ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»»

На правах рукописи

Скоринкин Даниил Андреевич

СЕМАНТИЧЕСКАЯ РАЗМЕТКА ХУДОЖЕСТВЕННЫХ ТЕКСТОВ ДЛЯ КОЛИЧЕСТВЕННЫХ ИССЛЕДОВАНИЙ В ФИЛОЛОГИИ (НА ПРИМЕРЕ РОМАНА «ВОЙНА И МИР» Л. Н. ТОЛСТОГО)

Резюме
диссертации на соискание ученой степени
кандидата филологических наук НИУ ВШЭ

Научный руководитель
кандидат филологических наук
А. А. Бонч-Осмоловская

Москва 2018

Общая характеристика исследования

Анализ художественных произведений с применением компьютерного моделирования и количественных методов — актуальное направление современных филологических исследований. Растущая доступность текстов в цифровой форме и совершенствование методов анализа текстовых данных открывают новые возможности для изучения литературного наследия. Современные инструменты лингвистического анализа позволяют автоматически извлекать не только грамматические, но и некоторые семантические и прагматические свойства текста.

Вместе с тем анализ современных работ в области количественного литературоведения указывает на наличие нерешенных проблем, которые тормозят развитие этого перспективного направления. В частности, большой сложностью остается извлечение из текста чистых структурированных данных, необходимых для количественного анализа и компьютерного моделирования. В данном случае можно говорить о том, что разные стороны художественного творчества имеют разную степень «доступности» для компьютерного моделирования и анализа. Так, некоторые аспекты стиля художественного текста могут быть выражены через частотности слов, получение которых является сегодня тривиальной задачей. В связи с этим активно развивается компьютерная стилистика и стилеметрия. В то же время автоматическое извлечение и масштабный количественный анализ, например, сюжетных мотивов средствами современного автоматического анализа текстов, по-видимому, на сегодняшний день невозможны.

Персонаж художественного произведения занимает на шкале доступности для компьютерного моделирования и анализа промежуточное положение. С одной стороны, образ персонажа порождается вполне конкретной последовательностью упоминаний в тексте, которую можно отграничить и проанализировать. Текстовые вхождения персонажа могут быть различными: варианты имени (Наташа, Натали, Наталья Ильинична), титулы и гоноративы (графиня, сударыня), анафорические упоминания (она, он), обращения в речи других персонажей (любезный, вы), ролевые и ситуативные характеристики (молодой человек, проезжающий, раненый) и др. Однако все эти вхождения имеют очевидные границы и конкретные лингвистические признаки (имя собственное, личное местоимение, гоноратив) — и поэтому оказываются доступными для выделения существующими средствами компьютерной лингвистики.

С другой стороны, точное извлечение из художественного текста всех упоминаний конкретного персонажа и связанных с ним признаков (прямой речи, совершаемых действий, взаимодействий с другими персонажами) является сложной задачей. Проблемы связаны с упомянутой выше вариативностью наименования персонажа, наличием

местоименной анафоры, сложностью автоматического определения авторства реплики. Характерно, что многие интересные и масштабные опыты по моделированию персонажей сегодня производятся на материале драматических текстов, где значительная часть указанных проблем отсутствует — реплики, как правило, обозначены явно через имя персонажа, анафора практически не используется.

Решением проблемы представляется переход к анализу системы персонажей художественных текстов с опорой на предварительно подготовленную **стандартизированную семантическую разметку**. Такая разметка позволяет хранить дополнительный смысловой слой (например, все упоминания одного персонажа в виде одной цепочки вне зависимости от конкретного выражения в тексте, или все вхождения прямой речи персонажа) без отрыва от исходного текстового материала. Этот слой является машиночитаемым, то есть может быть автоматически считан из документа программой и переведен в однозначные структурированные данные (например, таблицу всех вхождений прямой речи с однозначной припиской каждой реплики к идентификатору произносящего ее персонажа). Это позволяет производить количественный анализ системы персонажей без сложной предварительной подготовки данных. Однако сегодня, несмотря на появление работ, специально направленных на моделирование персонажа и системы персонажей, данные в большинстве случаев извлекаются напрямую из текста, с неизбежными при этом ошибками, без возможности воспроизведения результата и уточнения разметки, а обсуждение разметки не производится.

Диссертационное исследование посвящено созданию семантической разметки для книги «Война и мир» Л. Н. Толстого и апробации этой разметки для моделирования и количественного анализа системы персонажей произведения. **Актуальность** диссертационного исследования состоит в разработке инструментария автоматизированной разметки с использованием современных методов компьютерной лингвистики и демонстрации его применения для количественного исследования. В работе была предпринята попытка преодолеть разрыв между возможностями средств автоматического анализа языка — и задачами филологического исследования художественного произведения.

Объектом исследования является текстовая репрезентация системы персонажей в прозаическом художественном произведении крупной формы. **Предметом** исследования выступают параметрические характеристики персонажей, извлекаемые из текста с опорой на семантическую разметку. Выбор объекта и предмета принципиально отличает диссертацию от упомянутых выше стилеметрических работ. Выбирая в качестве объекта систему персонажей, мы стремились приложить компьютерно-лингвистические

инструменты к анализу одного из компонентов *содержания* художественного произведения, его *сюжета*, а не *стиля*.

Целью диссертации была поставлена разработка и апробация метода анализа системы персонажей художественного произведения с опорой на семантическую разметку текста. Для достижения указанной цели были решены следующие **задачи**:

1. Исследованы теоретические аспекты моделирования системы персонажей художественного произведения, произведен анализ существующих работ.
2. Осуществлена автоматическая разметка упоминаний персонажей в тексте книги «Война и мир». Выделенные вхождения связаны в единые кореферентные цепочки, соответствующие одному конкретному персонажу, при помощи уникального идентификатора персонажа.
3. Осуществлена полуавтоматическая разметка прямой речи персонажей в тексте книги «Война и мир».
4. На основе подготовленной разметки:
    a. Осуществлен статистический анализ прямой речи персонажей с использованием двух различных методов, произведено сравнение методов.
    b. Построены сети взаимодействия персонажей с использованием двух различных методов, произведено сравнение методов.

Новизна работы заключается, во-первых, в сравнении различных методов анализа на едином материале. Такой подход, будучи стандартным в лингвистике, до сих пор практически не применялся в количественном литературоведении. В частности, в диссертации были сопоставлены два метода анализа прямой речи персонажей (на основе лексического состава реплик и нелексических характеристик) и два метода сетевого анализа (на основе диалоговых взаимодействий и совместной встречаемости в тексте). Во-вторых, в адаптации методов современной компьютерной лингвистики (извлечение именованных сущностей, разрешение анафоры, извлечение событий) к исследованию художественного произведения на русском языке. В-третьих, в научный оборот введены новые количественные характеристики персонажей (интенсивность взаимодействия, параметрические характеристики произносимых реплик).

Теоретическая значимость диссертации состоит в сравнении различных методов количественного анализа и моделирования системы персонажей художественного произведения на открытом доступном для других исследователей материале (семантически размеченном тексте книги «Война и мир»). Результаты сравнения позволяют говорить о том, какие именно особенности и соотношения в системе персонажей высвечивает тот или

иной метод, и какие свойства он не фиксирует. Показаны ограничения ряда методов, которые ранее не учитывались или не проговаривались исследователями.

