АНАЛИТИЧЕСКАЯ ИНФОРМАЦИОННАЯ СИСТЕМА (АИС) «ЯЗЫКИ РОССИИ: СОЦИОЛИНГВИСТИЧЕСКИЙ ПОРТРЕТ» И ПЕРСПЕКТИВЫ ЕЕ РАЗВИТИЯ[1]

ANALYTICAL INFORMATION SYSTEM «THE LANGUAGES OF RUSSIA: SOCIOLINGUISTIC PORTRAIT» AND THE PERSPECTIVE OF ITS DEVELOPMENT

И. Л. Иткин

Институт проблем информатики РАН, г. Москва

joseph@e-tools.com

О. А. Казакевич

Научно-исследовательский вычислительный центр МГУ, г. Москва

kazak@orc.ru

Н. Г. Колесник

Институт языкознания РАН, г. Москва

socioling@mail.ru

Т. Б. Крючкова

Институт языкознания РАН, г. Москва

socioling@mail.ru

В. Ю. Михальченко

Институт языкознания РАН, г. Москва

socioling@mail.ru

И. В. Самарина

Институт языкознания РАН, г. Москва

ira_samarina@hotmail.com

В основе АИС «Языки России: Социолингвистический портрет» лежит реляционная база данных, содержащая информацию по функционированию 86 языков РФ. В работе освещаются вопросы информационного наполнения, программного обеспечения данной системы обработки социолингвистических данных, ее возможности и перспективы развития.

[1] Работы по созданию аналитической информационной системы обработки социолингвистических данных в Научно-исследовательском центре по национально-языковым отношениям (НИЦ НЯО) ИЯз РАН велись и ведутся под руководством В.Ю.Михальченко при финансовой поддержке РГНФ (проект «База данных "Языки России: Социолингвистический портрет"» № 01-04-12025в – демоверсию см. на сайте http://instling.narod.ru; проект «База данных “Языки России: динамика функционирования”», № 05-04-12485в – рук. В.Ю.Михальченко). Программное обеспечение создано при техническом содействии OrkHard Group (http://www.orkhard.com).

Введение

В 1993 г. между Институтом языкознания РАН и Международным центром исследований по языковому планированию (Centre international de recherche en aménagement linguistique (CIRAL) Университета им. Лаваля (Квебек, Канада) был заключен договор о создании одного из томов международной серии «Письменные языки мира: степень и способы использования», посвященной изучению письменных языков мира с точки зрения их функционирования в различных коммуникативных сферах. Руководил созданием серии известный канадский ученый Грант МакКоннелл. Одними из основных целей данного международного труда являются: 1) получение сопоставимых социолингвистических описаний языков мира и 2) создание социолингвистической базы данных. Очевидно, что необходимым условием для решения этих задач являются унифицированность представления и структурированность социолингвистических данных в описаниях языков. Для выполнения этого условия канадскими учеными была разработана подробная анкета[2], учитывающая все аспекты функционирования языков, а создаваемые для данного проекта описания языков представляют собой корпус ответов на пункты анкеты.

Коллективом НИЦ НЯО ИЯз РАН совместно с исследователями из республик и регионов РФ в рамках этого международного проекта было проведено описание 86 языков России по единой схеме, представляющей собой расширенный и учитывающей особенности функционирования языков в РФ вариант анкеты канадских ученых, (см. Карту описания языков [3, с. IL–L]), результаты которого представлены в труде «Письменные языки мира. Языки Российской Федерации. Социолингвистическая энциклопедия», изданном в двух книгах [2; 3]. В Книге 1 содержатся описания 32 функционально наиболее развитых письменных языков народов РФ, имеющих непрерывную письменную традицию [2]. В Книге 2 описываются 54 языка народов РФ с менее развитыми социальными функциями, среди них младописьменные языки, языки с возобновленной письменной традицией, новописьменные языки, бесписьменные языки [3]. Источниками социолингвистических данных послужили:

- опубликованные работы (особенно публикации последних двух десятилетий);
- полевые социолингвистические обследования отдельных регионов, проводившиеся разработчиками базы;
- данные предоставленные коллегами, занимающимися исследованием соответствующих языков;
- Интернет-публикации, размещаемые на сайтах субъектов Российской Федерации.

