ЛИНГВИСТИКА

УДК 81’33:81’23:81’27
DOI: 10.17223/19986645/64/1

К.И. Белоусов, Е.В. Ерофеева, Д.А. Баранов, Н.Л. Зелянская, С.А. Щебетенко

МНОГОПАРАМЕТРИЧЕСКИЙ АНАЛИЗ ЛИНГВИСТИЧЕСКИХ ДАННЫХ В ИНФОРМАЦИОННОЙ СИСТЕМЕ «СЕМОГРАФ» (НА ПРИМЕРЕ ИССЛЕДОВАНИЯ РЕЧЕВОГО ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ)[1]

Демонстрируются возможности информационной системы «Семограф» как инструмента анализа текстового контента при реализации сетевого подхода к организации научных исследований в лингвистике на примере многопараметрического анализа пользователей социальной сети. На основе классификации реплик и метаданных (пол и психологические характеристики) пользователей с помощью средств визуальной аналитики предложена модель взаимосвязей лингвистических параметров речи, социальных и психологических характеристик личности.

Ключевые слова: *сетевая наука, социальные интернет-сервисы, информационная система «Семограф», многопараметрический анализ, визуальная аналитика, графосемантическое моделирование.*

Введение

Объем информации, с которой работает современный ученый, требует привлечения не только информационных технологий и средств автоматизации отдельных сторон деятельности исследователя, но и формирования новых исследовательских стандартов. Эти стандарты касаются прежде всего открытости и доступности исходных данных, возможности их повторного использования (например, в метаанализе), быстроты / оперативности получения результатов и междисциплинарного характера исследований, требующих согласованной работы специалистов разных научных областей. Реализации новых стандартов научной деятельности наиболее органична сетевая форма организации исследований (сетевая наука).

Сетевая организация взаимодействия субъектов в профессиональной деятельности в целом признается наиболее эффективной [1, 2], и сетевые программные решения в настоящее время применяются во многих сферах. В частности, в образовательном сегменте (в основном в вузах) для реализации программ дистанционного обучения используются программные ре-

[1] Работа выполнена в рамках государственного задания ПГНИУ на 2020–2022 гг., номер темы FSNF-2020-0017 «Многопараметрическое моделирование процессов коммуникации пользователей социальных интернет-сервисов с использованием методов машинного обучения и визуальной аналитики».

шения Moodle, ILIAS, АСУ ВУЗ и др., а также образовательные порталы (Coursera, edX, Udacity, Stepic и др.), однако до сих пор создано не так много программных продуктов, в которых реализованы принципы сетевой организации научной деятельности. В то же время количественный анализ современных публикаций показывает, что 90% всех научных статей написаны двумя и более авторами, большинство статей принадлежат авторским коллективам от 6 до 10 человек из нескольких учреждений [3]. Это говорит о необходимости создания аналитических инструментов для работы исследовательских коллективов в рамках концепции сетевой науки.

Сетевая наука, с нашей точки зрения, должна обеспечивать следующие аспекты научной работы:

– распределенный в режиме реального времени научный процесс;

– организацию сетевого взаимодействия участников;

– системное управление работой коллектива;

– использование единых технологий обработки информации и общей базы данных;

– интеграцию результатов исследовательской работы каждого участника в единое информационное пространство.

В настоящее время в открытом доступе находится большое количество ресурсов, предоставляющих информацию об исследованиях в любых областях науки и техники, а также порталов, используемых для коммуникации между исследователями. Это веб-сайты научных журналов и ученых, базы знаний и базы данных (в том числе и результатов экспериментов, как, например, БД Reaxys), научные социальные сети и ресурсы, созданные для поддержки перспективных научных исследований, и др. Кроме того, существуют системы сетевой организации научных исследований, основывающихся на концепции Citizen science. Эти исследования проводятся группами волонтеров в сотрудничестве или под руководством ученых и / или научных организаций (список активных и завершенных проектов см. [4]). Однако для полного воплощения идеи сетевой науки помимо обилия ресурсов, почти неисчерпаемого объема научной информации и спектра форматов ее представления, внушительного числа участников научного пространства требуется распределенная аналитическая сетевая среда, в которой осуществляется онлайн-взаимодействие субъектов исследовательского процесса и его коррекция. При этом программные продукты, обеспечивающие такую среду, очевидно, должны быть направлены на конкретные научные области и задачи (хотя, возможно, и весьма широкие).

В лингвистике программные решения предназначены в первую очередь для обработки и анализа текстовых массивов (см., например, каталог лингвистических ресурсов NLPub (http://nlpub.ru)). Программные продукты, созданные для обработки и анализа текста, предлагают вполне достаточный инструментарий для социолога, маркетолога, контент-менеджера или специалиста в области машинного обучения, однако большей частью неприменимы для решения лингвистических задач. Кроме того, существую-

щие программные решения не предполагают командной работы в процессе анализа языкового / текстового материала.

В данной статье демонстрируются возможности информационной системы (ИС) «Семограф» (http://semograph.org) как инструмента анализа текстового контента при реализации сетевого подхода к организации научных исследований в лингвистике.

Информационная система «Семограф»

Цели информационной системы «Семограф»

Основная задача ИС «Семограф» – создание доступных и понятных широкому кругу лингвистов технологий, помогающих решать собственно научные задачи, поставленные в отдельном исследовании. «Семограф» может использоваться для анализа текстовых данных; создания и / или разметки языковых / текстовых корпусов; проведения, обработки и анализа данных психолингвистических, социолингвистических экспериментов; разработки классификаторов и тезаурусов, а также для решения других задач, возникающих при анализе языкового материала. Результаты анализа в ИС «Семограф» служат основой для построения лингвистических моделей.

ИС «Семограф» является открытой платформой, для работы требуется только выход в Интернет и современный браузер.

В «Семографе» реализованы следующие принципы:

– проведение полного цикла исследования, включая сбор материала, обработку и экспертный анализ данных, статистический анализ, построение моделей, основанных на принципах редактируемой визуализации;

– сетевая распределенность участников научного процесса, предполагающая возможность работы с разных машин над одним научным проектом группы исследователей, в том числе географически отдаленных от основного коллектива;

– многопользовательский режим работы в информационной системе, обеспечивающий в том числе и онлайн-взаимодействие участников научного проекта;

– методологический плюрализм, позволяющий исследователям, придерживающимся самых разных теоретико-методологических взглядов, использовать данный программный продукт.

Возможности информационной системы «Семограф»

Возможности ИС «Семограф» обеспечиваются архитектурой системы, в которой используются объекты (рис. 1) и модули (рис. 2).

Опишем объекты ИС «Семограф».

– **Проект** – рабочее пространство, в котором осуществляется реализация полного исследовательского цикла. Типовой проект включает массив контекстов, множество компонентов, систему полей и набор метаполей.

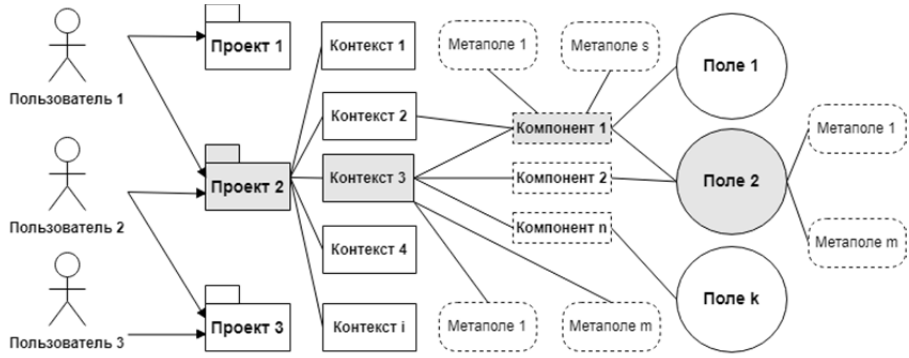


Рис. 1. Объекты информационной системы «Семограф»[1]

– **Пользователь** – исследователь или информант, участвующий в работе над проектом (или несколькими проектами). Пользователь может создавать проект и получать доступ к проекту в роли наблюдателя или редактора. Роль наблюдателя предполагает возможность просматривать материалы проекта. Роль редактора дает право пользователю выполнять определенные действия в проекте; результаты работы любого редактора сохраняются в базе данных проекта на сервере.

