УДК 81’33

О ПРИНЦИПАХ НОРМАЛИЗАЦИИ ТЕМАТИЧЕСКОЙ РАЗМЕТКИ КОРПУСА РУССКОГО РАССКАЗА XX ВЕКА

Маргарита Александровна Кирина
преподаватель департамента филологии,
младший научный сотрудник лаборатории языковой конвергенции
Национальный исследовательский университет
«Высшая школа экономики» – Санкт-Петербург
190068, Санкт-Петербург, наб. канала Грибоедова, 119–121. mkirina@hse.ru

В статье рассматривается проблема нормализации тематической разметки Корпуса русского рассказа XX в. Целью исследования стала разработка методологии, сочетающей в себе лингвистические и литературоведческие подходы к анализу текста, и стандартизация параметра «тема», выделяемого экспертным путем. В рамках исследования предлагается рассматривать тематику художественного произведения как социокультурный феномен, в связи с чем обсуждаются перспективы изучения влияния внетекстологических факторов на тематическое разнообразие текстов определенного исторического периода.

Ключевые слова: русский рассказ; тема художественного произведения; корпусная лингвистика; тематическая разметка; нормализация.

1. Введение

Квантитативный подход к анализу текста широко применяется в корпусной лингвистике. Как подчеркивает Е. Тогнини-Бонелли, преимуществом корпусно-ориентированного анализа текста является функциональное описание языка и социокультурного контекста. Предметами статистического анализа становятся частотность и распределение языковых единиц в тексте, а также установление связи между количественными характеристиками текста и порождаемыми им значениями, зачастую выходящими за рамки текста [Tognini-Bonelli 2001]. В этой связи интерес представляет такой параметр, как тема текста, т. е. его содержательные характеристики, и исследование этого параметра зависимости от внетекстологических факторов.

Тематика, выступая ключевым семантическим компонентом художественного произведения, на реализацию которого направлены все его идейные составляющие, может служить достоверным свидетельством о тех общественных настроениях, которыми характеризовался период создания текста [Лотман 1996]. В настоящей статье описываются основные принципы методологии, разработанной для анализа тематического компонента художественного произведения на материале малой русской прозы. Предлагаемая техника выделения и нормализации тематических компонентов художественных текстов была апробирована на данных аннотированного подкорпуса Корпуса русского рассказа 1900–1930 гг. [Мартыненко и др. 2018б; Sherstinova, Martynenko 2020] и планируется для использования в работе с тематической разметкой Корпуса русского рассказа XX века [Шерстинова, Кирина, Хлусова 2023]. Рассматриваемая концепция сочетает в себе корпусные и литературоведческие подходы к анализу текста, а также учитывает взаимосвязь между темой как семантическим компонентом текста и временем его создания, ранее обсуждавшуюся в [Shersinova, Skrebtsova 2020; Skrebtsova 2020].

2. Понятие «тема». Подходы к определению темы текста

2.1. Тема как литературоведческая категория

В литературоведении понятие «тема», обозначая структурный элемент содержания художественного произведения, часто выступает предметом споров. Так, некоторыми исследователями предлагается определять тему как «объект художественного отражения, то есть те жизненные характеры и ситуации (взаимоотношения характеров, а также взаимодействия человека с обществом в целом, с природой, бытом и т. п.), которые <…> переходят из реальной действительности в художественное произведение и образуют объективную сторону его содержания» [Есин 2000: 22]. Все темы произведения в совокупности составят тематику произведения.

Согласно А.Б. Есину, «для того, чтобы “выйти” непосредственно на тему, надо раскрыть характеры, воплощенные в персонажах» [там же: 23]. Под характерами здесь подразумевается определение на глубинном уровне тех отношений, в которых находятся герои произведения. Так, посредством анализа характеров персонажей выделяются конкретно-исторические темы – «характеры и обстоятельства, рожденные и обусловленные определенной социально-исторической ситуацией в той или иной стране» [там же: 24]. Причем такой анализ должен складываться из трех параметров: социального, временного и национального. При этом стоит отметить, что необходимо выделение не только конкретно-исторических, но и вечных, так называемых общечеловеческих, тем (например, любви, дружбы, смерти и т. п.). Такие темы «фиксируют повторяющиеся моменты в истории различных национальных обществ, …в жизни разных поколений, в разные исторические эпохи» [там же: 24]. В контексте настоящего исследования обращение к ним или, наоборот, отказ от них во время социальных потрясений также может служить характеристикой рассматриваемого периода в истории литературы.

Несколько иное определение темы дает Б.В. Томашевский: «тема (о чем говорится) является единством значений отдельных элементов произведений» [Томашевский 1996: 176]. При этом тему имеет не только произведение в целом, но и его отдельные, составные части: так, тема обладает свойством «разложения <…> на мельчайшие повествовательные единицы» [там же: 182]. Путем такого разложения произведения происходит выявление «системы мотивов, составляющих тематику данного произведения» и предотвращающих непреднамеренный распад произведения на части» [там же: 191]. В этом смысле задача выделения общей темы сводится к сбору этих равномерно распределенных тематических элементов воедино – иными словами, их суммированию.

