Проект 10-06-00532-а

О. А. Казакевич
Мультимедийный размеченный корпус текстов на говорах западных эвенков

Научно-популярное изложение результатов проекта

Богатство России – это не только ее обширная территория, не только ее недра и промышленный комплекс, прежде всего, это люди, живущие в стране, люди с их многоликой культурой и их языками. Язык – самое надежное хранилище культурного наследия и самый надежный механизм передачи этого наследия от поколения к поколению. Мы смотрим на мир сквозь призму языка, усвоенного в детстве; знакомство с каждым новым языком дает нам возможность увидеть новые грани большого мира, нас окружающего, обогатиться опытом многих поколений людей, говоривших на этом языке. Однако необходимо отдавать себе отчет в том, что над языковым и культурным многообразием нашей страны, как впрочем и всей планеты, сегодня нависла серьезная угроза: все чаще мы становимся свидетелями исчезновения малых языков под натиском языков более мощных – государственных, региональных или глобальных. А ведь с исчезновением каждого языка человечество теряет часть сокровищ, накопленных предшествующими поколениями, и в высшей степени нерачительно сложа руки смотреть, как гибнет наше наследие. Сохранение языкового и культурного многообразия нашей страны должно стать одной из стратегических задач современной науки (как, впрочем, и современной политики государства). И в этом могут помочь современные технологии.

Созданный при поддержке РФФИ Мультимедийный размеченный корпус текстов на говорах западных эвенков мы рассматриваем как наш скромный вклад в сохранение эвенкийского языка, одного из языков Сибири, находящегося сегодня под угрозой исчезновения. Это построенное по современным принципам компьютерное хранилище эвенкийских текстов, в котором пользователям обеспечивается не только быстрое нахождение нужного текста, но и нахождение заданного слова, словосочетания, предложения, грамматического признака или сочетания грамматических признаков внутри каждого текста, группы текстов или всего корпуса.

В корпус вошло 52 текста из обширного мультимедийного эвенкийского архива Лаборатории. Тексты записаны в 14 поселках на территории Эвенкийского, Таймырского, Туруханского и Енисейского района Красноярского края и Верхнекетского района Томской области в 1998-2011 гг. По жанру это в основном истории жизни и охотничьи рассказы, фольклорных текстов немного: для первой версии корпуса были отобраны тексты, отражающие, прежде всего, спонтанную речь. Большинство текстов имеют графическое, звуковое и видеопредставление. Тексты разбиты на предложения. Синхронизация графического, звукового и видео-представлений осуществлялась в программе ELAN. Графическое представление каждого предложения состоит из нескольких слоев: это фонетическая транскрипция, отражающая особенности локальных вариантов языка, с разбивкой слов на морфемы, поморфемные грамматические индексы (глоссы), приписываемые каждой морфеме, текст в официально принятой графике и русский перевод. Корпус размещен на Московском сервере языковых архивов LangueDOC http://languedoc.philol.msu.ru, где установлена программная платформа LAT (Language Archive Technology), разработанная в Институте психолингвистики им Макса Планка в Ниемегене (Нидерланды) специально для решения задач архивации языковых материалов, прежде всего материалов исчезающих языков (пример международной кооперации).

Для того, чтобы каждый текст и каждый элемент текста в корпусе легко было найти, корпус должен был быть размечен. С одной стороны, производилась так называемая метаразметка корпуса (каждому тексту приписывается набор признаков, по которым его можно будет впоследствии находить – время и место записи, имя рассказчика, его возраст, имя собирателя, имя расшифровщиков звуковой записи, жанр текста и т.д.); с другой стороны, делается внутренняя разметка каждого текста, прежде всего морфологическая (каждой морфеме, входящей в состав слов корпуса, посредством специального индекса (глосса) приписывается ее значение (лексическое для корней, словообразовательное или словоизменительное для аффиксов; служебные слова также получают свои индексы. Морфологическая индексация была наиболее временеемкой частью нашей работы, к тому же требовавшей достаточно высокой квалификации (знания эвенкийской грамматики, причем не одного, а по возможности всех когда-либо описывавшихся говоров западных эвенков, поскольку между ними существуют не только фонетические и лексические, но и структурные различия). Однако анализ «сложных мест» текстов, неизбежный в процессе индексации, принес свои научные плоды: в северо-западных эвенкийских говорах была обнаружена ранее никем не описанная грамматическую категорию, был также выявлен ряд черт, характерных для отдельных говоров или групп говоров, но ранее исследователями не отмечавшихся. Это не только углубляет наши знания о конкретных говорах конкретного языка, но и расширяет фактическую базу развития лингвистической типологии и теории языковых контактов, а также исследований изменения структуры языка в ситуации языкового сдвига, а всякое новое знание о языке – это ведь в первую очередь знание о нас самих: становление человека неразрывно связано со становлением его языка.

Мы рассчитываем, что пользователями корпуса станут исследователи, представляющие разные направления гуманитарной науки: лингвисты, фольклористы, этнологи, историки, а также и преподаватели эвенкийского языка и все, кто интересуется языком и культурой эвенков и современной историей нашей страны.