Министерство образования и науки РФ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(ФГБОУ ВО «АмГУ»)

ЛИНГВИСТИЧЕСКИЕ БАЗЫ ДАННЫХ

сборник учебно-методических материалов
для направления подготовки 45.03.03 – Фундаментальная
и прикладная лингвистика

Благовещенск, 2017

Печатается по решению
редакционно-издательского совета
филологического факультета
Амурского государственного
университета

Составители: Андросова С. В., Морозова О. Н.

Лингвистические базы данных: сборник учебно-методических материалов для направления подготовки 45.03.03. – Благовещенск: Амурский гос. ун-т, 2017.

ТЕМА 1. БАЗЫ ДАННЫХ

1. Базы данных. Основные понятия.

Существует несколько определений понятия «база данных». Наиболее широкая трактовка этого понятия такова. База данных — это совокупность определенным образом упорядоченных сведений о некоторых объектах. Под объектами при этом понимаются сведения, факты, события, процессы реальной жизни. Объект может быть материальным (студент, летчик, пчела, товар, телевизор, город и т.д.) и нематериальным. К числу последних относятся различные события (поход в цирк, встреча друзей, покупка автомобиля и т.п.), факты (конец урока, наступление лета, окончание университета и т.д.), процессы (выплавка стали, изготовление автомобиля, перевод текста и т.п.), числа, имена (номер зачетной книжки, имя ребенка) и т.д. В реальном мире каждый объект обладает определенными свойствами. В сознании человека свойствам объекта приписываются некоторые атрибуты (вес, скорость, длина, цвет и т.п.). Атрибутам приписываются определенные значения: батон весит 400 г, машина имеет скорость 90 км/ч и т.д.

2. Организация данных об объекте.

В базе данных атрибуты представляются элементами данных или просто данными, а значениям атрибутов ставятся в соответствие значения элементов данных или просто значения данных (данные). Рассмотрим, например, таблицу, содержащую информацию, которой можно охарактеризовать любого студента университета **Таблица 1**.

Т а б л и ц а 1. **Пример организации данных об объекте «Студент»**

Номер зачетной книжки	Ф.И.О. студента	Пол	Год рождения	Факультет	Группа	Размер стипендии
482835	Иванов И.И.	м	1992	ФФ	091	2300
516297	Сидорова А.А.	ж	1993	ЭкФ	974	1500

Это одна из разновидностей баз данных. Объектом в ней является студент. Он может быть однозначно описан атрибутами «Номер зачетной книжки», «Ф.И.О. студента», «Пол» и т.д. Эти атрибуты для каждого конкретного студента имеют определенные значения. Так, для студента Иванова И.И. атрибут «Номер зачетной книжки» имеет значение «482835», атрибут «Пол» — «м» и т.д. В компьютерной памяти атрибутам «Номер зачетной книжки», «Ф. И. О. студента», «Пол» и т.д. соответствуют элементы данных или просто данные, а значениям атрибутов «482835», «Иванов И.И.», «м» и т.д. — значения элементов данных или значения данных.

Таким образом, данное — это некоторый показатель, который характеризует заданный объект и принимает для конкретного экземпляра объекта некоторое значение.

Группу данных, образующих в таблице одну строку, называют записью об объекте или просто записью (ее также называют кортежем или сегментом). Кортеж, содержащий ***n*** данных, называется n-мерным кортежем или просто группой из n данных. В таблице «СТУДЕНТ» представлен семимерный кортеж или группа данных, состоящая из 7 элементов. Чтобы пользователь мог обратиться к кортежу, его необходимо идентифицировать. Поэтому одно из данных, значения которого не повторяются в базе данных, т. е. являются уникальными для каждого экземпляра объекта, выбирается в качестве ***идентификатора объекта, или первичного ключа записи***. Например, в описанном выше примере это может быть данное «Номер зачетной книжки». По идентификатору из базы данных можно получить

все сведения о студенте.
Как видно из приведенной выше базы данных «СТУДЕНТ», она представляет собой набор кортежей или записей. Если таких кортежей n (5, 6, ..., 10), то соответствующая таблица будет иметь n столбцов. Набор значений одного столбца называется полем или доменом (в нашем примере используются поля «Пол», «Ф.И.О. студента» и т.д.). Если несколько записей имеют одно и то же множество данных с однотипной информацией, то говорят, что эти записи имеют один формат.

3. Типы данных
При этом различают следующие типы данных, используемых в базах данных:
1)текстовые данные (для хранения текста ограниченного размера);
2)числовые данные (для хранения чисел любого типа);
3)дата/время (для хранения календарных дат и текущего времени);
4)денежные данные (для хранения денежных сумм);
5)счетчики (для хранения автоматически наращиваемых чисел);
6)логические данные (содержат одно из значений «да» или «нет» — «true» или «false»);
7)поле «MEMO» (для хранения больших объемов текста. В этом поле хранится не сам текст, а указатель на месторасположение текста);

!! При низкой скорости и малом объеме памяти компьютеров 80-х — нач. 90-х гг. для ускорения обработки баз данных требовалось, чтобы записи имели постоянный размер (заданный при проектировании конкретной таблицы базы). В то же время возникла необходимость вносить в базу данных текстовые блоки заранее неопределенной длины. Резервирование под это больших объемов дискового пространства было нецелесообразно. Это привело бы к неоправданному разрастанию файлов баз данных. Поэтому была предложена альтернатива с полем «MEMO». В данном поле и находилась ссылка на другой файл, в котором находились все текстовые примечания.

8)поле объекта «OLE» (содержит указатели на месторасположение в структуре файла базы данных мультимедийной информации — рисунков, звуков, фильмов и т.д.);
!! Помимо данных об объекте вставляется ссылка на ту программу, которая умеет работать с этим объектом (т.е. вызывает приложение, запускающее данный объект).
Примеры: 1) вставка электронной таблицы/диаграммы/графика в текстовый документ: при двойном щелчке на этой таблице вызывается соответствующее приложение, напр. MS Excel. Если на другом компьютере нет аналогичной программы, напр. Open Office Calc, то представление объекта будет невозможным; 2) электронный словарь с функцией воспроизвести звуковой облик слова словарной статьи.
9)гиперссылки (содержат URL-адреса Web-объектов сети Интернет).
Примеры расположения этих данных представлены в **таблице 2**.

Таблица 2. Пример расположения данных

№ п/п	Тема и диктор	Пол диктора	Общая длительность (сек.)	Общая длительность пауз (сек.)	Количество фонемоупотреблений	Дата записи	Примечания
1	Family (D1)	m	300	28	691	05.12.1996	Slow tempo
2	Town (D2)	m	300	20	783	01.10.1998	Medium tempo
3	Youth	f	120	10	564	16.09.1999	Fast tempo

№ п/п	Тема и диктор	Пол диктора	Общая длительность (сек.)	Общая длительность пауз (сек.)	Количество фонемоупотреблений	Дата записи	Примечания
	(D3)						
4	Dreams (D4)	m	200	20	800	12.03.2002	Fast tempo

Множество записей одного формата называют файлом. Множество файлов образуют базу данных. Это еще одно рабочее определение базы данных. Допустим теперь, что объект «студент» будет описан данными, содержащимися в таблице 47 (в памяти ПК это будет отдельный файл). Причем шифры специальностей и размеры стипендий можно также представить в виде таблиц (файлов в компьютерной базе данных).

4. Способы организации баз данных — типы моделей

В приведенном выше наиболее широком определении понятия «база данных» отмечалось, что это совокупность определенным образом упорядоченных сведений об объекте. Способ взаимосвязи входящих в базу данных записей называют моделью представления данных.

Типы моделей: модели объяснительного типа или объясняющие модели и модели воспроизводящего типа. От первых не ожидается порождения языкового продукта; эта модель должна непротиворечиво объяснять действие языка в целом, если он моделируется, или каких-либо его частей моделируемого явления, если моделируются эти части (Марчук, 200...: 21). Вторые обретают ценность, когда воспроизводимый ими результат подобен тому, который был бы получен в результате деятельности человека (там же).

До последнего времени чаще всего использовались три способа взаимосвязи записей, входящих в базу данных:

1) иерархический;
2) сетевой;
3) реляционный.

Иерархическая модель представляет взаимосвязь данных в виде иерархического дерева, состоящего из узлов. На самом верхнем уровне иерархии имеется только один узел — корень. Каждый узел, кроме корня, связан с одним из узлов на более высоком уровне, называемым исходным узлом для данного узла. Ни один элемент иерархической модели не имеет более одного исходного. Каждый элемент может быть связан с одним или несколькими элементами на более низком уровне. Они называются порожденными. Элементы, расположенные в конце ветви, т.е. не имеющие порожденных, называются листьями. Например, сведения о студентах нашего университета мог быть поданы в виде иерархической модели данных (см. рис. 1).

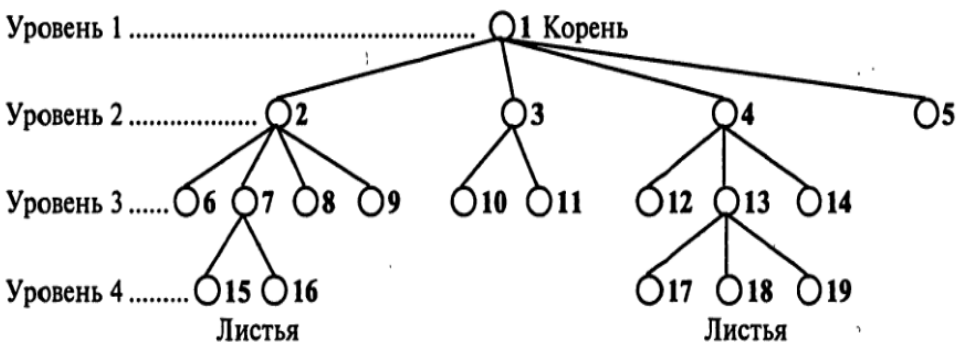


Рисунок 1. Иерархическая модель

В сетевой модели данных любое данное может быть связано с любым другим данным. При этом, как и в иерархическом представлении, можно использовать понятия корня, исходных и порожденных элементов, листьев. Порожденные элементы обычно располагаются ниже исходных. Здесь также можно говорить об уровнях (см. рис. 3). Однако в отличие от иерархической модели в сетевой модели у порожденного узла может быть несколько исходных узлов. Например, пусть необходимо построить базу данных для описания всех животных, которые находятся в зоопарке большого города. База данных должна содержать все сведения о животных, о тех, кто за ними ухаживает, и даже о директоре. Такую базу данных можно было бы представить в виде сетевой модели (см. рис. 3).

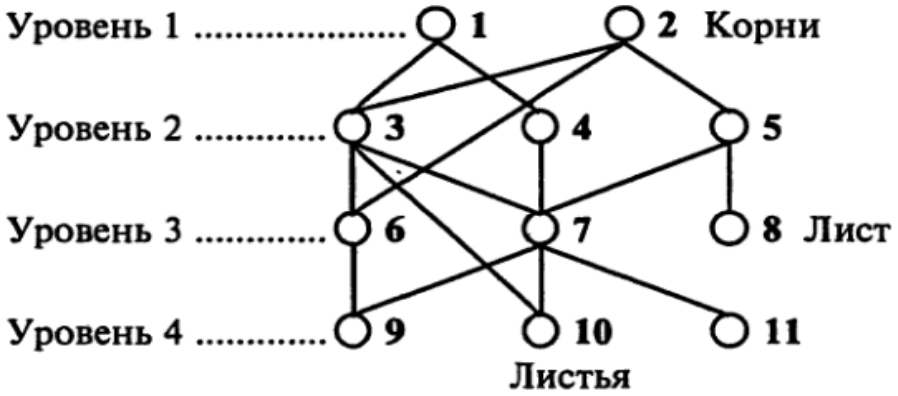


Рисунок 2. Сетевая модель

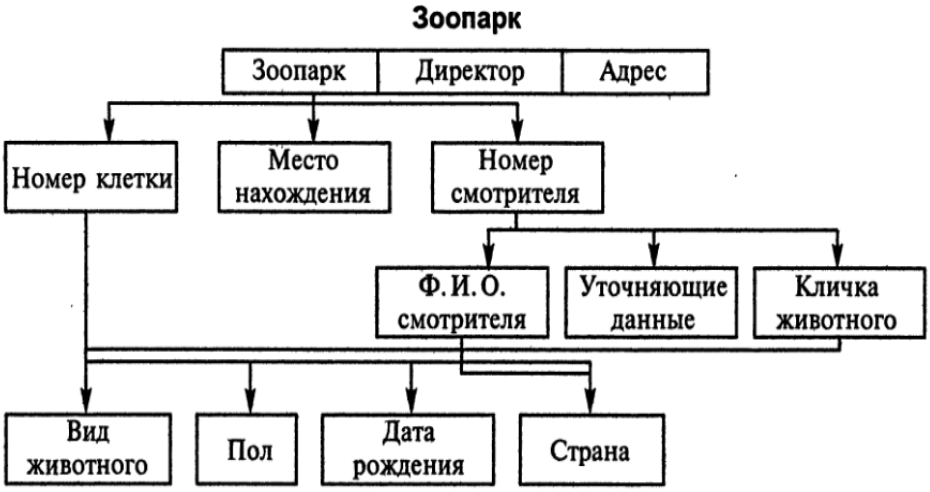


Рисунок 3. Схематичное представление сетевой модели

Добавлением в определённые позиции некоторых данных сетевое представление можно преобразовать в иерархическое (рис. 4).

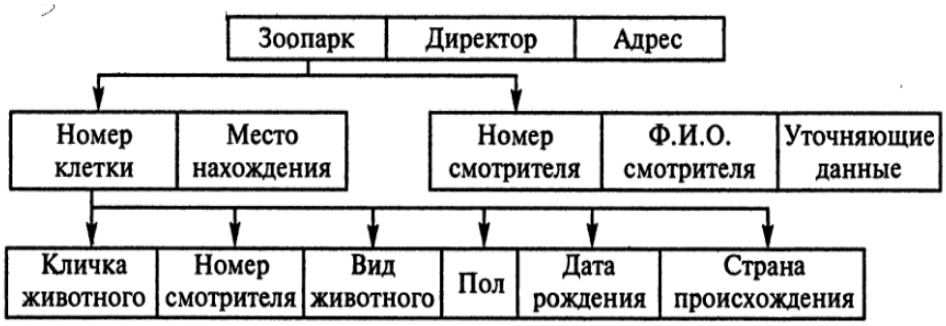


Рисунок 4. Преобразование сетевой модели в иерархическую

Данные в реляционной базе данных представляются в виде таблицы. Каждая таблица выражает отношения (relation) между включенными в нее данными. Как отмечалось выше, таблица-это набор кортежей (записей). Если кортежи являются n-мерными, т. е. таблица имеет n столбцов, то отношение называется отношением степени n. Столбец с номером j называется j-м доменом отношения.
Рассмотрим реляционную базу данных «Фонетисты и Фонологи». Как видно из таблицы 3, один домен может быть представлен двумя и более данными (год = год рождения + год смерти). Данное, значение которого идентифицирует наборы кортежей, как отмечалось выше, называется первичным ключом записи или идентификатором записи. В наборе кортежей «Фонетисты и Фонологи» таким ключом является данное «имя».

Таблица 3. **Реляционная база данных «Фонетисты и фонологи»**

Имя	Страна	Год	
		рождения	смерти
Щерба Л.В.	Россия	1880	1944
Трубецкой Н.С.	Россия	1890	1938
Jones D.	Великобритания	1881	1967
Соссюр Ф. де	Швейцария	1857	1913
Пасси П.Э.	Франция	1859	1940

Большая часть лингвистических информационных ресурсов может быть представлена в виде реляционных баз данных. Однако в последние годы к этим наиболее распространенным базам данных добавилось два новых типа:
1) объектно-ориентированные базы данных;
2) объектно–реляционные или гибридные базы данных.
Появление объектно-ориентированных баз данных связано с большинство современных средств разработки приложений, проектирования бизнес процессов и моделирования данных в той или иной мере являются объектно ориентированными. При этом понятие «объект», используемое в таких базах берется в полном объеме из объектно-ориентированного программирования. Такие объекты уже не делятся на составляющие (элементы данных), а как единое целое характеризуются некоторыми свойствами (атрибутами) и поведением (методами).
В гибридных базах данных объединение объектно-ориентированных и реляционных возможностей осуществляется путем использования специальных программных модулей, обрабатывающих определенные типы данных.
Сравнение основных технологий баз данных (см. табл. 4).