Предшественницей АИС «Языки России: Социолингвистический портрет» была База данных «Языки малочисленных народов России (ЯМАЛ)» (проект РГНФ № 96-04-12679), содержавшая информацию о 54 языках коренных малочисленных народов РФ [4]. Англоязычная версия этих материалов, при подготовке которой НИЦ НЯО ИЯз РАН сотрудничал с Токийским университетом, стала частью Всемирной базы данных «Исчезающие языки мира», создаваемой при поддержке Совета Европы и ЮНЕСКО.

В АИС «Языки народов России: социолингвистический портрет», разработанную НИЦ НЯО ИЯз РАН в 2001–2003 гг. (проект РГНФ № 01–04–12025в), вошли все социолингвистические данные, содержащиеся в двух книгах «Письменные языки мира. Языки Российской Федерации» [2; 3], и таким образом, информационное содержание БД «ЯМАЛ» теперь инкорпорировано в новую АИС, которая из известных БД является наиболее полной социолингвистической БД по языкам РФ. Она обладает существенно бóльшими функциональными возможностями по сравнению с БД «ЯМАЛ» и представляет собой инструмент для 1) проведения научных исследований, требующих привлечения большого информационного массива, 2) создания функциональной классификации языков, 3) разработки решений актуальных практических вопросов языковой политики в РФ и т.д.

1. Информационное наполнение АИС

В настоящее время в АИС «Языки России: социолингвистический портрет» представлена информация по 86 языкам народов РФ. Как указывалось выше, основу информационного наполнения АИС «Языки России: Социолингвистический портрет» составили социолингвистические данные, содержащиеся в труде «Письменные языки мира. Языки Российской Федерации. Социолингвистическая энциклопедия» [2; 3].

Важным принципом формирования информационного обеспечения нашей БД является обязательное указание на источник включаемой в базу информации и время, к которому относится данная информация, что, с одной стороны обеспечивает верифицированность данных, а с другой стороны, позволяет размещать в БД однотипную информацию, относящуюся к различным временным срезам.

Информационная структура БД представляет собой дерево, содержащее более 20 основных ветвей:

- названия языка;
- названия этноса;
- статистические и географические данные;
- общие сведения о языке;
- письменность и орфография, стандартизация языка;
- статус языка;

[2] Данная анкета уже была использована для описания языков Индии, Китая, Западной Африки и Западной Европы [1].

- история развития литературы;
- использование языка в религиозной практике
- категории литературы (публикации);
- использование языка в периодической печати;
- использование языка в образовании;
- использование языка в СМИ;
- использование языка в региональном правительстве;
- использование языка в центральном правительстве;
- использование языка в местных администрациях;
- использование языка в суде;
- использование языка в органах законодательной власти;
- использование языка в производственной сфере;
- использование языка в сфере обслуживания и торговли;
- источники информации;
- общие замечания,

Вложенность рубрик для для некоторых ветвей бывает достаточно глубокой. Общее количество пунктов рубрикатора превышает 1000 (см. рис. 3, на котором изображен основной вид *рубрикатора*).

2. Программное обеспечение

При разработке настольной версии программной системы использовалась Visual Studio .NET 2003 и SQL Server 2000 как механизм СУБД. В инсталляцию системы вошел файл в формате Access полученный экспортом из SQL Server.

Опыт эксплуатации созданной ранее в рамках проекта системы доступа к БД «Языки России: социолингвистический портрет» выявил необходимость разработки среды позволяющей не только выполнять запросы, но и анализировать их результаты, сравнивая и упорядочивая данные, в связи с чем на завершающем этапе были поставлены и решены задачи 1) улучшения структуры реляционной базы данных и сценариев переноса базы данных между различными СУБД; 2) создания средств обеспечения целостности данных; 3) разработки профессиональной программной аналитической информационной системы обработки данных.

Основное *окно пользовательского интерфейса* (рис. 1) АИС состоит из панели управления навигацией, рубрикатора и панели данных. «Дружественность» элементов пользовательского интерфейса обеспечена их схожестью с элементами широко распространенных программ (см. ниже).