– **Контекст** – языковая / речевая единица или набор единиц, выбранные исследователем при представлении и анализе данных. В качестве контекста могут выступать текст, текстовый фрагмент, наборы слов и / или словосочетаний (например, совокупность экспериментальных реакций, полученных от одного информанта, наборы ключевых слов и т.п.). Каждый контекст описывается определенным набором метаполей, актуальных для исследования.

– **Компоненты** – выбранные исследователем единицы, выделенные из анализируемого контекста. В качестве компонентов могут избираться любые входящие в контекст лингвистические единицы, от самых мелких (например, букв или слогов) до таких крупных единиц, как предложение или часть текста. Обычно контекст включает несколько компонентов, однако в предельном случае компонент может совпадать с целым контекстом. Каждый компонент, в свою очередь, может быть описан набором метаполей (как лингвистических, так и экстралингвистических). Компоненты, входящие в один контекст, автоматически считаются системой связанными.

– **Поле** – множество компонентов, объединенных каким-либо общим признаком (поля могут создаваться на основе любого заданного исследователем признака, как лингвистического – семантического, грамматического и т.п., так и экстралингвистического характера).

[1] На рисунке выделены те экземпляры классов объектов, на примере которых показаны структурные связи между объектами системы.

– **Метаполе** – структурированная единица системы, позволяющая ввести дополнительную информацию о контексте, компоненте или поле (такой информацией могут быть дата, имя автора, пол, возраст, образование и другие социальные характеристики информанта, адрес интернет-ресурса и мн. др., а также лингвистические характеристики компонентов, контекстов или полей). Каждое метаполе имеет набор возможных значений в одном из таких форматов, как строка, целое, дробное, дата, файл, ссылка. Метаполя используются в системе для создания выборок контекстов, компонентов или полей.

Информационная система состоит из нескольких модулей (рис. 2):

– модуля импорта данных (поисковый робот, парсер, созданный на основе Python-фреймворк Scrapy, поисковый сервер на основе Apache Solr, а также инструменты для импорта файлов с табличным типом организации данных);

– модуля системы управления проектами (добавление приглашенных пользователей системы к проектам; детализированная система прав доступа, создание открытых проектов, коммуникационная система, создание билетов с задачами, назначение исполнителей, ведение статистики по каждому участнику проекта с графическим представлением данных, тайм-менеджмент работы);

– исследовательского модуля: 1) широкий набор инструментов анализа языкового контента; 2) результаты анализа, представленные в виде семантических карт, таблиц, частотных распределений; 3) визуализация результатов посредством интегрированной в Семограф адаптивной мультиплатформенной системы научной визуализации SciVi [5], используемой в качестве основного инструмента визуальной аналитики;

– модуля экспорта результатов во внешние приложения, в частности R (статистическую среду анализа данных), Gephi (средство построения и анализа графов), а также в табличные форматы.

Этапы работы в ИС «Семограф»

Объекты и модули информационной системы в совокупности позволяют реализовать полный исследовательский цикл, включающий следующие этапы.

1. При работе в ИС «Семограф» руководитель исследовательской группы сначала создает проект и приглашает в него участников, наделяя их определенными правами. Таким образом формируется исследовательский коллектив.

2. Формирование массива данных. В качестве материала исследования могут выступать языковые данные разных типов: тексты или текстовые массивы, слова или группы слов, фразы или предложения, ответы информантов, полученные в лингвистических экспериментах, и т.д.

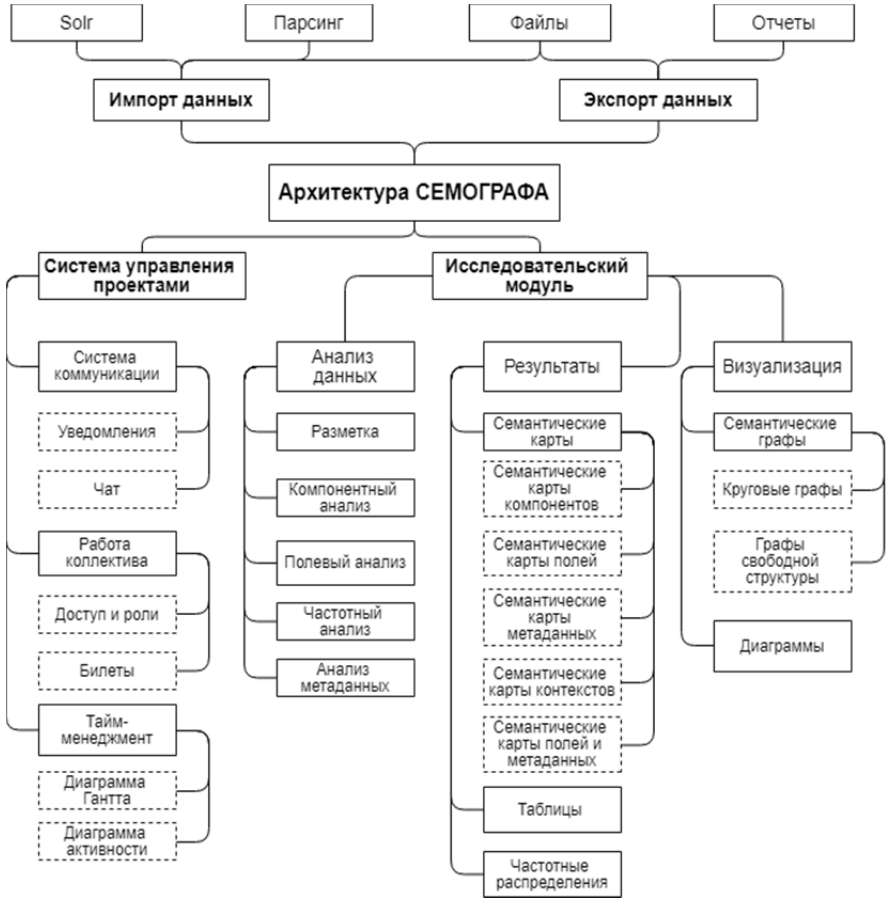


Рис. 2. Модули информационной системы «Семограф»

Данные могут вноситься в ИС «Семограф» несколькими способами:

– внесение данных вручную (данный способ может применяться, например, при онлайн-экспериментах, когда информанты вносят данные в процессе выполнения экспериментальных заданий);

– импорт с помощью загрузчика файлов (способ эффективен в тех случаях, когда анализируемый контент еще до работы в информационной системе был представлен в каком-либо офисном формате);

– импорт с помощью платформы полнотекстового поиска Solr (для импорта используются xml-файлы, хранящиеся в Solr).

На данном этапе формируются контексты проекта и им приписываются метаполя.

3. Выделение компонентов из контекстов. Компоненты могут выделяться автоматизированно (например, при импорте рефератов научных статей в качестве компонентов выделяются авторские ключевые слова) или вручную пользователем (например, вычленение словосочетаний из текста).

4. Проведение полевого анализа выделенных компонентов, т.е. объединение компонентов в поля (классы). Этот этап требует привлечения экспертов и их согласованной работы. ИС «Семограф» позволяет устанавливать не только отношения «многие к одному» между компонентами и полями (при которых компонент может входить только в одно поле), но и отношения «многие ко многим» (при которых компонент может входить в несколько полей одновременно). При этом ИС «Семограф» позволяет проводить не только простую, но и иерархическую (с системой вложенных полей) классификацию компонентов. Компоненты, входящие в одно поле, автоматически считаются системой связанными.

5. На основе проведенного деления на компоненты и классификации компонентов ИС «Семограф» автоматически генерирует семантическую карту – матрицу, отражающую совместную встречаемость двух единиц (компонентов или полей) в контекстах проекта или его выборках, организованных с помощью системы метаполей. Совместная встречаемость рассматривается как связь между компонентами и / или полями.

6. На основе семантической карты может быть построен семантический граф – графическая экспликация результатов анализа связей между полями и / или компонентами в виде неориентированного графа свободной структуры или кругового графа.

7. Последний этап исследовательского цикла – интерпретация полученной модели.

Описание различных аспектов работы ИС «Семограф» можно найти в [6, 7].

