**А. М. Лаврентьев¹, Ф. Н. Соловьев², М. И. Суворова (Ананьева)³
А. И. Фокина⁴, А. М. Чеповский⁴**

¹ *Институт истории представлений и идей нового времени ИЦНИ
и Высшей нормальной школы Лиона
паперть Р. Декарта, 15, Лион, 69007, Франция*

² *Институт физико-технической информатики
Заводской проезд, 6, Протвино, Московская обл., 142284, Россия*

³ *Федеральный исследовательский центр «Информатика и управление» РАН
пр. Академика Вавилова, 44, корп. 2, Москва, 119333, Россия*

⁴ *Национальный исследовательский университет «Высшая школа экономики»
ул. Мясницкая, 20, Москва, 101000, Россия*

*ananyeva@isa.ru, alexei.lavrentev@ens-lyon.fr, the0@yandex.ru,
aifokina@edu.hse.ru, achepovskiy@hse.ru*

НОВЫЙ КОМПЛЕКС ИНСТРУМЕНТОВ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА ДЛЯ ПЛАТФОРМЫ ТХМ И ЕГО АПРОБАЦИЯ НА КОРПУСЕ ДЛЯ АНАЛИЗА ЭКСТРЕМИСТСКИХ ТЕКСТОВ^{*}

Платформа ТХМ предоставляет широкие возможности корпусного анализа, такие как анализ соответствий, кластеризация, построение лексических таблиц, поиск сложных лексических конструкций, выделение подкорпусов по различным параметрам. По умолчанию платформа работает со словоупотреблениями в качестве структурных единиц анализа. Она интегрирована с единственным расширением TreeTagger, позволяющим проводить лишь морфологический анализ и лемматизацию словоупотреблений. Однако пользователь может сопровождать каждое словоупотребление набором дополнительных характеристик, позволяющих существенно усложнить анализ, сделать его более гибким. В настоящей работе описывается разработанный нами набор утилит, позволяющий, опираясь как на наши собственные программные решения, так и на готовые средства анализа, расширить и усложнить анализ корпусов в платформе ТХМ. Особого внимания заслуживают выделение псевдоосновы в словах текста с использованием метода структурных схем и выявление именных групп в структуре текста. Эти расширения позволяют повысить эффективность таких используемых ТХМ методов, как анализ специфичности и анализ соответствий. В порядке апробации излагаются результаты эксперимента по анализу корпуса, содержащего тексты, оцененные экспертами как экстремистские, и «нейтральные» тексты схожей тематики (религия, политика, идеология). Все тесты показывают ярко выраженное противостояние нейтральных и маркированных текстов и позволяют на основе полученных результатов продолжить работу по автоматическому и полуматематическому выявлению потенциально противоправных текстов.

Ключевые слова: корпусная лингвистика, автоматический морфологический анализ, автоматический синтаксический анализ, платформа ТХМ, анализ соответствий, специфичность, выявление экстремистских текстов.

^{*} Работа выполнена при поддержке РФФИ, гранты № 16-29-09546 и 16-07-00641.

Введение

В настоящее время возрастает востребованность средств автоматизированного анализа текстов, позволяющих упростить работу с текстовой информацией. Такая необходимость становится все более заметной с ростом объемов доступных текстов. Средства автоматизированного анализа текста находят свое применение в широком диапазоне и прикладных, и исследовательских задач, таких как поиск информации, коммуникация, ведение электронного документооборота, анализ потоков данных, классификация текстов, выявление языковых закономерностей, построение и применение моделей структуры языка.

Решение сложных прикладных задач, связанных с языком, подразумевает анализ достаточно больших объемов текста. Потребность в таком анализе привела к формированию самостоятельного раздела прикладной лингвистики – корпусной лингвистики. Корпусами называются наборы текстов, подобранных и обработанных в соответствии с исследовательской задачей, выступающие в качестве базы для анализа того или иного языкового аспекта. Важным отличием корпусов от простых наборов текстов является их аннотация. Слова и предложения снабжаются лингвистической информацией, такой как морфологические характеристики слов, синтаксическая разметка предложений, семантические характеристики слов, предложений и словосочетаний.

Однако задача автоматического аннотирования текстов грамматическими, синтаксическими и прочими характеристиками оказалась отнюдь не тривиальной. Дело в том, что на данный момент автоматические средства аннотации показывают сравнительно хорошее качество в основном при решении относительно простых задач (например, задачи частеречной разметки) и, как правило, для небольшого числа достаточно распространенных языков с нормированной орфографией.

