Создание и лингвистическая разметка звуковой словарно-грамматической базы данных по ительменскому языку

A sound lexico-grammatical database of the Itelmen language: creation and linguistic annotation

Долозова О. Н. (dolozova@gmail.com)

Санкт-Петербургский государственный университет

В докладе представлено описание звуковой базы данных по ительменскому языку, созданной на основе архивных аудиозаписей и словарных материалов. База данных включает в себя элементы лингвистической разметки, позволяющей осуществлять поиск лексем и словоформ по алфавиту, частеречной принадлежности, некоторым грамматическим и фонетическим признакам.

1. Задачи создания звуковой словарно-грамматической базы данных по ительменскому языку

Актуальность и значимость создания **электронных лингвистических ресурсов** по языкам, находящимся под угрозой исчезновения, неоднократно подчеркивалась во многих работах, посвященных этой проблематике[1]. Создание такого рода ресурсов позволяет говорить об эффективном достижении, как минимум, двух целей:

- Аккумуляция и систематизация материалов, существующих в разрозненном виде, что затрудняет доступ к ним для проведения исследований. Преобразование в электронный формат данных, существующих в рукописном виде, оцифровка аналоговых звуковых записей позволяет также сохранить эти материалы, не допустив их окончательной утраты.
- Создание дополнительной мотивации к изучению языка, благодаря представлению данных в современном компьютерном формате, обеспечивающем гибкость, доступность и привлекательность ресурса, а также за счет лингвистической классификации языковых данных по нескольким параметрам.

Наконец, поскольку речь идет о языке исчезающем[2], который не воспроизводится носителями, эти материалы представляют особую ценность. Будучи структурированы и организованы в систему, они дают возможность воссоздать утраченную на настоящий момент **звуковую материю** языка и проследить за произошедшими изменениями (звуковые записи демонстрируют аутентичное звучание речи на ительменском языке, которое современные единичные носители языка уже не способны воспроизвести). Несмотря на проблемное качество записи, эти материалы способны дать представление о звуковой форме языка и могут послужить своего рода «образцом для подражания» тем, кто будет изучать этот язык в дальнейшем.

Разработка описываемой базы данных по ительменскому языку велась в Институте филологических исследований СПбГУ в рамках проекта — Разработка национального фонда звучащей речи «Голоса народов России» при поддержке РГНФ, проект № 07-04-12163в. В ходе создания базы данных решались такие задачи как:

- перенос языковых данных на современные электронные носители, — систематизация, редактирование и структурирование языковых данных;

[1] См., например, LULCL 2005, Proceedings of the Lesser Used Languages and Computer Linguistics Conference, Bolzano, 27–28 October 2005, Ed. Izabella Ties — 336 p.

[2] По данным переписи 2002 г., ительменским языком владело 375 человек; впрочем, все они знали его значительно хуже, чем русский. По другим данным, уже в 1989 году языком владело менее 100 человек. (http://lingsib.iea.ras.ru/ru/languages/itelmen.shtml)

- разработка формата представления и способов подачи языкового материала;
- конвертация в формат базы данных имеющейся структурированной информации;
- разработка удобного и интуитивно понятного интерфейса, позволяющего осуществлять быстрый поиск необходимой информации по базе данных.

Материалом для описываемой звуковой словарно-грамматической базы данных послужили:

1) ительменские архивные полевые тетради[3], представляющие собой записи словарной программы;
2) соответствующие тетрадям звуковые записи, представленные на аналоговых носителях, которые в ходе выполнения проекта были оцифрованы и внесены в электронный каталог.

Языковые материалы этих двух типов были систематизированы, звуковые записи словаря соотнесены с расшифровками и транскрипциями. В качестве **программной среды** для реализации был выбран формат базы данных **Microsoft Access**, позволяющий реализовать поиск по различным параметрам в структуре реляционного типа. Этот формат также является достаточно гибким с точки зрения возможностей редактирования и пополнения базы новыми языковыми данными.

