На правах рукописи

ЯРОСЛАВЦЕВА Елена Игоревна

КОМПЬЮТЕРНАЯ БАЗА ДАННЫХ «ЯЗЫКИ МИРА» И ЕЕ ВОЗМОЖНЫЕ ПРИМЕНЕНИЯ

Специальность: 10.02.21 — прикладная лингвистика

А в т о р е ф е р а т
Диссертации на соискание ученой степени доктора филологических наук

Москва-2005

Работа выполнена в Институте языкознания РАН

Официальные оппоненты - доктор филологических наук,
профессор **Марчук Юрий Николаевич**

- доктор технических наук,
профессор **Леонтьева Нина Николаевна**

- доктор филологических наук
Тестелец Яков Георгиевич

Ведущая организация - Пермский политехнический институт

Защита состоится «14» апреля 2005 г.
в 13 00 часов на заседании специализированного совета Д.002.17.01 по защите диссертаций на соискание ученой степени доктора филологических наук при Институте языкознания РАН по адресу 125009, Москва, Большой Кисловский пер., 1/12.

С диссертацией можно ознакомиться в библиотеке Института языкознания РАН.

Автореферат разослан «03» марта 2005 г.

Ученый секретарь диссертационного совета
кандидат филологических наук А.В.Сидельцев

Общая характеристика работы

В Институте языкознания с начала 80-х гг. ведется разработка базы данных (БД), которая включает в себя свернутые, формализованные и определенным образом структурированные описания языков мира. В настоящее время БД «Языки мира», несмотря на свою незавершенность, находится в такой стадии разработки, когда она вполне может уже использоваться как инструмент лингвистического исследования, в чем и состоит ее основное назначение. В связи с этим главной целью данной диссертации является краткое описание устройства этой БД и полезных для лингвистики функций, которые она способна выполнять на данном этапе ее создания.

В период широкого распространения компьютеров, электронной почты, надежных электронных носителей информации и т.п. появляется реальная альтернатива для получения, хранения и распространения научных знаний. С этой целью необходимо решить ряд проблем, связанных с обеспечением эффективного и использования электроники в повседневной практике научной деятельности ученого. В русле этих проблем мной и моими коллегами предлагается комплекс компьютерных программ, способствующих как облегчению кропотливых и трудоемких процессов изучения особенностей различных языков, так и расширению круга пользователей энциклопедии «Языки мира» - компьютерная база данных «Языки мира».

В 2000 году к данному проекту присоединился Московский государственный лингвистический университет, где была создана Лаборатория типологических исследований (зав. лабораторией - А.И.Новиков). В соответствии с договором данная тема в настоящее время разрабатывается как совместная. (Виноградов В.А., Новиков А.И., Ярославцева Е.И. База данных «Языки мира» как инструмент лингвистического исследования. // Вопросы языкознания, 2003, № 3, с. 3-14).

Хочется надеяться, что по своему научному уровню и практическому значению эта работа займет свое место в ряду современных концепций баз данных по языкам мира - Ethnologue, созданный в Summer Institute of linguistics (www. sil. org/ethnologue/maps), базой данных лейпцигских исследователей в Max Planck Institute for Evolutionary Anthropology (www.emeld.org/workshop /2004/bibiko/bibiko-original.html), базой данных М. Драйера (wings.buffalo.edu/ linguistics/dryer/database), см. также проект Autotyp Дж. Николз и Б. Бикеля (www.unileipzig.de/~autotyp), представляющий собой попытку объединения нескольких наиболее автори-

тетных баз. Создаваемая в нашем институте база данных «Языки России: социолингвистический портрет» является более частной по сравнению с базой данных «Языки мира» (www.iling.narod.ru) и имеет ряд сходных с ней областей применения.

Целью данной диссертации является разработка и описание устройства компьютерной базы данных «Языки мира», включающей в себя свернутые, формализованные и определенным образом структурированные описания языков мира и перечень ее возможных применений в лингвистике и других областях знания.

