СРАВНИТЕЛЬНЫЙ АНАЛИЗ СПЕЦИАЛЬНЫХ КОРПУСОВ ТЕКСТОВ ДЛЯ ЗАДАЧ БЕЗОПАСНОСТИ[1]

Лаврентьев А.М.[2], Рябова Д.М.[3], Тихомирова Е.А.[4], Фокина А.И.[5], Чеповский А.М.[6], Шерстинова Т.Ю.[7]

Цель исследования: *разработка методики сравнения специальных корпусов текстов для последующего применения в задачах идентификации экстремистских текстов.*

Метод: *применялись частотные методы и показатель специфичности для анализа текстов в рамках корпусной платформы ТХМ.*

Полученные результаты: *разработана методика сравнительного анализа специальных корпусов текстов, которая позволяет выявлять неявные связи между корпусами разнородных текстов; показана возможность использования индекса специфичности для составления своего рода «профиля» подкорпуса (набора текстов); проведен сравнительный анализ корпуса текстов террористической, экстремистской направленности и корпуса русских рассказов первой трети двадцатого века; обнаружены взаимосвязи лексики противоправных и литературных текстов; показаны возможности использования корпусной лингвистики для исследования свойств экстремистских текстов с целью обнаружения противоправных ресурсов и сообщений в Интернете; показаны возможности использования как морфологических характеристик слов, так и псевдооснов словоупотреблений в анализе специфичности при корпусном анализе; результаты исследований показывают, что инструменты частотного анализа, предоставляемые платформой ТХМ, эффективны для прикладных задач, когда необходимо выявить неявные лексические совпадения различных корпусов текстов.*

Ключевые слова: *корпусная лингвистика, автоматический анализ текстов, платформа корпусного анализа, показатель специфичности, экстремистские тексты.*

DOI: 10.21681/2311-3456-2020-03-58-65

1. Введение

Актуальной проблемой государственной безопасности стала задача ограничения распространения экстремистской информации в Интернете. Необходимо своевременно выявлять подобные сайты и блокировать их.

Проблема обнаружения онлайн-экстремизма привлекает внимание исследователей с начала этого века. В настоящее время существует огромное количество литературы по этой теме, включая обширные обзоры существующих решений и методов с комплексным анализом тенденций, например работа [1], в которой отмечается, что наиболее популярными методами для выявления экстремистского контента являются текстовая классификация на основе ссылок. В [2] исследователи пришли к выводу, что автоматическое обнаружение вербовки террористов в Интернете является выполнимой задачей, если использовать методы классификации текстовой информации. Стоит отметить, что в большинстве наблюдаемых подходов, в качестве признаков классификации используют простые лексические или статистические атрибуты. Такие методы статистического анализа обычно применяются для обнаружения плагиата и идентификации автора [3]. В работах по автоматическому обнаружению противоправных текстов отмечается необходимость создания соответствующего программного обеспечения для экспертов [4, 5].

1 Работа выполнена при финансовой поддержке РФФИ в рамках научных проектов № 19-07-00806 и № 17-29-09173

2 Лаврентьев Алексей Михайлович, кандидат филологических наук, сотрудник Института истории представлений и идей нового времени Национального центра научных исследований Франции и Высшей нормальной школы Лиона. Лион, Франция. E-mail: alexei.lavrentev@ens-lyon.fr

3 Рябова Дарья Михайловна, магистр, Школа бизнес-информатики Национального исследовательского университета «Высшая школа экономики». Москва, Россия. E-mail: dmryabova@edu.hse.ru

4 Тихомирова Елизавета Алексеевна, доцент кафедры ИУ3 Московского государственного технического университета им. Н.Э. Баумана. Москва, Россия. E-mail: elizarti@bmstu.ru

5 Фокина Алина Игоревна, студент магистратуры Школы бизнес-информатики Национального исследовательского университета «Высшая школа экономики». Москва, Россия. E-mail: aifokina@edu.hse.ru

6 Чеповский Андрей Михайлович, доктор технических наук, профессор, профессор кафедры информационных технологий Российского университета Дружбы Народов (РУДН) (Москва) и профессор кафедры Информационной безопасности Национального исследовательского университета «Высшая школа экономики». Москва, Россия. E-mail: achepovskiy@hse.ru

7 Шерстинова Татьяна Юрьевна, кандидат филологических наук, доцент Департамента филологии Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург) и доцент Филологического факультета Санкт-Петербургского государственного университета, г. Санкт-Петербург, Россия. E-mail: tsherstinova@hse.ru

Для определения экстремистских текстов методами классификации необходим размеченный корпус текстов, а также наличие набора дифференцирующих признаков. Подобные корпуса позволяют повышать эффективность существующих методов обнаружения. В предыдущих работах [6, 7] нами были предложены различные расширения стандартной корпусной платформы для изучения специализированных наборов текстов, которые были применены для исследования корпуса противоправных текстов и решена задача выделения дифференцирующих признаков с целью применения методов классификации для выявления экстремистских текстов в Интернете. В данной работе разрабатывается методика сравнения различных по своей природе корпусов текстов с целью сравнения лексических характеристик текстов.

