УДК 81:0/9

ЭКСПЕРИМЕНТАЛЬНАЯ ЛИНГВИСТИКА И СЕТЕВАЯ НАУКА [1]

Константин Игоревич Белоусов
д. филол. н., профессор кафедры теоретического и прикладного языкознания
Пермский государственный национальный исследовательский университет
614990, г. Пермь, ул. Букирева, 15. belousovki@gmail.com

В статье рассматривается современное состояние экспериментальной лингвистики и ее основные задачи в контексте использования информационных технологий. Обсуждаются варианты решения проблем репрезентативности данных; классификации лингвистических экспериментов; сбора, обработки, анализа и хранения данных; методологической агрессии и плюрализма и др. а также показываются возможности и перспективы организации экспериментальных исследований с помощью технологий сетевой науки.

Ключевые слова: экспериментальная лингвистика; лингвистический эксперимент; сетевая наука; информационная система «Семограф»; информационные технологии в лингвистике; деятельностная модель; онлайн-модели.

1. Введение. Состояние и задачи современной экспериментальной лингвистики

Несмотря на то, что для современного состояния лингвистических исследований характерно частое обращение к методам научного эксперимента, в области экспериментального изучения языка и речевой деятельности сложилась ситуация «технологического запаздывания» и необоснованной фрагментации ее предметного поля. Можно говорить о том, что в отличие от корпусных исследований, использующих возможности IT-сферы и имеющих впечатляющие результаты в виде отдельных масштабных корпусов национальных языков и др. частных корпусов, экспериментальный подход пока остается вне поля современных информационно-технологических способов познания языковой реальности. В то же время многие отечественные лингвистические школы за десятилетия их существования накопили (и продолжают накапливать) огромные массивы информации, полученной экспериментальными методами и представленной, как правило, в бумажном формате в словарях, картотеках, анкетах экспериментов и др. Ценность данной информации огромна, поэтому появляется насущная задача введения данной информации в научный оборот.

Таким образом, основной задачей современной экспериментальной лингвистики, на наш взгляд, является создание доступных и понятных широкому кругу лингвистов технологий и созданных на их основе программных средств, помогающих лингвисту-экспериментатору решать собственно научные задачи, поставленные в каждом отдельном исследовании. Среди частных исследовательских задач можно выделить следующие:

1) разработка классификатора лингвистических экспериментов (для того чтобы с помощью создаваемых технологий охватить как можно больше экспериментальных исследований);
2) решение проблемы репрезентативности экспериментальных выборок и проблемы достоверности результатов в целом;
3) решение проблемы хранения данных и организации доступа к ним;
4) разработка инструментария сбора, обработки и анализа данных;
5) решение проблемы взаимосвязи научных результатов.

Одна часть этих задач группируется вокруг деятельности лингвиста-экспериментатора, вторая часть – вокруг деятельности испытуемых, а третья связана с совместной деятельностью исследователя и информанта.

2. Информационные технологии в сфере обработки и анализа языковых/текстовых данных

Применение информационных технологий в сфере обработки и анализа языкового материала многомерно, т.к. может рассматриваться с позиций: 1) области/сферы применения (научной, образовательной, прикладной), 2) анализируемых форм существования языка (устной речи, письменного/печатного/электронного текста), 3) анализируемых уровней языковой системы, 4) типа

информационных продуктов (баз данных, баз знаний, программных средств), 5) реализуемых методов и др.

Текстуальность современного человекоразмерного бытия и растущие возможности в области получения, переработки и хранения информации актуализировали разработки в сфере интеллектуального анализа текста. Исследования в области text mining в связи с появившейся возможностью обрабатывать огромные массивы текстовой информации, представленной в социальных медиа, привели к созданию многочисленных систем, связанных с мониторингом социальных сетей, блогов и СМИ. Среди отечественных систем можно назвать SemanticForce (http://www.semanticforce.net), Медиалогию (http://www.mlg.ru), IQbuzz (http://www.iqbuzz.ru), Buzzlook (http://buzzlook.ru), Constrim DataRetriever (http://www.cdretr.com). Среди зарубежных аналогов можно отметить Vennmaker (http://www.vennmaker.com), Socialmention (http://socialmention.com), Alterian (http://www.alterian.com), J.D. Power and Associates (http://businesscenter.jdpower.com), Salesforce (http://www.salesforce.com).

Разнообразие методов и подходов к обработке и анализу текста (распространенных особенно широко в англоязычном секторе ресурсов, каталогов и программных средств) рассчитано на целевую аудиторию, представленную преимущественно специалистами в областях прикладной математики, искусственного интеллекта, маркетинга, с одной стороны, и связей с общественностью, социологии и т.п. – с другой, а также прикладной лингвистики, ориентированной на решение задач в перечисленных областях.

Конечно, для нас прежде всего представляют интерес разработки в области семантического анализа русскоязычных текстов. На специализированных порталах, посвященных компьютерной лингвистике, в частности, на портале Диалог (http://www.dialog-21.ru) и отечественных каталогах лингвистических ресурсов NLPub (http://nlpub.ru) и др. представлены как перечни отдельных аспектов семантического анализа, так и программные средства, созданные для работы в данных областях.

Представители лингвистических наук из данного разнообразия технологий и программных продуктов могут использовать для своих исследовательских целей очень ограниченный набор инструментов, в частности, национальные и специализированные языковые корпусы и словари. Программные продукты, созданные для обработки и анализа текста, большей частью не применимы для решения языковедческих задач, так как навязывают лингвисту ограниченный набор инструментов, вполне достаточный для социолога/контент-менеджера или аналитика в сфере создания предметно ориентированных тезаурусов, но нерелевантный целям и методам языковедческого анализа.

