

Litera

Правильная ссылка на статью:

Северина Е.М., Ларионова М.Ч. — Проект Chekhov Digital: задачи и проблемы реализации семантической разметки текстов (на примере рассказа А. П. Чехова «Смерть чиновника») // Litera. – 2023. – № 10. DOI: 10.25136/2409-8698.2023.10.68862 EDN: IHSMSE URL: https://nbpublish.com/library_read_article.php?id=68862

Проект Chekhov Digital: задачи и проблемы реализации семантической разметки текстов (на примере рассказа А. П. Чехова «Смерть чиновника»)

Северина Елена Михайловна

ORCID: 0000-0001-6518-2771

доктор философских наук

профессор, Институт филологии, журналистики и межкультурной коммуникации, Южный федеральный университет

344006, Россия, г. Ростов-На-Дону, пер. Университетский, 93, каб. 5

✉ emkovalenko@sfedu.ru



Ларионова Марина Ченгаровна

ORCID: 0000-0002-2955-2621

доктор филологических наук

зав. отделом гуманитарных исследований, Южный научный центр РАН

344006, Россия, г. Ростов-На-Дону, ул. Чехова, 41

✉ chengarovna@yandex.ru



[Статья из рубрики "Интерпретация"](#)

DOI:

10.25136/2409-8698.2023.10.68862

EDN:

IHSMSE

Дата направления статьи в редакцию:

01-11-2023

Аннотация: В статье рассматривается модель подготовки машиночитаемой (семантической) разметки текстов для проекта Chekhov Digital на примере филологической интерпретации отдельных значимых элементов рассказа А. П. Чехова «Смерть чиновника» и представления этих сведений в явном виде с опорой на

стандарты цифровой публикации Text Encoding Initiative (TEI/XML). Выявлены значимые сущности для разметки корпуса текстов писателя, однако вопрос их репрезентации в тексте остается достаточно сложным. Проведена филологическая экспертиза таких аспектов, как «свойства, состояния и события; особенности характера» в отрывке из рассказа А.П. Чехова с точки зрения возможностей разметки TEI по сохранению филологических знаний в машиночитаемом формате. Одна из задач проекта Chekhov Digital – выйти за рамки простого оцифрованного текста и предусмотреть полезные для исследователя цифровые инструменты. Представлены элементы машиночитаемой разметки, позволяющие разметить значимые сущности в чеховских текстах для организации семантического поиска по корпусу текстов писателя, рассмотрены проблемы и исследовательские задачи, возникающие в процессе реализации такого рода междисциплинарных проектов в связи с необходимостью объединения усилий специалистов из разных областей знания. В рамках проекта реализуется принцип открытых исследовательских данных (Open Data), важнейшей задачей которого является создание научных сообществ вокруг данных. Работа над проектом привела к развитию научного сотрудничества между исследователями НИУ ВШЭ, ЮНЦ РАН и ЮФУ.

Ключевые слова:

проект Chekhov Digital, цифровое издание, Чехов, филологическая экспертиза, TEI, машиночитаемая разметка, семантический поиск, экспертная разметка, технологии автора, автоматическая обработка текста

Разработка проекта, разметка TEI подготовлены в рамках реализации проекта «Зеркальные лаборатории» НИУ ВШЭ, № 6.13.1-02/250821-1, тема «Конвергенция языковых пластов русского языка в зеркале цифровых решений».

Литературоведческий анализ и экспертная разметка подготовлены в рамках реализации ГЗ ЮНЦ РАН, № гр. проекта 122020100347-2.

Проект Chekhov Digital представляет собой семантическое издание Полного собрания сочинений и писем А. П. Чехова в 30 томах (далее I ПССиП)^[1], разрабатываемое в формате стандарта цифровой публикации Text Encoding Initiative (TEI/XML)^[2]. Академическое издание текстов произведений и писем писателя^[1], включая ранние редакции и различные варианты текста, состоит из двух серий: тексты произведений писателя I Сочинения (тт. I-XVIII) и Письма (тт. I-XII), для которых разрабатывается семантическая машиночитаемая разметка. Семантическое издание ПССиП позволяет включить тексты писателя в цифровой культурный контекст, открывает новые возможности для проведения академических исследований в цифровом формате и использования литературных текстов в цифровых проектах и приложениях^[3]. Разработка проекта осуществляется Центром цифровых гуманитарных исследований Института филологии журналистики и межкультурной коммуникации Южного федерального университета (ИФЖИМКК ЮФУ) совместно с Международной лабораторией языковой конвергенции НИУ ВШЭ и Отделом гуманитарных исследований Южного научного центра Российской академии наук ЮНЦ РАН.

