Корпусная лингвистика, компьютерная лексикография, мультимедийные технологии и исчезающие языки (Проект завершен – да здравствует новый проект!)[1]

О. А. Казакевич (kazak@orc.ru)
Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова

Л. М. Захаров (leon@philol.msu.ru)
Филологический факультет МГУ им. М.В. Ломоносова

И.В. Самарина (Ira_samarina@hotmail.com)
Институт языкознания РАН

Д.Л.Трушков (morion-tr@mtu-net.ru)
ООО ЛУКойл-нефтехим

В докладе предполагается рассказать о результатах только что завершенного проекта «Говоры северных селькупов: сопоставительное описание и база данных звуковых файлов» (грант РФФИ № 01-06-80363). Опыт реализации этого проекта позволяет нам сделать некоторые обобщения относительно принципов построения корпусов текстов и компьютерных словарей языков, функционирующих преимущественно или исключительно в устной форме при постоянном сужении сферы и объема их функционирования, с учетом возможностей, предоставляемых современными компьютерными технологиями. Существенной частью проекта были экспедиционные работы по сбору лингвистического и около-лингвистического (социолингвистического, этнологического) материала - документации исчезающего языка. При сборе материала мы также старались использовать современные технологии его фиксации: цифровые аудио- и видеозаписывающие устройства. Одной из своих задач мы считали представление в компьютерной базе всего спектра функционирования языка – от речи пожилых компетентных носителей (записи спонтанной речи и фольклора) до речи едва владеющих языком молодых людей. Это дает возможность лучше понять природу не только внешних изменений объема функционирования языка, но и изменений в языковой структуре, которые происходят буквально на наших глазах. В ходе работы над проектам был создан «Озвученный словарь говоров северных селькупов». Запись материалов для этого словаря, направленная, казалось бы, на сбор исключительно лексики, оказалась способом получения ценнейших данных о функционировании языка на всех уровнях. При этом, поскольку в работе со всеми информантами (а их было более 40 человек, примерно по 10 человек на каждый из четырех говоров) использовался один и тот же словник, данные эти легко сопоставимы как по возрастным группам в пределах одного и того же говора, так и между говорами. Таким образом, побочным продуктом проекта стало что-то вроде тестирования языковой компетенции носителей обследовавшихся говоров, что весьма важно для оценки реального положения языка

[1] Доклад подготовлен в рамках проекта «Мультимедийная база данных кетского языка», реализуемого при финансовой поддержке РГНФ, грант № 04-04-12028в.

и перспектив его дальнейшего функционирования. Доклад будет сопровождаться демонстрацией аудио- и видеоматериалов из компьютерной базы данных.

В заключение доклада предполагается рассказать о новом проекте «Мультимедийная база данных кетского языка», над которым группа начала работать в 2004 г. (грант РГНФ № 04-04-12028в) и который также связан с документацией исчезающих языков и организацией мультимедийного компьютерного архива.

Введение

Язык – это часть окружающей среды, разрушение которой весьма чувствительно для человека и может иметь весьма серьезные последствия для социума. В настоящее время во всем мире малые языки подвергаются мощному давлению со стороны доминирующих языков. В случае России доминирующим языком является русский. Поскольку в основе нашей цивилизации лежит многообразие, в том числе и языковое, сохранение этого многообразие необходимо для сохранения нашей цивилизации и избегания кризисных ситуаций. Только что (2003 г.) завершенный в НИВЦ МГУ им. М.В. Ломоносова проект «Говоры северных селькупов: сопоставительное описание и база данных звуковых файлов» (грант РФФИ № 01-06-80363) и был, собственно, направлен на сохранение языкового многообразия или, по крайней мере, памяти о нем для будущих поколений. За время реализации проекта собран уникальный видео- и аудиоматериал по говорам северных селькупов и разработан модельный образец организации этого материала. Некоторые результаты завершенного проекта представлены в докладе.

О ходе реализации проекта рассказывалось в публикациях семинара «Диалог» [Казакевич 2001; Казакевич и др. 2002]. Здесь мы сосредоточимся на общих итогах и «побочных результатах», а также на полученном опыте, который может быть использован при работе над аналогичными проектами. В заключение рассказывается о новом проекте, в работе над котором мы как раз и собираемся использовать наш опыт, полученный за три предшествующих года.

