

Богданова Н.А., к. филол. н.
(г. Минск, Беларусь)
МГЛУ
доцент кафедры фонетики
и грамматики немецкого языка

Bahdanava N.A., Cand. of Sc. (Philology)
(Minsk, Belarus)
MSLU

Associate Professor of Department of Phonetics and Grammar of German
e-mail: albertowna@mail.ru

ВОЗМОЖНОСТИ САТМА 6 ДЛЯ ИССЛЕДОВАНИЯ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

В тезисах освещается вопрос программных инструментов для аннотирования текстов и текстовых корпусов, описываются возможности САТМА 6 как браузерного инструмента для ручного тегирования текстов, алгоритм работы и возможности визуализации результатов исследования. Приводится пример анализа текста короткого рассказа. Г. Воман.

Ключевые слова: аннотирование; тег; тегсет; повествовательный регистр; описательный регистр.

POSSIBILITIES OF CATMA 6 FOR FICTION TEXT RESEARCH

The thesis deals with the issue of software tools for the annotation of texts and text corpora, describing the capabilities of CATMA 6 as a browser-based tool for manual text tagging, the algorithm of operation and the possibilities of visualization. An example of textual research on a short story of G. Woman is given.

Key words: annotation; tagging; tagset; narrative case; descriptive case.

Современные технологии, все глубже проникающее в гуманитарные науки, способствовали появлению такого направления как цифровое литературоведение. Особенностью этого направления является видение текста как набора структурированных данных, нелинейное прочтение текста, то есть построение на основе текста определенной модели с последующим анализом и интерпретацией данных. Литературоведение также является одним из направлений, где активно проводятся корпусные исследования на базе как аннотированных (размеченных), так и неаннотированных текстовых корпусов (Ю. С. Холод, Е. В. Парышева, Е. В. Шарапова, А. И. Горожанов).

Аннотирование текстовой информации является одним из наиболее фундаментальных направлений деятельности в лингвистике, вычислительной лингвистике и смежных областях: литературоведении, библиотековедении, цифровых гуманитарных науках. Аннотации – это вторичные данные, добавленные к первичным, текстовым. Являясь частью метаданных, они также включают общую информацию о первичных данных (автор/создатель, модальность, дата создания текста и т. д.). Лингвистические аннотации охватывают любые описательные или аналитические обозначения, применяемые к исходным языковым данным.

Существуют различные типы лингвистической разметки: морфологическая, синтаксическая, семантическая, анафорическая, просодическая и др. Большинство доступных и широко используемых текстовых корпусов имеет морфологическую или синтаксическую разметку. Одним из самых сложных видов разметки является семантическая. В настоящее время работа по аннотированию, как правило, ведется в веб-среде, т. е. в браузере, с использованием одного из веб-инструментов аннотирования (Brat, WebAnno, INCEpTION, CATMA). Текстовые данные, аннотируемые таким образом, могут быть веб-данными, которые, как правило, «неживые», поскольку привязка аннотаций к «живым» веб-документам, которые в любую минуту могут измениться, как незначительно, так и существенно, является технически сложной задачей.

Программный инструмент CATMA (6 версия), созданный в Техническом Университете г. Дармштат – средство цифрового исследования художественных произведений, позволяющее работать с текстами в браузере и создавать проекты для индивидуальных исследований и для групповой удаленной работы. Инструмент имеет четыре базовые вкладки: Проект, Теги, Аннотирование и Анализ. После регистрации, создания проекта и добавления метаданных проекта в рабочее окно загружается текст в форматах .txt, rtf, pdf. Далее создается система лингвистических помет – тегов (тегсет). CATMA позволяет создавать разноуровневую систему тегов путем добавления субтегов к основным тегам в неограниченном количестве и задавать цвета для каждого тега. Во вкладке “Аннотирование” происходит работа по разметке текста. Загруженный текст разбивается системой по умолчанию на несколько страниц, пользователь выделяет отрезки текста и выбирает тег для их аннотации. После завершения аннотирования фрагмент текста маркируется соответствующим цветом. Область аннотирования для работы с тегсетом конкретного проекта позволяет подключать и отключать тегсеты, добавлять, редактировать или удалять теги и субтеги, выполнять двойное аннотирование, аннотирование прерывных текстовых фрагментов.

