Филология: научные исследования
Правильная ссылка на статью:
Северина Е.М., Фёдоров Н.А. Проект Chekhov Digital: семантическая разметка параллельного корпуса переводов художественной прозы А. П. Чехова на немецкий язык // Филология: научные исследования. 2024. № 4. DOI: 10.7256/2454-0749.2024.4.70560 EDN: PXMQSB URL: https://nbpublish.com/library_read_article.php?id=70560

Проект Chekhov Digital: семантическая разметка параллельного корпуса переводов художественной прозы А. П. Чехова на немецкий язык

Северина Елена Михайловна

ORCID: 0000-0001-6518-2771

доктор философских наук

профессор, кафедра лингвистики и профессиональной коммуникации, Южный федеральный университет

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

emkovalenko@sfedu.ru

Фёдоров Никита Александрович

ORCID: 0000-0002-7436-2202

магистр, институт филологии, журналистики и межкультурной коммуникации, Южный федеральный университет

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

nfyodorov@sfedu.ru

Статья из рубрики "Автоматическая обработка языка"

DOI:

10.7256/2454-0749.2024.4.70560

EDN:

PXMQSB

Дата направления статьи в редакцию:

24-04-2024

Аннотация: В статье рассматриваются вопросы разработки принципов семантически размеченного параллельного корпуса переводов художественной прозы А.П. Чехова на немецкий язык в рамках проекта Chekhov Digital цифрового академического издания собрания произведений писателя в формате TEI (Text Encoding Initiative). Проект

параллельного корпуса ориентирован на создание цифровой инфраструктуры для изучения произведений писателя, позволяющей исследователям анализировать и сравнивать оригинальные тексты с их переводами. Были выявлены сложности, связанные с интерпретацией значимых элементов произведений писателя, спецификой их перевода на немецкий язык и семантической разметкой переводов художественной прозы, например, возникли сложности с определением границ и связей между элементами семантической разметки. Предложены пути их преодоления, включая использование цифровых методов и технологий обработки естественного языка. В проекте используются цифровые методы и технологии обработки естественного языка, стандарт цифровой публикации Text Encoding Initiative (TEI). Структура разметки текстов, основанная на стандарте TEI, делает документы машиночитаемыми, что позволяет разрабатывать инструменты сложного семантического поиска информации. Включение в проект Chekhov Digital параллельных корпусов переводов произведений А. П. Чехова на разные языки позволяет расширить исследовательские инструменты в области переводоведения, давая возможность сравнивать тексты переводов и оригиналов, обнаруживать сходства и различия в лексике, грамматике, стиле и культурных отсылках, а также автоматизировать рутинные процессы исследования, что делает значительно более эффективным поиск и анализ информации на больших объемах текстов. Результаты проекта будут вносить вклад в развитие цифровой гуманитарной среды, способствуя сохранению и популяризации литературного наследия А.П. Чехова. Создание семантически размеченного параллельного корпуса переводов будет иметь важное значение для литературоведов, лингвистов и переводчиков, позволяя им изучать специфику переводов произведений Чехова и развивать новые формы анализа и интерпретации текстов. Опыт, полученный в ходе проекта, будет ценным для будущих исследований и практических применений, демонстрируя эффективность цифровых технологий в гуманитарных исследованиях и образовании.

Ключевые слова:

проект Chekhov Digital, цифровое издание, Чехов, параллельные корпусы, Text Encoding Initiative, машиночитаемая разметка, семантический поиск, цифровые технологии, автоматическая обработка текста, парсинг

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

Введение

В современном гуманитарном знании цифровые технологии приобретают все большее значение, соответствуя глобальным тенденциям цифровизации различных сфер человеческой деятельности, с одной стороны, а с другой ▌позволяет развивать новые формы и исследовательские подходы в гуманитарных науках. Развитие цифровой среды меняет формы существования текстов, формируя новые подходы к форматам изданий литературных текстов, которые должны стать цифровыми, обеспечивая не только сохранность текстов как культурных объектов, но и доступ к ним как к носителям культурных смыслов и знаний, что требует преобразования филологических знаний в цифровой машиночитаемый формат. Обеспечить такое представление текста в виде связанных данных, выражающих прямую, явную и понятную для компьютерной обработки взаимосвязь сущностей возможно в рамках семантической разметки. В этом контексте

создание семантических изданий литературных произведений способствует систематизации творческого наследия писателей и оптимизации профессиональной деятельности филологов.

