О. А. Казакевич, Е. Л. Клячко
O. A. Kazakevich, E. L. Klyachko

СОЗДАНИЕ МУЛЬТИМЕДИЙНОГО АННОТИРОВАННОГО КОРПУСА ТЕКСТОВ КАК ИССЛЕДОВАТЕЛЬСКАЯ ПРОЦЕДУРА[1]

DEVELOPING A MULTIMEDIA MARKED TEXT CORPUS AS A RESEARCH PROCEDURE

Аннотация. В докладе представлены некоторые результаты недавно завершенного проекта, в ходе которого на базе мультимедийного эвенкийского архива лаборатории автоматизированных лексикографических систем НИВЦ МГУ при поддержке РФФИ был создан мультимедийный размеченный корпус текстов на говорах западных эвенков объемом около 35 тыс. словоупотреблений. Более подробно мы остановимся на морфологической разметке корпуса и некоторых ранее не описанных грамматических особенностях отдельных говоров, обнаруженных в процессе этой разметки.

Abstract. The paper presents some outcomes of the recently accomplished project, which resulted in a multimedia marked text corpus of West Evenki Dialects (about 35000 running words), created on the basis of the multimedia computer Evenki archive of Laboratory for Computational Lexicography, Research Computing Centre, Lomonosov Moscow State University, with financial support from Russian Foundation for Basic Researches. The morphological interlinearising of the texts and some specific grammar features of particular local dialects revealed in the course of the glossing will be in the focus.

1. Введение: общее описание корпуса

Мы представляем некоторые результаты недавно завершенного проекта, в ходе которого на базе мультимедийного

[1] Доклад подготовлен по результатам завершившегося в 2012 г. проекта «Мультимедийный размеченный корпус текстов на говорах западных эвенков», грант РФФИ 10-06-00532.

эвенкийского архива лаборатории автоматизированных лексикографических систем НИВЦ МГУ при поддержке РФФИ был создан мультимедийный размеченный корпус текстов на говорах западных эвенков объемом около 35 тыс. словоупотреблений. У всякого подобного проекта имеется две стороны: информационное обеспечение и программная реализация. Мы постарались воспользоваться существующими программными продуктами, сосредоточив основное внимание на информационном обеспечении. Поскольку мы строили размеченный корпус текстов на языке, сфера функционирования которого неуклонно сужается, а письменная традиция невелика, важной составляющей информационного обеспечения являлась фиксация текстового материала. Современные технические средства позволяют сравнительно легко делать это в полевых условиях в аудио- и видео-формате. Получение графического формата записи текстов и их смыслового представления в виде перевода на язык, используемый существенно шире, чем язык самих текстов (в нашем случае на русский язык), требует значительно больших усилий. К началу работы над проектом для многих западных эвенкийских говоров мы имели тексты, представленные не только в звуковом и видео форматах, но и в графическом формате в виде полевой расшифровки (фонетической транскрипции и близкого к пословному русского перевода), сделанной с помощью носителей соответствующего говора, и это создало нам «стартовый капитал». Однако в нашем материале имелись лакуны, которые мы постарались хотя бы отчасти заполнять. В ходе работы над проектом мы провели две экспедиции в Томскую область для сбора дополнительного материал по сымскому диалекту и на Таймыр для сбора материала по говорам таймырских эвенков. Таким образом нам удалось пополнить наш архив и сделать наш корпус более представительным в отношении локального многообразия говоров западных эвенков. Существенным для нас является наличие адекватного представления о языковой ситуации в местах записи текстов и знание языковой биографии каждого из

наших информантов. Помимо прочего, все это иногда помогало при работе с самими текстами.

То, что у нас получилось в результате трехлетней работы, мы рассматриваем как первую версию корпуса, которую в будущем предполагаем существенно пополнить. В настоящее время в корпус входит 52 текста с морфологической и дискурсивной разметкой (глоссами), что составляет примерно пятую часть всех эвенкийских текстов архива ЛАЛС. Тексты, вошедшие в корпус, представляют 14 локальных говоров западных эвенков. По жанру это в основном истории жизни и охотничьи рассказы, есть несколько диалогов и несколько фольклорных текстов: для первой версии корпуса было решено отобрать тексты, в которых представлена максимально спонтанная речь. Все это тексты устной речи, которые были записаны на территории Эвенкийского, Таймырского, Туруханского и Енисейского районов Красноярского края и Верхнекетского района Томской области в 1998–2011 гг. Большинство текстов корпуса имеют графическое, звуковое и визуальное представление. Лишь для нескольких текстов визуальное представление отсутствует. Каждый текст снабжен набором метаданных. Тексты разбиты на предложения. Синхронизация графического, звукового и визуального представлений осуществлялась в программе ELAN. Графическое представление каждого предложения состоит, по меньшей мере, из четырех слоев: это фонетическая (приближенная к фонематической, но отражающая особенности локальных вариантов языка) транскрипция с поморфемной разбивкой слов, поморфемные аннотации (глоссы: семантические, грамматические и дискурсивные), текст в официально принятой графике, и русский перевод.