Еще одним взглядом на проблему определения темы и ее непосредственного выведения из художественного текста является понимание темы как наиболее абстрактного «семантического аспекта описания», работающего в соответствии с моделью «тема – текст» [Жолковский, Щеглов 2013: 37]. А.К. Жолковский и В.А. Щеглов предлагают сворачивать темы в модели, оформляемые на так называемом метаязыке, путем «вывода»: между темой и текстом, таким образом, постулируется существование закономерного соответствия, выводимого логически и не являющегося «выжиманием» из текста его «идейной квинтэссенции» [там же: 38]. Следует сказать, что моделируется здесь не замысел автора, т. е. творческая, или, как ее еще называют, эстетическая составляющая любого художественного текста, а competence, что понимается как «чисто логическое соответствие между темами и текстами», выраженное в тексте имплицитно [там же: 62].

Таким образом, утверждается возможность экспликации темы текста до разной степени подробности, вплоть до получения самых элементарных понятий. Применение того или иного из предложенных вариантов разложения темы зависит от задач исследователя. Для тематической разметки Корпуса русского рассказа первой трети XX века, осуществляемой в рамках данного исследования, предлагается совмещение двух способов вывода темы: нелинейного и образного [там же].

2.2. Тема и читатель

Важную роль при определении темы имеет ориентированность на читательскую аудиторию. Данный фактор наряду с прочими учитывался при категоризации имеющихся списков. Выбор темы во многом мотивирован желанием автора вызвать интерес у читателя [Томашевский 1996: 177]. Стоит отметить, что это двустороннее общение: не только автор «пускает в ход некоторый набор норм» для создания определенного поля интерпретации текста потенциальным читателем, но и читатель привносит свой «багаж», то есть «комплекс социальных, исторических, культурных норм», влияющий на результат чтения и последующую интерпретацию [Компаньон 2001: 179].

Механизмы, с помощью которых происходит дешифровка художественной реальности, в свою очередь носят когнитивный характер [Stockwell 2002]. Интерпретация, таким образом, возможна исключительно благодаря активной работе воображения, которое неразрывно связано с пониманием текста через концептуальные метафоры. Отсюда и вырастает множественность значений, порождаемых текстов, поскольку результат интерпретации обуславливается не столько задумкой автора, сколько персональными категориями, выбор которых зависит от опыта читателя и его способности к эмпатии. По этой причине при выборе темы необходимо условиться о той аудитории, на которую интерпретация текста ориентирована.

2.3. Тема в компьютерной лингвистике

Одним из широко используемых подходов к автоматическому определению тематики коллекции текстов является тематическое моделирование [Daud et al. 2010: 280]. Согласно А.В. Коршунову и А.Г. Гомзину, под «темой» в рамках кластерного анализа понимается «результат би-кластеризации, то есть одновременной кластеризации и слов, и

документов по их семантической близости» [Коршунов, Гомзин 2012: 216]. На выходе модель дает вероятностное распределение языковых единиц, составляющих семантическое описание документа, и представляет их в виде вектора.

Все это отвечает конечной цели – тематической разметке корпуса и осуществлению в дальнейшем поиска в нем по параметру «тема», однако явно контрастирует с особенностями выделения тематики текста, обсуждавшимися выше. Стоит отметить, что финальной задачей тематического моделирования при таком подходе оказывается не столько порождение модели, сколько ее описание – определение темы, или, как принято говорить в отношении тематического моделирования, топика (англ. *topic*), на основании которого слова были объединены. Иными словами, исследователь должен «выяснить, с помощью каких скрытых структур вероятнее всего могли бы быть сгенерированы исходные документы» в результате работы алгоритма, и присвоить списку, если его интерпретация возможна, название [там же: 223].

Проблема интерпретируемости данных особенно остро встает при анализе литературных произведений, поскольку каждому произведению зачастую соответствует сразу несколько тем одновременно. Причем для описания художественных текстов имеют значение не только главные темы, но и так называемые «нишевые», часто игнорируемые алгоритмом при построении модели [O'Callaghan et al. 2015]. Более того, не всегда возможно определение темы художественного произведения в отрыве от его идейного содержания.

2.4. Тема как языковая категория

Во всех рассмотренных выше способах извлечения тематики произведения одним из главных свойств темы называют возможность разложения ее на составные элементы. Более того, немаловажной спецификой при формулировании темы является допущение о возможности темы быть свернутой до абстрактных языковых единиц и развернутой обратно – до самых подробных [Жолковский, Щеглов 2013]. Представление о теме как о суммирующем семантическом элементе роднит ее с понятием «категория», в особенности с той его интерпретацией, что принимается в когнитивной лингвистике. Существование базовых концептов в языке объясняется способностью человека категоризировать на самых разных уровнях обобщения [Fillmore 1982].

При этом важную роль в объяснении концептуальных категорий, существующих в языке, играет его метафоричность [Лакофф, Джонсон 2004; Lakoff 1987]. Человек постоянно пользуется метафорической понятийной системой, и она, по сути, проходит через всю его повседневную жизнь, определяет социальный и культурный опыт и находит конечное выражение в языке. Безусловно, одним главных источников большинства метафор становится художественная литература. Метафоры такого рода называются «образными», или «творческими» (англ. *novel metaphor*) [Лакофф, Джонсон 2004: 169].

Отличительной особенностью метафорических концептов, происходящих из литературного языка, является их «способность определять действительность» при помощи «связной сети следствий, высвечивающих одни свойства реальности и скрывающих другие» [там же: 187]. По этой причине именно контекст реализации метафоры во многом определяет то ее значение, которое оказывается релевантным в данный момент. Оценивать истинность метафоры предлагается исходя из подразумеваемых ею следствий и действий, которые проецируются на объективную реальность.