Таблица 4. **Сравнение баз данных**

Тип БД	Преимущества	Недостатки
Реляционная	Зрелая технология, высокая скорость, основана на стандартах	Не поддерживает мультимедиа, не ориентирована на сеть Интернет
Объектно-ориентированная	Обрабатывает мультимедиа, совместима с сетью Интернет	Не вполне зрелая технология, нет официально принятого стандарта разработки, недостаточная устойчивость работы
Гибридная	Возможность модернизации, обеспечиваемая производителем, совместимость со старыми приложениями, использующими реляционные БД	Нет официально принятого стандарта разработки, обслуживает только объекты, поддерживаемые производителем, разрозненная архитектура

ТЕМА 2. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ

Система управления базами данных (СУБД) — это совокупность программных средств, позволяющих осуществлять ведение баз данных (создание, обновление, удаление элементов данных и т.д.) и поиск в них информации. Практически все современные СУБД содержат средства для создания баз данных. Создавать базу данных можно тремя способами
1)с помощью специальных утилит, поставляемых отдельно от СУБД;
2)методом написания DDL-сценария (сценарий Data Definition Language);
3)использованием специальных средств, называемых CASE- средствами (средства Computer-Aided System Engineering). К их числу относят программные продукты ERwin, System Architect, Power Designer, ER/Studio и т.п.
С целью поиска информации и модернизации данных в СУБД используется специальный непроцедурный язык структурированных запросов SQL (Structured Query Language). Он является индустриальным стандартом. Его непроцедурность означает, что на этом языке можно указать, что нужно сделать с базой данных, но невозможно описать алгоритм этого процесса. Язык SQL включает группы операторов, ответственных за ведение баз данных и поиск в них нужной информации.
В настоящее время различают два основных типа СУБД:
1) настольные;
2) серверные.
Настольные СУБД, как правило, не содержат специальных приложений и сервисов, управляющих данными. Все взаимодействия с данными осуществляются здесь с помощью файловых сервисов операционной системы. Обработка извлекаемой из баз данных информации осуществляется в пользовательском интерфейсе. Сейчас такие СУБД стали сетевыми и многопользовательскими. Самые последние версии подобных СУБД приобрели визуальные средства проектирования форм, отчетов и приложений. Они дают возможность публиковать данные в сети Интернет, позволяют обращаться к данным серверных СУБД. Основной недостаток таких систем — снижение производительности и появление сбоев в работе при значительном увеличении объема хранимых данных и увеличении числа пользователей.
Существует более двадцати типов настольных СУБД. По степени сложности такие системы подразделяются на:
1) СУБД для обработки небольших объемов информации;
2)СУБД, ориентированные на конечного пользователя, который не умеет программировать или не желает тратить на это время;
3)сложные СУБД, ориентированные на разработку законченных
приложений.
Примерами простейших СУБД для обработки небольших объемов информации (первый тип СУБД) являются программы MS Schedule+ и MS Outlook из интегрированного пакета MS Office разных версий фирмы «Microsoft». Сюда же можно отнести и Lotus Organizer фирмы «Lotus». Такие программы позволяют хранить, изменять и управлять сведениями о контактах, встречах, событиях, собраниях, задачах. Они имитируют работу с обычной записной книжкой, представляющей собой адресный справочник контактных лиц, и настольным календарем, в котором отмечается расписание встреч и событий, список дел (задач и проектов).
Ко второму типу настольных СУБД относятся программы создания и обработки электронных таблиц (табличные процессоры), например: MS Excel, Lotus 1-2-3 фирмы «Lotus», Corel QuattroPro фирмы «Corel». Они не только создают базы данных в виде сложных таблиц и осуществляют в них поиск информации, но и могут проводить в таблицах различные вычисления.

Табличные процессоры могут содержать до нескольких десятков тысяч строк и 256 столбцов. В каждую ячейку можно поместить значение, число, текст, диаграмму, рисунок и скрытую формулу, выполняющую те или иные вычисления над явными данными. Программы создания и обработки электронных таблиц часто используются для построения таких баз данных, как перечни товаров на складах, картотеки клиентов, вклады в банках и т.д.
Пользователи, использующие настольные СУБД третьего типа, должны уметь программировать на специальных алгоритмических языках, встроенных в СУБД (например, VBA — Visual Basic for Applications, Paradox Application Language и т.д.). К числу таких СУБД относятся Paradox, dBASE, FoxBase, FoxPro, Clipper, MS Access и т.д.
Серверные СУБД используют архитектуру «клиент-сервер», суть которой заключается в централизованном хранении и обработке данных. Для этого используется так называемый сервер баз данных, выполненный как приложение или сервис операционной системы. Только этот сервер может манипулировать файлами базы данных, в которых хранится информация.
Серверные СУБД обладают следующими преимуществами по сравнению с настольными.
1. При выполнении запросов пользователей значительно снижается сетевой трафик (уменьшается время обработки запросов).
2. Имеется возможность хранения бизнес-правил (например, правил ограничений на значения данных).
3. Предоставлена возможность управления пользовательскими привилегиями и правами доступа к различным объектам базы данных (например, владелец некоторого объекта может предоставить другим пользователям право использовать его тем или иным способом).
4. Появляется возможность резервного копирования и архивации данных.
5. Предоставлена возможность параллельной обработки данных.
6. Обеспечена поддержка всех типов мультимедийных данных (текст, графические изображения, рисунки, видео- и аудиоинформация).

К числу наиболее известных серверных СУБД относятся программы Oracle, Informix, Ingres, Sybase, DB2, MS SQL Server.

Способы доступа к информации в базах данных

Существует несколько способов доступа к информации в базах данных. Они могут быть разделены на две категории:
1) способы доступа, использующие пециальный прикладной программный интерфейс API (Application Programming Interface);
2) способы, использующие универсальные механизмы доступа к данным. Программный интерфейс API представляет собой набор функций, вызываемых из клиентского приложения. Для настольных СУБД эти функции обеспечивают чтение/запись файлов базы данных. Для серверных СУБД интерфейс API инициирует передачу запросов серверу баз данных и получение от него результатов выполнения запросов (или кодов ошибок). Универсальный механизм доступа к данным реализуется в виде библиотек и дополнительных модулей, которые называют драйверами или провайдерами. Библиотеки обычно содержат некоторый стандартный набор функций. Дополнительные модули зависят от числа СУБД и реализуют непосредственное обращение к функциям клиентского API конкретной СУБД. Среди универсальных механизмов доступа к базам данных наиболее используемыми являются:

1) ODBC (Open Database Connectivity);
2) OLEDB (Object Linking and Embedding Database), использующий набор COM- интерфейсов (Component Object Model) и дающий возможность осуществить унифицированный доступ к данным из разных источников;
3) ADO (ActiveX Data Objects);

4) BDE (Borland Database Engine).
Первые три механизма доступа к базам данных по существу являются промышленными стандартами.
В практической работе с базой данных применяют три основных режима выдачи информации: 1) «on-line»; 2) «off-line»; 3) ИРИ (избирательное распространение информации).
В режиме «on-line» вся необходимая пользователю информация выдается мгновенно на экран ПК (она даже может быть записана в память компьютера, если это не запрещено условиями контракта по использованию базы данных).
В режиме «off-line» на экран выдаются только количественные данные о результатах поиска (например, сколько книг по грамматике английского языка найдено в базе данных, но сами названия книг не выдаются). Полная информация выдается пользователю за отдельную плату в распечатанном виде позже.
В режиме ИРИ запрос пользователя помещается в специальный каталог базы данных и автоматически обрабатывается при каждом обновлении информации, при этом информация передается пользователю, так же как и в режиме «off-line».

ТЕМА 3. ЛИНГВИСТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ

Основные понятия

Лингвистические информационные ресурсы — это одна из составляющих частей информационных ресурсов.

Под информационным ресурсом понимают некоторый интеллектуальный ресурс, результат коллективного творчества.

К пассивным формам информационных ресурсов относят книги, журналы, газеты, словари, энциклопедии, патенты, базы и банки данных и т. п. Активные формы включают алгоритмы, модели, программы, базы знаний.

Лингвистические информационные ресурсы — это множество определенным образом организованных речевых и языковых данных, находящихся на машинных носителях информации и используемых в различных сферах практической деятельности (образовании, промышленности, экономике, культуре, искусстве, издательстве и т.п.). Термин «лингвистические ресурсы» впервые был использован итальянским ученым Антонио Замполли в 1992 году. Он употребил его в докладе «Структура Европейского агентства языковых технологий», который сделал на конференции, посвященной проблемам создания основы для развития индустрии европейских языков. Выбор этого термина был связан с необходимостью выразить идею о том, что большие массивы лингвистических данных и описаний, используемые для создания и развития эффективных систем обработки текста и речи, играют такую же существенную, фундаментальную роль, как и железные дороги, автомобильные шоссе, электросети (электроэнергия), средства коммуникации для промышленности и экономического развития страны.

С современной точки зрения пассивные лингвистические информационные ресурсы включают следующее множество лингвистических данных (см. схему на рис. 5).

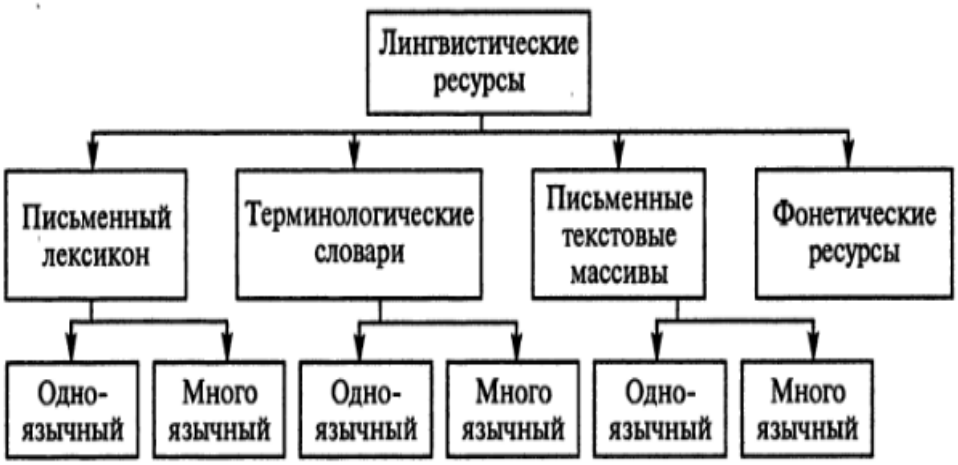


Рисунок 5. **Пассивные лингвистические ИР (Зубов, Зубова, 2004: 157)**

В самом общем виде лингвистические ресурсы — это своеобразные лингвистические базы данных, которые можно обновлять (добавлять новые данные, исключать или изменять старые) и в которых можно искать ту или иную информацию. Лингвистические ресурсы необходимы как пользователям ПК, так и различным компьютерным системам, связанным с обработкой текста и речи. В частности, они используются для распознавания речи и анонимных текстов, реферирования, аннотирования и перевода текстов, построения

диалоговых систем, автоматического анализа текста, синтеза речи и текста и т.д.

Проблемам создания лингвистических ресурсов ежегодно посвящается большое число международных и национальных научных конференций во всем мире. Создан ряд крупных организаций, объединяющих исследователей десятков стран, занимающихся разработкой лингвистических ресурсов. Наиболее известными из них являются LDC (Linguistic Data Consortium) (США), ELRA (European Language Resources Association) и TELRI (Transeuropean Language Resources Infrastucture) (Европа). Однако лингвистические традиции стран — участниц этих объединений, недостаточность финансирования, разобщенность участников приводят к тому, что лингвистические ресурсы, созданные отдельными коллективами одной страны, не всегда могут быть использованы в других странах. Все это ставит перед разработчиками лингвистических ресурсов ряд проблем, важнейшими из которых являются следующие:
1. Разработка единых стандартов создания ресурсов.
2. Разработка способов защиты лингвистических ресурсов от несанкционированного доступа.
3. Создание единых экспертных требований.
4. Планирование единой стратегии разработки лингвистических ресурсов.
5. Создание многофункциональных лингвистических ресурсов большого объема для использования в разных странах.

Письменный лексикон как простейшая составляющая лингвистических ресурсов

Как видно из схемы (рис. 5), письменный лексикон представлен лексическими ресурсами двух типов: 1) одноязычными и 2) многоязычными лексиконами (словарями).
В самом общем смысле словарь — это справочная книга, которая содержит слова (морфемы, словосочетания, идиомы и т.п.), расположенные в определенном порядке (различном в разных типах словарей). В нем может содержаться толкование значения описываемых единиц, а также различная информация о них. Как видно из этого определения, существуют различные типы словарей. Среди них выделяют:

а) по лексикографической форме основной единицы словаря:

1) для единицы, меньшей, чем слово:
- словари корней;
- словари морфем (приставок, суффиксов, окончаний);
- словари «-буквенных сочетаний;

2) для единицы, эквивалентной слову:
- словари словоформ;
- словари слов;

3) для единицы, большей, чем слово:
- словари словосочетаний;
- фразеологические словари;
- словари цитат;

б) по специфике отбираемой лексики:

1) словари, отражающие некоторые тематические и стилевые пласты лексики:
- терминологические словари;
- диалектные словари;
- словари просторечий;
- словари арго (табуированной лексики);
- словари языков писателей;

2) словари, содержащие различные разновидности слов:
- словари неологизмов;

- словари архаизмов;
- словари редких слов;
- словари сокращений;
- словари иностранных слов;
- словари имен собственных;

в) по способу описания основных единиц словаря выделяют словари, раскрывающие отдельные аспекты таких единиц и отношений между ними:

- этимологические словари;
- словообразовательные словари;
- грамматические словари;
- орфографические словари;
- орфоэпические словари (правила произношения);
- синонимические словари;
- антонимические словари;
- паронимические словари (содержат слова с частичным звуковым сходством при их семантическом различии, например: главный — заглавный, выплата — оплата и т. д.);
- частотные словари;
- словари рифм;
- словоуказатели;
- словари созвучий;

г) по расположению материала в словарях:

- идеографические словари, содержащие не слова, а рисунки со смыслом (например, глаз с капающей слезой обозначает «горе»);
- аналогические словари (в них слова располагаются не по алфавиту, а по смысловым ассоциациям — мебель: стол, стул, кровать, диван, кресло и т.п.);
- обратные словари (в них слова располагаются по алфавиту конечных букв);

д) по эпохе функционирования слова:

- исторические словари;

е) по назначению (адресату):

- словари трудностей какого-либо языка;
- словари ошибок;
- учебные словари.

Любой словарь может быть представлен в виде двухмерной реляционной базы данных, где по вертикали расположены единицы лексикона — части слов, слова, словосочетания — экземпляры объекта «СЛОВАРЬ», а по горизонтали — их некоторые характеристики (данные), упомянутые в приведенном выше определении понятия «словарь».

Простейшей лингвистической базой данных может служить частотно-алфавитный словарь словоформ какого-либо текста. Каждый экземпляр объекта (слово) описывается в нем одним данным — абсолютной частотой употребления слова в некотором тексте.

Более сложную организацию имеет словоуказатель. В нем кроме абсолютной частоты употребления словоформы в тексте указываются номера страниц и строк на странице, где встретилась данная словоформа. Словоформы располагаются, как правило, по алфавиту. Например, фрагмент словоуказателя к рассказу М. Шолохова «Судьба человека» выглядит так, как показано в таблице 5.

Таблица 5. **Пример словоуказателя**

№ п/п	Частота, *F*	Словоформа	Страница/строка	Страница/строка	Страница/строка	.....
...	.....	.....	.....	.....	.....	.....
56	1	*батареей*	51/29			
57	2	*батарей*	35/13	51/33		
58	3	*батарея*	34/31	35/9	52/38	
59	1	*батареями*	57/9			
60	2	*бегает*	41/4	55/11		
...	.....	.....	.....	.....	.....	.....

В других типах словарей в качестве характеристик XI, X2, X3, ... выступают толкования слов, их грамматические или стилистические особенности и т.д.

Еще более сложным типом словарей являются конкордансы. В них каждая словоформа текста характеризуется не только численными показателями (частотой, номером страницы, номером строки и т.д.), но и некоторым контекстом, в котором она употреблена. Как правило, этот контекст состоит из трех предложений: предложения, в котором встретилась словоформа; предложения, стоящего перед основным предложением, и предложения, стоящего после него. Предполагается, что контекст такого объема является достаточно полным и содержит внутри себя отрезок, законченный в смысловом отношении. Но возможны контексты и на уровне отдельных слов и словосочетаний. Конкордансы — ценный материал для научных исследований. Они помогают различать значения слов, широко используются в качестве примеров в других словарях (толковых, синонимических, антонимических, словарях писателей и поэтов и т.д.).

