Отчет за 2014 год по разделу 4: Новая усовершенствованная версия Базы знаний БД ИЯз РАН "Языки мира" как уникальный фактографический и программный аппарат для типологических, компаративных, квантитативных, квалитативных и прикладных лингвистических исследований

4.1 «Разработка новых методов поиска релевантных признаков языковой семьи в БД ИЯз РАН "Языки мира"»

В.Н. Поляков

В рамках данной подтемы в 2014 г. велась работа по уточнению и совершенствованию контента базы данных и ее программной среды и по формированию новых принципов и методов поиска релевантных признаков языковой семьи с использованием БД ИЯз РАН "Языки мира", которые заключаются в комбинировании генеалогических, географических, типологических и квантитативных данных по грамматическому строю языков.

4.1.1. Исходные данные

База данных «Языки мира» базируется на одноименном энциклопедическом издании [Языки мира 1993-н.в]. Статьи из энциклопедий являются источником сведений для языковых описаний, используемых в БД (т.н. рефератах, использующих бинарный формат описания данных).

БД «Языки Мира» характеризуется следующими параметрами:

1) Количество языков: 315 языков Евразии.

2) Описания языков представлены в формате списков языковых признаков.

3) Общее количество языковых признаков: 3824.

4) Содержит описание всех уровней языка (фонетика, морфология, морфосинтаксис, парадигматика, дейксис, синтаксис) и основные справочные данные о каждом языке.

Windows, рабочая версия (1-ая)

Windows-версия БД (версия 1) представляет собой 32-разрядное приложение, написанное на языке Delphi Pascal (версия 7). В качестве СУБД используется Borland DataBase Engine (BDE). Рабочая среда: Windows 95/98/2000/NT/XP. Объем инсталляционного варианта: 17,4 МБ. Объем программы вместе с БД (версия от 01.12.2006): 18,8 МБ.

Версия 1 базы данных представляет собой полномасштабное приложение, позволяющее вводить, удалять и редактировать список языков, модель реферата, рефераты языков. Кроме того, версия 1 позволяет выгружать и загружать рефераты языков в текстовом формате.

Windows, информационно-справочная версия (2-ая)

Доступна для скачивания: https://cloud.mail.ru/public/eaa329cb8c0b/DBlang/

Версия 2013 г. написана на языке C# с использованием библиотеки ASP.NET и, таким образом, требует для использования установленной Microsoft.NET Framework 2.0 и выше. Имеется возможность загрузки рефератов из текстовых файлов. Однако нет возможности пополнять список языков и список характеристик. Общий объем инсталляционной версии программного обеспечения – около 1 ГБ.

Программа предоставляет более удобный интерфейс для просмотра основных данных базы, включает ссылки на исходную статью о языке в энциклопедии (оцифрованную в pdf). Имеет более мощный поисковый аппарат, чем предыдущая версия.

В нее включены «Глоссарий», который дает расшифровку всех терминов модели описания языка, генетический указатель, географический указатель, содержащий наименование области распространения языка и географические координаты ее центра (по атласу ЮНЕСКО); перевод на английский язык признаков; англоязычное название языка; код языка, соответствующий принятому международному стандарту ISO 639-2 (портал Ethnologue). Программа имеет двуязычный интерфейс (русский, английский).

Следует отметить, что созданный в ходе предшествующей работы программный продукт обладает обширными техническими возможностями, которые удовлетворяют требованиям к работе с базой. Следующий этап развития проекта – создание его новой версии, усовершенствованной не только технически, но и информационно.

4.2 «Разработка методов визуализации материалов БД ИЯз РАН "Языки мира" и результатов проводимых исследований»

Г.А. Черкасова

В рамках данной подтемы был проведен анализ методов визуализации для представления данных, хранимых в Базе данных «Языки мира» в форме схем, графиков, диаграмм и т.п. и разрабатывались новые, более выразительные и действенные изображения для представления статической информации о языках, хранимых в БД ИЯз РАН "Языки мира, описанных следующими характеристиками:

Количество языков (отдельных рефератов) …….345
Количество признаков ……3800
Количество парадигм ……10
Количество оцифрованных томов энциклопедии ……..17
Количество переведенных томов энциклопедии ……....9
Признаковое пространство ……1311 тыс.

