А. В. Архипов

Документирование малых языков: научные и технические аспекты*

Документирование языков[1] – сравнительно молодая область лингвистики. Теснее всего она связана с *полевой лингвистикой*, то есть с изучением языков в среде их естественного обитания. Основная задача документирования – сбор и накопление первичных данных о малоизученных языках. Особенную актуальность она приобрела именно в последние годы, поскольку языковое разнообразие сейчас сокращается как никогда стремительно. По различным экспертным оценкам, в том числе по оценкам ЮНЕСКО, до конца XXI века исчезнут от 50 до 90% всех языков Земли. Естественно при этом, что наибольшая опасность угрожает малым языкам, на которых говорят в сравнительно небольших (менее 50 000 человек) языковых сообществах; такие языки обычно не имеют официального статуса и зачастую не обладают социальным престижем. Поэтому перед учеными встает задача сохранить как можно больше языковых данных для следующих поколений – как для будущих ученых, так и для самих носителей малых языков. В настоящее время многие фонды и академические организации прилагают значительные усилия к тому, чтобы сберечь наследие языков и культурных традиций, находящихся под угрозой исчезновения.

1. Задачи документирования языков

Документирование языка понимается как «долговременная многоцелевая фиксация языковых данных» [1].

Деятельность полевого лингвиста обычно бывает нацелена на *описание языка*, на непосредственное изучение языковых феноменов, в ходе которого могут решаться самые разнообразные конкретные исследовательские задачи. Специалисты по фонетике, например, могут изучать в некотором языке закономерности, связанные с ударением; специалисты по грамматике – правила построения и употребления глагольных форм; специалисты по семантике – систему цветообозначений и т.п. Описание языка опирается на абстрактные понятия и правила; оно является результатом лингвистического анализа, наблюдений, обобщений и догадок исследователя и поэтому всегда в значительной степени субъективно – трудно найти двух лингвистов, которые одинаково опишут один и тот же набор языковых явлений.

* Статья написана при поддержке РФФИ (грант №05-06-80351а) и NSF (грант №0553546).

[1] Англ. language documentation, или linguistic documentation. Используя термин «документирование», мы пытаемся подчеркнуть, что речь идет о деятельности, протяженной во времени, тогда как «документация» ассоциируется скорее с материалами, полученными в результате такой деятельности.

В отличие от описания языков цель документирования – сбор большого количества *исходных данных* («сырого» материала), которые позволят в дальнейшем изучать язык (то есть решать конкретные исследовательские задачи), даже если новые данные собрать будет уже нельзя. Основным видом речевой деятельности, фиксируемым при документировании, являются тексты различных жанров – рассказы, сказки, легенды, случаи из жизни, бытовые диалоги, песни и др. Большую ценность имеют записи спонтанной (неподготовленной) речи. Помимо этого, часто фиксируются также отдельные фразы, слова, грамматические формы слов (образцы склонения и спряжения).

Эти два вида деятельности взаимосвязаны, поскольку при решении любой исследовательской задачи лингвисты накапливают некоторое, подчас весьма значительное, количество исходных данных. Однако в подавляющем большинстве случаев эти данные остаются недоступными для других исследователей – лежат мертвым грузом в личных тетрадях или на кассетах, тогда как публикуются только результаты анализа. Задача же документирования – не только собрать материал, но и сделать его общедоступным, и притом доступным в удобной электронной форме, что значительно повышает эффективность работы каждого отдельного исследователя и позволяет избежать многократного повторения одной и той же работы разными людьми. Кроме того, необходимо обеспечить надежное долговременное хранение собранных данных и их периодический перевод на более современные носители и в более современные форматы.

Наконец, еще одна важнейшая задача, хотя она и может решаться независимо от предыдущих, – использование полученных материалов не только для изучения языка, но и для обучения языку, его поддержания или даже возрождения. Здесь особенно полезны и продуктивны тесное взаимодействие с носителями языка и их активное вовлечение в работу.

2. Из чего складывается языковая документация

Целью документирования, в идеале, является создание всеобъемлющего корпуса первичных данных, который полностью удовлетворит последующие поколения пользователей, какой бы аспект языка они ни захотели исследовать. (Естественно, на практике любой проект имеет свои ограничения.) Следовательно, должны учитываться любые разновидности языка – местные (диалекты, говоры), социальные, жанровые и пр.

