2. Результат анализа соответствий для глаголов

В качестве процедуры метода анализа соответствий в ТХМ используется интегрированная в платформу сторонняя библиотека, описанная в [11].

3. Корреляционный ранговый анализ

Сравнение подкорпусов текстов можно проводить попарным сравнением частотных словарей различных лингвистических характеристик (всей лексики в целом, отдельных частей речи), составленных для наборов текста каждого подкорпуса. Определяются ранги записей словаря по результатам сортировки по частоте встречаемости занесенной в словарь характеристики. Данные записи рассматриваются как случайные величины. Тогда для каждой пары словарей можно подсчитать коэффициенты попарной ранговой корреляции. Для каждой пары словарей рассматриваем одинаковое количество *n* записей, которое в реальных расчетах ограничивается некоторым заданным значением. Данное ограничение числа записей словарей обеспечивает отбрасывание характеристик с низкими частотами использования.

Если записи в словарях рассматриваются как выборки для двух случайных величин: $\boldsymbol{X^n} = \{\boldsymbol{X_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$, $\boldsymbol{Y^n} = \{\boldsymbol{Y_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$. Мера зависимости случайных величин X, Y определяется через средние значения выборок $\overline{\boldsymbol{X^n}}$ и $\overline{\boldsymbol{Y^n}}$. Для ковариации $\boldsymbol{cov(X^n, Y^n)}$ имеем:

$$\boldsymbol{cov(X^n, Y^n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X^n})(Y_i - \overline{Y^n})}, \quad (1)$$

а дисперсия может быть записана в следующей форме:

$$\boldsymbol{\sigma X^n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X^n})^2}}\ . \quad (2)$$

Для отсортированных по частотам словарей рассматриваются ранги $\boldsymbol{rgX^n}$ и $\boldsymbol{rgY^n}$. Коэффициент попарной ранговой корреляции определяется следующим образом [15]:

$$\boldsymbol{r = r(rgX^n, rgY^n) = \frac{cov(rgX^n,\ rgY^n)}{\sigma rgX^n \cdot \sigma rgY^n}}, \quad (3)$$

где числитель и знаменатель определяются формулами (1) и (2), соответственно.

Коэффициент попарной ранговой корреляции, определенный в (3), не накладывает никаких ограничений на порядок элементов, имеющих одинаковые значения. В случае наличия элементов, имеющих одинаковые значения, они могут быть упорядочены согласно произвольной перестановке [5]. Для целей однозначности вычисления коэффициента (3) используется ранг, усредненный по всем перестановкам. Равные по значению элементы получают одинаковое значение усредненного ранга, не зависящее от их перестановки. Если все частоты внутри каждого из анализируемых словарей не совпадают, формула (3) преобразуется в классическую формулу ранговой корреляции Спирмена [16].

Посредством коэффициента попарной ранговой корреляции сравниваются два частотных словаря. Анализируемые словари могут иметь достаточно большие размеры. Поэтому мы рассматриваем только первые 10000 (по убыванию частоты) элементы каждого из словарей.

Вычисляемый коэффициент попарной ранговой корреляции принимает значения на интервале [-1,1]. Близкие к **1** значения говорят о монотонной согласованности словарей. Близкие к –**1** говорят об обратном эффекте: если в одном словаре слово с частотой выше второго, то в другом – наоборот, его частота будет ниже. Если значение близко к 0, то словари несогласованы: соотношение между частотами слов в одном словаре независимо от соотношения с частотами в другом словаре. Результаты такого анализа определяют, близки ли тексты разных подкорпусов по тематической и психологической направленности.

4. Результаты исследований

Вычислительные эксперименты для описанного корпуса проводились двумя указанными выше методами: множественным анализом соответствий и корреляционным анализом частотных словарей подкорпусов.

Примеры результатов анализа соответствий для исследуемого корпуса представлены на рис. 1 и 2, для существительных и глаголов, соответственно. В качестве результатов представляются графики, имеющие две оси координат. В названии каждой из осей графика указан процент характерных (отличительных) параметров, выделенных в ходе процедуры анализа соответствий по корпусу. В качестве величин, которые характеризуют оси координат, используется численная характеристика степени отклонения набора признаков от указанной в названии оси процента вариации для конкретного подкорпуса. Пространственное удаление, расположение корпуса от начала отсчета координат свидетельствует о сильном отличии его набора признаков от признаков по корпусу.

Очевидно, что в графическом представлении результатов анализа соответствий для существительных и глаголов подкорпуса текстов разнесены по пространству диаграммы. Представленные результаты анализа соответствий для наборов признаков подкорпусов текстов (по авторам) исследуемого корпуса демонстрируют достаточно явное разделение по лингвистическим характеристикам.