2. Исследуемые корпуса текстов

В работе сравнивались два корпуса текстов на естественном языке: корпус противоправных текстов и корпус рассказов.

Корпус противоправных текстов. Для проведения исследований был использован корпус, в который вошли тексты экстремистской направленности, а также коллекция сходных по тематике, но нейтральных по стилю текстов, включая сообщения с оппозиционных и проправительственных политических блогов, разрешенные тексты религиозного содержания, новостные статьи [8, 9, 10]. Корпус противоправных текстов собирался вручную и насчитывает почти 3,3 миллиона словоупотреблений.

Корпус текстов содержит противоправные тексты семи категорий: терроризм, идеологические тексты, религиозная ненависть, сепаратизм, национализм, агрессия и призывы к беспорядкам, фашизм, а также нейтральные тексты со схожей лексикой.

Подкорпус терроризма состоит из текстов с сайтов, запрещенных в РФ организаций, где размещаются обращения членов этих группировок, идеологическая пропаганда. Идеологические тексты содержат религиозную тематику: в них говорится о превосходстве одной религии над другой, распространяются неверные толкования священных книг и возникают призывы о принятии другой веры. Тексты религиозной ненависти близки к идеологическим, но в них содержатся более активные призывы к действиям против представителей других конфессий, распространяются ложные сведения об агрессивных намерениях представителей других религий, формируется их негативный образ. Оскорбления различных этнических групп, угрозы в их адрес и идеи отделения субъектов РФ попали в подкорпус сепаратизма. «Национализм» содержит тексты о вражде к определенным этническим группам, их уничтожении и ограничении прав и свобод. Различные призывы к митингам, тексты о ненависти к власти, ее свержению и уничтожению попали в подкорпус агрессии и призывов к беспорядку. «Фашизм» состоит из текстов об идеях фашизма, распространении символики, поддержании идей неофашизма и геноцида.

Корпус рассказов. В работе рассматривается корпус русских рассказов первой трети XX в., который был разработан на филологическом факультете Санкт-Петербургского государственного университета при участии департамента филологии НИУ Высшая школа экономики в Санкт-Петербурге [11, 12, 13]. Корпус предназначен для проведения стилеметрических исследований русской прозы, а также для изучения изменений, которые произошли в русском языке в эпоху революций.

Произведения малой художественной прозы как нельзя лучше отражают происходящие события и изменения во всех сферах жизни человека: в культурной, политической, социальной и даже бытовой сферах. Это наиболее популярный литературный жанр, который охватывает практически все литературные направления. Несомненным его преимуществом можно считать и то, что рассказы четко улавливают и оперативно реагируют на изменения в сознании и культуре общества. В отличие от традиционных корпусов, ориентированных на тексты выдающихся писателей (одного или нескольких), данный корпус рассказов предназначен для исследования языка всей литературной системы, то есть содержит рассказы в том числе многих забытых и периферийных писателей, что позволяет в полной мере использовать весь потенциал художественной литературы для проведения исследований в области корпусной лингвистики.

Для аннотированной части корпуса из общего списка русских писателей, насчитывающего более 2800 персоналий, было отобрано 300. Общий объем аннотированной части – 310 рассказов, 1 млн. словоупотреблений. Корпус разбит на 4 временных отрезка: 1) начало XX века (1900-1913 гг.), 2) предреволюционные годы и Первая мировая война (1914-1916 гг.), 3) революционные годы – Февральская и Октябрьская революции и Гражданская война (1917-1922 гг.) и 4) постреволюционные годы (1923-1930 гг.). Корпус охватывает рассказы представительного числа авторов для каждого временного периода, позволяет проводить лингвистический и статистический анализ языка и стиля.

Предобработка корпусов текстов. Для всех исследуемых текстов определялись дифференцирующие признаки программным комплексом, созданным для расширения стандартной корпусной платформы для изучения специализированных корпусов текстов [7, 14, 15]. Характеристики текстов определялись процедурами автоматизированной обработки текстов на естественных языках, описанными в [16]. Для определения дифференцирующих признаков мы используем аналитический метод выделения псевдооснов, так как он позволяет обрабатывать отсутствующие в стандартных словарях формы. Используемый способ выделения псевдооснов представляет собой метод структурных схем. Использование широкого спектра характеристик позволяет анализировать текстовые конструкции, опираясь не только на точные словоформы, и тем самым повышает полноту и гибкость корпусного анализа. Выделяются не только словоформы, но и именные группы и глагольные группы целиком. В отличие от отдельных слов, выделенные именные и глагольные группы несут информацию о конкретных отдельных аспектах содержания текста. При предобработке всех текстов мы осу-