На основе конкордансов составляются контекстологические словари. Их теоретической основой является теория детерминант. В соответствии с ней каждое значение (перевод) многозначного слова определяется в контексте другими словами (детерминантами), с которыми сочетается исходное слово. Детальные сведения об устройстве и функционировании таких словарей приведены в соответствующих работах.
Если говорить об одноязычных словарях как части лингвистических ресурсов, находящихся на машинных носителях информации, то наибольшее их число разрабатывается в настоящее время в рамках организации ELRA. Так, по ее URL-адресам http://www.icp.grenet.fr/ELRA/home.html и http://www.elda.fr можно заказать Европейский португальский лексикон (ELRA-L0033 LusoEX European Portuguese Lexicon), содержащий 61000 основ (лемм) и 1600 соответствующих формообразующих парадигм. Там же можно найти японский компьютерный словарь (ELRA-L0036), в который включено 260 000 японских слов, английский компьютерный словарь (ELRA-L0037), состоящий из 190000 английских слов, и ряд других словарей. Известны созданные на машинных носителях французские словари общеупотребительной и специализированной лексики.
В многоязычных словарях дается перевод значения слова исходного языка на один или несколько иностранных языков. В качестве лингвистических информационных ресурсов чаще всего используются двуязычные словари. Наиболее известными из них являются серии электронных словарей МУЛЬТИЛЕКС, LINGVO и КОНТЕКСТ. Все они построены на основе известных бумажных словарей и служат для быстрого поиска переводных эквивалентов слов, уточнения перевода, поиска синонимов и целого ряда других операций. По указанным выше

URL-адресам организация ELRA предлагает целую серию двуязычных словарей. Например, японско-английский словарь (ELRA-M0023), включающий 230000 японских слов с их переводами на английский язык, англо-японский словарь, содержащий 160000 английских слов и соответствующих японских переводных эквивалентов.

Энциклопедии — это словари, содержащие характеристики не слова как такового, а обозначенного им предмета, факта или явления. Термин «энциклопедия» впервые был употреблен в 1541 году и обозначал набор систематизированных отрывков, описывающих различные разделы человеческого знания и искусства. Описание некоторого предмета или понятия в энциклопедии осуществляется в рамках словарной статьи. Например, статья, описывающая такой предмет, как буксир, в однотомном Советском энциклопедическом словаре выглядит так:

«Буксир (от гол. boegseren — 'тянуть') — самоходное судно для вождения (буксировки) несамоходных судов, плотов и т.п. Подразделяются на буксировщики (для вождения судов на тросе), кантовщики (для швартовки судов к причалу), толкачи (для буксировки судов толканием), спасатели (для оказания помощи судам в открытом море и их буксировки в порт-убежище)».

Чаще всего в энциклопедиях объясняются понятия, выражаемые именем существительным или именем собственным. Научные понятия часто выражаются именными словосочетаниями. В настоящее время энциклопедии являются важным ресурсом для создания баз знаний различных интеллектуальных задач. Энциклопедии могут использоваться в различных компьютерных моделях понимания текста, когда контекст не дает возможности распознать неоднозначную ситуацию.
Существует достаточно большое число различных энциклопедий на машинных носителях информации. Наиболее известна среди них энциклопедия «Britannica». Она включает 82000 статей и 700 дополнительных материалов, опубликованных с 1768 года. Не менее известны французские энциклопедии «Tons les savoire du Monde», «Le monde sur CD-ROM», «Versailles» и др. На русском языке изданы «Большая Энциклопедия Кирилла и Мефодия» и «Иллюстрированный Энциклопедический Словарь». Основная особенность электронных энциклопедий заключается в том, что их статьи содержат большое число иллюстраций, видео и звук.

Тезаурус — принципиально иной тип словарей. В нем в явном виде указаны семантические связи между определенной частью его лексических единиц. Как правило, такие словари строятся для текстов достаточно узкой проблемной области: вычислительной техники, музыки, кораблестроения, сельского хозяйства и т.д. При этом в простейшем случае между лексическими единицами тезауруса могут указываться следующие типы семантических связей: 1) синонимические; 2) антонимические; 5) родовые отношения; 4) видовые отношения, 5) ассоциативные отношения; 6) вхождение в состав более крупной семантической единицы; 7) состав (из более простых семантических единиц) и т.п. Например, в «Тезаурусе по теоретической и прикладной лингвистике» слово предложение представлено так:

ПРЕДЛОЖЕНИЕ
Синонимы: предикативная синтагма, фраза
Родовое понятие: высказывание, коммуникативная единица
Видовое понятие:
- главное предложение, придаточное предложение, вводное предложение;

- простое предложение, сложное предложение;
- повествовательное предложение, восклицательное предложение, вопросительное предложение;
- полное предложение, неполное предложение;
- распространенное предложение, нераспространенное предложение.

Вхождение в состав более крупной единицы: сверхфразовое единство, текст.
Состав (из более простых семантических единиц):
- главные члены предложения, второстепенные члены предложения;
- синтаксическая структура, синтаксическая группа, синтаксическая конструкция;
- знаки препинания;
- пропозиция, модальность.

Главное слово тезауруса называют дескриптором (слово предложение), а все нижестоящие слова и словосочетания — ключевыми словами (или словосочетаниями).

Первый тезаурус (тезаурус Роже) как средство «для облегчения выражения мыслей и помощи при написании сочинений» был создан в 1852 году в Англии. Особенно широко тезаурусы использовались в 70-х годах XX века, когда с целью поиска фактических данных об оборудовании (станках, автомобилях и т. п.) или технической литературы активно разрабатывались информационно-поисковые системы. Особой популярностью в те годы пользовался «Тезаурус научно-технических терминов». Находят применение тезаурусы и в наши дни. Так, в Москве в 1996 году был создан тезаурус по маркетингу и общеязыковой тезаурус русского языка. Он содержит 1600 дескрипторов и около 7000 распределенных по этим дескрипторам русских слов.

Терминологические словари и банки данных

Терминологическим словарем (ТС) называется словарь, основной единицей которого является термин. Нет единого подхода к определению этого понятия. В качестве рабочего определения примем следующее: «Термин — это слово или подчинительное словосочетание,имеющее специальное значение, выражающее и формирующее профессиональное понятие и применяемое в процессе познания и освоения научных и профессионально-технических объектов и отношений между ними».
В число единиц терминологических словарей в настоящее время входят и номены (номенклатурные единицы, идентификаторы). К ним, как правило, относят географические названия (океанов, морей, озер, гор, долин, городов, сел, полей), названия станков, приборов, аппаратов, фирм, систем, лекарств, медицинских препаратов и т. п. В качестве номенов выступают отдельные слова (Кавказ), словосочетания (Северный Ледовитый океан), буквенные символы (IBM), сочетания слов и символов (Pentium MMX), цифры, символы и цифры (IBM 486) и т.п..
Существуют различные подходы к классификации терминологических словарей. Так, по тематике, охватываемой лексикой терминологического словаря, различают:
1) многоотраслевые словари, включающие термины, относящиеся к различным научно-техническим отраслям (подъязыкам): машиностроению, судостроению, автомобилестроению и т.п.;
2) отраслевые (тематические) словари, содержащие термины, относящиеся только к одной теме, например вычислительной технике, оптике, органической химии и т.п.;
3) узкоотраслевые словари, содержащие термины по отдельным разделам некоторой более широкой темы. Например, существуют узкоотраслевые

словари по принтерам (тема «Вычислительная техника») и т.п.
По цели и назначению различают:
1) ТС, ориентированные на специалистов;
2) учебные ТС;
3) классификаторы;
4) словари (сборники) рекомендуемых терминов;
5) терминологические стандарты и т.д.
В приведенной классификации два первых наименования не требуют особых объяснений.
Классификатор — это словарь терминов с указанием различных типов иерархических отношений между ними. Подобные словари по принципу своей организации схожи со словарями-тезаурусами. Например, в классификаторе «Республика Беларусь» можно выделить следующие термины, которые сами являются классификационными рубриками: «Область», «Район», «Город», «Район в городе», «Улица» и т. п. При компьютерном использовании классификаторов их единицы получают определенные коды (схема 29).
Терминологические стандарты содержат термины вместе с их вариантами и рекомендациями о возможности употребления в тех или иных видах текстов.

По числу используемых естественных языков выделяют следующие терминологические словари:
1) одноязычные;
2) двуязычные;
3) многоязычные.

По способу упорядочения терминов различают следующие терминологические словари:
1) алфавитные;
2) неалфавитные (тематические);
3) частотные.

В неалфавитных (тематических) терминологических словарях термины расположены гнездами, т.е. сначала указывается заглавное слово, а затем перечень относящихся к нему ключевых слов, расположенных по алфавиту (см. рис. 6).

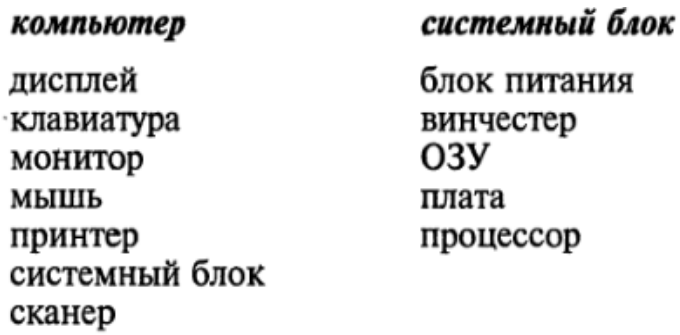


Рисунок 6. Схема терминологического словаря

Многообразие терминологических словарей объясняет тот факт, что объем информации, закладываемый в словарные статьи ТС, сильно отличается от словаря к словарю. В общем виде каждая статья включает термин и некоторые его характеристики (пометы): в частотных словарях — частоты употребления терминов, в классификаторах — коды понятий, в многоязычных словарях — адреса переводных эквивалентов и т.д. Есть и более сложная

организация словарной статьи, типичная для многоязычных терминологических словарей.
Основными источниками для составления словарей служат научные монографии и статьи по соответствующей проблемной области («вычислительная техника», «сельское хозяйство» и т.п.), соответствующие учебники и учебные пособия для вузов, рефераты и аннотации на научные издания, описания изобретений, патенты на изобретения, а также термины, зафиксированные в различных энциклопедиях. Упомянутая выше международная организация ELRA занимается созданием и терминологических словарей. Так, в ее рамках разработан технический терминологический словарь (ELRA-L0041), содержащий 80000 английских терминов и 120000 их японских переводных эквивалентов.

Терминологический банк данных — это некоторое множество терминологических словарей, относящихся к различным темам, определенным образом характеризующим какую-либо науку, производство или вид деятельности (медицину, образование, связь, транспорт и т.п.). Такие банки данных используются сегодня в основном для следующих видов работ:
1) создания новых терминологических словарей (в лексикографических целях);
2) перевода текстов с одного языка на другой (как человеком,так и с помощью компьютера);
3) создания учебников, учебных пособий, написания диссертаций и т. п. (в исследовательских целях).

Письменные текстовые массивы
Письменный текстовый массив или корпус текстов является в настоящее время основным понятием нового направления в лингвистике, называемого корпусной лингвистикой (corpus linguistics). Суть этого направления сводится к тому, что достоверные данные о фонетической, морфологической, синтаксической и семантической структуре языка и речи могут быть получены только из достаточно большого массива текстов. Впервые эти идеи были высказаны российским ученым профессором Р. Г. Пиотровским в докладе «Статистическое исследование лексики и грамматики текста с помощью электронно-вычислительной машины», прочитанном в Московском государственном педагогическом институте ино-странных языков в 1965 году. Дальнейшее развитие этой идеи, ее теоретическое обоснование и успешное практическое применение показаны в целом ряде крупных работ Р.Г.Пиотровского и его учеников .
В сегодняшнем понимании корпус текста — это совокупность текстов, являющаяся достаточной для обеспечения надежных научных выводов о некотором языке, диалекте или ином другом подмножестве языка. Такие письменные совокупности текстов могут быть использованы для решения большого числа лингвистических задач:
1) в лексикографии и лексикологии (для составления различных словарей, определения значений многозначных слов, выявления ассоциативных связей слов в тексте, выделения терминов и терминологических словосочетаний и т.п.);
2) в грамматике (для определения частоты употребления грамматических морфем в текстах различного типа, выявления наиболее употребляемых типов словосочетаний и предложений, определения значений синонимичных морфологических единиц, частоты употребления классов слов и т.д.);
3) в лингвистике текста (для дифференциации типов текста, создания конкордансов, выявления связи между предложениямив абзацах и между абзацами и т.д.);
4) при автоматическом переводе текстов (для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов терминологических и

фразеологических словосочетаний в параллельных текстах и т.д.);
5) в учебных целях (для выбора цитат, отдельных фрагментов произведений, примеров, используемых в процессе создания учебников и учебных пособий, и т.д.).

Существуют различные подходы к классификации корпусов текстов в зависимости от типа текстов, способов их организации, языка и т. д. С точки зрения их использования лингвистами наиболее значимы следующие виды корпусов текстов.
1. Исследовательские — создаются с целью изучения различных аспектов функционирования языка. Как правило, такие корпусы текстов содержат несколько десятков миллионов словоупотреблений.
2. Иллюстративные — служат для выделения из них лингвистических примеров, подтверждающих те или иные языковые (речевые, текстовые) факты, обнаруженные ранее иными лингвистическими приемами.
3. Статические — содержат тексты какого-то небольшого временного промежутка.
4. В динамические корпусы текстов включают письменные источники большого временного периода. Они предназначены для проведения различных диахронических исследований.
5. Корпусы параллельных текстов состоят из множества текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков. Это неоценимый материал для проведения сравнительно-сопоставительных исследований и для обучения переводу человека и компьютера.

Входящие в корпус тексты могут быть представлены в нем следующим образом:
1) в виде текстового архива в исходной форме, переносимой с письменных источников;
2) в виде банка данных, при этом тексты и их единицы предварительно определенным образом размечаются;
3) в виде базы данных информационно-поисковой системы, в этом случае тексты записываются в специальном формате, предназначенном для поиска текстов, их единиц и последующей статистической обработки.

Важной особенностью корпуса текстов является то, что это не просто множество случайным образом объединенных текстов того или иного языка. При его создании возникает целый ряд проблем.
Основными из них являются следующие:
1. Что должно являться основной единицей корпуса текстов?
2. Каков должен быть объем корпуса текстов (сколько единиц он должен содержать)?
3. Какие письменные текстовые источники должны быть представлены в корпусе текстов и в каком количестве?
4. Из какой исходной языковой области должны быть выбраны тексты, включаемые в состав корпуса?

Первые ответы на эти вопросы были даны в многочисленных исследованиях учеников профессора Р.Г.Пиотровского в 1965-1980 годах. Именно тогда были впервые использованы различные статистические приемы для оценки генеральной совокупности выборки, объема выборки, порции выборки (элементарной выборки) и т.п.
С сегодняшней точки зрения основной единицей корпуса текстов могут быть словоупотребления (обычно их называют словами), основы (корни, леммы) и предложения.

Объем создаваемого корпуса текстов в принятых единицах зависит от целей создания. Он может быть небольшим при изучении частоты употребления букв, буквосочетаний, звуков, звукосочетаний. Гораздо большим он должен быть при изучении лексики, морфологических

явлений. Неизмеримо большее количество единиц требуется при изучении синтаксических или стилистических особенностей текстов. Создаваемые сегодня корпусы текстов являются многофункциональными (т. е. могут быть использованы для изучения различных языковых явлений) и содержат миллионы слов, десятки тысяч лемм, тысячи предложений. Так, в апреле 2000 года по URL-адресам http://www.icp.grenet.fr/ELRA/home.html и http://www.elda.fr организация ELRA предлагала корпус французских текстов (ELRA-WOO20 PAROLE French Corpus), содержащий 20093099 слов. Она же предлагала корпус португальских текстов (ELRA-WOO24 PAROLE Portuguese Corpus), включающий 3000000 слов. По тому же URL-адресу можно найти португальский лексикон (ELRA-LOO33 LUSOLEX European Portuguese Lexicon), содержащий 61000 лемм и 1600 соответствующих морфологических парадигм.

Более трудной и менее определенной является задача конкретизации качественного состава корпуса текстов. При ее решении необходимо определить следующее.

1. Тексты каких функциональных жанров включать в корпус текстов (художественную прозу, драму, стихи, научные тексты, газеты, журналы, технические описания и т.д.)?
2. Тексты каких временных промежутков включать в корпус текстов (современные, 10-летней давности, 50-летней давности, древние и т.п.)?
3. Включать тексты только литературного языка или диалектические источники?

