Опыт создания корпусов дагестанских языков[1]

An experience of creation of the national corpus of Dagestan languages

Муталов Р. О. (mutalovr@mail.ru)

Дагестанский государственный университет,
Махачкала, Республика Дагестан

В докладе рассматриваются проблемы и перспективы национальных корпусов текстов шести литературных языков Дагестана, работы по созданию которых ведутся в Дагестанском государственном университете. Особое внимание уделено проблемам создания системы автоматической разметки текстов и переводу печатных текстов в электронный формат.

Корпусная лингвистика в настоящее время является одной из быстроразвивающихся областей компьютерной лингвистки. Созданы корпуса текстов большинства распространенных языков мира; по многим языкам создаются параллельные корпуса, диалектные корпуса, корпуса устной речи, корпуса поэтической речи и т.д. Однако, как в России, так и за рубежом, работы по созданию корпусов текстов языков малочисленных народов находятся пока на начальной стадии. Исследовательской группой под руководством А.Е.Кибрика разработаны конкретные требования к современным корпусам текстов для малых языков [Кибрик и др. 2007]. Высокий уровень результатов достигнут при составлении корпусов хиналугского и алюторского языков, электронной грамматики и электронного словаря арчинского языка.

В дагестанском государственном университете ведутся работы по созданию Национальных корпусов шести литературных языков Дагестана — аварского, даргинского, лезгинского, лакского, кумыкского, табасаранского [Муталов 2007]. Разработку национальных корпусов дагестанских языков можно отнести к первым опытам создания корпусов текстов малых языков России. Необходимость скорейшего ввода в информационное поле материала данных языков заключается в том, что эти языки отнесены специалистами к языкам, обреченным в будущем на исчезновение; поэтому одной из важных задач создания корпусов является содействие в решении проблемы сохранения и развития национальных языков Дагестана. Следует также отметить, что дагестанские языки, обладая сложной морфологической системой и являясь языками эргативного типа, представляют для лингвистической типологии особый интерес. Тексты с лингвистической разметкой станут базой для создания современных научных грамматик, словарей и учебников. Корпуса текстов в перспективе станут также основой для разработок параллельных русско-дагестанских корпусов.

Основную часть текстов национальных корпусов дагестанских языков — до 75%, будут составлять художественные произведения дагестанских авторов. Будут также представлены и другие тексты: драматургия, мемуарно-биографическая литература, журнальная публицистика и литературная критика, газетная публицистика, научные, учебные, религиозные, юридические тексты, деловые и бытовые тексты. Они должны стать базой создания национальных корпусов. Предполагается, что объем создаваемых корпусов будет составлять от 3 до 5 млн. словоупотреблений по каждому языку.

Создателям корпусов приходится сталкиваться с трудностями, связанными как с малым количеством текстов в электронном формате, так и отсутствием механизмов разметки текстов каждого языка. Создатели корпусов освоили навыки оцифровки текстов, их сканирования и последующей обработки, вычитки и исправления ошибок, овладели технологией создания корпусов текстов. Работа по созданию электронных коллекций текстов ведется в двух направлениях: сбор имеющих в электронном виде текстов и перевод в электронный формат печатных текстов. Собраны воедино тексты, уже имеющие электронный формат; это, как правило, произведения авторов последних лет, созданные в компьютерную эру. Одна-

[1] Работа поддержана РФФИ, грант № 07-06-00460-а.

ко подавляющее большинство текстов на дагестанских языках написано в 60–90-е годы прошлого века и представлено в печатном виде. Трудоемкую техническую работу по их сканированию удалось ускорить посредством применения высокоскоростного планшетного сканера А3, позволяющего сканировать, помимо стандартных изданий, широкоформатные литературные журналы-альманахи. Для распознавания отсканированных текстов была применена последняя, девятая версия FineReader. Распознаванию сложных графем современной графики дагестанских языков, состоящих из нескольких, порой 3–4 знаков, способствовало применение созданных для решения данной проблемы специальных шрифтов. При вычитке текстов особое внимание обращалось удалению в словах знаков переноса и устранению «жестких» концов строк. Создаются электронные архивы текстов, представляющих собой отсканированные, но не обработанные тексты. Вычитка данных текстов и исправление имеющиеся в них ошибок и опечаток составляют значительную по объему часть работы.

В составлении инвентаря грамматических помет, созданного на основе латинской графики, и в сокращениях, принятых для обозначения определенных грамматических категорий, создатели корпусов, хотя и следовали рекомендациям известных «Лейпцигских правил глоссирования», значительно дополнили и уточнили список морфологических категорий, встречаемых в дагестанских языках.

Проведены работы по созданию электронных библиотек и электронных словарей дагестанских языков, что в комплексе должно стать основой для проведения грамматической разметки текстов. Подготовлена также информация, необходимая для проведения метаразметки текстов — собраны сведения об авторах, внешних параметрах текстов, проведена их типизация.