Объектом исследования являются языки мира, а именно языки народов, населяющих сейчас (и населявших ранее) земной шар. Общее число от 2500 до 5000 (точную цифру установить невозможно, потому что различие между отдельными языками и диалектами одного языка условно) (Иванов Вяч. Вс. Языки мира. В кн.: Лингвистический Энциклопедический Словарь. М., 1990).

В качестве **предмета исследования** выступает компьютерная база данных «Языки мира». **Исследуемым материалом** служат статьи энциклопедии «Языки мира», работы языковедов - специалистов по отдельным языковым семьям, статьи Лингвистического энциклопедического словаря.

На защиту выносятся следующие положения:

1) Компьютерная база данных «Языки мира» является электронным аналогом создаваемой в Институте языкознания РАН энциклопедии.

2) Компьютерная база данных «Языки мира» может предоставить широкие возможности для лингвистических исследований (составление различных указателей, автоматизированный перевод базы данных, автоматизированный поиск информации в базе данных, получение формальной типологии языков).

3) Итеративно пополняемая модель реферата является одновременно и моделью всех языков, содержащихся в базе данных, а значит, после введения в базу всех известных науке языков, представляет собой структурную модель грамматики человеческого языка.

4) Система составления поисковых предписаний по поисковым запросам потребителей, дает возможность организовать многоаспектный поиск в базе данных.

5) Разработанный тезаурус грамматических категорий и явлений, составленный по оригинальной схеме словарной статьи, является словарем нового типа.

6) Географический и генетический указатели к базе данных позволяют без обращения к энциклопедическому изданию определять перечни языков, распространенных в пределах некой географической единицы, и находить генетические единицы, подчиняющие данную.

7) Созданная система автоматизированного перевода базы данных на английский (в принципе на любой другой) язык обеспечивает возможность широкого использования базы данных.

8) С помощью базы данных можно выявлять лакуны в описаниях языков и определять импликации языковых категорий и явлений.

9) Составленные программы сопоставления языков, основанные на разных критериях соответствия, позволяют создать так называемую формальную типологию языков - материал для верификации традиционной типологии.

10) Структура банка данных позволяет создать диалоговую вопросно-ответную систему по материалам базы данных.

Научная новизна работы состоит в том, что разработаны принципы и методы компьютерного представления информации о языке - компьютерный вариант энциклопедии «Языки мира».

Впервые создана компьютерная база данных, в которой представлены сведения о 330 языках Евразии.

По предложенной методике в будущем возможно введение в базу данных языков других семей, групп и подгрупп. Составление первичной формы описания какого-либо языка значительно упростится и станет более эффективным при использовании модели реферата, сформированной в базе данных.

Разработаны схемы словарных статей тезауруса грамматических категорий и явлений, географического и генетического указателей по имеющейся базе данных. По этим схемам составлены тезаурус и указатели.

Реализация проекта по созданию многоаспектной информационной системы стала возможной в результате интеграции опыта различных школ и направлений в лингвистике. Была проделана широкомасштабная работа по сбору и анализу конкретных материалов по языкам мира, найдены единые принципы описания языков различного типа - ключ к решению проблемы их сопоставимости. Для программной реализации разработанных алгоритмов был создан универсальный терминологический аппарат описания различных языковых явлений.

Начата и продолжается работа по созданию формальной типологии языков на основе формальных критериев.

Теоретическая значимость работы состоит и том, что разработан новый, нестандартный способ представления множества языковых фактов, относящихся к некоторому конкретному языку, которым является формализованный реферат описания данного языка. Он может быть вычленен из общей структуры базы данных.