Методы визуализации данных

Статистический график - это чертеж, на котором статистические совокупности, характеризуемые определенными показателями, описываются с помощью условных геометрических образов или знаков.

Значение графического метода в анализе и обобщении данных велико. Графическое изображение, прежде всего, позволяет осуществить контроль достоверности статистических показателей, так как представленные на графике они делают более очевидными имеющиеся неточности, связанные либо с наличием ошибок наблюдения, либо с сущностью изучаемого явления. При построении графического изображения должен быть соблюден ряд требований. Прежде всего, графики должны быть достаточно наглядными, так как весь смысл графического изображения как метода анализа в том и состоит, чтобы наглядно изобразить статистические показатели. Кроме того, график должен быть выразительным, доходчивым и понятным. Чтобы все эти требования выполнялись, каждый график должен включать ряд основных элементов: графический образ; поле графика; пространственные ориентиры; масштабные ориентиры; экспликацию графика

Существует множество графических изображений. В основу их классификации может быть положен ряд признаков: а) способ построения графического образа; б) геометрические знаки, изображающие статистические показатели и отношения; в) задачи, решаемые с помощью графического изображения.

По способу построения статистические графики делятся на **диаграммы** и **статистические карты**. Диаграммы - наиболее распространенный способ графических изображений. Диаграммы применяются для наглядного сопоставления в различных аспектах (пространственном, временном и др.) независимых друг от друга величин: территорий, населения и т.д.

При построении точечных диаграмм в качестве графических изображений применяются совокупности точек; при построении линейных - применяются линии. Основной принцип построения всех плоскостных диаграмм сводится к тому, что статистические величины изображаются в виде геометрических фигур и, в свою очередь, подразделяются на столбиковые, полосовые, круговые, квадратные, фигурные.

Статистические карты по графическому образу подразделяются на картограммы и картодиаграммы.

В зависимости от круга решаемых задач выделяют диаграммы сравнения, структурные диаграммы и диаграммы динамики.

На **столбиковых** диаграммах статистические данные изображаются в виде вытянутых по вертикали прямоугольников. Построение столбиковой требует применения вертикальной масштабной шкалы. Основания столбиков размещаются на горизонтальной линии, а высота столбиков устанавливается пропорционально изображаемым величинам. При построении столбиковых диаграмм необходимо выполнять следующие требования:

-шкала, по которой устанавливается высота столбика должна начинать с нуля;

-шкала должна быть непрерывной;

-основания столбиков должны быть равны между собой;

-наряду с разметкой шкалы соответствующими надписями следует снабжать сами столбцы.

Полосовые диаграммы состоят из прямоугольников, расположенных горизонтально. В этом случае масштабная шкала - горизонтальная ось. Принцип их построения тот же, что и в столбиковых.

Секторные диаграммы удобно строить следующим образом: вся величина явления принимается за сто процентов, рассчитываются доли отдельных частей в процентах. Круг разбивается на секторы пропорционально частям изображаемого целого. Таким образом, на 1% приходится 3,6 градуса. Для получения центральных углов секторов, изображающих доли частей целого, необходимо их процентное выражение умножить на 3,6 градуса. Секторные диаграммы позволяют не только разделить целое на части, но и сгруппировать отдельные части, давая как бы комбинированную группировку долей по двум признакам.

Секторные диаграммы выглядят убедительно при существенных различиях сравниваемых структур, а при небольших различиях она может быть недостаточно выразительна. Значительным преимуществом полосовых структурных диаграмм по сравнению с секторными является их большая емкость, возможность отразить на небольшом пространстве большой объем полезной информации.

Линейные диаграммы воспроизводят непрерывность процесса развития в виде непрерывной ломаной линии. Линейные диаграммы удобно использовать: когда целью исследования является изображение общей тенденции и характера развития явления; когда на одном графике необходимо изобразить несколько динамических рядов с целью их сравнения; когда наиболее существенным является сопоставление темпов роста, а не уровней.

Для построения линейных диаграмм используют систему прямоугольных координат. По оси абсцисс откладывается показатели, а по оси ординат – количественные характеристики отображаемых явлений или процессов. На оси ординат может использоваться масштабирование.