Если в области корпусной лингвистики есть существенные достижения в виде многочисленных корпусов, в том числе открытых (в частности, представленных на порталах http://www.ruscorpora.ru, http://opencorpora.org, http://nlpub.ru, http://lexrus.ru, http://spokencorpora.ru, http://rusling.narod.ru и мн. др.), то в сфере экспериментальной лингвистики присутствуют преимущественно ресурсы, представляющие результаты ассоциативных экспериментов (http://it-claim.ru/Projects/ASIS/index.htm, http://tesaurus.ru/dict/dict.php, http://adictru.nsu.ru). Однако работа с данными ресурсами сопряжена со следующими ограничениями:

1. Существующие ассоциативные словари фиксируют только обобщенные данные о реакциях (R) на слова-стимулы (S), например: *РОДИНА – мать* (65), *моя* (19); *зовет* (6); *любимая*, *одна*, *СССР* (5); *Отчизна* (4) и т.д. [Русский ассоциативный словарь: электр. ресурс]. Отсутствует возможность обращения к первичным данным, т.е. к реакциям отдельных информантов, что значительно сужает область применения данных ресурсов.

2. Существующие ресурсы не предполагают возможности проведения собственных экспериментов и интеграцию полученных данных с уже имеющимися.

3. Ассоциативные словари построены с использованием результатов одной из возможных методик – первой реакции на предлагаемый стимул. В то же время ассоциативные эксперименты могут проводиться с использованием других методик.

4. Представление материала в существующих ассоциативных словарях позволяет фиксировать только отношения типа S – R и R – S. В то же время отсутствует описание структуры отношений внутри реакций (например, в случае цепочечного эксперимента). При обращении к цепочечному эксперименту следует учитывать порядок реакций, который может прояснить значение реакции для многозначного слова.

5. Ассоциативные словари не позволяют исследовать в реакциях одних и тех же испытуемых все множество их реакций на предъявляемые стимулы. Методика проведения ассоциативных экспериментов с одной реакцией на стимул дает возможность получения от каждого информанта нескольких десятков пар S – R за сеанс. Для изучения проблем языкового сознания, языковой личности и др. особый интерес представ-

ляют «персонологические» (индивидуальные) распределения языкового материала, а не усредненные данные по всему корпусу.

6. Ассоциативные словари почти не используют метаданные, характеризующие реакции, информантов, особенности экспериментальной ситуации.

7. Отсутствуют инструменты обработки и анализа экспериментальных данных, в частности, множественной классификации реакций (как с использованием разных классификаторов, так и осуществляемой несколькими экспертами) и построения моделей, в том числе визуальных, на основе созданных классификаций и др.

8. Отсутствует привязка ко времени полученной экспериментальной реакции.

Обозначенные проблемы показывают теоретическую и прикладную уязвимость существующих информационных продуктов в данной сфере.

3. Деятельностная модель лингвиста-экспериментатора в контексте использования современных информационных технологий

Деятельностная модель лингвиста-экспериментатора может быть разработана на основе типичных целей, задач, этапов проведения каждого исследования и их инструментально-технологического решения.

3.1. Сбор, обработка и анализ данных

Применение современных ИТ позволяет поновому организовать работу исследователя (в том числе и в паре «исследователь – информант/ы») на каждом ключевом этапе эксперимента: планирования, сбора данных, их обработки и анализа.

Сбор данных, т.е. проведение эксперимента с привлечением информантов к работе над стимульным материалом, сегодня может осуществляться с помощью web-технологий в контексте развиваемых нами идей сетевой науки. В узком понимании **сетевая наука** – это распределенный в режиме реального времени научный процесс, предполагающий организацию сетевого взаимодействия участников и систему управления исследовательской деятельностью, использование единых технологий обработки информации и общей базы данных, интегрирующих результаты исследовательской работы каждого участника в создаваемое информационное пространство проекта; а также наука, изучающая организацию сетевого взаимодействия участников научного процесса.

Реализация описанного взаимодействия представителей разных профессиональных групп (например, «исследователь – информант/ы») осуществляется посредством **сетевой распределенности** участников научного процесса, предполагающей возможность работы каждого вовлеченного в проект участника с разных машин. Благодаря распределенности появляется возможность привлекать к работе над проектом экспертов и информантов, географически отдаленных от исследователя. Сетевая распределенность обусловливает работу в многопользовательском режиме, т.е. в режиме онлайн-взаимодействия участников научного проекта. При этом сетевая распределенность должна реализовываться в архитектуре иерархий прав доступа к отдельным инструментам и функциям. Так, в частности, информанты не должны иметь доступ к экспериментальным данным других информантов.

Обработка данных (поиск, сортировка, фильтрация, генерирование выборок, таблиц, графиков и пр.; импорт, экспорт и др.) и их анализ с помощью разнообразных инструментов классификации (в частности, в рамках компонентного, полевого или тезаурусного моделирования), методов математического моделирования, научной визуализации и др., реализованные в рамках концепции сетевой науки, имеют очевидные преимущества. Так, например, решаются все проблемы представления данных в существующих ресурсах-словарях результатов ассоциативных экспериментов. Кроме того, экспорт данных в табличные форматы, позволяет затем осуществлять их анализ вне самой сетевой среды (например, исследователь может использовать обычные десктопные программные продукты, предназначенные для статистической обработки данных, визуального моделирования и др.).

3.2. Проблема репрезентативности материала

Рост количества экспериментальных исследований порождает проблему их качества, достоверности результатов, выносимых на обсуждение научной общественности. Решение проблемы репрезентативности исходных данных можно осуществить несколькими способами.

1. Каждая реакция информанта обязательно должна описываться набором метаданных – параметров, характеризующих самого информанта (возраст, пол, образование, специальность/сфера занятости и т.п.); кроме того, должны описываться стимульный материал и особенности проведения эксперимента (цели, задачи, условия, время, затрачиваемое информантом) и др. Описанные таким образом данные можно рассматривать в контексте эпистемологической ситуации и в каждом отдельном случае принимать решение о возможности их использования.

