

Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований.

Часть 2. Машина PCA и опыт ее использования*

Ю. М. Кузнецова, И. В. Смирнов, М. А. Станкевич, Н. В. Чудова

Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия

Аннотация. Рассмотрены наиболее известные инструменты лингво-статистического анализа корпусов текстов и представлен новый инструмент анализа текстов для социо-гуманитарных исследований – машина PCA. Описаны архитектура и средства программной реализации машины PCA. Представлены результаты пилотажного исследования текстов с использованием PCA. Пилотажное исследование возможностей нового инструмента проведено на материале 142 текстов эссе, написанных на тему «Я. Другие. Мир». Корреляционный анализ показал наличие связей между текстовыми показателями, выявляемыми с помощью машины PCA (частота различных типов предикатов и типов синтаксиса), и данными психодиагностического исследования, в частности, уровнем агрессивности. Предложена схема интерпретации данных предикатного анализа, позволяющая описать специфику речемыслительной деятельности и картины мира в зависимости от групповых и личностных особенностей авторов текстов.

Ключевые слова: анализ корпусов текстов, программная архитектура, графовая база данных, семантико-синтаксические конструкции, социо-гуманитарные исследования, картина мира.

DOI 10.14357/20718594190305

Введение

Данная публикация является второй частью статьи, посвященной предлагаемой методологии применения средств автоматического анализа текста в исследованиях социальных и гуманитарных дисциплин [1]. В первой части представлен обзор в этой области и даны основные положения подхода, развиваемого авторами. Во второй части описаны функции машины PCA и ее архитектура; представлены данные пилотажного исследования, демонстрирующие возможности использования разработанной системы в интересах психологии.

Сегодня существует достаточно много инструментов для лингво-статистического анализа корпусов текстов. Рассмотрим коротко наиболее известные из них. ТХМ – платформа с открытым исходным кодом для разметки и анализа корпусов текстов [2]. В ТХМ можно работать с корпусами на десятках языков, в том числе русском. Все доступные функции ТХМ можно поделить на поисковые и аналитические. К поиску относятся сфокусированный поиск (определение интересующего события, к примеру, появление некоторой словоформы/леммы/конструкции в корпусе) и несфокусированный/панорамный поиск (рассматривается корпус в целом, к примеру, выводится

*Работа выполнена при финансовой поддержке РФФИ (№17-29-02247).

✉ Кузнецова Юлия Михайловна. E-mail: kuzjum@yandex.ru

общий частотный словарь). Также доступны анализ конкордансов (легкий переход от частотного словаря к конкордансу и от конкорданса к тексту корпуса), совместная встречаемость, построение графика распределения той или иной леммы/словоформы по корпусу. К основным аналитическим функциям относятся: вычисление специфичности (вероятности появления некоторой словоформы/леммы/конструкции в определенной части корпуса) и факторный анализ (картография корпуса), который позволяет оценить, насколько сильно отличаются друг от друга разные корпуса (в зависимости от задач исследования могут отличаться по жанрам/ времени публикации и т.д.).

Sketch Engine – коммерческая платформа для работы с текстовыми корпусами [3]. Позволяет работать с текстами на 60 языках, а также содержит в себе предзагруженные словари и корпуса, в том числе параллельные, на этих языках. Sketch Engine может получать всю словарную информацию о грамматических формах, вариантах употребления и устойчивых выражениях, в которых встречается некоторое слово, а также выводить примеры из интересующего корпуса. К другим функциям относится составление конкордансов и тезаурусов на основе общих употреблений (если два слова встречаются в похожих контекстах, значит понятия, которые они обозначают, тоже близки). NoSketch Engine — это ограниченная версия Sketch Engine с открытым исходным кодом.

IMS Open Corpus Workbench – еще одна платформа для анализа корпусов [4]. Поддерживает поиск по леммам, частям речи, семантическим меткам и границам фрагментов, позволяет строить частотные словари, анализировать конкордансы и совместную встречаемость словоформ и многое другое. Платформа предназначена для продвинутых экспертов в анализе корпусов, поскольку обладает сложным интерфейсом и требует понимания моделей данных и языка запросов. Некоторые компоненты IMS Open Corpus Workbench используются и в других платформах, например, TXM.

Рассмотренные инструменты незаменимы в таких областях, как лексикография, анализ дискурса, преподавание перевода, а также для разных языковых исследований. Однако многие из них имеют не очень удобный интерфейс и сложный язык запросов, что не позволяет быстро начать работу с ними. Кроме того, эти ин-

струменты предоставляют ограниченные возможности по работе с древовидными представлениями текстов, отображающими синтаксическую или семантическую информацию, и слабо ориентированы на работу с русским языком.