Основные характеристики описания проекта

Как уже отмечалось выше, работа с лингвистическим материалом в системе «Семограф» может проводиться с использованием лингвистических методов, выбранных исследователем в соответствии с конкретной исследовательской задачей. Каждый проект может быть охарактеризован с помощью формализованной системы описания, позволяющей на первом этапе знакомства с проектом распознать в нем реализованный исследовательский фрейм. В качестве основных параметров описания проекта выступают:

1. Объекты: Пользователь, Проект, Контекст, Компонент, Слово, Лексема, Поле, Метаполе, Фрагмент, Опорное слово.

2. Операторы (действия, производимые над объектами):

2.1. Операторы машинные: Индексация, Выборка.

2.2. Операторы данных: Права доступа, Разметка, Оценка (субъективная характеристика объекта на основе использования шкал), Добавление, Принадлежность, Классификация.

2.3. Операторы результатов: Классификатор (или тезаурус), Семантическая карта, Семантический граф, Семантическое расстояние, Матрицы переходов, Таблицы частотности, Таблицы соответствий.

3. Характеристики: Значение (например, метаполе имеет значение «25 лет»), Частота, Последовательность, Локализация, Тип (текст, целое,

дробное, дата, файл), Объем (например, объем выборки), Глубина (например, глубина тезауруса = количество уровней иерархии семантических полей), Временные интервалы.

4. Режим работы: онлайн / офлайн.

5. Статус проекта: открытый / закрытый (если открытый, то указывается название проекта), количество участников и роли.

6. Дополнительная информация (дата создания проекта, автор проекта, перечень лиц, имеющих доступ к проекту, время работы их с проектом и т.д.).

Из разных комбинаций основных параметров можно создавать разнообразные схемы работы с лингвистической информацией для решения широкого круга исследовательских задач. Рассмотрим далее использование ИС «Семограф» в организации сетевого научного исследования в области лингвистики на примере проекта “Social Network Analysis”. Данный проект выбран в качестве примера, так как многогранно представляет возможности работы с информационной системой «Семограф»: в исследовании участвует значительное количество экспертов, одновременно работающих с языковым материалом; привлекается внушительный объем анализируемого материала, автоматизированно собранного от большого числа пользователей соцсети; при обработке лингвистического материала используется многоуровневая система классификации; экстралингвистическая разметка анализируемого массива осуществляется с помощью сложной системы метаописания, характеризующей социальные, поведенческие и психологические параметры пользователей соцсети, а получаемые результаты соотносят самые разные стороны социальных, поведенческих, психологических и языковых характеристик информантов.

Проект “Social Network Analysis”

Характеристики проекта “Social Network Analysis”

1. Проект “Social Network Analysis”.

1.1. Тип проекта: закрытый.

1.2. Время создания проекта: 2017 г.

1.3. Материал проекта: 340 контекстов (каждый контекст представляет собой набор реплик – постов и комментариев пользователя соцсети), 19 179 компонентов (каждый компонент – одна из реплик пользователя).

2. Пользователи: эксперты (12 чел.).

3. Метаполя.

3.1. Метаполя контекстов: пол, возраст, количество пользовательских постов, количество постов друзей, количество друзей, количество набранных лайков (медиана), значения по пяти параметрам теста BFI (bfie, bfic, bfin, bfia, bfio), показатели самооценки.

3.2. Метаполя контекстов: используются для создания выборок.

4. Операторы: Права доступа: не ограничены для экспертов.

5. Режим работы: офлайн (данные экспортировались в систему с помощью загрузчика файлов).

6. Классификация: экспертная.

6.1. Поля: 43 поля.

6.2. Глубина иерархии: 4 уровня иерархии (компоненты принадлежат полю нижнего уровня иерархии и автоматически входят в поля более высоких уровней).

7. Операторы результата (результаты, которые можно получить на данном этапе работы с этим проектом):

7.1. Анализ метаполей (пол; возраст; BFI-параметры и т.д. и их сочетания);

7.2. Семантическая карта связи полей, семантическая карта связи полей и метаполей;

7.3. Семантический граф связи полей, семантический граф связи полей и метаполей;

8. Дополнительная информация: цель проекта – выявление зависимостей между языковыми, социальными и психологическими характеристиками пользователей социальных сетей.

Цели и задачи проекта "Social Network Analysis"

В современной науке существует запрос на создание фундаментальной концепции личности, которая бы позволила описывать, объяснять и прогнозировать речевое и неречевое поведение человека и социальных групп, включая группы пользователей социальных интернет-сервисов (Social Network Services – SNS) [8–13]. Несмотря на широкий спектр задач, решаемых в области исследования SNS, в открытых источниках пока не встречаются концепции комплексного анализа типов их пользователей, взаимосвязей между ними и моделей их поведения.

В исследовании ставилась **задача** разработки социокогнитивной модели пользователя социальной сети на основе многопараметрического анализа речевого поведения, социальных параметров и психологических характеристик личности.

Комплексное описание пользователей SNS должно основываться на моделях интеграции социального, поведенческого, психологического и языкового компонентов личности. В качестве социальных параметров рассматривается информация из профиля пользователя (пол, возраст, образование, сфера интересов, социальное окружение и др.); в качестве поведенческих – предпочтения (например, отмеченные как понравившиеся публикации и другие материалы, размещаемые в Сети) и т.п. Психологические параметры выявляются в результате психологического опроса, а языковые – на основе анализа комментариев пользователей.

Проведение психологического опроса

В качестве психологического опросника использовалась русская версия «Вопросника Большой Пятерки» (BFI – Big Five Inventory) [14, 15]; автор адаптированной русскоязычной версии С.А. Щебетенко [16].

В опросе участвовали студенты одного из российских университетов. Опрос проводился в лаборатории; форма опроса – письменное анкетирование в группах от 8 до 25 человек. Участников просили указать в бланке вопросника свои имя, фамилию и адрес электронной почты. Эта информация впоследствии использовалась для поиска профилей в социальной сети «Вконтакте».

Всего было идентифицировано 943 профиля.

Участникам сообщалось, что они могут отказаться от участия в исследовании, чем воспользовалось менее 1% предполагавшихся участников. Участникам гарантировалась анонимность. Доступ к идентификаторам черт личности имел только один из авторов данной статьи. Идентификаторы черт не передавались третьим лицам, включая соавторов данной статьи. После сбора лингвистических данных они были объединены с данными черт личности пользователей в одну матрицу, после чего идентификаторы были удалены из матрицы, а строки рандомизированы. Таким образом, материал был полностью обезличен.

Используемый опросник позволил получить данные о выраженности пяти психологических черт личности: экстраверсии / интроверсии, доброжелательности / враждебности, добросовестности / недобросовестности, нейротизма / эмоциональной стабильности, открытости новому опыту / консерватизма. Обработка полученных данных осуществлялась по стандартному ключу опросника BFI [17].

Каждая из психологических характеристик личности описывалась с помощью пятибалльной шкалы проявления двух противопоставленных признаков, вычисляемых на основе данных математического ожидания (M) и стандартного отклонения (SD): «++» – максимальное проявление признака (M±2SD), «+» – значимое проявление признака (M±SD), «0» – признак не выражен. Например, экстраверсия / интроверсия информанта может описываться как сильно выраженная экстраверсия («экстраверсия++» – M+2SD), или выраженная экстраверсия («экстраверсия+» – M+SD), или невыраженная экстраверсия / интроверсия («0»), или выраженная интроверсия («интроверсия+» – M-SD), или сильно выраженная интроверсия («интроверсия++» – M-2SD).

Значения математического ожидания (M) и стандартного отклонения (SD) для пяти шкал: экстраверсия / интроверсия: $M = 3.38$, $SD = 0.71$; доброжелательность / враждебность: $M = 3.47$, $SD = 0.58$, добросовестность / недобросовестность: $M = 3.34$, $SD = 0.65$, нейротизм / эмоциональная стабильность: $M = 3.06$, $SD = 0.73$, открытость новому опыту / консерватизм: $M = 3.76$, $SD = 0.64$.