DOI:10.21681/2311-3456-2020-03-58-65

ществляем автоматический морфологический анализ словоформ на основе словарной компьютерной морфологии, описанной в [16]. Для расширения характеристик текстов проводилось выделение глагольных групп (словосочетаний, главным словом которых является глагол), то есть установление связей выделенных именных групп с глаголами. Данная задача решается нами при помощи анализа глагольного управления. Каждому глаголу сопоставлен набор ограничений, накладываемых им на зависимые словосочетания. Такой набор есть парадигма глагольного управления, если мы отвлекаемся от самого глагола и рассматриваем набор его ограничений сам по себе. Двухбуквенные и трехбуквенные сочетания использовались нами как дополнительные признаки текстов на естественном языке. Все описанные признаки были включены в качестве характеристик корпусов текстов по аналогии с анализом словоформ в стандартном анализе на базе платформы TXM [7, 14, 15].

3. Применяемые методы корпусного анализа

Сравнительный анализ корпусов базируется на платформе TXM [17] (http://textometrie.org), являющейся предназначенным для корпусного анализа программным комплексом, обеспечивающим разнообразный анализ текстов. В платформе TXM инструментом количественной оценки «необычности» подкорпуса относительно всего корпуса является показатель специфичности [18]. Анализ специфичности позволяет составить своего рода «профиль» подкорпуса путем выявления наиболее характерных или нехарактерных для него лингвистических объектов (лексем, псевдооснов, именных и глагольных групп и других).

В данной работе мы используем анализ специфичности для отдельных характеристик текстов: слов (лемм), псевдооснов. Этот показатель характеризует «необычность» подкорпуса относительно всего корпуса. Суть специфичности заключается в том, что корпус T разбивается на несколько подкорпусов t_i. Специфичностью характеристики w (слов, псевдооснов) из выбранного подкорпуса t_k называется вероятность того, что в случайно выбранном подкорпусе t_m из корпуса T, где известен размер выбранного подкорпуса, характеристика w встретится столько же раз, сколько она встретилась в рассматриваемом подкорпусе t_k. Другими словами, специфичность соответствует вероятности из всего корпуса выбрать такой подкорпус с известным количеством употреблений характеристики w, чтобы это количество было равно числу ее употреблений в рассматриваемом подкорпусе.

Специфичность измеряется индексом специфичности, отображающим порядковую величину вероятности. В TXM специфичность представлена целой частью логарифма по основанию 10 (log_{10}). Если для рассматриваемой характеристики ее значение встречается в подкорпусе реже, чем в среднем по всему корпусу, то индекс специфичности может приобретать отрицательные значения. Принято считать, что значения характеристики является «нормальной» для рассматриваемого подкорпуса, если индекс специфичности принимает значения в пределах от -2 до 2. Такие значения характеристики не должны использоваться для описания особенности подкорпуса. С помощью анализа специфичности можно составить характеристику корпуса благодаря определению наиболее характерных и нехарактерных слов, словоформ, псевдооснов, буквосочетаний данного корпуса. Составленный «профиль» корпуса можно использовать для сравнения с другими текстами.

Процедура анализа в данной работе состояла в определении наиболее часто встречаемых значений характеристик в подкорпусах одного из корпусов и определения специфичности данных характеристик в другом корпусе. Для каждого из корпусов рассказов и противоправных текстов формируются таблицы используемых значений характеристик (для существительных, глаголов, прилагательных и псевдооснов). Вычисляются частоты встречаемости значений характеристик по каждому подкорпусу. Рассчитываются ранги и относительная частота значений характеристик в каждом из подкорпусов. Для наиболее частотных значений характеристик в подкорпусах противоправного корпуса строятся графики специфичности данных характеристик в корпусе рассказов.

4. Результаты сравнения корпусов текстов

Приведем некоторые из результатов сравнительных исследований. Примеры сопоставлений корпусов текстов, полученные на основе построения частотных словарей и вычислении коэффициентов специфичности, приведены на рисунках 1-3 и таблице 1.

Сравнительный анализ проводился в первую очередь для лексических характеристик «начальных форм слова» с морфологическими признаками: существительных, глаголов и прилагательных. Для выделенных в корпусе экстремистских текстов и корпусе рассказов начальных форм с указанными морфологическими признаками составлялись частотные словари. Для каждой морфологической характеристики выбирались начальные формы, которые имеют ранг менее 20 в частотных словарях соответствующих подкорпусов противоправного корпуса.

