

КОРПУСНЫЕ ИССЛЕДОВАНИЯ В ДАГЕСТАНЕ

(на русском языке)

Муталов Р.О. (mutalovr@mail.ru), профессор, Дагестанский государственный университет, Махачкала, Республика Дагестан, Россия

Доклад посвящен проблемам разработки Национальных корпусов аварского, даргинского, лезгинского, лакского, кумыкского, табасаранского языков. Проведены работы по созданию электронных библиотек и электронных словарей данных языков. Подготовлена также информация, необходимая для проведения метаразметки текстов – собраны сведения об авторах, определены внешние параметры текстов, проведена их типизация.

Основное внимание в ходе работы уделяется разработке механизмов лингвистической разметки. Предпринята попытка создания упрощенной системы автоматической разметки текстов. Исходя из лексического значения и параметров словоизменения, все слова распределены на несколько групп. В одну группу объединены слова с полностью идентичными грамматическими признаками и близкие по семантике. Каждой словоформе приписываются морфологические значения (исходная форма слова; принадлежность к той или иной части речи) и принадлежность к определенной семантической группе. Затем даются словоизменительные признаки словоформы.

Специально разработанная для решения данной проблемы программа заменяет словоформу в тексте другой словоформой, имеющей морфологическую и семантическую разметку. При запросе нужного слова появляются предложения с данной словоформой, имеющие метаразметку, а морфологические и семантические значения слова можно извлечь при нажатии курсора мыши на словоформу. Хотя при разметке возникают проблемы, связанные как с омонимией и сложностью морфологической структуры дагестанских языков, так и с современной орфографией, применение механизма автоматической разметки текстов для создания корпусов малых языков представляется обоснованным и эффективным.

В ходе доклада будет продемонстрирован фрагмент Национального корпуса даргинского языка.