Для облегчения же автоматического поиска в корпусе различных языковых конструкций, сопоставления их друг с другом, вычисления самых разнообразных статистических характеристик таких конструкций, параметрического выделения частей корпуса для анализа и сравнения требуется новое программное обеспечение, позволяющее успешно решать широкий круг этих задач. Начало их решению уже положено, описание широкого ряда методов, применяемых для такого анализа, приведено в [Чеповский, 2015].

В настоящей работе мы опираемся на программный комплекс – платформу ТХМ [Heiden, 2010], предлагающую широкий набор возможностей для корпусного анализа.

Вместе с тем мы также предлагаем ряд расширений, позволяющих дополнить и усложнить анализ корпусов, включающий: автоматический морфологический анализ словоформ и приведение их к канонической форме, выделение псевдооснов (аналога корня слова), выделение именных групп и комбинирование результатов работы предлагаемых расширений. Отдельного разъяснения заслуживает анализ и выделение псевдоосновы в словах текста с использованием метода структурных схем. Его преимущество в том, что такой подход позволяет не только анализировать текстовые конструкции, опираясь на точные словоформы, что в определенной степени ограничивает полноту и гибкость анализа, но и задавать более сложные запросы к корпусу, включающие различные комбинации псевдооснов и грамматических характеристик слова. Кроме того, метод структурных схем легко масштабируется. Так, в предлагаемом нами расширении уже поддерживается более 60 языков, и это значит, что в будущем поддержка может быть оперативно расширена на другие языки по мере приобретения их дискурсами своего особого места в межкультурной коммуникации.

Морфология словарная

При анализе содержания текста нередко бывает целесообразно и практически удобно рассматривать различные словоформы одной лексемы как разные употребления одного и того же (в лексическом отношении) слова. Один из подходов к решению этой задачи, предложенный в [Чеповский, 2015], состоит в автоматическом морфологическом анализе словоформ с последующим приведением их к канонической форме (т. е. лемме).

Возможность привести словоформу к канонической форме позволяет анализировать различные элементы словоизменительной парадигмы как одну и ту же структурную единицу

текста. Это, в свою очередь, позволяет более корректно проводить содержательный статистический анализ текста, например, путем рассмотрения частот лексем вместо частот отдельных словоформ.

Что касается реализации автоматического анализа словарной морфологии, используемого в настоящей работе, то он основан на словаре А. А. Зализняка [1977]. В словаре содержатся словоформы вместе с их грамматическими пометами, что полностью отвечает задачам нашего исследования. А морфологическая модель, используемая в нем, относит каждое слово к одной из 24 морфологических категорий, близких к части речи в традиционном понимании:

0. Неизменяемое слово
1. Существительное
2. Прилагательное
3. Глагол несовершенного вида
4. Предлог
5. Глагол совершенного вида
6. Количественное числительное
7. Порядковое числительное
8. Местоимение
9. Местоименное прилагательное
10. Собирательное числительное
11. Сокращение
12. Латинское слово
13. Аббревиатура
14. Фамилия
15. Имя
16. Отчество
17. Причастие
18. Союз
19. Наречие
20. Частица
21. Междометие
22. Топоним
23. Субстантивированное прилагательное

Каждая из этих морфологических категорий характеризуется набором грамматических категорий: род, падеж, число, наклонение и др.

Программная реализация словарной морфологии русского языка опирается на представление словаря в специализированной структуре данных «бор», что позволяет осуществлять поиск словоформ за линейное по числу букв словоформы время. Каждая словоформа, хранящаяся таким образом, содержит свои грамматические характеристики, а также информацию о ее канонической (начальной) форме и об основе словоформы.

Псевдоосновы

К сожалению, словарная морфология не позволяет анализировать неологизмы и жаргонизмы. Между тем в ситуации, когда требуется анализировать короткие тексты сети Интернет, принадлежащие к таким жанрам, как комментарии и сообщения, доля таких слов может быть значительной. По этой причине мы, помимо словарной морфологии, использовали и аналитический метод выделения псевдооснов – формального аналога корня слов, что также позволило избежать анализа частот отдельных словоформ.

Данный способ выделения псевдооснов представляет собой метод структурных схем, описанный подробно в [Болховитянов, Чеповский, 2011; Egorova et al., 2016]. Суть метода состоит в получении псевдоосновы словоформы путем рассмотрения и отбрасывания ее словоизменяемых аффиксов. Словообразовательные аффиксы считаются элементом корневой части и не отбрасываются. С каждым словом можно сопоставить отвечающую ему последовательность аффиксов. Такие последовательности называются *структурами* некор-

невой части слова. Отсюда происходит название метода. Как и в традиционном морфологическом анализе, аффиксы подразделяются на префиксы и суффиксы в соответствии с их позицией относительно корня слова. Псевдоосновой называется часть слова, не содержащая суффиксов и префиксов. Псевдооснова не всегда совпадает с основой слова в традиционном понимании, поскольку выделяется автоматически, и в ряде случаев реальные аффиксы словоформы не распознаются, либо за аффиксы принимается часть корня слова. Способ автоматического выделения псевдооснов состоит в сопоставлении рассматриваемой словоформы с множеством допустимых в языке структур некорневой части слова. Псевдооснова слова выделяется отбрасыванием всех соответствующих определенной структурной схеме аффиксов (т. е. допустимой в данном языке максимальной комбинации префиксов и суффиксов).

