УДК 81`322.2

В. В. Васильева, Э. С. Клышинский

МЕТОД ВЫДЕЛЕНИЯ СЮЖЕТНЫХ ЛИНИЙ, ОСНОВАННЫЙ НА АНАЛИЗЕ ОНОМАСТИКОНА ЛИТЕРАТУРНОГО ПРОИЗВЕДЕНИЯ

Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Поступила в редакцию 20.05.2024 г.
Принята к публикации 12.09.2024 г.
doi: 10.5922/vestnikpsy-2024-4-5

Для цитирования: *Васильева В. В., Клышинский Э. С.* Метод выделения сюжетных линий, основанный на анализе ономастикона литературного произведения // Вестник Балтийского федерального университета им. И. Канта. Сер.: Филология, педагогика, психология. 2024. №1. С. 47—60. doi: 10.5922/vestnikpsy-2024-4-5.

Одной из отраслей цифровых гуманитарных исследований является анализ структуры литературных произведений. Среди направлений исследований такого рода наиболее популярно создание социальной сети взаимодействия персонажей. Другую важную задачу составляет анализ структуры произведения, выделение фабульных единиц, их сравнение между собой. В статье предложен количественный метод выделения сюжетных линий, основанный на анализе употребления имен персонажей и названий локаций. В качестве единицы анализа используется авторское разделение текста на главы как наиболее адекватно выражающее общий замысел. Метод основан на поиске пересечений ономастиконов глав, пересечение оценивается с использованием коэффициента Дайса. Наличие метрики позволяет построить граф связности глав, из которого при помощи Лувенского алгоритма можно извлечь наиболее связанные фрагменты. Метод апробирован на материале текста романа М. А. Булгакова «Мастер и Маргарита». В первую очередь сюжетная линия Иешуа и Пилата в Ершалаиме была отделена от событий в Москве. Московские события разделены на три подсюжета: история массолитовцев и их отношения с мастером, московское Варьете и его сотрудники, история Маргариты. История Воланда и его свиты оказалась плотно связана с другими сюжетами и не выделилась в самостоятельный фрагмент.

Ключевые слова: анализ структуры произведения, литературное произведение, ономастикон произведения, обработка текстов на естественном языке

Введение

Начиная с конца XIX в. формализация структуры текста развивалась в форме анализа организации художественных произведений. В начале XX столетия под знаменем формализма появляются попытки моделирования литературных явлений и придания литературоведению

большей объективности. Одна из ранних работ такого рода принадлежит советскому фольклористу В. Я. Проппу. Его концепция сюжета волшебной сказки как сочетания функций персонажей (всего им выявлена 31 функция), изложенная в работе [1], послужила основой для разработки различных логических моделей сюжета. Формальный подход к анализу художественных текстов также был изложен в [2]. Ф. Моретти, возглавляющий сейчас Стэнфордскую лабораторию литературы (Stanford Literary Lab), ввел понятия «медленное чтение» (close reading) и «дальнее чтение» (distant reading) для обозначения противоположных подходов к анализу художественных текстов. «Медленное чтение» предполагает детальное изучение и интерпретацию отдельных текстов самим исследователем, в то время как «дальнее чтение» фокусируется на анализе больших коллекций или отдельных текстов как бы опосредованно, с опорой не на чтение, а на выделение сюжетных фрагментов или общих идей, в том числе используя различные вычислительные методы. Машина может извлекать важные параметры из текста и представлять их исследователю в удобном формате, снимая необходимость в ручной работе с данными. Например, машинные данные можно визуализировать на графиках или наглядно представить в виде таблицы. В своей следующей работе [3] Ф. Моретти сформулировал убедительные аргументы в пользу использования абстрактных моделей и количественных методов в анализе литературного текста. К числу абстрактных моделей он относит такие искусственные конструкты для описания, как графы (graphs), карты (maps) и деревья (trees). Данные методы актуальны и по сей день и активно применяются в рамках цифровых исследований в гуманитарной сфере благодаря направлению Digital Humanities (DH), или цифровая гуманитаристика. Работа Моретти оказала значительное влияние на сферу цифровых гуманитарных наук, вдохновив ученых на исследование литературы с использованием вычислительных и количественных подходов.

Сегодня мы наблюдаем третью волну количественных исследований литературы. Она обусловлена развитием IT и обилием данных в цифровой форме. Востребованность цифровых методов подчеркивается для разных гуманитарных дисциплин, в том числе для литературоведения. Для филологов и читателей анализ литературы с использованием цифровых технологий открывает возможность углубленно изучать произведение и обнаруживать неочевидные связи внутри текста. При этом, как указывают А. Володин и Б. Орехов, «цифровые исследовательские практики — это реальность любого ученого» [4, с. 19]. Следует отметить, что ключевые проблемы литературоведения остаются за пределами формального количественного анализа. Проблема состоит в высокой содержательной сложности объектов, требующих моделирования, среди которых можно выделить такие центральные категории литературоведения, как композиция произведения, сюжет, система персонажей. Тем не менее благодаря применению как формальных моделей, так и количественных методов появился спектр исследований художественного

текста, посвященных его графико-фонетическим, лексико-семантическим и композиционным особенностям (обстоятельный обзор современного состояния проблемы см., например, в [4]).