Связь с объектом отражения через установленную языком реальность сближает концептуальную метафору с тем пониманием темы художественного произведения, что признается рядом авторов в литературоведении. Следовательно, можно говорить о том, что тема, являясь содержательным ядром произведения, ведет себя точно так же, как языковая категория. Во-первых, будучи семантическим инвариантом, тема обладает свойством покрывать сразу несколько значений, причем не обязательно на основании наличия общих признаков, а значит, она способна и расширяться.

Во-вторых, тема, в зависимости от замысла автора, а также читательской интерпретации, «высвечивает» в конкретном, отдельно взятом произведении только часть тех значений, которые на самом деле в себя может включать. Данное наблюдение имеет решающее значение в рамках корпусного анализа, ставящего перед собой задачу не только выделения тем произведения, но и их каталогизации, и дает возможность объединять тематические составляющие разных произведений в одну категорию, общую для некоторой группы текстов.

3. Виды тематической разметки Корпуса русского рассказа

Начальный список тематических элементов был сформирован на материале аннотированного подкорпуса Корпуса русского рассказа первой трети XX века (Корпус-300), который содержит 310 текстов, написанных 300 авторами, общим объемом в 1 000 000 токенов [Мартыненко и др.

2018б: 180–197; Sherstinova, Martynenko 2020: 105–120]. Годы создания рассматриваемых художественных произведений приходятся на время социальных, политических и идеологических изменений в России (таких как, например, русско-японская война, Октябрьская и Февральская революции, Гражданская война). Для Корпуса были предложены два вида тематической разметки: автоматическая и ручная.

Автоматическое извлечение тематики произведений на основе неотрицательного матричного разложения осуществлялось Е.В. Замирайловой и О.А. Митрофановой на материале Корпуса русского рассказа первой трети XX века [Zamiraylova, Mitrofanova 2020]. Фокусом работы являлась динамика изменений распределения тем по трем временным периодам: 1900–1913 гг., 1914–1922 гг., 1923–1930 гг. Ниже приводятся некоторые из полученных кластеров [там же: 325–327]:

1. **Образ жизни, быт:**
   1) семья (*муж*, *ребенок*, *жена*, *мама*, *сестра*, *отец*, *дом*, *мальчик*, *кухня*);
   2) работа (*кабинет*, *деньги*);
   3) представители разных профессий (*купец*, *приказчик*, *извозчик*, *доктор*);
   4) «новая» жизнь (*товарищ*, *завод*, *гражданин*, *рабочий*);
   5) обычный образ жизни (*барин*, *старик*, *деревня*, *благородие*).
2. **Природа:** *пруд*, *река*, *ночь*, *солнце*, *куст*, *лес*, *волк*, *ветер*, *море*, *небо*, *ночь*, *берег*, *сосна*, *птица*, *зверь*, *лес*, *болото*.
3. **Военные, армия**: *солдат*, *офицер*, *пост*, *немец*, *немецкий*, *стрелять*, *рота*.
4. **Движение на поезде:** *вагон*, *пассажир*, *поезд*, *станция*, *ход*, *курс*.
5. **Деревня:** *батюшка*, *поп*, *церковь*, *Бог*, *святой*.

Для данной выборки рассказов из Корпуса русского рассказа первой трети XX века была также сделана экспертная, т. е. осуществленная вручную, тематическая разметка [Skrebtsova 2020]. По аналогии с компонентным анализом каждый рассказ рассматривалась как самостоятельная семантическая единица с неограниченным числом значений. Способ тематического тегирования нацелен на выявление всех семантических компонентов, которые способствуют развитию сюжета, связаны с действиями и мотивами героев и прямым образом влияют на конфликт произведения и его разрешение (см. подробнее [Sherstinova, Skrebtsova 2020]). При этом на один рассказ могло приходиться неограниченное количество тем. В соответствии с этой схемой было выделено 89 тем для 310 рассказов.

Неоспоримым преимуществом такого подхода является то, что он позволяет определить именно те темы, которые в большей мере соответствуют литературоведческому представлению об этой категории. Так, например, применяя классификацию А.Б. Есина [Есин 2000: 24], имеющиеся темы можно условно разделить на «вечные» и «конкретно-исторические». К первой категории можно отнести такие темы, как «любовь»; «смерть»; «дружба»; «искусство»; «предательство»; «серые будни, монотонная жизнь, быт»; «добро vs зло»; «христианский Бог»; «труд»; «духовное перерождение»; «сны vs явь»; «идеал vs реальность» и т. д. В основном они описывают межличностные отношения и внутренние конфликты, которые претерпевают герои произведений.

К конкретно-историческим, в свою очередь, будут отнесены следующие темы экспертной разметки: «русско-японская война»; «политическая борьба до революции»; «первая мировая война»; «октябрьская революция»; «новые порядки»; «еврейский вопрос» и т. д. Эти темы, как правило, подчеркивают социально-исторический подтекст, который имеют многие произведения в выборке, а также раскрывают связанные с этой тематикой сопутствующие явления, проявляющиеся на тематическом уровне: «казнь», «смерть от пуль», «эмиграция», «убийство на войне» и т. д.