Отметим, что ранг начальных форм слова в подкорпусах противоправного корпуса сильно отличается от ранга в словарях подкорпусов рассказов и противоправном корпусе в целом. Многие из часто употребляемых слов в противоправном корпусе, в подкорпусах рассказов превышают ранг в 100 или даже 1000. В целом из выбираемых начальных форм слова более половины имеют в словарях подкорпусов рассказов ранги, отличающиеся от рангов в словарях противоправного корпуса агрессии более, чем в 2 раза. Указанные свойства частотных словарей начальных форм слова не позволяет проводить формальное сравнение лексики двух разнородных корпусов.

Ситуация со сравнением корпусов меняется при рассмотрении специфичности во втором корпусе для выявленных в первом корпусе высокочастотных характеристик. На рисунке 1 приведены выборочные результаты диаграмм индекса специфичности в корпусе рассказов для наиболее часто употребляемых существительных в подкорпусе агрессии противоправного корпуса.

Красными линиями выделен показатель специфичности равный *2* и *-2*. Большинство столбцов находятся в пределах этих показателей. Это говорит о том, что слова являются «нормальными» для подкорпусов рассказов и не представляют «необычности». Однако есть несколько «необычных» слов в рамках рассматриваемых подкорпусов. Диаграмма демонстрирует то, что вычисление индекса специфичности позволяет найти сильные выбросы показателей частоты, не свойственные для корпуса в целом.

Похожие результаты продемонстрированы на рисунке 2, где приведены индексы специфичности в корпусе рассказов для наиболее часто употребляемых глаголов в подкорпусе идеологии противоправного корпуса.

Результаты демонстрируют возможности среди выбранных слов выбрать такие, которые являются специфичными для каких-то подкорпусов рассказов: они либо слишком часто употребляются в подкорпусе, либо слишком редко относительно всего подкорпуса, что можно увидеть по графикам специфичности.

Сопоставление корпусов проводилось по характеристике псевдооснова. Под псевдоосновой мы понимаем часть слова, не содержащая суффиксов и префиксов. Способ автоматического выделения псевдооснов состоит в сопоставлении рассматриваемой словоформы с множеством допустимых в языке структур некорневой части слова [14, 16]. Для коротких текстов социальных сетей, которые характеризуются особыми психолингвистическими свойствами и содержанием неологизмов и жаргонизмов, анализ псевдооснов может дать лучшие качественные результаты, чем анализ слов.

Для всех подкорпусов противоправного корпуса были составлены списки наиболее часто встречающихся псевдооснов и подсчитаны индексы специфичности для подкорпусов корпуса рассказов. Пример распределения индексов специфичности для трех подкорпусов корпуса рассказов, для некоторых наиболее частотных псевдооснов в подкорпусе сепаратизма (ранг меньше 100) противоправного корпуса показаны на диаграмме (Рисунок 3). Из графиков видно, что встречаются псевдоосновы, которые являются специфичными для ряда подкорпусов рассказов. Представляется логичным, что некоторые наиболее частотные для подкорпуса сепаратизма псевдоосновы проявляются как специфичные именно для подкорпуса рассказов революционного периода.

Приведем фрагмент сводной таблицы, составленный показателей специфичности для псевдооснов (Таблица 1). В ней представлен список наиболее часто употребляемых псевдооснов в подкорпусе терроризма, которые являются специфичными в одном из подкорпусов противоправных текстов и в одном из подкорпусов рассказов. Цифрами в таблице указан знак специфичности: «1» означает, что индекс специфичности в рассматриваемом подкорпусе больше 2, «-1» – индекс специфичности в подкорпусе ниже -2, а термин «норма» в ячейке говорит о «нормальности» выбранной псевдоосновы в данном подкорпусе.

Из таблицы видно, что псевдоосновы являются «специфичными» для различных подкорпусов рассказов. Однако в противоправном корпусе текстов большинство из выбранных псевдооснов являются положительно специфичными для подкорпуса идеологии. Причем в подкорпусе агрессии знак специфичности почти у всех псевдооснов точно противоположен знаку в подкорпусе идеологии. В подкорпусе терроризма и сепаратизма большинство псевдооснов являются нейтральными несмотря на то, что выбранные псевдоосновы являются наиболее частотными для подкопруса терроризма. Можно заметить, что далеко не все псевдоосновы являются характерными для подкорпуса терроризма, тогда как в других они либо нейтральны, либо имеют отрицательную специфичность.

Результаты наших исследований указывают на требуемую аккуратность при выборе характерных для данной тематики признаков из частотного списка. Но возможно формировать списки слов и псевдооснов, которые являются специфичными для подкорпусов рассказов и наиболее часто встречаются в противоправных подкорпусах.