Важно отметить, что различные элементы сюжета исследованы в разной степени. Так, достаточно популярным объектом для анализа является система персонажей: анализ реплик персонажей, автоматическое выделение взаимоотношений между персонажами, построение графа взаимодействий между ними. Потребность в подобном анализе можно объяснить тем, что система персонажей — ключевой элемент сюжета, важный для его понимания. При этом после прочтения произведения бывает тяжело удержать в памяти связь персонажа с другими героями или историю его появления в сюжете. Зачастую читатели схематично отмечают для себя подобные связи и обоснования в ходе «дальнего чтения». Аналогичные инструменты помогают исследователям обосновать и визуализировать выдвигаемые ими положения. Однако труднее объяснить, почему меньше работ посвящено автоматическому выделению системы локаций или визуализации сюжетных линий — их тоже может быть несколько в произведении и их связи также могут быть нетривиальными.

Мы предлагаем количественный метод анализа литературного текста, позволяющий разделить несколько сюжетных линий, объединенных автором в единый сюжет. Метод основан на анализе упоминаний локаций и действующих лиц в тексте произведения. В качестве единицы анализа мы предлагаем опираться на авторскую структуру произведения — систему глав. Метод был апробирован на материале романа М. Булгакова «Мастер и Маргарита».

Обзор существующих подходов

Объектом исследования в цифровом литературоведении являются такие элементы, как персонажи и их речь, локации, события. Анализ литературного произведения при этом зависит от того, какой элемент сюжета исследуется. В частности, для исследования персонажей и их взаимодействий в цифровых гуманитарных исследованиях традиционно пользуются методом построения графов социальных отношений [3; 5]. При таком подходе узлами графа выступают герои произведения, а ребра графа отражают связи между персонажами.

Комбинирование нескольких видов анализа литературных произведений в рамках подходов из области цифровых гуманитарных исследований дает возможность провести более полный анализ структуры произведения, что упрощает исследовательский процесс и помогает выявить новые интересные аспекты в тексте как исследователям, так читателям за счет его представления в наглядном или интерактивном формате. Например, в рамках совместного проекта «Живые страницы» компании Samsung Electronics, лингвистов группы Tolstoy Digital и школы лингвистики НИУ ВШЭ реализованы несколько нестандартных сценариев знакомства с романом Льва Толстого «Война и мир» и его героями — таймлайны с событиями и судьбами, карточки-описания

и «цитатники» персонажей, интерактивные карты с привязкой мест к эпизодам художественного произведения. Доступность данного произведения обеспечена циклом нескольких других связанных между собой проектов, таких как «Весь Толстой в один клик» и «Tolstoy Digital» [6], в рамках которых была произведена полная оцифровка собрания сочинений Толстого и семантическая разметка в TEI-формате соответственно. В настоящее время проект развивается в новом направлении, осваивая уже чеховское наследие («Chekhov Digital») [7].

В рамках нашей работы система отношений между персонажами представляется не исчерпывающей для анализа сюжета. Одни и те же герои могут действовать в разных частях повествования, отличающихся также и локациями. В [8] авторы попытались выйти за рамки построения графов на основе коммуникаций персонажей и извлечь из литературных текстов не только «кто», но также «где» (то есть локации), что контрастирует с предыдущими исследованиями, которые были сосредоточены на разрешении референтности топонимов, то есть привязке географических названий к географическим координатам [9—11]. Так, например, в работе [12] авторы продемонстрировали систему CHAPLIN, которая создает граф по пьесе Уильяма Шекспира, где узлы представляют персонажей и места, а ребра обозначают их взаимодействия.

Наше исследование лежит в области, близкой к тематическому моделированию, позволяющему сформировать представление о наличии нескольких тем повествования, которое отчасти близко к представлению о сюжетных линиях. Хотя тематическое моделирование обычно применяется для новостных и научных текстов, есть примеры использования такого анализа для художественных произведений, например для англоязычных поэтических текстов [13] и русскоязычных художественных текстов [14; 15].

В данной работе мы высказываем гипотезу о том, что анализ употреблений имен собственных позволяет автоматически сгруппировать части сюжета, связанные с определенными персонажами или локациями.