Реферат является результатом включения в него из модели реферата тех языковых фактов (строк), которые присущи данному конкретному языку. Поэтому реферат не может состоять из чего-то другого, отличного от того, что содержится в модели. Однако предусмотрено так называемое итеративное пополнение модели, т.е. включение в нее после ряда проверок и консультаций со специалистами тех строк, которых до этого в модели не было. Можно считать, что если реферат является основной единицей ввода, хранения и обработки в базе данных, то модель реферата - это инструмент формирования реферата, обеспечения его стандартности, унифицированности и тем самым формализованности. Поэтому она может рассматриваться как язык внутреннего представления информации в базе данных.

В диссертации вводятся и объясняются этапы работы: написание статьи по типовой схеме в энциклопедию, составление по ней реферата, введение его в компьютер, составление программы по обработке данных и т.п.

Практическая ценность работы состоит в том, что компьютерная база данных «Языки мира» может использоваться в разных областях знания, но в первую очередь в лингвистике. На ее основе уже созданы различные виды указателей к энциклопедии «Языки мира»: географический, генетический, предметный алфавитный и предметный систематический.

Разрабатываются теоретические и методологические основы создания диалоговой вопросно-ответной системы. Указатели дают отсылки от каждой строки модели к статье энциклопедии «Языки мира», где читатель сможет найти более подробную информацию об интересующем его аспекте. Это сокращает затрачиваемое читателем время, и, кроме того, дает каждое явление в контексте (в иерархической структуре), указывает его синонимы и англоязычный эквивалент. В этом плане особый интерес представляет тезаурус грамматических категорий и явлений, составлен-

ный по базе данных и грамматикой - перечень всех встретившихся в базе данных грамматических категорий и явлений.

В работе используются следующие **методы исследования:** различные языки программирования; первоначально комплекс программ для данной базы данных был разработан мной и Ю.П.Скоканом.

Они были написаны на языке Clipper и позволяли осуществлять ввод, хранение, инспекцию, редактирование и преобразование рефератов, а также ввод новых строк в модель реферата. Кроме того, этот программный продукт позволяет осуществлять автоматизированный перевод рефератов на английский язык. Программная реализация этих функций позволяет рассматривать процесс формирования базы данных как процесс ее постоянного расширения как по горизонтали (ввод новых языков), так и по вертикали (ввод новых строк в модель реферата).

В настоящее время первоначальное программное обеспечение было при моем участии перепрограммировано на языке Delphi и адаптировано под Windows. В этой второй версии программного обеспечения была полностью сохранена идеология первой версии. Дополнительно были реализованы функции БД, связанные с информационным поиском.

Также применялись методы и формулы математической статистики, логики и языка исчисления предикатов, лингвистические методы денотативного анализа текста, сравнительно-исторического языкознания, социолингвистики и психолингвистики, метод опроса и анализа потенциальных пользователей базы данных.

Перспективность исследования в части применения базы данных заключается в том, что, кроме введения в нее всех известных на настоящее время языков мира, формируемая и итеративно пополняемая модель языка даст представление о формальной структуре ЯЗЫКА вообще, точнее, о структуре его грамматической составляющей. Представляется также возможность исследования с ее помощью психолингвистических, лингво-палеонтологических и лингво-антропологических закономерностей, структурных особенностей разных грамматических категорий, и многое другое.

Разработаны требования к реферату, к его структуре и единицам, составлены и опробованы программы автоматического реферирования статей энциклопедии (его общей и индивидуальной части). Получены положительные отзывы на рефераты статей от их авторов.

Одним из наиболее перспективных путей представляется изучение грамматики различных языков, сопоставление таких грамматик, созда-

ние исчерпывающего (по возможности) перечня грамматических категорий и явлений, выработанных человечеством в ходе его "лингвистического" развития.

Создаваемая в Институте языкознания компьютерная база данных "Языки мира" основывается на т.н. "Модели реферата" (см. Журинская и др., 1986).

Апробация и внедрение: результаты работы по созданию и ведению базы данных были опубликованы в монографии, ряде статей в отечественных и зарубежных журналах, докладывались на конференциях и семинарах. (Институт языкознания РАН, МГЛУ, Пермский политехнический институт, Уфимский университет).