В русском языке префикс редко выполняет словоизменительную функцию, поэтому на практике выделение псевдооснов состоит в отбрасывании суффиксов. Например, в словоформе *находившихся* отбрасываются суффиксы *-и* и *-ви*, флексия *-их* и постфикс *-ся*, в результате чего получается псевдооснова *наход-*. В то же время отбрасывание похожих на аффикс сегментов приводит в ряде случаев к некорректным с точки зрения лингвистического анализа результатам. Например, для словоформы *верность* выделяется псевдооснова *вернос-*. Эти «побочные эффекты» автоматической обработки не влияют, однако, на качество статистического анализа.

Именные группы

Дополнительную информацию о специфическом содержании текста можно получить, анализируя не только словоформы, но и целые именные группы. Именная группа определяется нами как группа слов, у которой главное слово существительное, а другие слова связаны с ним подчинительными синтаксическими связями. Рассмотрение частотных именных групп и их сочетаний в совокупности с анализом отдельных словоупотреблений позволяет получить более полную картину семантических и стилистических характеристик текста, релевантных его содержанию.

Определенную сложность при выделении именных групп представляет множественность морфологических разборов при омонимии. В ходе анализа слов в предложении наш метод предполагает рассмотрение всего множества возможных морфологических разборов каждого слова.

Используемый нами алгоритм основан на [Белоногов, Богатырёв, 1974] и подробно описан в [Чеповский, 2015]. Он работает в три этапа, после того как текст разделен на предложения:

- 1) установление подчинительных синтаксических связей в предложении;
- 2) установление синтаксических связей внутри конструкций с однородными членами;
- 3) выделение именных групп.

Установление подчинительных синтаксических связей

Подчинительные связи устанавливаются между парами слов, где, учитывая порядок слов в предложении, одно – левое, второе – правое. Подчинительные связи устанавливаются в соответствии со следующими тремя правилами.

1. Если в паре слов существует морфологический разбор левого слова, где оно идентифицировано как существительное, и существует морфологический разбор правого слова, где оно – существительное в родительном падеже, то правое слово синтаксически подчинено левому.

2. Если в паре слов существует морфологический разбор левого слова, где оно – прилагательное, и существует морфологический разбор правого слова, где оно – существительное, и у пары слов совпадают падеж, род, и либо совпадает число, либо левое слово стоит в единственном числе, то левое слово подчинено правому.

3. Если в паре слов существует морфологический разбор левого слова, где оно – количественное числительное, и существует морфологический разбор правого слова, где оно – су-

существительное, и правое слово стоит в родительном падеже, а левое слово стоит либо в именительном, либо в родительном падеже, то правое слово подчинено левому.

Так, в конструкции *оценка состояния и перспектив российской экономики* на первом этапе выделяются следующие связи: *оценка > состояние*, *перспектива > экономика*, *экономика > российская*.

Установление синтаксических связей внутри конструкций с однородными членами

Пусть имеется множество связок однородных членов предложения. Элементы этого множества находятся между однородными членами предложения.

Список существительных, являющихся главными словами в цепочке однородных членов (сокращенно – «однородных существительных»), формируется в соответствии со следующими двумя правилами.

1. Существительные разделены связками из заданного в языке множества связок.
2. Существительные имеют совпадающие или пересекающиеся (по морфологическим разборам) значения падежа.

В каждом списке «однородных существительных» подчинительные связи могут быть изменены в соответствии со следующими двумя правилами.

1. Если последние два существительных списка разделены союзом «и», а у последнего слова существует зависимое существительное *w*, и если у предпоследнего слова нет ни одного зависимого существительного, то *w* считается подчиненным предпоследнему слову (так же, как и последнему). Если же у предпоследнего слова имеется хотя бы одно зависимое существительное, то на этом обработка списка однородных существительных прекращается.

2. Если все слова списка находятся в родительном падеже и в предложении есть слова, стоящие в именительном падеже, которым подчинено первое слово, то все прочие слова списка считаются подчиненными им.

В результате этого этапа в ранее процитированном примере будут выявлены дополнительные связи: *оценка > перспектива* и *состояние > экономика*.