Вычислительные эксперименты по исследованию частотных словарей подкорпусов заключались в подсчете коэффициента корреляции (3) для всех пар частотных словарей лингвистических характеристик текстов подкорпусов, перечисленных

Табл. 2

Сравнение словарей существительных

Подкорпус, номер	**1 А.П. Чехова**	**2 Н.В. Гоголя**	**3 М. Горького**	**4 А. С. Пушкина**	**5 И.С. Тургенева**
1.А. П. Чехова	1				
2. Н. В. Гоголя	-0.023	1			
3. М. Горького	0.069	-0.085	1		
4. А. С. Пушкина	-0.078	-0.014	-0.138	1	
5. И.С. Тургенева	0.046	0.011	-0.037	-0.007	1

Табл. 3

Сравнение словарей глаголов

Подкорпус, номер	**1 А.П. Чехова**	**2 Н.В. Гоголя**	**3 М. Горького**	**4 А. С. Пушкина**	**5 И.С. Тургенева**
1.А. П. Чехова	1				
2. Н. В. Гоголя	0.017	1			
3. М. Горького	0.047	-0.119	1		
4. А.С. Пушкина	-0.044	0.083	-0.133	1	
5. И.С. Тургенева	0.043	0.114	-0.101	0.034	1

в табл. 1. В качестве примеров приведены таблицы сравнения через коэффициент корреляции словарей существительных и словарей глаголов для исследуемых подкорпусов (табл. 2,3).

Значения коэффициентов корреляции для всех частотных словарей в табл. 2 и 3 колеблются около нулевого значения, что указывает на то, что частотные словари разных подкорпусов (текстов одного автора) несогласованы: частоты слов в одном словаре независимы от частот в другом. По сути, фиксируются принципиальные различия текстов по частотным характеристикам лексических характеристик. В целом результаты сравнения, представленные в табл. 2 и 3, показывают возможность рассматривать лексические единицы (части речи в наших исследованиях) как дифференцирующие признаки литературных произведений различных авторов.

Сравнение представленных результатов на рис. 1 и 2 с одной стороны и в табл. 2 и 3 с другой демонстрирует хорошую корреляцию результатов разными методами анализа подкорпусов текстов по разделению подкорпусов текстов (в нашем исследовании по авторам) по набору признаков (значений лингвистических характеристик).

Заключение

Применение анализа соответствий демонстрирует возможности разделения подкорпусов на основе анализа частот совместного появления значений переменных. Набор характерных (отличительных) параметров (например, лексических единиц), выделенных в ходе процедуры анализа соответствий, можно рассматривать как дифференцирующие признаки, которые можно использовать в самых различных задачах классификации, например, в задачах разделения текстов по тематике, задачах определения авторства. Корреляционный ранговый анализ показывает соответствие результатам анализа соответствий разделения подкорпусов текстов.

Приведенные примеры иллюстрируют возможность использования всех сформированных в исследованиях признаков двумя представленными методами (анализа соответствий и корреляционного рангового анализа) для решения задач классификации текстов.

Исследуемый корпус и методы его анализа демонстрируют возможности использования математических методов компьютерной лингвистики для создания наборов данных для машинного обучения с целью решения различных задач средствами методов искусственного интеллекта.

