УДК 81’322.2

МИР РУССКОГО РАССКАЗА СВОЗЬ ПРИЗМУ СОВРЕМЕННЫХ ЦИФРОВЫХ ТЕХНОЛОГИЙ

Татьяна Юрьевна Шерстинова
к. филол. н., доцент департамента филологии
Национальный исследовательский университет
«Высшая школа экономики» – Санкт-Петербург
Санкт-Петербург, 190068, наб. канала Грибоедова, д. 119–121. tsherstinova@hse.ru

В статье представлены результаты междисциплинарного исследования русской малой прозы цифровыми методами компьютерной лингвистики, digital humanities, а также современными методами искусственного интеллекта, связанными с обработкой текстовых данных. Исследование выполнено в рамках одного жанра – русского рассказа, который представляет собой наиболее распространенный жанр прозы. Рассматриваются тексты, написанные в 1900–1930 гг., причем изучаемая эпоха представляется как последовательность трех хронологических временных срезов. Представлены основные выводы, полученные в результате исследования тематики и персонажей малой русской прозы, а также ее эмоциональной окраски.

Ключевые слова: русский рассказ; тематика; персонажи; тональность; корпусная лингвистика; digital humanities; количественные методы.

В статье представлены результаты междисциплинарного исследования русского рассказа первых трех десятилетий XX в. с помощью современных компьютерных технологий – методов компьютерной лингвистики, digital humanities, а также современных методов искусственного интеллекта, связанных с обработкой текстовых данных. Исследование выполнено в рамках жанра рассказа, который представляет собой наиболее распространенный жанр прозы, и осуществлено на материале аннотированной выборки Корпуса русского рассказа 1900 – 1930 гг. [Корпус русского рассказа: электр. ресурс]. Этот электронный ресурс создавался специально для проведения лингвостатистических исследований прозы [Мартыненко и др. 2018а, 2018б], а также для моделирования «литературно-художественной системы» рассматриваемой эпохи [Мартыненко 1988, 2019; Тынянов 1929]. Такой подход подразумевает включение в исследование не только текстов всем известных классиков, но и литераторов «второго эшелона», а также малоизвестных и уже фактически забытых писателей, которые публиковались в 1900 – 1930 гг. [Sherstinova, Martynenko 2020].

Изучаемый период был выбран не случайно. Первые тридцать лет XX века оказались для нашей страны поистине драматическими, эта эпоха была насыщена войнами и революциями, в результате которых произошла смена социального строя государства. Очевидно, произошедшие в обществе изменения не могли не повлиять на литературный процесс, а также на тематику и эмоциональную составляющую художественных произведений. Представим некоторые выводы, полученные в результате проведенных в этом направлении исследований, выполненных на материале аннотированной части Корпуса русского рассказа 1900 – 1930 гг. (Корпус-300).

Корпус-300 состоит из 310 произведений 300 русских писателей, для большинства авторов он содержит по одному рассказу, выбранных рандомно [Sherstinova, Martynenko, 2020]. Все множество рассказов делится на три хронологических последовательных периода по дате их написания[1]:

1) довоенный период начала XX в. (1900 – 1913 гг.),
2) военно-революционный период (1914 – 1922 гг.),
3) раннесоветский период (1923 – 1930 гг.).

Такой подход позволяет проводить исследование в диахронии и изучать корреляцию языка, стиля, тематики и других аспектов литературных текстов в зависимости от происходящих в момент их создания исторических событий.

Можно предположить, что именно тематика художественных произведений должна меняться больше всего в результате значимых социально-исторических преобразований, поэтому первый аспект исследования посвящен анализу тем. Эмпирическое исследование тем русского рассказа стало возможным благодаря тематической разметке Корпуса русского рассказа, разработанной Т.Г. Скребцовой [Skrebtsova 2020] и осуществ-

ленной группой экспертов-филологов для всего Корпуса-300. В процессе тематического аннотирования каждый текст размечался вручную одним экспертом, в задачу которого входило приписать каждому рассказу после его прочтения одну или несколько тем из предложенного списка.