2. Для изучения влияния социальных параметров на речевую деятельность требуется со-

здавать сбалансированные выборки по данным параметрам (возраст, пол, образование, специальность/сфера занятости и др.) [Ерофеева Е.В, Ерофеева Т.И. 2010].

3. Из корпуса экспериментальных реакций автоматически вычленяется несколько случайных выборок. Результаты, относящиеся к наблюдаемым переменным каждой выборки, сопоставляются. Принадлежность выборок одной генеральной совокупности свидетельствует о репрезентативности экспериментального материала [Ярхо 2006]. В качестве варианта использования сбалансированных выборок можно предложить инструментарий автоматической генерации нескольких взвешенных выборок из рассматриваемого экспериментального материала и их последующее сопоставление. В основе автоматической генерации выборок можно использовать инструментарий фасетной классификации, широко распространенной в сфере информационного поиска и таргетирования Интернет-аудитории (например, [Smith 2008]).

4. Огромные массивы информации, полученной экспериментальными методами и представленной, как правило, в бумажном формате в словарях, картотеках, анкетах экспериментов должны быть введены в широкий научный оборот. Это позволит использовать данные, полученные с помощью обращения к тем же методам исследования, но в разное время и в разных регионах.

3.3. Проблема хранения данных и организация доступа к ним

Результаты экспериментов, внесенные исследователем в базу данных, должны бессрочно храниться в ней, а возможности экспорта в популярные офисные форматы деактуализируют проблему потери данных. Использование web-технологий позволяет 1) получать доступ к данным с любого устройства, имеющего выход в Интернет; 2) масштабировать нагрузки на ресурсы при обращении пользователей к программному средству.

Проблема доступа к данным в рамках концепции сетевой науки многоаспектна и включает в себя организацию доступа к собственным данным и предотвращение несанкционированного доступа к любым пользовательским данным. И так как сетевая наука предполагает распределенный в режиме реального времени научный процесс, организация этого научного процесса включает детализированную систему прав доступа, позволяющую пользователю:

- в закрытом режиме выполнять экспериментальные исследования, обрабатывать полученные данные, хранить результаты;
- привлекать к работе удаленных пользователей с разными правами (для участия в экспериментах, для анализа полученных данных и др.);
- делать результаты экспериментальных исследований доступными (под разными лицензиями) для других пользователей.

3.4. Проблема методологической агрессии и плюрализма

Недостатком многих существующих программных продуктов в области языковых/текстовых исследований является «вшитые» в них алгоритмы/схемы анализа, классификации, шкалы и т.п., заставляющие исследователей, не разделяющих те или иные концептуальные представления создателей программного средства, работать в рамках предлагаемых знаниевых структур. На наш взгляд, те или иные готовые решения должны не навязываться, а предлагаться исследователю. Конечно, полностью преодолеть методологическое давление на исследователя невозможно, т.к. программные средства содержат ограниченное количество шаблонов, форм представления данных, инструментов обработки и анализа. Однако эти ограничения можно рассматривать и как достоинство программного продукта, т.к. не навязывая исследователю априорных классификационных схем и шкал, фреймворк «требует» целенаправленной единообразной работы с данными.

4. Деятельностная модель информанта в эксперименте как основа классификации лингвистических экспериментов

Существующие классификации лингвистических экспериментов, широко распространенные в учебной и научной литературе, сложно использовать для создания программного продукта, так как они, как правило, классифицируют либо материал (например, эксперименты со словом, эксперименты с текстом), либо методы/методики (шкалирование, семантический дифференциал и т.п.) и т.п. (подробный обзор классификаций экспериментов со словом см. [Залевская 2011: 70-81])[2], и не в состоянии с общих позиций охватить всю предметную область экспериментальной лингвистики. На наш взгляд, для программной реализации экспериментальной деятельности необходимо 1) учитывать операционально-технологичес-кую составляющую деятельности информантов, 2) стремиться к наиболее широкому охвату существующих методов и методик, в границах которых осуществляется деятельность информантов.

Ниже предлагается следующая классификация лингвистических экспериментов:

Группа I. Реакция информанта на стимульный материал представлена **в языковой форме**. Стимульный материал может быть представлен в языковой и/или графической формах.

Класс 1. Реакция представляет собой языковую единицу или набор языковых единиц, соотносимых со стимулом.

1.1. Языковые единицы **создаются** информантом. Например: свободный ассоциативный эксперимент, направленный ассоциативный эксперимент с неограниченным количеством реакций; методика толкования значений слов; методики описания графических текстов; методика компрессии текста до цепочки ключевых слов [Сахарный 1989] и мн. др.

1.2. Языковые единицы **воспроизводятся (выбираются)** информантом из некоторого списка. Например: анализ времени отклика на распознавание значений слов в контексте или поиск соответствий эквивалентов трансформированных предложений ядерным предложениям [Miller 1962]; методика анализа текста посредством ключевых слов А.С. Штерн [Мурзин, Штерн 1991].

Класс 2. Реакция представляет собой редактирование стимульного материала.

2.1. Информант **не вносит** в стимульный материал **правку**, но делает его разметку. Например: методика анализа текста посредством ключевых слов (с помощью подчеркивания и под. инструментов) [Пищальникова 1999]; выявление определенных языковых единиц в большем наборе языковых единиц/тексте (возможно с использованием шума, создаваемого языковой графикой).

2.2. Информант **вносит правку** в стимульный материал. Например: методика дополнения текста; методика заполнения текстовых лакун [Белянин 1983].

Класс 3. Реакция представляет собой внесение новой языковой информации в сочетании с разметкой стимульного материала. Например, тематический анализ текста – выделение темы, микротем и отнесение лексического материала к выделенным микротемам [Белоусов, Ичкинеева 2011].

Класс 4. Реакция представляет собой результат структурирования/группировки языковых единиц.