Лингвистический анализ материала в проекте “Social Network Analysis”

В проекте Social Network Analysis материал исследования представляет собой данные о профилях участвовавших в психологическом опросе поль-

зователей и их тексты в социальной сети «ВКонтакте» (vk.com). Для сбора информации из социальной сети «ВКонтакте» была использована платформа API ВКонтакте (интерфейс, который позволяет получать информацию из базы данных с помощью http-запросов к специальному серверу). Стандартные средства API «ВКонтакте» позволяют собрать данные о профиле пользователя: книгах, фильмах, музыке и т.д., однако не предоставляют возможность получить все его комментарии одним запросом. Эта проблема была решена путем автоматического перебора комментариев к записям на личных страницах пользователя и его друзей и проверкой их авторства. Все полученные сведения были собраны в одной базе и обезличены. Общий объем материала – 19 179 автоматизировано собранных реплик 340 пользователей, прошедших психологический опрос.

На следующем этапе осуществлялась их загрузка в ИС «Семограф». На рис. 3 в качестве примера представлено окно контекста проекта «Social Network Analysis», в котором размещена вся собранная информация об одном пользователе. Сверху слева дано название контекста – в данном случае номер информанта (345). Ниже приводится сам контекст, который в данном проекте представляет собой совокупность реплик информанта в социальной сети «ВКонтакте» с пометами о времени их размещения. Справа показаны компоненты – отдельные реплики данного информанта. Снизу приводятся метаполя, приписанные данному контексту (пол, возраст информанта и его психологические параметры, выявленные в результате психологического опроса). Социальная информация часто представлена факультативно (если отмечена пользователем в профиле), психологические параметры отмечены у каждого информанта.

Для лингвистического анализа материала был разработан многоуровневый классификатор, учитывающий такие языковые параметры, как дейктические показатели, модальность, субъективно-оценочные значения, использование эмотиконов, бранной лексики и др. Процедура классификации состояла в приписывании каждой реплики к определенным ячейкам классификатора на основании представленности в данной реплике определенного языкового параметра.

При проведении классификации компонентов в проекте соблюдаются следующие принципы:

1) классификация проводилась несколькими экспертами (в процессе классификации вырабатывалась согласованная позиция всех экспертов по спорным вопросам);

2) каждое поле (класс) формировалось рядом лингвистических единиц, которые обладали общим признаком (данный признак может иметь любую природу, как лингвистическую – грамматическую, семантическую, синтаксическую, стилистическую и т.д., так и экстралингвистическую);

3) одна реплика могла быть отнесена к нескольким полям (если включала несколько лингвистических единиц, рассматриваемых в исследовании, например одновременно бранную лексику и эмотикон).

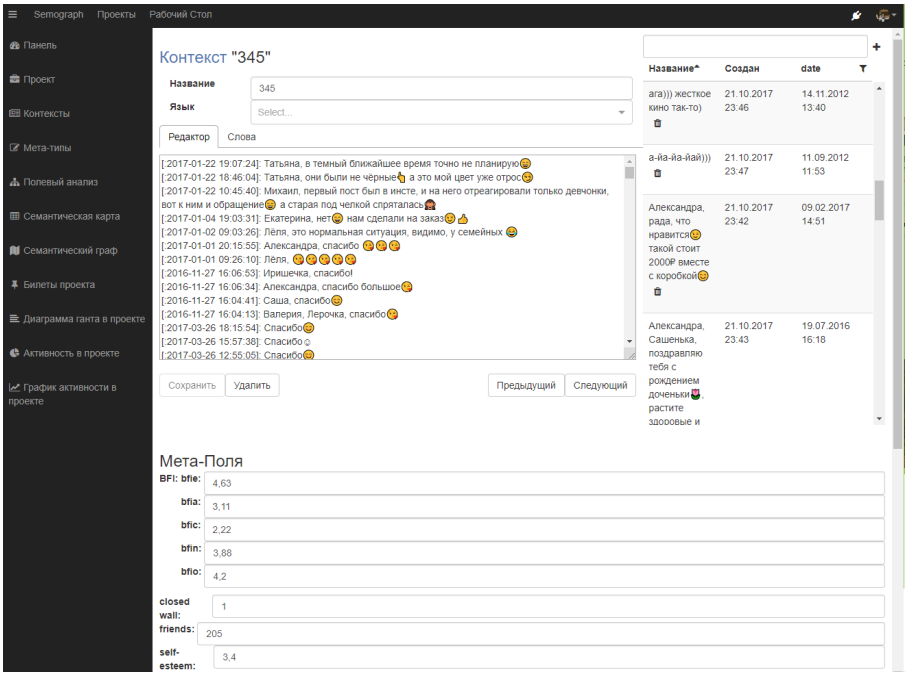


Рис. 3. Окно контекста проекта «Social Network Analysis» в информационной системе «Семограф»

В данном проекте при классификации использовалась иерархическая система вложенных полей (максимальное количество уровней – 4). Пример окна классификации приведен на рис. 4. Интерфейс классификации экспериментальных реакций состоит из трех столбцов: ПОЛЯ, КОМПОНЕНТЫ и КОНТЕКСТЫ.

В левом столбце ПОЛЯ приведены семантические поля; в среднем КОМПОНЕНТЫ – реплики пользователей социальной сети; в правом КОНТЕКСТЫ – множества реплик тех информантов, которые использовали данный компонент. В столбце КОМПОНЕНТЫ отражаются частотность употребления реплик во всем корпусе реакций (столбец C) и количество вхождений данной единицы в семантические поля (столбец F). Например, реплика *Блин, это и есть Курт)* (факт вхождения фиксируется под реакцией в виде списка семантических полей и цифрой «3» столбца F, передающего количество разных полей, в которые входит реакция). В данном примере реплика относится к полям БЛИН (ЭВФЕМИЗМЫ БРАНИ), БЛИЗКО (ПРОСТРАНСТВЕННЫЙ ДЕЙКСИС) и КОНЕЦ ВЫСКАЗЫВАНИЯ (указание на расположение ЭМОТИКОНА).

Возможности иерархической классификации позволяют каждому эксперту создавать и разрабатывать отдельную ветку классификации (например, отдельно от других экспертов анализировать ссылки или стилистику реплик). Таким образом, один и тот же материал рассматривается разными

экспертами с различных точек зрения и создается его многопараметрическая лингвистическая классификация. Разработанный классификатор, размеченный контент и социальные параметры пользователей составили базу данных «Речевые и неречевые параметры пользователей социальной сети» [18].

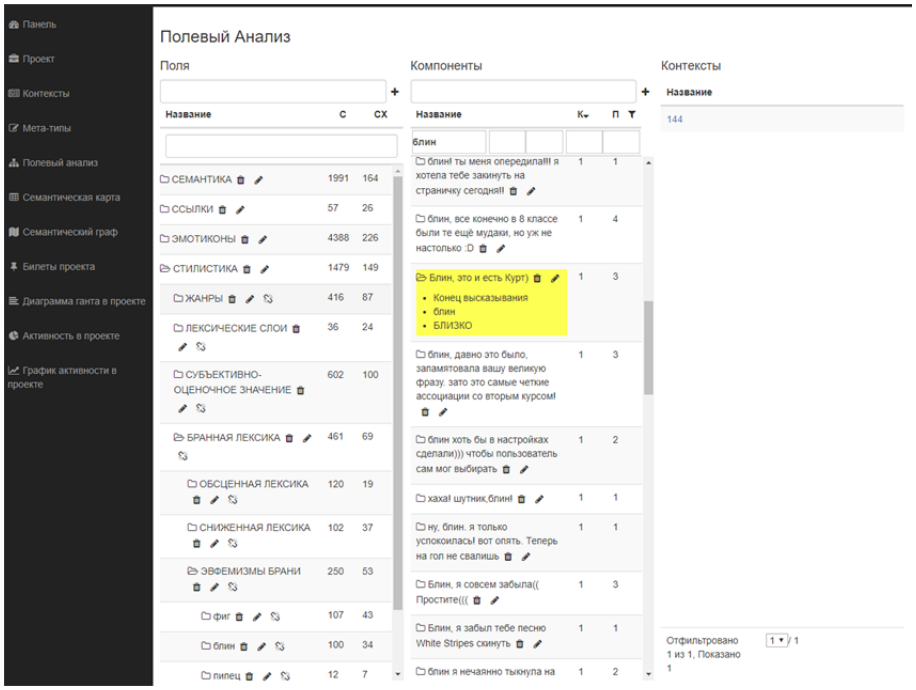


Рис. 4. Иерархическая классификация компонентов
в проекте «Social Network Analysis»

На следующем этапе результаты полевого анализа реакций информантов обрабатывались с помощью инструментария ИС «Семограф»: автоматически вычислялись объемы семантических полей и строилась таблица сопряженности, отражающая распределение семантических полей по выделенным метаполям, а также семантические графы, визуализирующие связи между социальными и психологическими параметрами информантов, с одной стороны, и особенностями их речевого поведения – с другой.