Использованные данные и их предобработка

В качестве материала для экспериментов рассматривался роман М. А. Булгакова «Мастер и Маргарита» в его советском издании. Критериями для выбора материала были насыщенность произведения различными локациями и персонажами и наличие нескольких линий повествований. «Мастер и Маргарита» — произведение многоплановое, богатое сюжетными коллизиями. Композиционно роман разделен на 32 главы (эпилог в исследовании не учитывался), которые, в свою очередь, распределены по двум частям. Особенность этого романа заключается в противопоставлении двух сюжетных линий: исторического повествования об Иешуа и Понтии Пилате в Ершалаиме, и описания разнообразных событий, происходящих с мастером и Маргаритой, а

также с Воландом и его свитой в Москве. Причем линия Понтия Пилата разворачивается параллельно событиям в Москве и изложена всего в четырех главах, которые расположены симметрично в двух частях.

О. А. Митрофанова отделяет линию Понтия Пилата и Иешуа Га-Ноцри, которая развивается параллельно основной линии событий в Москве [16]. Тем не менее информация о темах в тексте без привязки к структуре произведения представляется недостаточной для формирования представления о композиционных особенностях произведения.

Как отмечалось выше, для выделения логически связанных элементов сюжета нам необходимо извлечь из текста информацию о героях и локациях, в которых происходят действия. Для их отбора традиционно используют модель распознавания именованных сущностей (Named Entity Recognition, NER). Для решения данной задачи доступны готовые нейросетевые модели, разработанные для русского языка, например синтаксические анализаторы Spacy и Stanza или специальные модели, входящие в состав библиотек Natasha и DeepPavlov. В своей работе мы использовали модель DeepPavlov ввиду ее преимущества по качеству распознавания.

Важно также отметить, что в произведении встречаются персонажи без имени или с прозвищами: это один из главных героев — мастер, а также кот Бегемот, которого, как единственного кота в повествовании, не всегда называют по имени, и Фагот-Коровьев, носивший прозвища «регент» и «клетчатый» (заметим, что сами эти слова недостаточно часто употребляются в произведении для того, чтобы выделить их при помощи статистических методов). Подобные случаи не получится распознать NER-моделью, поскольку это не имена собственные и они не всегда называют человека. Таким образом, чтобы не потерять необходимую информацию, мы фиксировали факт наличия этих трех лемм — мастер, кот и регент — в списке лемм для каждой главы, а также считали количество их упоминаний. Частота упоминаний считалась и для каждой сущности, распознанной при помощи NER-модели, и для ее кореферента. Для каждой главы создавался частотный словарь, хранящий количество упоминаний каждой распознанной сущности с пометой о виде именованной сущности (персона, локация, организация и пр.).

При выделении сущностей по главам также важно было учитывать, что именование одного и того же персонажа или места может упоминаться в разных падежных формах. Также могут встречаться несколько вариантов номинации для одного персонажа или места, например для героя Ивана Бездомного это Иван, Иван Николаевич, Бездомный, Иван Николаевич Понырев. Поскольку стала задача определить, о ком идет речь в главе, все подобные упоминания требовалось обобщить в рамках одной сущности, сложив при этом количество упоминаний каждого именования. Анафорические и прономинальные упоминания сущности при этом не учитывались. Объединение флексий проводилось путем объединения сущностей с общей леммой. Для нахождения различных номинаций одной сущности использовались общие подстроки: если у двух сущностей есть общая подстрока, то они объединялись [17].

При этом важно учитывать, что, если название места выражено топонимом, который состоит из имени собственного и номенклатурного слова (улица, бульвар, река), последнее не может расцениваться как общая подстрока для объединения сущностей. Это не позволит объединить «Арбатский переулок» и «Патриарший переулок». Поэтому при поиске общей подстроки производилась проверка по собранному нами датасету из слов, которые могут выступать такими номенклатурными словами в составе топонимов.

В связи с этим был собран словарь слов, относящихся к описанию элементов городской среды: улицы, площади, деревья, лес, парк, стены, дома, дворцы. Для этого использовался Национальный корпус русского языка (НКРЯ), откуда были собраны выгрузки слов топологических категорий «вместилища» и «горизонтальные поверхности», а также выгрузки слов с таксономическим классом «пространство и место» и «здания и сооружения».

Метод группировки сюжетных линий

Мы опирались на идею о том, что линии повествования объединены определенными персонажами и локациями, что позволяет объединить схожие части, опираясь при этом на авторское членение произведения на главы.

В качестве меры сходства был выбран коэффициент Дайса. Данный коэффициент дает высокий вес общим элементам, что полезно и для нашей задачи, где существенно пересечение множества персонажей и локаций разных глав.