Литература

1. *Лаврентьев А.М., Смирнов И.В., Соловьев Ф.Н., Суворова М.И., Фокина А.И., Чеповский А.М.* Анализ корпусов текстов террористической и антиправовой направленности // Вопросы кибербезопасности. 2019. № 4(32). С.54-60. DOI: 10.21681/2311-3456-2019-4-54-60.
2. *Лаврентьев А.М., Смирнов И.В., Соловьев Ф.Н., Суворова М.И., Фокина А.И., Чеповский А.М.* Создание специальных корпусов текстов на основе расширенной платформы TXM // Системы высокой доступности. 2018. Т.14. № 3. С.76-81.
3. *Аванесян Н.Л., Соловьев Ф.Н., Тихомирова Е.А., Чеповский А.М.* Выявление значимых признаков противоправных текстов // Вопросы кибербезопасности. 2020. № 4 (38). С. 76–84. DOI: 10.21681/2311-3456-2020-04-76-84
4. *Фокина А.И., Чеповский А.А., Чеповский А.М.* Использование платформы TXM корпусного анализа для анализа текстов сообществ социальных сетей // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2023. Т. 21. № 2. С. 29-38. DOI: 10.25205/1818-7900-2023-21-2-29-38.
5. *Аванесян Н.Л., Зенькова В.В., Чеповский А.А., Чеповский А.М.* Анализ текстов сообществ социальных сетей // Успехи кибернетики. 2023. 4(2). С. 33–39. DOI: 10.51790/2712-9942-2023-4-2-05.
6. *Heiden S.* The TXM Platorm: Building Open-Source Textual Analysis Sofware Compatile with the TEI Encoding Scheme. In: 24th Pacific Asia Conference on Language, Information and Computation – PACLIC24 / Ed. by R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto and Y. Harada. Institute for Digital Enhancement of Cognitive Development. Waseda University, Sendai, Japan.2010. P. 389–398.
7. TXM public website. [Online] Available from: http://textometrie.org. (Дата обращения 23.01.2024)
8. *Schmid, H.* Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings International Conference on New Methods in Language Processing, Manchester, UK, Sept. 1994. P. 44–49.
9. *Лаврентьев А.М., Соловьев Ф.Н., Чеповский А.М.* Внедрение в TXM дополнительных инструментов автоматической обработки текста. В кн.: Труды международной конференции «Корпусная лингвистика – 2019». СПб.: Издательство Санкт-Петербургского университета. 2019. С. 55-62.
10. *Benzécri J.-P., Bellier L.* L'analyse des données. V. 2:L'analyse des Correspondances. Paris: Dunod. 1976. 616 p.
11. *Lê S., Josse J., Husson F.* FactoMineR:an R package for multivariate analysis // Journal of statistical software.2008. № 25 (1). P. 1-18.

12. *Чеповский А.М.* Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.: Национальный открытый университет «ИНТУИТ». 2015. 228 с.
13. *Lavrentiev A., Sherstinova T., Chepovskiy A., Pincemin B.* Using TXM Platform for Research on Language Changes over Time: The Dynamics of Vocabulary and Punctuation in Russian Literary Texts // Вестник Томского государственного университета. Филология. 2021. № 70. С. 69-89. DOI: 10.17223/19986645/70/5.
14. Прикладная статистика: Классификации и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин / Под ред. С.А. Айвазяна. М.: Финансы и статистика. 1989. 607 с.
15. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных. М.: Мир. 1989. 540 с.
16. *Деза Е.И., Деза М.М.* Энциклопедический словарь расстояний. М.: Наука. 2008. 444 с.

Аванесян Нина Левоновна. Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия. Аспирант. Область научных интересов: автоматическая обработка текста, информационная безопасность. E-mail: nlavanesyan@edu.hse.ru

Губина Ольга Вячеславовна. Российский университет дружбы народов им. Патриса Лумумбы, г. Москва, Россия. Студент. Область научных интересов: автоматическая обработка текста. E-mail: 1032201737@pfur.ru

Чеповский Андрей Михайлович. Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия. Профессор. Доктор технических наук. Российский университет дружбы народов им. Патриса Лумумбы, г. Москва, Россия. Профессор. Область научных интересов: автоматическая обработка текста, информационный поиск, кибернетика, информационная безопасность. E-mail: achepovskiy@hse.ru (Ответственный за переписку).

Corpus analysis methods for study of texts of prose literary works by various authors

N.L. Avanesyan[I], O.V. Gubina[II], A.M. Chepovskiy[I,II]

[I] National Research University «Higher School of Economics», Moscow, Russia.

[II] Peoples' Friendship University of Russia named after Patrice Lumumba, Moscow, Russia.

Abstract. This article is devoted to the application of corpora analysis mathematical methods for the research of Russian fiction texts. A corpus of prose texts of Russian XIX century fiction, consisting of five subcorpora, has been created for the research. Each subcorpora contains texts of one certain author. Using the example of the created corpora, the possibilities of using the correspondence analysis method integrated into the TXM platform as one of the tools of the statistical research method are demonstrated. As another method, we consider the analysis of pairwise rank correlation coefficients to compare the frequency characteristics of texts of different subcorps. The methods described give correlated results and make it possible to identify differentiating features. The methods described give correlated results and make it possible to identify differentiating features. The described method can be used both for linguistic and literary studies and for creating appropriate training text sets for artificial intelligence tasks.

Keywords: *corpus linguistics, TXM platform, correspondence analysis, correlation analysis*.
DOI: 10.14357/20790279240204 **EDN: IKHGUO**

References

1. *Lavrentiev A.M., Smirnov I.V., Solovev F.N., Suvorova M.I., Fokina A.I., Chepovskiy A.M.* Analysis of corpus of extremist texts and unlawful texts // Voprosi kiberbezopasnosti. 2019. № 4(32). P. 54–60. DOI: 10.21681/2311-3456-2019-4-54-60 [in Russian].
2. *Lavrentiev A.M., Smirnov I.V., Soloviev F.N., Suvorova M.I., Fokina A.I., Chepovskiy A.M.* Creating text corpora for special purposes on the basis of extended TXM platform // Sistemy vysokoy dostupnosti, 2018. Vol. 14. No. 3. P. 76–81. [in Russian].