Так как список тем для разметки оказался достаточно обширным (около 90), то для проведения статистического анализа и для представления на сайте [Корпус русского рассказа: электр. ресурс] он был нормализован. Для этого было выделено 30 обобщающих тем: 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 [Кирина 2020; Sherstinova, Kirina 2022]. Дистрибуция тем, выделенных экспертами за весь рассматриваемый период, представлена на рисунке 1 [Кирина 2020]. Можно видеть, что наиболее частотными темами русской прозы того времени являются взаимоотношения между людьми, среди которых отдельно выделятся любовь, а также тема смерти. Незначительно уступают им по частоте темы, связанные с жизнью общества (различные социальные группы, семья и социальные процессы). Достаточно значимыми для писателей являлись также темы денег, войны и человеческих пороков.

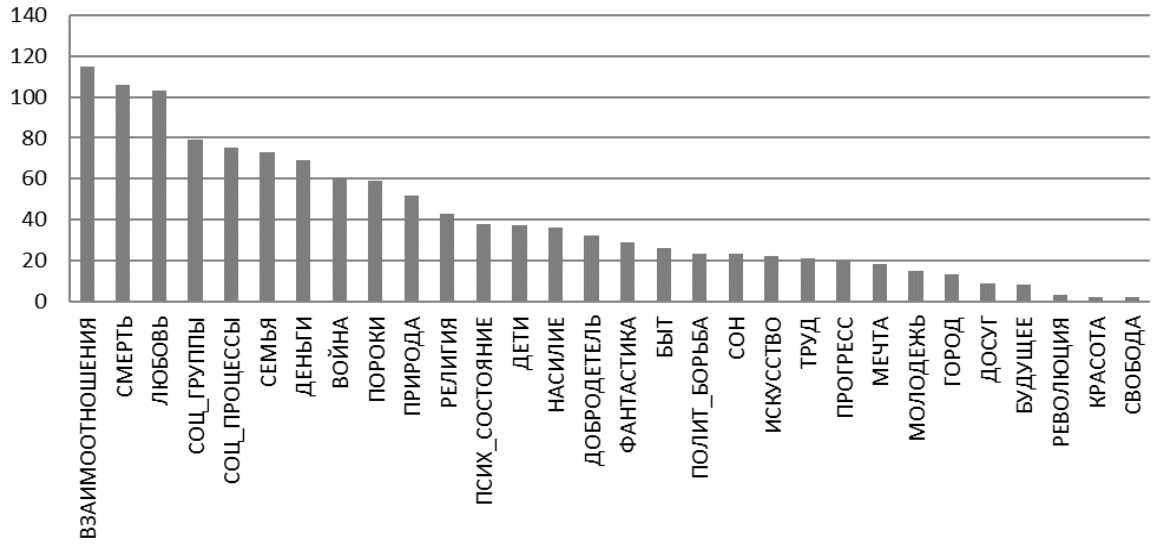

Рисунок 1. Дистрибуция тем русского рассказа 1900 – 1930 по результатам экспертной тематической разметки

Распределение данных экспертной тематической разметки по хронологическим периодам позволили увидеть динамику тематического разнообразия русской малой прозы, подробно описанную в [Скребцова 2020; Sherstinova, Skrebtsova 2020]. Как и следовало ожидать, можно говорить о том, что в целом тематика литературных произведений «идет в ногу со временем», так как в большинстве текстов находят свое отражение актуальные для соответствующего временного периода вопросы, например: смерть на войне; нищета, голод, разруха; новый социальный порядок; освоение новых земель и т. д. Можно отметить также и несколько отсроченное («большое видится на расстоянии») отражение в литературных произведениях важных для общества событий: так, тема Гражданской войны чаще всплывает в литературных произведениях раннесоветской прозы 1923 – 1930 гг., чем в военные 1918 – 1922 гг. [Скребцова 2020].

Наиболее значимые отличия между темами разных временных периодов следующие. Доля рассказов, посвященной такой важной для довоенной прозы темы как частная жизнь героев (романтическая любовь, брак, супружеская измена; дети; крах мечты; ложные надежды; разочарование) постепенно уменьшается в военно-революционный и раннесоветский периоды. Можно предположить, что интерес к личной жизни и внутренним переживаниям «более типичны для мирного благополучного времени, чем для напряженных драматичных периодов в истории страны, будь то война или послевоенная разруха. В тяжелые времена они вытесняются актуальной общественно-политической повесткой» [там же: 57].