Вкладка “Анализ” позволяет получать статистическую информацию об аннотированных и неаннотированных данных исследуемого текста: создавать алфавитно-частотный словарь токенов и тегов текста, осуществлять поиск токенов по задаваемым параметрам: по степени сходства токена, частотности, по шаблону слова или фразы, также показывает частотный список аннотированных фрагментов и осуществляет сортировку по тегам и их частотности. Поисковый запрос доступен также по тегам: по их частоте и наполняемости.

В разделе “Визуализация” возможно наглядно представить результаты поискового запроса: проанализировать ключевое слово в контексте (KWIC), визуализировать встречаемость одного или нескольких выбранных фрагментов в тексте с шагом в 10% (DISTRIBUTION), создать облако слов или тегов по выбранным параметрам (WORDCLOUD), проанализировать частотность встречаемости левого и правого окружения фокус-слова (DOUBLETREE). Создаваемые визуализации интерактивны, т.е. они могут быть изменены по заданным параметрам, а также позволяют возвращаться в аннотированный документ из любой точки визуализации.

С целью выявления предпочтений автора в нарративном развертывании сюжета рассказа, а именно — в соотношении репродуктивной и информативной части повествования, текст рассказа Г. Воман «Mein Freund, das neue Jahr» был обработан инструментами САТМА 6. На основе коммуникативных регистров Г. А. Золотовой был создан тегсет из четырех тегов, отражающих способ передачи информации в тексте: репродуктивный и информативный регистры, а также их описательные и повествовательные подрегистры. Текстовые фрагменты, представляющие собой иные коммуникативные регистры, в работе не аннотировались и не рассматривались.

Пример аннотированного фрагмента текста в рабочем окне САТМА 6 представлен на рисунке 1.

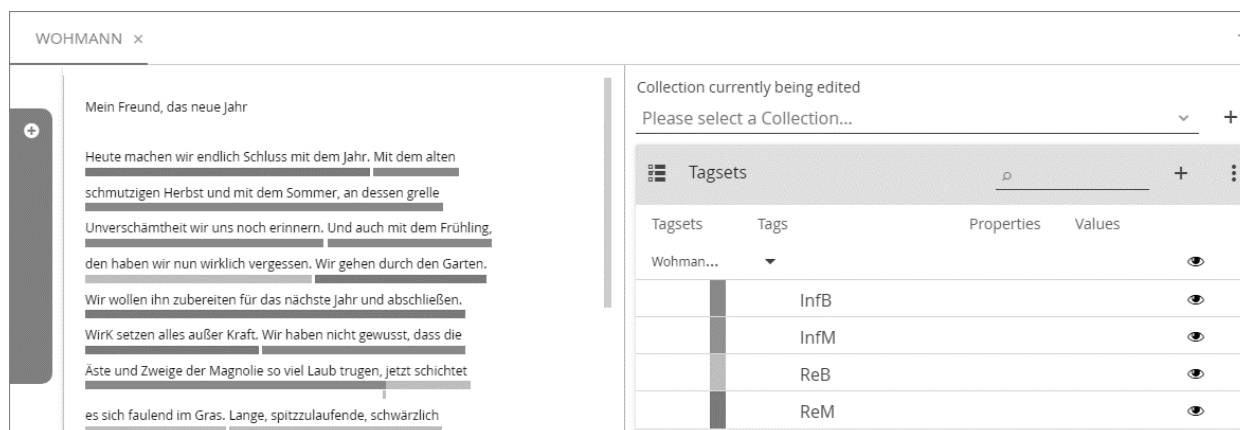


Рис. 1. Фрагмент аннотированного текста Г.Воман.

Всего было аннотировано 57 текстовых фрагментов, из которых 34 относятся к репродуктивному регистру 23 – к информативному. При этом интересен факт распределения подрегистров в данных регистрах: в репродуктивном регистре большая часть фрагментов относится к повествовательному подрегистру, а меньшая – к описательному (22 и 12 соответственно), а в информативном подрегистре, наоборот – большая часть фрагментов представляет собой описательный подрегистр (22 и 1 соответственно). Для нарративного развертывания данного текста характерно непосредственно наблюдаемое описание действий главного, второстепенного и прочего субъектов с одновременным, но несколько меньшим вниманием к непосредственно наблюдаемым описаниям главного, второстепенного, прочего субъектов, объектов, места, времени. При этом автор уделяет достаточно много внимания дополнительным описаниям главного, второстепенного, прочего субъектов, объектов, места, времени, непосредственно не участвующих в сюжетодвижении.