Институт филологии, журналистики и межкультурной коммуникации Южного федерального университета совместно с отделом гуманитарных исследований Южного научного центра Российской академии наук и Международной лабораторией языковой конвергенции Национального исследовательского университета «Высшая школа экономики» ведут работу над проектом Chekhov Digital, в рамках которого разрабатывается семантическое цифровое издание текстов Полного собрания сочинений и писем А. П. Чехова в 30 томах (ПССиП)[1] в соответствии со стандартом цифровой публикации Text Encoding Initiative (TEI)[2]. Для каждого текста ПССиП, включая редакционно-критические материалы, была разработана семантическая машиночитаемая разметка в формате TEI, также ведется работа по созданию цифрового индекса имен и названий, упоминаемых в текстах академического издания, включая комментарии и примечания. Цифровой индекс строится на основе существующих указателей, верифицируется полуавтоматически, дополняется информацией о реальных людях и объектах из внешних баз данных, таких как Wikidata [3]. Основная цель проекта – сделать семантическую разметку текстов общедоступной, разработать доступные цифровые инструменты работы с текстами ПССиП [4].

В проекте Chekhov Digital используется структура разметки текстов, основанная на стандарте цифровой публикации TEI (Text Encoding Initiative)[2], которая делает документы машиночитаемыми. Документы, размеченные в соответствии с принципами TEI, состоят из двух частей: TEI-заголовка с метаданными источника (например, описанием издания, названием, именем автора, языком текста и т.д.), и текстового модуля, включающего размеченную текстовую информацию. TEI позволяет учесть специфику текста, например, дополнительные метаданные для писем (адресат, дата и место написания и т.д.), особенности представления информации в тексте, разметить именованные сущности, биографические сведения и некоторые социальные категории (социальный статус, профессиональную принадлежность и т. п.). Размеченные таким образом тексты позволяют разрабатывать инструменты сложного семантического поиска информации [4, с. 158].

В рамках исследования разработаны принципы семантической разметки переводов художественной прозы А. П. Чехова на немецкий язык в рамках проекта Chekhov Digital, что позволит создать параллельный корпус переводов в формате TEI.

Включение в проект Chekhov Digital параллельных корпусов переводов произведений А. П. Чехова на разные языки позволяет расширить исследовательские инструменты в области переводоведения, давая возможность сравнивать тексты переводов и оригиналов, обнаруживать сходства и различия в лексике, грамматике, стиле и культурных отсылках, а также автоматизировать рутинные процессы исследования, что делает значительно более эффективным поиск и анализ информации на больших объемах текста. Таким образом, включение в проект Chekhov Digital параллельного корпуса рассказов А. П. Чехова и их переводов на немецкий язык является важной и перспективной задачей, которая способствует развитию языковых, литературных и переводческих исследований.

Создание параллельных корпусов является сложной задачей, которая требует решения ряда проблем, связанных с выравниванием текстов. Выравнивание (alignment) – это

процесс сопоставления фрагментов текста в исходном и целевом языках. Выравнивание может быть выполнено как вручную, так и с помощью автоматических инструментов, однако в любом случае этот процесс требует значительных временных и ресурсных затрат.

Как правило, соотнесение текстов в параллельных корпусах производится по предложениям, однако для автоматической разметки этот метод подходит не всегда: «Одна из наиболее существенных трудностей выравнивания заключается в том, что авторское членение текста на предложения и абзацы не всегда выдерживается в тексте перевода» [5, с. 289]. Тексты на разных языках могут иметь различную структуру и организацию в зависимости от множества лингвистических факторов. Например, предложения в одном языке могут быть длиннее или короче, чем в другом, или могут содержать различные грамматические конструкции. В немецком языке существует «рамочная конструкция» предложения [6, с. 141]: глагол занимает конечную позицию, а остальные элементы предложения располагаются вокруг него; эта конструкция характерна для немецкого языка и отличается от структуры предложений во многих других языках, где глагол обычно располагается в середине предложения. Это может затруднить процесс выравнивания и сделать его менее точным. Кроме того, выравнивание может быть осложнено наличием в тексте идиоматических выражений, фразеологизмов и других языковых особенностей, которые невозможно перевести дословно или сопоставить с текстом на другом языке.