4.3 «Терминосистема лингвотипологической БД ИЯз РАН "Языки мира": Проект синонимического терминологического русско-английского указателя» (на сопоставительном материале описаний языков в БД и Энциклопедии “Языки мира”)»

А. К. Зотова

На новом этапе – этапе практической реализации Интернет версии БД – актуальным становится вопрос о справочном аппарате, а именно указателей к ней. Как известно, БД «Языки мира» разрабатывалась специалистами Института языкознания РАН на основе оригинальных статей, написанных авторами для Энциклопедии “Языки мира”. Таким образом, в совокупности оба источника представляют собой результаты исследований нескольких поколений лингвистов. Помимо ценности содержащейся в них фактической информации эти исследования иллюстрируют широкую палитру школ, направлений и традиций описания языков – тюркологии, германистики, славистики и др. Этот тип информации представляет значительный самостоятельный, в том числе и исторический интерес.

Таким образом, разработка и создание дополнительных указателей может расширить и оптимизировать возможности использование БД в исследовательских и справочных целях.

Работа осуществляется одновременно в теоретическом и практическом направлениях. Материалом служат два основных источника - тексты статей о языках в Энциклопедии «Языки мира» и описания языков в виде рефератов БД «Языки мира».

Методика основывается на последовательном сопоставительном анализе всех типов описания языков (полного или частично структурированного (согласно одной из четырех типовым схем) текста статей энциклопедии на естественном языке) с несколькими вариантами модели описания языков в БД (в виде частично и/или полностью формализованного описания согласно модели реферата).

Качественная сравнительная оценка традиционных описаний языковых явлений (с учетом несовпадения принципов классификаций и различия школ - германистики, романистики или славистики) и терминосистемы компьютерной версии, позволит проследить динамику и основные тенденции развития терминообразовательных моделей (унификация с помощью использования латинизированных терминов, калькирование иноязычных, преимущественно английских и немецких и др.).

Результаты теоретического анализа терминологических расхождений между моделями описания явлений в семьях и группах языков могут быть использованы для решения практических задач:

- верификации сведений о языках и выявление адекватности их описания в БД «Языки мира»;
- повышения качества дальнейшей работы по пополнению и редактированию БД «Языки мира» на этапе перехода от текстового описания к формальному представлению данных о языках;
- оптимизации применения БД «Языки мира» в дидактических целях (на этапе интерпретации сведений из формальных описаний), в том числе с использованием иных, чем это предусмотрено типовой моделью реферата, терминов и классификаций.

В окончательном виде полученный материал предполагается использовать в качестве самостоятельного проекта - терминологического указателя в виде словаря синонимов (полных и/или частичных) и общего русско-английского/англо-русского терминологического указателя в качестве приложения к БД «Языки мира».

Основные выводы и результаты работы в 2014 г.:

1. Каждый из элементов сопоставляемой пары *статья Энциклопедии – реферат БД* может рассматриваться под разными углами зрения. По времени создания они могут отстоять друг от друга весьма значительно. Имеют место несовпадения принципов классификаций, традиций описания (различные школы) и используемая терминология.

2. Интерпретация терминологии, характерной для отдельных лингвистических традиций (например, германистики, романистики или славистики и др.), затрудняется тем, что ее расшифровка может содержаться в разных статьях или одной из общих статей тома о семье/группе языков.

3. Наблюдается общая современная тенденция использования латинизированных терминов и калькирования иноязычных (преимущественно английских и немецких).

4. В некоторых случаях встречается неконвенциональная узкоспециальная терминология.

5. Основными особенностями указателя являются:

- то, что он включает в себя не только унифицированную терминологию (принятую в БД), но и термины, традиционно используемые авторами, представляющими различные лингвистические школы;

- то, что он включает в себя определения, характеризующие уникальные языковые явления, свойственные отдельным языкам, в том числе описывающие их на разных этапах развития в диахронии (этот тип информации особенно важен для получения исторических сведений о языках (в том числе о мертвых), и о динамике языковых процессов/явлений);

……..