В рамках проекта Chekhov Digital разрабатывается цифровой ресурс, который включает в себя не только структурную разметку с опорой на стандарт цифровой публикации TEI/XML, но и разметку значимых сущностей в чеховских текстах, примечаниях и комментариях, что позволяет сделать документы машиночитаемыми и разработать

инструменты достаточно сложного семантического поиска информации в текстах произведений и писем писателя. В системе Text Encoding Initiative существует обширный список тегов для разметки самой разнообразной информации, в том числе внешней по отношению к тексту, – от специфики представления информации в разных типах текста (пьеса, рассказ, повесть или письмо) до имен собственных и таких социальных категорий, как социальный статус, профессиональная принадлежность и т.п. Такого рода информация создает трудности для технологий автоматической обработки текстов, поскольку может быть представлена разными способами, но описанная явно с помощью разметки TEI уменьшает сложность обработки с помощью компьютерных методов. При этом любая разметка – это форма интерпретации текста, поэтому цифровой формат представляет собой некоторый вариант интерпретации чеховских текстов. Размеченные таким образом тексты размещаются в открытом доступе на сайте проекта (<http://chekhov-digital.sfedu.ru/>), на котором дополнительно представлены некоторые цифровые исследовательские инструменты, такие как семантический поиск, инструменты визуализации и т. п..

С опорой на работы исследователей-литературоведов были выявлены значимые сущности для разметки корпуса текстов писателя – это «имена, названия; даты; времена года; цвета; свойства, состояния и события; особенности характера; природные явления; социальный статус; профессия; животные / растения; комментарии; примечания» [4]. При этом вопрос репрезентации этих сущностей в тексте остается достаточно сложным и требует разработки. Какие-то смыслы имеют вполне отчетливые лексические формы репрезентации, однако такие аспекты как «свойства, состояния и события; особенности характера» требуют рассмотрения конкретных текстов как с точки зрения филологического подхода, так и с точки зрения возможностей разметки TEI по сохранению филологических знаний в машиночитаемом формате. В качестве примера такого рода исследования мы рассматриваем отрывок из рассказа А.П. Чехова «Смерть чиновника»:

«В один прекрасный вечер не менее прекрасный экзекутор, Иван Дмитрич Червяков, сидел во втором ряду кресел и глядел в бинокль на "Корневильские колокола". Он глядел и чувствовал себя на верху блаженства. Но вдруг... В рассказах часто встречается это «но вдруг». Авторы правы: жизнь так полна внезапностей! Но вдруг лицо его поморщилось, глаза подкатились, дыхание остановилось... он отвел от глаз бинокль, нагнулся и... апчихи!!! Чихнул, как видите. Чихать никому и нигде не возбраняется. Чихают и мужики, и полицеймейстеры, и иногда даже и тайные советники. Все чихают. Червяков несколько не сконфузился, утерся платочком и, как вежливый человек, поглядел вокруг себя: не беспокоил ли он кого-нибудь своим чиханьем? Но тут уж пришлось сконфузиться. Он увидел, что старичок, сидевший впереди него, в первом ряду кресел, старательно вытирал свою лысину и шею перчаткой и бормотал что-то. В старичке Червяков узнал статского генерала Бризжалова, служащего по ведомству путей сообщения» [5, 164].

Слово «один» в первом предложении использовано в качестве неопределенного местоимения (некий, какой-то), призванного подчеркнуть в контексте рассказа типичность ситуации. Очевидно, что это не первый выход героя в театр, следовательно, местоимение становится маркером свойства, характеристики ситуации. Словосочетание «прекрасный вечер» является маркером, скорее свойства субъективного восприятия, чем объективной характеристикой времени суток; так его воспринимает Червяков. Выражение «не менее прекрасный экзекутор» представляет собой с семантической точки зрения, сложную конструкцию. Это снова самооценка героя, при этом сочетание

«прекрасного вечера» с «прекрасным экзекутором» выражает авторскую иронию: «экзекутор» в сочетании с двумя «прекрасными» образует контраст, поскольку слово «экзекутор», с одной стороны, означает невысокий социальный статус («экзекутор м. лат. I (исполнитель) чиновник при канцелярии или присутственном месте, на котором лежат полицейские и хозяйственные обязанности» [6, 662]), а с другой стороны – родственно «экзекуции», то есть наказанию, казни.