При решении этой задачи разработчики корпуса текстов обычно используют консультации специалистов по языкознанию и лингвостатистике либо метод анкет. Исходя из своего опыта исследований, специалисты определяют общий объем корпуса текстов, время издания текстов, число текстов и размер элементарной выборки, жанры отбираемых текстов и их количество, число элементарных выборок из каждого жанра.
Так, при создании одного из первых корпусов текстов, корпуса Брауна (The Brown Standard Corpus of American English), группа консультантов-ученых определила его объем в 1000000 словоупотреблений. Было решено, что он должен состоять из 500 текстов по 2000 словоупотреблений каждый. Тексты должны быть взяты из произведений американских авторов, изданных в США в 1961 году. При этом было рекомендовано отобрать 15 письменных жанров: 9 - информативная проза и 6 — художественная проза. Из каждого жанра было сделано от 6 до 80 элементарных выборок. Метод анкет в сочетании с опытом специалистов был использован при создании корпуса текстов The American Heritage Intermediate Corpus. Специалисты, ориентируясь на заданное время создания корпуса, определили его объем в 5000000 слов (словоупотреблений) и рекомендовали включить в него лексику из 22 разделов (жанров) детской и юношеской литературы на английском языке. Для конкретизации текстов в 221 школу США были разосланы анкеты с просьбой указать, какие тексты желательно включить в корпус. После изучения анкет был составлен список из 19000 названий книг. Из этого множества было отобрано 1045 текстов. На их основе было составлено 10000 элементарных выборок по 500 словоупотреблений каждая.

В упомянутый выше корпус португальских текстов, предлагаемый ELRA, вошли:
1) тексты газет (за 1996—1997 гг. 3 названия) — 65 %;
2) художественные книги (12 названий) — 20%;
3) периодические журналы (7 еженедельных изданий одного названия) — 5 %;
4) периодические журналы (8 названий) — 10%.

Последняя из четырех основных проблем, возникающих при создании корпуса текстов, связана с определением исходного объема языковой области, из которой должны быть

сделаны выборки. Она решается эмпирически, путем проведения предварительных экспериментов. Как показали результаты использования корпусов текстов для практических исследований, многие лингвистические задачи с их помощью не могут быть решены. Так, во многих языках нельзя установить принадлежность слова предложения к тому или иному грамматическому классу (имени существительному, глаголу и т. п.), что, в свою очередь, не позволяет определить частоту употребления грамматических классов слов, структуры предложений на уровне классов слов (частей речи), а значит, и употребительность таких структур. Без специальной подготовки текста не разрешимы омография морфем и слов, полисемия и целый ряд других важных задач. Поэтому в последние годы стали создаваться таггированные (размеченные) корпусы текстов (от англ. tag— 'индекс, помета'). Все слова такого корпуса получают некоторые буквенные или цифровые индексы, которые обозначают их грамматические, лексические, семантические или структурные признаки. Таких индексов может быть несколько. Принципы разработки подобных индексов (кодов) и их практического использования впервые также были предложены Р. Г. Пиотровским и его учениками. Например, слова предложения Этой весной опять расцвела акация получили следующий набор индексов :

Этой — МЖЕТ21 весной — СЖЕТ22 опять — Н22 расцвела — ГЖЕПЗЗ акация — СЖЕИЧ42.

Первый индекс у каждого слова указывает на класс слова (М — местоимение, С — имя существительное, Н — наречие, Г — глагол), второй индекс служит для обозначения рода, третий — числа, четвертый — падежа (или времени для глагола). Первая цифра в конце кода указывает на число слогов в слове, а вторая — на место ударного слога.
Естественно, «приписывание» таких индексов словам текстов — задача достаточно сложная и дорогостоящая. Для некоторых языков с определенной долей вероятности такое приписывание может быть сделано автоматически или полуавтоматически.

Существует огромное число подходов к разметке текстов и его единиц в корпусе текстов. Для этих целей можно выделить следующие группы признаков:

1) признаки для кодирования отдельного текста; признаки для кодирования предложения;
2) признаки для кодирования словосочетаний;
3) признаки для кодирования слова;
4) признаки для кодирования морфемы.

К первой группе признаков относятся следующие:
1) название текста (полное и краткое);
2) фамилия, имя и отчество автора;
3) псевдоним автора (если есть);
4) дата рождения автора (и смерти, если автор умер);
5) время создания текста (или его первой публикации);
6) место первой публикации;
7) тип текста (художественный, научный и т.п.);
8) графика текста (кириллица, латиница, книжно-славянская и т.п.);
9) наличие в тексте таблиц, графиков, рисунков;
10) указание на рецензии и критические материалы к тексту;
11) адрес места хранения текста;
12) другая информация, важная для проведения конкретных исследований.
Для кодирования отдельных предложений текста используются следующие лингвистические

признаки:

1) тип предложения (простое или сложное);
2) тип сложного предложения;
3) вид придаточного предложения;
4) число сегментов (формальных групп: группа подлежащего, группа сказуемого и т.п.);
5) структурная формула предложения на уровне классов слов и/или членов предложения;
6) другая информация.

Словосочетание может получить в корпусе текстов следующие признаки:
1) число слов в словосочетании;
2)тип (свободное, идиоматическое, фразеологическое);
3)тип по главному слову (именное, нагольное, предложно-именное и т.п.);
4)тип связи между главным словом и другими словами слово сочетания (согласование, управление, примыкание и др.);
5)другая информация.

Слово (словоупотребление) максимально может иметь следующую информацию:
1) признак класса слова;
2) тип слова (простое, сложное, составное);
3) набор грамматических признаков;
4) семантический признак;
5) синтаксический признак;
6) стилистический признак;
7) наличие приставки;
8) наличие суффикса;
9) наличие окончания;
10)структурная формула слова;
11)другая информация.

Для отдельной морфемы слова возможна такая информация:

1) число букв в морфеме;
2) тип морфемы (приставка, основа, суффикс, окончание);
3) семантический тип морфемы («отрицание», «уменьшительность» и т.п.);
4) другая информация.

Так как чаще всего основной единицей корпуса текстов является словоупотребление, то весь возможный объем информации, который может быть получен из таггированных текстов, зависит от того, насколько удачно проведено кодирование каждого отдельного словоупотребления. Пока, к сожалению, не разработаны единые обозначения для кодирования грамматических, семантических и стилистических признаков словоупотребления в создаваемых корпусах текстов.

С опорой на правильно проведенное кодирование всех словоупотреблений предложения во многих случаях может быть проведено автоматическое таггирование самого предложения. Особая проблема возникает при подготовке корпусов параллельных текстов. Она заключается в установлении соответствий между текстом оригинала и его переводами. Для решения такой задачи используется так называемый метод автоматического выравнивания текстов. Его суть заключается в параллельной сегментации оригинального текста и его

перевода по предложениям. При этом могут использоваться шесть возможных соответствий между предложениями обоих текстов.

1.Одно исходное предложение переводится одним предложением.
2.Два исходных предложения переводятся одним предложением.
3.Одно исходное предложение переводится двумя предложениями.
4.Два исходных предложения переводятся двумя предложениями, но внутренние границы этих предложений в тексте оригинала и тексте перевода не совпадают.
5.Предложение исходного текста не переводится.
6.Предложение в тексте перевода не имеет эквивалента в тексте оригинала.

Одна из проблем, возникающая при создании и эффективном использовании корпусов текстов, связана с созданием удобного для пользователя языка, позволяющего извлекать из корпуса максимум информации. По существу задача сводится к созданию специальной поисковой системы, способной функционировать в сети Интернет. Такая система должна иметь возможность найти все те тексты, предложения, словосочетания, слова и морфемы, которые обладают перечисленными выше признаками.

Еще больший объем информации позволяют получить параллельные корпусы текстов. С их помощью, дополнительно к вышесказанному, можно:

1) автоматически строить двуязычные и многоязычные переводные словари;
2) автоматически создавать и пополнять словари для систем машинного перевода;
3) автоматически устранять полисемию лексических единиц. Это становится возможным в связи с тем, что компьютер может использовать контекстное окружение многозначного слова, по длине превышающее предложение;
4) автоматически переводить терминологические и фразеологические единицы текста;
5) осуществлять полностью автоматический перевод в рамкахновых систем машинного перевода, называемых системами с переводческой памятью. Суть данного подхода заключается в том что в памяти компьютера накапливаются корпусы исходных текстов и их переводов, выравненных между собой на различных уровнях.

В процессе перевода такая система пытается отыскать переводимое предложение в массиве исходных параллельных текстов. Если оно найдено в исходном массиве текстов— оригиналов, то система выбирает перевод такого предложения или его части в массиве переведенных текстов.

Фонетические лингвистические ресурсы

Как видно из общей структуры лингвистических ресурсов , их составной частью являются также фонетические лингвистические ресурсы. При их создании возникают те же проблемы, что и при создании письменных текстовых массивов. Однако главная трудность создания фонетических лингвистических ресурсов связана с необходимостью транскрибирования устной речи. При этом возникают следующие проблемы:

1. Какой алгоритм использовать для транскрибирования?
2. Учитывать ли индивидуальные особенности произношения?
3. Учитывать ли весь устный текст или его фрагменты?
4. Учитывать ли диалектные варианты произношения слов?
5. Учитывать ли ударения в словах?
6. Учитывать ли просодические признаки произносимых фраз?

7. Отмечать ли слова, которые при прослушивании не распознавались?
8. Отмечать ли в записи для фонетического корпуса паралингвистические явления, сопутствующие речи (паузы, смех, бормотание, кашель и т. п.)?

В настоящее время общепринято, что для создания машиночитаемых фонетических корпусов используется транскрипция на основе орфографического представления звуков речи с дополнительными знаками, передающими (при необходимости) просодические, паралингвистические и другие особенности произношения.

Несмотря на трудности создания, в мире уже существует много достаточно представительных фонетических корпусов. Так, в 70-х годах XX века в США Х. Далем и его коллегами был создан «Корпус устной речи американского варианта английского языка». Он включал 1000000 словоупотреблений, взятых из записей психоаналитических сеансов. С каждой из 15 кассет, имевшихся в распоряжении составителей корпуса, было случайным образом отобрано 225 записей сеансов. Они содержали речь 8 женщин и 21 мужчины из 9 городов США. Отобранные записи были транскрибированы на основе стандартной английской орфографии. Диалектные варианты произношения не учитывались. Нераспознанные слова при записи обозначались буквой Z. Ударения и другие просодические характеристики речи также не учитывались.
В то же время при орфографической записи устной речи в качестве специальных комментариев отмечались паузы, смех, вздох, кашель и другие паралингвистические явления. Известен Международный машиночитаемый архив современного английского языка (The International Computer Archive of Modern English — ICAME). В последние годы предлагается коммерческий «Корпус разговорного профессионального американского варианта английского языка» (Corpus of Spoken Professional American English). Он включает 2000000 слов с индексами части речи при каждом слове (его можно найти по URL адресам www.athelstan.com и www.athel.com').
Существует несколько фонетических корпусов немецкой устной речи. Одним из первых является Фрейбургский корпус. Он создан на базе 820 магнитофонных записей устной немецкой речи тонца 60-х — начала 70-х годов XX века во Фрейбургском отделении Института немецкого языка. Корпус включает записи радиопередач, а также различных конференций, заседаний и других общественных мероприятий. Для транскрибирования было отобрано 122 текста различного объема (от 175 до 16390 словоупотреблений) общей длиной в 600000 словоупотреблений. Для приведения записей в машиночитаемый вид была разработана специальная система транскрипции, опирающаяся на стандартную немецкую орфографию. По указанным выше URL-адресам организация ELRA предлагает и фонетические корпусы текстов. Так, норвежский фонетический корпус (ELRA-SOO81 Norwegian Speech Dat (II) FDB-1000) содержит записи телефонных разговоров 1016 носителей норвежского языка (517 мужчин и 499 женщин). Известен датский корпусразговорной речи, содержащий 10000000 слов.

Фонетические корпусы текстов широко используются для решения следующих задач:
1) сопоставительного изучения устной и письменной форм языка;
2) изучения грамматических и лексических особенностей устной речи;
3) исследования фонетических особенностей диалектов;
4) построения частотных списков фонем и их сочетаний;
5) изучения акустических свойств речевых единиц и их использования в психолингвистических и лингвистических экспериментах;
6) создания компьютерных систем, распознавания и синтеза устной речи.

ТЕМА 4. ОПЫТ СОЗДАНИЯ ЛИНГВИСТИЧЕСКИХ БАЗ ДАННЫХ В АМУРСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ

(по материалам опубликованных статей)

Булатова Надежда Яковлевна
Институт лингвистических исследований Российской Академии Наук
Санкт-Петербург, Российская Федерация
Nadezhda Ya. Bulatova
Institute of Linguistic Studies of the Russian Academy of Sciences
StPetersburg, Russian Federation
bulatovany@gmail.com

Морозова Ольга Николаевна
Амурский государственный университет
г.Благовещенск, Российская Федерация
Olga N. Morozova
Amur State University
Blagoveshchensk, Russian Federation
morozova_olga06@mail.ru

ЗАДАЧИ И ПРИНЦИПЫ СОСТАВЛЕНИЯ ЗВУКОВОГО ЭВЕНКИЙСКО-РУССКО-АНГЛИЙСКОГО ТЕМАТИЧЕСКОГО СЛОВАРЯ[*]

GOALS AND PRINCIPLES OF COMPILING SOUND EVENKI-RUSSIAN-ENGLISH SUBJECT DICTIONARY

Аннотация
Звуковой эвенкийско-русско-английский тематический словарь разработан на материале находящегося под угрозой исчезновения селемджинского говора эвенкийского языка и направлен на решение глобальной проблемы по сохранению, развитию и исследованию языков малочисленных народов России. Словарный материал разделен на тематические блоки: «Явления природы», «Охота», «Жилище» и так далее. Объем словаря составляет около 2600 лексических единиц. Словарь разработан в двух версиях: печатной и электронной. Звуковая версия представлена в произнесении дикторов с сохранной апикальной эвенкийской артикуляцией. Электронный формат звукового словаря окажет положительное влияние на развитие мотивации и поддержание интереса у эвенкийской молодежи к изучению родной речи.

Abstract
The project of audio subject dictionary is worked out on the material of the endangered Selemdzha Evenki language. The project is aimed at the solution of the global problem of saving, surviving and researching of languages of minor peoples of Russia. The lexical material of the dictionary is divided on topics, e.g. «Natural phenomena», «Hunting», «Dwelling» etc. The dictionary contains about 2600 audiofiles in CD attached to the book. Words are pronounced by dictors with Evenki apical articulation. This electronic format of audio dictionary will have a positive influence on the development of motivation and interest of young Evenks in order to learn their native language.

Ключевые слова: эвенкийский язык, звуковой словарь, документация исчезающих языков мира.

[*] Работа выполнена при поддержке гранта РФФИ № 17-04-12004в

Keywords: the Evenki language, sound dictionary, documentation of the endangered languages.

1. Введение

Составление словарей, учебников и учебных пособий для коренных народов Севера, Сибири и Дальнего Востока всегда было одной насущных и востребованных задач, связанных с сохранением исчезающих национальных культурных и языковых традиций малочисленных этносов мира. Данная проблематика поднималась для эвенкийского языка во время круглого стола «Современные методы обучения эвенкийскому языку», проходившего в рамках региональной научно-методической видеоконференции, организованной Амурским государственным университетом [Морозова, 2013].

Разработанный «Звуковой эвенкийско-русско-английский тематический словарь: на материале селемджинского говора эвенкийкого языка» авторами Н. Я. Булатовой, О. Н. Морозовой, Г. А. Стручковым (см. обложку словаря на рис. 1) явился результатом кропотливой работы по сбору диалектной лексики эвенкийского языка, её перевода на русский и английский языки. Необходимо отметить, что данный словарь вносит весомый вклад в дело сохранения исчезающего говора малочисленного этноса Приамурья – селемджинских эвенков. Звуковая часть словаря (аудиофайлы словарной лексики) является уникальной, поскольку звуковых образцов селемджинского говора эвенкийского языка в аудиоархивах по языкам малочисленных народов РФ не имеется. Актуальность данной работы не вызывает сомнения, поскольку самое пристальное внимание диалектологов в последнее время направлено на лексику, номинирующую определённые формы народной культуры местного населения. При этом особый интерес представляют недостаточно изученные в лингвистическом отношении номинации предметов материальной культуры малочисленных народов Севера, в том числе наименования пищи, одежды, построек, орудий труда и так далее, поскольку именно они являются основными источниками информации о повседневной жизни и деятельности человека.