Панель управления навигацией имеет вид, близкий к панели навигации самого распространенного веб-броузера Microsoft Internet Explorer (см. рис. 2). Она позволяет управлять первичным (основным) критерием запросов (выбор языка / группы языков в выпадающем списке **Язык**), быстро возвращаться к уже просмотренной информации, перемещаться между результатами запросов (**На-**

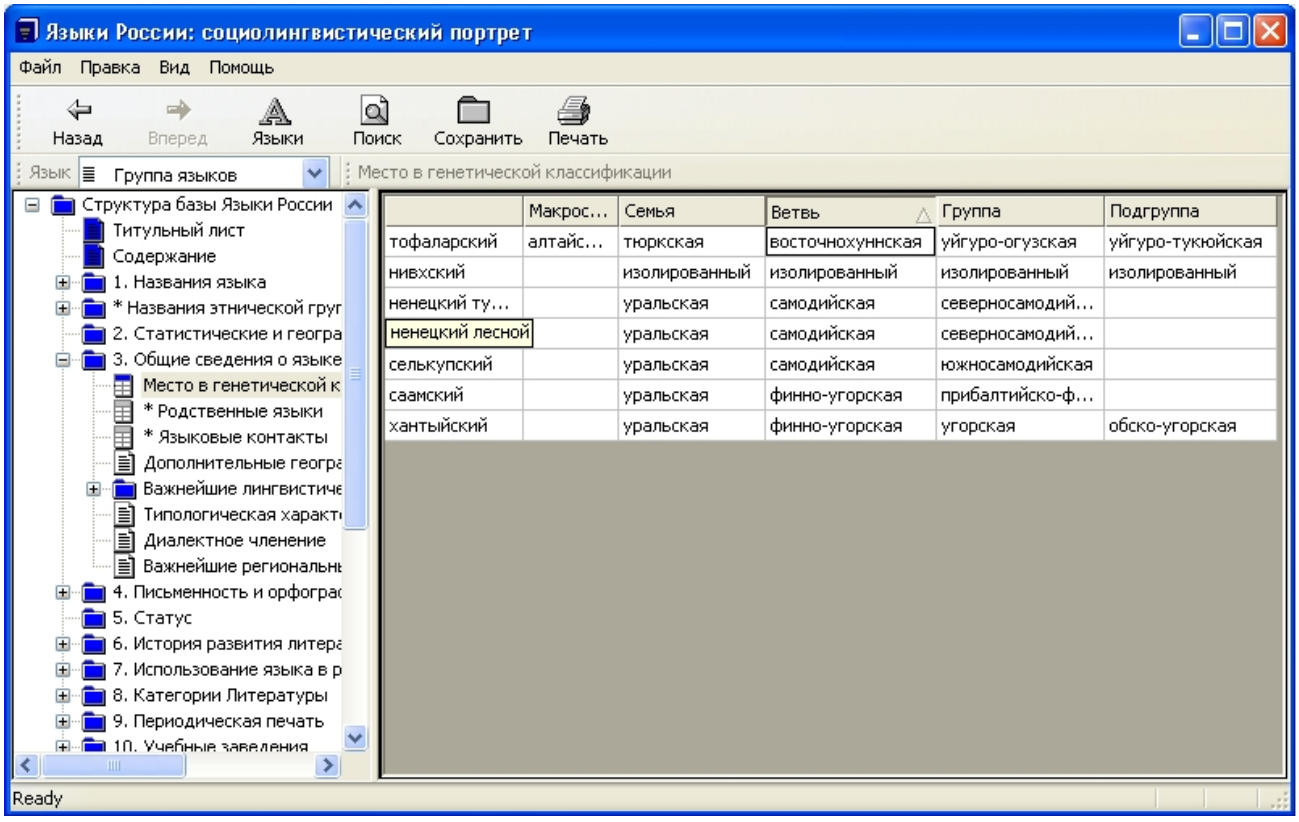

Рис. 1. Основное окно пользовательского интерфейса

зад/Вперед), детализировать основной и дополнительный критерии запросов (**Языки**), осуществлять полнотекстовый поиск по БД (**Поиск**), сохранять результаты выполнения запросов в файл (**Сохранить**) и выводить эти результаты на печать (**Печать**).