В нашей стране это не первый проект создания озвученных словарей. В Санкт-Петербурге созданы озвученные ненецкий и нганасанский словари и ненецко-русский разговорник (см. [Люблинская 2000]). Отличие нашей базы от названных проектов состоит прежде всего в том, что аудиоматериалы, в нее вошедшие, являются не чтением дикторами записанного заранее словника или текста, а не связанное с письменным текстом свободное (нередко спонтанное) произнесение слов, предложений, а иногда и целых текстов. Нам это отличие представляется весьма существенным. Кроме того, большое количество представленных в базе дикторов дает возможность проводить анализ вариативности произношения и границ этой вариативности.

2. Структура базы данных и программное обеспечение

Наиболее распространенной формой фиксации языкового материала в разнообразных компьютерных базах данных была и остается форма графическая. Однако современные компьютерные технологии позволяют хранить, воспроизводить и анализировать не только графическое изображение речи, но и ее звучание, а также видеозапись речевых актов. Создание комплексных компьютерных архивов исчезающих языков, включающих как графическое, так и аудиовизуальное представление языкового материала, является, на наш взгляд, оптимальным с точки зрения обеспечения сохранности и возможности последующего анализа этого материала.

Ядром базы является Озвученный словарь говоров северных селькупов. Объем словаря - 410 лексем. Для каждой лексемы в словаре зафиксированы ее троекратные произнесения от 32 информантов-дикторов разного пола и возраста, представляющих один их четырех ныне функционирующих северных селькупских говора (по восемь дикторов-носителей каждого из говоров). Для некоторых лексем даются также примеры употребления. Параллельно акустическому представлению приводится графическое представление материала в виде фонетической транскрипции, а словоформы, отличные от канонической словарной (именительный падеж единственного числа для существительных, простой инфинитив для глаголов), снабжены также и грамматическим комментарием.

Видеоряд в базе на данный момент ограничивается портретами всех дикторов, а также пейзажами мест распространения каждого из говоров. Для каждого диктора приводится также его лингвистическая биография.

Словарная база данных построена в виде гипертекстового документа с целью облегчения работы конечных пользователей, независимо от программного обеспечения их компьютеров. Первоначальный вариант базы

представлял собою объемную сводную таблицу с различными входами поиска (по говору, русскому, английскому или селькупскому слову) (см. [Казакевич и др. 2002]). Неотъемлемой частью базы были фонетические (звуковые) фрагменты, содержащие троекратное произнесение одного слова одним диктором. При увеличении объема информации стало очевидно, что двумерная сводная таблица не может быть адекватно воспринимаема пользователем (поскольку занимает несколько экранов по горизонтали и по вертикали), а большой объем аудиоинформации ведет к чрезмерному замедлению работы с базой через Интернет. В новой версии были сделаны изменения структуры базы, направленные, в первую очередь, на облегчение работы со словарем конечного пользователя. Во-первых, все таблицы выполнены с применением фреймовой архитектуры, что позволяет постоянно иметь перед глазами названия колонок и строк в больших таблицах. Во-вторых, большие сводные таблицы сформированы таким образом, чтобы пользователь мог быстро получить интересующий его элемент (слово, диктор, говор) при входе в базу по любому из возможных элементов. При этом, при необходимости, пользователь может, как и в предыдущей версии, получить всю сводную таблицу. В-третьих, в базу добавлены графические (текстовые) и фонетические (звуковые) фрагменты, состоящие из более чем одного слова (словосочетания и предложения). Поиск этих дополнительных фрагментов возможен из любого места базы, содержащего ключевое слово-вход (русское, английское или селькупское). Таким образом, за счет внесения изменений в структуру базы данных, удалось добиться увеличения скорости обработки запросов при резком увеличении объема самой базы. Это особенно заметно при доступе через медленные каналы связи, такие как коммутируемый доступ в сеть Интернет.