В период острых социальных катаклизмов (война, революция) наблюдается всплеск тем, связанных с обращением к Богу, к духовным ценностям, с одной стороны, и увеличение количества рассказов, сюжеты которых строятся на

мистических явлениях, видениях и предчувствиях – с другой стороны. Повышение интереса писателей к «сверхъестественной» тематике можно объяснить психологической потребностью обращения человека «к высшим силам» в тяжелые для жизни времена. В советский период, по известным причинам, проявление этих тем в прозе существенно спадает.

Что касается раннесоветской литературы, здесь появляются новые, не свойственные дореволюционным рассказам темы, такие как технический прогресс, массовое образование, эмансипация женщин, исследования и изобретения, новый образ жизни, сопоставление старого и нового жизненного уклада. Много рассказов посвящено описанию сельской жизни, в которой также произошли разительные перемены. На смену романтической любви, характерной для начала века, появляются тексты про любовь сексуальную, также можно отметить интерес советских писателей того времени и к другим «физиологическим» проявлениям жизни. Достаточно много рассказов, особенно в период «нового времени», посвящены также теме насилия [Sherstinova, Skrebtsova 2020].

В наше время для тематической кластеризации больших текстовых данных используются методы тематического моделирования (см., напр. [Митрофанова 2014]). Нужно сразу оговориться, что их основное предназначение состоит в выявлении тематических классов для специальных текстов, однако есть удачные примеры использования тематического моделирования и на текстах художественной литературы [Митрофанова 2019; Rhody 2012; Schöch 2017; Uglanova, Gius 2020]. Результаты применения алгоритмов тематического моделирования к аннотированной части Корпуса русского рассказа приведены в [Zamiraylova, Mitrofanova 2020] и [Кирина 2022]. Отдельное исследование посвящено тематической кластеризации рассказов, относящихся, согласно экспертной оценке, теме насилия [Gryaznova, Kirina 2021].

В Таблице 1 представлен список топиков, автоматически выделенных на материале Корпуса-300 с помощью метода NMF [Zamiraylova, Mitrofanova 2020]. Интересно, что, по полученным данным, по периодам различаются не только темы, но и количество автоматически выделяемых топиков: если для двух относительно «мирных» периодов (довоенного и раннесоветского) отмечается разнообразие тем (по 10 на каждый период), то в период 1914 – 1922 гг. их количество резко снижается до четырех. Можно отметить и некоторые общие темы хронологически смежных периодов – это тема «письма» в первый и второй периоды и тема «войны» во второй и третий.

Таблица 1

Результаты тематического моделирования Корпуса-300 русского рассказа с помощью метода NMF

Начало 20 века: 1900 – 1913 гг.	**Первая мировая война и революции: 1914 – 1922 гг.**	**Ранний советский период: 1923 – 1930 гг.**
Тема 1. ПИСЬМО	Тема 11. ПИСЬМО-2	Тема 15. ВСТРЕЧА В КОМНАТЕ
Тема 2. ТЮРЬМА	Тема 12. ВОЙНА	Тема 16. ДЕРЕВНЯ
Тема 3. НАВЕДЕНИЕ ПОРЯДКА В ДЕРЕВНЕ	Тема 13. ВЕРА	Тема 17. ЧЕЛОВЕК В ЛЕСУ
Тема 4. ДОМ В ЛЕСУ	Тема 14. ЛЮБОВЬ НА ПРИРОДЕ	Тема 18. ВОЙНА-2
Тема 5. СЕМЕЙНОЕ СЧАСТЬЕ		Тема 9. РАБОТА НА ЗАВОДЕ
Тема 6. ТОЛПА		Тема 20. ОХОТА
Тема 7. РОМАНТИЧЕСКОЕ СВИДАНИЕ		Тема 21. ЛЮДИ НА БОЛОТАХ
Тема 8. СВЯЩЕННИК С ПРИХОЖАНАМИ В ДЕРЕВНЕ		Тема 22. ПРАЗДНИК В ДЕРЕВНЕ
Тема 9. ЖИЗНЬ В ГОРОДЕ		Тема 23. ПУТЕШЕСТВИЕ НА ПОЕЗДЕ
Тема 10. ОБЕД		Тема 24. МАТЬ И ЕЕ РЕБЕНОК