Результаты проекта Social Network Analysis

Результаты анализа генерируются в ИС «Семограф», как уже упоминалось выше, в виде семантических карт и семантических графов разного типа. В настоящем проекте для визуализации используется круговой граф с кольцевой иерархической шкалой и дополнительной шкалой фильтрации.

Опишем структуру графа. На круговой шкале на одной из частей окружности (нижней) представлен небольшой набор выделенных языко-

вых параметров, на другой (верхней) – даны психологические параметры пользователей. «Психологические» вершины графа соединяются с «языковыми» вершинами; при этом толщина линии пропорциональна силе связи между параметрами и можно активировать связи отдельного узла графа (как «психологического», так и «лингвистического»), наводя на него курсор и выделяя левой клавишей мыши. Кроме того, результаты, представленные на графе, можно дополнительно отфильтровать по гендерной принадлежности информантов (шкала фильтрации по социальным параметрам находится ниже круговой шкалы). Интерактивный граф результатов проекта Social Network Analysis доступен по ссылке (http://graph.semograph.org/cgraph/psycho).

Покажем возможности ИС «Семограф» при моделировании взаимосвязей лингвистических параметров речи и социопсихологических характеристик личности на примере анализа употребления бранных слов, особенностей дейксиса и некоторых других лингвистических характеристик.

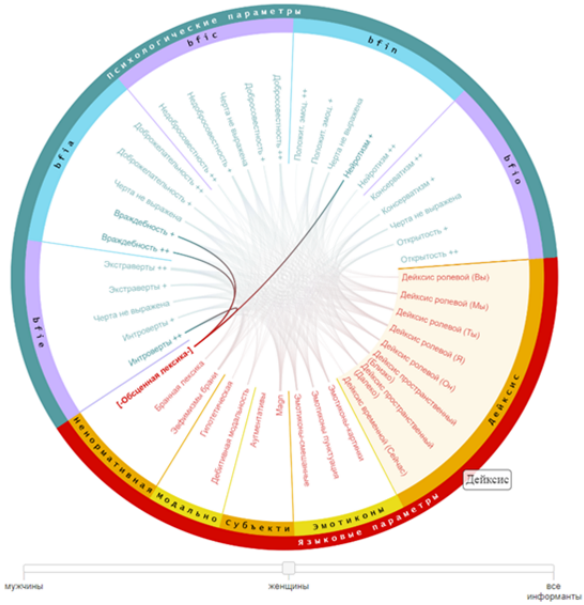
Социально-психологические параметры использования ненормативной лексики

При анализе употребления бранной лексики в открытой сетевой коммуникации рассматривались три группы этой лексики (обсценная лексика, брань и эвфемизмы брани) в речи мужчин и женщин.

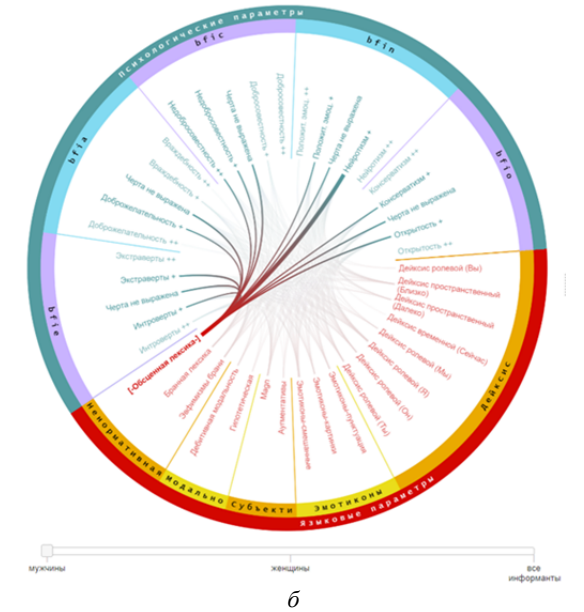
На рис. 5 показаны психологические характеристики женщин и мужчин, использующих в своей письменной публичной речи обсценную лексику. На графе видно, что женщины, обращающиеся к обсценной лексике, характеризуются ярко выраженной интроверсией, враждебностью и нейротизмом, т.е. сочетанием асоциальных психологических черт; у мужчин не наблюдается столь четкой привязки к психологическим характеристикам, среди мужчин, использующих обсценную лексику, встречаются люди с невыраженными чертами по BFI, с чертами, выраженными сильно и слабо, имеющими положительный или отрицательный знак. Таким образом, обсценную лексику в речи используют женщины с определенным ограниченным набором психологических характеристик; в отношении мужчин нельзя говорить об однозначном соответствии между обращением к обсценной лексике и психологическими чертами. Следовательно, обсценная лексика может выступать маркером психологических черт только для женщин.

На рис. 6 показано использование бранной лексики, на рис. 7 – эвфемизмов брани в зависимости от психологических характеристик информантов и их пола. Сравнение графов на рис. 5, *а*; 6, *а*; 7, *а* показывает, что для женщин, использующих обсценную лексику, брань и эвфемизмы брани в открытых постах социальной сети, спектр психологических характеристик тем шире, чем менее грубой оказывается лексика. Так, обсценную лексику использовали только женщины с ярко выраженной интроверсией, враждебностью и нейротизмом; бранную лексику используют женщины с выраженной враждебностью и нейротизмом, однако уже не только интроверты, но и экстраверты; эвфемизмы брани встречаются у женщин не

только с нейротизмом, но и с положительным эмоциональным фоном, не только у характеризующихся враждебностью, но и у тех, у кого данная психологическая черта не выражена, у интровертов и экстравертов.

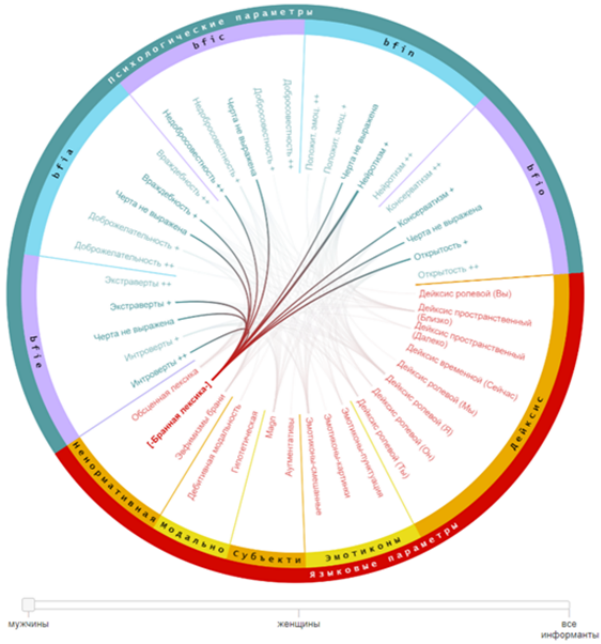


а

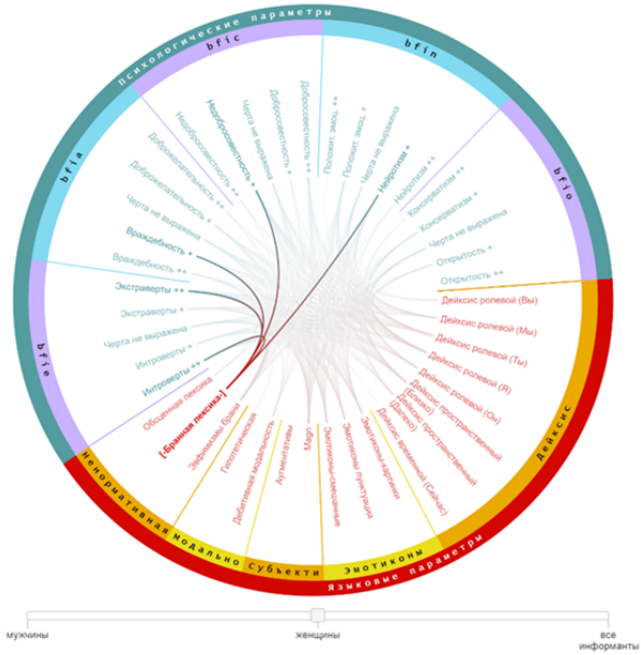


б

Рис. 5. Социально-психологические параметры использования обсценной лексики:
а – женщинами; *б* – мужчинами