Выделение именных групп

Все построенные подчинительные связи образуют граф – связанную структуру, узлами которой является множество слов предложения, составляющих каждую именную группу, а связями – множество подчинительных связей, где последнее рассматривается как объединение двух множеств: множества связей между парами существительных, и множество связей, где прилагательное подчинено существительному. Именными группами являются пути цепочки последовательно связанных подчинительными связями слов, частеречная принадлежность которых определяется следующими двумя правилами.

1. Если длина цепочки равна двум, то оба слова – существительные.
2. Если длина цепочки больше двух, то в цепочке есть хотя бы одно прилагательное, подчиненное существительному.

В итоге в приведенном выше примере будут выявлены следующие именные группы:

- *экономика > российский*
- *перспектива > экономика > российский*
- *состояние > экономика > российский*
- *оценка > перспектива*
- *оценка > состояние*
- *оценка > состояние > экономика*
- *оценка > перспектива > экономика*
- *оценка > перспектива > экономика > российский*
- *оценка > состояние > экономика > российский*

TreeTagger

В настоящей работе мы также использовали программный пакет TreeTagger [Schmid, 1994], предоставляющий возможность совместного морфологического анализа слов предложения на основе статистической модели путем сопоставления словоупотреблений, снабженных специальными метками, кодирующими морфологические характеристики. TreeTagger также предоставляет возможность лемматизации, т. е. определения начальных форм слов предложения. Преимуществом пакета является однозначность морфологического анализа, поскольку при сопоставлении меток словоупотреблений учитываются все слова предложения в совокупности. Однако при таком анализе существует риск ошибок, который возрастает, если текст содержит большое количество неологизмов и нестандартных написаний слов.

Платформа ТХМ

Платформа ТХМ, разрабатываемая с 2007 г., представляет собой программное обеспечение с открытым кодом, предназначенное для подготовки, обработки, анализа и публикации корпусов среднего размера (примерно до 10 000 000 словоупотреблений). В настоящее время ТХМ развивается в лаборатории Institut d'Histoire des Représentations et des Idées dans les Modernités (IHRIM)¹ Высшей нормальной школы Лиона и Национального центра научных исследований Франции и в лаборатории Edition, Littératures, Langages, Informatique, Arts, Didactique, Discours (ELLIAD)² университета Франш-Комте (г. Безансон). ТХМ распространяется свободно на условиях Стандартной общественной лицензии GNU (GPL) версии 2³ в виде приложения для операционных систем Linux, Windows и Mac OS, а также в виде приложения для интернет-портала (на основе технологии GWT⁴). Исходный код ТХМ доступен на портале sourceforge.net. ТХМ включает в себя поисковую машину CQP, платформу статистического анализа R⁵ и ранее упомянутый пакет автоматической морфологической разметки и лемматизации TreeTagger⁶. Платформа поддерживает широкий спектр форматов корпусов на входе (от простого текста до TEI XML⁷). Особенностью ТХМ является гармоничное сочетание инструментов количественного и качественного анализа. От статистических инструментов (таких, как факторный анализ и специфичность, которые будут представлены ниже) в ней всегда можно перейти к конкордансам значимых для анализа словоупотреблений и к более широкому контексту с возможностью сохранения особенностей оформления текста (шрифты, иллюстрации, расположение на странице) и доступа к факсимильному изображению оригинального документа.

Открытый код и модульная архитектура позволяют также интегрировать в ТХМ внешние инструменты обработки и анализа текстовых данных.

Интеграция с ТХМ

Все упомянутые средства были объединены нами в набор утилит, дающий возможность вычислять для текстовых корпусов ряд характеристик достаточно широкого круга языковых единиц, входящих в их состав, например:

- 1) словоформы, встреченные в исходном корпусе;
- 2) начальные формы слов по словарной морфологии;
- 3) морфологические характеристики слов по словарной морфологии;
- 4) начальные формы слов, полученные с помощью TreeTagger;
- 5) морфологические характеристики, полученные с помощью TreeTagger;
- 6) псевдоосновы слов;

¹ Институт истории представлений и идей нового времени, <http://ihrim.ens-lyon.fr>.

² Издание, Литература, Языки, Информатика, Искусства, Дидактика и Дискурс, <http://elliadd.univ-fcomte.fr>.

³ <https://www.gnu.org/licenses/old-licenses/gpl-2.0.html>

⁴ <http://www.gwtproject.org>

⁵ <https://www.r-project.org>

⁶ Требуется отдельной установки.