Поскольку наиболее важной частью создания корпусов является разметка текстов, основное внимание уделяется разработке механизмов лингвистической разметки текстов. Разметка текстов малых языков, как в России, так и за рубежом, делается большей частью вручную. По корпусам больших языков, количество словоупотреблений которых составляет несколько сот миллионов слов, разрабатываются специальные парсеры. Для создания парсеров по каждому корпусу дагестанских языков с объемом словоупотреблений в 3–5 млн. словоупотреблений, аналогичных парсеру Национального корпуса русского языка или корпусов других языков, естественно, не хватает ни людских, ни финансовых ресурсов. С другой стороны, не представляется также реальным проводить ручную разметку всех текстов корпуса. Поэтому была предпринята попытка создания собственной, упрощенной системы автоматической разметки текстов.

Для начала были созданы электронные словари, содержащие функционирующие в языке основные лексемы. Исходя из лексического значения и параметров словоизменения, все слова были распределены на несколько групп. В одну группу объединялись слова с полностью идентичными грамматическими признаками и близкие по семантике. Каждой словоформе приписываются следующие морфологические значения: исходная форма слова; принадлежность к той или иной части речи; семантическая группа. Затем даются словоизменительные признаки словоформы; для именных частей речи и наречия — это информация о классе, числе и падеже. Например, даргинскому слову *дудешлис* «отцу» приписывается значение мужского класса, дательного падежа, единственного числа. Для глаголов указывается информация о классе, числе, лице, виде, переходности, времени, наклонении; для отглагольных образований — причастия, деепричастия (деепричастия места) и масдара, помимо перечисленных признаков, указывается также и падеж. Здесь же дается информация о нестандартных словоформах; таковы, например, даргинские глаголы *гес* «дать», *хес* «принести», *кес* «привести».

Разработанная специально программа заменяет словоформу в тексте другой словоформой, имеющей морфологическую и семантическую разметку. При запросе нужного слова появляются предложения с данной словоформой, имеющими метаразметку, а морфологические и семантические значения слова можно извлечь при нажатии курсора мыши на словоформу — всплывают ее словоклассифицирующие и словоизменительные признаки. Здесь же появляется также информация о принадлежности слова к той или иной семантической группе.

Естественно, при автоматической разметке текстов и приписывании каждой словоформе морфологической информации возникает ряд проблем. К примеру, одна из них связана с омонимичностью классно-числовых показателей дагестанских языков. Следует отметить, что среди шести классных индикаторов в даргинском языке имеются три пары омонимичных показателей: классификатор *б* служит для обозначения среднего класса единственного числа и 3-го лица множественного числа мужского и женского классов; показатель *д* служит классификатором 1 и 2-го лиц множественного числа непереходных глаголов и 3-го лица среднего класса множественного числа. Омонимию классно-числовых показателей разметчик должен снимать вручную. В ряде дагестанских слов на современном этапе развития языков классные показатели окаменели; информация о выражении ими значения грамматического класса в таких случаях не дается.

При указании информации по категории числа имен существительных внимание обращается на группы слов, имеющих лишь форму единственного или множественного числа. Существительные, всегда функционирующие в форме единственного числа, указываются как слова «без формы множе-

ственного числа». Существительные же, имеющие лишь форму множественного числа, а также «собирательные существительные» указываются как существительные «нерасчлененной совокупности».

Хотя при разметке возникают проблемы, связанные со сложностью морфологической структуры дагестанских языков, или проблемы, создаваемые орфографическими правилами, (такими, как, например, слитное написание некоторых слов служебных частей речи с предшествующим словом), применение механизма автоматической разметки текстов для создания корпусов малых языков представляется обоснованным и эффективным. Она позволяет решать первостепенные задачи поиска лингвистической информации о слове. Система автоматической разметки текста была применена к нескольким текстам различных дагестанских языков, после чего пробные образцы размеченных текстов были перенесены на сервер Лаборатории лингвистических исследований ДГУ.

В перспективе предполагается продолжить работы по пополнению имеющихся электронных библиотек новыми текстами, а также по переводу в электронный формат печатных текстов. Оцифрованные тексты будут вычитаны, имеющиеся ошибки исправлены. Предстоит усовершенствовать и доработать систему автоматической разметки текстов. Будут продолжены работы по морфологической, семантической и экстралингвистической разметке текстов, а также ручное снятие омонимии. Размеченные подобным образом национальные корпуса и электронные словари дагестанских языков предполагается разместить в режиме открытого доступа в Интернет-сети.

Литература

1. *Кибрик А.Е., Архипов А. В., Даниэль М. А., Кодзасов С. В., Майерс Том, Нахимовский А. Д.* Технологии обработки языковых данных в документировании малых языков // М.: Материалы Международной конференции «ДИАЛОГ 2007» «Компьютерная лингвистика и интеллектуальные технологии», 2007.
2. *Муталов Р. О.* Корпусная лингвистика и перспективы ее развития в Дагестане // Махачкала: Современные проблемы кавказского языкознания, 2007. Вып. 7, С. 160–173.
3. *Плунгян В. А.* Зачем нужен Национальный корпус русского языка? неформальное введение. // М.: Национальный корпус русского языка: 2003–2005. Результаты и перспективы, 2005.