Языковая документация адресована самому широкому кругу пользователей. Для местных жителей записанные тексты интересны тем, что привязаны к их сегодняшней жизни, быту, знакомым людям и судьбам. Кроме того, сам факт, что их языком интересуются ученые, что эти материалы будут доступны во всем мире, несомненно, поднимает престиж языка, может способствовать пробуждению интереса к собственному языку у молодых. Для будущих поколений – это сохранение культурного наследия. Конечно же, собранные и обработанные материалы имеют огромную ценность для всех специалистов, изучающих язык в самых разных аспектах. Это лингвисты, которые занимаются языками данной языковой семьи; лингвисты-типологи, изучающие общеязыковые закономерности путем сопоставления далеких и непохожих друг на друга языков; социолингвисты, этнографы, фольклористы. Далее назовем тех, кто занимается образованием, поддержкой языков, языковым планированием. Надо также иметь в виду неизвестных в данный момент потенциальных пользователей и такие задачи, которые на момент документирования еще никто не ста-

вил. Вместе с тем задачи проекта должны быть четко очерчены, чтобы документирование не свелось к накоплению массы бесполезных данных.

Основой документации являются звукозаписи – это такой «сырой» материал, который, благодаря доступности и записывающих устройств, и программ для обработки звука, легко собирать, обрабатывать и воспроизводить. Видеозаписи содержат больше информации – они позволяют легче ориентироваться в ситуации общения и в целом рассказывают намного больше о культуре изучаемого языкового сообщества, и в силу этого вызывают в нем гораздо более живой отклик. Однако создание качественной видеозаписи – весьма трудоемкий процесс, требующий специальной подготовки; кроме того, видеозаписи труднее хранить и обрабатывать, а для поиска и воспроизведения нужного фрагмента обычно требуется дополнительная временна́я разметка. Наконец, материалом для документирования могут служить изображения (фотографии, рисунки) различных предметов быта и окружающей среды, карты местности, схемы родственных отношений, рисунки клейма для животных, изображения героев фольклора и т.п. Здесь граница между изучением языка и культуры размывается, и мы можем незаметно для себя перейти к сбору уже этнографического материала.

Хотя в фокусе внимания при документировании находятся первичные данные, едва ли возможно ограничить проект только сбором данных – грамматические очерки, словари, грамматический анализ текстов являются полноправными компонентами языковой документации. Здесь нужно отметить, что даже текст, опубликованный в письменном виде, не является «сырым» материалом, поскольку любая транскрипция звучащей речи уже есть результат лингвистического анализа. Лингвист, транскрибирующий текст, принимает огромное количество решений (зачастую явно не оговоренных): относительно способа записи тех или иных звуков, интонации, отражения речевых сбоев и других явлений, свойственных устной речи.

3. Документирование языков и информационные технологии

Из сказанного выше вытекает ряд технологических требований к конечному продукту документирования – корпусу собранных материалов. Так, оптимальный корпус языковой документации должен иметь большой объем (к тому же, быть открытым для пополнения) и содержать как текстовые, так и мультимедийные данные. Очень важно обеспечить высокое качество исходных записей, которое позволит, например, проводить спектральный анализ звука или изучать артикуляцию по видеозаписи.

3.1. Хранение данных

Если хранение больших объемов текстов не представляет особых проблем, то при работе со звуком и, в особенности, с видеофайлами нужно располагать значительным свободным пространством на дисках. Учитывая современные стандарты качества, следует предусмотреть около 600 мегабайт на каждый час стереозаписи, или около 300 Мб на час монозаписи[2]. Что касается видео, то здесь объемы на порядок выше –

[2] При записи с параметрами 44 кГц/16 бит («качество компакт-диска»).

десятиминутный фильм, записанный на цифровую камеру стандарта MiniDV, займет около 2,2 Гб на жестком диске. Для хранения, скажем, 30 часов записи в таком формате понадобится диск объемом не менее 400 Гб, что на порядок больше, чем у среднестатистического офисного компьютера. Для больших архивов проблема нехватки места встает еще более остро, и чаще всего они (как, например, архив программы DoBeS) принимают на хранение файлы в сжатых форматах MPEG-2 или MPEG-4. Но даже такие объемы данных трудно передавать пользователям через Интернет; один из выходов состоит в том, чтобы иметь копии видеоматериалов разного качества для хранения и для демонстрации.