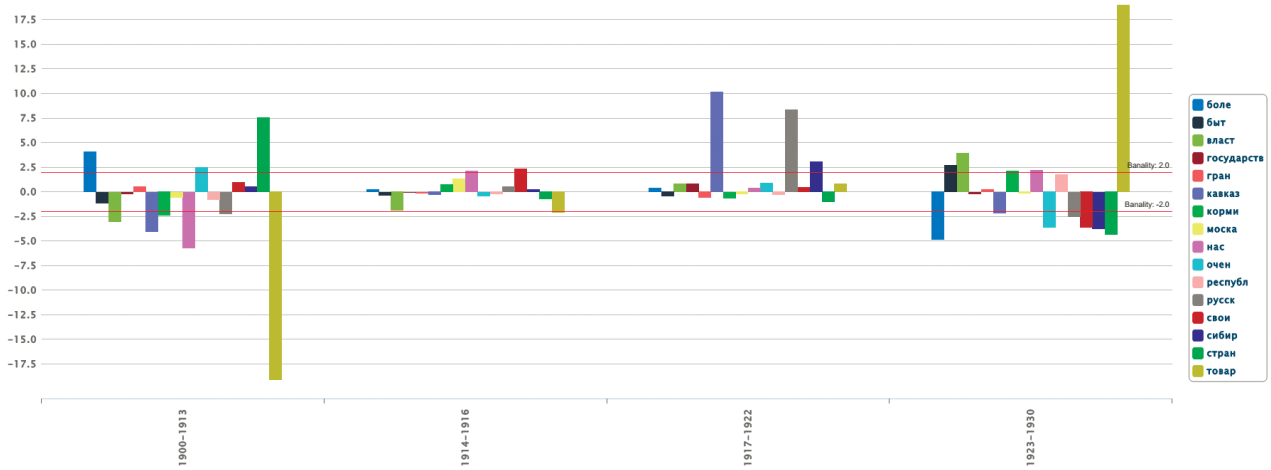

Рис. 1. Коэффициенты специфичности в корпусе рассказов для наиболее часто употребляемых существительных в подкорпусе агрессии противоправного корпуса

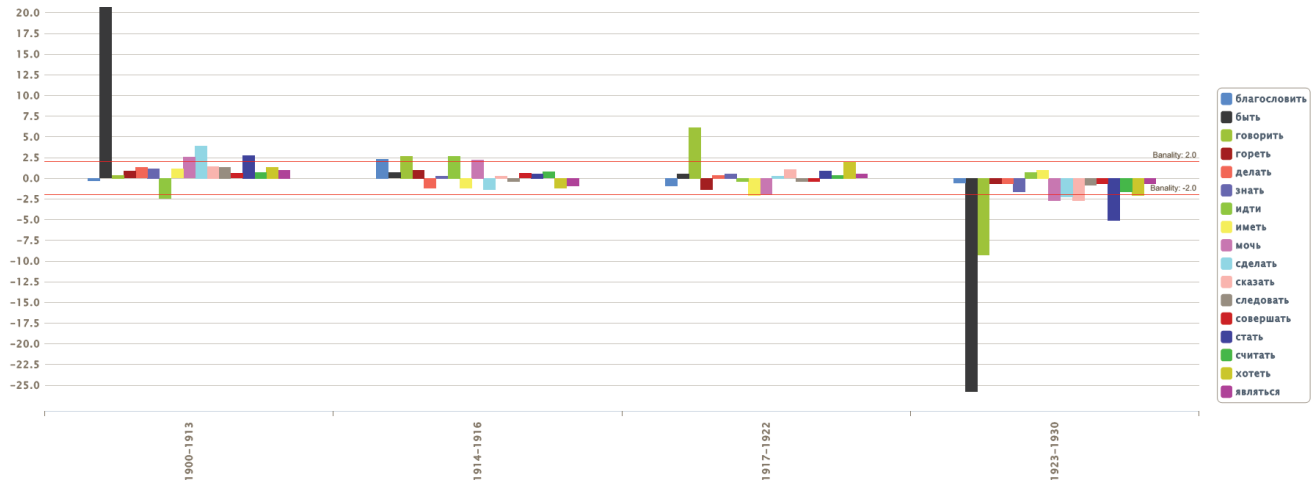

Рис. 2. Коэффициенты специфичности в корпусе рассказов для наиболее часто употребляемых глаголов в подкорпусе идеологии противоправного корпуса