2. Цель, задачи и принципы составления словаря

Цель данной статьи – дать описание подходам, используемым при составлении звукового эвенкийско-русско-английского тематического словаря. Для достижения поставленной цели необходимо следующее: 1) объяснить причину обращения к материалу селемджинского говора эвенкийского языка и его перевода не только на русский, но и английский языки; 2) дать описание требованиям, предъявляемым к дикторам при озвучивании словарных единиц; 3) рассмотреть параметры применяемых при записи речи и программировании компьютерных программ; 4) раскрыть принципы и подходы при составлении лексической части словаря; 5) выявить целевые группы потенциальных пользователей словарём.

Предлагаемый читателю звуковой тематический словарь является трёхъязычным переводным словарём; он отличается от других аналогичных словарей некоторыми особенностями, в зависимости от той цели, какую ставит перед собой обращающийся к словарю читатель, и от того, насколько он знает эвенкийский язык, с одной стороны, русский и английский – с другой.

Перевод эвенкийской и русской лексики на английский язык обусловлен статусом английского языка как средства международного общения. В последнее время зарубежные исследователи проявляют активный интерес к тунгусо-маньчжурским языкам. Мы учитываем и тот факт, что у коренных малочисленных народов России появилась возможность участвовать в грантах, предоставляемых англоязычными и иными фондами поддержки исчезающих языков мира. Кроме этого, такая подача материала поможет учащимся в

изучении как эвенкийского, так и английского языка.

Обращение к селемджинскому говору эвенкийского языка обусловлено малоизученностью речи селемджинских эвенков [Булатова, 1987, с. 11]. Не до конца выявлены в данном говоре лексемы, общие для всех диалектов эвенкийского языка, а также диалектизмы, являющиеся языковыми фактами селемджинского говора. Звуковой архив словарных лексем в произнесении дикторов, родившихся на селемджинской земле и усвоивших родной язык в условиях кочевой жизни, является не только уникальным, но пригодным для многоуровневых лингвистических исследований. Созданию такого звукового архива способствовало жёсткое соблюдение требований к получению высококачественной записи речи. Запись речи производилась в условиях Лаборатории экспериментально-фонетических исследований при кафедре иностранных языков Амурского государственного университета. Запись производилась с микрофона через микшерный пульт на звуковую плату компьютера при помощи программы AUDACITY. Устройство с микрофоном было закреплено на голове на расстоянии 2,5 см. от лица говорящего. При записи использовались обязательные для получения фонетически представительного материала параметры оцифровки: частота дискретизации 44 кГц, разрядность – 16 бит, моно. В ходе записи был сформирован корпус словарного материала, где каждая звучащая лексема записана в троекратном произнесении для фиксации интонационной вариативности восходящего, ровного и нисходящего тонов. Массив словарного материала впоследствии был сегментирован на отрезки, равные звучанию троекратной подачи словарной лексемы. Полученные лексемы были записаны в отдельные файлы формата wav.

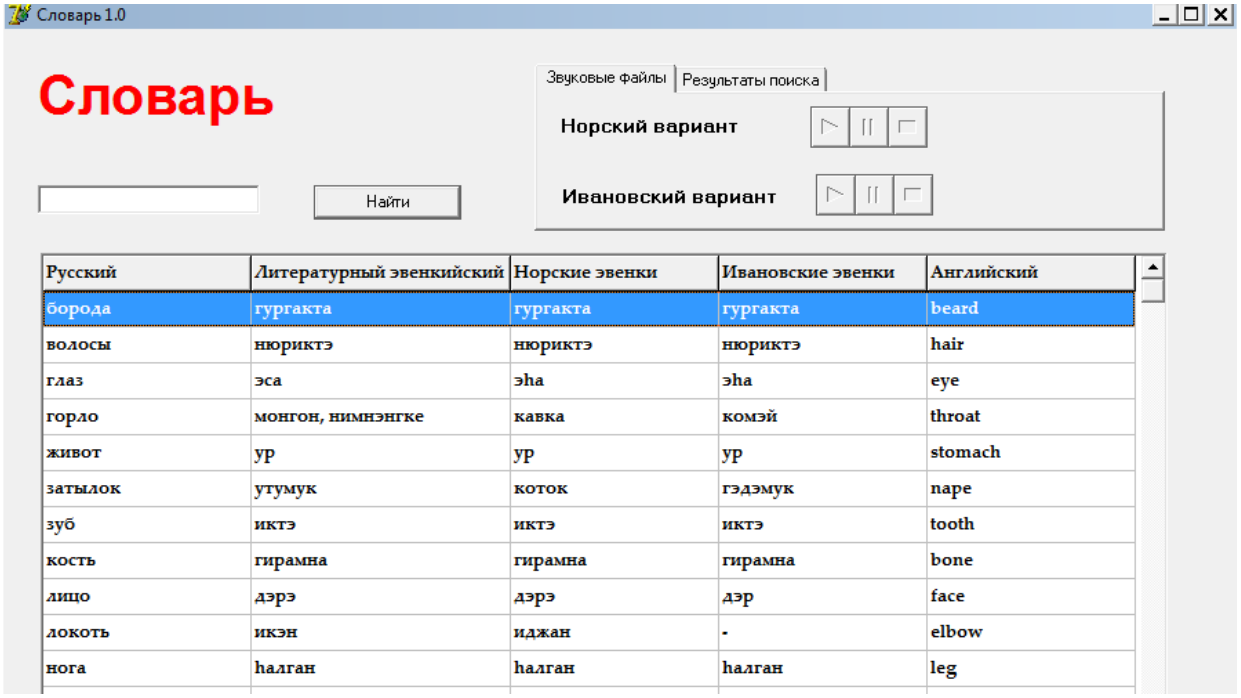
При отборе потенциальных дикторов словаря авторы придерживались принципа сохранности в их речи апикального характера артикуляции. В настоящее время ни для кого не является секретом дорсальный русский уклад речи у современных эвенков, когда кончик языка находится у нижних зубов, а спинка языка поднята к твердому нёбу. Если же обратиться к данным литературных источников, то выясняется, что в конце 30-х годов XX в. артикуляционный уклад речевых органов у эвенков был именно апикальным, когда кончик языка направлен вверх к альвеолам [Матусевич, 1960, с. 138]. Из 16 возможных дикторов-носителей селемджинского говора эвенкийского языка были выбраны мужчина и женщина, параметры социолингвистических портретов которых отвечают следующим требованиям: 1) усвоение родного языка в кругу семьи с младенческого возраста; 2) непосещение детских комбинатов и иных дошкольных учреждений с русскоязычным персоналом; 3) коммуникация только на родном языке до поступления в школу; 4) возраст старше 60 лет; 5) отсутствие дефектов речи и слуха; 6) сохранность апикального уклада произносительных органов, свойственных эвенкийскому языку [Морозова, 2015, с. 76].

3. История создания тестовой версии словаря

Первая тестовая программная версия звукового словаря была разработана в феврале 2013 г. на основе оболочки Borland Delphi 6.0 и содержала только тематический блок лексики «Человек» [Морозова, et al., 2014, с. 186]. При практическом использовании данной версии необходимо было набрать в строке поиска требуемое слово на одном из предлагаемых языков (русском, эвенкийском (литературном эвенкийском, селемджинском говоре (норская или ивановская разновидности), английском). После данной команды искомое слово появлялось в верхней строке списка слов. При выделении строки существовала возможность звукового проигрывания слова в произнесении эвенков-жителей с. Ивановского Селемджинского района или п. Майский Мазановского района Амурской области (см. рис. 2).

Однако тестовая версия звукового словаря показала недостатки компьютерной оболочки Borland Delphi 6.0, которая не поддерживает фонетические шрифты и специальные диакритические знаки, необходимые для отображения долгот эвенкийских гласных на письме, а также не имеет возможности отображать графему ӈ для носового заднеязычного

/ŋ/. Написание же знака долготы для эвенкийских гласных мы считаем традиционным, нормативным и обязательным, хотя отсутствие теоретических оснований для обозначения долготы гласных в эвенкийском языке оставляет за пользователями и специалистами-лингвистами альтернативу не обозначать или обозначать долготу посредством надчёркивания черты над долгими гласными [Бурыкин, 2004, с. 233]. С целью избежать вышеуказанные недостатки, было решено использовать алгоритмический язык C++, достоинством которого является, помимо поддержки знаков фонетического алфавита, его совместимость не только с Microsoft Windows, но и другими операционными системами.



Русский	Литературный эвенкийский	Норские эвенки	Ивановские эвенки	Английский
борода	гургакта	гургакта	гургакта	beard
волосы	нюриктэ	нюриктэ	нюриктэ	hair
глаз	эса	эһа	эһа	eye
горло	монгон, нимнэнгке	кавка	комэй	throat
живот	ур	ур	ур	stomach
затылок	утумук	коток	гэдэмук	nape
зуб	иктэ	иктэ	иктэ	tooth
кость	гирамна	гирамна	гирамна	bone
лицо	дэрэ	дэрэ	дэр	face
локоть	икэн	иджан	-	elbow
нога	налган	налган	налган	leg

Рисунок2. **Тестовая версия звукового словаря**

4. О пользовании словарём

Текст словаря имеет общепринятую для многоязычных словарей структуру, в ней выдержан табличный способ подачи материала, наиболее удобный для использования его читателями. Кроме того, в словаре имеются ссылки на известные словари эвенкийского языка, что говорит о лексикографической выверке авторами полученного словарного материала.

Весь словарный материал поделён на тематические разделы: «Возгласы», «Время», «Животный мир», «Жилище. Предметы быта», «Инструменты», «Приспособления, принадлежности», «Люди. Родственники», «Неживая природа», «Одежда и обувь», «Оленеводство», «Охота», «Профессии», «Продукты», «Растения», «Религия», «Тело человека», «Явления природы», «Глаголы», «Местоимения», «Наречия», «Прилагательные. Причастия», «Существительные», «Числительные» [Булатова et al., 2017].

Внутри каждого раздела лексика дана в алфавитном порядке в соответствии с выбираемым языком. Исключение составляют разделы «Время» и «Числительные». В разделе «Время» указаны лексемы в последовательности: годы, сезоны, месяцы, дни, время суток. В разделе «Числительные» лексика дана, начиная от числительных первого и второго десятка, наименований десятков, сотен, тысяч. Затем идут порядковые числительные и прочие.

Особенность тематического раздела «Охота» состоит в том, что все глаголы, относящиеся к данной области хозяйственной деятельности, приведены в разделе «Глаголы».

Необходимо особо отметить раздел «Существительные», куда были помещены отвлечённые, абстрактные существительные, названия металлов, болезней, учреждений и прочие, которые не могли быть причислены ни к одному из имеющихся тематических разделов. Кроме того, слово, имеющее разные значения, может встречаться в разных разделах, например, *кӯрчэк* в значении «молоко взбитое» отнесено в «Продукты», в значении «мутовка для приготовления взбитого молока» – в раздел «Кухонные принадлежности».

В случае регулярных фонетических соответствий даются оба варианта слова, например, *эмэ̄нму-мӣ*, *эмэ̄мму-мӣ* «остаться». В некоторых случаях оставлено произношение диктора, участвовавшего в аудиозаписи словаря, поскольку предпочитаемый им вариант слова был в его речи частотным, например, *соӈинтугин* (ср. лит. эвенк. *согинтугин*).

Пользователи электронной версии словаря могут загрузить его в любом браузере компьютера (Internet Explorer, Google Chrome, Yandex, Mozilla Firefox, Apple Safary и др.). Электронную версию словаря отличает удобное перелистывание страниц, быстрая навигация по тематическим разделам, озвучивание эвенкийской лексики при наведении курсора. Имеется возможность приближения текста словарной лексики, что удобно при сниженном зрении читателя. Кроме того, в электронную версию добавлены функции быстрой навигации и поиска слов во всём массиве словаря, печати выбранных страниц.

Целевыми группами пользователей словарём являются учащиеся эвенкийских школ, учителя эвенкийского языка, преподаватели и студенты филологических факультетов вузов, а также специалисты в области тунгусо-маньчжуроведения и алтаистики. Звуковая версия словаря будет интересна не только эвенкам, проживающим в нашей стране, но и эвенкам и орочонам Внутренней Монголии и провинции Хэйлунцзян Китайской Народной Республики.

5. Заключение

В заключение хотелось бы отметить возможную востребованность звуковых образцов словаря для создания различного рода образовательной и лингвистической продукции, в том числе компьютерных программ на электронных носителях на материале эвенкийского языка, пригодных для размещения в сети Интернет и использования on-line, учебных пособий и грамматических справочников эвенкийского языка, а также экспериментальных материалов для различного рода лингвистических, литературоведческих и культурологических исследований.

Рисунок1. Обложка словаря с диском

Список литературы

1. Булатова Н. Я. Говоры эвенков Амурской области [Текст] / Н. Я. Булатова. – Л., 1987. – 168 с.

2. Булатова Н. Я. Звуковой эвенкийско-русско-английский тематический словарь. Звуковой русско-эвенкийско-английский тематический словарь [Текст] / Н. Я. Булатова, О. Н. Морозова, Г. А. Стручков / отв. ред. О. Н. Морозова, науч. ред. Н. Я. Булатова. – Благовещенск : Изд-во ОДЕОН. 2017. – 184 с.

3. Бурыкин, А. А. Язык малочисленного народа в его письменной форме (на материале эвенского языка) [Текст] / А. А. Бурыкин. – СПб. : Петербургское Востоковедение, 2004. – 384 с.

4. Матусевич, М. И. Очерк системы фонем ербогоченского говора эвенкийского языка на основе экспериментальных данных [Текст] / М. И. Матусевич // Учёные записки ЛГУ. Сер. филол. наук. – 1960. – Вып. 40. – № 237. – С. 132–169.

5. Морозова, О. Н. Артикуляторно-акустические характеристики переднеязычного глухого смычного /t/ в эвенкийского языке [Текст] / О. Н. Морозова // Теоретическая и прикладная лингвистика. – 2015. – Вып. 1. – № 1. – С. 74–85.

6. Морозова О. Н. Концепция и структура звукового тематического словаря эвенкийского языка [Текст] / О. Н. Морозова, А. Лаврилье, В. С. Гибалин // Вестник Амур. гос. ун-та. – 2014. – Вып. 64. – С. 189–196.

7. Морозова О. Н. Материалы круглого стола «Современные методы обучения эвенкийскому языку» [Текст] / О. Н. Морозова // Материалы региональной научно-методической видеоконференции «Образовательный процесс обучения иностранному языку на неязыковых отделениях вуза в условиях перехода на двухуровневую систему образования». – Благовещенск : Амур. гос. ун-т, 2013. – С. 138–140.

References

1. Bulatova, N. Ya. (1987). Govory evenkov Amurskoy oblasti. Materialy issledovaniya. [Accents of Amur Region Evenks. Research materials]. Leningrad : Nauka Press.

2. Bulatova, N. Ya., Morozova O. N., Struchkov G. A. (2017). Zvukovoy evenkiysko-russko-angliyskiy tematicheskiy slovar' na materiale selemdzhinskogo govora evenkiyskogo yazyka. [Sound Evenki-Russian-English Subject Dictionary]. Blagoveshchensk : Odeon Press.

3. Burykin, A. A. (2004). Yazyk malochislennogo naroda v ego pis'mennoy forme (na materiale evenkiyskogo yazyka) [A minority language in its written form (Based on Evenki)]. St. Petersburg : Asian-Studies-in-Petersburg Press.

4. Matusevich, M. I. (1960). Ocherk sistemy fonem erbogochenskogo govora evenkiyskogo yazyka na osnove eksperimental'nykh dannykh [The outline of Erbogochen accent of the Evenki language phonological system]. Uchenye zapiski LGU. Ser. filol. Nauk [Scientific papers of Leningrad State University. Series Philological Sciences], 40 (237), 132–169.

5. Morozova, O. N. (2015). Artikulyatorno-akusticheskie kharakteristiki peredne-yazychnogo glukhogo smychnogo /t/ v evenkiyskom yazyke [Articulatory and acoustic features of the fore-lingual voiceless plosive consonant /t/ in the Evenki language]. Teoreticheskaya i prikladnaya lingvistika [Theoretical and Applied Linguistics], 1 (1), 74–85.

6. Morozova, O. N., Lavrillier, A., Gibalin, V. S. (2014). Kontseptsiya i struktura zvukovogo tematicheskogo slovarya evenkiyskogo yazyka [The concept and structure of the Evenki subject dictionary]. *Vestnik Amurskogo gosudarstvennogo universiteta* [Vestnik of the Amur State University], 64, 189–196.