Говоря о выборе формата представления данных, следует заметить, что в мировой практике документирования и обработки языкового материала, сопровождаемого аудио или видеозаписью, сложились определенные стандарты. Имеется специализированный инструментарий, широко используемый и хорошо зарекомендовавший себя, в частности, такие программы как Toolbox[4] и ELAN[5]. Впрочем, как отмечается в статье J. Good[6], универсальных рекомендаций о том, какое программное обеспечение следует использовать при работе с тем или иным типом языковых данных, не существует. В нашем случае использование этого инструментария не позволило бы эффективно решить весь спектр поставленных задач, или создавало бы дополнительные сложности для потенциальных пользователей. Причины, по которым был выбран универсальный формат баз данных Microsoft Access, а не специализированные лингвистически ориентированные инструменты Toolbox и ELAN, перечислены ниже.

- В создаваемой базе данных исходной единицей хранения информации являются лексемы и соотносимые с ними производные словоформы, которые характеризуются по заданному набору признаков. Такой инструментарий как Toolbox и ELAN, хотя и позволяет создавать, в том числе, словарные материалы, однако в первую очередь ориентирован на обработку текстов как линейно разворачивающейся во времени звучащей цепи (ELAN) и последующий анализ составляющих элементов (ELAN, Toolbox). В нашем случае сегментация словоформ не осуществляется, все классификации оперируют словоформами как целостными единицами.
- Набор признаков, релевантных для осуществления поиска в создаваемой базе данных, и способ их организации предполагает особый тип наглядного представления данных, а также возможность соотнесения единицы определенного типа с несколькими единицами другого типа (например, наличие нескольких вариантов произнесения той или иной словоформы). Таким образом, более востребованным оказывается иерархический способ организации данных, а не последовательно линейный.
- Система поисковых запросов организована с учетом двух типов адресата: лингвисты-исследователи (на них ориентирована более специальная лингвистическая разметка) и представители языкового сообщества, те, кто с помощью базы данных сможет изучать язык (для них созданы интуитивно понятные поисковые запросы, содержащие общеизвестную лингвистическую терминологию — например, существительное, словоформа).
- Более широкое распространение формата Microsoft Access в нашей стране, более простой интерфейс, не требующий длительного специального изучения, как в случае с программами ELAN и Toolbox, которые имеют достаточно сложную структуру и ориентированы на более подготовленных пользователей компьютера.

2. Структура создаваемой базы данных

Итак, как уже было отмечено ранее, ключевыми **единицами хранения информации** в нашей базе данных являются **словоформы** и **лексемы.**

Лексемы репрезентированы исходными словарными формами лексических единиц[7]. В базе дан-

---

[3] Записи осуществлялись в ходе лингвистической полевой экспедиции в сентябре-октябре 1984 г. в поселке Ковран. Исследователи — д. филол. н., проф. А. П. Володин, д. филол. н. А. С. Асиновский

[4] http://www.sil.org/computing/toolbox/

[5] http://www.lat-mpi.eu/tools/elan/

[6] Jeff Good, Data and language documentation. To appear in Peter Austin and Julia Sallabank (eds.), Handbook of Endangered Languages. Cambridge: Cambridge University Press. http://www.acsu.buffalo.edu/~jcgood/publications.html

[7] В некоторых случаях для глагольных форм (особенно тех, которым в переводе соответствует целое словосочетание) в качестве исходной выступает не форма инфинитива, а личная форма (как правило, 1-е или 3-е лицо настоящего времени)

ных они соотнесены со всеми встретившимися в материалах соответствующими производными формами. Таким образом, в качестве единицы, по которой осуществляется поиск, может выступать: — *лексема*: в этом случае в ответ на поисковый запрос мы получаем список всех соотнесенных с ней производных словоформ, имеющихся в БД; — *словоформа*: в ответ на поисковый запрос мы получаем подробную информацию об интересующей нас словоформе и ссылку на соответствующую ей лексему.

Специфика экспедиционной программы, в соответствии с которой осуществлялась запись, заключалась в том, что конкретные грамматические формы фиксировались лишь выборочно. При подготовке базы данных в ряде случаев были добавлены исходные формы, которых не было в расшифровках и звуковых материалах. Эти формы не сопровождаются звучанием и транскрипцией, но являются полноценной единицей хранения информации, обеспечивающей более удобный систематизированный поиск.