На защиту выносятся **следующие положения**:

1. Современные средства автоматической обработки текста могут использоваться для извлечения и структурирования значимой информации о системе персонажей художественного произведения в форме семантической разметки.
2. Подготовленная разметка позволяет осуществлять анализ системы персонажей с применением количественных методов (анализ частотностей, многофакторный статистический анализ, корреляционный анализ, сетевой анализ).
3. Выбор конкретного метода анализа данных, получаемых из разметки, влияет на то, какие именно свойства системы персонажей будут отражены в полученной модели. Сети персонажей, построенные на основе диалогового взаимодействия, отличаются от тех, что были получены при помощи метода совместной встречаемости.

Практическая значимость работы заключается, во-первых, в создании семантической разметки книги «Война и мир». Разметка опубликована и доступна для использования другими исследователями, в т.ч. за рубежом. Так как разметка сделана на основе международного формата кодирования текстов TEI, она дает возможность производить подсчеты и манипуляции с семантическими элементами толстовского текста (персонажами, фактами речевой активности) даже без знания русского языка. Разметка содержит идентифицированные упоминания персонажей, в том числе анафорические, и реплики прямой речи персонажей с однозначным указанием адресанта и адресата. Во-вторых, подготовленные визуализации сетевой структуры персонажей могут использоваться в педагогическом процессе. Материалы диссертации были использованы в рамках образовательной программы Лицея НИУ ВШЭ (2017/2018 уч. г.), в курсе по цифровым методам в гуманитарных науках в Университете Хельсинки (2018 г.), в лекциях на школах Центра цифровых гуманитарных исследований НИУ ВШЭ (2016–2018 гг.).

Апробация работы: результаты исследования были представлены на международной конференции молодых филологов в Тарту (Тарту, 26 апреля 2015 и 1 мая 2017), международной конференции Digital Humanities 2015 — Annual Conference of the Alliance of Digital Humanities Organizations (Сидней, 3 июля 2015), международной конференции Digital Humanities 2016 — Annual Conference of the Alliance of Digital Humanities Organizations (Краков, 13 июля 2016), международной конференции TEI Conference and Members' Meeting 2016 (Вена, 30 сентября 2016), международной конференции 6th AIUCD Conference 2017 (Рим, 24 января 2017), международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог» (Москва, 30 мая 2015 и 1 июня

2017 г.), международной научно-практической конференции «Информационные технологии в гуманитарных науках» (Красноярск, 20 сентября 2017 г.), всероссийской конференции «Естественнонаучные методы в цифровой гуманитарной среде» (Пермь, 17 мая 2018 г.). В рамках диссертационного исследования подготовлены 7 публикаций, в том числе в изданиях, индексируемых Scopus/Web of Science — 2 публикации, в изданиях списка ВАК (кроме Scopus/Web of Science) — 2 публикации.

Работа включает в себя введение, 3 главы, заключение, библиографию и приложение.

Основное содержание работы

Глава 1 посвящена истории и теории построения модели персонажа художественного произведения, а также описанию модели разметки, осуществленной в работе. В первом разделе главы описываются подходы к анализу и формализации понятия персонажа, созданные в докомпьютерную эпоху. В этом разделе показано, что рассмотрение художественного персонажа как типового объекта с набором параметров (т.е. модели) возникло уже в античности — примерами являются «Поэтика» Аристотеля и «Характеры» Теофраста. У Горация как особенно важный параметр выделена речь персонажа и ее соответствие роли в произведении. Нормативно-дидактические рекомендации античных классиков служили ориентиром для позднейших литературных течений, ориентированных на античные образцы, в первую очередь для классицизма XVII–XVIII вв. В XIX веке параллельно со становлением теории литературы происходит постепенный переход от дидактизма к анализу техники и средств конструирования персонажа. В начале XX века в анализе художественного персонажа выделились два основных направления. Одно из них, которое можно условно обозначить как «формальное», рассматривало персонажа как некоторый авторский прием или инструмент. При этом отрицался психологизм персонажа. Персонаж рассматривался в утилитарном ключе, как средство группировки мотивов в тексте. Для формальной школы было характерно понимание персонажа как накапливающейся совокупности приемов, динамической переменной в тексте. В дальнейшем эта идея была унаследована структурализмом, который редуцировал персонажа до цепочки фрагментов текста. Для Р. Барта персонаж — набор символов, объединенных именем, компактная переменная для хранения смыслов, цепочка фрагментов текста. Хотя в дальнейшем структуралисты отошли от такого радикального понимания персонажа, их идеи оказались созвучны современным практикам компьютерных исследований художественного произведения, где предполагается именно выделение в тексте фрагментов на основе имени и исследование системы персонажей с опорой на эти

текстовые вхождения. Другие идеи структурализма и семиотической школы, которые оказались актуальны для цифровых исследований литературы, — это актантные модели А. Ж. Греймаса и Ц. Тодорова, а также понимание персонажа как «пересечения структурных функций» с набором («парадигмой») дифференциальных признаков, предложенное Ю. М. Лотманом.

Параллельно складывался альтернативный подход к анализу и построению модели персонажа. В нем персонаж рассматривался как более или менее полноценная модель личности, созданная писателем и имеющая связь с личностью писателя. Так, в концепции М. М. Бахтина персонаж осознавался как полноценный мыслящий и нравственный субъект, носитель сознания, чувств, желаний. Также у Бахтина обнаруживается идея возникающей «автономности» персонажа. Позже с близких психологических позиций рассматривали персонажа многие советские литературоведы.

Удачно соединить формалистские и структуралистские взгляды на героя как на последовательность упоминаний с психологическими подходами удалось Л. Я. Гинзбург в работе «О литературном герое» [Гинзбург, 1979]. С одной стороны, Л. Я. Гинзбург — вполне в духе кумулятивных интерпретаций формалистов — определяла героя как «серию последовательных явлений одного лица в пределах данного текста» [Гинзбург, 1979: с. 87]. С другой стороны, понимание персонажа у Л. Я. Гинзбург не сводилось к формалистской редукции: «литературный герой моделирует человека». С психологическими подходами концепцию Л. Я. Гинзбург также роднит внимание к прямой речи персонажей, ее использованию писателями как наиболее непосредственного приема психологической характеристики персонажа.

Второй раздел диссертации посвящен практикам компьютерного моделирования свойств персонажей и системы персонажей. В этом разделе показано, как идеи докомпьютерного литературоведения актуализируются в современных практиках компьютерного анализа персонажа (системы персонажей). В компьютерных исследованиях выделяются два подхода: моделирование и анализ персонажей на основе лингвистических параметров их прямой речи и сетевой анализ.

Работ, специально посвященных компьютерному моделированию системы персонажей, сравнительно мало. Их анализ показал, что для исследования прямой речи персонажей используются стилеметрические подходы, т.е. сравниваются частотность и дистрибуция словоформ. Такой подход позволяет выделить группы похожих и группы противопоставленных друг другу персонажей, а также выявить лексические пласты, являющиеся значимыми характеристиками создаваемых речевых образов. Основной проблемой является выделение материала для анализа из основного текста, т.е.

разграничение речи и нарратива. В большинстве случаев исследователи не публикуют разметку, что затрудняет воспроизведение исследований. Создание разметки также практически не обсуждается — исключением является лишь работа, посвященная более простым с точки зрения компьютерной обработки текстам драматических произведений.

Сетевой анализ как способ компьютерного моделирования системы персонажей имеет сравнительно бо́льшее распространение. При этом анализ работ, посвященных сетевому анализу художественных произведений, позволяет проследить эволюцию этого метода от технического эксперимента к полноценному инструменту формального филологического исследования. Одновременно происходит формирование стандарта работы с данными: обоснование метода выделения отношения между персонажами и публикация данных в виде разметки, которая дает возможность другим заинтересованным исследователям воспроизвести построение сети и повторить анализ. В последние годы появляются работы, опирающиеся на стандартизированную семантическую разметку художественного текста. Однако, как и в случае с исследованием речи персонажей, разметке производится в первую очередь для драматических текстов, тогда как для сетевого анализа прозаических текстов применяются невоспроизводимые «одноразовые» эвристики.