Таким образом, выравнивание текстов является необходимым шагом в процессе, который позволяет сравнивать и анализировать тексты с достаточной точностью. Переводы рассказов А. П. Чехова на немецкий язык, отобранные нами для анализа и разметки, представляют довольно точную интерпретацию оригинальных текстов, но лучше соотносятся по абзацам, чем по предложениям, в связи с обозначенными выше особенностями синтаксиса немецкого языка.

Создавая параллельные корпусы текстов, оснащенные семантической разметкой, исследователи стремятся к автоматизации и универсализации структуры этой разметки. Создатели семантического издания Chekhov Digital ориентируются на выявленную в результате анализа значимую лексику [7]. Мы полагаем, что для оптимизации разметки параллельного корпуса необходимо обратить внимание на выделение значимой лексики для текстов на двух языках, что позволит разрешить некоторые трудности выравнивания.

Материал исследования

На страницах сайта электронной библиотеки немецкоязычных текстов Projekt Gutenberg-DE были собраны 84 текста переводов рассказов А.П. Чехова на немецкий язык[8]. Собранный корпус представляет собой все переводы, находящиеся в свободном доступе в данной электронной библиотеке. Оригинальные тексты произведений писателя из Полного академического собрания сочинений представлены в рамках проекта Chekhov Digital [9].

На данный момент общий корпус исследуемых текстов составляет 168 текстов: 84 текста на русском языке (оригинальные тексты), 84 текста переводов, выполненных А. Элиасбергом (46), В. Чумиковым (12), К. Хольмом (5); авторство 21 текста переводов не обозначено ни на сайте Projekt Gutenberg-DE, ни в других сборниках переводов рассказов А.П. Чехова, находящихся в свободном доступе в сети Интернет [10].

С помощью стилометрического анализа [11] были выявлены элементы авторского стиля разных переводчиков в этих текстах, в связи с чем предполагается, что они были переведены совместно [12].

Размеченные тексты проекта Chekhov Digital свободно распространяются по лицензии Creative Commons Attribution Non-Commercial (CC BY-SA), т.е. разрешено свободное использование произведения и его разметки, при условии указания авторства и сохранения условий использования [9]. Материалы и разметка сайта «Projekt Gutenberg-DE» распространяются по лицензии Creative Commons Attribution Non-Commercial (CC BY-NC), т. е. разрешено свободное использование произведения и его разметки, при условии указания авторства, но только в некоммерческих целях.

Большая часть исследуемых рассказов принадлежат к раннему периоду творчества А. П. Чехова – с 1883 по 1887 годы: 2-й том ‖ 14 рассказов; 4-й том – 12 рассказов; 5-й – 18 рассказов; 6-й – 19 рассказов.

Результаты и обсуждение

Для подготовки корпуса важную роль играют цифровые технологии, позволяющие автоматизировать процесс сбора и систематизации текстовых данных. Для решения этой задачи был использован парсинг (англ. *parsing*; *web scraping*) ‖ автоматизированный сбор и систематизация информации из открытых источников с помощью программного обеспечения [13, с. 33]. Для решения этой задачи использовалась библиотека Beautiful Soup ЯП Python, которая позволила также проверять, соответствует ли разметка текста определенным шаблонам, разрабатываемым в проекте [7], и вносить необходимые корректировки.

С другой стороны, большое значение для создания параллельного корпуса в рамках проекта Chekhov Digital имеет аннотация текстов, то есть их разметка по грамматическим, синтаксическим и семантическим признакам, что делает текст машиночитаемым для автоматической интерпретации. Для выявления значимых элементов из текста для такого рода разметки были использованы цифровые методы анализа: тематическое моделирование для определения набора скрытых тем в исследуемых рассказах А. П. Чехова, а также их ключевой лексики; анализ тональности текстов для выявления эмоционального контекста; стилометрический анализ для изучения сходств и различий в стилях автора и переводчиков прозы писателя, исследование соавторства в переводах рассказов и т.д.

