ТЕХНОЛОГИИ ОБРАБОТКИ ЯЗЫКОВЫХ ДАННЫХ В ДОКУМЕНТИРОВАНИИ МАЛЫХ ЯЗЫКОВ[1]

DIGITAL PROCESSING OF LINGUISTIC DATA FOR MINORITY LANGUAGES DOCUMENTATION

Кибрик А.Е. *(kibrik@philol.msu.ru),* ***Архипов А.В.*** *(arxipov@philol.msu.ru),*
Даниэль М.А. *(daniel@qub.com),* ***Кодзасов С.В.*** *(sankod@philol.msu.ru),*
Московский государственный университет
Майерс Том *(tommyers@dreamscape.com), N-Topus Software*
Нахимовский А.Д. *(adnakhimovsky@colgate.edu), Колгейтский Университет*

В докладе описан новый формат электронного лингвистического документирования малых языков. В рамках проекта по документации разрабатывается стандарт представления языкового материала и программная среда для создания и использования мультимедийных языковых ресурсов для малых языков.

1. Введение

В докладе описан новый формат электронного лингвистического документирования, представляющего собой важный этап стандартизации представления языковых данных малых языков. Стандарт, разрабатываемый в рамках проекта по документации нескольких малых языков России, позволит всем заинтересованным пользователям получать данные о малом языке в едином формате, в режиме свободного доступа через Интернет.

В сферу интересов проекта входят следующие языки: арчинский (лезгинский, нахско-дагестанский), нганасанский (самодийский, уральский; работой руководит сотрудник ИЯ РАН В.Ю. Гусев), хиналугский (нахско-дагестанский), алюторский (чукотско-камчатский).

2. Задачи документирования

В ходе документирования языков создаются три основных вида электронных ресурсов — корпуса текстов, репрезентативные фонетические базы данных (для арчинского, хиналугского, алюторского) и словари (для хиналугского и алюторского). Кроме того, для арчинского и алюторского языков, в связи с пожеланиями носителей этих языков, будет разработана практическая орфография (для арчинского языка эта работа уже почти завершена в сезоне 2006 года).

2.1. Репрезентативная фонетическая база данных

Под репрезентативной фонетической базой данных подразумевается реляционная база данных, содержащая аудио- и видеозаписи специальным образом отобранных слов и высказываний. При этом не преследуется цель создать аудиословарь (т. е. покрыть весь словарный запас языка или его лексически существенный фрагмент); имеется в виду, что отобранные слова и высказывания составляют репрезентативную выборку с точки зрения фонологической и просодической системы данного языка.

2.1.1. Материал фонетической базы данных

База содержит видео- и аудиозаписи произнесений отдельных слов (изолированных словоформ, в отдельных случаях — словосочетаний) в исполнении нескольких дикторов разного пола и возраста, а также их транскрипции. В фонетической базе для арчинского языка помимо фонемной и фонетической записи в системе МФА содержатся также: (i) запись в практической орфографии, принятой в публикациях МГУ по дагестанским языкам; (ii) запись в практической орфографии, принятой в арчинском словаре, создаваемого в Университете Суррея;

[1] Работа поддержана РФФИ, грант № 05-06-80351а, и программой Documenting Endangered Languages Program, National Science Foundation, грант # 0553546.

(iii) запись во вновь созданной кириллической орфографии. Кроме этого, включается вспомогательная информация о слове (перевод на русский и английский языки, частеречная помета и т. п.).

Словник для фонетической базы подбирается таким образом, чтобы покрывать все фонемы языка в наиболее типичных контекстах, как, например, начало / конец / середина слова, ударный / предударный / заударный слоги; согласные в интервокальной позиции, стечения согласных, зияния гласных и т. п. Для сравнения скажем, что словник для фонетически одного из наиболее богатых языков Кавказа — арчинского языка — включает около 400 словоформ (не обязательно принадлежащих к базовой лексике или наиболее частотной зоне словаря). Естественно, при возможности объём словника можно увеличивать, единственное существенное ограничение здесь диктуется количеством рабочего времени, необходимого для обработки записей — как для составления и выверки транскрипций, так и для технической обработки файлов.

В ещё большей мере это ограничение актуально для фразового компонента фонетической базы. При подготовке арчинского материала было решено отобрать для фонетической базы не более полусотни наиболее типичных реплик, иллюстрирующих основные типы фразовой просодии (различные типы сообщений, вопросов, побуждений; клишированные фразы и т. п.). Ставить в качестве задачи исчерпывающее исчисление просодических типов в настоящее время нецелесообразно, поскольку эта задача не решена и для таких языков, как русский[2]. Одна из главных причин состоит в том, что количество иллокутивно-модальных и логико-коммуникативных параметров, влияющих на фразовую просодию, исчисляется десятками, что даёт чрезвычайно много потенциальных комбинаций.