Панель рубрикатора (рис. 3) – основной способ выбора запросов – представляет собой элемент управления в виде дерева. Вид рубрикатора напоминает пользователям внешний вид оглавления справочных систем, используемых в Microsoft Windows, и программу просмотра файловой системы Windows Explorer. Пункты панели рубрикатора представляют собой несколько модифицированный вариант Карты описания языков [3, с. XL–L], в строгом соответствии с которой даны описания всех языков в обеих книгах труда «Письменные языки мира: Языки Российской Федерации. Социолингвистическая энциклопедия» [2; 3]. Двойной щелчок по пункту рубрикатора выполняет запрос к БД. Пункт рубрикатора в виде папки содержит подпункты, но самому пункту не соответствует никакой запрос. Пункт в виде белой страницы означает, что результат запроса имеет вид гипертекста (HTML) и зависит от выбранного значения в списке **Язык** на панели навигации. Пункт в виде синей страницы означает, что результат запроса имеет вид гипертекста и информация относится ко всем языкам. Пункт в виде «решетки» (#) означает, что результатом запроса будет таблица.

На *панели данных* (рис. 4) отображаются результаты запросов. В зависимости от выбранного пункта рубрикатора данные принимают вид таблицы или html-документа. Табличное представление имеет вид сходный с таблицами Microsoft Excel и