1. Устройство машины PCA

Машина PCA представляет собой инструмент лингво-статистических и корпусных исследований, ориентированный, прежде всего, на работу с русским языком. Этот инструмент реализует поисковые и аналитические функции для корпусов текстов или одного текста, при этом текст представляется в виде сетевой структуры произвольной сложности, что позволяет моделировать все уровни языка и любые виды разметки. Сетевое представление текста позволяет искать в нем конструкции сложной структуры, что отличает машину PCA от многих аналогичных инструментов, в которых структурные отношения между элементами языка не моделируются. В нашем исследовании представление текста основано на реляционно-ситуационной модели текста, поэтому в основном работа ведется с семантико-синтаксическими конструкциями.

1.1. Функции машины

К основным функциям машины PCA относятся:

1. поиск для заданного пользователем слова или списка слов из словаря всех предикатных слов (глаголов и девербативов), для которых слово является аргументом и заполняет ту или иную семантическую роль в тексте;
2. поиск всех семантических ролей для заданного слова;
3. поиск враждебной лексики, деловой/управленческой лексики, маркеров модальности и любых других слов из заданных словарей;
4. поиск и отображение в тексте предикатных слов с заданным значением (движения, разрушения, мыслительной активности...) и их аргументов;
5. отображение в тексте всех предикатов со всеми аргументами, в том числе теми, для которых роли не установлены;
6. поиск для заданного слова всех однокоренных слов разных частей речи;

7. поиск для заданного слова синонимов, антонимов, гипонимов, конверсивов и т.д. для заданного слова.

Результаты поиска представляются, в том числе, в виде конкордансов с указанием для искомой конструкции контекстов заданной длины и ссылок на тексты, в которых содержится конструкция. От конкорданса можно перейти к просмотру вхождения конструкции непосредственно в тексте.

Кроме поисковых функций, машина РСА реализует функции вычисления частотных и статистических характеристик для результатов поиска, корпусов или отдельных текстов. Например, при поиске по заданному слову (функция 1) выводится упорядоченный по частотности список предикатных слов, для которых заданное слово является аргументом, и упорядоченный по частотности список семантических ролей, заполняемых заданным словом при указанных предикатных словах. Так же как и при работе с реляционными базами данных, в машине РСА возможна группировка результатов поиска по различным полям и вычисление количества найденных результатов, среднего и т.п. с упорядочиванием по убыванию или возрастанию. С помощью таких функций можно вычислять, например, следующие показатели по отдельным текстам или коллекциям:

- количество слов и предложений;
- частотность слов, лемм, частей речи и любых других конструкций;
- частотность отдельных синтаксисов;
- количество незаполненных аргументов при предикатах в отдельно взятом тексте;
- доля враждебной лексики, деловой/управленческой лексики, маркеров модальности от всей лексики;
- частотность совместной встречаемости заданных слов (на каком расстоянии друг от друга, в какой форме).

Отдельное место в машине РСА занимают функции вычисления психолингвистических характеристик текстов и корпусов. К таким характеристикам относятся, например:

- отношение количества глаголов к количеству прилагательных;
- частотность местоимений первого лица множественного числа;
- частотность глаголов прошедшего времени первого лица единственного числа;

- отношение числа инфинитивов к общему числу глаголов.

В дальнейшем в машине РСА предполагается реализовать широкий набор функций для управления корпусами, создания иерархии или сети корпусов, атрибуции текстов, редактирования словарей и тезаурусов, а также служебные функции для авторизации пользователей, настройки прав доступа к корпусам и текстам и т.п.

Одной из главных функций машины РСА является возможность сравнения текстов или коллекций по их частотным и статистическим характеристикам с определением достоверности различий и выделением наиболее значимых разделяющих характеристик. Предполагаются и другие виды количественного и качественного анализа, включая факторный и кластерный анализы, вычисление специфичности и т.п. с отображением результатов в виде различных графиков и изображений.

1.2. Программные компоненты машины

Машина РСА состоит из следующих программных компонентов.

–Лингвистический процессор - выполняет морфологический, синтаксический и семантический (реляционно-ситуационный) анализ текстов из корпусов и текста поискового запроса.

–Программное ядро – реализует пользовательские и служебные функции. Состоит из следующих модулей:

- редактор словарей – в модуле реализованы функции для создания, редактирования и удаления словарей. Функции модуля доступны только через интерфейс эксперта;
- корпусный менеджер – модуль обеспечивает функции для редактирования корпусов, а именно: создание нового корпуса, добавление документов в корпус, удаление документов из корпуса, атрибуция корпусов и текстов. Функции данного модуля также доступны только через интерфейс эксперта;
- поисково-аналитический модуль – содержит реализацию поисковых и статистических функций, описанных выше. Доступен через пользовательский интерфейс.

–Модуль работы со словарями – предназначен для создания, удаления и редактирования словарей.