2.1.2. Возможности поиска и анализа информации

Если репрезентативная фонетическая база данных представляет ценность уже как иллюстрация репертуара фонетических средств языка, то современные технологии обработки данных дают возможность превратить её в по-настоящему мощный исследовательский инструмент.

Основными форматами записи, по которым осуществляется поиск слов в базе, являются фонемная и фонетическая транскрипции. Каждая фонема (аллофон) характеризуется цепочкой транскрипционных символов (в кодировке Unicode UTF-8), а также множеством различительных признаков, таких как *cons* ‘согласный’, *stop* ‘смычный’, *eject* ‘абруптивный’, *phar* ‘фарингализованный’ и т. п. Пользователь может вводить собственные признаки, приписывая их произвольному подмножеству фонем. Разработанный специально для фонетической БД язык запросов на основе регулярных выражений позволяет осуществлять поиск слов как по конкретной фонеме (аллофону), так и по различительным признакам. Более того, в запросе можно задать сочетание фонем (аллофонов), в том числе разрывное, контекст начала или конца слова или слога, позицию ударения, а также комбинировать все перечисленные условия отбора.

Приведём примеры некоторых запросов:

• t’	все слова с абруптивным t’
• {eject,dental,stop}	все слова с абруптивным зубным смычным (= t’)
• a{phar}	все слова, содержащие *a* перед фарингализованным
• (#|-){cons vowel}(#|-)	все слова со слогом типа CV
• #(m|r)	все слова, начинающиеся на *m* или на *r*
• #{cons}*{cons}#	все слова, начинающиеся и заканчивающиеся на согласный

3. Корпус глоссированных текстов

Языковые данные, которыми привыкла пользоваться современная лингвистическая типология и в целом теоретическая лингвистика, существуют главным образом в виде зафиксированных на письме слов, фраз и (реже) текстов. Преобладающая часть исходных данных, которые ложатся в основу грамматических описаний, обычно остается недоступной широкому кругу исследователей, публикуются в основном выводы и обобщения, сделанные на их основе. Такая ограниченность в материале уже давно не удовлетворяет запросам большинства лингвистов, хотя может быть успешно преодолена благодаря развитию компьютерных технологий и всё бóльшей их доступности. Разработанный на современном уровне корпус текстов для малого языка должен отвечать следующим требованиям:

• корпус может публиковаться на бумаге, но обязательно должен существовать в виде электронной базы данных, допускающей обновление и пополнение;

[2] В базе данных «Интонация русского диалога», созданной под руководством С. В. Кодзасова, собрано около тысячи вопросительных реплик, побуждений и сообщений, но и такой массив, покрывающий все существенные параметры варьирования, иллюстрирует лишь наиболее важные сочетания их значений.

- тексты должны быть представлены не только в транскрипции, уже представляющей результат определенной интерпретации исследователем языковых данных, но и в исходном виде, т. е. в (видео- и) аудиозаписях;
- тексты должны нести максимум лингвистической и иной разметки (аннотации), в том числе обязательно поморфемное глоссирование, а также, в зависимости от ресурсов, которыми располагает исследовательская группа, дополнительные слои морфонологической, просодической, синтаксической, частеречной, семантической разметки; текстовые, метатекстовые и энциклопедические комментарии;
- аннотация должна быть в высокой степени стандартизована для того, чтобы облегчить поиск сходных явлений в разных языках, описанных разными исследователями;
- пользователь должен иметь возможность выбрать для отображения только интересующие его компоненты (слои) информации;
- корпус должен поддерживать обработку самых разнообразных поисковых запросов пользователя, включая поиск по различным слоям разметки (например — и в первую очередь — по грамматическим глоссам и пр.);
- корпус должен являться открытым Интернет-ресурсом, чтобы любой заинтересованный пользователь мог легко к нему обратиться.

3.1. Источники материалов

Электронные корпуса исследуемых языков будут пополняться в общем случае из двух источников — новых текстов и текстов, записанных ранее. При этом тексты должны быть отглоссированы и, там где это возможно, сопровождаться не только аудио-, но и видеофиксацией. Для арчинского, нганасанского, хиналугского и алюторского языков существует значительный объем уже обработанных текстов, однако ни по одному из этих языков нет удовлетворительных аудиоматериалов; лишь по нганасанскому языку имеется большой объем магнитофонных записей 1980-х — 1990-х гг. и небольшое количество записей последнего времени; записей в формате WAV нет. В связи с этим было принято решение там, где это возможно, перезаписать имеющиеся обработанные тексты с новыми дикторами, обеспечив одновременную аудио- и видеофиксацию в соответствии с современными стандартами. Летом 2006 года была успешно проведена запись 29 арчинских текстов из публикации 1977 г. общим объёмом около 1700 предложений. Тексты читались дикторами во вновь созданной кириллической орфографии.

