А. Е. Кибрик,
А. В. Архипов

Изучение малых языков на ОТиПЛе МГУ: опыт и вызовы современности

1. Опыт ОТиПЛа по изучению малых языков

Цель документирования – сбор и долговременное хранение большого количества исходных языковых данных, которые позволят в дальнейшем изучать язык в разных его аспектах, даже если новых данных собрать будет уже нельзя. Эта задача особенно актуальна для малых языков, находящихся под угрозой исчезновения. Основу языковой документации составляют тексты различных жанров – рассказы, сказки, легенды, случаи из жизни, бытовые диалоги, песни и др., – записанные на высококачественную аппаратуру (аудио и видео) и снабженные многоуровневым лингвистическим анализом (фонетическим, грамматическим, семантическим) и комментариями.

В 2005 году на филологическом факультете МГУ имени М. В. Ломоносова отмечался 45-летний юбилей ОТиПЛа– отделения теоретической и прикладной лингвистики[1]. В свое время отделение стало колыбелью московских лингвистических экспедиций. Первая пробная экспедиция состоялась в 1967 году в лакский язык (Дагестан). С тех пор под руководством А. Е. Кибрика прошло более 40 экспедиций в языки Дагестана, Азербайджана, Грузии, Абхазии, Тувы, Камчатки, Памира, Поволжья. Результатами последующих экспедиций стали грамматические описания ряда языков России и СНГ, прежде малоизученных: «Фрагменты грамматики хиналугского языка» (1972), фундаментальная четырехтомная грамматика арчинского языка со словарем и текстами (1977), в последние годы – грамматики цахурского (1999), алюторского (2000) и багвалинского (2001) языков; а так-

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

же двухтомное «Сопоставительное изучение дагестанских языков» (1988, 1990). Сейчас многие лингвисты, прошедшие через ОТиПЛовскую школу, проводят собственные экспедиции. Способствовать изучению и сохранению наследия малых языков стало прямой задачей недавно созданного отдела лингвокультурной экологии Института мировой культуры (ИМК) МГУ[1].

2. 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[2]. В 2007 г. завершился трехлетний проект РФФИ «Малые языки и народы: существование на грани» под руководством директора ИМК МГУ, академика Вяч. Вс. Ива́нова, посвященный разработке стандартов записи и комплексной репрезентации текстов на бесписьменных языках.

Текст на малоизученном языке – сложный лингвистический объект, и для его полноценного представления может использоваться много компонентов (слоёв) 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). Исследователи разных языков, принадлежащие к различным школам, практикуют различные способы записи, используя к тому же разные 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, доступного 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 вклю-

[1] http://www.imk.msu.ru/Structure/Linguistics/linguistics.html
[2] http://www.philol.msu.ru/~languedoc/

чает словарные и текстовые материалы языков различных семей. Во втором выпуске сборника продолжается издание алеутско-русского словаря, составленного Вяч. Вс. Ивановым на основе русско-алеутского словаря о. Якова Нецветова. Текстовые материалы включают фрагмент легенды на африканском языке пулар-фульфульде и около 20 текстов разных жанров на малых языках России (уральские языки: водский, энецкий, селькупский и арчинский язык нахско-дагестанской семьи) и Азербайджана (удинский язык нахско-дагестанской семьи). Это издание примечательно тем, что собранные под одной обложкой тексты на разноструктурных языках впервые представлены в едином и унифицированном формате морфологического глоссирования (за исключением текста на пулар-фульфульде, представленного в дискурсивной транскрипции без морфологических глосс). Кроме того, предложен особый формат размещения различных компонентов текста на книжном развороте (текст в практической орфографии; комментированный перевод на русский язык; разбитый на морфемы текст в фонологической транскрипции), ориентированный на многопрофильную аудиторию – носителей языка, этнографов и культурологов, лингвистов.

3. 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 финансирует полевые исследования и обеспечивает техническую поддержку проекта. Первоначально планировалось документирование четырех языков в 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 северно-кавказской 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 арчинском – около 1200 человек, на хиналугском – около 3000. 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, по-

скольку 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.

4. К созданию электронного архива материалов по малым языкам

Актуальной задачей является создание централизованного электронного архива, где хранились бы материалы по малым языкам, собранные разными исследователями. Такой архив должен пополняться за счет не только вновь собранных материалов, но и оцифровки/реставрации старых. Как и для любого хранилища большого количества данных, особую важность для такого архива имеют расширенные поисковые возможности. Для обеспечения поиска необходима большая и трудоемкая работа по унификации формата материалов, а также по снабжению их метаданными – вспомогательной информацией о происхождении, содержании и формате данных.

В настоящее время проходит подготовка к созданию в МГУ электронного архива, который сможет принимать материалы не только от лингвистов МГУ, но в перспективе и от любых российских исследователей. Архив будет основан на технологиях, разработанных в рамках программы DoBeS[4], что позволит хранить не только текстовые, но также аудио- и видеоматериалы, обеспечить обращение к данным через Интернет, расширенный поиск по метаданным и по содержанию (в т. ч. для глоссированных текстов – с учетом слоев глоссирования), контролировать права доступа к данным по желанию их создателей.

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