Третий раздел главы 1 описывает осуществленную в работе семантическую разметку текста книги «Война и мир». Разметка выполнена в соответствии с международным стандартом машиночитаемой семантической и структурной разметки текстов TEI. Этот стандарт широко используется в гуманитарных исследованиях за рубежом, масштабно применяется для сохранения культурного наследия. Средствами TEI были размечены в машиночитаемом виде два слоя текста, которые необходимы для моделирования системы персонажей при помощи предлагаемого нами метода: непосредственные наименования персонажей в тексте (в виде имен собственных, нарицательных именных групп, а также личных и возвратных местоимений) и прямая речь персонажей. Для разметки применялась система извлечения именованных сущностей из текста ABBYY Compreno, собранные вручную списки имен и дополнительные эвристики. По результатам проведенной оценки на случайной выборке из 50 абзацев итоговая точность идентификации именованных персонажей составила 94%, итоговая полнота — 67%, итоговая F-мера (гармоническое среднее точности и полноты) — 78,2%. Таким образом, в разметке есть неидентифицированные упоминания персонажей, но достаточно мало неверно идентифицированных упоминаний, что делает ее пригодной для дальнейшего исследования. Наличие пропущенных упоминаний, по нашим наблюдениям, не критично, т.к. это, как правило, рядом с пропущенным находится несколько выявленных упоминаний того же героя. Используемые в диссертации методы

количественного анализа в таком случае учитывали упоминание персонажа, и результат не искажался.

Глава 2 посвящена исследованию системы персонажей «Войны и мира» через количественный анализ прямой речи с опорой на подготовленную семантическую разметку. При этом использовались и сравнивались друг с другом два различных метода анализа прямой речи.

Первый метод — Delta, [Burrows, 2002] — относится к методам стилеметрии, используемым для статистической атрибуции авторства. Метод опирается на распределение частотностей наиболее частотных слов (лемм) в речи персонажей. Размер списка слов устанавливался в ходе эксперимента, принципиально схожего с опытами по атрибуции авторства. Речь, принадлежащая одному персонажу, рассматривалась как аналог корпуса произведений одного автора. Реплики каждого персонажа были случайным образом разделены на две сопоставимые по размеру коллекции — обучающую и тестовую. Далее производилась классификация документов в тестовой коллекции с использованием списка лемм различной длины. Наиболее успешная классификация была осуществлена на основе 130 наиболее частотных слов. В связи с этим именно список из 130 наиболее частотных лемм использовался для построения системы персонажей произведения Л. Н. Толстого при помощи стилеметрии.

Ограничением стилеметрического метода является чувствительность к размеру текста. Проанализировав работы по стилеметрии, мы определили нижнюю границу размера корпуса речи каждого персонажа как 1000 словоупотреблений. В этой части исследования участвовали только те персонажи, чей совокупный объем прямой речи превышает 1000 словоупотреблений. Подсчет на основе подготовленной разметки показал, что таких персонажей 16, два из них — Наполеон и Рамбаль — были исключены, т.к. их речь передана преимущественно на французском языке. В результате в эксперименте исследовалась речь следующих персонажей: Андрей Болконский, Пьер Безухов, Наташа Ростова, Николай Ростов, Марья Болконская, Василий Курагин, старый князь Николай Болконский, Денисов, граф Илья Ростов, Кутузов, графиня Наталья Ростова, Долохов, А. М. Друбецкая, А. П. Шерер

В ходе эксперимента каждый персонаж был представлен как вектор относительных частотностей слов (лемм) в его речи. Вектор для каждого персонажа содержал частотности лемм из общего для всех персонажей списка в 130 лемм. Таким образом, все исследованные персонажи были представлены как точки в 130-мерном пространстве и сгруппированы по геометрической близости в этом пространстве. Для снижения размерности и визуализации близости применялись методы снижения размерности: многомерное шкалирование и метод

главных компонент. Для отображения группировки персонажей применялась иерархическая кластеризация.

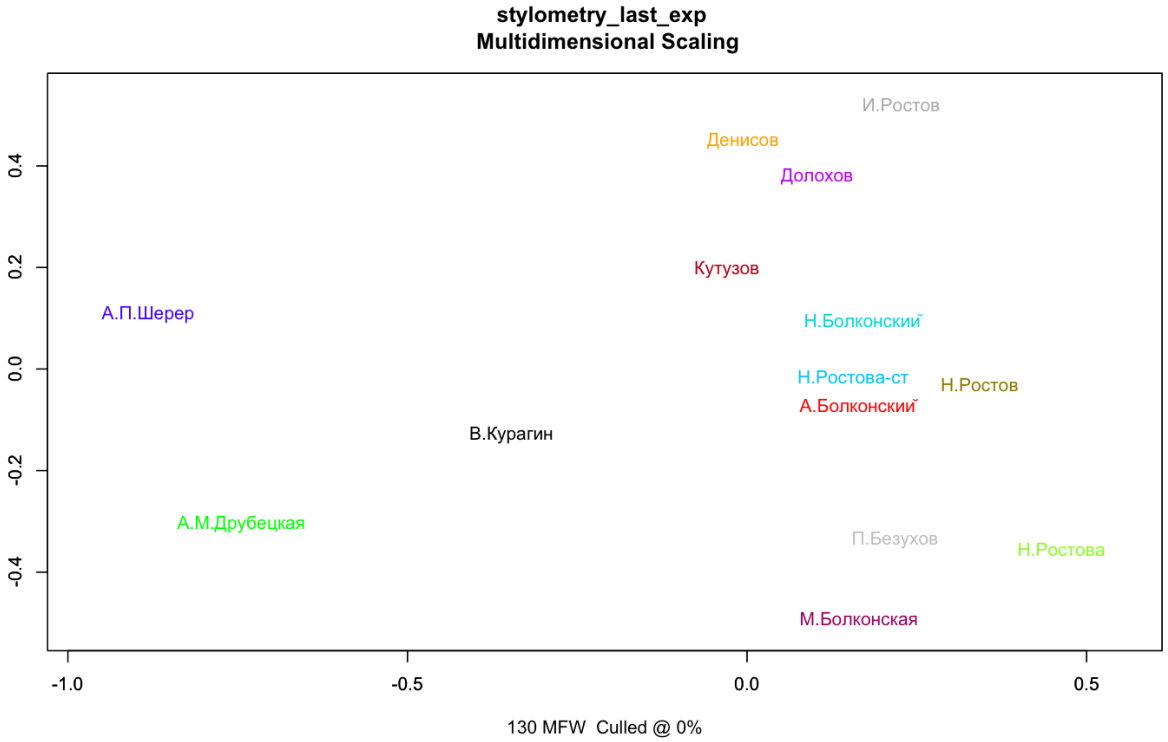

Рис. 1. Визуализация пространства 14 персонажей «Войны и мира» при помощи метода многомерного шкалирования

Наиболее явная граница в пространстве персонажей, построенном на основе стилеметрических измерений речи, проходит между группой из А. П. Шерер, А. М. Друбецкой и Василия Курагина — и всеми прочими персонажами, вошедшими в наш список. Группу выделяют все использованные методы многомерного анализа.