Используемая терминология:

Реферат - сокращенный вариант статьи энциклопедии «Языки мира», содержащий все основные сведения об описываемом языке и представленный в виде специальной формализованной записи, другими словами, это последовательность наименований языковых явлений, записанных в отдельных строках и связанных между собой определенными отношениями (в основном отношениями подчинения и соподчинения). Каждая позиция (раздел) типовой схемы статьи о языке содержательно соответствует в модели классу языковых явлений.

Модель - эффективное средство стандартизации процесса реферирования. Рефераты статей о языках представляют собой свернутые, формализованные и определенным образом структурированные описания языков, отличающиеся стадартизованностью и унифицированностью. Поэтому референту предоставляется право дополнять список характеристик, задаваемый моделью. В этом случае предусмотрен режим добавления строки в модель. Она сначала создавалась априорно, на основе знаний и опыта лингвистов в разных областях языкознания и при опоре на имеющиеся в энциклопедии описания языков. Эта модель стала эффективным средством стандартизации процесса реферирования.

Итеративное пополнение модели - добавление в модель тех строк, которых не было в ней на момент составления реферата о некотором языке, но которые необходимы для его описания.

Класс, аспект, подаспект, характеристика - структурные единицы реферата разных уровней, соответствующие темам, подтемам, субподтемам и микротемам денотатной структуры (см. А.И.Новиков. Семантика текста и ее формализация. М., 1983).

Тема, это предмет описания (некоторый язык), *подтема* - позиция (раздел) типовой схемы описания языка, *субподтема* - подаспект того аспекта рассмотрения языка, который задан позицией типовой схемы. *Микротема* соответствует конкретному языковому явлению.

Лакуна ~ отсутствие языкового явления или его описания, отмечается в рефератах языков специальными графическими знаками - 0 - явление отсутствует, О - явление не описано.

Ведущие точки - способ отражения иерархии элементов, составляющих модель, специальная формализованная запись. Каждый следующий уровень иерархии имеет на одну "ведущую точку" больше, чем предшествующий.

Вес — или весовая категория - цифровое обозначение значимости некоторого элемента для решения конкретной задачи; чем больше вес, тем значимее элемент.

Грамматикон - универсальный, конкретно-языковой и частные - наборы грамматических категорий и явлений в модели языка, в рефератах конкретных языков и в отдельных классах модели;

Фонематикон, ономастикон, вербатикон, просодикон, нумерикон, птотикон, дейктикон, фонотактикон, фонотипикон, морфотипикон, партикон, парадигматикой, слово-форматикон, дериватикон, сентенсикон, комплексикон, графикон и т.п. - названия частных таксонов грамматикона.

Кластер - «пучок» характеристик языка, группы языков, класса, аспекта модели.

Классифицирующие элементы реферата - элементы, подчиняющие себе другие классы, аспекты, подаспекты.

Фактографические элементы реферата - элементы, стоящие на самых нижних уровнях иерархии, характеристики, редко подаспекты.

Поисковый запрос - интересующий пользователя базы данных вопрос, касающийся ее содержимого.

Поисковое предписание - формальная запись запроса специальными символами.

Дисплей, панель экрана, реперные точки, мемо-поле — термины информатики и программирования.

Банк данных - система программных, языковых, организационных и технических средств, предназначенных для централизованного накопления и коллективного использования данных.

Содержание работы

Работа состоит из предисловия, введения, двух глав, заключения, выводов, списка литературы и трех приложений.

В предисловии излагается история создания и становления прикладной лингвистики как относительно нового направления в языкознании, которая со второй половины 20 века стала иногда называться **«вычислительной лингвистикой», «инженерной лингвистикой»,** или **«автоматической лингвистикой».** Она рассматривала методы решения лингвистических задач с помощью вычислительной техники.

Во введении освещается история создания компьютерной базы данных «Языки мира» - электронного аналога энциклопедии «Языки мира».