4.1. Разнообразные эксперименты с классификацией, в том числе и с возможностью создания вертикально развернутых иерархических структур. Например: классификация лексики определенных тематических групп [Залевская 2011]; перечисление лексики определенной тематической/семантической группы и т.п. [Ерофеева, Пепеляева 2011].

4.2. Создание композиций языковых единиц из их некоторого набора. Например: методика рассыпного текста (при ослабленной цепной связи между предложениями) [Москальская 1981].

Группа II. Реакция информанта на стимульный материал представлена **в числовой форме**. Стимульный материал может быть представлен в языковой и/или графической формах.

Класс 1. Реакция представляет собой редактирование непосредственно стимульного материала. В качестве инструментов разметки используются числовые характеристики, приписываемые отдельным единицам стимульного материала. Например, моделирование процесса восприятия текста [Рубакин 1977].

Класс 2. Использование числовых данных для придания категориям веса, определяющего его значимость в общей системе категорий. Например, экспертный анализ и оценка соответствия языкового/текстового материала первичным образцам (компрессия текста, перевод и др.) [Сорокин 1985].

Класс 3. Использование числовых данных для оценки языкового материала по шкалам. Например: методы шкалирования, метод семантического дифференциала.

5. Информационная система «Семограф» как инструмент сетевой экспериментальной лингвистики

Информационная система (ИС) «Семограф» (http://semograph.com) является инструментом зарождающейся в настоящее время сетевой науки. Несмотря на то что ИС предназначена для извлечения знаний о предметных областях из информационных массивов, ее функциональность постоянно расширяется и на сегодняшний день дает возможность проведения экспериментальных исследований нескольких типов.

5.1. Сбор экспериментальных данных в ИС «Семограф»

ИС позволяет проводить эксперименты, относящиеся к классу 1 «Реакция представляет собой языковую единицу или набор языковых единиц, соотносимых со стимулом». На рис. 1 отображено окно редактирования контекста с данными одной анкеты, полученной в направленном цепочечном ассоциативном эксперименте (в реакциях информантов сохраняется авторское написание).

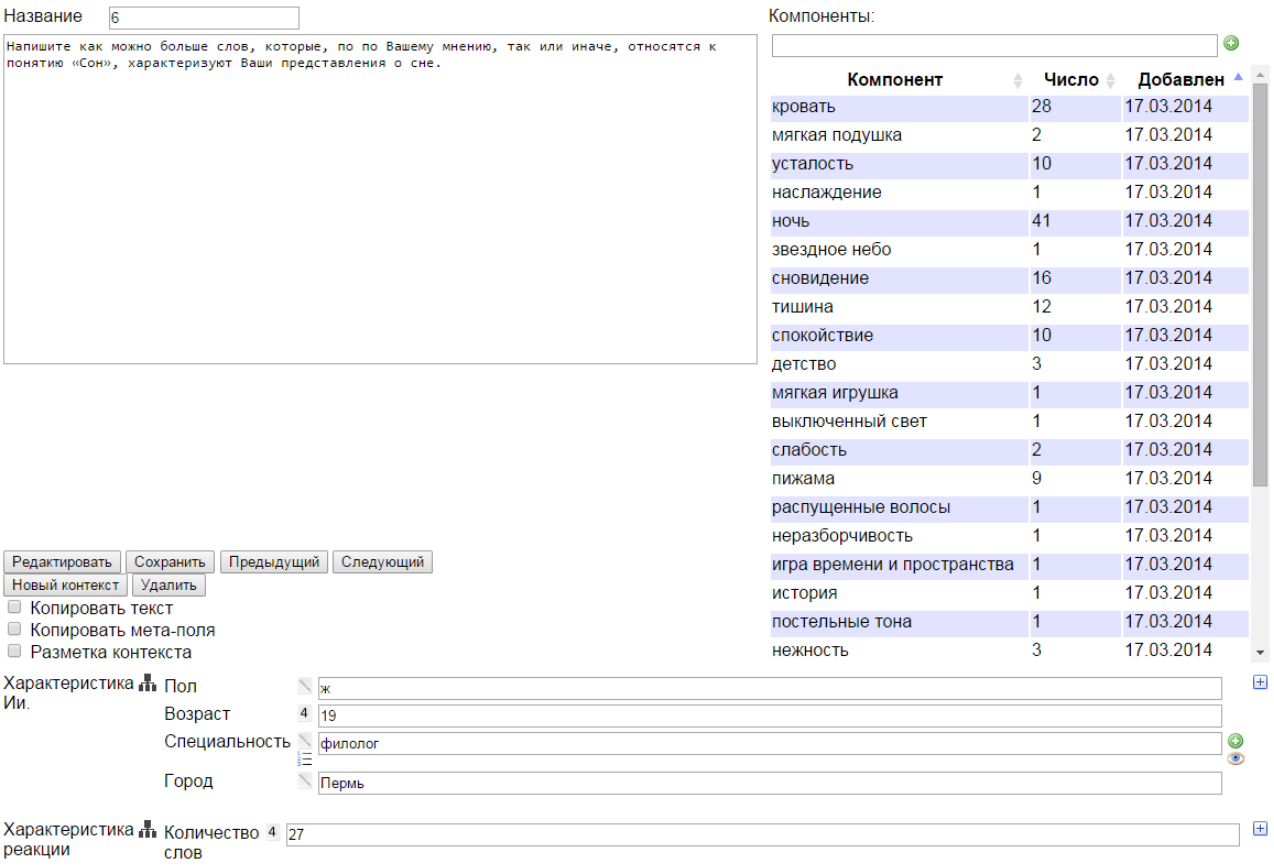

Рис. 1. Окно редактирования контекста
(реакции Ии. 6 в направленном цепочечном ассоциативном эксперименте)

В поле КОНТЕКСТ отображено задание для информанта в эксперименте. В поле КОМПОНЕНТ – реакции информанта. Поле ЧИСЛО отображает общее количество одних и тех же реакций во всем проекте. Поле ДОБАВЛЕН представляет время (в окне отображается только дата, но есть возможность вывести время в формате с точностью до секунды) добавления реакции. Полученные реакции можно сортировать по значению каждого поля (КОМПОНЕНТ, ЧИСЛО или ДОБАВЛЕН). Время добавления важно как для идентификации значения конкретной реакции по окружающему ее контексту, так и для определения временных интервалов между реакциями одной тематической группы и при смене тематических групп.