Уже на этом примере видно, что стандартная морфологическая разметка покажет только специфику языковых форм, поэтому для разметки семантики требуется совершенно другой подход, в котором будут учтены соответствующие маркеры I свойства, характеристики ситуации (типичность); свойства субъективного восприятия – времени и профессии/социального статуса; авторской иронии, построенной по принципу контраста: профессия – прекрасная, по самооценке героя, но дающая невысокий социальный статус, и намекающая на возможность наказания.

«Иван Дмитрич Червяков» – не просто имя, 1) «Иван» отражает *национальный стереотип* (В. Г. Короленко вспоминал слова Чехова о герое пьесы «Иванов»: «Иван Иванович Иванов. Понимаете? Ивановых тысячи ... обыкновеннейший человек, совсем не герой...» [7, 143]); 2) «Дмитрич» разговорной формой подчеркивает *незначительность социального статуса*; 3) «Червяков» заранее, еще до совершения основного действия, предсказывает *раболепие и чинопочитание героя*. Имя становится *свойством*, характеризует героя как ничтожного, мелкого человека.

Однако, и это все еще первое предложение рассказа, Червяков сидит во втором ряду кресел, прямо за генералом. Сочетания сущностей (действие, место, состояние) образуют контраст между социальной незначительностью героя и его высокой самооценкой – место приобретает значение *свойства характера*. Он ощущает себя значительным лицом. Вспомним, что он «экзекутор», то есть обладает властью над нижестоящими, но одновременно «Червяков», зависимый от вышестоящих, «одной стороной бытия такое лицо всегда в ряду бессловесно трепещущих, другой – в числе безапелляционно вершащих суд» [8, 27-28]. Экзекутор Червяков – это, таким образом, контекстуальный оксюморон, а прекрасный экзекутор Червяков – оксюморон, возведенный в степень.

Использование специализированных алгоритмов извлечения именованных сущностей (Named Entity Recognition, NER) для русского языка (например, библиотека SlovNet, <https://github.com/natasha/slovnet>) дает возможность автоматически разметить в текстах фамилию, имя, отчество героя, но информация о маркерах *национального стереотипа, социального статуса и характера* героя должна быть внесена дополнительно. Например, с помощью алгоритмов NER из текста извлечены имена персонажа, для разметки которых могут быть использованы следующие теги: `<person xml:id="Червяков" subtype="personage"> <persName full_name="Червяков Иван Дмитриевич"> <forename type="first"> И в а н </forename> <forename type="patronym"> Д м и т р и ч </forename> <surname> Ч е р в я к о в </forename> </person>`.

Разметка имени персонажа позволяет обращаться к его описанию в любом месте текста с использованием атрибута `ref="#Червяков"` или `xml:id="Червяков"`, при этом описание может включать и внешние по отношению к тексту сведения, например, социально-экономический статус персонажа маркируется тегом `<socecStatus>`; статус/состояние I тегом `<state>`; занятость/профессия I `<occupation>`; события I `<event>`; а с помощью тега `<trait>` может быть указано свойство, в котором отмечен как характер героя

(например, «ничтожный»), так и стереотипность его имени. Причем эти теги с соответствующими атрибутами могут быть использованы как в описании героя, помечая его универсальные характеристики (<trait>) на протяжении всего текста, так и в описываемой в тексте конкретной ситуации/событии (<state>), что значительно уменьшает сложность автоматической обработки текста. Контекстуальный оксюморон также может быть некоторым универсальным образом помечен в тексте, например, тегом, который используется для сегментации текста и классификации элементов: <seg type = "оксюморон"> прекрасный экзекутор, <person xml:id="Червяков" > Иван Дмитрич Червяков</person></seg>. Разметка текста – это всегда некоторая форма его интерпретации, поэтому мы полагаем, что в цифровой публикации литературных текстов очень важно опираться на филологическую экспертизу.

Но первое предложение еще не окончено. Червяков «глядел в бинокль на "Корневильские колокола"». Обычно спектакль «смотрят», оперу «слушают», однако использование глагола «глядел» в сочетании с предлогом «на» и названием комической оперы вместо слов «пьеса», «спектакль» и пр., организует пространство по принципу «сверху вниз». То есть Червяков, который находится «на верху блаженства», будто с высоты взирает на «Корневильские колокола», располагается метафорически над ними, что подчеркивает не столько его наслаждение представлением, как может показаться, сколько его амбициозность. Так простое действие – «глядел на» (сущность-действие) в совокупности с названием комической оперы становится характерологическим свойством героя | амбициозность (сущность-свойство), для разметки которого можно использовать тег <trait type = "character" key="амбициозный">.