Корпус документации должен быть рассчитан на долгосрочное хранение данных. Это подразумевает, во-первых, регулярное создание резервных копий; во-вторых, усилия по поддержанию необходимого оборудования (магнитофонов, камер, дисководов, компьютеров) и носителей (кассет, дисков, бумажного архива) в рабочем состоянии; в-третьих – перевод данных из устаревших форм хранения в более современные: с магнитной ленты и бумажных карточек в компьютерные файлы, из доюникодовских шрифтов в Юникод и т.п.

3.2. Доступ к данным

Материалы должны быть доступны самому широкому кругу пользователей, что в наше время означает в первую очередь их выдачу в электронной форме через Интернет. С другой стороны, для разных групп пользователей необходимо предусмотреть разные формы представления – более простые в управлении и облегченные по содержанию для неспециалистов, для носителей языка; более сложные и подробные для лингвистов и других специалистов. Кроме того, для разных пользователей могут быть предусмотрены разные права доступа.

Что очень важно, вся система должна как можно меньше зависеть от конкретной информационной среды (оборудования, программного обеспечения, шрифтов и т.п.). Человек в любой точке мира должен видеть на экране одно и то же, независимо от марки компьютера, операционной системы и других технических параметров. В этом отношении перспективно использование бесплатного ПО с открытым кодом (ср. офисный пакет OpenOffice.org, браузер Mozilla Firefox), открытых файловых форматов, таких как XML; насущной необходимостью является использование шрифтов, поддерживающих кодировку Unicode.

3.3. Метаданные

Метаданные – это вспомогательные «данные о данных», облегчающие хранение и поиск материалов. Это важное понятие в отечественной практике до сих пор почти не применяется. Различают метаданные нескольких типов, в том числе классификационные (заглавие текста, участники разговора, автор записи, время и место записи, название или код языка…); описательные (касающиеся содержания записи); структурные (описывающие внутреннюю структуру документа – например, двуязычный словарь); технические (формат файла, размер файла, кодировка символов…); административные (дата последнего изменения, сведения об авторских правах, ограничения на доступ и распространение…).

Существуют два международных стандарта лингвистических метаданных: OLAC (Open Language Archives Community) и IMDI (ISLE MetaData Initiative).

4. Опыт Московского университета по документированию малых языков

Два года назад на филологическом факультете МГУ им. М. В. Ломоносова отмечался 45-летний юбилей ОТиПЛа – отделения теоретической и прикладной лингвистики[3]. В свое время отделение стало колыбелью московских лингвистических экспедиций. Заведующий кафедрой ТиПЛ, член-корреспондент РАН Александр Евгеньевич Кибрик занимается изучением малых языков России (СССР) с 1967 года. За это время под его руководством прошло более 40 экспедиций в языки Дагестана, Азербайджана, Грузии, Абхазии, Тувы, Камчатки, Памира, Поволжья. В настоящее время многие ученики А. Е. Кибрика как на ОТиПЛе, так и в других научных центрах проводят собственные экспедиции. А. Е. Кибрик также заведует недавно созданным отделом лингвокультурной экологии Института мировой культуры (ИМК) МГУ[4], чьей задачей является изучение и сохранение наследия малых языков.

4.1. Разработка стандартов для представления текстов

Начиная с 2005 года, группа сотрудников ОТиПЛа и ИМК МГУ применяет накопленный за долгие годы опыт работы с языком «в поле» в новых проектах по документированию[5]. Трехлетний проект РФФИ «Малые языки и народы: существование на грани», под руководством директора ИМК МГУ, академика Вячеслава Всеволодовича Иванова, уже подходит к завершению. В рамках проекта вырабатываются стандарты записи и комплексной репрезентации текстов на бесписьменных языках.