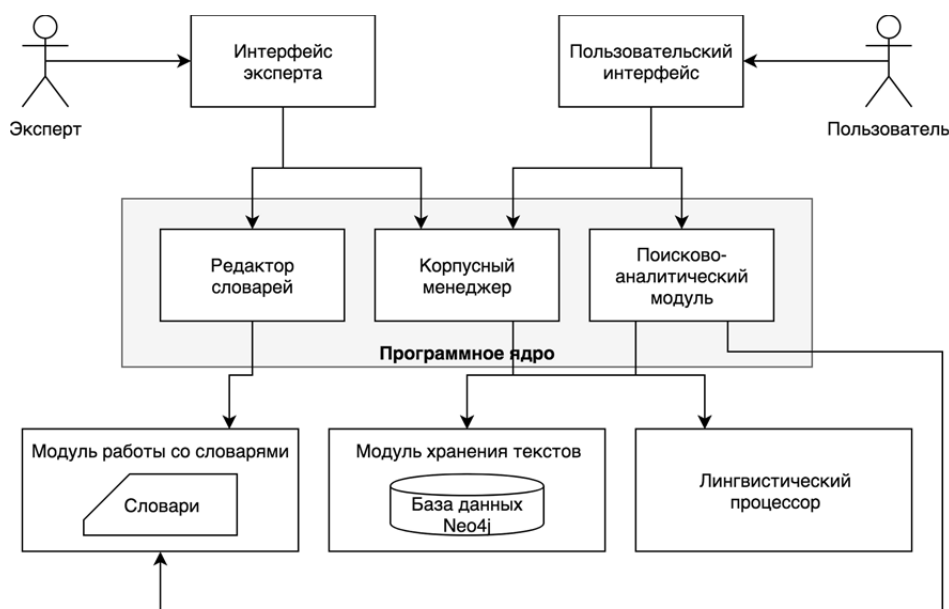


Рис. 1. Программная архитектура машины RSA

–Модуль хранения сетевой структуры текстов на основе графовой базы данных. Редактирование данных в базе происходит через редактор корпусов, а поиск – через поисково-аналитический модуль.

–Пользовательский интерфейс – клиентское приложение, позволяющее использовать основные функции машины PCA. Реализовано как веб-приложение и доступно через браузер.

–Интерфейс эксперта – веб-приложение, реализующее функции администрирования: редактирование словарей, редактирование корпусов.

На Рис. 1 показана программная архитектура машины PCA.

Результаты синтаксического и семантического анализа представляют собой деревья зависимостей. Для их эффективного хранения целесообразно использовать графовые базы данных. В машине PCA используется самая распространенная реализация графовой БД Neo4j¹. Данные в графовой БД Neo4j хранятся в атрибутах узлов и связей. Для каждого типа узлов задаются собственные атрибуты. База данных машины RSA содержит следующие типы узлов их атрибуты.

- Корпус: название корпуса.
- Документ: порядковый номер документа, название документа.
- Текст: текст документа без форматирования.

- Предложение: порядковый номер предложения.

- Клауза: порядковый номер клаузы.

- Слово:

- смещения (позиции первого и последнего символа слова в тексте);
- порядковый номер слова;
- изначальная форма слова;
- нормальная форма слова;
- морфологические признаки.

- Роль: база данных содержит следующие типы связей:

- содержит (contains); эта связь устанавливается между следующими типами узлов: корпус и документ, документ и текст, документ и предложение, предложение и клауза, клауза и слово;
- синтаксическая связь (SLINK); эта связь устанавливается между узлами типа слово;
- аргумент (argument); эта связь устанавливается между узлами типов слово и роль; содержит атрибут с названием роли, к которой идет связь;
- предикат (predicate); эта связь устанавливается между узлами типов слово и роль;
- следующее слово (next); эта связь устанавливается между узлами типа слово.

На Рис. 2 показана иерархия типов узлов, сформированная типом связи 'содержит'. На