Однако выполнение корпусной разметки в соответствии с данной методикой хотя и возможно, но крайне затруднительно: во-первых, ввиду длины списка, во-вторых, вследствие сильной раздробленности некоторых из выделяемых тем. По сути, это взаимосвязанные проблемы, т. к. учет большого количества частных случаев приводит к увеличению позиций в конечном списке тем, что в свою очередь затрудняет использование его при добавлении новых текстов. Оказывается проблематичным и манипулирование элементами имеющегося списка, т. к. они находятся на разных и трудно сопоставимых уровнях абстракции. По этой причине требуется применение еще одного, дополнительного способа систематизации тематических элементов – в данном случае предлагается их категоризация. Иными словами, предпринимается попытка свертки сюжета произведения до абстрактных лексических единиц.

4. Принципы нормализации тематической разметки

Нормализация исходных тем, предложенных экспертом [Skrebtsova 2020], осуществлялась в соответствии с особенностями выделения тематики художественного произведения, рассмотренными выше, а также принципами свертывания и развертывания тем [Жолковский, Щеглов 2013]. Движение от меньших тематических элементов к большим позволило вывести темы на один уровень абстракции и таким образом сделать список тем более однородным. Количество тематиче-

ских элементов оценивалось как критерий значимости выделения той или иной тематической категории. Кроме того, важно отметить, что при нормализации предлагалось ориентироваться на широкую публику, т. е. на массового читателя, иначе говоря, при выборе объединяющего тега предпринималась попытка усреднить восприятие художественного текста на историческом, культурном, географическом, а также субъективно-психологическом уровнях [там же].

Таким образом, задача состояла в объединении имеющихся подтем, где требуется, в одну категорию. При этом операция не сводилась до простого поиска тематических элементов, которые логически можно объединить в один кластер, – важно было также провести и различия между тематическими категориями. Полученное в результате логического обобщения семантическое ядро и становилось «тегом». Кроме того, учитывалась историческая специфичность вводимых автором тематик, т. к. во многом она указывает на ключевые особенности произведений исследуемого периода [Мартыненко и др. 2018а].

В соответствии с этими замечаниями были проанализированы уже выделенные с помощью автоматического и ручного анализов темы художественных текстов, вошедших в тестовую выборку из 310 рассказов Корпуса русского рассказа первой трети XX века. На основании сравнения данных списков предложен новый – список тегов. В качестве основы была выбрана экспертная разметка как содержавшая большее количество тематических элементов.

Всего было выделено 30 тегов: БУДУЩЕЕ, БЫТ, ВЗАИМООТНОШЕНИЯ, ВОЙНА, ГОРОД, ДЕНЬГИ, ДЕТИ, ДОБРОДЕТЕЛЬ, ДОСУГ, ИСКУССТВО, КРАСОТА, ЛЮБОВЬ, МЕЧТА, МОЛОДЕЖЬ, НАСИЛИЕ, ПОЛИТ_БОРЬБА, ПОРОКИ, ПРИРОДА, ПРОГРЕСС, ПСИХ_СОСТОЯНИЕ, РЕВОЛЮЦИЯ, РЕЛИГИЯ, СВОБОДА, СЕМЬЯ, СМЕРТЬ, СОН, СОЦ_ГРУППЫ, СОЦ_ПРОЦЕССЫ, ТРУД, ФАНТАСТИКА.

С помощью средств Microsoft Access была собрана тестовая версия базы данных русских рассказов первой трети XX века, размеченных в соответствии со списком тегов (Рис. 1).

Stories w/ Tags

StoryCode	story_name	author	year	TagNum	tag_name
S001	Курортный муж	Амфитеатров	1911	tag_3	ВЗАИМООТНОШЕНИЯ
S001	Курортный муж	Амфитеатров	1911	tag_26	СОН
S001	Курортный муж	Амфитеатров	1911	tag_24	СЕМЬЯ
S002	Рассказ о семи повешенных	Андреев	1908	tag_24	СЕМЬЯ
S002	Рассказ о семи повешенных	Андреев	1908	tag_15	НАСИЛИЕ
S002	Рассказ о семи повешенных	Андреев	1908	tag_25	СМЕРТЬ
S002	Рассказ о семи повешенных	Андреев	1908	tag_25	СМЕРТЬ
S002	Рассказ о семи повешенных	Андреев	1908	tag_16	ПОЛИТ_БОРЬБА
S002	Рассказ о семи повешенных	Андреев	1908	tag_3	ВЗАИМООТНОШЕНИЯ
S002	Рассказ о семи повешенных	Андреев	1908	tag_27	СОЦ_ГРУППЫ
S002	Рассказ о семи повешенных	Андреев	1908	tag_26	СОН
S002	Рассказ о семи повешенных	Андреев	1908	tag_20	ПСИХ_СОСТОЯНИЕ
S003	Своя минута	Анненкова-Бернар	1913	tag_3	ВЗАИМООТНОШЕНИЯ
S003	Своя минута	Анненкова-Бернар	1913	tag_6	ДЕНЬГИ
S003	Своя минута	Анненкова-Бернар	1913	tag_2	БЫТ
S003	Своя минута	Анненкова-Бернар	1913	tag_24	СЕМЬЯ
S003	Своя минута	Анненкова-Бернар	1913	tag_13	МЕЧТА
S003	Своя минута	Анненкова-Бернар	1913	tag_10	ИСКУССТВО
S004	В деревне	Арцыбашев	1906	tag_3	ВЗАИМООТНОШЕНИЯ
S004	В деревне	Арцыбашев	1906	tag_25	СМЕРТЬ
S004	В деревне	Арцыбашев	1906	tag_27	СОЦ_ГРУППЫ
S004	В деревне	Арцыбашев	1906	tag_28	СОЦ_ПРОЦЕССЫ
S004	В деревне	Арцыбашев	1906	tag_15	НАСИЛИЕ
S005	Занятые люди	Ауслендер	1912	tag_5	ГОРОД
S005	Занятые люди	Ауслендер	1912	tag_25	СМЕРТЬ
S005	Занятые люди	Ауслендер	1912	tag_12	ЛЮБОВЬ
S005	Занятые люди	Ауслендер	1912	tag_17	ПОРОКИ
S006	Прекрасное воспитание	Аверченко	1911	tag_24	СЕМЬЯ
S006	Прекрасное воспитание	Аверченко	1911	tag_7	ДЕТИ