Примечателен здесь и выбор пьесы, на которую «глядел» Червяков. Сюжет комической оперы Р. Планкета связан с темой социального возвышения: героиня оперы в финале оказывается дочерью прежнего владельца замка Корневиль. Ее приемный отец – богат и скряга. Так Чехов намекает на тайные мечты своего героя. А для исследователя это еще один аргумент в пользу того, что в литературном произведении, особенно чеховском, нет ничего случайного, всякая деталь функциональна.

Таким образом, первые предложения могут быть размечены следующим образом с использованием описанных выше тегов: «В один прекрасный вечер не менее <seg type = "оксюморон"> прекрасный экзекутор, <trait type = "character" key="ничтожный"> <person xml:id="Червяков"> Иван Дмитрич Червяков</person> </trait>, </seg> <trait ref="#Червяков" type = "character" key="амбициозный"> сидел во втором ряду кресел и глядел в бинокль на <name xml:id="Q959242"> "Корневильские колокола" </name>. Он глядел и чувствовал себя на верху блаженства </trait>».

В данной разметке использован идентификатор (ID) из БД Викиданные, в которой Q959242 является ID записи о комической опере Робера Планкетта «Les cloches de Corneville (Q959242)» (URL: <https://www.wikidata.org/wiki/Q959242>). Такой формализованный подход дает возможность уже в самом начале текста описать основные характеристики героя – его ничтожность и одновременно амбициозность, которые и определяют драматический финал.

Следующее предложение начинается противительным союзом «но». Встает вопрос, отмечать ли при разметке служебные слова. Разбираемый случай свидетельствует, что делать это необходимо. Первый абзац рассказа построен по принципу антитезы: отсутствие события – событие, наслаждение и самолюбование – падение, к чему читателя готовит предшествующий текст. Чем «выше», тем «ниже», от «экзекутора» к «Червякову». Еще один пример, как сущность-место превращается в сущность-свойство.

«В рассказах часто встречается это "но вдруг". Авторы правы: жизнь так полна внезапностей!» Эти два предложения усиливают антитезу, подчеркивая неожиданность, спонтанность события. Но они имеют и другие художественные функции: 1) маркируют жанр (рассказ – это повествование о событии, как правило, необычном, «внезапном», выходящем за рамки повседневности), 2) апеллируя к многочисленным авторам рассказов, указывают на жанровое клише, 3) вводят в рассказ фигуру автора-повествователя, сигнализируют о смене точки зрения, благодаря чему теперь Червяков из субъекта превращается в объект, неслучайно Чехов избирает в качестве завязки сюжета именно чихание – действие, которое почти невозможно контролировать. Актером становится не сам герой, а его отдельные части: «лицо поморщилось», «глаза подкатились», «дыхание остановилось» (совершенно гоголевский прием, ср. повесть Н. В. Гоголя «Нос»). Глаголы несовершенного вида сменяются глаголами совершенного вида, что демонстрирует быстроту смены состояний. Этот ряд завершается звукоподражанием «апчи!!!», которое вообще не имеет лица, рода и пр.: действие (сущность), оно же событие (сущность), совершается само по себе, без участия деятеля. А герой совершает действия, направленные на разрушение образа, созданного первой частью антитезы: он отводит от глаз бинокль и наклоняется, таким образом осуществляя движение (сущность-действие) сверху вниз, что одновременно является реализацией метафоры «смотреть свысока, сверху вниз» (сущность-особенность характера).

Автоматическая морфосинтаксическая разметка текста позволяет получить токены (словоформы), леммы (исходные формы), теги частей речи и грамматических признаков для каждой леммы, а также дерево синтаксических отношений с вершинами и зависимыми (полный список отношений UD v.2 и их определений см. <https://universaldependencies.org/u/dep/index.html>). Для разметки использована программа автоматического морфологического и синтаксического анализа UDPipe 2 [9]. В нашем исследовании использована модель UDPipe, обученная на данных синтаксического корпуса зависимостей UD-SynTagRus 2.6 для русского языка (SynTagRus, [10]). Этот корпус в настоящее время является стандартным русскоязычным набором данных для обучения современных нейросетевых парсеров (UDpipe, Stanford NLP, Turku NLP, DeepPavlov и др.). При этом некоторые конструкционные особенности синтаксиса можно использовать для семантической разметки. Например, в предложениях «Но вдруг...», «В рассказах часто встречается это но вдруг» конструкция «но вдруг» размечается как «ADV (root/conj) + CCONJ (cc)», выполняя функцию согласования (соединения), причем «но» ставится всегда в зависимое положение от «вдруг» [11], а в предложении «Но вдруг лицо его поморщилось глаза подкатились дыхание остановилось... он отвел от глаз бинокль нагнулся и... апчи» эта конструкция распадается и каждое слово становится зависимым от глагола «поморщилось», образуя предикативные связи: «VERB (root) + CCONJ (cc)» | конструкция согласования (поморщилось + но), и «VERB (root) + ADV (advmod)» | конструкция, меняющая значение глагола (поморщилось + вдруг). Такого рода изменения в использовании конструкций можно рассматривать как маркеры в изменении семантики, однако для использования в алгоритме автоматической разметки это требует дальнейшего изучения. В свою очередь с помощью автоматически полученных тегов частей речи и грамматических признаков для каждой леммы может быть размечена закономерность смены состояний, когда глаголы несовершенного вида сменяются глаголами совершенного вида. Но художественные функции и метафора, о которых речь шла выше, не могут быть размечены автоматически, а следовательно, должны быть размечены экспертно, например: <seg type = "metaphor"> он отвел от глаз бинокль нагнулся и </seg>. Разметка может быть дополнена конкретным значением метафоры, но даже в таком