¹ <https://neo4j.com/>



Рис. 2. Иерархия типов узлов

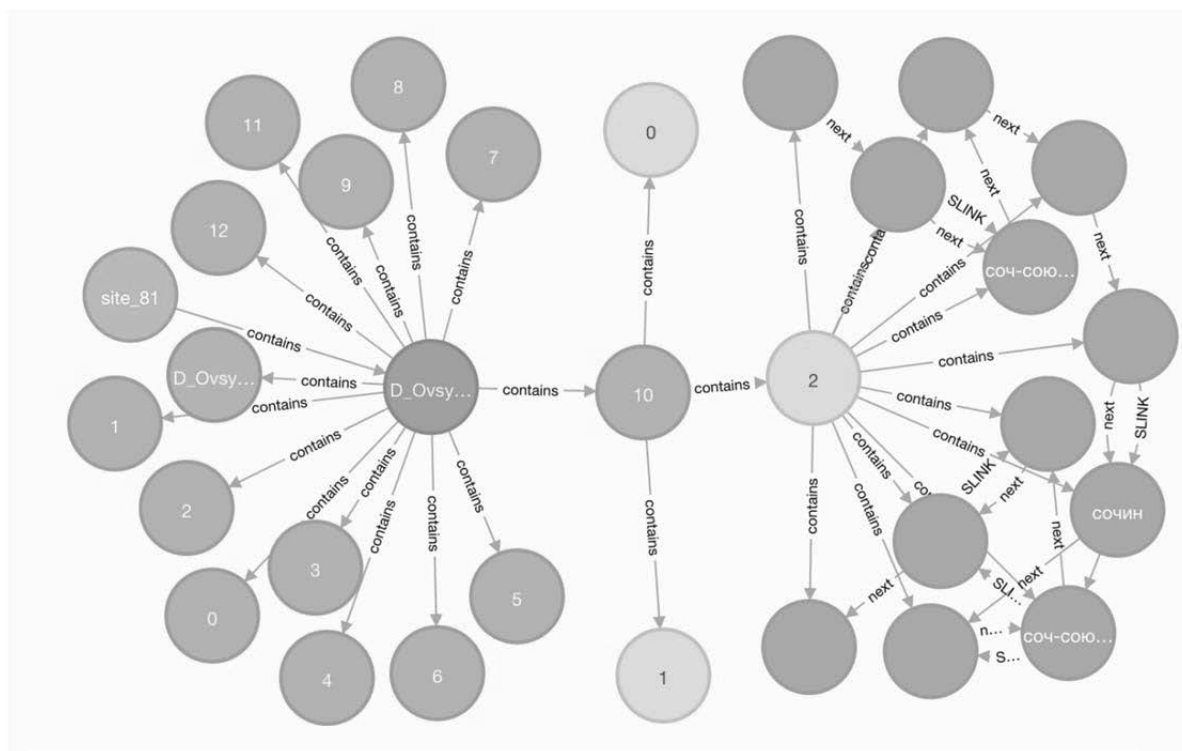


Рис. 3. Графовая структуры текста в интерфейсе Neo4j

Рис. 3 приведен пример графовой структуры текста с раскрытием структуры одного из предложений и структуры одной из его клауз.

В Neo4j реализован собственный декларативный язык запросов Cypher, который позволяет легко писать запросы-шаблоны для поиска подграфов в графах. Синтаксис языка Cypher похож на SQL из которого были заимствованы многие ключевые слова (WHERE, ORDER BY и т.д.). Например, запрос для функции «Поиск всех семантических ролей для заданного слова» выглядит следующим образом:

```

MATCH (d:document)-[:contains]->(s:sentence)-
  [:contains*]->(w:word)-[:a:argument]-(:role)
WHERE w.sDictForm = 'слово'
RETURN id(d), s.idx, w.idx, a.role_id
  
```

Ключевое слово 'MATCH' задает область поиска. В скобках указывается тип искомого объекта. Запись (d:document) означает, что найденный объект типа 'document' будет инициализирован в

переменную 'd'. В квадратных скобках указывается тип связи. Запись '(d:document)-[:contains]->(s:sentence)' указывает, что необходимо найти документы, содержащие предложения. Запись '(s:sentence)-[:contains*]->(w:word)' запрашивает предложения, у которых слова могут быть дочерним объектом более чем через 1 элемент. Предложения содержат клаузы, а клаузы – слова. Поэтому в данном примере, вместо '(s:sentence)-[:contains]->(c:clause)-[:contains]->(w:word)' указано '(s:sentence)-[:contains*]->(w:word)'. Запись '(w:word)-[:a:argument]-(:role)' означает, что необходимо вывести аргументы, имеющие роль.

Компоненты машины PCA разработаны как самостоятельные программные средства, поэтому могут быть заменены или использованы в других системах. Например, замена лингвистического процессора не является трудозатратной задачей. Необходимо лишь привести результат анализа к рабочему формату.

В машине PCA реализован программный интерфейс REST API, который позволяет создавать на основе инструмента различные пользовательские интерфейсы и использовать инструмент в веб-сервисах.

Машина PCA ориентирована на работу с большими корпусами текстов (несколько десятков миллионов словоупотреблений), поэтому основные вычисления выполняются на стороне сервера, а используемые вычислительные ресурсы масштабируются по мере необходимости. Такой подход обеспечивает высокую производительность и возможность работы с большими корпусами текстов.

Программный код, реализующий основные функции машины PCA, предполагается выложить в открытый доступ.

2. Пилотажное исследование текстов с использованием машины PCA

Машина PCA в режиме пилотажного исследования была использована для анализа текстов эссе, написанных на заданную тему: «Я. Другие. Мир». Тексты были написаны в электронном виде испытуемыми, входящими в одну из групп: 1) пациенты психиатрической клиники с диагностированными психическими заболеваниями (депрессивные расстройства, шизофрения), ср. возраст – 25 л., 18 чел.; 2) студенты-москвичи технических специальностей, ср. возраст – 23,6 л., 12 чел.; 3) студенты-москвичи гуманитарных специальностей, ср. возраст – 18,5 л., 71 чел.; 4) студенты гуманитарных специальностей из г. Кургана, ср. возраст – 22 г., 22 чел.; 5) взрослые жители Москвы, ср. возраст – 37 л., 21 чел. Всего в настоящем пилотажном исследовании приняли участие 142 человека.