Оставшаяся группа из 11 персонажей распадается на две подгруппы. Одна из подгрупп объединяет главных неисторических персонажей книги: Пьера Безухова, Андрея Болконского Николая Ростова, Наташу Ростову, княжну Марью Болконскую. В другой подгруппе — в основном статичные персонажи второго плана: Денисов, князь Николай Андреевич Болконский, Кутузов, граф Илья Ростов графиня Наталья Ростова (обозначена как «Н.Ростова-ст»), Долохов.

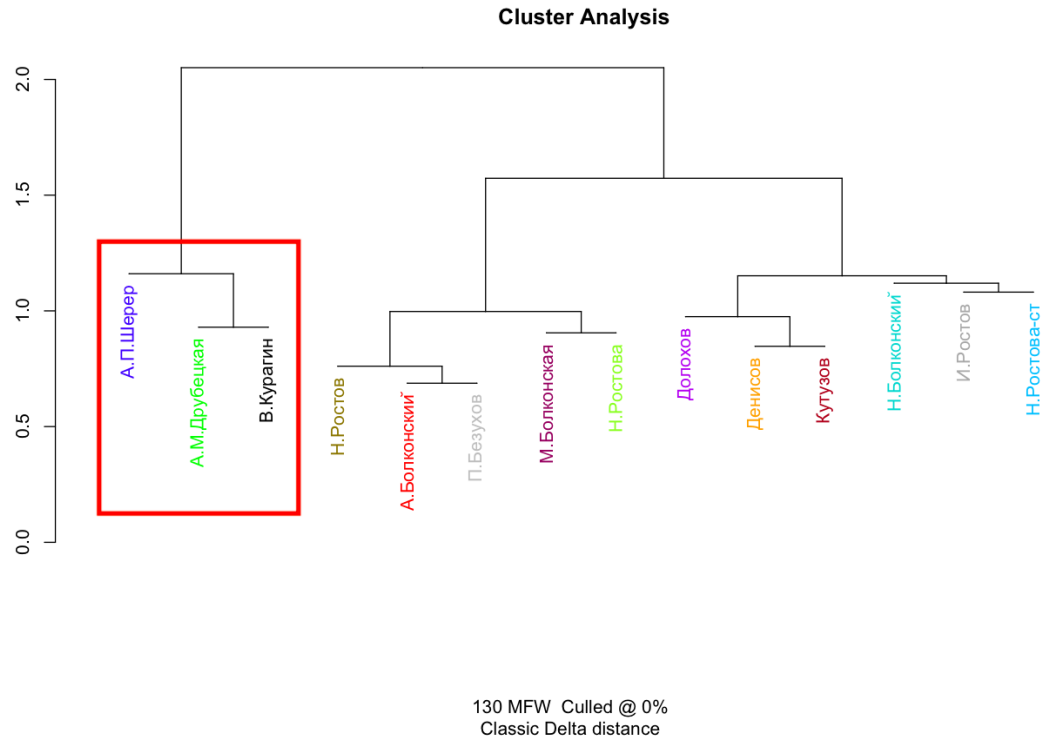

Рис. 2. Иерархическая кластеризация речи 14 персонажей «Войны и мира» на основе метрики Delta с использованием 130 наиболее частотных слов

Второй опробованный в работе метод является альтернативой стилеметрии. В нем целенаправленно используются характеристики речи, не связанные с ее лексическим составом. Эти характеристики подобраны с целью формализовать различия между героями в их манере говорить. При чтении «Войны и мира» можно заметить, что стили речи персонажей различны по своей экспрессии, непосредственности, темпераменту. Например, чрезвычайно жива и непосредственна речь Ростовых, в особенности Наташи, Пети и графа Ильи Андреевича:

1) Соня! что̀ ты?.. Что́, что́ с тобой? У-у-у!...
2) Что? Кому?.. Шутишь!
3) Я, я... я поеду с вами!

Речь князя Василия, напротив, построена с несвойственной устной речи правильностью:

4) Помните, что вы будете отвечать за все последствия, — строго сказал князь Василий, — вы не знаете, что вы делаете.

Для статистического анализа речь каждого персонажа была представлена в виде пяти количественных параметров (т.е. пятимерного вектора):

1. Доля восклицательных реплик
2. Доля вопросительных реплик
3. Доля знаков препинания в репликах (отношение числа знаков препинания к числу слов)

4. Частотность дискурсивных маркеров (в первую очередь частиц и междометий)
5. Читабельность текста (readability), рассчитанная на основе 5 наиболее известных метрик при помощи API сайта http://ru.readability.io/: индекс Флеша-Кинкэйда, индекс Колман-Лиау, метрика SMOG, Automatic Readability Index, формула Дэйла-Чэлла. Эти метрики опираются на среднюю длину слова (подсчет числа слогов) и предложения в тексте.

Группировка персонажей по формальным признакам, представленная при помощи метода главных компонент на рис. 3, частично совпадает с группировкой на основе стилеметрии.

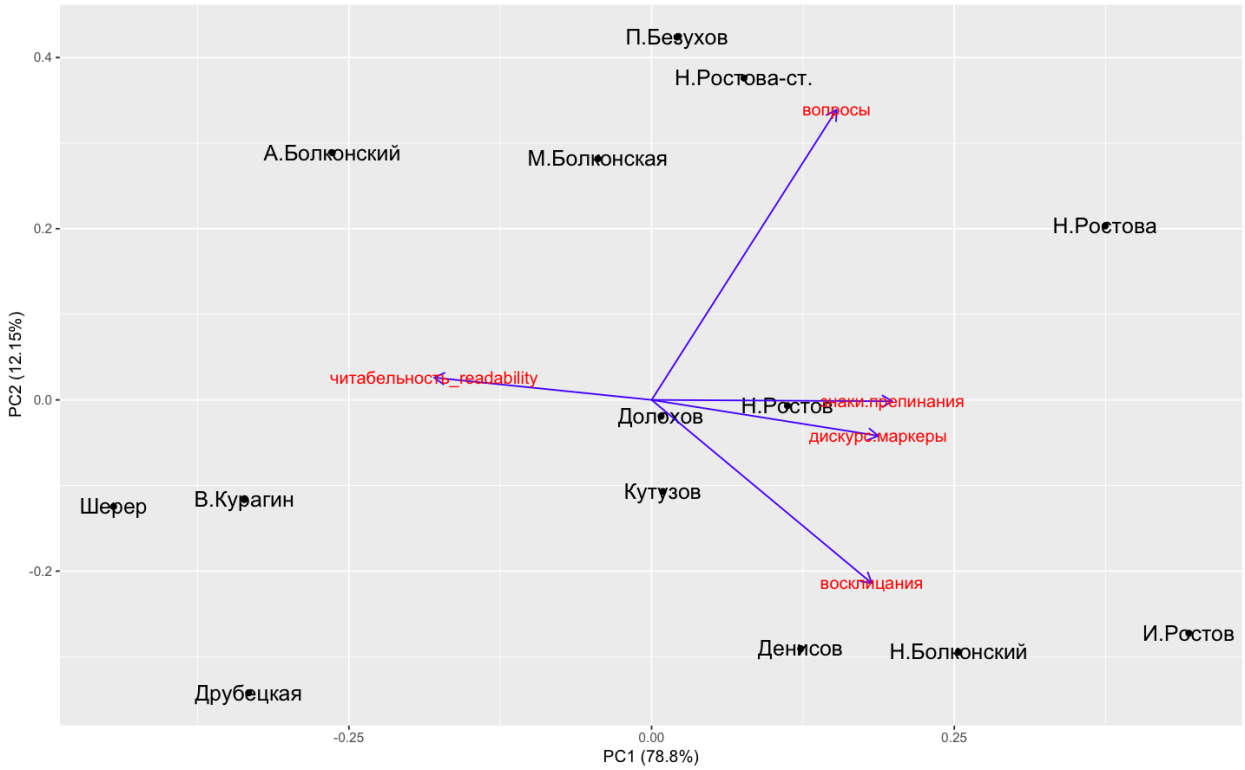

Рис. 3. Визуализация метода комплексной оценки нелексических параметров прямой речи при помощи метода главных компонент