Существенным элементом в подготовке текстов к загрузке на сервер является их сопроводительная разметка — создание метаописания каждого текста, которое позволяет впоследствии вести в корпусе поиск текстов по различным параметрам. Параметры метаразметки текстов можно разделить на четыре

группы: 1) данные о тексте, 2) данные о рассказчике в случае монолога или о собеседниках в случае диалога или полилога, 3) данные о тех, кто записал и обработал (расшифровал, выверил, проиндексировал и т.д.) текст, 4) данные о месте и времени записи текста. К характеристикам текста относятся его название, диалектная принадлежность, жанр, сюжет и мотив в случае, если мы имеем дело с фольклорным текстом. К данным о рассказчике относится его имя (фамилия, имя, отчество), возраст, место рождения, место постоянного проживания к моменту записи текста, краткая лингвистическая биография. В разделе о тех, кто работал с текстом, указывается, кто сделал аудио- и видеозапись текста, информант, помогавший в расшифровке текста, и лингвист, работавший с этим информантом, лингвист, проверивший полевую расшифровку, лингвист, снабдивший текст морфологической аннотацией и т.д.

Корпус размещен на Московском сервере языковых архивов LangueDOC (**http://languedoc.philol.msu.ru**), использующем программную платформу LAT (Language Archive Technology). В качестве платформы для редактирования и просмотра корпуса текстов на сервере LangueDoc используется набор инструментов TLA. Система поиска Trova позволяет проводить поиск по метаданным и по содержанию текстов.

Таким образом, в корпусе обеспечивается как поиск текстов по определенным параметрам метаданных (поселок, говор, наречие, информант, жанр и т.д.), так и внутритекстовой поиск на уровне звукового или графического представления. В звуковом представлении параметром поиска является время звучания, в графическом представлении возможен поиск в любом слое по отдельному параметру или набору параметров (морфема, слово, словосочетание, глосса, набор глосс, переводной эквивалент и т.д.), Возможен внутритекстовой поиск по любому подмножеству текстов корпуса, которое задается с помощью метапараметров.

Корпус рассчитан на широкий круг пользователей: ими могут стать исследователи, представляющие разные направления гуманитарной науки, прежде всего лингвисты – тунгусоведы,

типологи, компаративисты, социолингвисты, специалисты по малым языкам Сибири, а также школьные и вузовские преподаватели эвенкийского языка, для которых аннотированные тексты могут стать полезный дидактическим материалом.

2. Морфологическая разметка корпуса как исследовательская процедура

Важной составляющей работы с текстом, отчасти предваряющей морфологическую разметку, отчасти идущей параллельно с ней, является выверка и при необходимости корректировка полевой расшифровки (фонетической транскрипции и русского перевода) аудиозаписи эвенкийских текстов архива. Иногда и после неоднократного прослушивания в тексте остаются «темные места», прояснить которые удается в процессе поморфемного разбора и индексирования каждой выделенной морфемы.

Морфологическая индексация (глоссирование) текстов — это наиболее временеемкая часть подготовки материала корпуса, к тому же требующая достаточно высокой квалификации (знания эвенкийской грамматики, причем не одного, а (по крайней мере, в общих чертах) всех диалектов и говоров западных эвенков, поскольку между локальными вариантами существуют не только фонетические и лексические, но и структурные различия). Были проиндексированы как словоизменительные, так и деривационные морфемы. Эвенкийский язык имеет богатейший набор словообразовательных аффиксов, и при работе с текстами мы старались учесть максимальное их количество. При разработке системы поморфемной индексации мы исходили из Лейпцигских правил глоссирования, дополняя исходно составленный список общеупотребительных глосс обозначениями эвенкийской деривационной специфики (BUSH ‘кустарник’ (имеющий X в качестве плода, X – производящая основа) PRGRN ‘перегринатив (‘пойти за X’, X – производящая

основа), и т.д.)[2]. Для глоссирования текстов использовалась программа SIL Fieldworks Language Explorer (FLEx).

В процессе работы мы время от времени сталкивались с формами или явлениями, ранее в соответствующих говорах (или ни в одном из говоров) не замечавшимися и/или не описанными, с не зафиксированными в словарях лексическими единицами, с ранее не отмечавшимися значениями известных лексических единиц или морфологических показателей. Мы также обнаруживали различия говоров в предпочтении употребления тех или иных отчасти синонимичных грамматических или лексических форм. Таким образом, процесс глоссирования текстов оказался не технической, а исследовательской процедурой, и обнаружение новых лексических единиц, форм, значений форм и даже новой категории в отдельных говорах или группах говоров в общем-то хорошо описанного (по крайней мере, на морфологическом уровне) эвенкийского языка стало одним из важных результатов нашей работы.

Приведем только несколько примеров.