По итогам работы машины PCA в тексте каждого испытуемого были выделены предикаты и для каждого из них (глагола в различных формах) определены связанные с ним синтаксемы, или роли аргументов (соответствующие формы существительных и местоимений). Фрагмент таблицы с полученными данными приведен в Табл. 1.

2.1. Связи данных машины PCA и личностных особенностей

На первом шаге было проведено корреляционное исследование связи психологических

Табл. 1. Выявленные для испытуемых № 1 (гр. 1) и № 24 (гр. 2) предикаты и роли для объекта «Я»

Испытуемый	Роль аргумента	Аргумент	Предикат
1	Адресат	Мне	Дает
	Объект	Меня	Радует
	Субъект	Я	Люблю
	Субъект	Я	Люблю
	Субъект	Я	Люблю
24	Авторизатор	Мне	Кажется
	Объект	Меня	Касается
	Субъект	Я	Хотел
	Субъект	Я	Привык

особенностей и частоты использования в эссе предикатов и семантических ролей, выявленных с помощью метода PCA. Авторам эссе было предложено заполнить ряд опросников: опросник агрессивности Басса-Перри (BPAQ) в адаптации С.Н. Еникколопова и Н.П. Цибульского; опросник конструктивного мышления С. Эпштейна (ОКМ) в адаптации С.Н. Еникколопова и С.В. Лебедева; опросник нарциссических черт личности (НЧЛ) Н. М. Клепиковой, О. А. Шамшиковой; опросник «Способы совладающего поведения» (ССП) Р. Лазаруса в адаптации Т.Л. Крюковой и Е.В. Куфтык; опросник «Стиль саморегуляции поведения» (ССПМ) В.И. Моросановой; тест жизнестойкости (ТЖ) С. Мадди в адаптации Д.А. Леонтьева и Е.И. Рассказовой; опросник черт характера (ОЧХ) В.М. Русалова и О.Н. Маноловой; пятифакторный личностный опросник (5PFQ) Р. МакКрае и П. Коста в адаптации А.Б. Хромова, а также русскоязычные версии методик: Personal Need for Structure Thompson, Naccarato, Parker, & Moskowitz (шкала Потребности в структуре, ШПС), Multidimensional scale of anomie Heydari, Davoudi, & Teymoore (Шкала аномии, ША), New Personal Fable Scale Lapsley, Fitzgerald, Rice, & Jackson (опросник Личный миф, ЛМ) в адаптации Ю.М. Кузнецовой. Для групп больных и здоровых вычисление корреляций проводилось отдельно, использовался критерий ранговой корреляции Спирмена.

В условиях психологического тестирования (так называемая ситуация экспертизы) задача, решаемая нашими испытуемыми, являлась по сути задачей на самопрезентацию, включающей презентацию своего отношения к другим людям и системы взглядов на мир в целом. Соответственно, интерпретировать полученные

зависимости необходимо с точки зрения того, как различно протекает процесс самопрезентации в тексте в условиях наличия/отсутствия психиатрических проблем и наличия определенного уровня измеренных в обследовании психологических особенностей.

Продемонстрируем возможности предлагаемого подхода на материале такой психологической черты как физическая агрессивность (измерялась с помощью опросника Басса-Перри). Обнаружено, что в группе больных влияние склонности к физической агрессии сказывается на употреблении семантических ролей разного типа более сильно, чем в группе здоровых – в первой группе получено 6 значимых корреляций, в группе здоровых – одна. Можно констатировать, что группа психически больных более однородна в отношении влияния агрессивности на порождаемые человеком речевые конструкции, чем группа здоровых. Интерпретация этого факта требует дальнейших исследований, поскольку обнаруженное отличие может быть обусловлено:

- более высокой значимостью ситуации экспертизы для больных, заставляющей тех из них, кто чувствует в себе большую склонность к физической агрессии действовать по некоему негласному правилу – писать об одном и не писать о другом. Возможно, при дальнейшем делении уже внутри группы здоровых будет обнаружена подгруппа, так же обладающая каким-либо «правилом», согласно которому высокоагрессивные будут акцентировать внимание на определённых смысловых сторонах своих представлений о мире (возможно даже – на тех же, что и высокоагрессивные больные);

- наличием у больных более выраженного воздействия склонности к физической агрессии на речевую деятельность в целом, что и нашло проявление в росте приверженности к определенным семантико-синтаксическим конструкциям при росте склонности к физической агрессии.