Сопоставление экспертной и автоматической тематической разметки представлено в [Sherstinova et al. 2020], однако корреляция между темами экспертной разметки и топиками, полученными в результате автоматического тематического моделирования, оказалась весьма невысокой. По нашим наблюдениям, в целом тематическая кластеризация намного лучше отражает общий фон произведения (в частности, место действия рассказа – например, приморский город, монастырь или театр), чем особенности драматургии сюжета [там же]. Другим важным аспектом мира литературных произведений являются его персонажи. Наполнение базы данных героев русского рассказа 1900 – 1930 гг. было выполнено студентами-филологами, обучающимися в НИУ ВШЭ в Санкт-Петербурге. Согласно инструкции, размечались все персонажи, которые произносят хотя бы одну реплику, при этом в базу данных заносилась информация об имени, поле, возрастной группе, социальном происхождении, семейном положении, профессии и других характеристиках, которые можно было выделить, а также о том, является ли персонаж главным или второстепенным героем. В результате было получен список из 2 190 персонажей и посчитана стати-

стика распределения социальных характеристик героев в зависимости от года написания рассказа[2] [Иванова 2021].

Каждый рассказ в среднем содержит 7,06 «говорящих» персонажей. При этом среднее значение несколько колеблется от периода к периоду: 7,12 в довоенную эпоху, минимально (6,23) в военно-революционную и максимально (7,94) в советский период.

Оказалось, что персонажи мужского пола встречаются примерно в три раза чаще, чем женского (1609 против 576 соответственно[3]), причем эта доля также варьирует в зависимости от периода: если для «мирных» периодов отношение количества мужских персонажей к женским очень близко (2,66 и 2,70), то в военно-революционную эпоху оно увеличивается до 3,07.

Распределение персонажей по возрастным группам оказалось на удивление похожим для разных временных периодов (см. Табл. 2), особенно поражает совпадение показателей для среднего возраста, что наводит на мысль о существовании некоего инварианта этих пропорций в прозе. Кажется интересным также сравнить их с реальной демографической картиной общества того времени. Впрочем, здесь стоит отметить, что при осуществлении ручной разметки «средний возраст» указывался при разметке по умолчанию, когда не было оснований отнести персонажа к молодежной или пожилой группе, что может объяснять более высокие доли этой возрастной группы в выборке, но никак не стабильность этой доли в разные эпохи.

Что касается незначительных отличий между периодами, то можно выделить несколько более высокий процент детей и подростков в довоенный период, молодежи – в военно-революционный, и «ветеранов» – в советский.

Таблица 2

Распределение героев русского рассказа по возрастным группам

Период	Дети, %	Подростки, %	Молодежь, %	Средний возраст, %	Пожилые, %	Всего, чел.
1900 – 1913	3,23	4,49	20,65	59,13	12,08	712
1914 – 1922	2,27	2,58	22,88	59,24	11,36	660
1923 – 1930	2,69	3,06	20,17	59,78	14,18	818
В целом	**2,74**	**3,38**	**21,14**	**59,41**	**12,65**	**2190**

Распределение персонажей по их социальному статусу в обществе рассматривалось по трем категориям: высокий, средний, низкий. В целом для всей выборки они покрывают 12,74%, 45,98% и 40,64% соответственно. Однако с точки зрения хронологической динамики, наблюдается очевидная и объяснимая тенденция: доля персонажей с высоким социальным статусом резко падает от 1-го периода к 3-му (21,63% → 14,55% → 3,55), в то время как доли героев среднего и, особенно, низкого социального статуса растут (42,69% → 45,30% → 49,39% и 35,53% → 38,33% → 46,94% соответственно). Такая «гегемония» персонажей невысокого социального статуса является вполне закономерной для молодого государства «рабочих и крестьян».

Наконец, с точки зрения рода деятельности и профессий героев русской малой прозы, которые оказалось возможным определить для 57,58% всех персонажей, самыми распространенными оказались следующие: военнослужащие (11,69%), рабочие (8,95%), работники конторы (3,52%), студенты разных специальностей (3,29%) и прислуга (2,88%). Дистрибуцию долей этих профессий по периодам можно проследить по данным Таблицы 3 – эти цифры хорошо коррелируют как с исторической ситуацией, так и с основными тематическими тенденциями малой прозы, описанными выше. Так, налицо повышение доли военнослужащих и духовенства в период войн и революций, доли рабочих – в советский период, в то время как доли прислуги и полицейских, как «пережитки» прошлой дореволюционной России, в советский период существенно сокращаются.