В рамках проекта записываются также новые тексты, с жанровой точки зрения восполняющие пробелы уже имеющегося корпуса. Поэтому в арчинском языке акцент был сделан на записи диалогов, а в нганасанском языке, корпус которого в основном состоит из фольклорных текстов, помимо диалогов будут записываться также слабо представленные на сегодняшний день в корпусе бытовые рассказы.

3.2. Компоненты разметки (аннотации) текстов и её стандартизация

Несмотря на то, что большинство специалистов убеждены в необходимости пользоваться при письменной фиксации текстов не только переводом на язык-посредник, но и строкой поморфемного глоссирования, единого унифицированного формата глоссирования и, в целом, представления текстов сейчас не существует. Различия касаются не только инвентаря грамматических глосс, но также и количества и состава необходимых слоёв репрезентации. Это связано как с объективными научно содержательными проблемами (неизоморфность грамматической структуры различных языков, различия в степени прозрачности морфонологических процессов и т. п.), так и с организационными (отсутствие единого координирующего центра или стандарта). Одной из главных задач нашего проекта является выработка такого стандарта, который бы позволил организовать возможно более полный и унифицированный корпус текстов на малых языках, к участию в котором мог бы подключиться широкий круг исследователей. В рамках проекта готовится к печати сборник «Малые языки и традиции: существование на грани. Вып. 2», включающий приведённые к единому формату тексты на нескольких языках: арчинском, удинском, водском, селькупском, энецком, пулар.

Конечно, такая стандартизация ограничена разнообразием естественного языка; тем не менее некоторые шаги в этом направлении вполне возможны. Одно из самых очевидных, развитых и полезных направлений — это стандартизация морфологического глоссирования. Здесь мы в основном следуем рекомендациям так наз. «Лейпцигских правил глоссирования» — как в том, что касается используемых символов-разделителей, так и в сокращениях, принятых для обозначения тех или иных грамматических категорий. Необходимо также иметь стандартные способы для отображения различных дискурсивных явлений (фальстарты, хезитация, паузы; реплики разных говорящих, переключение кода), разного рода комментариев, для совмещения разных стилей перевода (более буквального и более идиоматичного), и т. п.

Разрабатываемый стандарт является «полужёстким» и имеет модульную структуру, т. е. жёстко регламентирует набор слоёв репрезентации и средств оформления, из которых лингвист, пополняющий корпус, выбирает необходимые ему модули. Например, для нганасанских текстов, в отличие от арчинских, необходим уровень глубинного морфонологического представления. Специалист также может вносить свои дополнения (например, нестандартные глоссы для не включенных в стандарт категорий), описывая новые элементы в соответствии с принятой системой. Предусмотрена возможность выбора между краткими и полными вариантами грамматических глосс в зависимости от употребительности категории в конкретных языках.

3.3. Проблемы фиксации спонтаннной речи

Известно, что организация устного дискурса принципиально отличается от структуры обычного письменного текста. Любой носитель языка, а вслед за ним и исследователь при транскрипции и анализе записи в той или иной степени проводит нормализацию материала. Специалисты по грамматике также в большинстве своём привыкли обращаться к нормализованному языковому материалу, из которого устранены явления, ограниченные устной речью (фальстарты, хезитации и т. п.). Таким образом, конечной символической репрезентацией текстовых материалов на малых языках чаще всего являются именно нормализованные тексты. Это, однако, создает большие сложности, если мы ставим своей целью предоставить пользователю корпуса исходные аудиоматериалы, разумным образом соотнесенные с символьным представлением текста.

Действительно, соотношение между исходной звуковой дорожкой и нормализованным текстом далеко не тривиально, зачастую даже не линейно. Перестройки синтаксических конструкций и другие «нарушения» (с точки зрения письменного языка) линейного порядка синтаксических единиц, частые в устном дискурсе, делают невозможным прямую увязку отрезков аудиотрека с синтаксическими группами нормализованного текста. В связи с этим оказывается необходимым введение промежуточной символической репрезентации, отражающей специфику устного дискурса — дискурсивной транскрипции. С одной стороны, она жестко синхронизирована с аудиофайлом, с другой — представляет в символьном виде значительную часть элементов окончательной нормализованной записи и может быть легко соотнесена с последней.

