МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12–06–06061–г),
гранта РГНФ (проект № 12–04–14154–г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:
В. А. Баранов, д-р филол. наук, проф.
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

И741 **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

МОДЕЛИ И ТЕХНОЛОГИИ РЕАЛИЗАЦИИ ТЮРКСКИХ КОМПОНЕНТ В РУССКО-ТАТАРСКОЙ ЛЕКСИКОГРАФИЧЕСКОЙ БАЗЕ ДАННЫХ1

О. А. Невзорова1, А. Р. Гатиатуллин1, Р. А. Гильмуллин1,
Б. Э. Хакимов2
1НИИ «Прикладная семиотика» Академии Наук Республики Татарстан;
2Казанский (Приволжского) федеральный университет, Казань

The article is presented the base models of new Russian-Tatar lexicographical database. Lexicographical database consists of interrelated components (Russian and Tatar) with an independent structure. The components are merged by semantic codes at the level of lexical equivalents. Each component contains a grammatical, semantic and derivational information. Component of the new Turkic language will have a structure similar to the structure of the Tatar component. The article discusses the basic design problems of new Turkic components and also technologies of extension of lexicographic database by adding new components.

Для теоретических исследований по компьютерной лингвистике и раз-работки лингвистических приложений (систем машинного перевода, инфор-мационного поиска и др.) большое значение имеет создание специализиро-ванных лексикографических ресурсов с детализированными лингвистиче-скими аннотациями. Функциональные особенности таких лексикографиче-ских ресурсов определяются кругом потенциальных задач, для решения ко-торых они проектируются.

В настоящее время создано значительное количество многофункцио-нальных лексикографических ресурсов. Для тюркских языков можно отме-тить параллельный онлайн-словарь тюркских языков Турецкого лингвисти-ческого общества (Turk Dil Kurumu) [Türk Dil Kurumu, 2012], проект машин-ного фонда башкирского языка [Машинный, 2012], электронные словари ка-захского языка [Русско-казахский, 2011]. Для татарского языка еще в 90-х годах прошлого века была разработана концепция машинного фонда [Буха-раев, 1995] и разработаны различные электронные словари (электронные словари ABBYY Lingvo, интернет-словари языка татарских писателей [Ка-занский, 2012] и др.). Однако актуальной является задача разработки специа-лизированных лексикографических ресурсов для различных теоретических и прикладных целей, таких, как сравнительное изучение тюркских языков, лингвистическое обеспечение систем машинного перевода и других при-кладных лингвистических технологий.

В НИИ «Прикладная семиотика» Академии наук Татарстана ведется разработка русско-татарской лексикографической базы данных [Невзорова, 2012] на основе русско-татарского словаря объемом около 50 000 слов под редакцией Ф.А. Ганиева [Русско-татарский, 1997]. Новая лексикографиче-ская база данных (ЛБД) ориентирована, прежде всего, на приложения в об-ласти автоматической обработки текстов (аннотирование корпусов текстов, машинный перевод и информационный поиск). Разработанная модель ЛБД позволяет расширять данный лексикографический ресурс на другие тюрк-ские языки. Подключение новых тюркских компонент выполняется на осно-ве моделей и технологий, разработанных для татарской компоненты базы данных.

Лексикографическая база данных состоит из взаимосвязанных русской и татарской компонент, имеющих независимую структуру и объединяемых при помощи семантических кодов на уровне лексических эквивалентов. Компонента для другого тюркского языка будет иметь структуру, подобную структуре татарской компоненты, и связь с русской и татарской компонен-тами будет осуществляться с помощью семантических кодов.

Каждая из компонент содержит грамматическую, семантическую и словообразовательную информацию. Грамматическая часть татарской компо-ненты (Т-компоненты) представляется двумя словарями: словарем основ и словарем окончаний. Словарь основ содержит такие параметры, как семан-тический код, словарная форма основы, морфологическая форма основы, морфологический и морфонологический типы основы. Морфологическая форма используется при порождении поверхностных форм словоформы путем присоединения аффиксальных морфем и может не совпадать со словар-ной формой основы за счет каких-либо внутренних изменений (например, чередований или отсутствия символов).

Одно из назначений семантического кода — обеспечение связи различных компонент, что особенно важно при расширении лексикографической базы на другие тюркские языки. Морфологический и морфонологический типы основы необходимы для связи со словарем окончаний и используются программами морфологического анализа и генерации.

Процесс добавления в лексикографическую базу данных новой компоненты для другого тюркского языка будет состоять из двух этапов: разработ-ка словаря основ и разработка словаря окончаний.