⁷ <http://www.tei-c.org>

7) именные группы, составленные из словоформ (п. 1), вместо отдельных словоупотреблений;

8) именные группы, составленные из начальных форм (п. 2), вместо отдельных словоупотреблений;

9) именные группы, составленные из морфологических характеристик (п. 3), вместо отдельных словоупотреблений;

10) именные группы, составленные из начальных форм, полученных с помощью Tree-Tagger (п. 4), вместо отдельных словоупотреблений;

11) именные группы, составленные из морфологических характеристик (п. 5), вместо отдельных словоупотреблений;

12) именные группы, составленные из псевдооснов (п. 6), вместо отдельных словоупотреблений.

Корпуса с вычисленными характеристиками преобразуются нами в формат для импорта пакетом TXM.

Специфичность подкорпуса в лингвостатистическом аспекте

Удобным инструментом количественной оценки «необычности» специального подкорпуса относительно всего корпуса является показатель *специфичности* [Lafon, 1980].

Лингвостатистическую суть понятия специфичности можно разъяснить следующим образом. Пусть имеется разбиение корпуса T на m подкорпусов. Специфичностью слова w в одном рассматриваемом подкорпусе относительно всего корпуса является вероятность того, что в *случайно* выбранном из корпуса T подкорпусе известного размера слово w встретится ровно столько раз, сколько оно встретилось в рассматриваемом подкорпусе. Иначе говоря, это вероятность случайно выбрать аналогичный рассматриваемому подкорпус из всего корпуса с известным количеством вхождений слова w , равным числу его вхождений в рассматриваемый подкорпус.

Можно показать, что максимума такая вероятность достигает тогда, когда относительная частота слова в рассматриваемом подкорпусе совпадает с относительной частотой этого же слова во всем корпусе. Таким образом, малое значение специфичности слова (или другой единицы анализа) говорит о том, что рассматриваемый подкорпус, с точки зрения распределения в нем слова w относительно распределения этого же слова во всем корпусе, – «необычный», высокое значение вероятности – о том, что подкорпус «обычный». «Индекс специфичности» представляет собой порядковую величину вероятности. Иными словами, индекс специфичности равен 2, если вероятность составляет 1 из 100, а индекс, равный 10, соответствует вероятности $1/10^{10}$. Индекс может быть отрицательным, если искомая единица встречается в подкорпусе реже, чем в среднем по корпусу. Считается, что индекс специфичности в пределах ± 2 является «банальным» и не должен использоваться для характеристики подкорпуса.

Анализ специфичности позволяет составить своего рода «профиль» подкорпуса, выделенного на каких-либо внешних основаниях (например, автор, жанр, тематика или идеологическая направленность текста) путем выявления наиболее характерных или нехарактерных для него словоформ (лексем, псевдооснов, именных групп и т. п.). В дальнейшем этот «профиль» может быть использован для диагностики нового текста, нуждающегося в анализе.

Анализ соответствий

Другим подходом к анализу разделенного на части (подкорпуса) по определенному критерию корпуса является анализ соответствий. Пусть имеются две «категориальные» (т. е. принимающие конечное множество неупорядоченных значений, например: пол, знак зодиака, падеж и т. п.) переменные X и Y и стоит задача определить, являются ли они независимыми. Подход анализа соответствий состоит в анализе частот совместного появления значений этих переменных на предмет равномерного распределения за счет поиска наиболее информативного представления таблицы частот совместного появления переменных X и Y в виде на-

бора точек на плоскости, где точки отвечают либо значениям переменной X , либо значениям переменной Y .

Методика анализа соответствий, используемая ТХМ, была предложена Ж.-П. Бензекри [Benzecri, 1979] и имплементирована в пакете FactoMineR для платформы R [Lê et al., 2008].

Описание корпуса

Проверка эффективности разработанных утилит была проведена на корпусе из 709 текстов общим объемом 900 000 словоупотреблений. Специфика корпуса состоит в том, что он включает в свой состав тексты противоправной направленности, т. е. оцененные экспертами как «экстремистские», и тексты схожей тематики, не имеющие такой окраски («нейтральные»). «Экстремистские» тексты разделяются на 7 тематических групп, каждая из которых рассматривается как отдельный подкорпус: «агрессия», «идеология», «национализм», «религия», «сепаратизм», «терроризм» и «фашизм». Принципы формирования специального корпуса и прочая релевантная информация подробно изложены в [Ананьева и др., 2016; 2017]. В соответствии с поставленными задачами корпус был проанализирован с использованием двух обозначенных выше функций ТХМ – специфичность и анализ соответствий, результаты их анализа можно представить в графической интерпретации. Детально были рассмотрены следующие лексические объекты: леммы, псевдоосновы и именные группы, по которым первоначально были сделаны предположения о схожести результатов статистического анализа.