3. Сбор и первичная обработка материала для базы данных

Остановимся подробнее на сборе лексического материала для аудиоархива и построении «Озвученного словаря говоров северных селькупов». Во время трех полевых сезонов (2001-2003 гг.)[2] сбор лексики проводился по единому эталонному словнику, организованному по тезаурусному принципу и содержащему в основном базовую лексику селькупского языка. Объем словника – около 2000 единиц. При отборе лексики для словника за основу был взят селькупско-русский словарь среднетазовского говора тазовско-туруханского диалекта, подготовленный по материалам 1970-х годов и опубликованный в [Очерки 1993]. При составлении словника учитывались в первую очередь особенности селькупской, а не русской лексики: в состав словника включались названия предметов материальной культуры, представителей фауны и флоры среды обитания селькупов, реалий, играющих важную роль в их повседневной жизни, несмотря на то, что в лексической системе русского языка данные единицы нередко относятся к периферии. В состав словника включен 100-словный список Сводеша[3]. Языком-посредником при сборе материала являлся русский язык. Кроме того, для каждого элемента словника приводятся английские эквиваленты, что расширяет круг потенциальных пользователей собранными материалами.

При записи материалов для озвученного словаря селькупскому информанту-диктору в качестве стимула предъявлялось русское слово. В случае если информант не мог вспомнить селькупский эквивалент слова, ему давался второй стимул – селькупское слово среднетазовского говора. Особенно эффективным этот стимул оказывался при работе с пожилыми информантами, не слишком хорошо знающими русский язык: поскольку говоры взаимопонятны, информант узнавал слово, но комментировал неправильность произношения, а затем произносил то же слово по-своему, так, как считал правильным, то есть по нормам своего говора. Кроме того, только с селькупской подсказкой можно было хоть как-то получить произнесение от молодых информантов, слабо владеющих своим этническим языком.

Несмотря на использование современной звукозаписывающей техники, получение качественного представления «звучащего» словаря селькупского языка на CD и в Интернете связано с определенными трудностями.

1. Информанты не являются профессиональными дикторами. Поэтому они не могут произносить весь материал с одинаковой громкостью и темпом. Даже простое задание повторить слово по-селькупски три раза с паузами для некоторых информантов является сложной задачей. Если привести громкость к приемлемому значению возможно, так же как и вырезать лишние паузы, то разбить паузой слитно

[2] Экспедиция 2001 г. финансировалась РФФИ, грант № 01-06-88020; экспедиции 2002 и 2003 гг. финансировались РГНФ, гранты № 02-04-18019 и 03-04-18007.

[3] Единственным словом из 100-словного списка Сводеша, не вошедшим в словник, является слово *семя*, поскольку селькупский эквивалент слова отсутствует.

произнесенные слова невозможно из-за явления коартикуляции — качество последнего звука в слове изменяется под воздействием первого звука следующего слова.

2. Информанты, хорошо владеющие языком — это представители старшего поколения. К сожалению, дикция этих возрастных информантов не всегда удовлетворительна. Более молодое поколение хуже владеет языком и в словаре возникает много пропусков.

3. При записях больших словарей (а это многочасовая работа) из-за постоянного лимита времени неизбежны и ошибки исследователя, проводящего запись (пропуски отдельных слов и даже фрагментов словаря). Не всегда удается отслеживать уровень записи и корректную работу микрофона. Оптимальный вариант работы — два исследователя (один следит за качеством и техническими параметрами записи) с одним информантом — далеко не всегда осуществим на практике.

4. Поскольку записи проводятся отнюдь не в студийных условиях, работе мешает и «бытовой» шум (к примеру, работа холодильника, когда запись производится в доме информанта), и шум окружающей среды (когда запись производится на улице, на природе).

5. Качественная в лингвистическом плане запись возможна только при условии комфортного состояния информанта. Поэтому постоянные просьбы повторить, переговорить слово (из-за плохого произнесения или постороннего шума) не всегда желательны.

Работа по подготовке материалов для словарной базы («нарезка» звуковых файлов — каждый файл это троекратное повторение слова по-селькупски) затрудняется из-за поиска нужного слова в записи (иногда лучшего варианта слова), необходимой коррекции громкости (в некоторых случаях) и вырезания пауз между словами. Данный этап работы был наиболее трудоемким и потребовал большого количества времени.