а



б

Рис. 6. Социально-психологические параметры использования бранной лексики: *а* – женщинами; *б* – мужчинами

Рис. 5, *б*; 6, *б*; 7, *б*, показывающие психологическую обусловленность использования грубой лексики у мужчин, демонстрируют широкий спектр психологических черт вне зависимости от степени грубости используемой лексики. Это свидетельствует о том, что мужчины используют и обсценную лексику, и брань, и эвфемизмы брани вне зависимости от психологических характеристик, не различая стилистически эти пласты лексики, в то время как женщины чувствуют стилистическую разницу между эвфемизмами брани и собственно бранными словами, и последние используются только женщинами с определенными психологическими чертами.

Таким образом, мы видим, что психологические параметры не являются универсальными для мужчин и женщин: для каждого гендера наблюдается свой набор психологических характеристик, связанных с использованием грубой лексики. Причины этого лежат, очевидно, в истории русской культуры и традициях: ранее обсценная лексика была табуирована и могла использоваться только мужчинами при совершении обрядовых действий, позднее ее использование уже не было табуировано, но строго ограничивалось мужским обществом [19, 20].

Лингвистические параметры экстраверсии / интроверсии

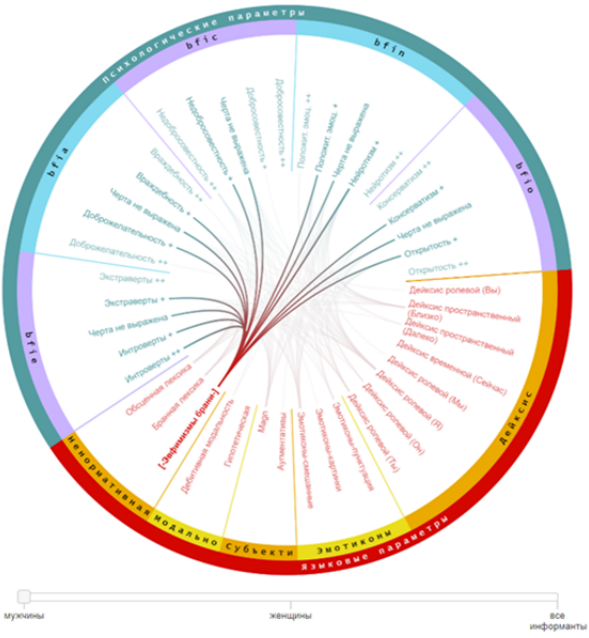
Обратимся теперь к лингвистическим характеристикам речи интровертов и экстравертов (рис. 8–9). Здесь рассматриваются только наиболее выраженные типы экстраверсии и интроверсии: «интроверты ++» у мужчин и женщин, «экстраверты++» у женщин и «экстраверты+» у мужчин (среди информантов-мужчин не встретилось экстравертов типа «++», что само по себе является интересным фактом взаимодействия социальных и психологических параметров).

Особенности речевого поведения интровертов-женщин и интровертов-мужчин (рис. 8) оказываются весьма сходными: и те, и другие говорят в своей публичной речи в социальных сетях о себе (ролевой дейксис «Я»), употребляют высшую степень проявления признака (Magn), а также эвфемизмы брани. Наблюдается лишь небольшая разница между мужчинами и женщинами в том, что при выборе дейктических категорий женщины более открыты ближайшему окружению и используют показатель «ТЫ», в то время как мужчины обозначают ближайшее пространство (пространственный дейксис «близко»).

Что касается экстравертов (рис. 9), то здесь наблюдается совсем иная картина. Лингвистические черты речи экстравертов-женщин практически ничем не ограничены: из рассмотренных лингвистических параметров женщинами-экстравертами не используется только 4 – обсценная лексика, дебитивная модальность, аугментативы. Спектр характеристик речи экстравертов-мужчин уже. Мужчины-экстраверты активно пользуются всеми видами грубой лексики, словами, выражающими высшую степень проявления признака (Magn), эмотиконами, состоящими из знаков препинания или смешанными (но не отдельными эмотиконами-картинками), и роле-

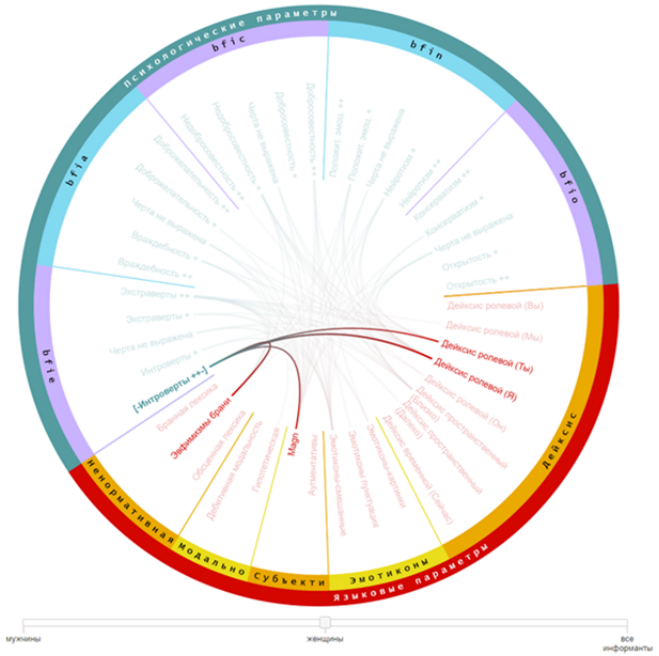
вым дейксисом, избегая социально маркированной формы («ВЫ»), а также пространственного и временного дейксиса.

а

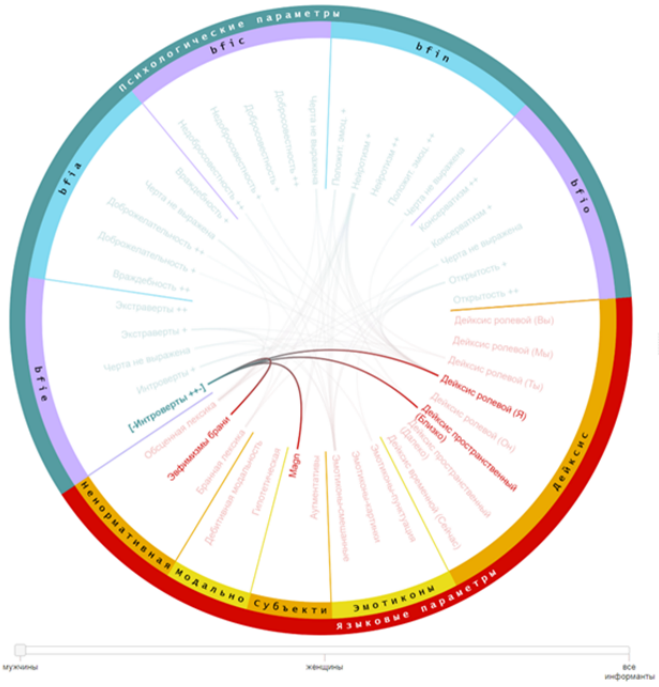


б

Рис. 7. Социально-психологические параметры использования эвфемизмов брани: *а* – женщинами; *б* – мужчинами