Таблица 3

Распределение самых частотных профессий персонажей русского рассказа, %

Период	Профессия или род занятий							
	военный	рабочий	работник конторы	студент	прислуга	медицина	духовенство	полиция
1900 – 1913	5,34	6,88	4,63	4,49	4,92	3,09	1,83	3,65
1914 – 1922	18,94	6,06	2,88	3,94	3,03	2,58	4,55	1,21
1923 – 1930	11,37	13,08	3,06	1,71	0,98	2,20	1,47	1,59
В целом	**11,69**	**8,95**	**3,52**	**3,29**	**2,88**	**2,60**	**2,51**	**2,15**

Главные герои составляют в среднем 21,83% от всех литературных персонажей. Доля главных героев несколько выше в военно-революционном периоде (25,75%), для довоенного и раннесоветского периодов составляет 20,37% и 19,93% соответственно. На один рассказ приходится в

среднем 1,55 главных героев, причем эта доля в разных периодах не сильно отличается: минимум наблюдается для довоенного времени (1,45), а максимум – для советского (1,60).

Если рассматривать только главных героев литературных произведений, то в 50% случаев это люди среднего возраста, 31,78% относятся к молодежной группе, 10,88% – пожилые, 4,39% – подростки и 2,30% – дети[4]. Женщины составляют меньше четверти всех главных героев (в среднем – 22,39%), мужчины – 77,20%, а 0,42% приходится на оккультных персонажей, не обладающих категорией пола.

По социальному статусу и профессиональной принадлежности в анализируемых временных периодах среди главных героев можно проследить те же тенденции, которые были получены в целом для всех персонажей: это увеличение доли героев низкого и среднего социального происхождения (32,41% → 31,76% → 44,79% и 39,31% → 46,47% → 52,76% соответственно) для счет резкого уменьшения доли героев высокого статуса (28,28% → 20,00% → 2,45%). Среди самых «востребованных» литераторами профессий также оказываются военные (12,13%) и рабочие (9,21%); 6,49% всех главных героев составляют студенты. Четвертый по частоте ранг неожиданно занимают «осужденные» или «преступники» (4,19%), при этом главных геров, «сидящих за решеткой в темнице сырой», особенно много (9,66%) в русской прозе довоенного имперского периода, что еще раз подчеркивает социальный раскол дореволюционной России.

В заключительной части статьи хочется сказать несколько слов и о тональности, или эмоциональной окраске, русского рассказа. Есть разные подходы к измерению тональности, основанные как на правилах и словарях, так и на машинном обучении [Прикладная и компьютерная лингвистика 2016]. В исследовании Ю.С. Созонтовой [Созонтова 2020] был применен метод с использованием словаря эмоционально-окрашенных слов для русского языка «РуСентиЛекс» [Лукашевич, Левчик 2017], содержащий 12 тыс. слов и выражений, каждое из которых охарактеризовано с точки зрения тональности: позитивная (positive), негативная (negative), нейтральная (neutral) или неопределеная оценка, которая зависит от контекста (positive/negative).

В отличие от описанных выше исследований, второй, военно-революционный, период в свою очередь поделен на три подпериода:

1) 1914 – 1916 гг. – начало Первой мировой войны,
2) 1917 – 1918 гг. – революционный период,
3) 1919 – 1922 гг. – Гражданская война,

что позволяет более точно рассмотреть специфику употребления эмоционально-окрашенных слов в непростой период острых социальных катаклизмов.

В Таблице 4 приведены основные статистические данные, связанные с анализом тональности исследуемой прозы: 1) общее количество всех слов для каждого исторического периода; 2) общее количество эмоционально-окрашенных слов; 3) количество уникальных тональных слов для каждого периода; 4) доля эмоционально-окрашенной лексики в текстах (отношение количества тональных слов к общему количеству слов); 5) количество слов с негативной тональностью; 6) количество слов с позитивной тональностью; 7) собственно тональность, определяемая как отношение количества негативно-окрашенных слов к количеству позитивно-окрашенных слов; 8) показатель разнообразия тональных слов, посчитанный как количество уникальных эмоционально-окрашенных слов, поделенное на общее количество эмоционально окрашенных слов (чем ближе к единице этот показатель, тем разнообразнее эмоционально-окрашенная лексика в текстах анализируемого периода, а чем меньше его значение, тем чаще тональные слова повторяются) [Созонтова 2020].