7. Morozova, O. N. (2013). Materialy kruglogo stola «Sovremennye metody obucheniya evenkiyskomu yazyku» [Proc. of the round-table discussion «Modern methods of the Evenki language teaching»]. Materialy regional'noy nauchno-metodicheskoy videokonferentsii «Obrazovatel'nyy protsess obucheniya inostrannomu yazyku na neyazykovykh otdeleniyakh vuza v usloviyakh perekhoda na dvukhurovnevuyu sistemu obrazovaniya» [Proc. of the Regional scientific and methodological video conference «Educational process of foreign language teaching at non-linguistic department»] (pp. 138–140). Blagoveshchensk : Amur State University Press.

С.В.Андросова, О.Н.Морозова, М.В.Артемчук

Амурский государственный университет

Корпус звучащей эвенкийской речи

Аннотация
В эпоху глобализации при господстве информационных технологий проблема экологии языков стоит особенно остро. Эвенкийский язык – один из языков, находящихся под угрозой исчезновения. Эвенкийско-русский билингвизм для подавляющего большинства эвенков Амурской области уже стал русско-эвенкийским, а количество эвенков, свободно говорящих на родном языке, мизерно. Информационные технологии как часть благ цивилизации, запустившие этот деструктивный процесс, имеют прекрасный потенциал его замедлить и, возможно, даже обернуть вспять через создание электронных аннотированных звуковых корпусов свободного доступа. Особый статус эвенкийского языка и недоступность или отсутствие звуковых подробно аннотированных материалов по говорам его восточного наречия послужили импульсом для начала работы над соответствующим корпусом.
Нами создана информационная система (веб-приложение написано на языке Ruby с использованием инструментария Rails) свободного доступа к базе данных по селемджинскому, джелтулакскому и зейскому говорам. Интерфейс базы данных представлен стандартными страницами: главная, новости, корпус, о проекте, ресурсы и контакты. На главную страницу систематически выкладываются новости о пополнении базы новыми речевыми образцами и о мероприятиях, связанных с эвенкийским и близкородственным орочонским языками, в частности об олимпиаде по указанным языкам, проводимую на базе Амурского государственного университета. На странице «Корпус», по настоянию самих дикторов, дается необходимая персональная информация (ФИО, профессия, род, говор). Корпус предназначен для широкого круга пользователей – от наивных носителей языка до методистов, ведущих обучение данному языку, и исследователей-лингвистов, занимающихся изучением данного языка. Нами предложены технические решения ряда неоднозначных вопросов, связанных с особенностями сегментации и транскрибирования. Одна из таких проблем – наличие в ряде случаев поствокального турбулентного и импульсного шума на переходном участке от гласного к согласному. В настоящее время ведется активная работа над созданием аналогичного корпуса звучащей орочонской речи.

Ключевые слова: исчезающий язык, тип пользователя, аннотирование, транскрипция, сегментный уровень, супрасегментный уровень

1. Введение

Эпоха изоляции, которая в течение многих веков поддерживала языковое многообразие, канула в Лету. Ей на смену пришла эпоха глобализации, до зубов вооруженная информационными технологиями. В этой реальности проблема экологии языков стоит особенно остро. Ни для кого не секрет, что примерно каждые две недели на нашей планете исчезает один язык. Еще совсем недавно лингвистическое многообразие было представлено 6800 языками мира, теперь их уже на сотню меньше. Эвенкийский язык – один из тех, которые грозят исчезнуть через два поколения. Поголовный билингвизм, многообразие говоров при своеобразии письменной формы и сложностях стандартизации на лексическом и фонетическом уровнях языка – факторы, неуклонно ускоряющие данный процесс. Численность эвенкийского этноса катастрофически сокращается. Эвенкийско-русский билингвизм для подавляющего большинства эвенков уже стал русско-эвенкийским, а количество эвенков, свободно говорящих на родном языке, мизерно. Такая ситуация типична для Амурской области российского Дальнего востока (см. подробнее об этом в работе О. Н. Морозовой [Морозова, 2014]). Количество представителей данного этноса, которые учились говорить на эвенкийским языке с младенчества в кругу семьи и далее общались преимущественно на нём, ничтожно мало по сравнению с числом тех, которые изучали свой национальный язык лишь со школьной скамьи как иностранный с довольно ограниченным количество часов, отведённых на освоение данной учебной дисциплины. Языковая ситуация в местах компактного проживания амурских эвенков носит диспропорциональный, несбалансированный характер, выражающийся в абсолютном функциональном доминировании русского языка [Процукович, 2015, с. 88-89]. Возможно, повышение социального статуса и благосостояния, важное

для любого человека в принципе, многим эвенкам видится в том, чтобы стать частью русскоязычной цивилизации. Вместе с тем, эвенки ощущают, что блага её сомнительны, если цена этому – забвение национального языка.

Информационные технологии как часть благ цивилизации, запустившие этот деструктивный процесс, имеют прекрасный потенциал его замедлить и, возможно, даже обернуть вспять. Создание звуковых корпусов для исчезающих языков можно по праву назвать первостепенной задачей современной прикладной лингвистики. Эта задача активно выполняется как ведущими отечественными, так и зарубежными специалистами. Достаточно обратиться к проектам по звуковому корпусу эвенкийского языка, разрабатываемым российскими специалистами [Афанасьева, Раднаева, 2011; Казакевич, 2013, 2016]. Между тем, общедоступный подробно аннотированный корпус говоров восточного наречия эвенкийского языка отсутствует. Таким образом, особый статус эвенкийского языка и недоступность или отсутствие звуковых подробно аннотированных материалов по говорам его восточного наречия побудили нас начать работу над соответствующим корпусом.

2. Корпус звучащей эвенкийской речи (аннотированный)

2.1. Технические решения и интерфейс

Нами создана информационная система (интерфейс написан на языке Ruby) свободного доступа к базе данных по говорам восточного наречия эвенков, проживающих на территории Амурской области [Речевой корпус …, 2016–2017] (более подробно с техническими решениями можно ознакомиться в нашей предыдущей работе [Морозова et al., 2017]).

К настоящему моменту интерфейс базы данных представлен стандартными страницами: главная, новости, корпус, о проекте, ресурсы и контакты (см. рис. 1).

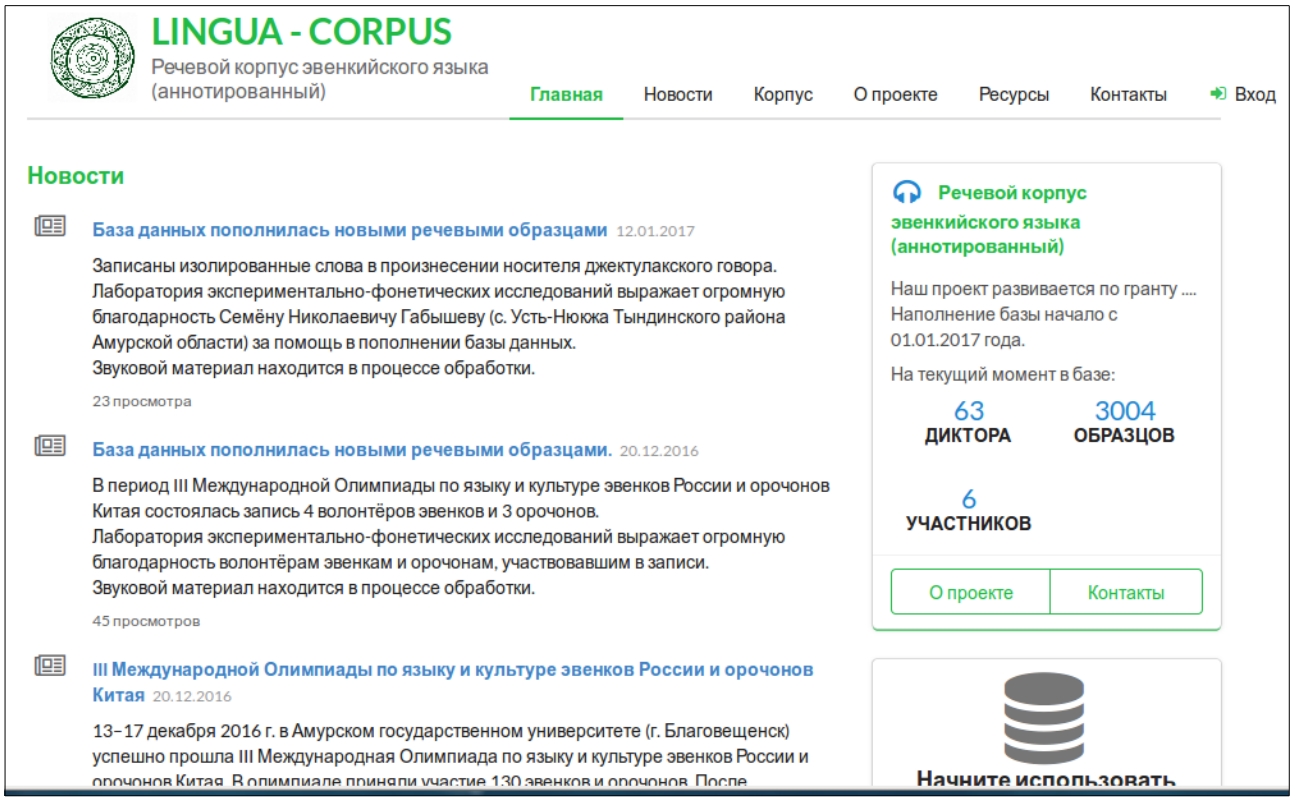


Рис. 1. Главная страница Корпуса
Fig. 1. Corpus main page

На главную страницу систематически выкладываются новости о пополнении базы новыми речевыми образцами и о мероприятиях, связанных с эвенкийским и близкородственным орочонским языками. В частности, отражены данные по ежегодной олимпиаде по языку эвенков России и близкородственному языку орочонов КНР, которая сравнительно недавно приобрела международный статус.

Здесь же представлена информация о количестве задействованных дикторов и полученных от них речевых образцов. Следует отметить, что активное пополнение базы данных осуществляется именно во время проведения данной олимпиады, а не только в ходе полевых экспедиций.

На странице «Корпус» представлены карточки дикторов. Обычно участие дикторов в речевых корпусах анонимно. Однако наши дикторы настаивали на внесение персональных данных в корпус и на том, чтобы эти данные были общедоступны. Часть дикторов также не возражали против размещения их фотографий (см. рис. 2). В каждой карточке приведены ФИО, профессия диктора, название эвенкийского рода и говор, количество аннотированных образцов на определенную дату.

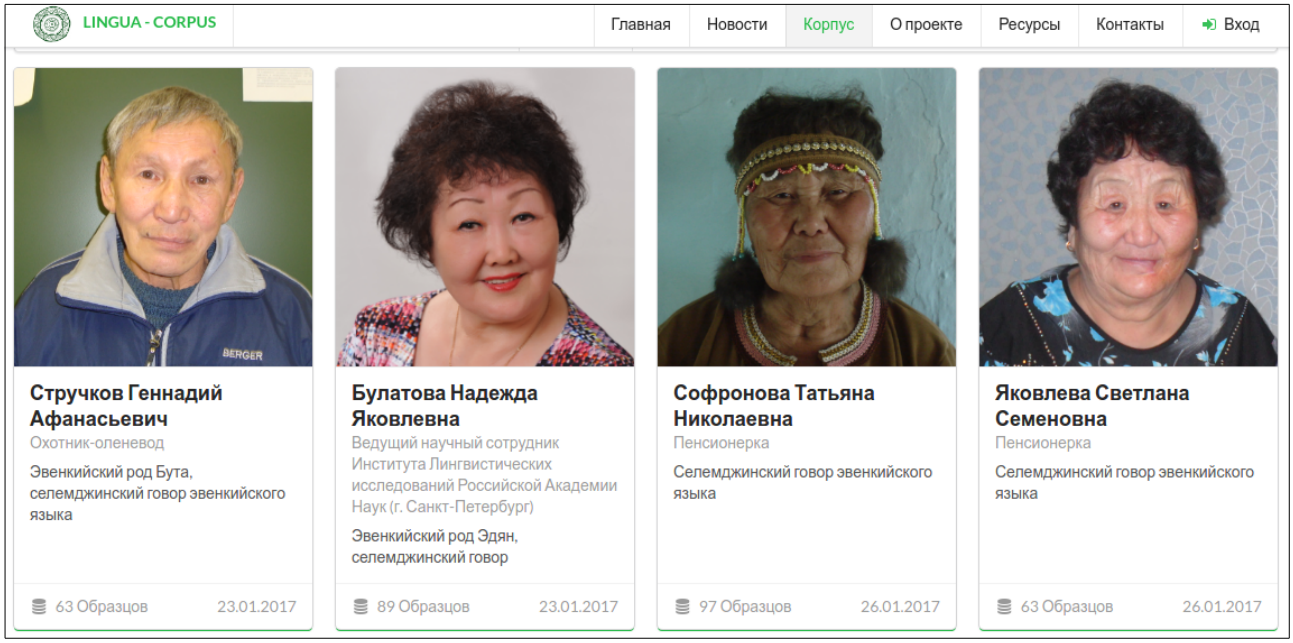


Рис. 2. Карточки дикторов
Fig. 2. Subjects' pages

Имеется специальная страница «О проекте», на которой кратко представлена его миссия, указаны используемые говоры, типы речевого и текстового материала и типы речевого сигнала, дана информация об участниках проекта (см. рис. 3-4).

Рис. 3. Общая информация о проекте
Fig. 3. General project info

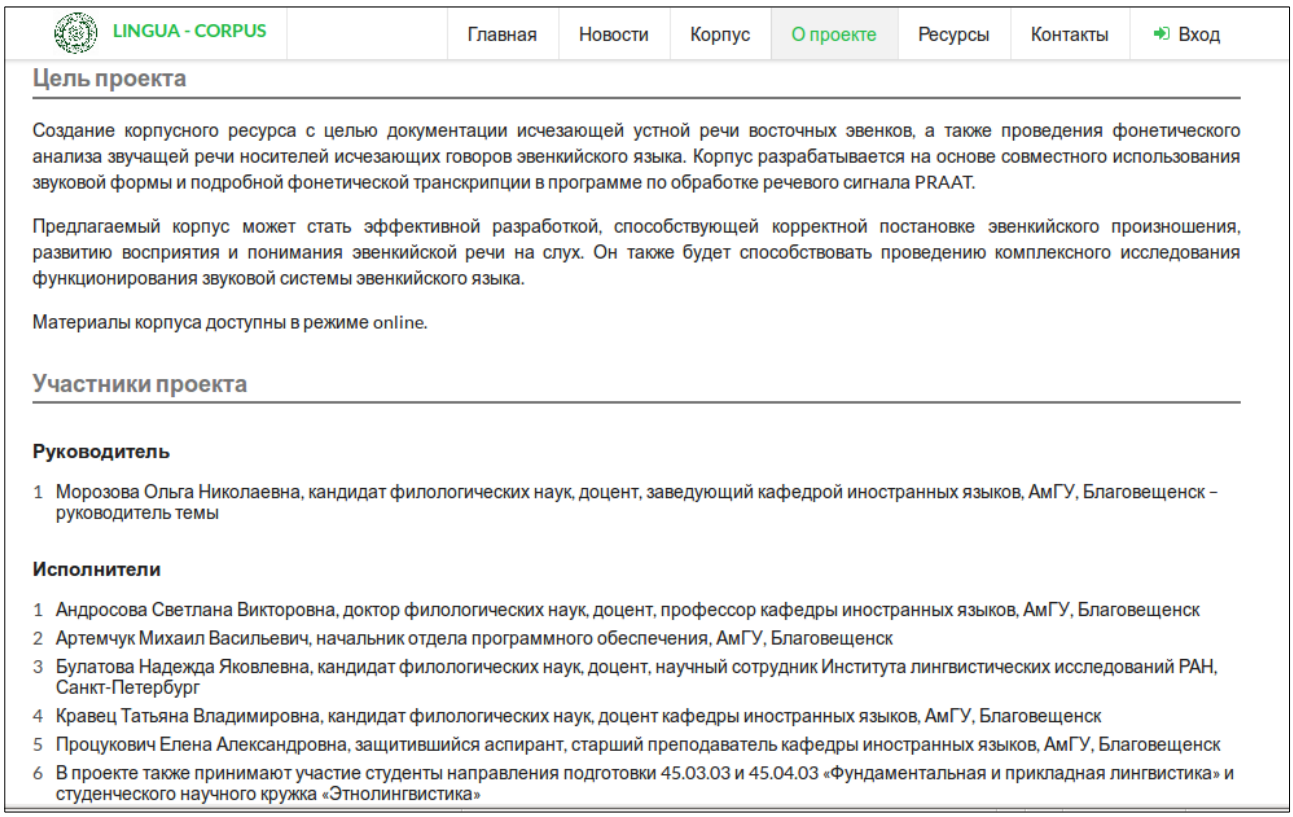


Рис. 4. Цель и участники проекта
Fig. 4. Aim and participants

Важно подчеркнуть, что в корпусе представлен весьма разноплановый материал. Во-первых, это начитанные изолированные слова в троекратном повторении, что в определенной степени имитирует просодию фразы с конечной деклинацией. Имеются и более мелкие единицы – начитанные слоги. Во-вторых, это начитанные одно- и многосинтагменные фразы, часть из которых организована по типу довольно популярных в экспериментальной лингвистике рамочных конструкций. В-третьих, это связные тексты, как начитанные так и спонтанные, как диалогического, так и монологического характера. Наконец, имеются и образцы эвенкийского устного народного творчества (сказания, героический эпос), произведенные теми немногими эвенками-старожилами, которые все еще прекрасно владеют языком своих предков.