Таким образом, семантическая разметка и создание параллельного корпуса – задачи, требующие, по нашему мнению, цифрового анализа ключевой лексики, эмотивных элементов, синтаксических особенностей текстов, маркеров авторского стиля и т.д. При создании параллельного корпуса переводов рассказов А. П. Чехова на немецкий язык для проекта Chekhov Digital мы опираемся на полученные результаты и опыт использования цифровых методов для разработки функционального инструментария работы с творческим наследием писателя.

В качестве языка разметки произведений на сайте проекта Chekhov Digital используется стандарт TEI (Text Encoding Initiative). Это международный стандарт для представления текстов в цифровом формате, используемый в гуманитарных науках, лингвистике и издательском деле, основными преимущества которого являются: гибкость и расширяемость; междисциплинарность; поддержка многоязычности; совместимость с

другими стандартами [2]. В целом TEI предоставляет мощный и гибкий инструмент для представления текстов в цифровом формате, который может быть использован в широком спектре дисциплин и приложений. Кроме того, гибкость и расширяемость позволяют настраивать его для решения конкретных задач, в том числе в рамках проекта Chekhov Digital, через создание новых тегов или модификации уже существующих.

Поскольку TEI совместим со многими другими языками разметки, это позволяет разработать систему перевода тегов из языка разметки HTML в формат TEI. Но для этого необходимо изучить код источника, определить соответствия тегов и зафиксировать их. Подобным образом была проведена структурная разметка корпуса оригинальных (русскоязычных) работ А. П. Чехова для проекта Chekhov Digital, включая письма, рассказы, пьесы, очерки и т.д. Источником HTML-размеченных текстов стало ЭНИ «Чехов» (электронное научное издание), размещенное на сайте Фундаментальной электронной библиотеки [14]. Из HTML разметки ее страниц была собрана информация о текстах для TEI-документов: названия произведений, годы написания, библиографические описания, объемы текстов (количество страниц), номер тома, наличие заголовков, различных форматирований, сносок и т.п. Однако, разметка на разных сайтах различается, в связи с чем возникают определенные трудности, если в коде отсутствует необходимая информация о произведениях. К таким сайтам относится Projekt Gutenberg-DE [8], содержащий тексты переводов рассказов А. П. Чехова на немецкий язык. С помощью его разметки можно соотнести оригинальные и переведенные тексты по абзацам и предложениям, однако такая информация, как библиографическое описание, год публикации и перевода, объем текста (количество страниц) и т.д. отсутствует, что затрудняет процесс автоматизации сбора метаданных об источнике перевода и включения этой информации в TEI-документ. Поэтому информация о текстах переводов была собрана вручную и использована в качестве универсального справочника для заполнения метаданных в TEI-документах.

Тем не менее, были найдены некоторые соответствия HTML разметки сайта Projekt Gutenberg-DE и разметки TEI проекта Chekhov Digital: например, тег HTML: *<meta name="title" content="..."/>* соответствует тегу TEI: *<title>*, тег HTML: <meta name="author" content="..."/> – тегу TEI *<author>* и т.д., поэтому некоторую метаинформацию удалось собрать со страниц сайта Projekt Gutenberg-DE. Например, произведения на немецком языке здесь представлены в сборниках, поэтому метаинформация в разметке HTML представлена для всего сборника, следовательно, при автоматической разметке названия произведения (тег *<title>*) необходимо использовать информацию из тега *<h3>*, который маркирует заголовок третьего порядка (название произведения в сборнике). Кроме того, в тегах *<meta name="editor">* или *<meta name="translator">* отмечены все переводчики, принявшие участие в выпуске сборника, но для отдельного произведения сборника следует использовать тег *<h5>* (заголовок пятого порядка), в котором указано имя переводчика данного текста.