3. *Avanesyan N.L., Solovev F.N., Tikhomirova E.A., Chepovskiy A.M.* Identifying the significant features in illegal texts. Voprosy kiberbezopasnosti, 2020. No. 4 (38). P. 76–84. (in Russ.) DOI: 10.21681/2311-3456-2020-04-76-84.
4. *Fokina A.I., Chepovskiy A.A., Chepovskiy A.M.* Using TXM Platform of Corpus Analysis for Text Analysis of Social Media // Vestnik NSU. Series: Information Technologies. 2023. Vol. 21. No. 2. P. 29–38. DOI: 10.25205/1818-7900-2023-21-2-29-38. [in Russian].
5. *Avanesyan N.L., Zenkova V.V., Chepovskiy A.A., Chepovskiy A.M.* Analysis of Social Media Community Posts. Russian // Journal of Cybernetics. 2023;4(2):33–39. DOI: 10.51790/2712-9942-2023-4-2-05.
6. *Heiden S.* The TXM Platorm: Building Open-Source Textual Analysis Sofware Compatile with the TEI Encoding Scheme In: 24th Pacific Asia Conference on Language, Information and Computation – PACLIC24 / Ed. by R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto and Y. Harada. Institute for Digital Enhancement of Cognitive Development. Waseda University. Sendai. Japan.2010. P. 389−398.
7. TXM public website. [Online] Available from: http://textometrie.org. [accessed:23.01.2024],
8. *Schmid H.* Probabilistic Part-of-Speech Tagging Using Decision Trees/ In: Proceedings International Conference on New Methods in Language Processing. Manchester. UK. Sept. 1994. P. 44–49.
9. *Lavrentiev A.M., Solovev F.N., Chepovskiy A.M.* Implementation in the TXM Platform of Additional Instruments of Automatic Text Processing. In: Proceedings of the international conference "Corpus linguistics – 2019" St. Petersburg University Publishing House. 2019. P. 55-62. [in Russian]
10. *Benzécri J.-P., Bellier L.* L'analyse des données. V. 2: L'analyse des Correspondances. Paris: Dunod. 1976. 616 p.
11. *Lê S., Josse J., Husson F.* FactoMineR: an R package for multivariate analysis // Journal of statistical software.2008. № 25 (1). P. 1-18.
12. *Chepovskiy A.M.* Informatsionnyye modeli v zadachakh obrabotki tekstov na estestvennykh yazykakh [Information Models for the Problems of Natural Text Processing]. 2nd ed. Moscow: Natsional'nyy otkrytyy universitet "INTUIT". 228 p. [in Russian]
13. *Lavrentiev A., Sherstinova T., Chepovskiy A., Pincemin B.* Using TXM Platform for Research on Language Changes over Time: The Dynamics of Vocabulary and Punctuation in Russian Literary Texts // Vestnik Tomskogo Gosudarstvennogo Universiteta, Filologiya. 2021. Vol. 70. P. 69-89. DOI: 10.17223/19986645/70/5.
14. Applied statistics: Classifications and dimensionality reduction / S.A. Ayvazyan, V.M. Buhshtaber, I.S. Enucov, L.D. Meshalkin. / Ed. S.A. Ayvazyan. – M.: Finansy I statistika. 1989. 607 p.
15. *Bendat J., Piersol A.* Prikladnoy analiz sluchainikh dannikh. Moscow: Mir, 1989. 540 p. [in Russian].
16. *Deza Elena, Deza Michel Marie.* Dictionary of Distances. Moscow: Nauka. 2008. 444 p. [in Russian].

Avanesyan Nina L. Postgraduate student, National Research University «Higher School of Economics» (Moscow, Russia). Author of 5 publications. Research interests: natural language processing, information security. E-mail: nlavanesyan@edu.hse.ru

Gubina Olga V. Student, Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University, Moscow, Russia). Research interests: natural language processing, E-mail: 1032201737@pfur.ru

Chepovskiy Andrey M. Doctor of Technical Science, professor of chair of computer security in National Research University «Higher School of Economics» (Moscow, Russia) and professor of chair of mathematical modeling and artificial intelligence in Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University, Moscow, Russia). Author of 120 publications. Research interests: natural language processing, information retrieval, cybernetics, information security. E-mail: achepovskiy@hse.ru