Каждая **словоформа** соотнесена с **переводным эквивалентом** на русском языке и охарактеризована по нескольким **грамматическим и фонетическим признакам**.

В базе данных реализовано **5 основных типов поисковых запросов,** которые в обобщенном виде релевантны для любого языка и могут быть конкретизированы в зависимости от того, какие **грамматические и фонетические особенности** представляют интерес. Поисковые запросы в базе данных основаны на предварительно осуществленной разметке, представленной в электронных таблицах, на основе которых была произведена конвертация в формат базы данных.

2.1. Поиск всех лексем (словоформ)

Создан поисковый запрос, по которому может быть получен упорядоченный по алфавиту список всех лексем или всех словоформ, имеющихся в базе данных. При этом для каждой лексемы предусмотрены:

- возможность просмотра всех соответствующих ей словоформ, сопровождаемых переводом на русский язык и образцами звучания (рис. 1);
- возможность перехода к подробному описанию каждой словоформы, включающему в себя перевод, транскрипцию, образцы звучания, грамматические и фонетические признаки, лексический комментарий, возможные варианты орфографической записи, метаданные (номер записи, шифр звукового файла, код информанта) (рис. 2).

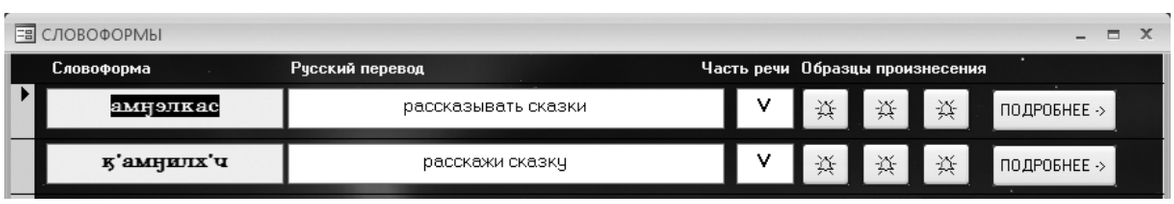

Рис. 1. Просмотр всех словоформ для заданной лексемы

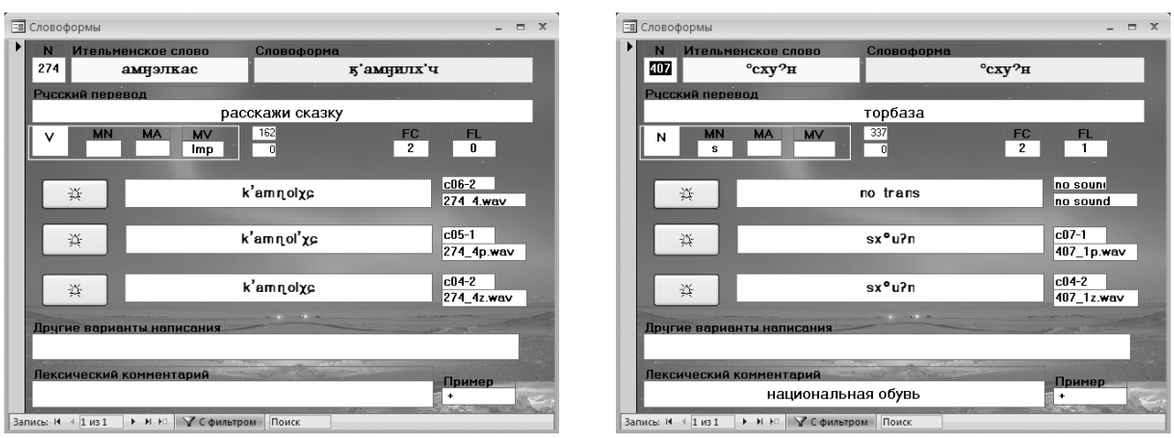

Рис. 2. Примеры описания словоформ по заданным в базе данных параметрам

2.2. Поиск всех лексем (словоформ), начинающихся на определенную букву ительменского алфавита

Результат поиска — список лексем (словоформ) на заданную букву, имеющихся в базе данных. Поиск реализован благодаря представлению списков лексем и словоформ в алфавитном порядке. В базе данных представлена не вся алфавитная последовательность, а только те символы, которым соответствует информация (рис. 3).