Благодаря развитой системе прав доступа информантам в эксперименте не доступен просмотр 1) реакций других информантов, 2) количества однотипных реакций (значения в поле ЧИСЛО).

В нижней части окна расположены метаполя, характеризующие данного информанта и его реакцию. Метаполя для каждого проекта могут быть разными (ср. с рис. 2).

ИС позволяет также проводить эксперименты класса 3 «Реакция представляет собой внесение новой языковой информации в сочетании с разметкой стимульного материала».

На рис. 2 представлено окно редактирования контекста с осуществленным тематическим анализом текста И.А. Бунина «Русь». В данном эксперименте информантам предъявляется текст до 300 словоформ и ставятся задачи: 1) прочитать текст, определить его тему; 2) выделить микротемы текста; 3) распределить слова текста по выделенным микротемам. Количество групп и слов в группах произвольно (подробнее об эксперименте см.: [Белоусов 2009: 136–150]).

В поле КОНТЕКСТ приведен текст. В поле КОМПОНЕНТ представлены микротемы (сказка, Вера в Бога, цвета родины и др.), выделенные информантом, зарегистрированным в ИС под логином thema, к которым привязаны слова/фрагменты из самого текста (в данном эксперименте используются другие возможности ИС – индексация текста и привязка проиндексированных слов к компонентам, в данном случае – к микротемам).

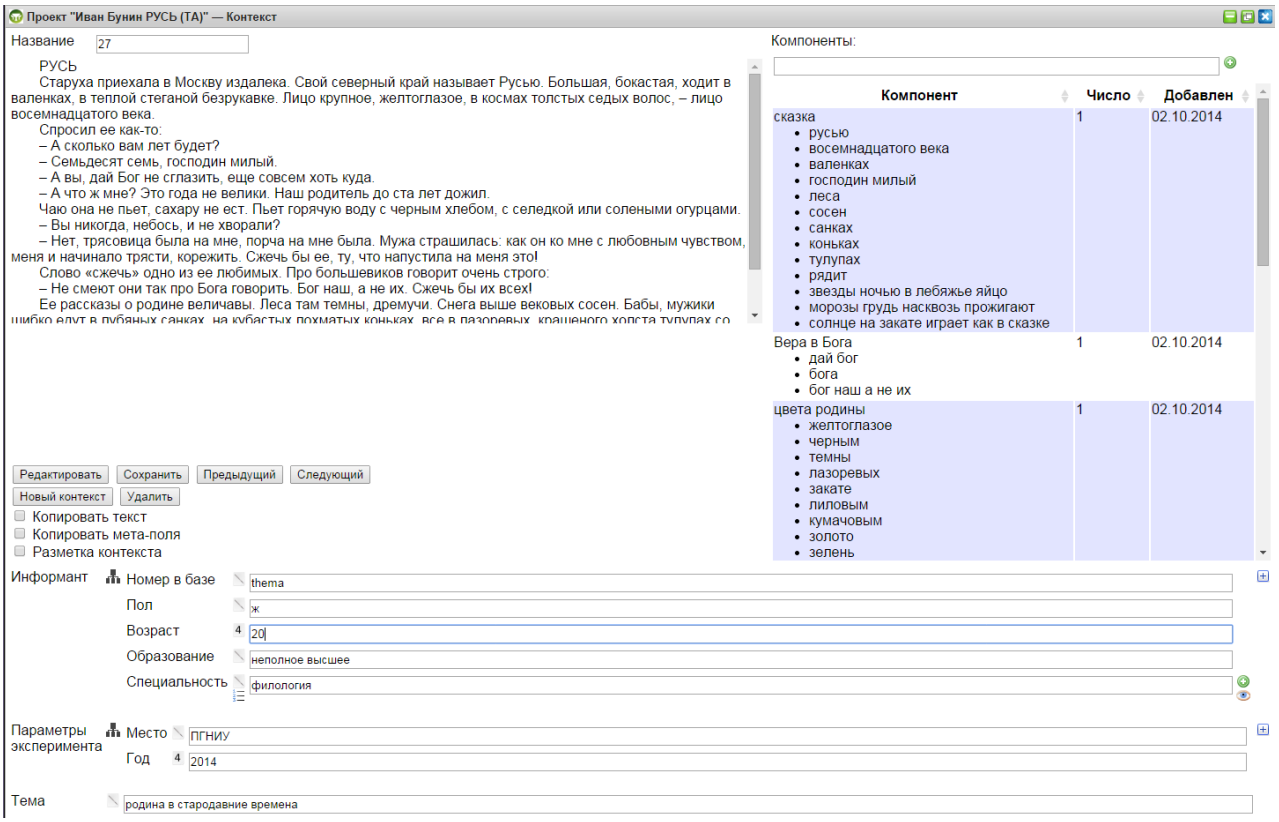

Рис. 2. Окно редактирования контекста (тематический анализ Ии. thema текста И.А. Бунина «Русь»)

Структура метаполей в данном проекте отличается от той структуры, которая приведена на рис. 1.

Из остальных классов лингвистических экспериментов на сегодняшний день доступны отдельные возможности проведения экспериментов класса 4 «Реакция представляет собой результат структурирования/группировки языковых единиц» и класса 2 «Реакция представляет собой редактирование стимульного материала».

5.2. Некоторые возможности обработки данных в ИС «Семограф»

В ИС «Семограф» обработка и анализ осуществляются с использованием двух типов единиц: 1) семантических полей (далее – С-полей), 2) метаполей. Отдельно можно рассматривать временные параметры работы информанта/эксперта.