Рисунок 1. Пример отображения параметра «тема» в базе данных Корпуса русского рассказа

Чтобы проиллюстрировать принцип объединения нескольких тематических элементов в один тег, рассмотрим в качестве примера тег СМЕРТЬ, в который были объединены шесть тем, выделенные экспертом: «неожиданная смерть», «естественная смерть», «смерть от эпидемии», «убийство», «смерть на войне», «казнь или расстрел», а также «самоубийство». Согласно Таблице 1, распределение элементов внутри этой категории может быть структурировано в зависимости от причины смерти: естественная, насильственная и самоубийство. К теме смерти также отнесены «мысли о смерти» и «страх перед смертью», которые ранее отмечались как сопутствующие.

Подобным образом анализировались и объединились в теги и другие темы, вошедшие в экспертный список [Skrebtsova 2020].

Таблица 1

Внутренняя структура тематической категории «смерть»

tag_25 СМЕРТЬ		
Естественная	**Насильственная**	**Самоубийство**
естественная смерть, смерть от эпидемии, неожиданная смерть (несчастный случай, обморожение и т. д.)	казнь, смерть на войне, расстрел, убийство	самоубийство, попытка самоубийства
мысли о смерти, страх перед смертью		

4.1. Количественный анализ нормализованных тематических тегов

Обратим внимание на то, что некоторые теги оказались более востребованными, чем другие. Наиболее частотными на протяжении всего временного периода, который охватывает рассказы, опубликованные с 1901 по 1930 гг., стали темы, маркированные тегами: ВЗАИМООТНОШЕНИЯ (115 рассказов), СМЕРТЬ (106 рассказов), ЛЮБОВЬ (103 рассказа). Лидерство этих категорий неудивительно, т. к. именно они относятся к там называемым «вечным» темам художественной литературы. Такие категории, как СВОБОДА (2 рассказа), КРАСОТА (2 рассказа) и РЕВОЛЮЦИЯ (3 рассказа), затрагиваются в слишком малом количестве рассказов, поэтому, возможно, их нужно объединить с другими категориями или же вовсе убрать.

Отдельно стоит прокомментировать сохранение некоторых, казалось бы, частных тематических элементов. Так, в список тегов был включен тег СОН, соответствующий экспертной теме «сны vs явь». Во-первых, это категория действительно обладающая определенной структурой. Исследователи выделяют несколько видов «литературного сна», которые при необходимости могут использоваться в качестве подтем: сон-кошмар, сон-предупреждение, сон-желание, сон-гротеск, сон-видение [Федунина 2013]. Во-вторых, выделение данной темы может быть полезно для исследования литературы в исторической перспективе, т. к. известно, что мотив сна занимают особое место в литературной традиции. Более того, отмечается, что сон в литературе XX в. претерпевает ряд изменений по сравнению с литературой XIX в. [там же].

При выделении категорий учитывался также и фактор времени создания текстов. В связи с этим было выделено сразу несколько тем, которые связаны с социальными и политическими изменениями: ПОЛИТ_БОРЬБА («политическая борьба до революции»; «террор»; «наказания за политические преступления – тюрьма, ссылка, каторга»), СОЦ_ПРОЦЕССЫ («предреволюционные волнения»; «новые порядки и социальные роли»; «старое vs новое»; «просвещение масс»; «женская эмансипация»; «эмиграция») и СОЦ_ГРУППЫ («казачество»; «жизнь крестьян»; «рабочий класс»; «межнациональные отношения и люди разных национальностей»; «еврейский вопрос»). В отдельные категории были объединены темы, связанные с войной и революцией: ВОЙНА («русско-японская война»; «Первая мировая война»; «гражданская война») и РЕВОЛЮЦИЯ («Октябрьская революция»).

Другими тегами, которые включили в себя только одну исходную тему (как и в случае с революцией), оказались следующие: БУДУЩЕЕ («светлое будущее»), БЫТ («серые будни, скука, монотонный быт»), ДОСУГ («встреча Рождества и Нового года, елка») и СВОБОДА («чувство свободы»). Однако не исключается возможность последующего расширения данных категорий, их конкретизации. В частности, тема свободы, так же, как и тема сна, является сквозной в истории литературы. В имеющейся разметке она выделяется в общем, но потенциально может быть разделена на несколько подкатегорий, например: «свобода внутренняя» и «свобода политическая».

4.2. Статистическая проверка зависимости количества тематических тегов от длины рассказа

При выделении тематики произведения часто предлагается обозначать одну или две главные темы [Есин 2000]. Однако определение только главных тем не только столь же затруднительно, как и определение, собственно, темы, но и препятствует получению полной картины о тематическом разнообразии анализируемой коллекции текстов. В этой связи статистически, путем анализа средних, проверяется, как распределились тексты в выборке Корпуса-300 в соответствии с количеством выделенных в них тем.