Так, здесь еще более отчетливо выделяется группа В. Курагина, А. М. Друбецкой и А. П. Шерер. Важное отличие второго метода состоит в том, что мы можем интерпретировать этот результат. Левый нижний угол визуализации соответствует речи, которая содержит малую долю вопросительных и восклицательных реплик, в которой мало знаков препинания (в расчете на одно слово) и которая определяется как трудночитаемая инструментами измерения читаемости. Полностью противоположному типу речи — с высокой долей восклицаний и вопросов, большой долей знаков препинания и дискурсивных маркеров, и одновременно высокой читаемостью — соответствует правый верхний угол визуализации. Как можно видеть на рис. 3, обладателем такого противоположного типа речи является единственный персонаж — Наташа Ростова.

При сравнении с методом стилеметрии можно увидеть, что метод оценки нелексических параметров дает более дробную картину с разделением персонажей на большее число групп. Это же подтверждают результаты иерархической кластеризации на основе данных второго метода, представленные на рис. 4 и рис. 5.

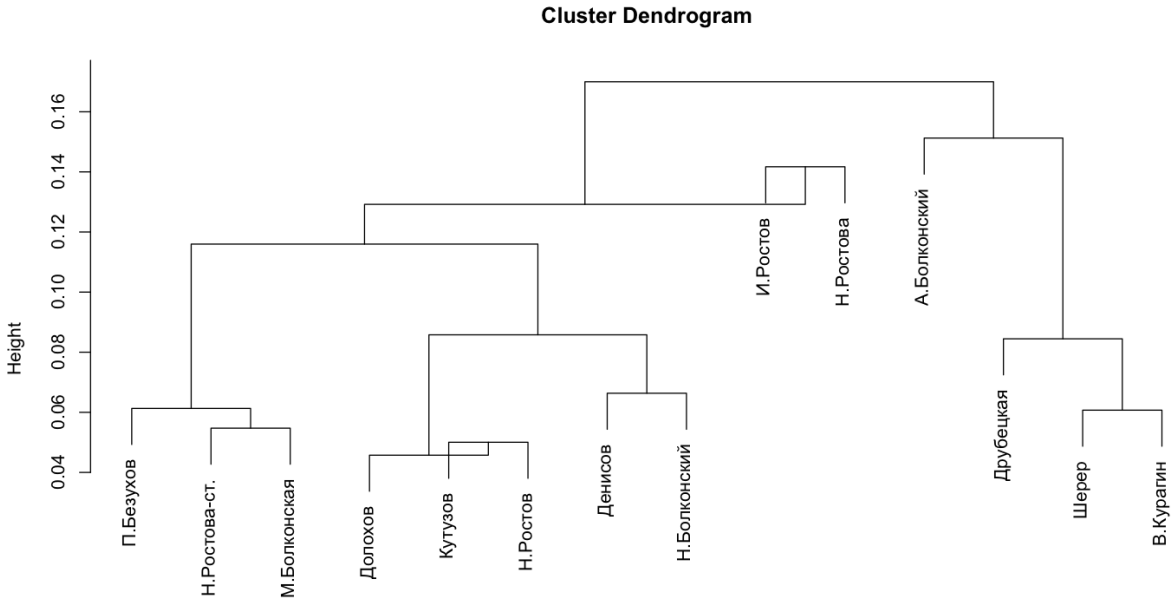

Рис. 4. Иерархическая кластеризация речи 14 персонажей «Войны и мира» на основе нелексических параметров текста

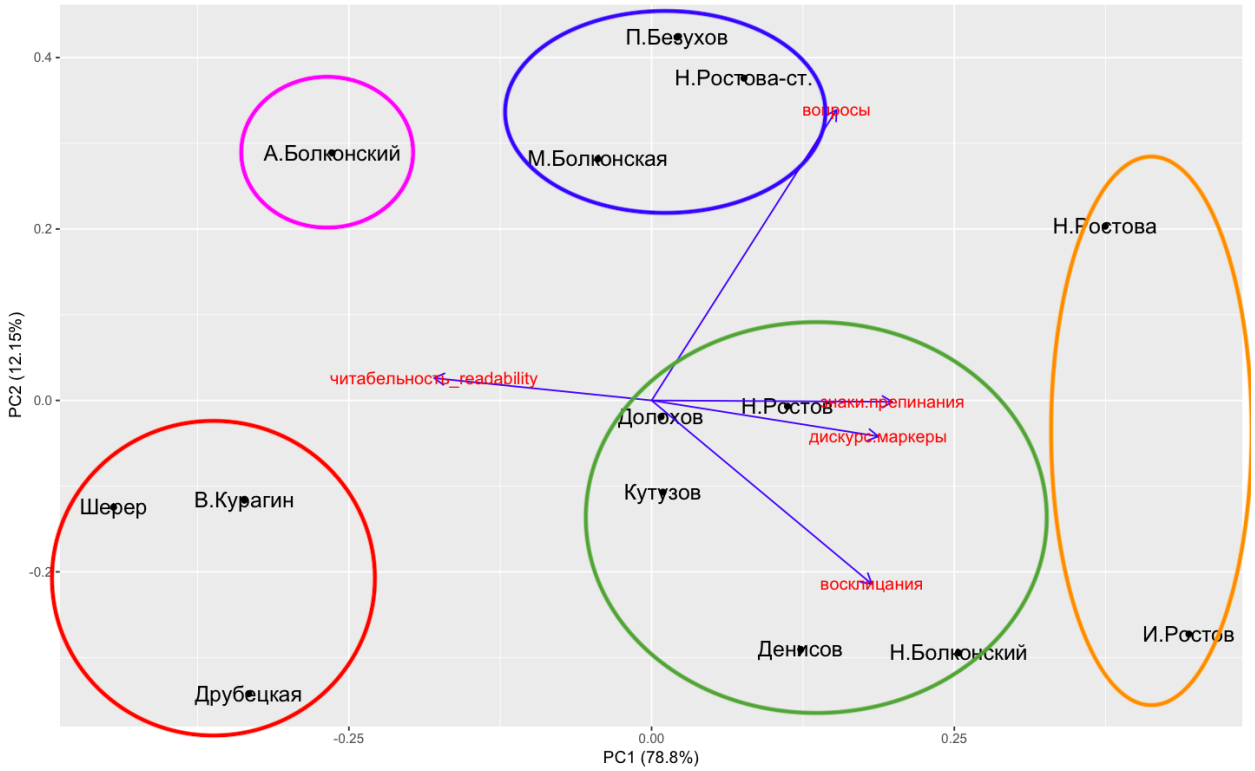

Рис. 5. Комбинированная визуализация пространства 14 персонажей на основе нелексических признаков (метод главных компонент + верхний уровень иерархической кластеризации)

Кластеризация методом нелексических признаков в большей мере улавливает некоторые эмоционально-темпераментные свойства персонажа, а также степень формализации его речи. Так, Николай Болконский — резкий, холерический и прямодушный персонаж, генерал в опале; Денисов резок и прямодушен изначально, остальные же черты приобретает по ходу действия: в конце «Войны и мира» Денисов — «отставной, недовольный настоящим положением генерал», критически настроенный по отношению к правительству «за свои неудачи по службе». Василий Курагин и Анна Шерер — два представителя высшего петербургского света, влиятельных при дворе, расчетливых и циничных. Илья Ростов и Наташа Ростова воплощают крайнюю степень ростовской непосредственности. Заметим, что в альтернативной классификации Илья Ростов не попадает в круг военных персонажей (Кутузов, Н. Болконский, Николай Ростов, Денисов, Долохов). Хотя граф Илья Ростов и связан с этим «мужским военным пространством» п, с миром Кутузова и Н. Болконского, по своему речевому темпераменту он схож не с этими героями, а с Наташей — что и фиксирует метод построения системы персонажей на основе нелексических признаков.

