А.В.ГУСЕВА., М.П. КОНЦЕВОЙ
БрГУ им. А.С. Пушкина (г. Брест, Беларусь)

МОДЕЛИРОВАНИЕ ЛИНГВИСТИЧЕСКОЙ РАЗМЕТКИ В OPEN CORPORA

Open Corpora (http://opencorpora.org) – проект создания аннотированного лингвистического корпуса русскоязычных текстов. Open Corpora предназначен для исследователей языка (применяющих корпусный инструментарий) и для разработчиков систем автоматической обработки текста (изучающих, редактирующих и использующих корпусные базы в целях создания нового инструментария для лингвистического исследования). В Open Corpora моделирование морфологической, синтаксической, семантической разметки осуществляется самими пользователями на основе технологии краудсорсинга (сетевой организации волонтерской работы сообщества над какой-либо задачей ради достижения общих благ). Open Corpora предоставляет возможность участия в различных видах разметки: морфологической (tagging, part-of-speech tagging), сопоставляющей каждому слову в тексте его словарную форму с указанием грамматических характеристик слова; разметке сущностей (выделению и тегированию в текстах онимов и названий различного типа).

Осуществление разметки предполагает базовое владение русским языком. Разметка текстов в рамках проекта Open Corpora силами сетевого сообщества может быть использована в образовательных целях в контексте языковой и лингвистической подготовки учащихся и, как показывает опыт такого использования, обладает значительным и разноплановым лингводидактическим потенциалом, а именно: реализует практическое взаимодействие учащихся с корпусными технологиями как одним из наиболее эффективных и современных инструментов лингвистического исследования; предполагает повторение и закрепление грамматики русского языка при непосредственном проведении учащимися разметки предлагаемых текстов; осуществляется на основе современных информационных и коммуникационных технологий, осваиваемых учащимися в конкретной практической деятельности: практическое знакомство и использование программного обеспечения организации коллективной удаленной работы над снятием морфологической неоднозначности (морфологический словарь, полуавтоматический токенизатор); открывает перед учащимися существенное различие между машинным и человеческим подходами к решению трудно формализуемых и алгоритмизируемых задач в области естественного языка; предполагает на выходе получение не просто учебного, но законченного общественно-востребованного продукта, что открывает возможности социально значимой продуктивной деятельности.