Текст на малоизученном языке – сложный лингвистический объект, и для его полноценного представления может использоваться много компонентов (слоев) информации, включая несколько различных транскрипций (более или менее подробных), несколько вариантов перевода (дословный, идиоматичный, литературный), комментарии различного рода (языковые, ситуационные, энциклопедические), а также различные аспекты грамматического анализа (морфологический, синтаксический). Во многих случаях, особенно для традиционных фольклорных, обрядовых текстов, важное значение также имеет визуально-антропологический компонент их исполнения. Исследователи разных языков, принадлежащие к различным школам, практикуют различные способы записи, используя к тому же различные технические средства. Для того, чтобы сделать возможным обмен данными между специалистами по разным языкам, облегчить автоматический поиск нужной информации в большом корпусе текстов, необходима стандартизация всех этих компонентов. Глобальная цель проекта – сделать возможным создание большого унифицированного корпуса текстов на малых языках Российской Федерации, доступный исследователям различных культур и языков.

Отделом лингвокультурной экологии ИМК МГУ издается также сборник «Малые языки и традиции: существование на грани» (в 2005 году вышел 1-й выпуск [2], 2-й готов к печати). Сборник посвящен проблемам документирования малых языков и включает словарные и текстовые материалы языков различных семей.

[3] http://www.philol.msu.ru/~otipl/new/main/index.php

[4] http://www.imk.msu.ru/Structure/Linguistics/linguistics.html

[5] http://www.philol.msu.ru/~languedoc/

4.2. Пять языков Евразии

Четырехлетний международный проект NSF «Пять языков Евразии» начался в мае 2006 года. Руководитель – Александр Нахимовский, профессор Колгейтского университета (США). Проект объединяет усилия лингвистов из Москвы и Петербурга. Американская сторона помимо финансирования полевых исследований обеспечивает техническую поддержку (разработка специального программного обеспечения). Первоначально планировалось документирование четырех языков России и одного языка в Азербайджане. Дополнительный грант NSF в 2007 году дал возможность включить в исследование еще один язык.

К настоящему времени состоялись экспедиции по документированию двух одноаульных языков северокавказской семьи: арчинского (с. Арчи, Дагестан) и хиналугского (с. Хиналуг, Азербайджан), и нганасанского языка самодийской группы уральской семьи (пос. Усть-Авам, п-ов Таймыр). Арчинский и хиналугский – бесписьменные языки, число говорящих на каждом из них – около 1200 человек. Сегодня эти языки находятся в относительной безопасности, пока еще не происходит сокращения населения и каждым из них дети овладевают с рождения. Однако их судьба все же вызывает опасение, поскольку традиционный уклад жизни в селах разрушается. Положение нганасанского языка намного хуже: среди этнических нганасанов в полной мере владеют языком только люди старше 50 лет.

В рамках проекта создаются ресурсы нескольких видов: электронные корпуса текстов, фонетические базы данных, словари. Все ресурсы имеют электронный формат и будут опубликованы в Интернете. Тексты – рассказанные носителями языка истории, легенды, бытовые диалоги и т.п. – записываются на высококачественную аудио- и видеоаппаратуру. Затем следует трудоемкий процесс перевода и многоуровневого фонетического и грамматического анализа. Когда текст, наконец, готов, пользователь может одновременно слышать звук, видеть рассказчика на экране и читать не только его слова, но и всевозможные лингвистические комментарии. Кроме того, лингвисты смогут осуществлять поиск отдельных слов, морфем или грамматических конструкций сразу во всей коллекции текстов. Фонетические базы данных необходимы для детальных исследований фонетики языка, они включают примеры на все звуки языка в различных вариантах их произнесения, записанные от нескольких дикторов.

Кроме того, для арчинского и хиналугского языков были созданы проекты письменности – на основе аварской кириллицы для арчинского, на основе азербайджанской латиницы для хиналугского. При этом арчинская письменность была сразу же использована нашими коллегами из Англии (среди них и выпускница ОТиПЛа М. Чумакина) при создании трехъязычного арчинско-англо-русского словаря.