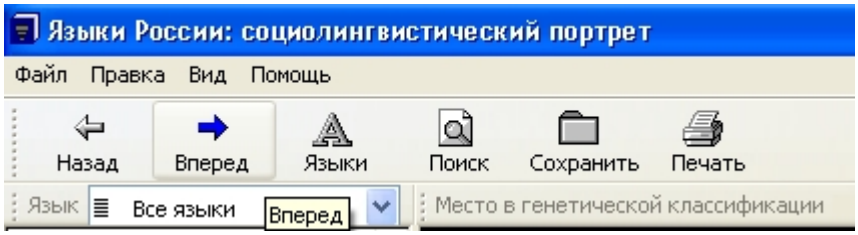

Рис. 2. Панель управления навигацией

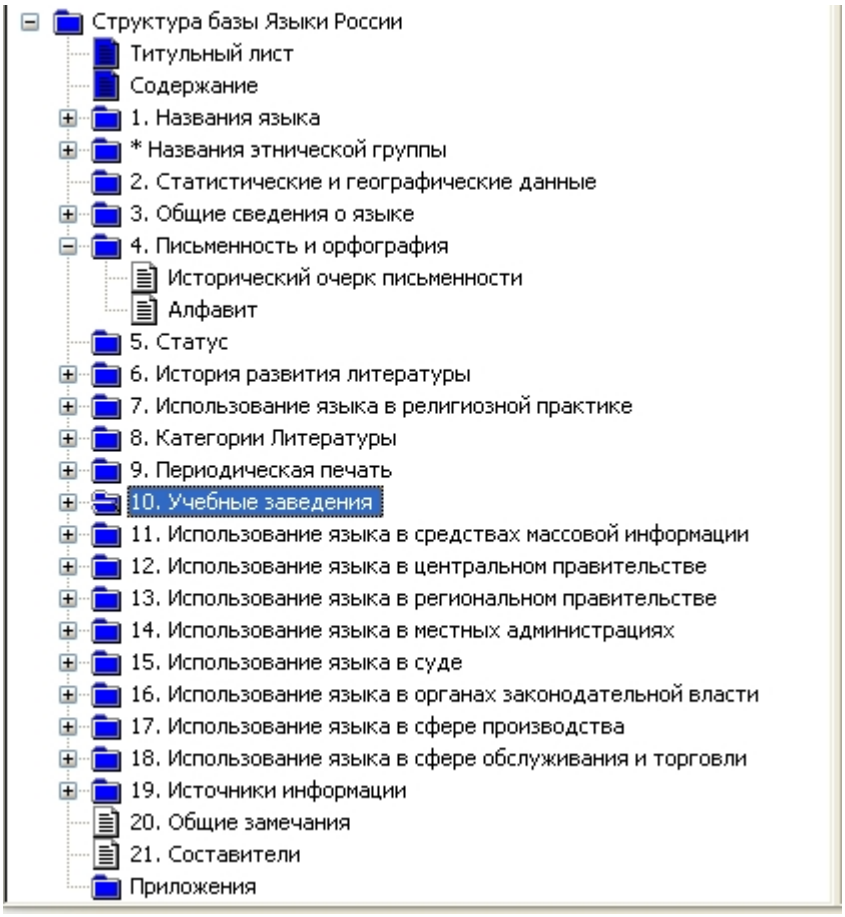

Рис. 3. Рубрикатор

Публикации

	Период	Количество
ненецкий лесной	1994–2002	6
ненецкий тундровый	1870–2002	ок. 150
нивхский	1932–1998	22
нивхский	1932–1937	5
нивхский	1982–1998	17
орокский	до 2002	0
орочский	до 2002	0
рутульский	1991–1998	более 20
саамский	1878–1995	ок. 25
селькупский	1870–2002	34
тиндинский	до 2002	0
тофаларский	до 2002	нет данных
удэгейский	1931–1995	более 10
удэгейский	1931–1936	9
удэгейский	1989–1995	3
ульчский	до 2002	0
хантыйский	1870–2002	ок. 150

Рис. 4. Панель данных

позволяет упорядочивать данные удобным для анализа образом, в том числе сортировать по любой из колонок. Представление в виде html-документа отображает информацию с учетом шрифтов и форматирования, а также позволяет переключаться по ссылкам (например по интернет-адресам организаций).

В целях обеспечение целостности базы данных был осуществлен переход к использованию Microsoft SQL Server 2000 и проделана работа по обеспечению самосогласованности данных, в том числе приведение таблиц реляционной базы данных к нормализованному виду, проверка ссылочной целостности, целостности на уровне доменов и обеспечение правильности агрегативных значений.

АИС «Языки России: социолингвистический портрет» разработана в трех вариантах:

- *демонстрационная версия* (содержит данные по ограниченному набору языков, недоступна часть пунктов рубрикатора, закрыты некоторые функциональные возможности);
- *профессиональная версия* (содержит полный набор данных; в отличие от обычной версии позволяет осуществлять полнотекстовой поиск по всей БД, а также влиять на основной критерий запросов – язык – с помощью дополнительных критериев, таких как этнос, регион, религия, генетическая принадлежность и т.д.);
- *экспертная версия* (в дополнение ко всем возможностям профессиональной версии позволяет динамически менять рубрикатор, создавая новые запросы; доступно редактирование данных, хранящихся в имеющихся таблицах, и создание новых таблиц; в настоящий момент заканчивается разработка этой версии системы).

Системные требования. Для нормальной работы АИС конфигурация компьютера должна соответствовать следующим параметрам: процессор не ниже Pentium-450, объем оперативной памяти не менее 96 Mb, свободное место на диске – не менее 15 Mb для обычной версии и не менее 50 Mb для профессиональной. Кроме того, требуется наличие на компьютере операционной системы семейства Microsoft Windows NT – Windows 2000, Windows XP или Windows 2003 Server, а также Microsoft Access 2000 или Microsoft Access 2002 и Microsoft Internet Explorer версии 5.0 и выше. Экспертная версия программной системы может использовать любой OLE DB провайдер, например Microsoft SQL Server 2000, Oracle и т.д.

3. Перспективы развития АИС «Языки России: социолингвистический портрет»

Коллектив разработчиков АИС «Языки России: социолингвистический портрет» (НИЦ НЯО ИЯз РАН) ставит перед собой задачу модификации созданной системы. Предполагается:

- Пополнить базу данных а) результатами переписи 2002 г.; б) новыми материалами по функционированию языков коренных народов, а также диаспор[3]; в) картографическими и графическими данными; г) данными полевых социолингвистических обследований; д) аудио- и видеоматериалами; е) образцами текстов.
- Произвести изменения в структуре БД в связи с пополнением ее данными нового типа.
- Расширить систему запросов, в том числе разработать пользовательский интерфейс, который позволит пользователю самому формировать новые типы запросов.
- Перевести большую часть компонентов имеющейся программной системы на платформу Microsoft .NET и создать на основе базы данных XML Web-сервиса для доступа к социолингвистической информации,
- Создать отдельный сайт, описывающего доступ к этому сервису и содержащий примеры получения социолингвистических данных в формате XML и преобразования их в содержимое Интернет-страницы.