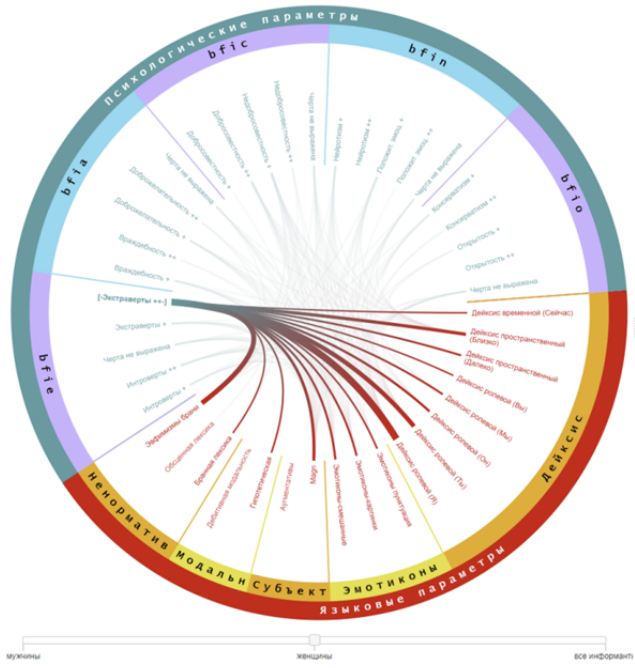


а



б

Рис. 8. Лингвистические параметры интроверсии:
а – у женщин; *б* – у мужчин



а

б

Рис. 9. Лингвистические параметры экстраверсии:
а – у женщин; *б* – у мужчин

Как видим, для интровертов – как мужчин, так и женщин – характерна бóльшая избирательность языковых средств, чем для экстравертов; при этом у женщин данная черта проявляется значительно ярче.

В целом проведенный анализ показывает сложную взаимозависимость психологических, социальных и речевых характеристик личности.

Заключение

В настоящее время существует большое количество программных продуктов, так или иначе связанных с анализом языкового материала, однако в лингвистических исследованиях они используются относительно редко, поскольку не дают тех методологических, технологических и организационных возможностей, которые необходимы для языковедческого анализа. В этой связи очевидна необходимость в создании доступных и многофункциональных инструментов для решения широкого спектра лингвистических задач, связанных с моделированием фрагментов языковой и социокультурной действительности на основе анализа языковых / текстовых массивов.

В статье показаны возможности анализа лингвистического материала на платформе информационной системы «Семограф». «Семограф» позволяет организовать удаленную многопользовательскую работу над проектами и управлять работой коллектива. Инструментарий включает методы создания многоуровневых систем классификации, контент-анализа и частотного анализа текстов и текстовых корпусов, осуществление лингвистической и экстралингвистической разметки текстов и др. В системе осуществляется полный цикл исследования, включая сбор материала, обработку и экспертный анализ данных, построение моделей, основанных на принципах редактируемой визуализации.

Работа системы показана на примере научного проекта «Social Network Analysis», посвященного разработке социокогнитивной модели пользователя социальной сети на основе многопараметрического анализа речевого поведения, социальных параметров и психологических характеристик личности. Применение «Семографа» позволило сопоставить социальные, психологические и языковые характеристики пользователей социальной сети, а разработанные средства визуализации предоставили возможность на основе анализа имеющихся и отсутствующих связей между отдельными языковыми и психологическими параметрами получить значимую для предметной области информацию. Так, были выявлены различия в использовании средств ролевого (в том числе социально маркированного) и пространственного дейксиса у пользователей-экстравертов и интровертов. Интересны различия в использовании бранной и обсценной лексики в письменной речи пользователей, имеющих ярко выраженные черты консерватизма и открытости и мн. др. Кроме того, речевая вариативность может объясняться не только психологическими различиями, но и гендерными (в частности, публичное использование обсценной лексики имеет связи с разными

психологическими характеристиками в мужской и женской группах пользователей), точнее, взаимодействием психологических и социальных параметров.

Развитие современных информационных технологий позволяет сохранять результаты речевой деятельности отдельных людей и социума в целом – различные варианты спонтанных, полуспонтанных и подготовленных письменных текстов. В силу системного характера когниции и больших объемов доступного текстового материала появляется возможность анализировать не только собственно лингвистические характеристики текста, но и отражение в них психологических и социальных черт говорящих. Информационная система «Семограф» позволяет, с одной стороны, работать с большими массивами текстов, используя лингвистическую и экстралингвистическую разметку, с другой стороны, применять сетевую модель организации исследований, что в совокупности дает преимущества при создании моделей фрагментов языковой и социокультурной действительности.

Литература

1. *Кастельс М.* Галактика Интернет: Размышления об Интернете, бизнесе и обществе. Екатеринбург : У-Фактория, 2004. 328 с.

2. *Пурдехнад Д.* Открытые инновации и социальные сети // Проблемы управления в социальных системах. 2012. Т. 4, № 7. С. 22–27.

3. *Cooke N.J., Hilton M.L.* (Eds.). Enhancing the Effectiveness of Team Science / Committee on the Science of Team Science; Board on Behavioral, Cognitive, and Sensory Sciences; Division of Behavioral and Social Sciences and Education; National Research Council. Washington DC : The National Academies Press, 2015. 256 p.

4. *Citizen* science. URL: https://en.wikipedia.org/wiki/List_of_citizen_science_projects (date of access: 03.08.2018).

5. *Рябинин К.В., Баранов Б.Д., Белоусов К.И.* Интеграция информационной системы Семограф и визуализатора SciVi для решения задач экспертного анализа языкового контента // Научная визуализация. 2017. № 4. С. 67–77.

6. *Белоусов К.И.* Теория и методология полиструктурного синтеза текста. М. : Флинта : Наука, 2009. 216 с.

7. *Baranov D.A., Belousov K.I., Ichkineeva D.A., Zelyanskaya N.L.* The network organization of experimental research in linguistics: opportunities and prospects // Procedia – Social and Behavioral Sciences. 2015. Vol. 214. P. 958–964.

8. *Liu D., Baumeister R.F.* The Big Five Personality Traits, Big Two Metatraits and Social Media: A Meta-Analysis // Journal of Research in Personal. 2017. Vol. 70. P. 229–240.

9. *Morrison M.A., Cheong H.J., McMillan S.* Posting, Lurking, and Networking: Behaviors and Characteristics of Consumers in the Context of User-Generated Content Morrison // Journal of Interactive Advertising. 2013. Vol. 13, № 2. P. 97–108.

10. *Nadkarni A.* Why Do People Use Facebook? // Personality and Individual Differences. 2012. Vol. 52, № 3. P. 243–249.

11. *Pentina I., Zhang L.* Effects of Social Support and Personality on Emotional Disclosure on Facebook and in Real Life // Behaviour and Information Technology. 2017. Vol. 36, № 5. P. 484–492.

12. *Wang X., Li Y.* Users' Satisfaction with Social Network Sites: A Self-Determination Perspective // Journal of Computer Information Systems. 2015. Vol. 56, № 1. P. 48–54.

13. *Zuniga H.G. de, Diehl T., Huber B., Liu J.* Personality Traits and Social Media Use in 20 Countries: How Personality Relates to Frequency of Social Media Use, Social Media News Use, and Social Media Use for Social Interaction // Cyberpsychology, Behavior, And Social Networking. 2017. Vol. 20, № 9. P. 540–552.

14. *John O.P., Donahue E.M., Kentle R.L.* The Big-Five Inventory-Version 4a and 54. Berkeley, CA : Berkeley Institute of Personality and Social Research; University of California, 1991.

15. *John O.P., Naumann L.P., Soto C.J.* Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues // O.P. John, R.W. Robins, L.A. Pervin (eds.). Handbook of personality: Theory and research. New York, NY : Guilford Press, 2008. P. 114–158.

16. *Shchebetenko S.* Reflexive Characteristic Adaptations Explain Sex Differences in the Big Five: but not in Neuroticism // Personality and Individual Differences. 2017. Vol. 111. P. 153–156.

17. *Shchebetenko S.* "The best man in the world": Attitudes toward personality traits // Psychology. Journal of the Higher School of Economics. 2014. Vol. 11, № 3. P. 129–148.

18. *База* данных «Речевые и неречевые параметры пользователей социальной сети»: Свидетельство о государственной регистрации базы данных, охраняемой авторскими правами / Баранов Д.А., Белоусов К.И., Боронникова Н.В., Ерофеева Е.В., Зелянская Н.Л., Константинов И.М., Обухова И.А., Руденко Е.С., Русинова И.И., Худякова Е.С. М. : Федеральная служба по интеллектуальной собственности. Внесена в реестр баз данных, регистрационный № 2018621839 от 20.11.2018.