Для создания параллельного корпуса переводов произведений А. П. Чехова на разные языки в проекте Chekhov Digital используется тег *<choice>*, который в TEI используется для представления альтернативных вариантов разметки одного и того же фрагмента текста. Этот тег позволяет закодировать неоднозначность или неопределенность в тексте и предоставить читателю или исследователю различные варианты интерпретации. При этом тег *<choice>* может использоваться не только для выделения альтернативных вариантов разметки одного и того же фрагмента текста, но и для выделения целых сегментов текста, которые могут содержать различные его варианты. В этом случае тег *<choice>* может содержать дочерние элементы, содержащие большие фрагменты текста.

В качестве фрагментов текста в нашем случае выступают абзацы или предложения (в зависимости от варианта соотнесения размеченных текстов: по абзацам или по предложениям – в разных случаях оптимальным может быть как первый, так и второй).

Существуют разные способы выделения двух альтернативных вариантов текста целиком, один из них ‖ каждый вариант представлен тегом абзаца * *, который содержит текст версии. Например, каждый элемент * * имеет уникальный идентификатор, заданный с помощью атрибута *@xml:id*, в котором указывается вариант текста – оригинальный или переведенный на определенный язык (de, en, bg и т.д.):

<choice>

Молодая рыжая собака — помесь такса с дворняжкой — очень похожая мордой на лисицу, бегала взад и вперед по тротуару и беспокойно оглядывалась по сторонам. Изредка она останавливалась и, плача, приподнимая то одну озябшую лапу, то другую, старалась дать себе отчет: как это могло случиться, что она заблудилась?

 Ein junger rotbrauner Hund – eine Kreuzung von Dachs und Dorfköter –, dessen Schnauze der eines Fuchses sehr ähnelte, lief auf dem Trottoir hin und her und schaute sich unruhig nach allen Seiten um. Zuweilen blieb er stehen, hob winselnd bald die eine, bald die andere seiner frierenden Pfoten und suchte sich darüber Rechenschaft zu geben, wie es doch passieren konnte, daß er sich verirrt hatte?

</choice>

Второй способ: каждый вариант представлен элементом *<seg>* (от *'segmentation'*), который содержит текст версии. Элемент *<seg>* имеет идентификаторы *<type>* и *<subtype>*, задающие тип фрагмента: «оригинал» («*orig*») или «перевод» («*translated*») – и язык перевода («en», «de», «fr») соответственно:

<choice>

<seg type="orig">*Молодая рыжая собака — помесь такса с дворняжкой — очень похожая мордой на лисицу, бегала взад и вперед по тротуару и беспокойно оглядывалась по сторонам. Изредка она останавливалась и, плача, приподнимая то одну озябшую лапу, то другую, старалась дать себе отчет: как это могло случиться, что она заблудилась?*</seg>

<seg type="translated" subtype="de">*Ein junger rotbrauner Hund – eine Kreuzung von Dachs und Dorfköter –, dessen Schnauze der eines Fuchses sehr ähnelte, lief auf dem Trottoir hin und her und schaute sich unruhig nach allen Seiten um. Zuweilen blieb er stehen, hob winselnd bald die eine, bald die andere seiner frierenden Pfoten und suchte sich darüber Rechenschaft zu geben, wie es doch passieren konnte, daß er sich verirrt hatte?*</seg>

</choice>

Для разметки параллельного корпуса был выбран элемент *<seg>* как более предпочтительный по ряду причин. Во-первых, его структура позволяет более точно закодировать альтернативные варианты текста, соотносимые нами по параграфам и знаменательной лексике, поскольку его назначение – выделение целых сегментов или фрагментов текста. В отличие от атрибута *xml:id*, который позволяет создавать ссылки на различные элементы документа и может использоваться, например, для присвоения

ссылок именованным сущностям или другим единичным элементам текста, тег *<seg>* имеет более узкую направленность. Выбрав его в качестве основного для разметки альтернативных сегментов текста, мы избежим проблем, когда один и тот же тег выполняет несколько разнонаправленных функций.