Вообще, с арчинским языком связано много интересных событий. Так, обычно вся деятельность по изучению малых языков происходит по инициативе самих ученых: от местных властей и местных жителей инициатива исходит крайне редко, хотя местные вузы и научные центры, как и местная администрация, обычно рады таким экспедициям и оказывают посильную поддержку. Часто случается, что когда экспедиция уже приехала, среди местных жителей находятся один-два энтузиаста, готовых работать с лингвистами день и ночь ради сохранения родного языка и родной культуры. Но в данном случае нас ждал сюрприз: в один прекрасный день в МГУ пришел арчинец Максуд Садиков (сейчас он возглавляет Институт теологии и религиоведения в Махачкале) и обратился с просьбой создать письменность для арчинского языка – что в конечном итоге и было сделано.

4.3. Проблемы

Самая острая проблема, с которой приходится сталкиваться, помимо финансовой, – нехватка временных и человеческих ресурсов. Языков, ждущих своего исследователя, много, а квалифицированных специалистов, к тому же способных отдать львиную долю своего времени такой работе, очень мало. Необходимо не только дорогостоящее оборудование – нужны программисты, инженеры по звуку и видео, нужны люди, которые смогут заниматься ведением архивов, и все это очень трудоемкая деятельность. Поэтому так важно, так жизненно необходимо обучать лингвистов (не только начинающих, но и вполне зрелых) хотя бы основам современных технологий обработки информации, чтобы они могли максимально обеспечивать себя сами.

5. Фонды, финансирующие документирование малых языков

В России финансовую поддержку лингвистических экспедиций и проектов по документированию осуществляют два государственных фонда: РГНФ – Российский гуманитарный научный фонд (www.rfh.ru) и РФФИ – Российский фонд фундаментальных исследований (www.rfbr.ru).

За рубежом существует несколько крупных программ по документированию языков. Обычно такие программы финансируют исследования по всему миру, принимая заявки от ученых из любых стран для осуществления проектов в любой части света.

Программа **DOBES, Германия-Нидерланды** (Dokumentation Bedrohter Sprachen)[6]. Программа существует с 2000 года и финансируется Фондом Фольксваген[7]. Научная и техническая база программы – Институт психолингвистики имени Макса Планка (Неймеген, Голландия)[8].

Программа **ELDP, Англия** (Endangered Languages Documentation Programme). «ELDP – составная часть проекта Ханса Розинга «Исчезающие языки»[9]. Она осуществляется под руководством международного комитета и реализуется Школой восточных и африканских исследований Лондонского Университета[10]. В рамках программы планируется предоставление исследовательских грантов на сумму в общей сложности 15 млн фунтов стерлингов за десятилетний период»[11].

Программа **DEL, США** (Documenting Endangered Languages)[12] – многолетняя совместная программа Национального научного фонда (National Science Foundation)[13] и Национального гуманитарного фонда (National Endowment for the Humanities)[14], которой недавно был

[6] http://www.mpi.nl/DOBES/dobesprogramme/

[7] http://www.volkswagenstiftung.de/index.php?id=3&L=1

[8] http://www.mpi.nl/

[9] HRELP (Hans Rausing Endangered Languages Project) http://www.hrelp.org/

[10] SOAS (School of Oriental and African Studies), http://www.soas.ac.uk/

[11] http://www.hrelp.org/grants/apply/information/russian/index.html

[12] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12816

[13] http://www.nsf.gov/index.jsp

[14] http://www.neh.gov/

придан постоянный статус. Именно в рамках этой программы проводится проект «Пять языков Евразии» (см. выше). Исследования возглавляет Национальный музей естественной истории Смитсоновского Института (National Museum of Natural History, NMNH)[15].

Литература

1. *Gippert J., Himmelmann N. P., Mosel U. (eds.).* Essentials of Language Documentation. // (Trends in Linguistics. Studies and Monographs; 178) – Berlin, New York: Mouton de Gruyter, 2006.

2. Малые языки и традиции: существование на грани. Вып. 1. Лингвистические проблемы сохранения и документации малых языков под ред. А. Е. Кибрика. М.: Новое издательство, 2005.

3. *Кибрик А. Е., Архипов А. В., Даниэль М. А., Кодзасов С. В., Майерс Т., Нахимовский А. Д.* Технологии обработки языковых данных в документировании малых языков // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.) – М.: Изд-во РГГУ, 2007, с. 231–235.

[15] http://www.mnh.si.edu/