Наиболее значимым отличием двух методов представляется разница в позиции князя Андрея Болконского. Стилеметрия относит его к группе персонажей, в которую входят все прочие протагонисты. Альтернативный метод обособляет князя Андрея: из-за сложности его речи он не может быть отнесен к той же группе, что Пьер, Наташа или Николай. Этот результат примечателен, ведь в книге Л. Н. Толстого можно обнаружить и явные свидетельства авторского указания на формализм речи Болконского. Так, трижды для описания реплик князя Андрея применяется наречие «сухо»: в разных ситуациях он «сухо» обращается к Ипполиту, маленькой княгине, княжне Марье. Для описания речи иного персонажа это наречие применяется лишь однажды — «сухо» обращается к русскому посланнику Наполеон. Таким образом «сухость» речи князя Андрея выглядит как авторское указание на высокомерие (неизжитое тщеславие). В этом контексте результаты метода количественного анализа речи могут рассматриваться как раскрытие авторского приема, при помощи которого создается специфический портрет персонажа. Речь князя Андрея на фоне других персонажей изображается Л. Н. Толстым как более формальная, чем у Марьи, Николая, и тем более Наташи. Заметим, что в дни после помолвки с князем Андреем именно избыток жизни в Наташе на фоне Болконского вызывает тревогу у ее матери и брата: «Ее материнское чутье говорило ей, что чего-то слишком много в Наташе и что от этого она не

будет счастлива» [Л. Н. Толстой, 1980 (а): с. 290]; «Что значила улыбка Николая, когда он сказал: «уж выбран»? Рад он этому или не рад? Он как будто думает, что мой Болконский не одобрил бы, не понял бы этой нашей радости» [Л. Н. Толстой, 1980 (а): с. 278].

Глава 3 посвящена моделированию системы персонажей при помощи сетевого анализа. В первом разделе главы описываются два основных метода извлечения сетей персонажей из текста художественного произведения: метод соседства (совместной встречаемости внутри определенного отрезка текста), и диалоговый метод, когда связь устанавливается на основе обмена репликами. Также в первом разделе описываются три основные меры центральности персонажа, основанные на положении в структуре сети: взвешенная степень, центральность по посредничеству и центральность собственного вектора. Центральность собственного вектора как наиболее сбалансированная метрика используется в диссертации в качестве основной меры центральности.

Подготовленная в диссертации семантическая разметка позволяет применять оба основных метода извлечения сетей персонажей из текста. Во втором разделе главы сопоставлены две сети персонажей, построенные при помощи двух методов, и показаны различия методов.

На первом этапе исследования сети строились для всего произведения целиком. Были построены две взвешенные сети персонажей: на основе диалогового метода (далее — Д-сеть) и на основе метода соседства (далее — С-сеть). Узлами Д-сети стали все персонажи, которые хотя бы однажды являются адресатами или адресантами прямой речи. Узлами С-сети — однозначно идентифицированные именованные персонажи, которые хотя бы раз встретились в одном предложении с другим идентифицированным персонажем.

В каждой сети на основе метрик центральности было выделено и исследовано ядро каждой сети. На рис. 5 представлены такие ядра — по 10 наиболее центральных персонажей сети по метрике центральности собственного вектора, выделенные из Д-сети и С-сети соответственно.

Рис. 5. Д-сеть (слева) и С-сеть (справа), 10 наиболее центральных персонажей по метрике центральности собственного вектора; размер узла пропорционален центральности собственного вектора

Д-сеть отличается от С-сети по составу центральных персонажей. Так, в ядре Д-сети нет военно-исторических персонажей — Кутузова, Александра I, Наполеона; вместо них туда попали Денисов, Василий Курагин, графиня Ростова. Схожие отличия наблюдаются и при сравнении ядер сетей, полученных с использованием двух других метрик центральности.

Для обобщения центральностей в сетях в работе используются обратные ранговые значения. По каждой из трех метрик центральности был определен список из 10 наиболее центральных персонажей. Далее за 1 место в списке персонаж получал 10 баллов, за второе место — 9 баллов, за 3 место — 8 баллов и т.д. Показатели по всем трем сетям были суммированы. Таким образом для каждого персонажа получен его совокупный ранг в каждой сети (рис. 6).

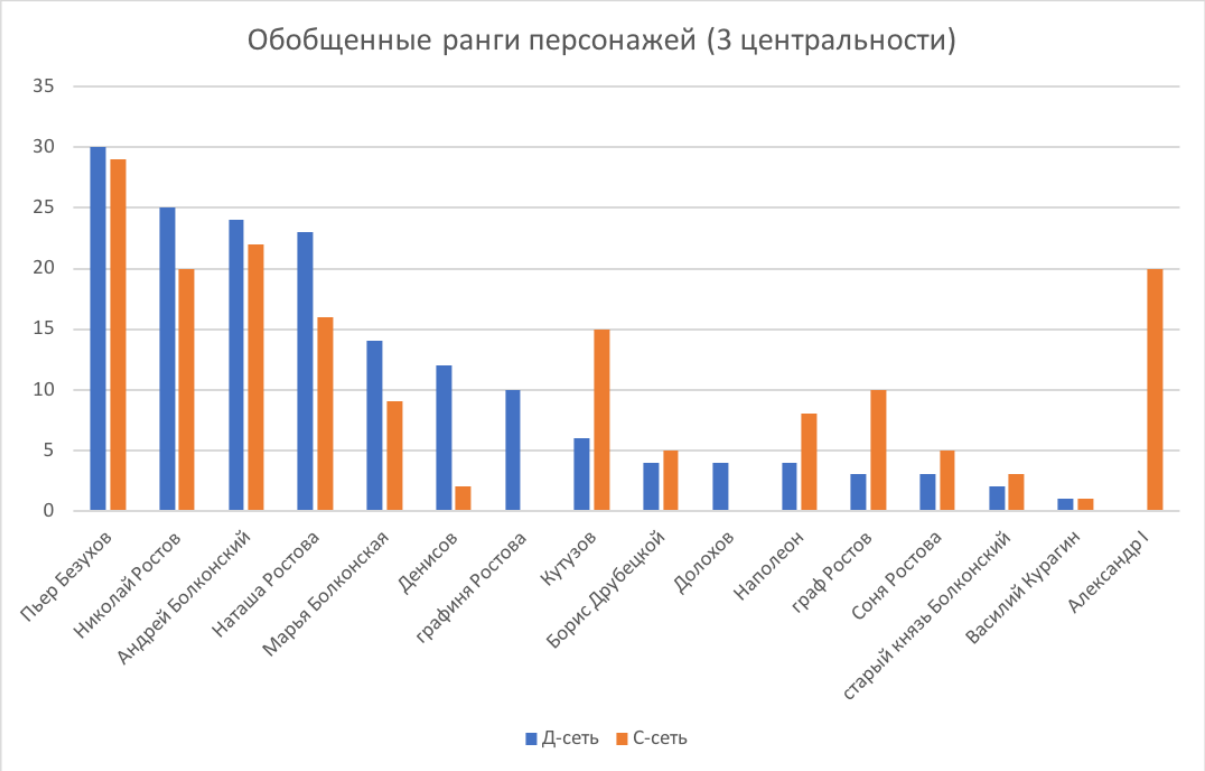

Рис. 6. Обобщенные ранги персонажей в двух сетях

Далее в главе 3 исследовалась структура сообществ в обеих сетях. С применением алгоритма оптимизации модулярности [Blondel *et al.*, 2008] обе сети были разделены на сообщества (кластеры) персонажей. Результаты разделения одним и тем же алгоритмом с одинаковыми настройками оказались различны для Д-сети и С-сети.