Во-вторых, опираясь на необходимость выравнивания по ключевой лексике, выделенной в процессе цифрового исследования, мы стремимся сперва разделить тексты на небольшие отрезки, внутри которых будет проще соотнести значимые элементы. В таком случае мы можем использовать тег *<seg>* для выделения отдельных абзацев в каждом тексте, обозначения языка перевода и т.д. Затем внутри этих фрагментов мы можем выровнять значимую лексику, используя атрибут *xml:id* или другие, имеющие конкретное назначение (имена, даты, фразеологизмы и т.д.).

Заключение

Семантическое издание Chekhov Digital ‖ динамично развивающийся проект по цифровизации литературного наследия А. П. Чехова, который ориентирован на создание новых исследовательских инструментов для изучения и анализа его произведений. Один из таких инструментов ‖ параллельный корпус переводов художественной прозы писателя на немецкий язык, с помощью которого появится возможность обнаруживать в текстах переводов и оригиналов сходства и различия в лексике, грамматике, стиле, культурных отсылках. Разработка семантической разметки для параллельного корпуса переводов рассказов А. П. Чехова на немецкий язык является важным шагом в развитии этого инструмента. Опыт использования цифровых методов для разработки функционального инструментария работы с творческим наследием Чехова в рамках проекта Chekhov Digital является ценным для будущих исследований и практических применений. Он демонстрирует возможность эффективного использования цифровых технологий в гуманитарных исследованиях и образовании.

В целом, проект Chekhov Digital является перспективным направлением развития цифрового гуманитарного образования и исследований, которое способствует сохранению и популяризации творческого наследия А. П. Чехова. Создание параллельного корпуса переводов рассказов писателя на немецкий язык расширяет функциональность проекта и его применимость в различных областях знаний, включая языкознание, литературоведение и переводоведение.

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 4.7.0. Last updated on 16th November 2023. TEI Consorti-um. URL: http://www.tei-c.org/Guidelines/P5/ (дата обращения: 24.03.2024).
3. Wikidata // URL: https://www.wikidata.org/wiki/Wikidata: Main_Page (дата обращения: 10.04.2024).
4. 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.
5. Добровольский Д. О., Кретов А. А., Шаров С. А. Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005, 263-296.
6. Калинина Е. Э. Рамочная конструкция предложения как результат генезиса порядка слов индоевропейских языков // Международный научно-исследовательский журнал.

2017. №5-1 (59). С. 141-143.
7. 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 URL: https://e-notabene.ru/fmag/article_33970.html
8. Projekt Gutenberg-DE. URL: https://www.projekt-gutenberg.org/ (дата обращения 10.04.2024).
9. Chekhov Digital. Семантическое издание текстов А. П. Чехова. URL: https://chekhov-digital.sfedu.ru/ (дата обращения: 10.04.2024).
10. Tschekhow Anton Pawlowitsch Gesammelte Erzählungen: Geschichten in Grau + Kleine Erzählungen + Lustige Geschichten + Von Frauen und Kindern + Duell… // Sharp Ink Publishing. 2023.
11. Eder M. Rolling stylometry //Digital Scholarship in the Humanities. 2016. 31(3). P. 457-469.
12. Федоров Н. А. Проект Chekhov Digital: стилометрический анализ текстов перевода рассказов А.П. Чехова на немецкий язык // Инновации в науке и трансформация парадигмы современного образования: сборник научных трудов. Казань: ООО "САНТРЕМ", 2024. С. 147-150.
13. Меньшиков Я. С. Преимущества автоматического сбора данных в сети интернет над ручным сбором данных // Universum: технические науки. 2022. №10-1 (103). С. 33-36.
14. ЭНИ «Чехов» / ФЭБ. URL: https://feb-web.ru/feb/chekhov/default.asp (дата обращения 10.04.2024).