Описанные выше проблемы вариативности наименования персонажей и локаций остаются актуальными и после разрешения их на уровне отдельных глав, потому что в одной главе, например, Ивана Бездомного могут называть Иваном и Иваном Николаевичем, тогда как в другой при упоминании о нем используется только псевдоним Бездомный. Возникает необходимость унифицировать именования, поскольку для подсчета пересечений одинаковые сущности должны быть записаны единообразно. Очевидно, что это актуально для персонажей и локаций с составными именами — ФИО и названий топонимов, включающих номенклатурное слово (например, где-то «Садовая улица», а где-то просто «Садовая»).

С этой целью был применен алгоритм Левенштейна для вычисления расстояния между строками и определения степени их схожести. Для героев и локаций с многословными именами совершался поиск близких именований следующим образом: если строки оказываются более близки к полному названию, чем определенный порог, то они считаются именами одного персонажа или локации. Далее в частотном словаре по каждой главе происходило объединение сущностей: все наименования, которые входят в какой-либо получившийся список, объединяются под полным названием сущности, к которой относится

этот список. Так, варианты «Михаил Александрович Берлиоз», «Михаил Александрович», «Берлиоз» объединяются под названием «Михаил Александрович Берлиоз».

После приведения именований сущностей по всем главам к единому написанию для всех возможных пар глав произведения считался коэффициент Дайса. Затем на основе рассчитанных значений близости для глав строился связный граф, в котором узлами выступают номера глав, а ребра отражают их взаимную близость.

Получившийся на данном этапе граф содержит информацию о связях разных глав в произведении, однако для того, чтобы понять, какие главы ближе друг к другу, этого недостаточно. Возможным решением могут стать алгоритмы для выделения сообществ на графах. Они позволяют выявлять группы узлов с плотными внутренними связями, что имеет практическое применение в различных областях. В рамках нашего исследования использовался Лувенский алгоритм [18], в результате работы которого получаются группы вершин (глав), предположительно принадлежащие одной линии повествования в произведении.

Результаты экспериментов

Описанный нами метод позволил выявить некоторые композиционные особенности текста романа «Мастер и Маргарита».

Наш метод разделил произведение на два слабо связанных кластера, один из которых представляет собой сюжетную линию Понтия Пилата и Иешуа Га-Ноцри в Ершалаиме и состоит из глав 2, 16, 25, 26. Остальные главы относятся к событиям в Москве.

Как отмечает А. З. Вулис [19], в первой части романа (главы 1—18) сюжетные линии сходятся к мастеру, а во второй (главы 19—32) — к Маргарите. Первая часть насыщена сюжетными поворотами больше, чем вторая. С участием Маргариты во второй части сюжет сосредотачивается вокруг стремления героини найти своего возлюбленного.

Среди глав, относящихся к событиям в Москве, удалось выделить три группы, которые соответствуют авторскому делению произведения на две части. Так, вторая часть была выделена полностью, за исключением глав 25 и 26, входящих в историю Иешуа, и глав 27, 28, 32, содержащих описание бед, совершенных свитой Воланда при расставании с Москвой. Первая часть, за исключением глав 2 и 16, входящих в историю Иешуа, соответствует двум сообществам, описывающим историю массолитовцев и Варьете. Результаты группировки показаны на рисунке 1. Заметим, что главы 27, 28 и 32 оказались больше связаны с историей Варьете, проходящей через две части романа. Это наводит на мысль о том, что изменение параметров Лувенского алгоритма могло привести к выделению еще одного сообщества, посвященного Воланду.

Таким образом, мы можем утверждать, что наш метод позволяет корректным образом разделить граф произведения на отдельные сюжетные линии, используя информацию об именах героев и названиях мест и организаций. Полученное разделение на подсюжеты в целом согласуется с мнением других авторов [19]. Сами связи между главами внутри выде-

ленных сообществ можно увидеть на рисунках 2—5. Здесь ширина линии показывает значение коэффициента Дайса, то есть степень связанности двух глав. Заметим, что некоторые главы расположены далеко от центра сообщества. Это позволяет неявно оценить степень связанности главы с выделенной сюжетной линией. Так, глава 5 «Дело было в Грибоедове» несколько отстоит от повествования о Бездомном, являющемся центральной фигурой в истории массолитовцев, что хорошо видно на рисунке 3. Аналогично главы 28 и 32 выделяются из подсюжета с сотрудниками Варьете, а главы 21 и 23 — из истории Маргариты.