Появившаяся возможность получать качественные записи в полевых условиях создала предпосылки для уточнения звукового состава исследуемого языка и качественного описания просодического строя. Однако следует признать, что здесь мы только в начале пути (см. доклад [Захаров, Казакевич 2004] в настоящем сборнике, а также [Казакевич, Захаров 2001; Захаров, Казакевич 2003].)

4. «Побочные продукты» проекта

Запись материалов для Озвученного словаря селькупских говоров, направленная, казалось бы, на сбор исключительно лексики, оказалась способом получения ценнейших данных о функционировании языка на всех уровнях. При этом, поскольку в работе со всеми информантами (а их было более 40 человек, примерно по 10 человек на каждый из четырех говоров) использовался один и тот же словник, данные эти легко сопоставимы как по возрастным группам в пределах одного и того же говора, так и между говорами. Таким образом, побочным продуктом проекта стало что-то вроде тестирования языковой компетенции носителей обследовавшихся говоров, что весьма важно для оценки реального положения языка и перспектив его дальнейшего функционирования.

Размещение мультимедийных материалов по малому языку в Интернете, помимо прочего, объективно работает на повышение престижа этого языка в глазах его носителей, прежде всего, молодежи, что немаловажно для сохранения внутрисемейной передачи языка от родителей к детям там, где эта передача еще существует, а также для повышения мотивации детей, изучающих язык в школе.

Наконец, созданная база данных может быть использована в качестве учебного материала

- для общей лингвистической подготовки студентов филологических факультетов университетов;
- для подготовки специалистов по селькупскому языку;
- для преподавания селькупского языка в школе.

В настоящее время разработана программа дальнейшего развития базы данных и создания серии учебных продуктов, непосредственно с этой базой связанных. На выходе мы предполагаем получить:

1. мультимедийную словарную базу данных (словарные статьи с примерами, звук, видеоряд) селькупского языка на CD, объем базы 1500 словарных статей;

2. учебное пособие по селькупской лексикологии как приложение к мультимедийной словарной базе;

3. комплект учебных видеофильмов, состоящий из серии фильмов «Фольклор северных селькупов» (6–10 фильмов по 30 минут; первый фильм серии знакомит с жанрами селькупского фольклора, историей записи, персонажами, образным строем, элементами традиционной картины мира как она предстает

сквозь призму фольклорных текстов, остальные фильмы представляют собой образцы селькупского фольклора в исполнении современных сказителей на селькупском языке с русскими титрами);

4. сборник селькупских фольклорных текстов, представленных в серии учебных фильмов, с переводом на русский язык и грамматическими комментариями, желательно также с иллюстрациями.

5. Мультимедийная база данных кетского языка как попытка остановить мгновение языковой реальности

Выбор кетского языка для очередной мультимедийной разработки не случаен. Сегодняшнее положение этого языка, последнего из некогда довольно многочисленной семьи енисейских языков, таково, что если не создать качественный аудио- и видеоархив архив фиксаций его функционирования в ближайшие несколько лет, останется мало что фиксировать. Кеткий язык представлен тремя говорами: северным, южным и центральным. Младшим носителям северного говора сегодня под 60, носители центрального говора тоже в основном старше 50. Несколько лучше ситуация с южным говором, однако и там естественная передача языка от родителей к детям прекратилась уже более двух десятилетий назад. Стоит ли говорить, что все владеющие кетским языком владеют также русским, то есть являются билингвами.

Мультимедийная база данных кетского языка – это попытка зафиксировать стремительно исчезающую языковую реальность, которая становится призрачной, и трансформируется в квази-компетенцию у младших носителей. Хотя полевая работа по сбору кетского лингвистического материала в настоящее время ведется исследователями Томского государственного педагогического университета, Института филологии СО РАН (Новосибирск) и немецкими лингвистами, аудиозапись звучащей речи до сих пор, насколько нам известно, носила вспомогательный характер, и основной формой фиксации материала была графическая. Кроме того, практически во всех проводившихся до настоящего времени полевых исследованиях кетского языка в качестве информантов использовались в основном компетентные носители старших поколений, а «искаженная» речь более молодых, не слишком компетентных носителей оставалась без внимания. При построении кетской базы одной из наших задач мы считаем фиксацию речи носителей языка разных поколений, в разной степени владеющих языком. Это даст возможность выявить динамику изменения языковой структуры в ситуации сужения сферы функционирования языка и сплошного билингвизма его носителей. Работу над организацией базы данных мы предполагаем вести параллельно с ее наполнением, причем наполняться база должна будет в основном за счет сбора новых лингвистических материалов.