2.3. Поиск словоформ по признаку принадлежности к определенной части речи

Поисковый запрос в базе данных построен таким образом, что в виде отдельных списков могут быть получены все словоформы существительных, прилагательных, глаголов, наречий, местоимений.

2.4. Поиск определенных грамматических форм (типов словоформ)

Осуществляется за счет указания дополнительных грамматических признаков для различных частей речи. В базе данных могут быть получены списки: существительных единственного числа, существительных множественного числа, существительных в форме диминутива, существительных в формах косвенных падежей, существительных собирательных, глагольных форм императива, форм суперлатива для прилагательных (рис. 4).

Рис. 3. Просмотр всех лексем на заданную букву алфавита

Рис. 4. Окно поиска по частям речи и типам словоформ

2.5. Поиск по фонетическим характеристикам

В базе данных предусмотрена возможность поиска по двум видам фонетических характеристик: наличие/отсутствие лабиализации («огубленности»), а также количество согласных в консонантных сочетаниях. Таким образом, может быть получен список словоформ, характеризующихся наличием гармонии лабиализации (рис. 5).

В виде отдельного списка, упорядоченного по алфавиту, могут быть представлены словоформы, содержащие консонантные сочетания, включающие от 3 до 6 согласных (рис. 6).

Звуковая составляющая базы данных организована следующим образом: каждой словоформе соответствует от одного до трех звуковых файлов, репрезентирующих произнесение различных информантов — носителей ительменского языка.

Всего в БД представлены записи от 3 информантов — все они — жители одного населенного пункта, примерно одного возраста, среди них — 2 женщины и 1 мужчина, поэтому представленная вариативность носит, скорее, идиолектный характер. В исходных записях словоформа в исполнении одного диктора произносилась 2–3 раза — в базе данных этот принцип сохранен, поскольку некоторые повторы звучат более четко.

Каждому варианту звучания соответствует *реальная фонетическая транскрипция*, записанная в символах IPA (международного фонетического алфавита). В базу данных были внесены варианты

Рис. 5. Фрагмент списка словоформ, характеризующихся наличием лабиализации

Рис. 6. Фрагмент списка словоформ, содержащих консонантные сочетания, состоящие из 3 согласных и более

транскрипции, зафиксированные исследователями в ходе экспедиционной работы, с некоторыми уточнениями и унификацией символов.

Представленные транскрипции могут стать предметом отдельного более детального исследования, в результате которого они могут подвергнуться корректировке. К сожалению, возможности корректировки ограничены низким качеством записи. Для того чтобы внесение уточнений в транскрипцию стало возможным, необходимо осуществить дополнительную реставрацию записи.

Орфографическая запись, представленная в таблице, максимально приближена к современной орфографической норме ительменского языка[8]. Впрочем, в ряде случаев представлены также варианты орфографической фиксации, которые могут более точно отражать те или иные произносительные особенности. Эти варианты внесены в специально созданное в БД поле «Другие варианты написания», предназначенное для внесения комментариев. В поле «Лексический комментарий» представлены некоторые пояснения к значению, комментируются специфические реалии.

Интерфейс базы данных интуитивно понятен и реализован на русском языке. Разработана удобная система непротиворечивой **аннотации** для осуществления поиска в базе данных. В большинстве своем названия полей интуитивно понятны любому пользователю системы, часть из них приводится на русском языке: *ительменское слово, основное значение, словоформа, русский перевод, часть речи, образцы произнесения, другие варианты написания*, *лексический комментарий*. Для некоторых полей, содержащих более специальную грамматическую и фонетическую информацию, использовались условные обозначения в символах латиницы.

В создаваемой базе данных исследователям открыт доступ к редактированию, добавлению и удалению языковых данных, а также изменению форм и типов поисковых запросов. Общая схема аннотации и структура базы данных может стать основой для организации материала по другим языкам, а типы созданных поисковых запросов могут быть конкретизированы в соответствии со спецификой описываемого языка.

---

[8] Эта норма отражена, в частности, в издании Володин А. П., Халоймова К. Н. Словарь ительменско-русский и русско-ительменский: Пособие для уч-ся нач. шк. — Л., 1989. — 255 с.