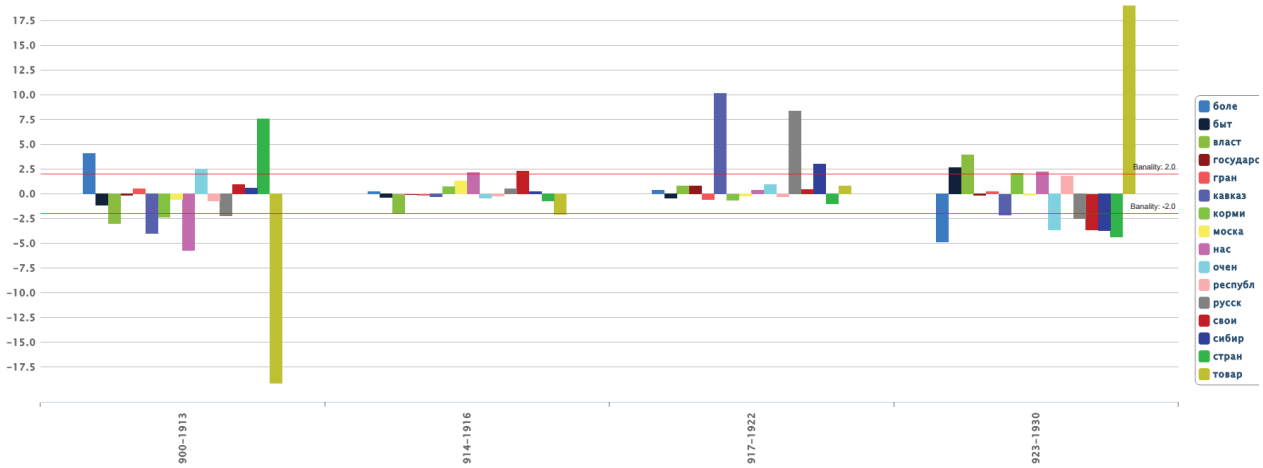

Рис. 3. Коэффициенты специфичности корпуса рассказов для наиболее частотных псевдооснов в подкорпусе сепаратизма противоправного корпуса

Таблица. 1

Знак специфичности в подкорпусах для псевдооснов, имеющих ранг меньше 100 в подкорпусе терроризма и специфичных в одном из корпусов рассказов

	1900-1913	1914-1916	1917-1922	1923-1930	Агрессия	Терро-ризм	Идеология	Сепаратизм
борьб	1	-1	норма	норма	норма	1	-1	норма
брать	норма	норма	1	норма	-1	норма	1	норма
был	норма	норма	1	норма	-1	1	норма	-1
вас	-1	норма	норма	1	-1	норма	1	норма
власт	-1	норма	норма	1	1	норма	-1	1
говор	норма	1	1	-1	-1	норма	1	норма
жизн	1	норма	-1	-1	-1	норма	1	-1

	1900-1913	1914-1916	1917-1922	1923-1930	Агрессия	Терро-ризм	Идеология	Сепаратизм
пути	-1	1	норма		-1	норма	1	норма
революц	-1	-1	1	1	1	1	-1	норма
свои	норма	1	норма	-1	-1	норма	1	норма
стран	1	норма	норма	-1	1	норма	-1	1
тольк	норма	1	норма	-1	1	норма	норма	норма

5. Выводы

Предложена и опробована методика сравнительного анализ специальных корпусов текстов, которая позволяет выявлять неявные связи между корпусами разнородных текстов.

Анализ специфичности позволяет составить своего рода «профиль» подкорпуса, определенного на основании некоторого свойства (тематики текста, психологической направленности текста) путем выявления наиболее характерных или нехарактерных для него признаков (начальных форм слова, псевдооснов). Этот «профиль» может быть использован для диагностики нового текста.

Данная работа в совокупности с работами [6, 7] констатирует возможность использования корпусной лингвистики для исследования свойств экстремистских текстов с целью обнаружения противоправных ресурсов и сообщений в Интернете.

Литература

1. Agarwal, S., et al, 2015, Open source social media analytics for intelligence andsecurity informatics applications. International Conference on Big Data Analytics. Hyderabad, Telangana State, India, pp. 21 37.
2. Scanlon, J.R., Gerber, M.S. Automatic detection of cyber-recruitment by violent extremists. Security Informatics, 2014, Vol. 3, No.1, pp. 1 10.
3. Zurini, M. Stylometry metrics selection for creating a model for evaluating the writing style of authors according to their cultural orientation. Informatica Economica, 2015, Vol. 19, No.3, pp. 107-119.
4. Hawkins, R. C. II, & Boyd, R. L. Such stuff as dreams are made on: Dream language, LIWC norms, and personality correlates. Dreaming, 2017, 27(2), 102-121.
5. Latov Y., Grishchenko L., Gaponenko V., Vasiliev F. Mechanisms of Countering the Dissemination of Extremist Materials on the Internet // Big Data-driven World: Legislation Issues and Control Technologies. – Springer, 2019. – P. 145-162.
6. 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
7. 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.
8. Ананьева М. И., Девяткин Д. А., Кобозева М. В., Смирнов И. В., Соловьев Ф. Н., Чеповский А. М. Исследование характеристик текстов противоправного содержания // Труды Института системного анализа Российской академии наук. 2017 Т. 67 № 3 С. 86-97.
9. Ананьева М. И., Кобозева М. В., Соловьев Ф. Н., Поляков И. В., Чеповский А. М. О проблеме выявления экстремистской направленности в текстах // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2016. Т. 14. № 4. С. 5-13.
10. Chepovskiy A., Devyatkin D., Smirnov I., Ananyeva M., Kobozeva M., Solovyev F. Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts), in: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017 Institute of Electrical and Electronics Engineers Inc., 2017 P. 188-190.
11. Мартыненко Г. Я., Шерстинова Т. Ю., Попова Т. И., Мельник А. Г., Замирайлова Е. В. О принципах создания корпуса русского рассказа первой трети XX века // Труды XV Международной конференции по компьютерной и когнитивной лингвистике «TEL 2018». – Казань, 2018. – С. 180–197.
12. Мартыненко Г. Я., Шерстинова Т. Ю., Мельник А. Г., Попова Т. И. Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) / Компьютерная лингвистика и вычислительные онтологии. Выпуск 2 (Труды XXI Международной объединенной конференции «Интернет и современное общество, IMS-2018, Санкт-Петербург, 30 мая - 2 июня 2018 г. Сборник научных статей»). — СПб: Университет ИТМО, 2018. С. 99-104.
13. Martynenko, G.Y., Sherstinova, T.Y. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski’s Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552. pp. 105–120.