Тем не менее, один факт мы можем констатировать с уверенностью – в условиях создания самопрезентационного текста склонность к физической агрессии и у здоровых и у больных связана с акцентированием внимания на причинах. В обеих группах получена значимая корреляция этого аспекта агрессивности с частотой использования каузативных конструкций, правда, для здоровых значимо указание на

объект каузирующего воздействия (семантическая роль «каузат», например, «способствовать вступлению России в ВТО»), а для больных – указание на саму причину (семантическая роль «каузатив». Например, «Я не из гордости – из горести так прямо голову держу». (Ахмадулина), «И чего вы так беспокоитесь? Неужто из самолюбия, что вас женщина первая бросила, а не вы ее?» (Достоевский). Можно предположить, что для человека, осознающего свою склонность к физической агрессии, эта значимость фактора причины в представлениях о мире связана со спецификой физической агрессии как поведенческой реакции на стимул: каузация, как логическое ядро стимульно-реактивного механизма, оказывается включена в картину мира тех, кто привык отвечать физическим действием на возникающие препятствия. Отметим, что речь не идет о сознательном выборе при применении определенных речевых конструкций, поскольку использование тех или иных семантических ролей и синтаксисом реализуется на неосознаваемом уровне психолингвистических механизмов разворачивания внутренней речи во внешнюю, т.е. на уровне, аналогичном уровню психофизиологических механизмов.

2.2. Анализ представленности в текстах предикатов и синтаксисом определенного типа

Материал, полученный в результате работы машины PCA, может быть интерпретирован не только путем сопоставления с другими характеристиками, но и сам по себе. В зависимости от задач исследования могут использоваться различные схемы как организации сбора данных, так и их обработки. На полученном нами материале кратко обрисовем некоторые возможности нового метода анализа текстов.

Из всех аргументов, выявленных в 142 эссе наших испытуемых, с целью упрощения обработки результатов пилотажа, были выделены два, соответствующих теме: «Я» и «Другие» (Люди / Человек), которые, таким образом, выступили в данном случае в качестве исследуемых объектов. Объект «Я» был зафиксирован в разных ролях во всех эссе всего в 1097 случаях, объект «Люди / Человек» – в 530 случаях. В Табл. 2 приведены данные о распределении ролей для двух выделенных объектов. Они

Табл. 2. Количество случаев употребления аргументов «Я» и «Люди/Человек» в различных семантических ролях на 142 эссе

Роль аргумента	Количество случаев для объекта «Я»	Количество случаев для объекта «Люди/Человек»
Абстинатив		1
Авторизатор	103	8
Агенс	1	13
Адресат	40	18
Владелец информации	5	
Высший предмет иерархии	1	1
Генератив	2	1
Делибератив	2	13
Дестинатив	10	3
Дименсив		1
Директив	3	5
Донор	7	3
Инструментатив	1	
Каузатив	2	2
Комитатив		3
Ликвидатив		1
Локатив	31	15
Медиатив		1
Объект	111	76
Объект-орудие	2	3
Партитив	1	1
Пациенс	3	
Поссесив	3	1
Посессив-субъект-отправитель	2	
Предикатив	5	14
Предмет сравнения	1	3
Социальная категория	1	
Субъект	759	309
Субъект неактивных действий		1
Субъект-соучастник		28
Темпоратив	1	
Транзитив		1
Эстиматив		4

свидетельствуют о существовании определенных тенденций различий в употреблении объектов «Я» и «Люди /Человек» в разных семантических ролях. Очевидными такие различия являются, например, для ролей «агенс» (преимущественно другие люди), «авторизатор», «адресат» и «субъект» (преимущественно «Я») или «субъект-соучастник» (которым может быть только некто другой, но не «Я»). Можно видеть, что данные различия отражают, прежде всего, степень активной субъектности, которая

вполне закономерно атрибутируется в основном объекту «Я».

Следующий шаг обработки полученных данных может быть связан с выявлением групповой специфики ролевых позиций исследуемых объектов. Сравним, например, частоту встречаемости объекта Я в той или иной семантической роли для двух групп – больных (группа 1) и здоровых испытуемых (объединенные группы 2-5). Частота ролей «субъект» (5,17 на одно эссе у больных и 5,37 – у здоровых) и

«объект» (0,78 в обеих группах) не зависит от наличия психического заболевания. В то же время «Я» в синтаксической роли «авторизатор» психически больными использовалось в два раза реже, чем здоровыми (0,33 на одно эссе против 0,78), а в роли «адресат» - в два раза чаще (соответственно, 0,56 и 0,24 на одно эссе). Данные различия могут быть интерпретированы с точки зрения того когнитивного содержания, которое стоит за синтаксическими ролями. Вероятно, выраженное преимущество субъектности над объектностью в случае, когда речь идет о самом себе, является соотношением настолько базовым, что на него не может оказать влияния наличие психического заболевания (что делает диагностически ценными варианты индивидуального отклонения от такого соотношения). В то же время роль «адресат» (согласно определению Г.А. Золотовой, «компонент, обозначающий лицо..., к которому обращено информативное, донативное или эмоциональное действие» [2, с. 430]) отражает ту пассивно-принимательную, зависимую позицию, которая, по-видимому, более органична для самоопределения в социальном взаимодействии для людей с психическими отклонениями, чем для здоровых. Здоровым же, более чем больным, свойственно переживать себя в позиции ««автора» оценки, восприятия, рече-мысли» (там же), то есть, несколько упрощая, субъектом самостоятельной ментальной деятельности. Увеличение количества наблюдений (прежде всего, за счет привлечения новых испытуемых с психиатрическим статусом) позволит применить статистические методы для оценки опи-

санных тенденций, однако и на настоящем этапе проявляется содержательная наполненность и принципиальная интерпретируемость полученных с помощью машины РСА данных.