Анализ корпуса

Вследствие понижения размерности матрицы, используемой в ходе анализа соответствий, изначально состоящей из 7 столбцов, соответствующих условному делению текстов экстремистской направленности на подкорпуса, есть возможность интерпретировать пространственное расположение подкорпусов относительно друг друга как оценку, т. е. как характеристику, указывающую на сходство или различие маркированных подкорпусов между собой и по отношению к «нейтральному» подкорпусу. Расположение вблизи пересечения осей координат следует интерпретировать как нейтральность по отношению к корпусу в целом, основную массу которого составляют «нейтральные» тексты. Отдаление же от начала координат, напротив, указывает на отклонение от идеологически и политически нейтрального текста (т. е. от его «нормальности») в пользу одной из позиций, отмеченной как противоправная по результатам экспертного анализа соответствующего текста.

Анализ соответствий демонстрирует пространственное расположение подкорпусов на основе анализа частот совместного появления значений переменных. Как можно заметить из рис. 1, точка «нейтрального» подкорпуса находится вблизи точки пересечения осей координат. В то же время «противоправные» подкорпуса в большинстве своем удалены от «нейтрального» и располагаются в непосредственной близости друг от друга, образуя полевой кластер (или «облако») единиц. Это противопоставление наиболее четко проявляется в первом, наиболее важном, измерении (ось 1 на графике), отвечающем за 50 % вариаций в корпусе. При изменении параметров корпуса (увеличение / уменьшение объема нейтрального подкорпуса, исключение пунктуации, повышение минимальной частотности) противопоставление нейтральных и маркированных текстов в этом измерении остается стабильным. Во втором измерении, отвечающем за 13 % вариаций корпуса (ось 2), результаты анализа сильно зависят от конфигурации корпуса и нуждаются в дополнительной интерпретации. На данном графике тексты, отнесенные к категории «фашизм», противостоят всем остальным, образующим компактное «облако» вблизи центральной оси. Это происходит за счет высокой частотности таких лексем, как «раса», «наследственность», «нордический» и «отбор», характерных для фашистского дискурса.

Аналогичную картину можно наблюдать и на рис. 2, где «нейтральный» подкорпус опять же находится в отдалении от каждого из «экстремистских» на первой (горизонтальной) оси. Подкорпуса «идеология», «религия», «терроризм», «национализм», «агрессия» и «сепаратизм», выявленные на основе экспертной оценки, близки по своим показателям. Немалую значимость имеет также близость расположения «религии» и «идеологии» как на рис. 1, так

и на рис. 2. В целом, можно отметить, что «облако» подкорпусов в первой четверти осей координат в анализе по леммам и псевдоосновам дает схожую картину, однако во втором случае наблюдается их чуть большее рассеивание относительно оси ординат. Подкорпус «фашизм», находящийся в четвертой четверти координатной плоскости, близок ко всем потенциально экстремистским по 1-й оси и противостоит им по 2-й оси за счет частотности уже упомянутых лексем фашистского дискурса.

Специфичность (рис. 3), в отличие от анализа соответствий (см. рис. 1, 2), вычисляется для отдельных значений слова и показывает оценку появления того или иного слова (или другой структурной единицы) в избытке либо недостатке в определенном подкорпусе по сравнению с его «родительским» корпусом и иными подкорпусами. Таким образом,

Рис. 1. Анализ соответствий по леммам

Рис. 2. Анализ соответствий по псевдоосновам

Рис. 3. Специфичность по именным группам

можно отобрать наиболее специфичные лексемы (псевдоосновы, именные группы и т. д.) определенного подкорпуса и сравнить их индексы специфичности в каждом из остальных подкорпусов. На графике каждый подкорпус представлен набором смежных полос (см. рис. 3). Каждое свойство рассматриваемой единицы, в данном случае это именные группы, изображено штрихом одного и того же цвета в различных частях диаграммы, соответствие цветов конкретным именным группам дано справа от диаграммы. Две красные линии со значениями 2.0 и -2.0 ограничивают «область банальности» вблизи оси индекса, равного нулю; бары, которые не выходят за ее пределы, следует считать «банальными», т. е. неспецифичными для данного подкорпуса.

В приведенном нами случае представлены значения специфичности по убыванию для подкорпуса «идеология». Как можно судить из результатов, показанных на рис. 3, индексы специфичности псевдооснов в данной части во многом схожи с аналогичными значениями подкорпуса «религия»: они имеют по несколько близких индексов специфичности для именных групп «агрессия» и «национализм», и один для группы «терроризм». В то же время для «нейтрального» подкорпуса все результаты вычисления индекса специфичности для тех же самых именных групп расположены в области отрицательных значений, что свидетельствует об их низкой специфике, т. е. о статистически значимой редкости появления данных именных групп в нейтральной части корпуса.