общем виде облегчит поиск метафор в тексте. Звукоподражание может быть отмечено в тексте следующим образом `<vocal who="#Червяков"> апчихи!!! </vocal>`, что позволит организовать поиск по такого рода элементам.

Далее в повествование опять вступает рассказчик, именно ему принадлежит комментарий о спонтанности, случайности чихания: *«Чихнул, как видите. Чихать никому и нигде не возбраняется. Чихают и мужики, и полицеймейстеры, и иногда даже и тайные советники. Все чихают»*. Глагол *«не возбраняется»* с отрицательной частицей в безличном предложении, где под действующим лицом подразумевается некая надличная сила или стихия, как полагал А. А. Потебня^[12], контрастно коррелирует с *«экзекутором»*, который *«бранит»*, *«возбраняет»*. Чихание соотносится исключительно с физиологией, а не с социальным статусом, что Чехов подчеркивает перечислением социальных ролей и усилительной частицей *«даже»*. Однако при внимательном прочтении оказывается, что социальная тема прорывается во фрагменте *«и иногда даже и тайные советники»*, где сочетание *«чихания»* с чином *«тайный (советник)»* дезавуирует мысль об общедоступности чихания, поскольку тайные советники и чихают тайно, что подчеркивается наречием *«иногда»* (сущность-время) и усилительной частицей *«даже»* (сущность-степень действия) – ирония. Такой прием напоминает вышивание стебельчатым швом: новый стежок начинается с середины предыдущего. Так и у Чехова: одна тема начинается *«внутри»* другой и каждый новый *«стежок»* захватывает обе темы.

Это видно и в следующем примере: *«Червяков нисколько не сконфузился, утерся платочком и, как вежливый человек, поглядел вокруг себя: не обеспокоил ли он кого-нибудь своим чиханьем?»* *«Нисколько не сконфузился»* отрицательно коррелирует с глаголом *«утерся»*, поскольку он обозначает не только *«вытерся»*, но и *«получил что-то неприятное, оскорбительное»*, ср. фразеологизм *«утереть нос»*, то есть *«проявить превосходство»*. Условно говоря, физиология проявила превосходство над самоощущением героя, его внутренней *«высотой»*, он вынужден метафорически *«утереться»*. То есть все-таки Червяков сконфузился, потому и проверяет, не обеспокоил ли кого-нибудь чиханием. Характеристика *«вежливый человек»* в таком случае звучит иронически. А следующее предложение: *«Но тут уж пришлось сконфузиться»*, – логически, *«стебельчатым швом»*, вытекает из предыдущих, чему способствует и повтор глагола *«skonфузиться»* (действие-состояние), теперь уже в прямом, не в отрицательном значении, то есть Червяков рухнул с пьедестала, на который сам себя возвел.

С помощью языка разметки TEI мы можем как отметить звукоподражательные элементы, так и социальный статус чихающих, что позволит облегчить поиск такого рода элементов: `<<vocal who="#Червяков"> Чихнул </vocal>, как видите. <vocal> Чихать </vocal> никому и нигде не возбраняется. <vocal> Чихают </vocal> и <socecStatus type="мужик"> мужики </socecStatus>, и <socecStatus type="полицмейстер"> полицеймейстеры </socecStatus>, и иногда даже и <socecStatus type="тайный советник"> тайные советники </socecStatus>. Все <vocal> чихают </vocal>»`.