Так как основные единицы модели – это метаполя и С-поля, то моделируемые отношения между ними могут быть следующие

Отношения типа «С-поле : С-поле». Этот тип отношений позволяет генерировать семантическую карту (далее – С-карту) и семантический граф (далее – С-граф). С-карта отражает совместное присутствие двух С-полей в одном и том же контексте с учетом подобной встречаемости во всех контекстах выборки или корпуса. Полагается, что если две единицы (например, слова) присутствуют в одной реакции информанта, то их можно считать связанными между собой. Поэтому можно сделать вывод и о связи между С-полями, в которые входят указанные компоненты. С-карта автоматически генерируется на основе подсчета количества связей между полями в пределах всей выборки (С-карта может быть построена не только между С-полями, но и на более низком уровне – уровне компонентов, например конкретных слов в реакциях информантов). С-граф представляет собой графическую экспликацию связей между выделенными С-полями в С-карте.

Фреймворк ИС позволяет генерировать С-карту и С-граф как на материале всего корпуса экспериментальных реакций, так и на материале отдельных выборок, созданных в результате фильтрации значений метаполей (например, по полу, образованию, возрасту, городу и др. характеристикам, взятым как отдельно друг от друга, так и в любых комбинациях).

Отношения типа «Метаполе : Метаполе». Данный тип отношений позволяет выявлять распределение значений одних типов метаданных по значениям других типов метаданных. Например, рассматривать параметры экспериментальной выборки в контексте «пол – возраст» (см. табл. 1).

Таблица 1

Параметры соответствия значений метаполей «пол» и «возраст» в экспериментальной выборке (на материале проведенного эксперимента)

Пол	Возраст																								
	18	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**40**	**41**	**42**	**49**	**50**	**53**	**54**	**55**
Ж	1	16	12	3	1	0	0	0	1	3	1	2	1	2	3	1	2	3	0	2	1	2	1	2	1
М	1	2	3	1	1	2	2	2	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0

Отношения типа «С-поля : Метаполя». Результатом анализа данных типов отношений становится распределение С-полей по значениям тех или иных метаполей, например, значениям пола, возраста, образования, города проживания информантов и т.д. Данные распределения передаются таблицей соответствий и в дальнейшем могут использоваться, в частности, при применении метода анализа соответствий [Боровиков 2003].

Временные параметры работы информанта/эксперта над материалом. В таблице 2 приведен фрагмент статистики пользовательских действий информанта astafuque в рамках тематического анализа текста И.А. Бунина «Русь» (о возможности временного анализа в ИС «Семограф» см. также: [Белоусов 2013]).

Таблица 2

Статистика пользовательских действий (фрагмент)

Пользователь	Действие	Объект	Контекст	Интервал	Микротема	Слова	Порядок слова
astafuque	save	contextHasComponent	19	7	Описание старухи		
astafuque	save	collocation	19	9	Описание старухи	большая	12
astafuque	save	collocation	19	3	Описание старухи	бокастая	13
astafuque	save	collocation	19	18	Описание старухи	ходит в валенках	14 15 16
astafuque	save	collocation	19	5	Описание старухи	теплой стеганой в	
astafuque	remove	collocation	19	7	Описание старухи	теплой в стеганой	
astafuque	save	collocation	19	13	Описание старухи	лицо крупное желтоглазое	21 22 23

В первом столбце отображается имя того, кто произвел действие (имя информанта/эксперта); во втором – совершенное действие (добавление или удаление категории); в третьем – объект, над которым было совершено действие; в четвертом – номер контекста (каждому информанту соответствует свой контекст); в пятом – временной интервал, характеризующий совершенное действие; в шестом – название микротемы; в седьмом – приписанные к микротеме слова; в восьмом – порядок (номер) слова в тексте.

На следующих этапах анализа сгенерированную таблицу можно использовать в статистических пакетах программ для выявления закономерностей, характеризующих деятельность информантов/экспертов.

5.3. Генерация С-карты и С-графа

На примере тематического анализа текста И.А. Бунина «Русь» покажем некоторые возможности работы с С-картой и С-графом.

Слова, находящиеся в одной микротеме, связаны между собой. Исходя из того что разные интерпретации порождают разные сценарии анализа и синтеза текста, каждое слово может стать компонентом довольно широкого поля микротем. **Сила семантической связи** между двумя словами в тексте показывает, насколько часто в реакциях информантов два слова данного текста являются компонентами одной микротемы. Семантическая связность каждого слова с остальными словами текста, рассмотренная в контексте всех микротем, отображается с помощью

С-карты. Очевидно, что связь между словами может иметь либо закономерный, либо случайный характер. Семантическая связь становится закономерной в том случае, когда два слова частотно включаются информантами в одну микротему. Порог значимости может быть установлен статистически. В таблице 3 представлен фрагмент С-карты текста И.А. Бунина «Русь» (вся таблица имеет размер 169×169).

Таблица 3

Фрагмент С-карты текста И.А. Бунина «Русь»

	космах	небось	в	шибко	едут	лубяных	кубастых	аршинными	кумачовым	лебяжье	яйцо
космах	—	2	17	1	1	1	1	1	1	2	1
небось	2	—	2	1	1	1	1	1	1	1	1
в	17	2	—	1	1	1	1	1	1	1	1
шибко	1	1	1	—	20	19	13	16	9	11	11
едут	1	1	1	20	—	19	13	15	8	11	11
лубяных	1	1	1	19	19	—	19	17	9	14	14
кубастых	1	1	1	13	13	19	—	15	9	12	12
аршинными	1	1	1	16	15	17	15	—	9	12	12
кумачовым	1	1	1	9	8	9	9	9	—	13	13
лебяжье	2	1	1	11	11	14	12	12	13	—	30
яйцо	1	1	1	11	11	14	12	12	13	30	—

Примечание. На пересечении *i*-ой строки и *k*-го столбца располагается ячейка, в которой отмечается сила семантической связи между *i*-ым и *k*-ым словами. Например, в ячейке, находящейся на пересечении строки ШИБКО и столбца ЛУБЯНЫХ зафиксировано значение 19. Это означает, что данные слова были отнесены информантами в одну микротему 19 раз (из 206 возможных микротем, выделенных 34 информантами).