Таблица 4

Эмоциональная окраска русского рассказа 1900 – 1930 гг.

Период	Всего слов	Эмоц. окраш. слов	Уник.	Доля эмоц. окраш. слов	Негативные	Позитивные	Тональность	Разнообразие
1900 – 1913	607526	29915	3163	0,050 (5,0%)	18526	10035	**1,85**	0,106
1914 – 1916	238010	11044	2034	0,046 (4,6%)	6461	4106	**1,57**	0,184
1917 – 1918	44898	2185	769	0,049 (4,9%)	1547	563	**2,75**	0,356
1919 – 1922	147521	6151	1606	0,041 (4,1%)	4018	1773	**2,27**	0,261
1923 – 1930	465281	19068	2982	0,040 (4,0%)	12999	5297	**2,45**	0,156

Полученные данные показывают, что максимальное значение тональности (2,75) приходится на революционные 1917 – 1918 гг., в которых отрицательно-окрашенная лексика превышает положительно-окрашенную почти в три раза. Напрашивается интересная гипотеза о доле эмоционально окрашенных слов в русской прозе: по данным Табл. 4, средний объем слов ненулевой тональности варьирует от 4% до 5%, что наводит на мысль о существовании некоего инварианта частоты упо-

требления эмоциональной лексики в прозе. Однако эта гипотеза требует дополнительной проверки.

Таким образом, серия проведенных квантитативных исследований подтверждает исходные предположения о том, что исторический фон, во время которого создаются литературные произведения, находит свое отражение в тематике, эмоциональном настроении и персонажах малой прозы. Кажется целесообразным продолжение подобных исследований – как в рамках изучения прозы той же исторической поры, подключая большее количество авторов и большее число произведений от одного автора, так и в рамках расширения выработанной методологии на другие временные периоды (например, на рассказы всего XX в.). Но если для методов автоматической обработки данных (тематического моделирования, анализа тональности) это относительно легко осуществимо, то для ручной экспертной разметки обработка больших массивов текстов все еще остается сложнореализуемой за счет своей высокой трудоемкости. Возможно, методы машинного обучения с учителем позволят преодолеть этот барьер в обозримом будущем и позволят получить более достоверное квантитативное описание литературного мира для значимых исторических эпох.

Благодарности

Создание Корпуса русского рассказа 1900 – 1930 было осуществлено в 2018 – 2020 гг. при поддержке Российского фонда фундаментальных исследований, проект № 17-29-09173 офи_м «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

Публикация подготовлена в результате проведения исследования по проекту № 21-04-053 «Методы искусственного интеллекта для филологических исследований» в рамках Программы «Научный фонд Национального исследовательского университета “Высшая школа экономики” (НИУ ВШЭ)» в 2021 – 2022 гг.

Примечания

[1] Если дата написания текста неизвестна, то отнесение рассказа к конкретному периоду осуществлялось по дате первой публикации.

[2] Статистика по персонажам посчитана для 309 рассказов Корпуса-300.

[3] Для четырех персонажей из выборки пол определить невозможно.

[4] В выборку Корпуса-300 включались только рассказы для взрослых, детская литература не рассматривалась.

Список литературы

Иванова О.Ю. Дистрибуция персонажей малой русской прозы (на материале рассказов 1900 – 1930 гг.): вып. квалиф. работа бакалавра / НИУ ВШЭ. СПб., 2021. 52 с.

Кирина М.А. О тематической компоненте базы данных русского рассказа // Тр. XVI Междунар. Конф. по компьютерной и когнитивной лингвистике «TEL 2020». Казань, 2020. (в печати)

Кирина М.А. Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник Новосибирского государственного университета. Лингвистика и межкультурная коммуникация. 2022. № 20(2). С. 93–109.

Корпус русского рассказа 1900 – 1930 гг. [Электронный ресурс]. URL: https://russian-short-stories.ru (дата обращения: 18.10.2022).

Лукашевич Н.В., Левчик А.В. Словарь оценочных слов русского языка. [Электронный ресурс]. URL: labinform.ru/pub/rusentilex/rusentilex_2017.txt (дата обращения: 28.10.2022).