Количество тем в исходной экспертной разметке варьирует от 1 до 13. Количество тегов в свою очередь представлено в диапазоне от 1 до 9.

Как известно, частота и распределение языковых элементов в тексте напрямую связаны с теми значениями, которые текст порождает [Tognini-Bonelli 2001: 177]. Предположительно, то же верно и по отношению к теме: число языковых элементов, из которых текст состоит, должно соотноситься с числом выделяемых в нем тем. Таким образом, количество тем предлагается рассматривать как фактор, зависящий от длины текста.

Для того чтобы проверить гипотезу, рассказы были токенизированы средствами пакета tidytext для R [Silge, Robinson 2017]. Сначала рассматривалось наличие главного эффекта – количество тематических тегов, приходящееся на среднюю длину соответствующих текстов. Согласно Рисунку 3, можно сделать предположение о наличии зависимости полученного количества тегов от длины рассказа. Так, рассказы длиной от 2000 токенов до 4000 токенов характеризуются сравнительно небольшим количеством тегов – их насчитывается до пяти; далее же тематика рассказов растет пропорционально их объему.

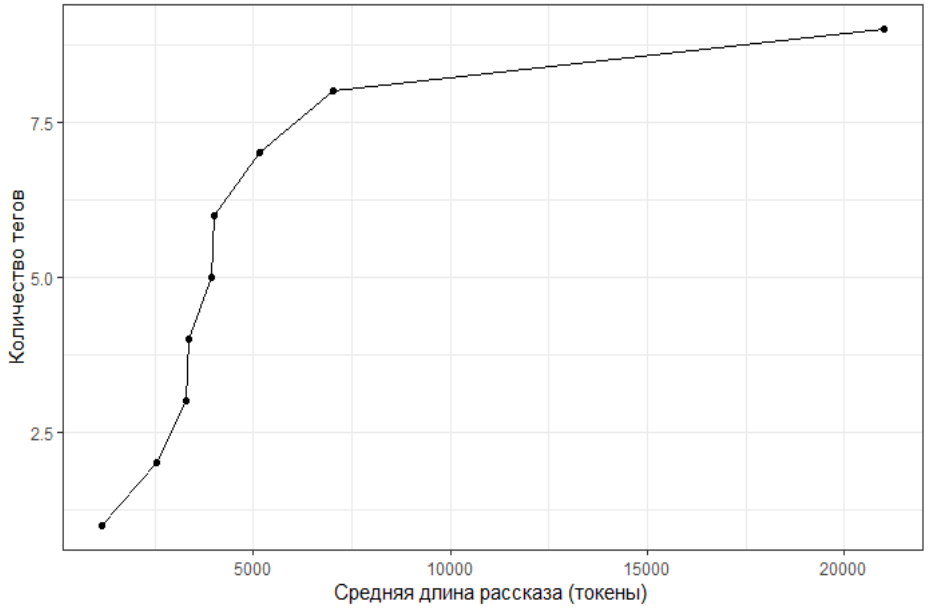

Рисунок 2. Взаимосвязь между количеством тегов и средней длиной рассказа

Результаты регрессионного анализа зависимости количества тегов от средней длины рассказов, которые они описывают подтвердили статистическую значимость ($p\text{-value} < 0.001$) этой зависимости (длина текста рассчитывалась аналогично – в токенах).

4.3. Сравнение автоматической и экспертной тематической разметки

Полученные результаты стандартизированной разметки могут использоваться для валидации результатов компьютерного моделирования тематики русского рассказа [Sherstinova et al. 2020]. В результате сопоставления экспертных тематических тегов с выделенными автоматически тематическими распределениями были сделаны следующие наблюдения. Наибольшее соответствие топиков и тем выявлено для тегов ВОЙНА, ПРИРОДА, СЕМЬЯ, РЕЛИГИЯ, в то время как проблематичными для автоматического определения оказались теги СМЕРТЬ, ВЗАИМООТНОШЕНИЯ, СОЦИАЛЬНЫЕ ГРУППЫ. Таким образом, лучше всего и, что важно, приближенно к человеческой оценке выделяются те тематические элементы, которые составляют фон действия произведения. Темы, которые необязательно могут эксплицитно развертываться в рассказе, – например, тема «социальные группы», при раскрытии которой в центре повествования будет скорее описание какого-то конфликта, нежели описание самой социальной группы – сложнее выявить и интерпретировать исходя из результатов тематического моделирования.

5. Тематическая разметка Корпуса русского рассказа XX века: предварительные результаты

Полученный список тегов и сформулированные принципы выделения тематики предлагается использовать и при увеличении уровней тематической разметки. Обсудим предварительные результаты тематического аннотирования Корпуса русского рассказа XX века – цифрового литературного ресурса, представляющего собой расширение Корпуса русского рассказа 1900–1930 гг. и разрабатываемого в настоящее время [Шерстинова, Кирина, Хлусова 2023].

На данный момент аннотированная выборка рассказов 1931–2000 гг. включает в себя 642 рассказа, для которых экспертным путем было выделено более 2 000 тематических элементов. Из них выделяется 600 уникальных тем, при этом около 400 не встречались ранее, например: «космос», «внеземные технологии», «русско-чеченская война», «холодная война», «репрессии», «сталинское время», «советско-финляндская война», «жизнь в колхозе», «совхоз», «бандитизм», «садоводство» и др.

Наиболее частотной темой «нового периода», ожидаемо, оказалась конкретно-историческая тема – «Вторая мировая война» (см. Рис. 3).