Визуализация С-карты осуществляется в виде С-графа. На рис. 3 представлен С-граф текста И.А. Бунина «Русь» (для построения графа использовалось свободно распространяемое программное средство Gephi – http://gephi.org).

Плотная семантическая сеть визуально представляет действия механизмов глобальной семантической связности текста («...нет такого компонента который бы не был связан хотя бы с одним другим компонентом текста...» [Мурзин, Штерн 1991: 11]). В то же время, не все связи между словами являются значимыми, поэтому требуется введение критериев «порога значимости» (вариантов обнаружения «порога значимости» может быть несколько, в частности, превышение показателя частотности семантической связи 0,05; плотности связей графа и др.). На рис. 4 показан С-граф анализируемого текста с ребрами (семантическими связями), превышающими показатель 0,1. Значения «порога значимости» могут варьироваться для установления оптимального соотношения количества вершин и ребер графа.

Заметим, что С-графы, как и С-карты, могут генерироваться на материале не только корпуса экспериментальных реакций, но и отдельных его выборок, создаваемых с помощью инструментов множественной и вложенной фильтрации значений метаданных (например, женщины; женщины 30–40 лет; женщины 30–40 лет с высшим образованием и мн. др.). Данный инструментарий позволяет сопоставлять семантические структуры и делать выводы о влиянии на них социальных (и других) параметров. На следующем этапе осуществляется интерпретация С-графа.

С-карта и С-граф, представленные в таблице 3 и на рис. 3, строятся из низкоуровневых единиц – слов, взятых непосредственно из реакций информантов. В то же время ИС позволяет генерировать С-карту и С-граф из высокоуровневых единиц – С-полей, состоящих их низкоуровневых единиц и являющихся результатом их классификации.

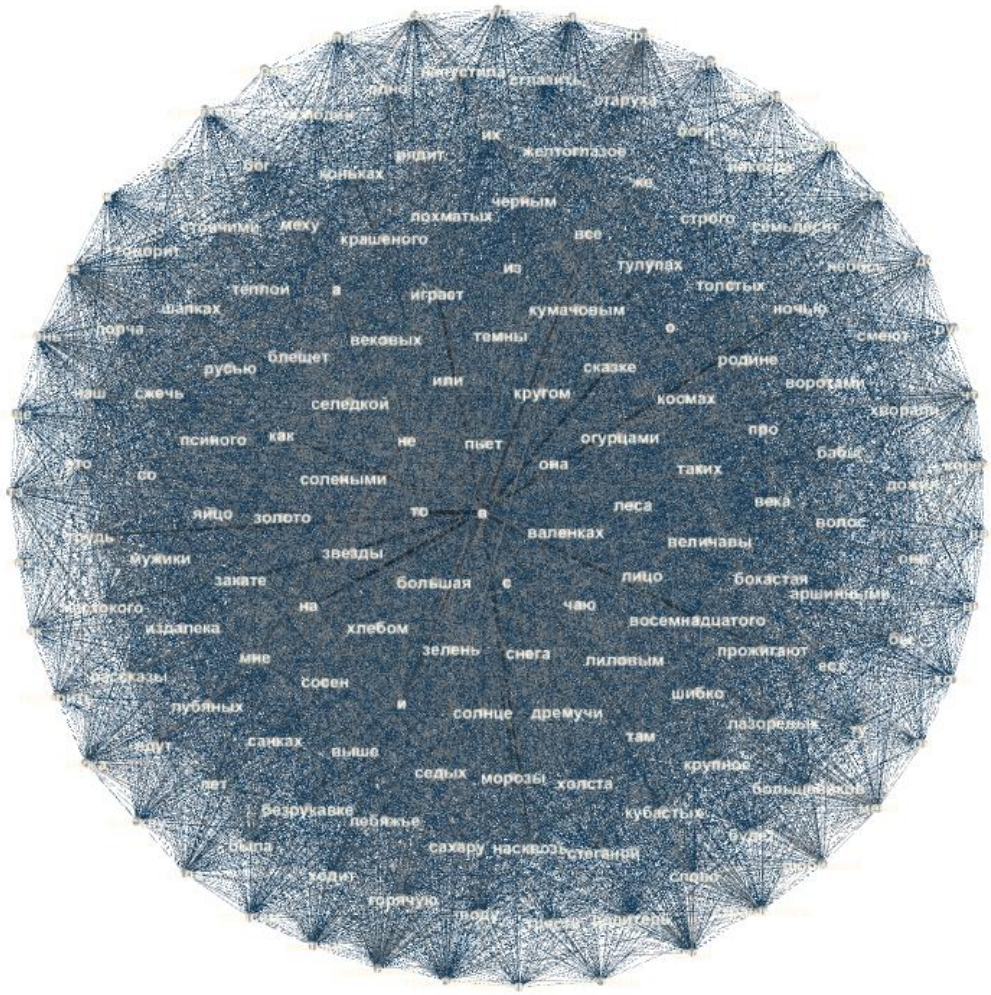

Рис. 3. Полный С-граф текста И.А. Бунина «Русь»

