

М.Ч. Ларионова, доктор филологических наук

Южный научный центр Российской академии наук,

Ростов-на-Дону, Российская Федерация

Е.М. Северина, доктор философских наук

Южный федеральный университет, Ростов-на-Дону, Российская Федерация

M.Ch. Larionova, Doctor of Philology

Southern Scientific Center of the Russian Academy of Sciences,

Rostov-on-Don, Russian Federation

E.M. Severina, Doctor of Philosophy

Southern Federal University, Rostov-on-Don, Russian Federation

**ПРОЕКТ CHEKHOV DIGITAL: ФИЛОЛОГИЧЕСКАЯ ЭКСПЕРТИЗА
И СЕМАНТИЧЕСКАЯ РАЗМЕТКА ТЕКСТОВ ПРОИЗВЕДЕНИЙ
А.П. ЧЕХОВА**

**PHILOLOGICAL EXPERTISE AND SEMANTIC ANNOTATION
OF A.P. CHEKHOV'S WORKS IN THE CHEKHOV DIGITAL PROJECT**

Доклад посвящен семантической разметке текстов произведений А.П. Чехова, основанной на стандарте TEI/XML, в проекте Chekhov Digital. В исследовании рассматриваются проблемы автоматической разметки текстов, а также необходимость филологической экспертизы для интерпретации художественных приемов писателя. Проект открывает новые возможности для академических исследований, интеллектуального анализа текста и цифровых гуманитарных приложений.

Ключевые слова: Chekhov Digital, семантическая разметка, TEI/XML, цифровая филология, машиночитаемый текст, филологическая экспертиза.

The report addresses the issue of semantic annotation of A.P. Chekhov's literary texts using the TEI/XML standard within the Chekhov Digital project. The study examines the challenges of automatic text annotation as well as the necessity of philological expertise for the interpretation of the writer's artistic techniques. The project opens new opportunities for academic research, intellectual text analysis, and digital humanities applications.

Key words: Chekhov Digital, semantic annotation, TEI/XML, digital philology, machine-readable text, philological expertise.

Проект Chekhov Digital [Chekhov Digital] направлен на создание цифрового издания Полного собрания сочинений и писем А.П. Чехова (ПССиП) в 30 томах [Чехов 1974–1983] и разрабатывается в формате стандарта цифровой публикации Text Encoding Initiative (TEI/XML) [TEI 2025]. Цель исследования – разработка машиночитаемой семантической разметки текстов произведений писателя, которая позволит интегрировать чеховское наследие в цифровой культурный контекст и открывает новые перспективы для академических исследований и цифровых гуманитарных приложений.

В рамках проекта разрабатывается структурная разметка, обеспечивающая идентификацию текстовых элементов (жанровая принадлежность, редакции, комментарии); а также семантическая разметка, включающая идентификацию значимых сущностей: имена, даты, социальный статус, характеристику персонажей, природные явления и др.

Исследование основывается на материале корпуса текстов писателя, включая произведения и письма, их ранние редакции и варианты текста. С опорой на литературоведческие исследования были выявлены значимые сущности для разметки чеховского корпуса [Северина 2020]. Однако вопрос их репрезентации остается сложным и требует комплексного подхода. В этом контексте был рассмотрен текст рассказа А.П. Чехова «Смерть чиновника» [Чехов 1975].

Для выявления значимых сущностей использованы методы автоматической обработки текста, включая Named Entity Recognition (NER) [<https://github.com/natasha/slovnet>], морфологический и синтаксический парсинг (UDPipe 2) [Straka, Straková, Hajič 2019]. Анализ выявил сложности автоматической обработки текста, включая: различные способы репрезентации одной и той же сущности (например, социальный статус может проявляться в имени, местоимении, описании действий персонажа), невозможность автоматического распознавания иронии без филологической интерпретации, сложность разметки переносных значений и многослойных художественных приемов [Ларионова, Шепелева 2019]. Для корректной интерпретации художественных приемов, таких как ирония, оксюморон, метафора, контекстные связи и др., была проведена филологическая экспертиза.

Основные результаты: представлен метод семантической разметки, учитывающий специфику текста А.П. Чехова и позволяющий анализировать как грамматические (морфология, синтаксис), так и семантические элементы текста; показано, каким образом текст А.П. Чехова пронизан контекстными связями и лексико-семантическими особенностями, требующими экспертной разметки (например, фамилии персонажей несут смысловую нагрузку, меняющиеся точки зрения влияют на интерпретацию событий). Разработана система тегирования текстов, позволяющая фиксировать ключевые аспекты: социальный статус персонажа; характеристики героя (например, «ничтожность» Червякова в «Смерти чиновника»); маркировка метафор и оксюморонов; фиксация звуковых элементов (например, «апчхи!!!» как действие без субъекта). Такой подход позволяет сделать филологические знания машиночитаемыми.

Включение в семантическую разметку информации о художественных приемах и значимых сущностях, таких как имена, социальный статус, профессии и характеристики персонажей позволяет разработать цифровые инструменты для сложного семантического поиска, визуализации и анализа текстов [Chekhov Digital].

Разметка Чеховского корпуса позволяет разрабатывать интеллектуальные инструменты анализа текста, полезные для литературоведов, историков и лингвистов. Размеченные тексты могут использоваться в образовательных проектах, цифровых гуманитарных исследованиях и интегрироваться в цифровые сервисы [Сав-

чук и др. 2024]. В перспективе – разработка алгоритмов машинного обучения, обученных на размеченных данных, которая позволит автоматизировать семантический анализ текстов.

Исследование открывает новые перспективы для цифровой филологии. Основные направления дальнейших исследований включают расширение семантической разметки, обучение алгоритмов машинного обучения и развитие междисциплинарного сотрудничества. Проект Chekhov Digital подчеркивает важность междисциплинарного подхода, объединяющего специалистов в области цифровой филологии, лингвистики, информатики и литературоведения. Такое сотрудничество необходимо для автоматической обработки текстов и сохранения филологических знаний в машиночитаемом формате.

Список литературы

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. URL: <https://doi.org/10.23683/1995-0640-2019-1-36-41>.

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Доница О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0 : новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34.

Северина Е. М., Ларионова М. Ч. Новые филологические практики : семантическое издание текстов А. П. Чехова. // Филология : научные исследования. 2020. № 10. С. 13–21.

Цифровой проект Chekhov Digital. URL: <http://chekhov-digital.sfedu.ru/> (дата обращения: 05.02.2025).

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.

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.

Straka M., Straková J., Hajič J. UDPipe at SIGMORPHON 2019 : Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging // Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. 2019. P. 95–103.

TEI Consortium, eds. TEI P5 : Guidelines for Electronic Text Encoding and Interchange. Version 4.9.0. Last updated on 24th January 2025. TEI Consortium. URL: <http://www.tei-c.org/Guidelines/P5/>.