При построении словаря основ главная проблема заключается в практическом отсутствии двуязычных словарей между языками тюркского семей-ства (за небольшим исключением в виде татарско-турецкого и турецко-татарского словарей). В то же время имеется множество тюркско-русских двуязычных словарей, таких, как русско-казахский, русско-узбекский, рус-ско-киргизский и др., которые и могут быть положены в основу проектиро-вания новых тюркских компонент ЛБД.

Создание словаря основ для новой компоненты состоит из двух этапов: автоматического и ручного. Во время автоматического этапа производится сопоставление содержимого русской компоненты ЛБД с входами словарных статей двуязычного русско-тюркского словаря. При обнаружении совпа-дающих лексем в тюркской компоненте создается новая запись с тем же се-мантическим кодом, что и у совпавшей лексемы в русской компоненте ЛБД. Если для русской лексемы русско-тюркского словаря аналогичная лексема в русской компоненте лексикографической базы не находится, то для новой записи тюркской компоненты генерируется новый уникальный семантиче-ский код.

После завершения автоматического этапа сопоставления построенная БД требует дополнительной филологической экспертизы. Процесс разработ-ки словаря окончаний новой тюркской компоненты начинается с построения таблицы соответствий между аффиксальными морфемами добавляемого и татарского языков. Затем разрабатываются морфотактические правила для генерации всех возможных словоформ нового языка с учетом принятых ог-раничений на количество аффиксальных морфем (в настоящей версии не бо-лее 5). При разработке морфотактических правил параллельно происходит выделение всех морфологических и морфонологических типов основ нового языка. В завершении все указанные виды окончаний автоматически генери-руются и заполняются в таблицы БД со структурой, аналогичной структуре татарской компоненты ЛБД.

Реализация дополнительных тюркских компонент в ЛБД позволит получить эффективный инструмент автоматизации сравнительных исследова-ний и разработки многоязычных лингвистических технологий, прежде всего, технологий многоязычного поиска и машинного перевода для близкородст-венных языков внутри тюркского семейства.

Для автоматизации сравнительно-типологических исследований тюркских языков и реализации интегрированных тюркских компонент в лексикографической базе данных необходима общая модель морфологии. Подобная модель может быть создана путем расширения модели татарской морфологии, реализованной в ЛБД. Морфологические параметры словоформ в тюркских языках можно подразделить на простые и сложные [Хакимов, 2011]. Простые параметры представлены единственным значением (возможно наличие, либо отсутствие параметра). Сложные параметры представлены двумя и более альтернативными значениями и могут быть обязательными или факультативными. Факультативные параметры отличаются от обязательных тем, что могут отсутствовать в словоформе. Все простые параметры являются факультативными. Описание тюркских грамматических категорий по данным критериям позволяет отразить специфику каждого языка в рамках универсальной классификации. В свою очередь, усиление горизонтальных и межуровневых связей между разноязычными компонентами дает возможность установления более сложных соответствий, например, в тех случаях, когда одна и та же функция в родственных языках реализуется единицами разных уровней (морфемы, слова, словосочетания и др.).

Список литературы

Türk Dil Kurumu, 2012 — Türk Dil Kurumu [Электронный ресурс]. URL: http://tdk.org.tr (дата обращения: 10.03.2012).

Машинный, 2012 — Машинный фонд башкирского языка [Электронный ресурс]. URL: http://mfbl.ru (дата обращения: 15.02.2012).

Русско-казахский, 2011 — Русско-казахский и казахско-русский словарь [Электронный ресурс]. URL: http://sozdik.kz (дата обращения: 21.10.2011).

Бухараев, 1995 — Бухараев Р.Г., Сафиуллина Ф.С., Сулейманов Д.Ш. и др. К концепции Машинного Фонда Республики Татарстан // Татарский язык и новые информационные технологии. Серия: Интеллект. Язык. Компьютер. Вып.2. Казань, 1995. С. 20–35.

Казанский, 2012 — Казанский лингвографический фонд [Электронный ресурс]. URL: http://klf.ksu.ru (дата обращения: 5.04.2012).

Невзорова, 2012 — Невзорова О.А., Салимов Ф.И., Хакимов Б.Э., Гатиатуллин А.Р. Организация информационного поиска в русско-татарской лексикографической базе данных. // Прикладная лингвистика в науке и образовании: Сборник трудов VI международной научной конференции. СПб., 2012.

Русско-татарский, 1997 — Русско-татарский словарь / под ред. Ф.А. Ганиева. М.: ИНСАН, 1997.

Хакимов, 2011 — Хакимов Б.Э., Гильмуллин Р.А. К разработке морфологического стандарта для систем автоматической обработки текстов на татарском языке // Системный анализ и семиотическое моделирование: материалы всероссийской конференции с международным участием (SASM-2011). Казань, 2011. С. 209–214.