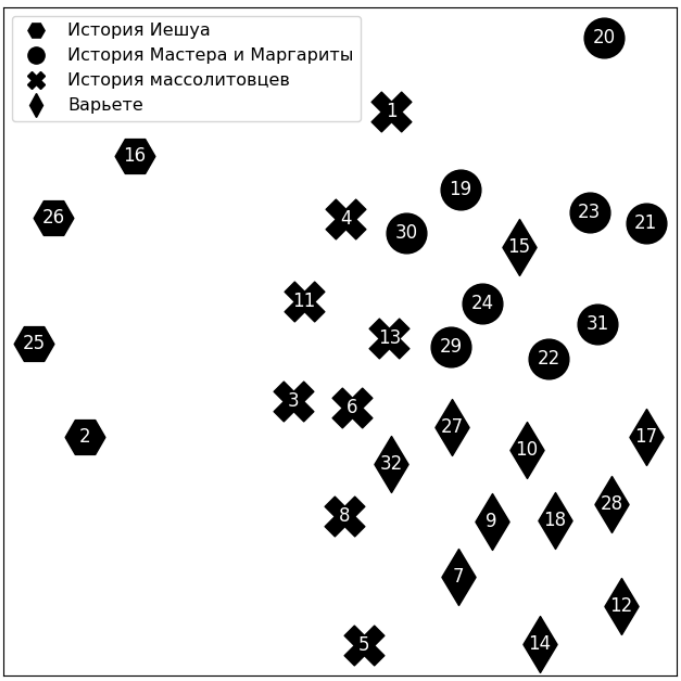

Рис. 1. Разделение глав романа М. А. Булгакова «Мастер и Маргарита» при помощи Лувенского алгоритма на графе связей по именам собственным. Близость вершин может показывать близость глав по списку персонажей и локаций