14. Лаврентьев А. М., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М. Новый комплекс инструментов автоматической обработки текста для платформы TXM и его апробация на корпусе для анализа экстремистских текстов // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2018 Т. 16 № 3 С. 19-31.
15. Соловьев Ф. Н. Автоматическая обработка текстов на основе платформы TXM с учетом анализа структурных единиц текста // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18, №1. С. 74–82.
16. Чеповский А. М. Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.: Национальный открытый университет «ИНТУИТ», 2015.
17. Heiden S. The TXM Platorm: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme // 24th Pacifc Asia Conference on Language, Information and Computation - PACLIC24 / Ed. 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. URL: http://halshs.archiies-ouiertes.fr/halshs-00549764
18. Lafon P. Sur la variabilité de la fréquence des formes dans un corpus // Mots. 1980. № 1. P. 127-165.

Рецензент: *Баранов Александр Павлович, доктор физико-математических наук, академик Академии криптографии России, заместитель Генерального директора Главного научного инновационного внедренческого центра, г. Москва, Россия. E-mail:baranov.ap@yandex.ru*

COMPARATIVE ANALYSIS OF SPECIAL TEXT CORPORA FOR SECURITY-RELATED TASKS

Lavrentiev A.M.[8], Raybova D.M.[9], Tikhomirova E.A.[10], Fokina A.I.[11], Chepovskiy A.M.[12], Sherstinova T.Yu.[13]

The purpose of the study: *development of a technique for comparing special text corpora for subsequent use in the identification of extremist texts*

Method: *frequency methods and a specificity indicator for text analysis of the corpus platform TXM were used.*

Results: *a methodology for comparative analysis of special text corpora has been developed, which makes it possible to identify implicit links between corpora of heterogeneous texts; the relationships between the vocabulary of illegal and literary texts were revealed; the possibility of using the specificity index to compile a "profile" of a text subcorpus was shown; comparative analysis of the corpus of extremist texts and the corpus of Russian stories of the first third of the twentieth century was made; the relationships between the vocabulary of illegal and literary texts were revealed; the possibilities of using corpus linguistics to study the properties of extremist texts in order to detect illegal Internet resources and messages were shown; the possibilities of using both morphological characteristics of words and pseudo-bases of word occurrences in the analysis of specificity on corpus data have been examined; research results showed that the frequency analysis tools provided by the TXM platform are effective for applications when it is necessary to identify implicit lexical matches between different text corpora.*

Keywords: *corpus linguistics, automated text analysis, corpora analysis platform, specificity score, extremist texts*

8 Alexei Lavrentiev, Ph.D. (Philology), IHRIM Research Lab, CNRS & ENS de Lyon. Lyon, France. E-mail: alexei.lavrentev@ens-lyon.fr

9 Darya Ryabova, master, School of Business Informatics of National Research University Higher School of Economics Moscow. Moscow, Russia. E-mail: dmryabova@edu.hse.ru

10 Elizaveta Tikhomirova, Associate Professor, Department IC3 of Bauman Moscow State Technical University. Moscow, Russia. E-mail: elizarti@bmstu.ru

11 Alina Fokina, master's student, School of Business Informatics of National Research University Higher School of Economics. Moscow, Russia. E-mail: aifokina@edu.hse.ru

12 Andrey Chepovskiy, Dr. Sc. (Eng.), Professor, professor of Department of Information technology of Peoples Friendship University of Russia (RUDN University), Moscow, Russia and professor of Department of Information Security of National Research University Higher School of Economics. Moscow, Russia. E-mail: achepovskiy@hse.ru

13 Tatiana Sherstinova, Ph.D. (Philology), associate professor of the Department of Philology, National Research University Higher School of Economics, St. Petersburg, Russia, and associate professor of the Faculty of Philology, St. Petersburg State University. St. Petersburg, Russia. E-mail: tsherstinova@hse.ru

