



Корпус подростковых пересказов школьной литературы

Анастасия Файер-Максимова, Джамия Эркенова
15.02.2024



Общая идея проекта

Корпус устных пересказов школьной литературы учениками 9-ых классов во внеучебных условиях

Гипотеза: во внеучебных условиях подростки будут использовать больше сленга в пересказах литературных произведений



Общая идея проекта: исследовательская проблема

- изучение лексикона подростков старшей школы
- изучение проникновения сленга в бытовую лексику
- выделение стратегий пересказа и расставления акцентов в нём
- порождение текстов в стилистике конкретного подсообщества



Общая идея проекта: актуальность

Корпусов, изучающих сленг подростков (в том числе русскоязычных), практически нет. Основные работы, посвящённые сленгу, опираются на собственноручно собранный материал.

При этом существует много словарей современного (разного времени современности) сленга, но нам интереснее корпус, так как в нём есть контекст использования слов.



Ресурсная база проекта

Тексты из программы школьной литературы, изучаемые в 9 классе и ранее

Ожидаемый объём: 20-50 пересказов, в идеале больше

- Аудиозаписи через зум или очно,
- Предварительная автоматическая расшифровка,
- Исправление ошибок автоматической расшифровки
- Семантическая разметка

Ориентировочно корпус только текстовый, аудио (и видео, если будет) останется у нас в качестве исходников.



Подготовительные задачи проекта

- Выбрать конкретные произведения для пересказа, знакомые 9-классникам (предположительно, “Евгений Онегин” и “Герой нашего времени”)
- Записать первый пересказ и разметить его самостоятельно
- Определить формат записи пересказов (самостоятельно или с помощью других школьников для большей свободы пересказывающего)



Подготовительные задачи проекта

- Договориться с учителями о возможности провести эксперимент со школьниками
- Найти инструмент автоматической расшифровки аудио
- Запланировать разметку текстовых данных (в первую очередь, семантическую)



Возможное развитие проекта

- Добавить дискурсивную разметку на основе аудио-файлов
- Увеличить количество пересказов
- Расширить коллекцию пересказанных текстов



Пользовательские истории: использование сленга

Кто: Александра, студентка-лингвистка, 20 лет, пишет дипломную работу об использовании сленга подростками.

Цель: собрать примеры использования сленга подростками.

С чего начала: страница со строкой поиска слов в пересказах ИЛИ страница со списком сленговых выражений с ссылками на пересказы, в которых они употребляются

Характер: методичный, она последовательно изучает корпус и неоднократно к нему возвращается



Пользовательские истории: структура пересказа

Кто: Виталий, доктор филологических наук, 50 лет, пишет научную работу об особенностях построения устных пересказов подростками.

Цель: собрать разные пересказы одних и тех же произведений

С чего начал: со страницы с перечислением произведений и соответствующими им ссылками на пересказы

Характер: методичный, будет неоднократно пользоваться корпусом



Пользовательские истории: генерация речи

Кто: Алина, ML-разработчица, 30 лет, хочет обучить модель использовать подростковый стиль речи

Цель: собрать примеры подростковой речи

С чего начала: со страницы с пересказами

Характер: спонтанный, в ее интересах собрать как можно больше данных, возможно автоматизирует процесс и выгрузит все пересказы с помощью краулера, скорее всего воспользуется корпусом один раз



Пользовательские истории: обучение

Кто: Илья, школьник, 15 лет, узнал о корпусе и захотел почитать пересказы произведений, которые прошел в школе

Цель: прочесть пересказы

С какой страницы начал: со страницы с перечислением произведений и соответствующими им ссылками на пересказы

Характер: спонтанный, перескакивает с пересказа на пересказ, скорее всего воспользуется корпусом один раз из любопытства



Пользовательские истории: слова-паразиты

Кто: Алла, лингвистка-исследовательца, 25 лет, пишет диссертацию об исследовании употребления слов-паразитов в речи

Цель: собрать примеры употребления слов-паразитов подростками

С какой страницы начала: страница со строкой поиска слов в пересказах, страница с частотой использования слов

Характер: методичный, она неоднократно возвращается к корпусу