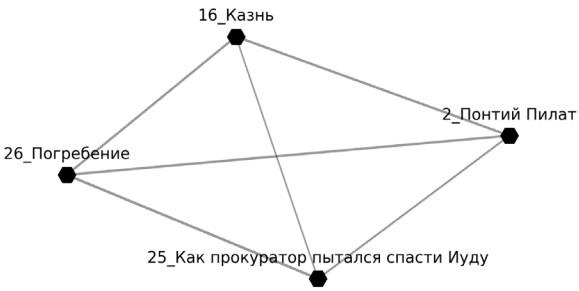

Рис. 2. Граф связей глав истории Иешуа

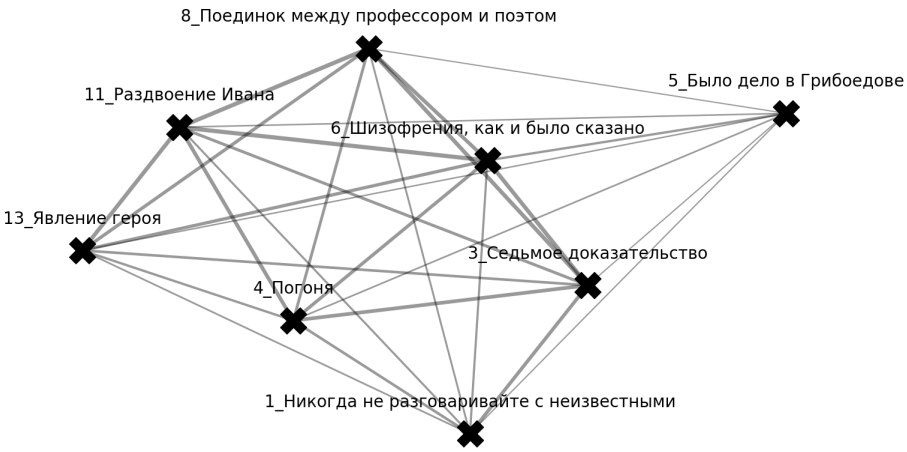

Рис. 3. Граф связей глав истории массолитовцев

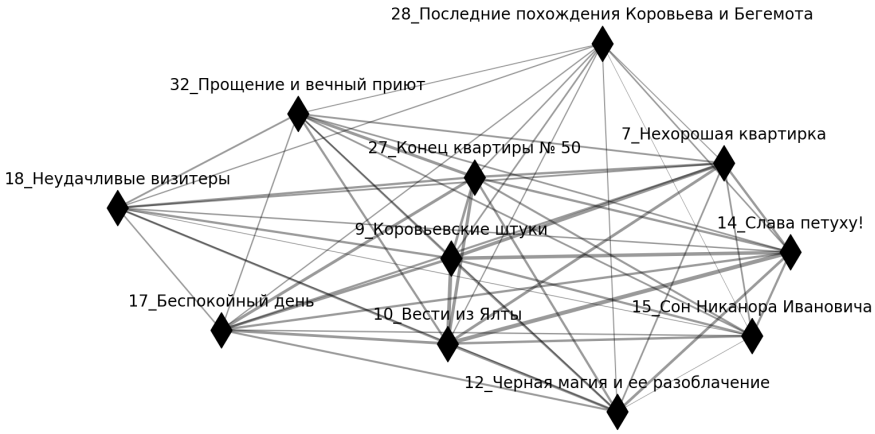

Рис. 4. Граф связей глав истории сотрудников Варьете

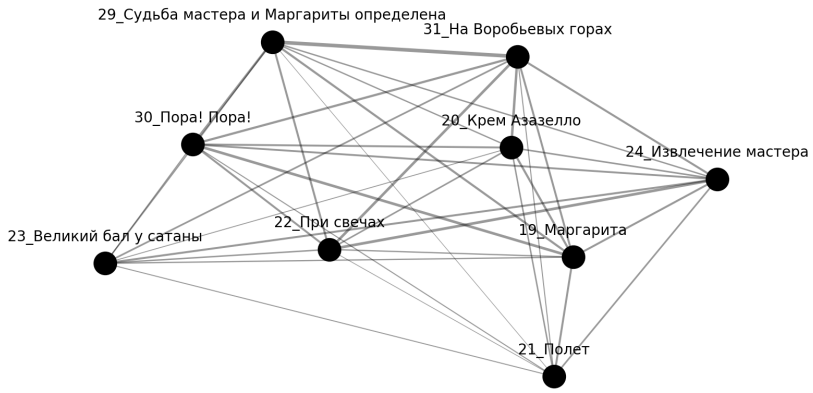

Рис. 5. Граф связей глав истории Маргариты и мастера

На рисунке 6 представлена матрица связей всех глав между собой, оттенками серого показана степень связанности глав между собой. Хорошо видно, что история Иешуа (левый верхний угол матрицы) практически не связана по именам собственным с другими главами. Аналогично выделяются и другие истории, хотя они имеют значительно больше связей между собой.

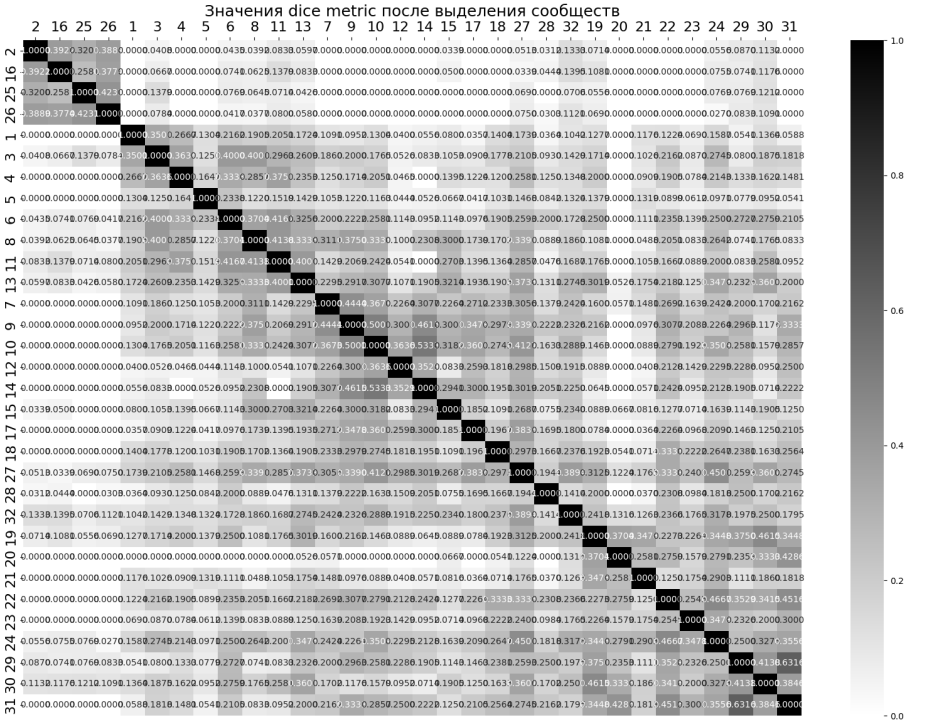

Рис. 6. Значения коэффициента Дайса
для выделенных сообществ (подсюжетов)

Заключение

В данной работе был предложен метод поиска подсюжетов в рамках литературного произведения, использующий информацию об ономастиконе произведения (именах, прозвищах и других обозначениях персонажей, названиях организаций, топонимах в широком смысле). Линейная структура произведения часто задает необходимость последовательного чередования сюжетных линий отдельных героев, которые связываются по ходу изложения общей истории или в конце произведения. Наша гипотеза заключалась в том, что общность мест или персонажей позволит определить связанность этих чередующихся фрагментов. Для выделения сюжетных линий использовалось авторское деление произведения на части (главы).

