

Лексический атлас русских говоров Удмуртии в Интернет

И. С. Соломенников

Удмуртский государственный университет, Ижевск, Россия

The technologies of storage and processing of the dialect material collected under special programs for the lexical atlas of Russian national dialects are considered. The realization of the system — the Lexical Atlas of the Russian National Dialects intended for remote work with electronic depositaries of data — and also the developed technique of construction of the linguistic-geographic maps of the territories where data was collected are described.

В сборе диалектного материала по программе Лексического атласа русских народных говоров (головная организация — Институт лингвистических исследований РАН, Санкт-Петербург) Удмуртский государственный университет участвует с 1990 года¹. За это время по ее разделам обследованы все районы Удмуртской республики, собрано огромное количество материала. Долгое время основным средством хранения информации являлись тетради, в которые данные заносятся при обследовании информаторов. Филологами УдГУ уже предпринимались попытки создания электронного хранилища для собираемого диалектного материала, но развития они по ряду причин не получили.

В то же время задача электронного хранения, редактирования и обработки диалектных данных является актуальной, так как в противном случае научная ценность собранного материала не может быть использована в полной мере из-за сложности работы с ним. Именно поэтому возникла идея разработать и создать систему на основе современной реляционной СУБД, которая бы обеспечивала

¹ Баранов и др. 1998 — Баранов, В. А. Методические указания к диалектологической практике / В. А. Баранов, С. А. Никифорова, Н. Е. Якименко. — Ижевск : Удм. гос. ун-т, 1998. — 82 с.

*Современные информационные технологии и письменное наследие:
от древних рукописей к электронным текстам*

ввод и хранение, а также необходимую обработку и картографирование материалов диалектных обследований.

Основой такой системы является хранилище данных, разработка которого было первым этапом работы. Было принято решение хранить в базе данных не только сам диалектный материал, но и всю необходимую для полноценного анализа дополнительную информацию — о месте, времени и других обстоятельствах фиксации данных. Это позволяет проводить обработку с учетом всех лингвистических и экстралингвистических свойств и значений. Все собираемые данные в базе кодируются в формате Unicode, что обеспечивает возможность хранения материала не только на русском, но и на любом другом языке.

После создания базы была разработана интерфейсная часть системы. Чтобы обеспечить возможность работы с ней удаленным пользователям, было решено сделать доступ к данным через Интернет. При этом все формы для ввода реализуются стандартными средствами HTML и доступны на сайте в сети Интернет. Такой подход имеет несколько существенных достоинств:

- возможность работы независимо от платформы клиентского компьютера (Windows, Linux, MacOS);
- нетребовательность к ресурсам компьютера пользователя;
- простота системы в обновлении и поддержке.

Так как в функции интерфейсной части системы входят ввод данных в систему и их обработка (за хранение отвечает СУБД), хорошим решением является разделение ее на две части: административную (ограниченный доступ для ввода и редактирования материалов) и пользовательскую (общедоступный инструмент для проведения исследований по хранимым данным).

Функциональность административной части системы предусматривает:

- ввод в базу данных собираемых диалектных материалов вместе с определенной дополнительной информацией о них;
- занесение в систему и редактирование вопросников программы ЛАРНГ;
- ведение дополнительных справочников, необходимых для функционирования атласа;
- доступ для авторизованных пользователей.

При реализации обращения к справочникам необходимо сделать работу максимально удобной, что было достигнуто оформлением их на страницах в виде таблиц с данными. При такой организации интерфейса работа с системой становится максимально простой и легкой в освоении даже для неспециализированного пользователя. Ввод материала в систему также максимально упрощен и позволяет осуществлять наполнение базы достаточно быстро.

Под инструментом проведения исследований понимается модуль, позволяющий формировать произвольные запросы с учетом всех свойств и значений данных. Этот инструмент имеет средства для упорядочивания, выборки, операций с выборками, для построения словников, для отображения результатов в различной форме. Основное его предназначение состоит в автоматизации картографирования лексики русского диалектного языка и динамического визуализирования результатов обработки материала на карте.

Разработанная технология позволяет пользователю при построении лингвогеографических карт самостоятельно определять как картографируемый материал, так и обозначения, которые будут использоваться на карте. Обработка материала проводится в несколько шагов, которые позволяют пользователю не только выбрать материал, но и удалить данные, которые не представляют интереса.

Результатом проделанной работы является создание лингвогеографической информационной системы с доступом к данным через Интернет. Данная разработка масштабируется на любые территории и на любые типы и объемы информации.

Учет регионального компонента на примере элементов шорской культуры и программы MS WORD

О. А. Соседко

Кузбасская государственная педагогическая академия,
Новокузнецк, Россия

Чаще всего разработчики программного обеспечения учитывают или популярные, или глобальные национальные приоритеты. В частности, русифицировано большинство программных продуктов для российских пользователей. Но мир языков многообразен. Например, в кузбасском регионе есть малая народность — шорцы, обладающая своей культурой и языком. Поэтому в настоящее время необходимо адаптировать программные продукты и для нужд представителей малых народов.

Чтобы понять, какие проблемы возникают перед пользователем-шорцем, нужно познакомиться с шорским алфавитом.

Современный шорский алфавит основан на кириллице, но имеет ряд специальных букв, в большинстве сходных с буквами тюркских алфавитов. Эти буквы имеют дополнительные элементы — «хвостики», «черточки» или «двоеточие» над буквой.

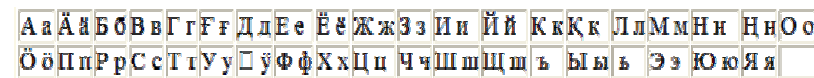


Рис. 1. Современный шорский алфавит

Как видно, проблемой является ввод спецбукв, так как для этих букв нет соответствующих клавиш на клавиатуре.

Проанализируем возможные решения проблемы.

1. Изменить наклейки тех клавиш, символы которых не используются в шорском языке. Перепрограммировать эти клавиши, чтобы при нажатии появлялись на экране эти спецсимволы. Иначе говоря, можно разработать по аналогии с русификатором программу «шорификатор». Недостаток этого способа состоит в том, что за