Марина Кустова
Доклад на семинаре НУГ

Разработка тестов по английскому языку на основе студенческих работ в корпусе REALEC

RETM — инструмент, который генерирует тесты по английскому языку на основе письменных работ самих студентов. Работы с исправленными ошибками берутся из корпуса REALEC.

Что такое REALEC?

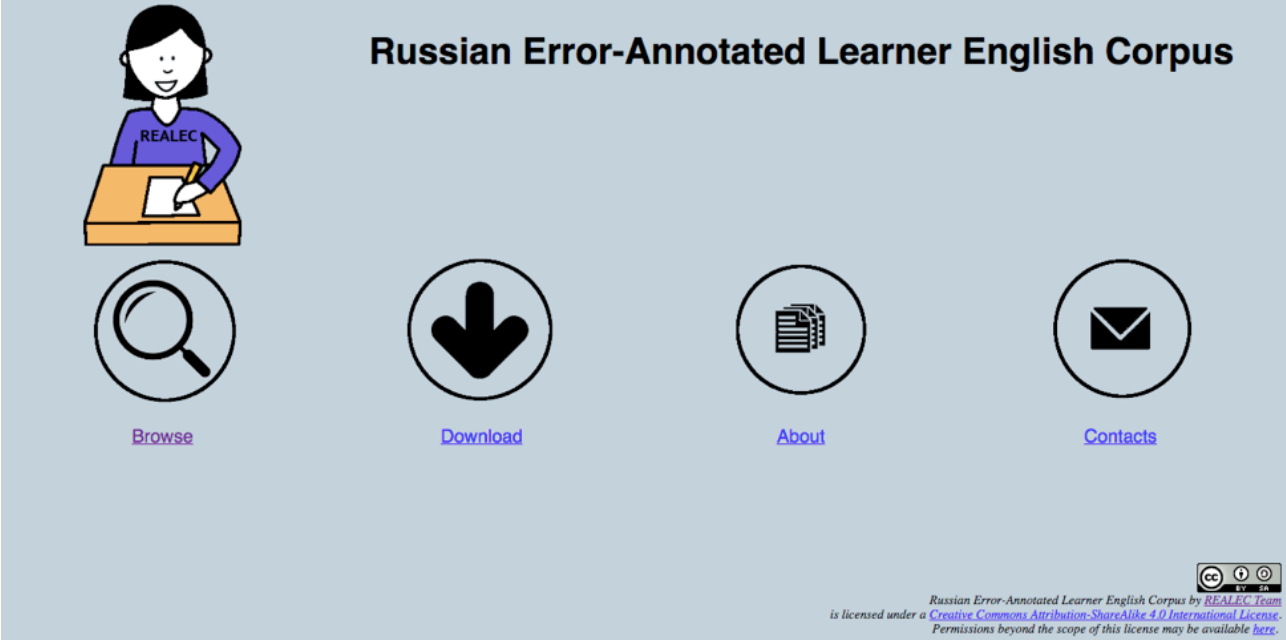

Рис. 1. Главная страница корпуса realec.org

Корпус студенческих эссе на английском языке. Все работы написаны студентами ВШЭ. Ошибки в каждом эссе исправлены преподавателями английского языка. Для разметки и исправления ошибок используется brat.

Что такое brat?

brat rapid annotation tool
Инструмент аннотирования, позволяющий создавать собственные схемы разметки и добавлять аннотации сразу на нескольких уровнях.

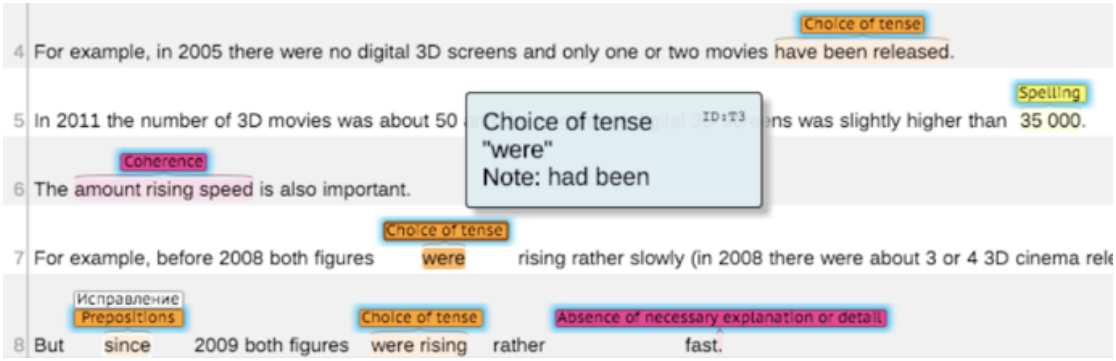

Рис. 2. Пример разметки в brat

Как аннотируются тексты в REALEC?

В корпусе есть подробная классификация ошибок, составленная опытными преподавателями и включающая более 150 тегов для ошибок различных видов. Кроме типа ошибки и исправления, аннотация в REALEC может также включать информацию о предполагаемой причине ошибки и о том, насколько она искажает исходный посыл текста.

Как из аннотации получаются тесты?

Тесты генерируются на основе исходного текста и того слоя аннотации, который включает тип ошибки и исправление, предложенное преподавателем. Кроме того, если пользователю нужен тест не на все, а на одну или несколько определённых тем (например, артикли или время глагола), он сообщает программе соответствующие теги ошибок.

Сначала RETM разделяет неисправленный текст на предложения. Это нужно для того, чтобы каждый вопрос в тесте имел вид «исправьте ошибку в предложении». Все последующие действия программа проводит с каждым предложением по отдельности.

Если в предложении есть ошибка с нужным тэгом, программа исправляет все ошибки, кроме этой. После этого программа объединяет предложение, в котором исправлены все ошибки, кроме одной, и исправление для этой ошибки, в пару. Каждая такая такая пара представляет собой будущее тестовое задание, где предложение с ошибкой — это вопрос, а исправление — правильный ответ. Если в предложении несколько ошибок требуемого вида, то действия повторяются для каждой из них.

После того, как вышеописанные действия выполнены для каждого предложения в тексте, в памяти программы хранится некоторое количество пар «вопрос-ответ». Программе остаётся только сохранить их в формате, пригодном для дальнейшего использования. Для RETM таким является особый XML-формат, используемый в системе Moodle, где в дальнейшем будут проводиться тесты.