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Цифровизация галопирующим образом заполняет все сферы деятельности. Не является исключением и мир гуманитарный. Как отмечает в начале своего труда автор, «в современном гуманитарном знании цифровые технологии приобретают все большее значение, соответствуя глобальным тенденциям цифровизации различных сфер человеческой деятельности, с одной стороны, а с другой позволяет развивать новые формы и исследовательские подходы в гуманитарных науках. Развитие цифровой среды меняет формы существования текстов, формируя новые подходы к форматам изданий литературных текстов, которые должны стать цифровыми, обеспечивая не только сохранность текстов как культурных объектов, но и доступ к ним как к носителям культурных смыслов и знаний, что требует преобразования филологических знаний в цифровой машиночитаемый формат». С данным утверждением сложно не согласиться, оно объективно, оно конструктивно. В данной статье анализу подвергается проект Chekhov Digital, который курирует Институт филологии, журналистики и межкультурной коммуникации Южного федерального университета совместно с отделом гуманитарных исследований Южного научного центра Российской академии наук и Международной лабораторией языковой конвергенции Национального исследовательского университета «Высшая школа экономики». Считаю, что грамотная оценка проекта необходима для дальнейшего совершенствования указанного цифрового ресурса. Статья достаточно грамотно выстроена, определенное чередование теоретического уровня с практическим дает возможность читателю (даже не подготовленному) получить необходимый срез оценки. Методика исследования соотносится с рядом современных наработок. Суждения имеют научный характер, серьезных нарушений не выявлено. Стиль соотносится с

академическим письмом: например, «в проекте Chekhov Digital используется структура разметки текстов, основанная на стандарте цифровой публикации TEI (Text Encoding Initiative), которая делает документы машиночитаемыми. Документы, размеченные в соответствии с принципами TEI, состоят из двух частей: TEI-заголовка с метаданными источника (например, описанием издания, названием, именем автора, языком текста и т.д.), и текстового модуля, включающего размеченную текстовую информацию. TEI позволяет учесть специфику текста, например, дополнительные метаданные для писем (адресат, дата и место написания и т.д.), особенности представления информации в тексте, разметить именованные сущности, биографические сведения и некоторые социальные категории (социальный статус, профессиональную принадлежность и т. п.)», или «создание параллельных корпусов является сложной задачей, которая требует решения ряда проблем, связанных с выравниванием текстов. Выравнивание (alignment) – это процесс сопоставления фрагментов текста в исходном и целевом языках. Выравнивание может быть выполнено как вручную, так и с помощью автоматических инструментов, однако в любом случае этот процесс требует значительных временных и ресурсных затрат» и т.д. Целевая составляющая сводится к следующему: «в рамках исследования разработаны принципы семантической разметки переводов художественной прозы А. П. Чехова на немецкий язык в рамках проекта Chekhov Digital, что позволит создать параллельный корпус переводов в формате TEI». Статистические данные вводятся в работу с учетом максимальной открытости: «на страницах сайта электронной библиотеки немецкоязычных текстов Projekt Gutenberg-DE были собраны 84 текста переводов рассказов А.П. Чехова на немецкий язык. Собранный корпус представляет собой все переводы, находящиеся в свободном доступе в данной электронной библиотеке. Оригинальные тексты произведений писателя из Полного академического собрания сочинений представлены в рамках проекта Chekhov Digital. На данный момент общий корпус исследуемых текстов составляет 168 текстов: 84 текста на русском языке (оригинальные тексты), 84 текста переводов, выполненных А. Элиасбергом (46), В. Чумиковым (12), К. Хольмом (5); авторство 21 текста переводов не обозначено ни на сайте Projekt Gutenberg-DE, ни в других сборниках переводов рассказов А.П. Чехова, находящихся в свободном доступе в сети Интернет…». Считаю, что работа имеет практическую направленность, результаты можно использовать далее при формировании / создании цифровых платформ. Выводы по работе ориентированы на основной блок, автор отмечает, что «в целом, проект Chekhov Digital является перспективным направлением развития цифрового гуманитарного образования и исследований, которое способствует сохранению и популяризации творческого наследия А. П. Чехова. Создание параллельного корпуса переводов рассказов писателя на немецкий язык расширяет функциональность проекта и его применимость в различных областях знаний, включая языкознание, литературоведение и переводоведение». Думается, работа в указанных сегментах может быть продолжена далее. Констатирую: тема данного исследования раскрыта, цель достигнута, общие требования издания учтены, текст не нуждается в серьезной правке и доработке. Рекомендую рецензируемую статью «Проект Chekhov Digital: семантическая разметка параллельного корпуса переводов художественной прозы А. П. Чехова на немецкий язык» к публикации в журнале «Филология: научные исследования».