Сегментация записанного материала производится студентами направления подготовки «Фундаментальная и прикладная лингвистика», а также аспирантами направления «Языкознание и литературоведение» профиля «Теория языка» в ЛЭФИ АмГУ под руководством высококвалифицированных специалистов по акустическому анализу речевого сигнала. Сегментация производилась с опорой на стандартные принципы, изложенные, в частности, в работе П. А. Скрелина [Скрелин, 1999].

Корпус предназначен для широкого круга пользователей – от наивных носителей языка до методистов, ведущих обучение данному языку, и исследователей-лингвистов, занимающихся изучением данного языка. В помощь пользователям имеется страница с ресурсами свободного доступа (см. рис. 5), куда помещены две мультиплатформенные (для операционных систем Windows, Linux и Mac) программы для обработки и анализа звукового сигнала. Здесь же в ресурсах для начинающих специалистов, интересующихся акустическим анализом, предлагается соответствующее пособие [Андросова, 2014], в котором кратко изложены основы работы в PRAAT и базовые акустические сведения с анализом примеров осциллографической и спектрографической картины конкретных звуков – согласных и гласных – и некоторых их модификаций. Многие эти сведения носят общефонетический характер и могут быть применены и для анализа реализаций эвенкийских фонетических единиц.

Рис. 5. Ресурсы
Fig. 5. Resources

2.2. Проблемы аннотирования и их решения

2.2.1. Общие принципы

Один из важных, на наш взгляд, общих принципов – оптимальное количество уровней в аннотации. Оно естественным образом определяется характером материала – меньшее количество уровней для более мелких единиц и большее – для более крупных (ср. рис. 6 с пятью интервальными уровнями и рис. 7 – с шестью). На уровнях с фонемного по аллофонный использовалась транскрипция разной степени детализации. Были использованы знаки международного фонетического алфавита (далее – МФА).

2.2.2. Просодия

При аннотировании разметка просодических особенностей заключалась в обозначении участков пауз (на данный момент – без указания типа паузы) и упрощенной фиксации мелодического компонента. Словесное ударение не отмечалось ввиду того, что авторы настоящей статьи, осознавая всю дискуссионность вопроса (см., напр., возможность музыкального характера эвенкийского ударения, выявленная на материале дисиллабов и трисилабов, начитанных с интонацией перечисления [Андреева, 2001]), придерживаются мнения об отсутствии такового в эвенкийском языке. Основанием этому служат результаты недавно проведенного нами эксперимента на материале рамочных конструкций и списка изолированных слов с целью выявить отличия в реализации фонетических характеристик ударности в изолированно произнесенном слове и слове в потоке речи в зависимости от структуры предложения [Морозова et al., 2014, 89–91]. Анализ мелодических характеристик показал наличие разных типов движения основного тона на одном и том же слове, но реализованном на начальном / срединном / конечном участках фразы и, таким образом, не подтвердил наличие в эвенкийском языке музыкального ударения. Возможно, такая разница с ранее полученными результатами Т. Е. Андреевой обусловлена, во-первых, разницей говоров, во-вторых, особенностью организации экспериментального материала. Все же, представляется необходимым говорить о тоновой природе словесного ударения в эвенкийском языке и сходстве типа акцентуации со слогоморфемными языками Юго-Восточной Азии и языками Африки (в которых морфема и слог не совпадают) с большой осторожностью.

В связи с вышеизложенным, производилось только транскрибирование мелодического компонента. Поскольку набор транскрипционных знаков МФА в PRAAT для обозначения особенностей движения основного тона (далее – ОТ) весьма ограничен, пришлось прибегнуть к тире и разнонаправ-

ленным слэшам для обозначения соответственно ровного и восходящего и нисходящего тонов. На данном этапе не ясно, возможна ли организация разметки ОТ, единая для всех речевых образцов. Так, для троекратно повторенных изолированных слов оказалось достаточным указать с помощью разнонаправленных знаков «\» и «/» нисходящее или восходящее движение ОТ на участке, соответствующем слову (см. разграничительные метки и знаки для ОТ на уровне «intonation» на рис. 6 и 7). Однако такая стратегия оказалась неподходящей для материала героического эпоса, прежде всего, в силу специфики певческой речи: разные движения ОТ на разных слогах в пределах одного слова и нередкие случаи разнонаправленного движения ОТ в пределах одного слога. Поэтому для таких речевых образцов принято решение производить мелодическую разметку не на участках слов, а на участках слогов (см. разграничительные метки и знаки для ОТ на уровне «intonation» на рис. 7).

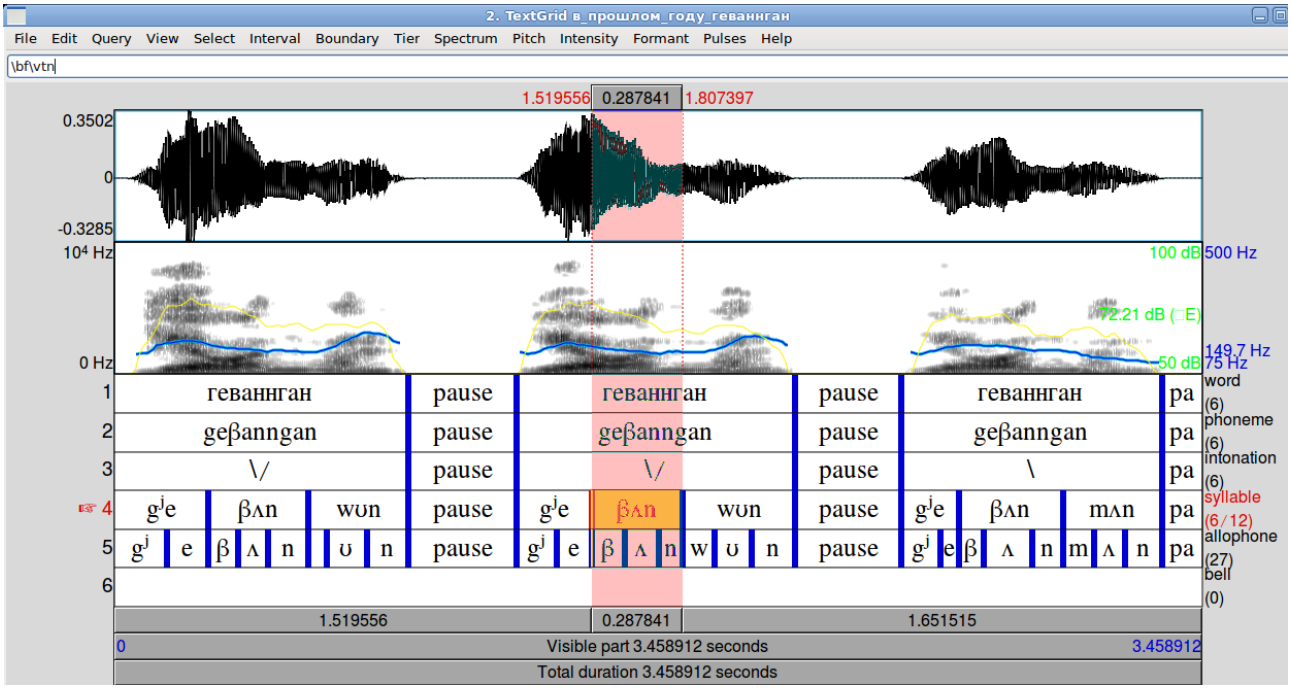


Рис. 6. Аннотация слова *геваннган* (= в прошлом году)
Fig. 6. Annotation of the word *gevanngan* (= last year)

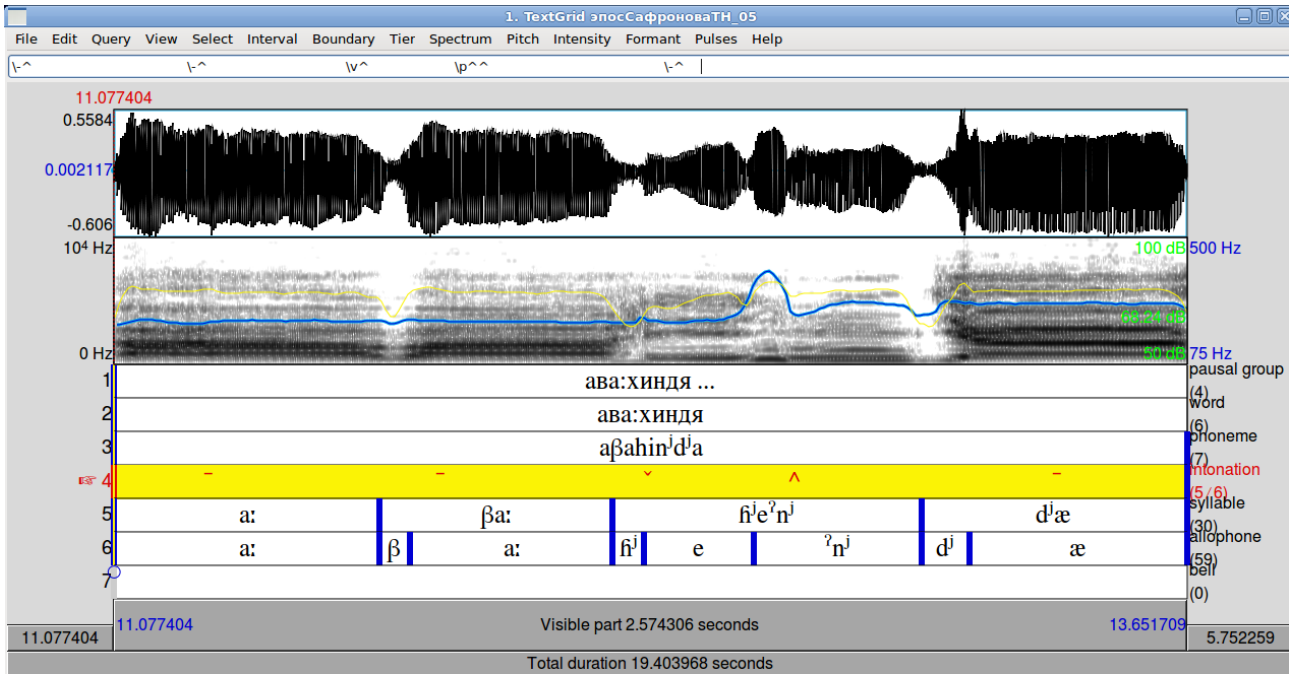


Рис. 7. Фрагмент аннотации героического эпоса
Fig. 7. A piece of heroic epic poem annotation

2.2.3. Сегментные единицы

Проблемы аннотирования на сегментных уровнях (фонемный и аллофонный) связаны с особенностями формирования письменной формы языка и перипетий становления эвенкийской орфоэпии. Одна из таких проблем – орфография и каноническая фонемная транскрипция на основе экающего произношения с предыдущим твердым согласным в примерах типа *сикс͡э*, вступающие в противоречие с акающим произношением с предыдущим мягким согласным: /sʲiksɛ/ в западных говорах vs /sʲiksʲæ/. Сложно надеяться на вариантную орфографию – *сикс͡э* vs *сикс͡я* – однако остается вопрос о вариантности фонемного или аллофонного состава этого и подобных слов. Помимо этого, при троекратном повторении изолированных слов диктор иногда мог реализовывать их с несколько отличающимся фонемным составом (ср. три реализации последнего слога в слове *геваннган* на рис. 6). Возможное объяснение этому заключается в том, что дикторы уже недостаточно хорошо помнят некоторые слова своего национального языка.

Другая проблема, с которой пришлось столкнуться – наличие турбулентного и импульсного шума, возможно, гортанного, в целом ряде случаев при переходе от гласного к следующему согласному (см. участки, отграниченные метками слева и справа на рис. 8–10). Данная особенность была отмечена в речи всех записанных дикторов. В связи с обнаруженным шумом в перспективе следует поднять вопрос об особенностях фонетического слога в изучаемых говорах восточного наречия эвенкийского языка. На данном этапе решено считать указанный шум окончанием гласного и, возможно, признаком плотного примыкания. Необходимо выяснить, насколько систематично он реализуется и какими факторами определяется его наличие или отсутствие.

Кроме того, пришлось столкнуться с совершенно естественным для речи на любом языке аллофонным варьированием, которое для эвенкийского языка описано очень неполно. Среди часто встречающихся феноменов были отмечены имплозивность смычно-взрывных согласных, спирантизация заднеязычного звонкого взрывного, многочисленные случаи варьирования сонорных по локусу, систематические вокализации поствокальных сонорных и многое другое.

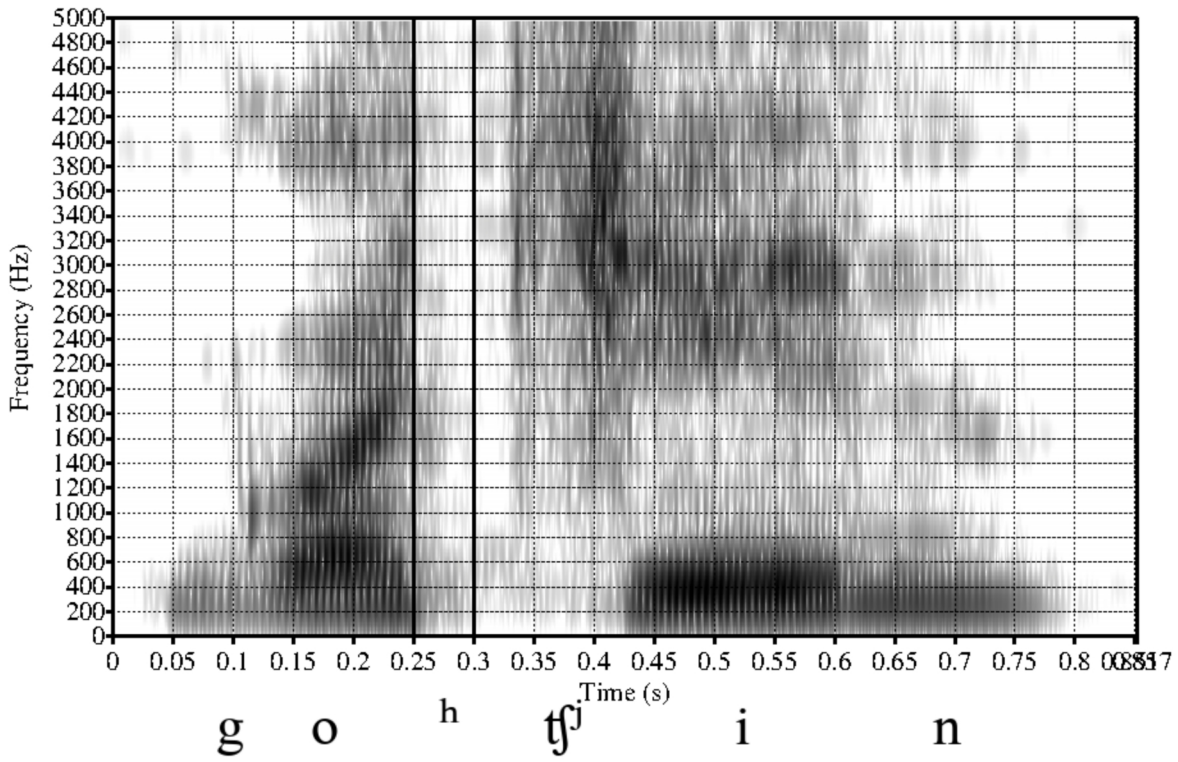


Рис. 8. Поствокальный шум в слове *гочин* (= в следующем году)

Fig. 8. Post-vocal noise in the word *gochin* (= next year)