В Д-сети крупнейшие сообщества сгруппировались вокруг одного из главных персонажей. В отдельное сообщество выделилось русское военное командование. Еще одно сообщество образовала группа французских военачальников с Наполеоном в центре.

В С-сети были выделены четыре крупных сообщества. Самое крупное объединило практически всю военную составляющую системы персонажей: русское военное командование (ставка и адъютанты), включая Андрея Болконского, Наполеона и его окружение, «простых» военных персонажей (Тушина, Тимохина). Исключение составили те военные персонажи, которые попали в зону притяжения семьи Ростовых: Денисов, Долохов. Николай Ростов также оказывается в С-сети частью ростовского кластера. Еще два сообщества С-сети — семья и дворня Болконских (за исключением князя Андрея), а также смешанный разнородный кластер вокруг Пьера.

Далее в работе были проанализированы сети персонажей для отдельных частей «Войны и мира». Это позволило убрать большую часть хронологических наслоений и получить сети, в которых извлечение сообществ дало более интересный для анализа результат. Были построены Д-сети и С-сети для 15 основных частей книги и отдельно — для первой части эпилога. Подробно проанализированы три наиболее показательные пары сетей: для первой части, в которой происходит экспозиция героев и фиксируется начальное состояние системы персонажей; для второй части третьего тома «Войны и мира», в конце которой происходит Бородинская битва (о Бородинском сражении как узловой точке «Войны и мира» см. [Великанова, 2003]); эпилога, в котором завершаются описываемые в книге события. На примере этих наиболее показательных фрагментов продемонстрированы различия двух методов построения сетей.

Особенно ярко различия проявляются в сети для второй части третьего тома.

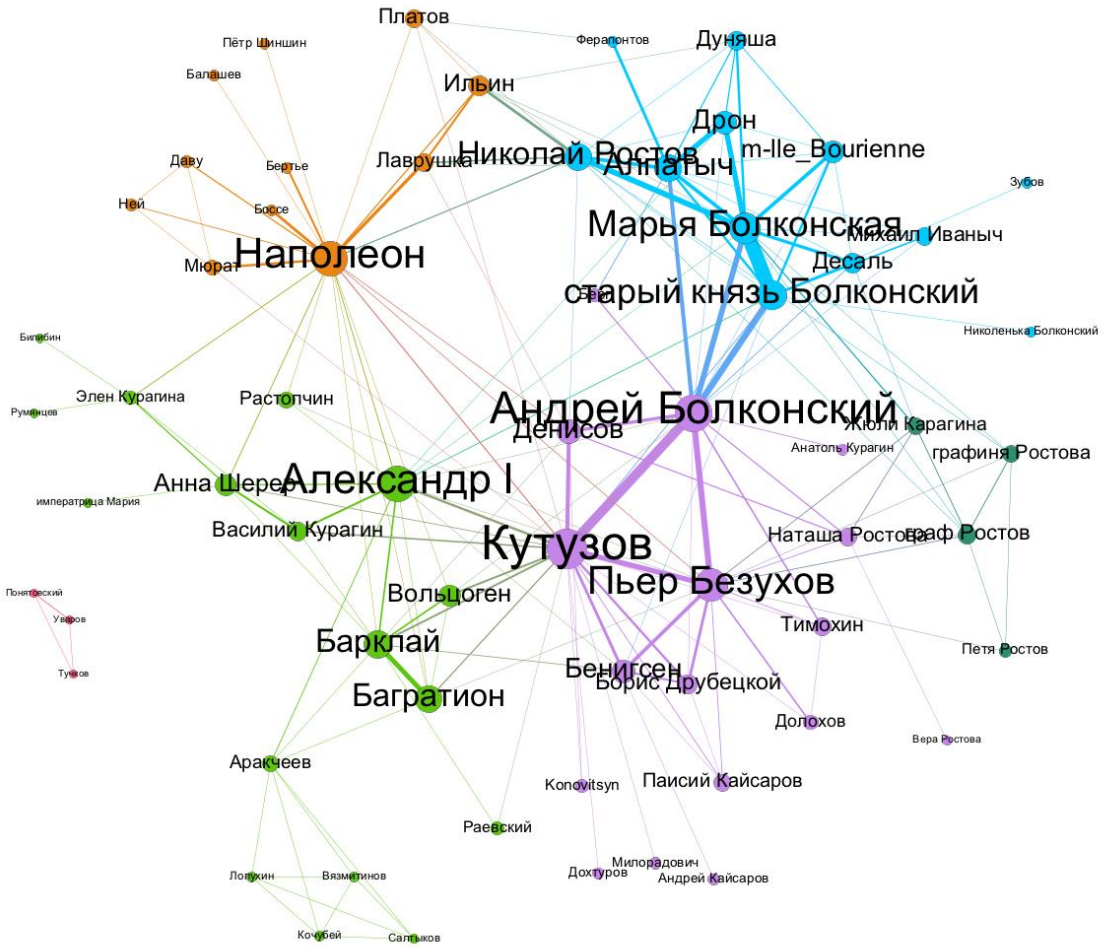

Рис 7. Сеть персонажей второй части третьего тома «Войны и мира», полученная методом соседства (C-сеть). Размер узла пропорционален центральности собственного вектора. Цветами обозначены результаты кластеризации графа