Мартыненко Г.Я. Основы стилеметрии. Л.: Изд-во ЛГУ, 1988. 176 с.

Мартыненко Г.Я. Методы математической лингвистики в стилистических исследованиях. СПб.: Нестор-История, 2019. 296 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, Т.Ю. Шерстинова, А.Г. Мельник, Т.И. Попова // Компьютерная лингвистика и вычислительные онтологии. Вып. 2: Тр. XXI Междунар. объединенной конф. «Интернет и современное общество, IMS-2018 / Университет ИТМО. СПб., 2018а. С. 97–102.

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

Митрофанова О.А. Моделирование тематики специальных текстов на основе алгоритма LDA // Тр. XLII Междунар. филологической конф. / С.-Петерб. гос. ун-т. СПб., 2014. С. 220–233.

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. СПб., 2019. С. 387–394.

Прикладная и компьютерная лингвистика / под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. М.: ЛЕНАНД, 2016. 320 с.

Скребцова Т.Г. Динамика тем русских рассказов начала XX века // Философия и гуманитарные науки в информационном обществе. 2020. № 3. С. 45–60.

Созонтова Ю.С. Анализ тональности русской художественной литературы на материале Корпуса русского рассказа первой трети XX века: вып. квалификационая работа бакалавра / НИУ ВШЭ. СПб., 2021. 54 с.

Тынянов Ю.Н. Архаисты и новаторы. Л.: Прибой, 1929. 598 с.

Gryaznova E., Kirina M. Defining Kinds of Violence: A Comparison of Topic Modelling with Latent Dirichlet Allocation and Principal Component Analysis for Russian Short Stories of 1900 – 1930 // Proceedings of International Conference "Internet and Modern Society". 2021. P. 281–290.

Rhody L.M. Topic Modelling and Figurative Language // Journal of Digital Humanities. 2012. Vol. 2(1). P. 19–35.

Schöch C. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama // Digital Humanities Quarterly. 2017. Vol. 11, No. 2. [Электронный ресурс]. URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (дата обращения: 28.10.2022).

Sherstinova T. et al. Topic Modelling with NMF vs. Expert Topic Annotation: the Case Study of Russian Fiction / T. Sherstinova, O. Mitrofanova, T. Skrebtsova, T. Zamiraylova, M. Kirina // Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. Mexico City, 2020. LNAI 12469. P. 134–151.

Sherstinova T., Kirina M. Normalization Issues in Digital Literary Studies: Spelling, Literary Themes and Biographical Description of Writers // Communications in Computer and Information Science, CCIS 1503. 2022. Pp. 332–346.

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 // R. Piotrowski's Readings in Language Engineering and Applied Linguistics: Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). CEUR Workshop Proceedings. 2020. Vol. 2552. P. 105–120.

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 // Proceedings of the International Conference "Internet and Modern Society" IMS-2020. CEUR Workshop Proceedings. 2020. Vol. 2813. P. 117–128.

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

Uglanova I., Gius E. The Order of Things: A Study on Topic Modelling of Literary Texts // Proceedings of the CHR 2020: Workshop on Computational Humanities Research. CEUR Workshop Proceedings. 2020. P. 57–76.

Zamiraylova E., Mitrofanova O. Dynamic Topic Modeling of Russian Prose of the First Third of the XXth Century by Means of Non-Negative Matrix Factorization // R. Piotrowski's Readings in Language Engineering and Applied Linguistics: Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). CEUR Workshop Proceedings. 2020. Vol. 2552. P. 321–339.

THE WORLD OF RUSSIAN SHORT STORIES FROM THE PERSPECTIVE OF MODERN DIGITAL TECHNOLOGIES

Tatiana Yu. Sherstinova
Associate Professor, the Philological Department
National Research University Higher School of Economics – St. Petersburg

The article presents the results of the interdisciplinary study of the Russian short prose by methods of computational linguistics, digital humanities, as well as by up-to-date methods of artificial intelligence used for text data processing. The study was carried out within a Russian short story genre as the most common prose genre. Literary texts written in the 1900 – 1930s are considered, the studied era being presented as a sequence of three chronological time periods. The main conclusions made while studying the main themes and characters of the Russian short prose, as well as its sentiment, are presented.

Keywords: Russian short story; themes; characters; sentiment; corpus linguistics; digital humanities; quantitative methods.