Рисунок 3. Частотные тематические элементы в русском рассказе 1931–2000 гг. (до нормализации)

При этом любопытно, что рассказы, связанные с этой тематикой, получая наибольшее распространение в 40-е гг., продолжают встречаться и в послевоенный период (Рис. 4). Вероятно, это связано с публикационной «отложенностью», характеризующей исследуемый временной отрезок.

Стабильно встречаются в течение всего рассматриваемого периода так называемые «общечеловеческие» темы: «любовь», «дружба», «брак», «дети», «труд». Данное наблюдение подтверждает тенденцию, замеченную при анализе тематической наполненности рассказов первой трети XX в. Темы, так или иначе отображающие взаимоотношения между людьми, пользуются популярностью в художественной литературе. Любопытно, что распределения некоторых семантически похожих тематических тегов приближены друг к другу (Рис. 5, 6).

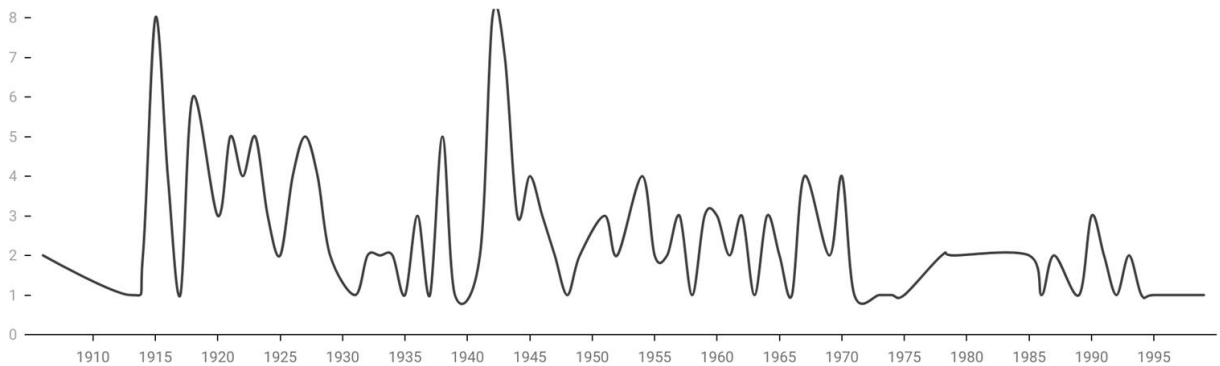

Рисунок 4. Распределение тематического тега ВОЙНА по годам (по данным первичной нормализации)

Рисунок 5. Распределение тематического тега СЕМЬЯ по годам (по данным первичной нормализации)

Рисунок 6. Распределение тематического тега ЛЮБОВЬ по годам (по данным первичной нормализации)

Интересно проследить, как изменится тематическая картина русского рассказа при завершении аннотирования всей выборки текстов и насколько увеличится процент уникальных тегов.

6. Заключение

В результате исследования были изучены основные способы, на основе которых возможно выделение темы художественного произведения. Методы извлечения тематики, рассмотренные выше, хотя и отличаются по инструментарию, объединены одной общей идеей: тема обладает свойством разложения на составные элементы. Причем художественная литература, в отличие от других текстов, например, академических, представляет случай более сложной тематической организации.

Применение того или иного из предложенных вариантов разложения темы зависит от задач исследователя. В настоящей работе предлагаются новые по отношению к анализу литературного текста принципы формулирования тем – по аналогии с языковыми категориями. Определение темы художественного произведения является проблематичным главным образом потому, что сформулировать одну и ту же тему можно по-разному, и это различие мотивировано сразу несколькими факторами. Во-первых, тема обладает свойством сворачиваться до разных уровней абстракции. Во-вторых, выбор темы во многом зависит от отдельного читателя. Обсуждаемый подход предполагает определенную системность в формулировании темы. Так, в сочетании с выделением конкретно-исторических тем тегирование, ориентированное на усредненное, или «общечеловеческое», читательское восприятие позволяет сделать аннотацию не только более универсальной и применимой для междисциплинарных исследований, но и доступной для пользователей разного уровня.

Полученный список тематических тегов для Корпуса русского рассказа первой трети XX века можно считать первой версией каталога тем художественных произведений рассматриваемого периода. Ориентация на экспертную разметку при его составлении позволила учесть большее количество тем, в особенности тех, представление о которых можно получить только с опорой на читательское восприятие. Однако увеличение исследовательского материала, в нашем случае расширение Корпуса русского рассказа на весь XX в., ставит необходимость адаптации списка тегов для работы с прежде не выделявшимися тематическими элементами.

Появление структурированного набора данных о темах рассказов всего XX в. может открыть новые перспективы для исследования

«произведения как на части, выражения чего-то более значительного, чем самый текст: <…> психологического момента или общественной ситуации» [Лотман 1996]. Кроме того, станет возможным изучение влияния внетекстологических факторов на сюжеты и, соответственно, тематику произведений для моделирования социокультурных процессов, происходивших в российском обществе на протяжении XX в.

В дальнейшем планируется нормализовать разметку не только по темам, но и, где необходимо, по подтемам, а также автоматизировать данный процесс с использованием результатов тематического моделирования и экспертной аннотации [Sherstinova, Moskvina, Kirina 2021]. Помимо этого, представляется целесообразным рассмотреть в контексте количественного анализа промежуточный этап свертывания текста – сжатый пересказ, находящийся между двумя крайними этапами свертывания – исходным произведением и его темой. Это должно позволить получить более точное представление о специфике данного процесса и определить оптимальное число выделяемых тем для рассказа в зависимости от его длины.