Другим направлением анализа может быть выявление закономерностей, связанных с типами использованных в текстах предикатов, при которых находятся исследуемые объекты. С учетом семантики глаголов и глагольных форм, определенных машиной РСА как предикаты, относящиеся к объектам «Я» и «Люди/Человек», были выделены следующие группы: 1) предикаты состояния: *становиться, являться, жить* и т.п., 2) ментальной деятельности: *представлять, видеть, знать* и т.п., 3) коммуникации и социального взаимодействия: *написать, говорить, поделиться* и т.п., 4) акциональные: *выйти, учиться, работать* и т.п., 5) аффективные: *любить, бояться, нравиться* и т.п. Эти группы только частично отражают принятую в лингвистике классификацию предикатов, выделены окказионально в связи с конкретными исследовательскими интересами, и могут быть изменены как в сторону большей дробности (например, группу аффективных предикатов можно разбить на предикаты положительных и отрицательных эмоций), так и слиты (например, могут быть объединены группы предикатов ментальной деятельности и аффективных, акциональных и предикатов социального взаимодействия) или вообще созданы на других основаниях в соответствии с задачами исследования. На Рис 4 представлены сравнительные данные о распределении предикатов, связанных с объектами «Я» и «Люди/Человек»,



Рис. 4. Сравнительные данные о распределении предикатов

по пяти сформированным нами группам. Они подтверждают описанную выше общую тенденцию к преимущественному приписыванию активности себе, но не другим людям.

Далее можно (как это было сделано для синтаксических ролей) сравнивать частоты предикатов из разных групп, встречающихся в эссе психически больных и здоровых испытуемых, что дает возможность выявлять особенности предикации, связанные с наличием болезни. Возможно также сопоставление двух категорий анализа – типов предикатов и синтаксических ролей. В этом случае будет получена информация о том, при каких именно предикатах объекты «Я» и «Люди/Другие» выполняют определенные роли. Оставив из всех выявленных только наиболее часто встречающиеся ролевые позиции «субъект» и «объект», получим для каждой группы наших испытуемых четыре варианта: Я-субъект, Я-объект, Другие-субъект, Другие-объект.

Сопоставление с типом (то есть, выделенной группой) предикатов позволяет определить, субъектом каких именно действий (ментальных, акциональных, аффективных, коммуникативных) или состояний преимущественно рассматривают авторы текстов себя и других людей, по отношению к каким действиям они склонны ощущать себя в пассивной позиции объекта и в каких случаях приписывают такую позицию другим людям.

Среди интересных фактов, обнаруженных при реализации данной схемы обработки данных, можно назвать характерную только для испытуемых из группы студентов технических специальностей склонность редуцировать образ «Люди/Человек» как объекта всего лишь двух видов деятельности: прежде всего, «Другие» испытывают *состояния*, которым атрибутируется качество активности по отношению к испытывающим их людям; кроме того, «Другие» могут быть объектами ментального воздействия со стороны окружающих. По-видимому, в таком образе «Других» находит свое отражение известная специфика коммуникативной компетентности представителей технических профессий, которая связана с недостаточно высоким уровнем эмпатии и в целом интереса к внутренней жизни человека. Такая интерпретация подтверждается тем, что в данной группе также отсутствует коммуникативная составляющая у образа Я-объект. Иначе говоря, себя

студенты технических специальностей ощущают как объект, прежде всего, эмоционального влияния со стороны окружающих, а затем – акциональных и ментальных воздействий, но не выделяют в качестве предмета рефлексии саму ситуацию общения, хотя и фиксируют возникающие в ходе общения ощущения подверженности постороннему влиянию. Можно отметить также специфическое снижение значимости собственной коммуникативной активности (позиция Я-субъект) у психически больных испытуемых по сравнению с переживанием себя в качестве объекта коммуникативной деятельности окружающих (Я-объект), что контрастирует с обратным соотношением для всех групп здоровых испытуемых (они ощущают себя в большей степени субъектом общения, нежели объектом чужой коммуникации).

Не останавливаясь на других результатах, подчеркнем следующий момент: благодаря применению машины РСА выявление описанных тенденций, находящих взаимное подтверждение или обоснование в психодиагностических данных, и касающихся глубинного устройства картины мира человека, возможно без обращения к рефлексии испытуемого. С помощью нового метода диагностика особенностей картины мира возможна на основе анализа текстов, порождаемых не только специально в ситуации исследования, но и спонтанно, в частности, в ходе сетевой коммуникации.