Выводы

Проведенный нами эксперимент по интеграции инструментов автоматической обработки текста (ограниченный выделением псевдооснов и именных групп) и платформы корпусного анализа ТХМ показал, что такая интеграция не только технически выполнима, но и позволяет расширить возможности статистического анализа текстов.

Сопоставляя полученные результаты по графическому анализу соответствий по леммам и псевдоосновам и специфичностей по именным группам, можно сделать следующие общие выводы:

- для лемм, псевдооснов и именных групп результаты в целом схожи;
- нейтральный подкорпус противостоит каждому из «экстремистских»;
- высокой схожестью обладают подкорпуса «религия» и «идеология».

В силу выявленных особенностей и противопоставленности нейтрального подкорпуса остальным, сформированный корпус может быть использован для машинного обучения в зада-

чах классификации текстов на предмет выявления потенциально противоправного содержания с целью их углубленного экспертного анализа.

В ходе дальнейших исследований мы планируем расширить набор применяемых методов статистического и качественного анализа корпуса, с тем чтобы создать статистически надежную и лингвистически мотивированную процедуру выявления экстремистской направленности текстов.

Список литературы

Ананьева М. И., Кобозева М. В., Соловьев Ф. Н., Поляков И. В., Чеповский А. М. О проблеме выявления экстремистской направленности в текстах // Вестн. НГУ. Серия: Информационные технологии. 2016. Т. 14, № 4. С. 5–13.

Ананьева М. И., Девяткин Д. А., Кобозева М. В., Смирнов И. В., Соловьев Ф. Н., Чеповский А. М. Исследование характеристик текстов противоправного содержания // Тр. Ин-та системного анализа РАН. 2017. Т. 67, № 3. С. 86–97.

Белоногов Г. Г., Богатырёв В. И. Автоматизированные информационные системы. М.: Сов. радио, 1973.

Болховитянов А. В., Чеповский А. М. Методы автоматического анализа словоформ // Информационные технологии. 2011. № 4 (176). С. 24–29.

Зализняк А. А. Грамматический словарь русского языка. М.: Русский язык, 1977.

Чеповский А. М. Информационные модели в задачах обработки текстов на естественных языках. 2-е изд., перераб. М.: Национальный открытый университет «ИНТУИТ», 2015.

Benzécri J.-P. L'analyse des données: l'analyse des correspondances. 2nd ed. Paris: Dunod, 1979. Vol. 2.

Egorova E., Chepovskiy A., Lavrentiev A. A structural pattern based method for automated morphological analysis of word forms in a natural language // Journal of Mathematical Sciences. 2016. Vol. 214. No. 6. P. 802–813.

Heiden S. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme // 24th Pacific Asia Conference on Language, Information and Computation – PACLIC24 / Eds. R. Otaguro, K. Ishikawa, H. Umemoto, K. Yoshimoto, Y. Harada. Institute for Digital Enhancement of Cognitive Development. Sendai, Japan: Waseda University, 2010. P. 389–398.

Lafon P. Sur la variabilité de la fréquence des formes dans un corpus // Mots. 1980. № 1. P. 127–165.

Lê S., Josse J., & Husson F. FactoMineR: an R package for multivariate analysis // Journal of Statistical Software. 2008. № 25 (1) P. 1–18.

Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. 1994. URL: <http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf>

Alexei M. Lavrentiev¹, **Fedor N. Solovyev**², **Margarita I. Suvorova (Ananyeva)**³
Alina I. Fokina⁴, **Andrey M. Chepovski**⁴

¹ *IHRIM Research Lab, CNRS & ENS de Lyon
 15 parvis René-Descartes, Lyon, 69007, France*

² *Institute of Physical and Technical Informatics
 6 Zavodskoy pass., Protvino, Moscow Region, 142284, Russian Federation*

³ *Federal Research Center – Computer Science and Control, RAS
 44 Academician Vavilov Ave., Moscow, 119333, Russian Federation*

⁴ *National Research University – Higher School of Economics
 20 Myasnitskaya Str., Moscow, 101000, Russian Federation*

*alexei.lavrentev@ens-lyon.fr, the0@yandex.ru
 ananyeva@isa.ru, aifokina@edu.hse.ru, achepovski@hse.ru*

A NEW TOOLKIT FOR NATURAL TEXT PROCESSING WITH THE TXM PLATFORM AND ITS APPLICATION TO A CORPUS FOR ANALYSIS OF TEXTS PROPAGATING EXTREMIST VIEWS

TXM platform provides a wide range of corpus analysis tools including correspondence analysis, clustering, lexical table construction, and parametrized subcorpus selection. The default structural unit of analysis for TXM is a token. The only TXM extension available by default is TreeTagger which performs automated morphological analysis and lemmatization during the corpus import process. However, it is possible to supply each token with a number of features enabling a more advanced text analysis. In this work we present a number of tools developed for even a more extensive, complex and flexible corpus analysis with TXM relying both on the tools previously developed by our team and on publicly available software libraries. We focus in particular on a stemming technique that uses a word structural pattern method and on noun phrase recognition that together make it possible to perform more sophisticated and powerful queries and analyses of the corpus not limited to word forms.