При этом вопрос о необходимости сохранять в формате разметки все элементы филологического анализа остается открытым, например: метафору *«утерся платочком»* следует разметить для поиска с помощью тега `<seg type="metaphor">`, но стоит ли размечать ироническую характеристику персонажа *«вежливый человек»*, вводя дополнительные теги, тем более, что возникает неоднозначность в определении границ контекста, в котором ирония будет распознаваема при автоматическом поиске.

Глагол *«увидел»* в следующем предложении соотносится с начальным *«глядел на (Корневильские колокола)»*. Но теперь это простое действие растерянного человека.

Точка зрения опять перешла к герою. Он осознает, что впереди, в первом ряду (отсылка ко второму ряду Червякова), сидит «старичок» (не «старик», именно «старичок» в глазах сконфузившегося Червякова), при этом генерал, и он не «утирается», а «вытирает» лысину и шею и бормочет что-то, что Червяков принимает на свой счет, то есть опять метафорически «утирается», еще больше «конфузится». Фамилия генерала тоже говорящая, «ассоциируется с глаголами "брызгать", "брезговать" или "брюзжать", "докучать", "надоедать"» [13, 36]. То есть генерал не так безобиден, как может показаться [см. 14]. Можно предположить, что название ведомства, в котором служит генерал, – «путей сообщения» – избрано Чеховым не случайно, в нем, кроме прямого, можно усмотреть переносный смысл: пути сообщения – это пути коммуникации. Именно эта коммуникация в рассказе не состоится.

Такого рода анализ показал, что необходимо обращать внимание на изучение глагольных форм и нестандартные формы существительных («старичок»), характеристики которых могут быть автоматически извлечены из текста с помощью морфосинтаксического парсера (см. выше) и размечены автоматически. Фамилия генерала может быть извлечена из текста с помощью алгоритмов NER и размечено с учетом его социального статуса «статский генерал» : `<person xml:id="Бризжалов" subtype="personage"> <persName> Бризжалов </persName> <socecStatus type="статский генерал"> статский генерал </socecStatus> </person>`. В этом случае к описанию персонажа в тексте можно обращаться с использованием атрибута `ref="#Бризжалов"` или `xml:id="Бризжалов"`.

Таким образом, первый абзац является в полном смысле завязкой рассказа: в нем намечены основная сюжетная линия, характеристики героев и предсказывается драматический финал. Этот фрагмент пронизан внутренними связями, перекличками, подобными тем, которые В. Б. Катаев назвал «драматическими рифмами», образующими «единое резонирующее пространство» [15, 3-4]. Чихание в рассказе оказывается не только физиологическим, но и социальным актом. «Пристальное прочтение» даже одного фрагмента выявляет художественную «технология» писателя, в том числе и природу присущего его рассказам лаконизма.

Сложность семантической разметки определяется многими аспектами, в том числе «технологией» автора, к которой можно отнести и наличие переносных значений у слов и конструкций; и невозможность автоматического распознавания иронии, которая не имеет вербальных способов выражения; и «сложносоставность» художественных приемов, когда ирония дополняется метафорой, контекстуальным оксюмороном и пр.; и прием «стебельчатого шва», когда одна тема начинается «внутри» другой и каждый новый «стежок» захватывает обе темы; и т. д.. Кроме того, сама «технология» обработки имеет сложности, связанные с автоматическим выделением сущностей в тексте, ограничением формальных приемов их разметки и самого языка разметки, а также необходимостью привлечения экспертов на разных этапах разметки, что предполагает наличие у исследователей специализированных знаний в области цифровой филологии.

Разметка текста – это всегда некоторая форма его интерпретации, поэтому мы полагаем, что в цифровой публикации литературных текстов очень важно опираться на филологическую экспертизу, с помощью которой могут быть выделены соответствующие категории для разметки. Например, наше исследование показало, что необходимо классифицировать и размечать такие художественные приемы в тексте, как метафора или антитеза и оксюморон, выявлять и размечать чеховские «говорящие» фамилии с их свойствами и характеристиками, переносные смыслы чеховского текста, иронические

характеристики людей и ситуации, и многое другое.