Подробная дискурсивная транскрипция предполагает детальный фонетический и интонационный анализ и требует огромных временных и человеческих ресурсов, которыми полевые исследователи чаще всего не располагают. Поэтому в рамках проекта используется упрощенная дискурсивная транскрипция, отражающая основные характеристики дискурса — членение на дискурсивные единицы, фальстарты, хезитации, паузы,— упрощающие синхронизацию с аудиофайлами; такие важнейшие для устной речи характеристики, как движения тона (нефонологические) или точная длительность пауз и филлеров, в упрощенной транскрипции не отображаются. Её назначение — позволить пользователю корпуса следить за фонетической субстанцией текста по ее символьному представлению и соотносить фрагменты спонтанной речи с нормализованной записью. Те пользователи, которые захотят «углубить» дискурсивную транскрипцию, во многом смогут это сделать на основе имеющихся в файле звуковых, транскрипционных и переводных данных.

4. Электронный словарь

Формат полной документации предполагает также создание словаря с орфографическим входом слова, фонологической и фонетической записью, видео- и аудиотреком, переводом (переводами) и по возможности полной (в зависимости от степени изученности языка) семантической и грамматической информацией о слове. Для арчинского языка такая работа уже осуществляется (практически завершена) М. Чумакиной (университет Суррея). Электронный словарь существует и для нганасанского языка, хотя и неполный и в несколько упрощенном формате. В рамках проекта предполагается создать электронные словари хиналугского и алюторского языков.

5. Интегрированная среда компьютерной обработки языковых данных

Не секрет, что такие узкоспециальные некоммерческие проекты, как документирование малых языков, малопривлекательны для крупных разработчиков программного обеспечения. Поэтому лингвисты, в отличие, к примеру, от офисных работников, лишены возможности использовать созданные специально для их нужд готовые программные пакеты, автоматизирующие большинство встающих перед ними задач. Кроме того, обычные объёмы финансирования заставляют выбирать наименее экономически обременительные инструменты работы. Вместе с тем, приложения общего профиля, такие как текстовые редакторы, не удовлетворяют специфическим запросам, возникающим при лингвистической обработке языкового материала.

В качестве решения названных проблем в рамках проекта предлагается интегрированная среда для компьютерной обработки языковых данных при документировании языков. Эта среда многокомпонентна и строится на следующих принципах:

- Все компоненты среды — бесплатные (Freeware), большинство распространяются с открытым кодом (OpenSource)
- Используются открытые форматы данных (XML, OpenDocument (ODF), простой текст [plain text])
- Вся среда в целом поддерживает платформы Windows, Macintosh и Linux
- Совместная работа и доступ к данным: компоненты среды могут работать как все на одном компьютере (индивидуальное рабочее место), так и на разных компьютерах, связываясь через Интернет (совместная работа на удаленных машинах).

Часть компонентов среды разрабатывается непосредственно для лингвистических целей, в том числе участниками проекта; большинство доступны от крупных мировых разработчиков. Интеграция их в единую систему возможна благодаря использованию открытых форматов данных, в особенности формата XML, позволяющего автоматизировать различные преобразования структурированных документов с помощью языка описания трансформаций, XSLT. В ряде случаев для выполнения определённых функций можно использовать различные приложения в зависимости от предпочтений пользователя.

Среда обеспечит возможность автоматического экспорта отглоссированных текстов из программы Toolbox в формат XML, который служит посредником между всеми формами представления данных, используемых в приложениях среды — офисными приложениями OpenOffice, реляционными базами данных, приложениями для синхронизации и просмотра мультимедийных материалов и др.

В состав среды входят:

Приложение	*Назначение*
Field Linguist's Toolbox (ранее Shoebox)	Приложение для автоматизации глоссирования текстов и ведения словаря. Собственный формат (простой текст с пометами)
Приложение для связи Toolbox с остальными компонентами системы: BoxReader/Writer	Преобразование файлов из формата Toolbox в формат XHTML (документы, структурированные согласно правилам XML и готовые для отображения в веб-браузере) и обратно
Браузер, напр. Mozilla Firefox	Просмотр готовых файлов XHTML/XML, работа с корпусом текстов и другими приложениями
HTML-редактор, напр. NVu	Редактирование текстовых файлов различных форматов (XHTML, XML, XSL...)
XSLT-процессор, напр. встроенный в Windows XP	Преобразования XML-файлов
Офисный пакет OpenOffice.org (бесплатный пакет офисных приложений, функционально аналогичный Microsoft Office)	Редактирование и печать текстовых документов. Используется открытый формат OpenDocument (ODF), основанный на XML. Возможна интеграция с базами данных. Встроенный экспорт в PDF
Система управления базами данных MySQL	Хранение данных и обеспечение доступа к созданным ресурсам
Веб-сервер, напр. Apache Tomcat	Обработка динамических веб-страниц (JSP) при обращении к базам данных
Приложение Apache Ant	Запуск приложений на Java
Аннотатор (MannX, ELAN)	Приложение для разметки мультимедийных материалов и их синхронизации с текстом

Таблица 1. Состав интегрированной программной среды