Благодарности

Публикация подготовлена в рамках работы по проекту «Русская литература в социальном измерении: компьютерная платформа СОЦИОЛИТ», поддержанного Научным фондом НИУ ВШЭ, в 2023 г.

Список источников

Корпус русского рассказа 1900–1930 гг. [Электронный ресурс]. URL: https://russian-short-stories.ru (дата обращения: 01.10.2023).

Список литературы

Есин А.Б. Принципы и приемы анализа литературного произведения. М.: Флинта, Наука, 2000. С. 22–26.

Жолковский А.К., Щеглов Ю.К. К понятиям «тема» и «поэтический мир» // Щеглов Ю.К. Избранные труды / сост. А.К. Жолковский, В.А. Щеглова. М.: РГГУ, 2013. С. 37–78.

Компаньон А. Демон теории. М.: Изд-во им. Сабашниковых, 2001. 336 с.

Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. № 23. С. 215–242.

Лакофф Д., Джонсон М. Метафоры, которыми мы живем / под ред. и с предисл. А.Н. Баранова. М.: Едиториал УРСС, 2004. 256 с.

Лотман Ю.М. О поэтах и поэзии. СПб.: Искусство-СПБ, 1996. 846 с.

Мартыненко Г.Я. и др. Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) / Г.Я. Мартыненко, А.Г. Мельник, Т.Ю. Шерстинова, Т.Ю. Попова // Компьютерная лингвистика и вычислительные онтологии. 2018а, №2. С. 97–102.

Мартыненко Г.Я. и др. О принципах создания корпуса русского рассказа первой трети XX века / Г.Я. Мартыненко, Т.Ю. Шерстинова, А.Г. Мельник, Е.В. Замирайлова // Труды XV Международной конференции по компьютерной и когнитивной лингвистике «TEL 2018». Казань, 2018б. С. 180–197.

Томашевский Б.В. Теория литературы. Поэтика. М.: Аспект Пресс, 1996. 334 с.

Федунина О.В. Поэтика сна. М.: Intrada, 2013. 196 с.

Шерстинова Т.Ю., Кирина М.А., Хлусова Я.К. Корпус русского рассказа как база для проведения социолингвистических исследований русской литературы // Информационные технологии в гуманитарных исследованиях: матер. науч. конф. Красноярск, 2023 (в печати).

Daud A. et al. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou, F. Muhammad // Frontiers of Computer Science in China. 2010. Vol. 4. Pp. 280–301.

Fillmore C. Towards a Descriptive Framework for Spatial Deixis. London: John Wiley, 1982. Pp. 31–59.

Lakoff G. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind? Chicago: University of Chicago, 1987. 613 p.

O'Callaghan D. et al. An Analysis of the Coherence of Descriptors in Topic Modeling / D. O'Callaghan, D. Greene, J. Carthy, P. Cunningham // Expert Systems with Applications. 2015. Vol. 42(13). Pp. 5645–5657.

Sherstinova T. et al. Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction / T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, M. Kirina // Lecture Notes in Computer Science. 2020. Vol. 12469. Advances in Computational Intelligence: MICAI 2020. Pp. 134–151.

Sherstinova T., Martynenko G. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century // CEUR Workshop Proceedings, 2020. Vol. 2552: R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Pp. 105–120. [Электронный ресурс]. URL: http://ceur-ws.org/Vol-2552 (дата доступа: 01.10.2023).

Sherstinova T., Moskvina A., Kirina M. Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian // 29th Conference of Open Innovations Association (FRUCT). IEEE, 2021. Pp. 313–323.

Sherstinova T., Skrebtsova T. Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930 // CEUR Workshop Proceedings. 2020. Vol. 2813: Proceedings of the International Conference "Internet and Modern Society" (IMS-2020). Pp. 117–128.

Silge J., Robinson D. Text mining with R: A tidy approach. E-book, O'Reilly Media, Inc., 2017. 194 p.

Skrebtsova T.G. Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories // CEUR Workshop Proccedings. 2020. Vol. 2813: Proceedings of the International Conference "Internet and Modern Society" (IMS-2020). Pp. 265–276.

Stockwell P. Cognitive Poetics: An Introduction. London: Routledge, 2002. 193 p.

Tognini-Bonelli E. Corpus Linguistics at Work. Amsterdam: John Benjamins Publishing Company, 2001. 224 p.

ON THE PRINCIPLES OF NORMALIZATION OF THEMATIC ANNOTATION IN THE CORPUS OF RUSSIAN SHORT STORIES OF THE 20TH CENTURY

Margarita A. Kirina
Lecturer, Department of Philology
Junior Research Fellow, Linguistic Convergence Laboratory
National Research University Higher School of Economics – St Petersburg

The article discusses the problem of normalization of the thematic annotation of the Corpus of Russian Short Stories of the 20-th century. The aim of the research was to develop a methodology that combines linguistic and literary approaches to text analysis, in order to standardize the "theme" parameter, identified by expert. The study proposes to consider the theme of literary works as a socio-cultural phenomenon, and discusses the prospects for studying the influence of extra-textual factors on the thematic diversity of texts of a certain historical period.

Keywords: Russian short stories; literary theme; corpus linguistics; thematic annotation; normalization.