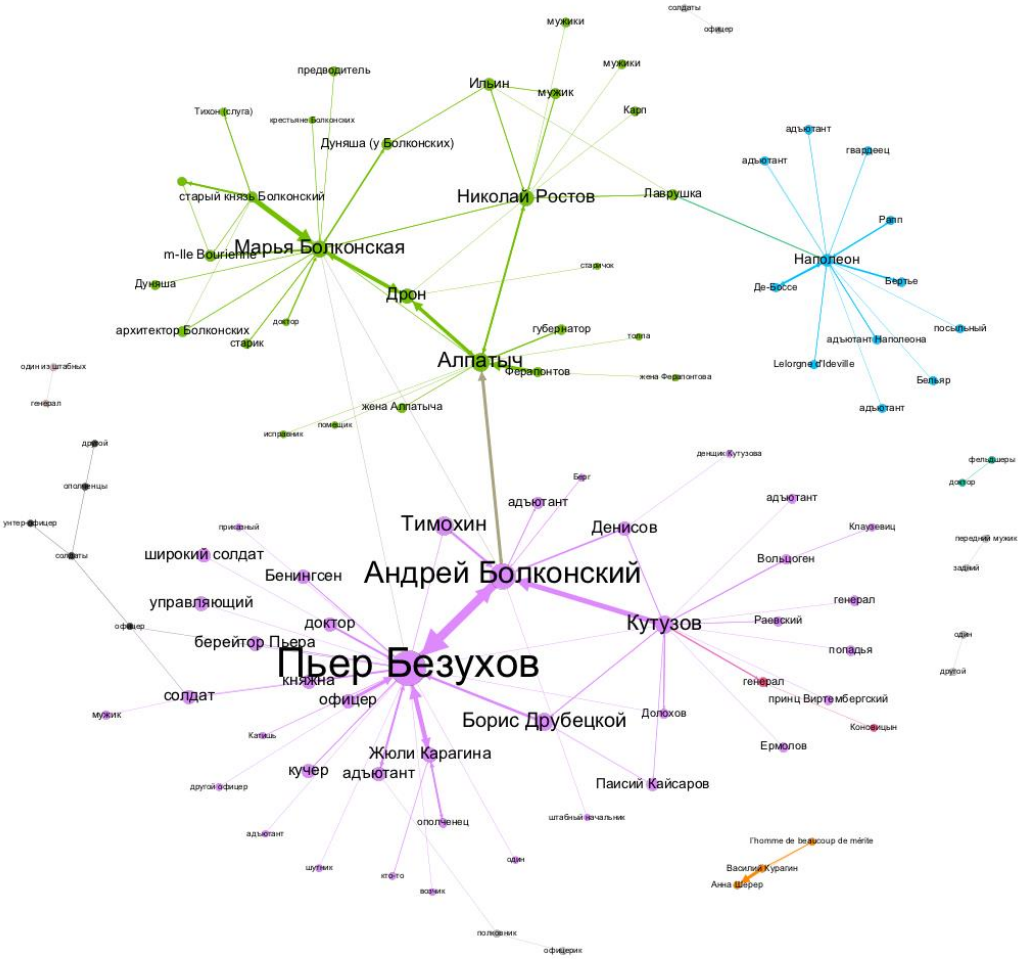

Рис. 8. Сеть персонажей второй части третьего тома «Войны и мира» на основе диалогового взаимодействия (Д-сеть). Размер узла пропорционален центральности собственного вектора. Цветами обозначены результаты кластеризации графа

В С-сети выражено противостояние двух сил на Бородинском поле в кульминационной точке произведения. Здесь (рис. 7) Кутузов и Наполеон оказываются двумя центрами двух крупнейших кластеров. В ядро С-сети по центральности входит и Александр I. В то же время в Д-сети (рис. 8) Кутузов и Наполеон — при схожей структуре выделяемых групп — остаются на вторых ролях. Здесь центральным оказывается Пьер, наблюдающий за сражением.

Видно также, что Д-сеть более детально отображает разные пространства «Войны и мира» — здесь А. П. Шерер и Василий Курагин образуют полностью изолированную группу (говоря на языке теории графов, отдельную компоненту) петербургского кружка. В С-сети те же светско-петербургские персонажи из-за частых упоминаний императора Александра I оказываются частью большого военно-политического кластера.

Также в главе 3 было произведено сопоставление Д-сетей и С-сетей в динамике по структурному параметру плотности сети. Плотность сети определяется как отношение числа связей в сети к максимально возможному их числу (т.е. все связаны со всеми). Динамика плотности двух сетей заметно отличается. В Д-сети чрезвычайно плотен по сравнению с другими частями эпилог — так вновь проявляется смещение информации о персонажах, которую моделирует данная сеть, в сторону семейно-бытового плана книги Л. Н. Толстого. С-сеть демонстрирует взаимосвязь между плотностью и сменой военных и мирных событий. Взаимосвязь была подтверждена при помощи корреляционного анализа, где плотность С-сети показала высокую (76,7%) корреляцию с тем, какие события — военные или мирные — описываются в соответствующей части.

Таким образом, в главе 3 показаны различия между тем, какие свойства системы персонажей многопланового литературного произведения могут отображать сети, построенные на основе двух разных методов. Д-сети, извлекаемые из диалогов, хорошо отражают прямые межличностные контакты, но упускают фоновые появления персонажей. В случае с таким многоплановым объектом, как «Война и мир», это может приводить к искажениям: диалоговые сети отображают в первую очередь семейно-бытовой план книги; военно-историческая часть сюжета лучше отражается в С-сети, построенной методом соседства. В то же время Д-сеть способна более точно отражать отдельные социальные группы в плотных частях сети персонажей.

В **заключении** приводятся основные результаты диссертационного исследования:

1. Произведен анализ работ по формализации понятия персонажа и компьютерному моделированию системы персонажей одного или нескольких произведения. Анализ работ выявил проблемный участок в современных цифровых исследованиях системы персонажей: сложность получения чистых структурированных данных о персонажах произведения напрямую из текста. Современным решением проблемы становится создание и публикация семантической разметки текста
2. Осуществлена семантическая разметка книги Л. Н. Толстого «Война и мир». В тексте произведения размечены упоминания персонажей (25,6 тыс. идентифицированных упоминаний) и вхождения прямой речи персонажей (6,3 тыс.) с однозначным указанием адресата и адресанта реплики. Разметка проводилась в автоматизированном режиме, использованный подход показал высокую точность определения персонажа.
3. Осуществленная разметка была использована для апробации метода моделирования системы персонажей и сравнения различных подходов к такому моделированию — как существующих, так и предложенного в настоящей работе метода анализа

нелексических признаков прямой речи.

4. Полученная разметка позволила применить к исследованию «Войны и мира» основные используемые сегодня методы компьютерного моделирования системы персонажей: количественный анализ прямой речи и сетевой анализ. Результаты такого анализа, кратко изложенные выше, могут быть воспроизведены на основе опубликованной разметки.
5. Полученная разметка позволила сравнить разные подходы внутри каждого метода, чего ранее не осуществлялось. Сравнение различных подходов на материале общеизвестного текста — книги «Война и мир» Л. Н. Толстого — показало, как выбор конкретного метода количественного анализа речи персонажей или метода извлечения сети из текста влияет на свойства получаемой модели персонажей.

Основное содержание работы отражено в следующих публикациях:

- Skorinkin D. Extracting Character Networks to Explore Literary Plot Dynamics // Komp'juternaja lingvistika i intellektual'nye tehnologii: po materialam ezhegodnoj mezhdunarodnoj konferencii «Dialog». 2017. Vyp. 16 (23): V 2 t. P. 257-270.
- Bonch-Osmolovskaya A., Skorinkin D. Text mining War and Peace: Automatic extraction of character traits from literary pieces// Digital Scholarship in the Humanities. 2017 Vol. 32, Issue supplement 1, , p. i17–i24
- Скоринкин Д. А., Бонч-Осмоловская А. А. «Особые приметы» в речи художественных персонажей: количественный анализ диалогов в «Войне и мире» Л. Н. Толстого // Электронный научно-образовательный журнал «История». 2017. Т. 7. № 7 (51)
- Скоринкин Д. А. Электронное представление текста с помощью стандарта разметки TEI // Вестник Московского университета. Серия 9: Филология. 2016. № 5. С. 90-108
- Skorinkin D., Mozhaev E. TEI markup for the 90-volume edition of Leo Tolstoy's complete works, in: TEI Conference and Members' Meeting 2016 Book of Abstracts. Wien: Austrian Centre for Digital Humanities. 2016. P. 107-109
- Skorinkin D. Digital Edition of the Complete Works of Leo Tolstoy // 6th AIUCD Conference Book of Abstracts. Rome. 2017. P. 264-267
- Bonch-Osmolovskaya A., Skorinkin D., Sidorova E. Verbal Identity of a Fictional Character: a Quantitative Study with a Machine Learning Experiment // Digital Humanities 2016. Conference Abstracts.Kraków: Jagiellonian University, 2016. P. 747-749.