Таким образом, разработка экспертной разметки, т. е. сохранение филологических знаний в машиночитаемом виде, ставит ряд задач перед исследователями. Например, реализация расширенного поиска для семантических изданий предполагает разметку более широкого спектра значимых элементов, таких как цвет/запах/вкус; природные явления; пространство и т. п., которые будут интересны филологам-исследователям. Но вопрос репрезентации этих сущностей в тексте остается достаточно сложным и требует дополнительных исследований. Кроме того, существует опасность предзаданной интерпретации, которая сохранится в текстах в машиночитаемом виде, но в то же время она может стать основой нового типа комментария, объединяющего тексты и смыслы, усложняющего и обогащающего понимание текста, давая возможность его включения в более широкий контекст: исторический, культурный, биографический и т.д.. Использование экспертной разметки для обучения алгоритмов машинного обучения кажется очень перспективной задачей для разметки других текстов, создания приложений в различных областях знания, связанных с текстом. И этот список далеко не исчерпан. Каждый из этих вопросов важно рассматривать с точки зрения проблем, которые требуют своего осмысления и решения.

Наша задача – поставить вопрос о сложности реализации такого рода междисциплинарных проектов и необходимости объединения усилий специалистов из разных областей знаний.

Библиография

1. Чехов А. П. Полное собрание сочинений и писем: В 30 т. / АН СССР. Ин-т мировой лит. им. А. М. Горького. М.: Наука, 1974-1983.
2. TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.6.0. Last updated on 4th April 2023. TEI Consortium. URL: <https://tei-c.org/guidelines/p5/>
3. 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. <https://doi.org/10.29025/2079-6021-2022-2-153-165>.
4. Северина Е.М., Ларионова М.Ч. Новые филологические практики: семантическое издание текстов А. П. Чехова. Филология: научные исследования. 2020, № 10, с. 13-21. DOI: 10.7256/2454-0749.2020.10.33970.
5. Чехов А. П. Смерть чиновника // Чехов А. П. Полное собрание сочинений и писем: В 30 т. Сочинения: В 18 т. Т. 2. М.: Наука, 1975. С. 164–166.
6. Даль В. И. Толковый словарь живого великорусского языка: В 4 т. Т. 4. М.: Русский язык, 1989.
7. Короленко В. Г. Антон Павлович Чехов // Чехов в воспоминаниях современников. М.: Гослитиздат, 1960. С. 135–148.
8. Бердников Г. П. А. П. Чехов. Идеиные и творческие искания. М.: Художественная литература, 1984.
9. M. Straka, J. Straková, J. Hajič (2019): UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging. In: Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 95-103, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-950737-36-9.

10. Дяченко П. В., Л. Л. Иомдин, А. В. Лазурский, Л. Г. Митюшин, О. Ю. Подлесская, В. Г. Сизов, Т. И. Фролова, Л. Л. Цинман. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Труды Института русского языка им. В.В. Виноградова. М., 2015. № 6. С. 272-300.
11. M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, & C. D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA). URL: https://nlp.stanford.edu/pubs/USD_LREC14_paper_camera_ready.pdf
12. Потебня А. А. Из записок по русской грамматике: В 4 т. Т. 3. М.: Просвещение, 1968.
13. Болотова Е. А. Антропонимическая языковая мозаика «говорящих» имен и фамилий // The Scientific Heritage. 2019. № 4. С. 35–39.
14. Ларионова М.Ч., Шепелева О.А. Отчего умер Червяков? Традиционная культура в рассказе А.П. Чехова «Смерть чиновника» // Известия ЮФУ. Филологические науки. 2019. № 1. С. 36–41. <https://doi.org/10.23683/1995-0640-2019-1-36-41>.
15. Катаев В.Б. «Степь»: драматургия прозы // Таганрогский вестник. Материалы международной научно-практической конференции «"Степь" А. П. Чехова: 120 лет». Вып. 3. Таганрог: ООО «Изд-во Лукоморье», 2008. С. 3–9.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.

Со списком рецензентов издательства можно ознакомиться [здесь](#).

Цифровая обработка текста на данном историческом этапе является продуктивной задачей, ориентирующей исследователей на более пристальное внимание к языку. Следовательно, языковая система оценивается и формально, и содержательно, это дает возможность объемнее представлять механизмы функционирования естественной знаковой парадигмы. Рецензируемая статья сформирована как аналитический обзор проекта «Chekhov Digital», который представляет собой семантическое издание Полного собрания сочинений и писем А. П. Чехова в 30 томах (далее ПССиП), разрабатываемое в формате стандарта цифровой публикации Text Encoding Initiative (TEI/XML). Как отмечается в начале труда, «разработка проекта осуществляется Центром цифровых гуманитарных исследований Института филологии журналистики и межкультурной коммуникации Южного федерального университета (ИФЖИМКК ЮФУ) совместно с Международной лабораторией языковой конвергенции НИУ ВШЭ и Отделом гуманитарных исследований Южного научного центра Российской академии наук ЮНЦ РАН». Автор дает достаточно подробный алгоритм работы системы, описывает нюансы, комментирует тонкости: «в системе Text Encoding Initiative существует обширный список тегов для разметки самой разнообразной информации, в том числе внешней по отношению к тексту, – от специфики представления информации в разных типах текста (пьеса, рассказ, повесть или письмо) до имен собственных и таких социальных категорий, как социальный статус, профессиональная принадлежность и т.п. Такого рода информация создает трудности для технологий автоматической обработки текстов, поскольку может быть представлена разными способами, но описанная явно с помощью разметки TEI уменьшает сложность обработки с помощью компьютерных методов. При этом любая разметка – это форма интерпретации текста, поэтому цифровой формат