Мы определили связанность глав как степень пересечения ономастикона этих глав, задаваемую при помощи коэффициента Дайса. Это позволяет построить граф связей между главами. Далее при помощи ал-

горитма выделения сообществ в графе (в нашем случае — Лувенского алгоритма) выделяется подмножество глав, связанных между собой общими именами персонажей или локаций. Заметим, что для нас важны как слова автора, описывающие действия героев, так и любые упоминания о герое или месте, которые свидетельствуют о наличии сюжетной связанности. По этой причине мы не проводим фильтрацию текста и не удаляем прямую речь[1]. Помимо этого мы не разделяем роли персонажей в повествовании, считая, что протагонист и антагонист могут иметь одинаковую важность для сюжетной линии. Даже редкие упоминания мест и героев могут быть важны. Подобное можно видеть, например, в пьесе Т. Стоппарда «Розенкранц и Гильденстерн мертвы», где основные герои «Гамлета» становятся «задником» произведения, появляются лишь изредка, но задают основную канву повествования.

Заметим, что метод будет плохо работать в случае, если авторское деление произведения на части не совпадает с разделением на сюжетные линии. Подобный прием (клиффхэнгер) часто используется авторами для введения неожиданного поворота сюжета или для удержания внимания читателя. Кроме этого, наш метод плохо приспособлен для произведений, написанных от первого или второго лица, а также изложенных как поток сознания. Заметим, что существует не так много литературных текстов, в которых совершенно не используются имена собственные, поэтому можно будет скорее выделить части с более активным их использованием и противопоставить частям с «привычной» обстановкой, не требующей имен. Для подобных фрагментов хороший результат должны давать также упоминавшиеся в обзоре методы тематического моделирования.

При этом предложенный метод может использоваться для выделения фабульных единиц, повторов тематического рисунка, основанного на упоминании имен и мест, или обнаружения «чеховского ружья», когда более раннее упоминание имени собственного или иного обозначения места либо персонажа может служить ключом к пониманию более поздней части.

Метод был проверен на материале произведения М. А. Булгакова «Мастер и Маргарита». Мы смогли получить результаты, согласующиеся с мнением других авторов (например, [16; 19]). Текст произведения был разделен на несколько сюжетных линий. Полностью разделились история Иешуа Га-Ноцри и Понтия Пилата в Ершалаиме и повествование о московских событиях. В последнем выделились три подсюжета: исто-

[1] Отделение упоминаний героев и мест от их прямого участия в сюжете представляет собой отдельную проблему. Помимо прямой речи в произведении могут встречаться авторские описания сути диалога героев, включающие имена третьих лиц или мест («По его лицу я понял, что мою кандидатуру они обсудили и утвердили сообща с Габдракиповым». Г. Жженов, «Прожитое»), что делает полное решение задачи удаления упоминаний затруднительным. Такую же проблему составляют кореферентные ссылки, разрешение которых могло бы существенно повысить точность предложенного метода. Однако существующие методы не обеспечивают нам достаточной точности для литературных произведений, в связи с чем было принято решение отказаться от их применения.

рия массолитовцев, история, связанная с московским Варьете, и история Маргариты. История Воланда и его свиты переплетена с последними тремя сюжетами и не была выделена отдельно. Выявленные сюжетные линии соотносятся с авторским делением романа на две части, первая из которых относится к мастеру и МАССОЛИТу, а вторая — к Маргарите и Воланду, причем история Воланда и Варьете вплетена в обе части.

Список литературы

1. *Пропп В. Я.* Морфология сказки. М., 1969.

2. *Moretti F.* Conjectures on World Literature // New Left Review. 2000. №1. P. 54—68.

3. *Moretti F.* Graphs, Maps, Trees: Abstract Models for Literary History. L., 2005.

4. *Цифровые* гуманитарные исследования : монография. Красноярск, 2023.

5. *Moretti F.* Network theory, plot analysis // New Left Review. 2011. №68. P. 80—102.

6. *Bonch-Osmolovskaya A., Skorinkin D., Pavlova I. et al.* Tolstoy semanticized: Constructing a digital edition for knowledge discovery // Journal of Web Semantics. 2019. Vol. 59 (2). Art. №100483.

7. *Северина Е. М. Бонч-Осмоловская А. А., Кудин А. М.* Цифровые филологические практики: проект «Chekhov Digital» // Актуальные проблемы филологии и педагогической лингвистики. 2022. №2. С. 153—165. doi: 10.1016/j.websem.2018.12.001.

8. *Lee J., Yeung C. Y.* Extracting Networks of People and Places from Literary Texts // Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation / ed. by R. Manurung, F. Bond. Faculty of Computer Science, Universitas Indonesia, 2012. P. 209—218.

9. *Li H., Srihari R. K., Niu C., Li W.* Location normalization for information extraction // COLING 2002: The 19th International Conference on Computational Linguistics. 2002. URL: https://aclanthology.org/C02-1127 (дата обращения: 01.05.2024).