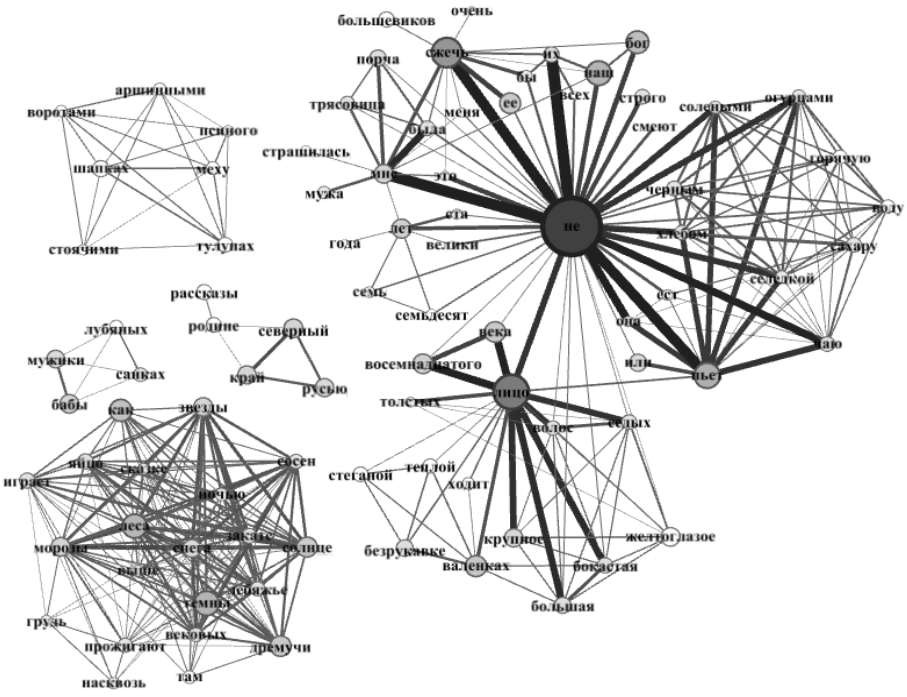

Рис. 4. С-граф текста И.А. Бунина «Русь»

Примечание. Размер вершины пропорционален частотности слова, а толщина ребра – силе связи между словами (частоте совместного присутствия двух слов в микротемах).

6. Заключение

Дальнейшее развитие экспериментальной лингвистики будет связано с разработкой и внедрением в исследовательский процесс современных технологий сбора и анализа данных в рамках имеющихся возможностей сетевой науки. Полагаем, что сетевые технологии позволят не только применить новые стандарты качества к научным результатам, но и создадут основу для их коммерческого использования в областях искусственного интеллекта, поисковых технологий, автоматического перевода, а также в маркетинге и в сфере управления и принятия решений.

Примечания

[1] Исследование выполнялось при финансовой поддержке Российского гуманитарного научного фонда (проект № 12-34-01087).

[2] Мы не рассматриваем аппаратурные эксперименты (например, методы регистрации движений глаз; психофизиологической активности и т.п.), которые требуют других информационно-технологических решений.

Список литературы

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

Белоусов К.И. Временные модели когнитивной деятельности (на материале экспертного анализа текстового контента) // Вестник Пермского университета. Российская и зарубежная филология. 2013. Вып. 4(24). С. 72–77.

Белоусов К.И., Ичкинеева Д.А. Графосемантическое моделирование структурного синтеза тематического пространства текста // Филология и человек. 2011. №1. С. 26–38.

Белянин В.П. Экспериментальное исследование психолингвистических закономерностей смыслового восприятия текста: дис. … канд. филол. наук. М., 1983. 271 с.

Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. СПб.: Питер, 2003. 688 с.

Залевская А.А. Значение слова через призму эксперимента / Твер. гос. ун-т. Тверь, 2011. 240 с.

Ерофеева Е.В., Ерофеева Т.И. Человек и текст: антропоцентрический подход к исследованию // Вестник Пермского университета. Российская и зарубежная филология. 2010. Вып. 4(10). С. 28–33.

Ерофеева Е.В., Пепеляева Е.А. Структура семантического поля «Человек» в сознании носителей русского языка // Вестник Пермского университета. Российская и зарубежная филология. 2011. Вып. 1(13). С. 7–19.

Ичкинеева Д.А. Аналитическая и синтетическая стратегии членения семантического пространства текста // Филология и человек. 2011. № 4. С. 171–178.

Москальская О.И. Грамматика текста. М.: Высш. шк., 1981. 183 с.

Мурзин Л.Н., Штерн А.С. Текст и его восприятие. Свердловск: Изд-во Урал. ун-та, 1991. 172 с.

Пищальникова В.А. Психопоэтика. Барнаул: Изд-во Алт. ун-та, 1999. 176 с.

Рубакин Н.А. Психология читателя и книги: краткое введение в библиологическую психологию. М.: Книга, 1977. 264 с.

Русский ассоциативный словарь [Электронный ресурс]. URL: http://tesaurus.ru/dict/dict.php (дата обращения: 25.10.2014).

Сахарный Л.В. Введение в психолингвистику: курс лекций. Л.: Изд-во Ленингр. ун-та, 1989. 184 с.

Сорокин Ю.А. Психолингвистические аспекты изучения текста. М.: Наука, 1985. 168 с.

Ярхо Б.И. Методология точного литературоведения: избр. тр. по теории литературы. М.: Языки слав. культур, 2006. xxxii, 927 с.

Miller G.A. Some psychological studies of grammar // American Psychologist. 1962. V. 17. P. 748–762.

Smith G. Tagging: People-powered Metadata for the Social Web. Berkeley: New Riders, 2008. 216 p.

EXPERIMENTAL LINGUISTICS AND NETWORK SCIENCE

Konstantin I. Belousov
Professor of Theoretical and Applied Linguistics Department
Perm State University

In article the current state of experimental linguistics and its main objectives in the context of use of information technologies is considered. Versions of the solution of the following problems are discussed: representativeness of data; classifications of linguistic experiments; collecting, processing, analysis and data storage; methodological aggression and pluralism, etc. Opportunities and prospects of the experimental researches organization by means of network science technologies are shown.

Key words: experimental linguistics; network science; information system «Semograph»; information technologies in linguistics; classification of linguistic experiments; activity model; online models.