представляет собой некоторый вариант интерпретации чеховских текстов. Размеченные таким образом тексты размещаются в открытом доступе на сайте проекта (<http://chekhov-digital.sfedu.ru/>), на котором дополнительно представлены некоторые цифровые исследовательские инструменты, такие как семантический поиск, инструменты визуализации и т.п.». Далее на материале цифровой обработки рассказа А.П. Чехова «Смерть чиновника» проиллюстрирован основной процесс. Кроме чисто технических звеньев, дана и аналитика механизма. Автор старается полновесно раскрыть суть оценки: например, «уже на этом примере видно, что стандартная морфологическая разметка покажет только специфику языковых форм, поэтому для разметки семантики требуется совершенно другой подход, в котором будут учтены соответствующие маркеры свойства, характеристики ситуации (типичность); свойства субъективного восприятия – времени и профессии/социального статуса; авторской иронии, построенной по принципу контраста: профессия – прекрасная, по самооценке героя, но дающая невысокий социальный статус, и намекающая на возможность наказания», или «использование специализированных алгоритмов извлечения именованных сущностей (Named Entity Recognition, NER) для русского языка (например, библиотека SlovNet, <https://github.com/natasha/slovnet>) дает возможность автоматически разметить в текстах фамилию, имя, отчество героя, но информация о маркерах национального стереотипа, социального статуса и характера героя должна быть внесена дополнительно. Например, с помощью алгоритмов NER из текста извлечены имена персонажа, для разметки которых могут быть использованы следующие теги: Иван Дмитрич Червяков» и т.д. Основная задача по ходу работы решена, цель, таким образом, также достигнута. На мой взгляд, было бы правильно вывести полный анализ рассказа, он не такой уж и большой, было бы интересно и полновесно. Разметка текста, как отмечает автор статьи, это некоторая форма интерпретации; да, отчасти с этим можно согласиться, принять во внимание. Материал можно использовать в качестве образчика аналитической процедуры оценки разметки текста, цифровой обработки высказывания. Стиль данного труда тяготеет к собственно научному типу, термины / понятия используются унифицировано. Работе присуща внутренняя логика, логика открытого движения мыслей. Например, «сложность семантической разметки определяется многими аспектами, в том числе «технологией» автора, к которой можно отнести и наличие переносных значений у слов и конструкций; и невозможность автоматического распознавания иронии, которая не имеет вербальных способов выражения; и «сложносоставность» художественных приемов, когда ирония дополняется метафорой, контекстуальным оксюмороном и пр.; и прием «стебельчатого шва», когда одна тема начинается «внутри» другой и каждый новый «стежок» захватывает обе темы; и т. д.. Кроме того, сама «технология» обработки имеет сложности, связанные с автоматическим выделением сущностей в тексте, ограничением формальных приемов их разметки и самого языка разметки, а также необходимостью привлечения экспертов на разных этапах разметки, что предполагает наличие у исследователей специализированных знаний в области цифровой филологии». Формально итоги подведены, но автор оговаривает, что указанная проблема должна разбираться далее. Это, на мой взгляд, весьма хороший итог: «разработка экспертной разметки, т. е. сохранение филологических знаний в машиночитаемом виде, ставит ряд задач перед исследователями. Например, реализация расширенного поиска для семантических изданий предполагает разметку более широкого спектра значимых элементов, таких как цвет / запах / вкус; природные явления; пространство и т. п., которые будут интересны филологам-исследователям. Но вопрос репрезентации этих сущностей в тексте остается достаточно сложным и требует дополнительных исследований». Список источников достаточен, цитации / ссылки даются с учетом требований издания. С учетом сказанного

нет оснований не допустить текст к публикации, рекомендую статью «Проект Chekhov Digital: задачи и проблемы реализации семантической разметки текстов (на примере рассказа А. П. Чехова «Смерть чиновника»)» к открытой печати в журнале «Litera».