Якутское заимствование *usta-/usten-* определяемое в словарях[3] как ‘длина, расстояние’, в экондинских текстах, записанных нами в Эконде, регулярно используется в значении ‘весь’:

(1) Эконда, Х.Г.Н.

ustə-wa *d'əptilə-l-wo guruhi-l-wo d'uwū-d'a-ŋkī-tin d'əptilə-l-wo oro-r-d'i həti-l-d'i*

весь-ACC продукт-PL-ACC груз-PL-ACC перевозить-IPFV-PSTITER-3PL продукт-PL-ACC олень-PL-INSTR сани-PL-INSTR

---

[2] Общий список деривационных и словоизменительных морфем и глосс был составлен участниками проекта Е.Л. Рудницкой, Н.К. Митрофановой.

[3] *Василевич Г. М.* Эвенкийско-русский словарь М.: Гос. изд-во иностранных и национальных словарей, 1958; Сравнительный словарь тунгусо-маньчжурских языков / Под ред. В.И. Цинциус. Том 1, 2. Л.: Наука, 1975; 1977.

‘Все продукты, грузы возили, продукты оленями, санями’.

В текстах из Хантайского Озера и Игарки часто используется не фиксировавшееся ранее местоимение *uŋun* ‘этот’:

(2) Хантайское Озеро, У.Л.Ф.

uŋun-mə *dəvit-pa ha-Ø-ndə*

этот-ACC Дэвид-ACC знать-NFUT-2SG

‘Его, Дэвида, знаешь?’

Показатель *-d’a-*, ранее рассматривавшийся как показатель имени деятеля, в текстах на говорах северного наречия выступает скорее как показателем причастия:

(3) Чиринда, Е.В.Х.

gu= goro-lok dundə-duk əmə-fkī-l bi-sō-l aŋi-l tarə gramota-dū ***us'it-nā-d'i-l*** *bəjə-l anan lusa-l əmə-fkī-l bi-so-l oro-r-d'i*

SLIP далекий-ELAT земля-ABL приехать-PHAB-PL быть-PANT-PL это-PL тот грамота-DATLOC **учить-PRGRN-PSIMN-PL** человек-PL специально русский-PL прийти-PHAB-PL быть-PANT-PL олень-PL-INSTR

‘Из далекой земли приезжали какие-то эти, грамоте учить специально люди русские приезжали на оленях’.

Представляют интерес встречающиеся повсеместно случаи нестандартного функционирование показателя множественного числа ***-l-***, который нередко употребляется с названиями объектов, для которых контекст предполагает единичность: это могут быть названия поселков, учреждений, месяцев. Если в каждом конкретном случае мы можем предложить некоторое объяснение подобного употребления, то общего объяснения нам сформулировать пока не удалось:

(4) Чиринда, Е.В.Х.

jehe-duk ə-duk bəjə-l bi-s'o-l hurkōkō-r ***jekəndə-l-dūk***

Ессей-ABL что-ABL человек-PL быть-PANT-PL парень-PL **Эконда-PL?-ABL**

‘Из Ессея, из этого самого тоже люди были, парни, из Эконды’.

В сымском диалекте были обнаружены формы отрицательных причастий с нестандартным показателем *–wVnu* (5):

(5) Сым, Б.Г.П.
taduk bi ə-čə-w ***ičə-wonu*** *nuŋan-ma-n*
потом 1SG NEG-PST-1SG **видеть-PNEG2** 3SG-ACC-PS3SG
'Потом я не видел его'.

При проверке описанных ранее закономерностей диалектного варьирования списка эвенкийских глаголов, нестандартно образующих форму небудущего времени[4], были получены результаты, не во всем совпадающие с прежними описаниями: в текстах обнаруживается гораздо большая вариативность форм.

3. Заключение

Грамматическая разметка корпуса текстов на языке, для которого этот процесс не автоматизирован и, в силу отсутствия общепринятого стандарта, вряд ли может быть полностью автоматизирован, отнимает немало времени, но при этом повышает квалификацию лингвистов, ею занимающихся. причем не только углубляя их знания об устройстве конкретного языка, с материалами которого они работают, но и давая более широкий взгляд на устройство человеческого языка вообще.

Очевидно, что мультимедийный корпус, где представлена широкая панорама современных эвенкийских говоров, открывает возможность более пристально рассмотреть как фонетические, так и грамматические различия между отдельными говорами. Наличие в корпусе звукового ряда обеспечивает пользователям возможность верификации графической записи текстов[5], что, на

---

[4] *Василевич Г.М.* Очерки диалектов эвенкийского (тунгусского) языка. Л.: Гос.Уч.-пед. Изд-во Мин.прос. РСФСР, 1948; Константинова О.А. Эвенкийский язык. М. – Л., 1964.

[5] По сути, любая графическая запись звучащей речи – это ее интерпретация.

наш взгляд, весьма существенно, особенно если иметь в виду, что некоторые эвенкийские говоры, представленные в корпусе, имеют печальную перспектива в недалеком будущем сохраниться исключительно в аудиозаписях.