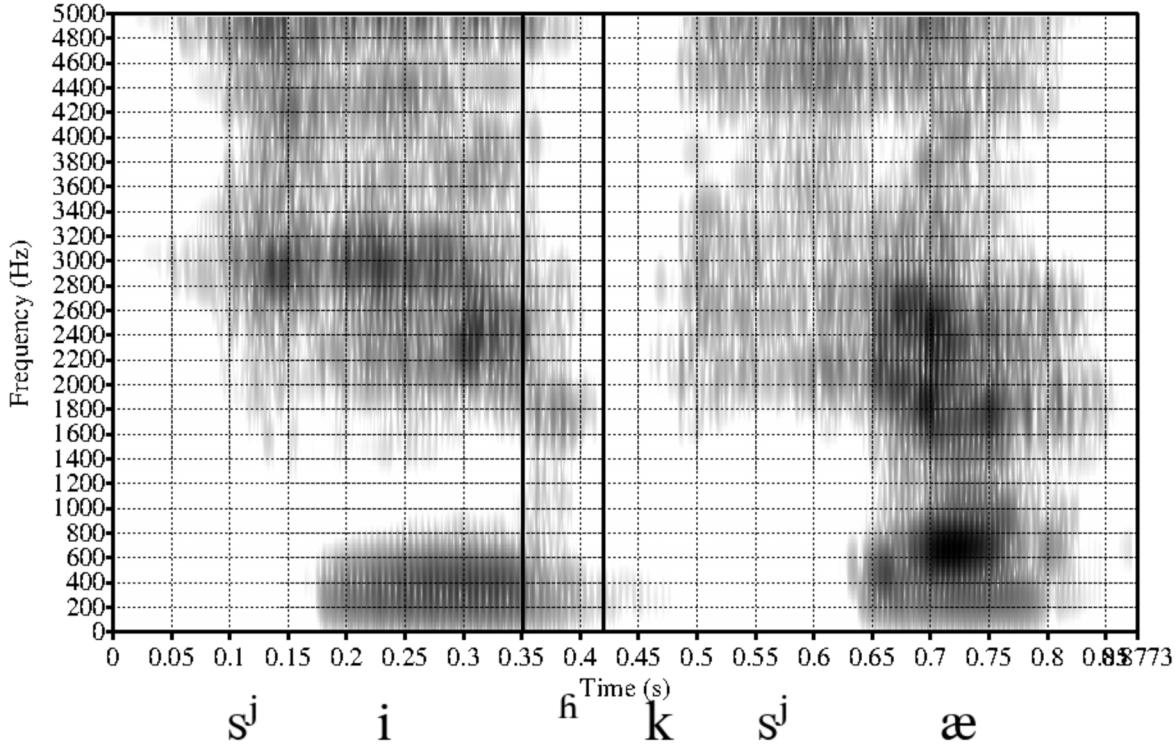


Рис. 9. Поствокальный шум в слове *сиксэ* (= вечером)
Fig. 9. Post vocal noise in the word *sikse* (= in the evening)

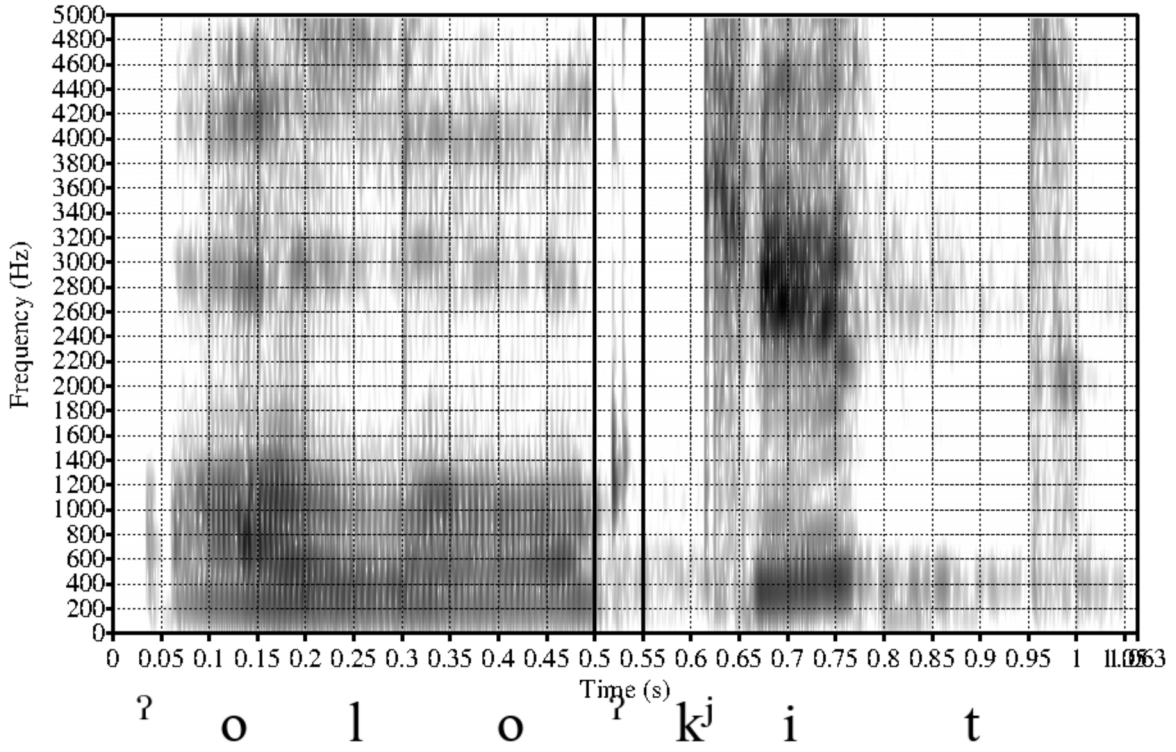


Рис. 10. Поствокальный шум в слове *олоκит* (= брод)

Fig. 10. Post vocal noise in the word *olokit* (= ford)

3. Заключение и перспективы

Предложенные в проекте корпуса по трем говорам восточных эвенков технические и лингвистические решения не претендуют на то, чтобы считаться единственно возможными, а ответы на поставленные непростые вопросы – полными. Однако смеем надеяться, что проделанная работа внесет вклад в ревитализацию эвенкийского языка и стимулирует желание представителей среднего и молодого поколений эвенков активно его использовать.

В настоящее время также ведется активная работа над созданием аналогичного корпуса звучащей орочонской речи. Поставленная задача чрезвычайно актуальна, поскольку данный язык, близкородственный эвенкийскому, также находится на грани исчезновения, будучи, несмотря на все усилия китайского правительства, практически вытесненным доминирующим китайским языком [Мэн Шусянь, 2017; Ян Лихуа, 2016]. На данный момент осуществляются накопление и акустический анализ речевых образцов по аналогии с эвенкийским материалом.

Список литературы

1. Андреева Т. Е. Словесное ударение в эвенкийском языке (экспериментально-фонетическое исследование на материале говоров эвенков Якутии). Новосибирск: Наука, 2001. 151 с.

2. Андросова С. В. Акустический анализ речевого сигнала : учеб.-метод. пособие. Благовещенск : Изд-во Амур. гос. ун-та, 2014.

3. Афанасьева Е. Ф., Раднаева Л. Д. Звуковой корпус данных современного эвенкийского языка // Проблемы изучения и сохранения языков и культур народов России. Материалы секции XXXIX Международной филологической конференции 15–20 марта 2010 г., Санкт-Петербург [отв. ред. Л. Д. Раднаева]. Санкт-Петербург, 2011. С. 3–8.

4. Казакевич О. А. Мультимедийный размеченный корпус текстов на говорах западных эвенков [Электронный ресурс]. Режим доступа : http://languedoc.philol.msu.ru .

5. Казакевич О. А., Клячко Е. Л. Создание мультимедийного аннотированного корпуса текстов как исследовательская процедура [Электронный ресурс] // Труды Международной конференции «Корпусная лингвистика-2013». СПб. : С.-Петербургский гос. ун-т, Филологический фак., 2013. С. 292–300. Режим доступа : corpora.phil.spbu.ru›Works2013/Казакевич.pdf

6. Кравец Т. В. Создание мультимедийного корпуса звучащей речи амурских эвенков: цели, задачи, методы и перспективы // Теоретическая и прикладная лингвистика. 2016. Вып. 2. № 1. С. 41–49.

7. Морозова О. Н., Андросова С. В., Артемчук М. В. Разработка корпуса звучащей эвенкийской речи // Анализ разговорной русской речи (АР^3-2017) : тр. Седьмого междисциплинарного семинара / науч. Ред. Д. А. Кочаров, П. А. Скрелин. СПб, 2017. С. 72–77.

8. Морозова О. Н., Лаврилье А., Болелая А. Н. Некоторые особенности реализации гласных и согласных эвенкийского языка и обозначение на письме // Актуальные проблемы фонетики и методики преподавания иностранных языков. Благовещенск. Амурский гос. ун-т, 2014. С. 84–97.

9. Морозова О. Н. Уровень владения эвенкийским языком в Зейском, Селемджинском и Мазановском районах Амурской области // Обучение иностранному языку на современном этапе студентов высших и средних общеобразовательных учреждений на современном этапе. Материалы Всероссийской 90 научно-методической видеоконференции. Благовещенск: Амурский гос. ун-т, 2014. С. 207–214.

10. Мэн Шусянь. Общее описание орочонского языка в Китае // Теоретическая и прикладная лингвистика. 2017. Вып. 3. № 1. С. 67–86.

11. Процукович Е. А. Уровень владения эвенкийским языком в местах компактного проживания эвенков Амурской области // Теоретическая и прикладная лингвистика. 2015. Вып. 1. № 2. С. 85–93.

12. Речевой корпус эвенкийского языка (аннотированный). 2016–2017. URL : https://linguacorpus.amursu.ru/

13. Скрелин П. А. Сегментация и транскрипция. СПб. : Изд-во С.-Петерб. ун-та, 1999.

14. Ян Лихуа. Исследование современного применения орочонского языка в провинции Хейлунцзян (КНР) // Филологические науки. Вопросы теории и практики. 2016. № 8. Ч. 2. С. 189–194.

S.V.Androsova, O.N.Morozova, M.V.Artemchuk

Amur State University
androsova.sv@amursu.ru , morozova_olga06@mail.ru , artemchuk.mv@amursu.ru

The corpus of oral Evenki speech

Abstract
In the era of globalization with the dominant role of information technologies we are urged to address the issue of language ecology. The Evenki language is the one which is listed among endangered languages. Evenki-Russian bilingualism for the vast majority of the Evenki people in the Amur region has already become Russian-Evenki, and the Evenks who can fluently speak their native language have become scarce. Although IT as a crucial part of civilization benefits have triggered the destructive process, they have a great potential to slow it down or even turn it around through developing on-line free-access oral speech corpora. Endangered language status and the lack of annotated oral speech materials on Eastern Evenki dialects gave us an incentive to start developing such a corpus.
We have developed an information system with the web application written in Ruby on Rails. The system provides free access to the database on Selemdzha, Dzheltulak, and Zea Evenki dialects. The interface is comprised by a number of standard pages: the main page, news, corpus, project information, resources and contacts. In the main page a user can find the information concerning new speech samples added to the database and about the events where the Evenki and Orochon languages are involved. These include the Olympiad in Evenki and Orochon hosted at the Amur State University. The Corpus page provides necessary personal information about the recorded subjects: name, occupation, clan affiliation, spoken dialect. This information is published upon the subjects desire. Some of them even provided photos to be placed on their pages. The corpus is meant for a wide range of users – from native Evenki speakers without any linguistic training to specialists in methods of teaching Evenki and researchers who explore it. We offered a number of technical solutions of certain challenging issues concerning segmentation and transcription peculiarities. One of those was post-vocal turbulent and impulse-like noise in the transition vowel-consonant phase. At present we are also working on developing the Orochon oral speech corpus on the analogy with the Evenki one.

Keywords: endangered Evenki language, user type, annotation, transcription, segmental level, suprasegmental level

References

1. Andreeva T. E. Slovesnoe udarenie v evenkiyskom yazyke (eksperimental'no-foneticheskoe issledovanie na materiale govorov evenkov Yakutii) [Lexical stress in the Evenki language (Experimental phonetic study of the Evenki dialects of Yakutia)]. Novosibirsk: Nauka, 2001.
2. Androsova S. V. Akusticheskiy analiz rechevogo signala [Speech acoustics]: A handbook. Blagoveshchensk: Amur State University Press, 2014.
3. Afanas'eva E. F., Radnaeva L. D. Zvukovoy korpus dannykh sovremennogo evenkiyskogo yazyka [Sound corpus of modern Evenki] // In L. D. Radnaeva (ed.) Proceedings of XXXIX International Philological Conference, Session “Problemy izucheniya i sokhraneniya yazykov i kul'tur narodov Rossii” [Challenges of exploring and preserving languages and culture of the peoples of Russia]. March 15–20, 2010. St-Petersburg, 2011. Pp. 3–8.
4. Kazakevich O. A. Mul'timediynyy razmechennyy korpus tekstov na govorakh zapadnykh evenkov [Multimedia annotated corpus of texts in Western Evenki dialects]. URL : http://languedoc.philol.msu.ru .
5. Kazakevich O. A., Klyachko E. L. Sozdanie mul'timediynogo annotirovannogo korpusa tekstov kak issledovatel'skaya protsedura [Developing multimedia annotated corpus of texts as a research procedure] // Proceedings of the International Conference «Korpusnaya lingvistika-2013» [Corpus Linguistics-2013]. St-Petersburg State University, Philological Faculty, 2013. Pp. 292–300. URL: corpora.phil.spbu.ru›Works2013/Kazakevich.pdf
6. Kravets T. V. Sozdanie mul'timediynogo korpusa zvuchashchey rechi amurskikh evenkov: tseli, zadachi, metody i perspektivy [Creating multimedia corpus of oral speech of the Amur Evenki: Objectives, tasks, methods and perspectives] // Teoreticheskaya i prikladnaya lingvistika [Theoretical and Applied Linguistics]. 2016. Vol. 2 (1). Pp. 41–49.
7. Morozova O. N., Androsova S. V., Artemchuk M. V. Razrabotka korpusa zvuchashchey evenkiyskoy rechi // In D. A. Kocharov, P. A. Skrelin (eds.) Proceedings of the 7th Interdisciplinary Seminar “Analiz razgovornoy russkoy rechi (AR3-2017)” [Russian oral speech analysis (AR3-)]. St-Petersburg State University, 2017. Pp. 72–77.
8. Morozova O. N., Lavril'e A., Bolelaya A. N. Nekotorye osobennosti realizatsii glasnykh i soglasnykh evenkiyskogo yazyka i oboznachenie na pis'me [Some peculiarities of Evenki vowels and consonants realization and the way to write them] // Aktual'nye problemy fonetiki i metodiki prepodavaniya inostrannykh yazykov [Challenging issues of phonetics and foreign languages teaching methods]. Blagoveshchensk, Amur State University, 2014. Pp. 84–97.
9. Morozova O. N. Uroven' vladeniya evenkiyskim yazykom v Zeyskom, Selemdzhinskom i Mazanovskom rayonakh Amurskoy oblasti // Obuchenie inostrannomu yazyku na sovremennom etape studentov vysshikh i srednikh ob-

shcheobrazovatel'nykh uchrezhdeniy na sovremennom etape. Materialy Vserossiyskoy 90 nauchno-metodicheskoy videokonferentsii. Blagoveshchensk: Amurskiy gos. un-t, 2014. S. 207–214.

10. Meng Shuxian. Obshchee opisanie orochonskogo yazyka v Kitae [General description of the Orochon language in China] // Teoreticheskaya i prikladnaya lingvistika [Theoretical and Applied Linguistics]. 2017. Vol. 3 (1). Pp. 67–86.

11. Protsukovich E. A. Uroven' vladeniya evenkiyskim yazykom v mestakh kompaktnogo prozhivaniya evenkov Amurskoy oblasti [The level of the Evenki language skills in the Amur Evenks residences] // Teoreticheskaya i prikladnaya lingvistika [Theoretical and Applied Linguistics]. 2015. Vol. 1 (2). Pp. 85–93.

12. Rechevoy korpus evenkiyskogo yazyka (annotirovannyy). 2016–2017. URL : https://linguacorpus.amursu.ru/

13. Skrelin P. A. Segmentatsia i transkriptsia [Segmentation and transcription]. – St-Petersburg : St-Petersburg University Press, 1999.

14. Yan Likhua (2016). Issledovanie sovremennogo primenenia orochonskogo yazyka v provintsii Heiluntszian (KNR) [Study of the contemporary use of the Oroqen language in the province Heilongjiang (People's Republic of China)]. *Filologicheskie Nauki. Voprosy teorii i praktiki* [Philological Sciences. Issues of Theory and Practice], 8 (62). Part 2, 189–194.