References:

1. Agarwal, S., et al, 2015, Open source social media analytics for intelligence andsecurity informatics applications. International Conference on Big Data Analytics. Hyderabad, Telangana State, India, pp. 21–37.
2. Scanlon, J.R., Gerber, M.S., Automatic detection of cyber-recruitment by violent extremists. Security Informatics, 2014, Vol. 3, No.1, pp. 1–10.
3. Zurini, M. Stylometry metrics selection for creating a model for evaluating the writing style of authors according to their cultural orientation. Informatica Economica, 2015, Vol. 19, No.3, pp. 107-119.
4. Hawkins, R. C. II, & Boyd, R. L. Such stuff as dreams are made on: Dream language, LIWC norms, and personality correlates. Dreaming, 2017, 27(2), 102-121.
5. Latov Y., Grishchenko L., Gaponenko V., Vasiliev F. Mechanisms of Countering the Dissemination of Extremist Materials on the Internet // Big Data-driven World: Legislation Issues and Control Technologies. – Springer, 2019. – P. 145-162.
6. Lavrentiev A. M., Smirnov I. V., Solovyev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M. Analis korpusov tekstov terroristicheskoi i antipravovoy napravlennosti // Voprosi kiberbezopasnosti. 2019. № 4(32). S. 54-60. DOI: 10.21681/2311-3456-2019-4-54-60
7. Lavrentiev A. M., Smirnov I. V., Solovyev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M. Sozdaniye spetsial’nyh korpusov tekstov na osnove rasshirennoy platformy TXM // Sistemy vysokoy dostupnosti. 2018. T. 14. № 3. S. 76-81.
8. Anan’yeva M. I., Devyatkin D. A., Kobozeva M. V., Smirnov I. V., Solov’yev F. N., Chepovskiy A. M. Issledovaniye harakteristik tekstov protivopravnogo soderzhaniya // Trudy Instituta sistemnogo analiza Rossiyskoy akademii nauk. 2017 T. 67 № 3 S. 86-97. (in Russian).
9. Anan’yeva M. I., Kobozeva M. V., Solov’yev F. N., Polyakov I. V., Chepovskiy A. M.. The problem of detection of extremist texts // Vestnik NSU. Series: Information Technologies. 2016. Vol. 14. № 4. S. 5-13.
10. Chepovskiy A., Devyatkin D., Smirnov I., Ananyeva M., Kobozeva M., Solovyev F. Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts), in: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017 Institute of Electrical and Electronics Engineers Inc., 2017 P. 188-190.
11. Martynenko, G., Sherstinova, T., Melnik, A., Popova, T., Zamirailova, E. O principakh sozdaniya korpusa russkogo rasskaza pervoj treti XX veka // Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics ‘TEL 2018’. Kazan. pp. 180–197. (in Russian).
12. Martynenko, G.Ya., Sherstinova, T.Yu., Melnik, A.G., Popova, T.I. (2018) Metodolo-gicheskie problemy sozdaniya Komp’yuternoj antologii russkogo rasskaza kak yazykovogo resursa dlya issledovaniya yazyka i stilya russkoj khudozhestvennoj prozy v ehpokhu revolyucionnykh peremen (pervoj treti XX veka)// Computational linguistics and computational ontologies. Issue 2 (Proceedings of the XXI International United Conference The Internet and Modern Society, IMS-2018, St. Petersburg, May 30 - June 2, 2018 Collection of scientific articles), ITMO University, St. Petersburg. Pp. 99–104. (in Russian).
13. Martynenko, G.Y., Sherstinova, T.Y. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski’s Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552. pp. 105–120.
14. Lavrentiev A. M., Solovyev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M. Novyy kompleks instrumentov avtomaticheskoy obrabotki teksta dlya platformy TXM i yego aprobatsiya na korpuse dlya analiza ekstremistskih tekstov // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Lingvistika i mezhkul’turnaya kommunikatsiya. 2018 T. 16 № 3 S. 19-31.
15. Soloviev F. N. Embedding Additional Natural Language Processing Tools into the TXM Platform. Vestnik NSU. Series: Information Technologies, 2020, vol. 18, no. 1, p. 74–82. (in Russian)
16. Chepovskiy A. M. Informatsionnyye modeli v zadachah obrabotki tekstov na yestestvennyh yazykah. Vtoroye izdaniye, pererabotannoye. M.: Natsional’nyy otkrytyy universitet “INTUIT”, 2015.
17. Heiden S. The TXM Platorm: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme // 24th Pacifc Asia Conference on Language, Information and Computation - PACLIC24 / Ed. 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. URL: http://halshs.archiies-ouiertes.fr/halshs-00549764
18. Lafon P. Sur la variabilité de la fréquence des formes dans un corpus // Mots. 1980. № 1. P. 127-165.