10. *Purves R. S., Clough P., Jones C. B. et al.* The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet // International Journal of Geographical Information Science. 2007. Vol. 21 (7). P. 717—745. doi: 10.1080/13658810601169840.

11. *Jones C. B., Purves R. S.* Geographical information retrieval // International Journal of Geographical Information Science. 2008. №22 (3). P. 219—228.

12. *Marazzato R., Sparavigna A. C.* Extracting Networks of Characters and Places from Written Works with CHAPLIN // ArXiv. URL: https://arxiv.org/abs/1402.4259 (дата обращения: 10.05.2024).

13. *Rhody L. M.* Topic Modeling and Figurative Language // Journal of Digital Humanities. 2012. Vol. 2 (1). URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ (дата обращения: 10.05.2024).

14. *Митрофанова О. А.* Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Корпусная лингвистика — 2015 : тр. междунар. конф. СПб., 2015. С. 332—343.

15. *Mitrofanova O. A., Sedova A. G.* Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose) // Information Technology and Computational Linguistics (ITCL-2017). Association for Computing Machinery, 2017.

16. *Митрофанова О. А.* Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М. А. Булгакова. // Корпусная лингвистика — 2019 : тр. междунар. конф. СПб., 2019. С. 387—394.

17. *Zykova V. I., Klyshinsky E. S.* Remus, Lupin and Moony Walk in a Bar… Grouping of Proper Names Related to the Same Denotation in Large Literary Texts Collections // Computational Linguistics and Intellectual Technologies : Proceedings of the International Conference «Dialogue 2023». URL: https://www.dialog-21.ru/media/5882/zykovaviplusklyshinskyes105.pdf (дата обращения: 10.05.2024).

18. *Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E.* Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment. 2008. Vol. 10. Art. № P10008. doi: 10.1088/1742-5468/2008/10/P10008.

19. *Вулис А. З.* Роман М. А. Булгакова «Мастер и Маргарита». М., 1991.

Об авторах

Варвара Вячеславовна Васильева – студ., Национальный исследовательский университет «Высшая школа экономики», Россия.

E-mail: varvaravasilyeva22@gmail.com

https://orcid.org/0009-0007-0814-7874

Эдуард Станиславович Клышинский – канд. техн. наук, доц., Национальный исследовательский университет Высшая школа экономики, Россия.

E-mail: eklyshinsky@hse.ru

https://orcid.org/0000-0002-4020-488X

SPIN-код 2185-5292

V. V. Vasilyeva, E. S. Klyshinsky

AN ONOMASTICON-BASED QUANTITATIVE METHOD FOR IDENTIFYING OF STORYLINES IN A LITERARY WORK

National Research University Higher School of Economics, Moscow, Russia

Received 20 May 2024
Accepted 12 September 2024
doi: 10.5922/vestnikpsy-2024-4-5

To cite this article: Vasilyeva V. V., Klyshinsky E. S., 2024, An onomasticon-based quantitative method for identifying of storylines in a literary work, *Vestnik of Immanuel Kant Baltic Federal University. Series: Philology, Pedagogy, Psychology*, № 4. P. 47—60. doi: 10.5922/vestnikpsy-2024-4-5.

One of the branches of digital humanities research is the analysis of the structure of literary works. Among the research directions in this field, the creation of a social network of character interactions is particularly popular. Another important task is the analysis of the structure of a work, the identification of narrative units, and their comparison. This article proposes a quantitative method for identifying plot lines based on the analysis of character names and location names. The author's division of the text into chapters is used as the unit of analysis,

as it most adequately reflects the overall intent. The method is based on finding intersections of the onomastic content of chapters, with intersections evaluated using the Dice coefficient. The presence of a metric allows for the construction of a chapter connectivity graph, from which the most connected fragments can be extracted using the Louvain algorithm. The method has been tested on M. A. Bulgakov's novel The Master and Margarita. *Primarily, the plotline of Yeshua and Pilate in Yershalaim was separated from the events in Moscow. The Moscow events were divided into three subplots: the story of the MASSOLIT members and their relationship with the Master, the Moscow Variety Theater and its employees, and Margarita's story. The story of Woland and his retinue was closely intertwined with the other plots and did not form an independent fragment.*

Keywords: analysis of a literary work, literary work, onomasticon-based, natural language processing

The authors

Varvara V. Vasilyeva, Student, National Research University Higher School of Economics, Russia.

E-mail: varvaravasilyeva22@gmail.com

https://orcid.org/0009-0007-0814-7874

Dr Eduard S. Klyshinsky, Associated Professor, National Research University Higher School of Economics, Russia.

E-mail: eklyshinsky@hse.ru

https://orcid.org/0000-0002-4020-488X

SPIN code 2185-5292