Описанные здесь варианты обработки и интерпретации данных, получаемых с помощью машины РСА, не исчерпывают возможности нового метода и приводятся лишь в качестве примера ее использования при решении разнообразных задач социо-гуманитарных исследований, ориентированных на выявление индивидуальных, групповых и популяционных особенностей текста как продукта речемыслительной деятельности и стоящей за ним картины мира автора.

Заключение

В отличие от большинства предлагаемых сейчас гуманитариям моделей анализа текста и средств извлечения информации о состоянии его автора, методы искусственного интеллекта, в первую очередь метод РСА, позволяют выйти за рамки относительно легко формализуемых компонентов текста (преимущественно лекси-

ческих, грамматических и морфологических) и частотного анализа. Разрабатываемый в ИСА ФИЦ ИУ РАН инструмент лингво-статистических корпусных исследований отличается от аналогов возможностью работы с лингвистическими конструкциями произвольной структуры и сложности, моделирующими все уровни языка (включая синтаксис, семантику, дискурс). Хранение сетевого представления текста с помощью графовой базы данных делает процесс поиска и анализа таких сложных структур быстрым и эффективным, а поддерживаемый язык запросов позволяет быстро создавать поисковые шаблоны для любых лингвистических конструкций или их сочетаний и получать различные статистики для результатов поиска.

С точки зрения своего психологического содержания данные, получаемые с помощью машины RSA, соотносятся с компонентами речемыслительной деятельности, которые отражают глубинные структуры ментальной организации личности – субъективную картину мира. Картина мира задает те правила, по которым в воспринимаемой реальности вычленяется система взаимодействующих определенным образом объектов, что на вербальном уровне определяет организацию речевых единиц и их комплексов. Описание речевой организации в терминах предикатно-синтаксического анализа является, таким образом, способом объективации

картины мира субъекта. Предлагаемый метод анализа текстов позволяет оперировать такими показателями (исследуемые объекты, семантические роли, типы предикатов) и в таких их выражениях (популяционные, групповые, индивидуальные) и сочетаниях, которые делают возможным проведение широкого спектра исследований в интересах множества социогуманитарных дисциплин.

Литература

1. Ениколопов С.Н., Кузнецова Ю.М., Смирнов И.В., Станкевич М.А., Чудова Н.В. Создание инструмента автоматического анализа текста в интересах социогуманитарных исследований. Ч.1 //Искусственный интеллект и принятие решений. 2019. №2.
2. 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. – Institute for Digital Enhancement of Cognitive Development, Waseda University. 2010. С. 389-398.
3. Kilgarriff A. et al. The Sketch Engine: ten years on //Lexicography. 2014. Т. 1. №. 1. С. 7-36.
4. Evert S., Hardie A. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. 2011.
5. Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. Изд. 3. М.: Едиториал УРСС, 2006. С. 430.

Creating a Text Analysis Tool for Socio-Humanitarian Research. Part 2. RSA Machine and the Experience of Using It

Y. M. Kuznetsova, I. V. Smirnov, M. A. Stankevich, N. V. Chudova

Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. The second part of the work describes the most known tools for linguistic-statistical analysis of text corpora and introduces RSA machine - a novel text analysis tool for socio-humanitarian research. This tool works with network representation of text and allows finding the constructions with complex graph structure in texts. RSA machine implements following features: search of constructions by query, computation of frequencies and statistical characteristics for search results, corpora or individual texts, comparing texts using statistical and frequency features. This paper describes the RSA machine architecture and developing tools. We present the results of pilot research of RSA machine using 142 texts examples written by people with different psychology and demographic characteristics. Some of them (18) were diagnosed with mental disorder. The performed correlation analysis revealed some relations between extracted texts attributes (e.g. frequency of predicate types) and results of psychological analysis performed by experts.

Keywords: text corpora analysis, software architecture, graph database, semantic-syntactic constructions, socio-humanitarian research, worldview.

DOI 10.14357/20718594190305

References

1. Kuznetsova Y.M., Smirnov I.V., Isakov V.A., Stankevich M.A., Chudova N.V. Sozdanie instrumenta avtomaticheskogo analiza teksta v interesah socio-gumanitarnyh issledovanij Ch. 1. [Creating a tool for automatic text analysis in the interest of socio-humanitarian research. P.1.] (In print).
2. 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. – Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010. – P. 389-398.
3. Kilgarriff A. et al. The Sketch Engine: ten years on //Lexicography. – 2014. – V. 1. – №. 1. – P. 7-36.
4. Evert S., Hardie A. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. – 2011.
5. Zolotova G.A. Sintaksicheskij slovar': Repertuar ehlementarnyh edinic russkogo sintaksisa [Syntax dictionary: The repertoire of elementary units of Russian syntax]. M.: Editorial URSS, Ed. 3., P. 430, 2006.