Заключение

Таким образом, выдвигается идея создания универсальной базы данных «Языки России: динамика функционирования» по функционированию всех языков Российской Федерации, банка социолингвистических данных, содержащего по возможности максимально полную и сопоставимую, ин-

[3] Данные по функционированию языков диаспор в России в АИС «Языки России: социолингвистический портрет» не представлены.

формацию в разных временных срезах. Сопоставимость информации обеспечится единой Картой описания языков, которая будет, соответственно, модифицирована в связи с добавлением в БД новых типов данных.

АИС «Языки России: динамика функционирования» позволит сопоставлять и обрабатывать социолингвистические данные, в том числе 1) сопоставлять языки на основе объема их социальных функций; 2) осуществлять хранение и классификацию количественных параметров социальных функций языков России; 3) прослеживать тенденции изменения языковой ситуации и языковой жизни, а также послужит инструментом для исследования функционирования языков народов России и построения функциональной классификации языков.

Разработка БД будет производиться на основе программного обеспечения компании Microsoft, в частности новой платформы разработки по Microsoft .NET. Предполагается использование продуктов семейства Visual Studio 2005 Express, а непосредственно в проекте будут использованы Visual Web Dev и Visual C#. В качестве сервера базы данных в настоящее время используется MSDE, который со временем может быть заменен на SQL Server Express 2005.

Будущее глобальной информационной сети Интернет лежит в области стандартизованного обмена данными между различными приложениями и сервисами. В настоящее время общепризнанными стандартами организации распределенных межпрограммных коммуникаций стали язык XML и технология Web-сервисов, однако реальные приложения только начинают создаваться. Предполагается что использование технологии ADO.NET для доступа к имеющейся базе данных «Языки России: динамика функционирования» даст возможность создать XML Web-сервис. Этот сервис обеспечит доступ к данным не только для конечных пользователей, но и для настольных и Web-приложений, а также других Web-сервисов, которые появятся в сети Интернет.

Проект может использоваться как конечными пользователями, так и создателями Интернет-ресурсов получающих социолингвистические данные в автоматическом режиме. Веб-сервис, предоставляющий доступ к данным, будет размещен в сети Интернет. Поскольку Web-сервисы основаны на использовании общедоступной сети Интернет и стандартного протокола HTTP, то обеспечение доступа к данным посредством XML Web-сервиса позволит использовать их в различных Интернет-ресурсах. В частности возможно создание программных систем и сайтов, посвященных отдельным аспектам информации хранящейся в БД «Языки России: динамика функционирования».

Список литературы

1. The Written Languages of the World: A Survey of the Degree and Modes of Use. 1988–1998. Vol. I–V).
2. Письменные языки мира. Языки Российской Федерации. Социолингвистическая энциклопедия. Книга 1. М.: Academia, 2000. 656 с.
3. Письменные языки мира. Языки Российской Федерации. Социолингвистическая энциклопедия. Книга 2. М.: Academia, 2003. 848 с.
4. Аношкина Ж.Г., Казакевич О.А. Банк социолингвистических данных «Языки малочисленных народов России (ЯМАЛ)» // Методы социолингвистических исследований. М., 1995. С. 7–26.
5. Михальченко В.Ю. Социолингвистический портрет письменных языков России: методы и принципы исследования. Приложение // Методы социолингвистических исследований. М., 1995.
6. Mikhalchenko V.Yu. Endangered Languages of Russia: an Informational Database // Studies in Endangered Languages. Tokyo, 1998.
7. Языки Российской Федерации и нового зарубежья. Статус и функции. М., 2000. 400с.