4.3.1.Анализ терминологии БД:

На предварительном этапе была проведена тщательная экспертиза, как формата представления, так и фрагментов содержания БД. Кроме того, были проанализированы источники неточностей и ошибок, связанные с *адаптацией* описания явлений и признаков, имеющихся в Энциклопедии к требованиям реферата и затем модели реферата в его окончательном виде. Требования унифицированного описания приводят к заданной в модели реферата формальной классификации и искусственному отнесению признаков к той или иной позиции, приводя тем самым к неадекватному представлению информации. В этом случае, уточнение терминологии с помощью ряда возможных синонимов и английских эквивалентов может стать одним из инструментов решения проблемы *артефактов.*

4.3.2. Разработка лексикографической модели указателя:

Формирование словника включает в себя: фильтрациию терминов всего корпуса описания в формате модели реферата: выделение ***фактографической*** информации - собственно терминов/терминов-синонимов, и эквивалентов ***классифицирующей*** информации - элементов формулировок, которые используются в рефератах в виде отдельных строк. Соотношение общего числа строк модели реферата (3821) и строк, относящихся к каждому из этих типов информации, составляет приблизительно 1:10, т.е. 400 и 3400 соответственно.

4.3.3. Анализ преимуществ:

Достоинства: облегчает поиск, так как позволяет сохранить исходную (измененную/падежную и др.) форму термина на русском языке.
Недостатки: избыточность и повторяемость терминов

4.4 «Систематизация контента БД ИЯз РАН "Языки мира"»

В. В. Дьячков

В рамках данного направления исследований в 2014 году была продолжена систематизация контента базы данных, которая была начата в 2013 году.

Систематизация контента БД шла в направлении повышения ее значимости как справочного ресурса для лингвистов-типологов, в котором бы в максимально стандартизированной и приемлемой для специалиста форме были представлены данные о различных языковых семьях. В 2013 году была начата работа по проверке контента существующей версии базы данных, в которой находилось свыше 3800 языковых признаков, описывающие более 300 языков мира, а также исправление ошибок и неточностей, допущенных при заполнении БД.

Конечной целью подобной работы является: 1) исключение из текущего варианта БД информации, которая признана устаревшей или неточно сформулированной; 2) подготовка контента для создания новой версии базы данных. Стоит отметить, что новая версия базы данных планируется к выпуску как интернет-ресурс и должна максимально соответствовать предъявляемым к ней критериям полноты и достоверности представляемой информации.

Главным направлением работы в 2014 году было исправление ошибок и неточностей, которые имеются в старом варианте базы данных. Стоит выделить следующее виды ошибок: 1) терминологические ошибки (употребление устаревших терминов, дублирование терминов, за которыми стоит одинаковое содержание); 2) ошибки родо-видового отнесения терминов к разделам; 3) термины с неочевидным содержанием (например, случаи, когда один и тот же термин служит для обозначения разных явлений или когда значение термина расплывчато). Отдельно также стоят термины, которые могут быть пересмотрены в связи с научными исследованиями (например, при пересмотре семантики конкретного грамматического показателя в конкретном языке). Последняя группа терминов в расчет пока не принимается.

В 2014 году в целом работа по исправлению ошибок и неточностей была завершена. В ходе работы были просмотрены целиком основные разделы базы данных, соответствующие фонологическому и морфологическому уровню языка (признаки БД, относящиеся к синтаксическому уровню, были просмотрены ранее). Среди всех 3800 признаков были выделены в отдельные группы признаки, которые подлежат удалению, объединению с другими признаками или переименованию. Для тех признаков, которые требуют модификации в новом варианте базы данных, был определен характер этой модификации: для тех терминов, которые требуют переименования, были определены замещающие термины. При этом многочисленные спорные случаи, которые не могут быть разрешены волевым решением исполнителя проекта, также фиксировались; характер модификации этих терминов должен быть определен путем консультаций со специалистами по конкретным языковым семьям.

Параллельно с этим также шла работа по терминологической унификации базы данных, которая на данном этапе также представляет собой один из аспектов усовершенствования БД (в базе есть глоссарий).

Приняты также некоторые решения о реструктуризации основного списка признаков. В частности, приняты решения о модификации способа представления фонологической структуры слога (в силу неудобства старого способа представления) и набора

аспектуальных категорий, по которым характеризуется аспектуальная система языка (в силу крайней запутанности и неоднородности старых обозначений граммем).

Результаты работы на данный момент оформлены в виде файлов, содержащих инструкции по модификации каждого признака базы данных. Эти инструкции должны быть реализованы в процессе создания новой версии базы данных, которую планируется сделать качественно новым интернет-ресурсом, в гораздо большей степени ориентированным на потребности исследователей-типологов.