Как и в только что завершенном селькупском проекте ядром базы станет озвученный словарь кетских говоров. Словарные материалы, загружаемые в базу, будут представлены в двух видах: звуковом и графическом (транскрипция и, возможно, впоследствии принятая графика). Кроме того, в базу войдут кетские тексты. Для каждого текста предполагается наличие трех представлений: графического (транскрипция, впоследствии, возможно, также принятая графика), аудиозаписи и видеозаписи процесса порождения текста. В дальнейшем предполагается снабдить тексты грамматической индексацией, однако в рамках данного проекта это будет сделано лишь выборочно для демонстрации возможностей дальнейшего развития системы. На начальном этапе предполагаемый объем словаря – 300–400 лексем. Для каждой лексемы будет обеспечен выход на примеры ее употребления в текстах. Озвученный словарь кетских говоров представляется нам основой для проведения инструментальных акустических исследований кетской фонетики (до настоящего времени подобных исследований, насколько нам известно, не проводилось). В частности, весьма перспективным представляется нам анализ природы кетских тонов в ударных слогах. Наличие в базе данных наряду со звуковыми материалами видеоматериалов представляется весьма полезным для составления адекватного описания артикуляторной базы носителей разных говоров. Кроме того, видеоряд облегчает исследователям восприятие звучащей речи и ее адекватное транскрибирование. Наличие видеоряда дает также дополнительную информацию о функционировании языка в рамках конкретного языкового коллектива.

Очевидно, что реализация проекта потребует экспедиционной работы по сбору соответствующего аудио- и видеоматериала. Одним из инструментов экспедиционной работы станет кетско-русский словник-тезаурус, составление которого ведется сейчас по материалам существующих кетских словарей, прежде всего «Сравнительного словаря енисейских языков» Г. К. Вернера [Werner 2002].

По окончании проекта будет расширена фактическая основа для возможных обобщений и рекомендаций относительно оптимальной организации мультимедийной базы данных исчезающего языка как хранилища лингвистической информации и инструмента ее (этой информации) анализа.

Литература

1. Захаров Л.М., Казакевич О.А. Инструментальный анализ фонетического строя исчезающего языка // Компьютерная лингвистика и интеллектуальные технологии. Труды Международного семинара Диалог'2002 (Протвино, 6-11 июня 2002 г.). Т. 1. Теоретические проблемы. М.: Наука, 2002. С. 183-187.
2. Казакевич О.А. Мультимедийная база данных исчезающего языка // Труды Международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Том 1. Аксаково, 2001. С. 108-110.
3. Казакевич О.А., Захаров Л.М. Экспериментальное исследование фразовой интонации при кодовом переключении // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог'2003 (Протвино, 11-16 июня 2003 г.). М.: Наука, 2003. С. 250-253.
4. Казакевич О.А., Самарина И.В, Трушков Д.Л. Озвученный словарь говоров исчезающего языка // Компьютерная лингвистика и интеллектуальные технологии. Труды Международного семинара Диалог'2002 (Протвино, 6-11 июня 2002 г.). Т. 2. Прикладные проблемы. М.: Наука, 2002. С. 245-249.
5. Люблинская М. Озвученный словарь как инструмент сохранения и исследования фонетики малых языков // Congressus nonus internationalis fenno-ugristarum. 7.-13.8.2000 Tartu. Pars 2. Tartu, 2000. С. 344-345.
6. Очерки 1993 - Кузнецова А.И., Казакевич О.А., Иоффе Л.Ю., Хелимский Е.А. Очерки по селькупскому языку. Тазовский диалект. Том 2. М., 1993.
7. Werner H. Vergleichendes Wörterbuch der Jenissej-Sprachen. Bd. 1-3. Wiesbaden: Harrassowitz Verlag, 2002.