The structural pattern stemming method is based on a set of specific language rules that allow separating a word stem from all affixes. The recognition of noun phrases is based on rules allowing the detection of subordination and coordination relations among nouns. These extensions result in the improvement of performance of statistical tools used by TXM, such as specificity scores and correspondence analysis.

The new set of tools has been tested on a corpus including texts marked as «extremist» by experts along with «neutral» texts in similar domains. The corpus of approximately 900,000 words is divided into eight subcorpora: neutral texts oppose seven thematic subcorpora considered as extremist (namely aggressive, fascist, ideological, nationalistic, religious, separatist, and terroristic). The specificity analysis detects the words (or other structural units) that are significantly more or less frequent in a given subcorpus compared to the entire corpus. The specificity score for selected units can be compared across all the subcorpora in order to verify their difference or similarity. The correspondence analysis produces a chart where the subcorpora are represented as points in a two-dimensional space based on their similarity as to the frequency of selected units.

All tests demonstrated a significant difference between neutral texts, on one side, and marked, on the other. Two «extremist» subcorpora, religious and ideological, demonstrated similar results and can probably be merged. These facts encourage further research on fully automatic or computer-aided expert recognition of extremist texts.

Keywords: corpus linguistics, automated morphological analysis, automated syntactic parsing, TXM platform, correspondence analysis, specificity, detecting extremist texts.

References

Anan'eva M. I., Devyatkin D. A., Kobozeva M. V., Smirnov I. V., Solov'ev F. N., Chepovskiy A. M. Issledovanie kharakteristik tekstov protivopravnogo sodержaniya [Research on peculiarities of text with illicit content]. *Proceedings of the Institute for Systemic Analysis of the Russian Academy of Sciences*. Moscow, Federal'nyy issledovatel'skiy tsentr «Informatika i upravlenie» Rossiyskoy akademii nauk, 2017, vol. 67, no. 3, p. 86–97. (in Russ.)

Anan'eva M. I., Kobozeva M. V., Solov'ev F. N., Polyakov I. V., Chepovskiy A. M. O probleme vyyavleniya ekstremistskoy napravlennosti v tekstakh [The Problem of Detection of Extremist Texts]. *Vestnik NSU. Series: Information Technologies*, 2016, vol. 14, no. 4, p. 5–13. (in Russ.)

Benzécri J.-P. L'analyse des données: l'analyse des Correspondances, 2nd ed. Paris, Dunod, 1979, vol. 2.

Bolkhovityanov A. V., Chepovskiy A. M. Metody avtomaticheskogo analiza slovoform [Methods of automatic word-form analysis]. *Informatsionnye tekhnologii*, 2011, no. 4 (176), p. 24–29. (in Russ.)

Chepovskiy A. M. Informatsionnye modeli v zadachakh obrabotki tekstov na estestvennykh yazykakh [Information models in natural language text processing problems]. 2nd ed. Moscow, Natsional'nyy otkrytyy universitet «INTUIT», 2015. (in Russ.)

Egorova E., Chepovskiy A., Lavrentiev A. A structural pattern based method for automated morphological analysis of word forms in a natural language. *Journal of Mathematical Sciences*, 2016, vol. 214, no. 6, p. 802–813.

Heiden S. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. 24th Pacific Asia Conference on Language, Information and Computation – PACLIC24. Eds. R. Otaguro, K. Ishikawa, H. Umemoto, K. Yoshimoto, Y. Harada. Institute for Digital Enhancement of Cognitive Development. Sendai, Japan, Waseda University, 2010, p. 389–398.

Lafon P. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1980, no. 1, p. 127–165.

Lê S., Josse J., & Husson F. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*, 2008, no. 25 (1), p. 1–18.

Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK. 1994. URL: <http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf>

Zaliznyak A. A. Grammaticheskiy slovar' russkogo yazyka [Grammatical dictionary of the Russian language]. Moscow, Russkiy yazyk, 1977. (in Russ.)

For citation:

Lavrentiev Alexei M., Solovyev Fedor N., Suvorova (Ananyeva) Margarita I., Fokina Alina I., Chepovskiy Andrey M. A New Toolkit for Natural Text Processing with the TXM Platform and its Application to a Corpus for Analysis of Texts Propagating Extremist Views. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2018, vol. 16, no. 3, p. 19–31. (in Russ.)

DOI 10.25205/1818-7935-2018-16-3-19-31