19. *Мокиенко В.М.* Русская бранная лексика: цензурное и нецензурное // Русистика. 1994. № 1/2. С. 50–73.

20. *Успенский Б.А.* Мифологический аспект русской экспрессивной фразеологии (статья первая) // Studia Slavica Hungarica. 1983. Vol. 29. P. 33–69.

The Multi-Parameter Analysis of Linguistic Data in the Information System Semograf (On the Example of the Study of Social Network Users' Speech)

Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology. 2020. 64. 6–29. DOI: 10.17223/19986645/64/1

Konstantin I. Belousov, *Elena V. Erofeeva*, *Dmitriy A. Baranov*, *Natalya L. Zelyanskaya*, Perm State University (Perm, Russian Federation). E-mail: belousovki@gmail.com / elenerofee@gmail.com / baranov@semograph.com / zelyanskaya@gmail.com

Sergei A. Shchebetenko, Higher School of Economics (Moscow, Russian Federation). E-mail: shebetenko@rambler.ru

Keywords: network science, social network-services, information system Semograph, multiparameter analysis, visual analytics, semantic graph modeling.

The aim of this article is to demonstrate the capabilities of the information system Semograph (http://semograph.org) as a tool for text content analysis when implementing a network approach to the organization of scientific research in linguistics. Semograph can be used for the analysis of text data, creation and/or annotation of language/text corpora, conducting, processing and analysis of psycholinguistic and sociolinguistic experiments, development of classifiers and thesauri, and solving other problems that arise when analyzing language material. Semograph implements the principles of a full research cycle, network distribution of research participants, a multi-user mode of operation and methodological pluralism. The possibilities of network organization of work in Semograph are shown on the example of a multiparametric analysis of speech behavior, social parameters and psychological characteristics of users of the social network VKontakte. The total volume of the automatically collected material is 18,126 utterances of 340 users who have completed a psychological survey of BFI, according to which results of the severity of the five psychological personal traits (extraversion vs. introversion, agreeableness vs. antagonism, conscientiousness vs. lack of direction,

neuroticism vs. emotional stability, openness vs. closedness to experience) are determined. For the analysis of the text material, a multi-level hierarchical classifier was developed that allows each expert-linguist to create and develop a separate classification branch (thus, the same material is considered by different experts from different points of view, and its multi-parametric linguistic classification is created). This classification and specific user metadata (gender, psychological characteristics, etc.) provide the basis for constructing a model of interrelations between linguistic parameters of speech and socio-psychological characteristics of a person by means of interactive visual analytics. The article demonstrates these interrelations on the example of differences in the use of role and spatial deixis tools by extroverts and introverts, abusive and obscene lexical unites by users with a strong tendency for closedness and openness to experince, etc. The resulting model shows that the speech variability of texts is due to the interaction of psychological and gender characteristics of the informants, rather than a single act of these factors. In general, the article demonstrates that the information system Semograph allows, on the one hand, analyzing large arrays of texts with linguistic and extra-linguistic annotations, on the other hand, applying a network model of research organization that in the aggregate gives advantages in constructing models of fragments of linguistic and sociocultural reality.

References

1. Castells, M. (2004) *Galaktika Internet: Razmyshleniya ob Internete, biznese i obshchestve* [The *Internet Galaxy*: *Reflections* on the *Internet*, *Business*, and *Society*]. Translated from English. Yekaterinburg: U-Faktoriya.

2. Pourdehnad, D. (2012) Open Innovations and Social Networking. *Problemy upravleniya v sotsial'nykh sistemakh – Problems of Governance*. 4 (7). pp. 22–27. (In Russian).

3. Cooke, N.J. & Hilton, M.L. (eds) (2015) *Enhancing the Effectiveness of Team Science*. Washington DC: The National Academies Press.

4. Wikipedia. (2018) *Citizen Science*. [Online] Available from: https://en.wikipedia.org/wiki/List_of_citizen_science_projects. (Accessed: 03.08.2018).

5. Ryabinin, K.V., Baranov, B.D. & Belousov, K.I. (2017) Integration of Semograph Information System and SciVi Visualizer for Solving the Tasks of Lingual Content Expert Analysis. *Nauchnaya vizualizatsiya – Scientific Visualization*. 4. pp. 67–77. (In Russian). DOI: 10.26583/sv.9.4.07

6. Belousov, K.I. (2009) *Teoriya i metodologiya polistrukturnogo sinteza teksta* [Theory and Methodology of Multistructural Text Synthesis]. Moscow: Flinta: Nauka.

7. Baranov, D.A., Belousov, K.I., Ichkineeva, D.A. & Zelyanskaya, N.L. (2015) The network organization of experimental research in linguistics: opportunities and prospects. *Procedia – Social and Behavioral Sciences*. 214. pp. 958–964. DOI: 10.1016/j.sbspro. 2015.11.681

8. Liu, D. & Baumeister, R.F. (2017) The Big Five Personality Traits, Big Two Metatraits and Social Media: A Meta-Analysis. *Journal of Research in Personal*. 70. pp. 229–240.

9. Morrison, M.A., Cheong, H.J. & McMillan, S. (2013) Posting, Lurking, and Networking: Behaviors and Characteristics of Consumers in the Context of User-Generated Content Morrison. *Journal of Interactive Advertising*. 13 (2). pp. 97–108.

10. Nadkarni, A. (2012) Why Do People Use Facebook? *Personality and Individual Differences*. 52 (3). pp. 243–249.

11. Pentina, I. & Zhang, L. (2017) Effects of Social Support and Personality on Emotional Disclosure on Facebook and in Real Life. *Behaviour and Information Technology*. 36 (5). pp. 484–492.

12. Wang, X. & Li, Y. (2015) Users' Satisfaction with Social Network Sites: A Self-Determination Perspective. *Journal of Computer Information Systems*. 56 (1). pp. 48–54.

13. Zuniga, H.G. de, Diehl, T., Huber, B. & Liu, J. (2017) Personality Traits and Social Media Use in 20 Countries: How Personality Relates to Frequency of Social Media Use, So-

cial Media News Use, and Social Media Use for Social Interaction. *Cyberpsychology, Behavior, And Social Networking*. 20 (9). pp. 540–552.

14. John, O.P., Donahue, E.M. & Kentle, R.L. (1991) *The Big-Five Inventory-Version 4a and 54*. Berkeley, CA: Berkeley Institute of Personality and Social Research; University of California.

15. John, O.P., Naumann, L.P. & Soto, C.J. (2008) Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In: John, O.P., Robins, R.W. & Pervin, L.A. (eds) *Handbook of Personality: Theory and Research*. New York, NY: Guilford Press. pp. 114–158.

16. Shchebetenko, S. (2014) "The best man in the world": Attitudes toward personality traits. *Psychology. Journal of the Higher School of Economics*. 11 (3). pp. 129–148.

17. Shchebetenko, S. (2017) Reflexive Characteristic Adaptations Explain Sex Differences in the Big Five: but not in Neuroticism. *Personality and Individual Differences*. 111. pp. 153–156.

18. Baranov, D.A. et al. (2018) *Baza dannykh "Rechevye i nerechevye parametry pol'zovateley sotsial'noy seti": Svidetel'stvo o gosudarstvennoy registratsii bazy dannykh, okhranyaemoy avtorskimi pravami* [Database "Speech and Non-Speech Parameters of Social Network Users": Certificate of State Registration of a Database Protected by Copyright]. Moscow: Federal Service for Intellectual Property. Registration No. 2018621839 of 20 November 2018.

19. Mokienko, V.M. (1994) Russkaya brannaya leksika: tsenzurnoe i netsenzurnoe [Russian Obscene Words: The Censored and the Obscene]. *Rusistika*. 1/2. pp. 50–73.

20. Uspenskiy, B.A. (1983) Mifologicheskiy aspekt russkoy ekspressivnoy frazeologii (stat'ya pervaya) [The Mythological Aspect of Russian Expressive Phraseology (Article One)]. *Studia Slavica Hungarica.* XXIX. pp. 33–69.