У истоков создания энциклопедии и БД «Языки мира» стояла член-корреспондент РАН В.Н.Ярцева - автор и руководитель проекта «Энциклопедия «Языки мира»».

Первоначально в разработке проекта по созданию базы данных принимал участие Борис Владимирович Якушин, но к сожалению, недолго, так как он преждевременно скончался.

В работе по составлению рефератов статей энциклопедии «Языки мира» и введению их в базу данных, а также в составлении модели реферата принимали участие О.И.Романова, Я.Г.Тестелец, А.К.Валентей, М.Е.Алексеев, Н.Б.Бахтин, В.А.Виноградов, А.В.Дыбо, В.П.Калыгин, И.Ш.Козинский, М.С.Полинская, Н.В.Рогова, Н.К.Рябцева, Д.И.Эдельман.

Ю.П.Скоканом при моем участии в качестве консультанта по лингвистическим вопросам был разработан комплекс программ, позволяющий осуществлять ввод, редактирование и корректировку информации. На основе данного программного продукта было введено и отредактировано около 200 описаний языков Евразии как на русском, так и на английском языках усилиями автора.

Были намечены задачи и этапы создания компьютерной базы данных «Языки мира».

Основной единицей энциклопедии «Языки мира» служит статья. Объем статей колеблется и может составлять нескольких печатных листов. Алфавитный принцип организации энциклопедии был выбран как наиболее универсальный. Но он не достаточен, поскольку назначение энциклопедии «Языки мира» не может быть сведено только к задаче обеспечения поиска статей по алфавиту. Энциклопедия должна обеспечить решение самых различных исследовательских задач, что связано с обращением не только к самим статьям, но и к тем сведениям, которые

содержатся в различных местах одной статьи и в разных статьях, находящихся в разных алфавитных зонах энциклопедии. Таким образом, для решения исследовательских задач на базе энциклопедии «Языки мира» нужен многоаспектный поиск, который не обеспечивается алфавитным расположением статей, вследствие чего такой поиск потребитель должен осуществлять сам. Учитывая значительный объем энциклопедии (она состоит из многих томов), следует предположить, что такой поиск будет связан со значительными трудностями, а в некоторых случаях будет и невозможен.

Все это сделало необходимым создание нами специального справочного аппарата, дополняющего основное издание. Традиционной формой такого справочного аппарата служат различного рода указатели, отражающие тот или иной дополнительный аспект поиска и тем самым компенсирующие в некоторой степени недостатки алфавитного принципа организации словаря или энциклопедии. Справочный аппарат может отразить только какие-то отдельные аспекты, а не всё необходимое, тем более что в момент его создания невозможно предугадать все потребности науки, которые могут возникнуть в будущем.

Компьютерная база данных обеспечит:

1) Эффективное обобщение знаний специалистов в области различных языков;

2) Создание электронного варианта энциклопедии «Языки мира»;

3) Удобное и экономное распространение энциклопедических знаний о языках в нашей стране и за границей.

В первой главе «Компьютерная база данных «Языки мира» описывается типовая схема статьи о языках, по которой пишутся статьи энциклопедии, ставятся задачи создания справочного аппарата к ней, выдвигается гипотеза о представлении статьи о языке в виде свернутых, формализованных и определенным образом структурированных описаний - рефератов.

Издание «Языки мира» представляет собой совокупность статей, посвященных описанию языков. Особенностью статей данного издания является то, что все они написаны по заранее созданной типовой схеме, обеспечивающей их единообразие при многоаспектном характере излагаемого в них материала. Это позволяет считать такое описание энциклопедическим, а само издание - энциклопедией.

Энциклопедия «Языки мира» является изданием уникальным не только по полноте охвата языков, но и по глубине теоретического рассмотрения их различных аспектов. Она рассчитана не только на специалистов по сравнительно-историческому и типологическому языкознанию, но и на исследователей в смежных дисциплинах, а также на использование при решении широкого круга исследовательских задач.

Нами была поставлена задача создания новой технологии, для которой необходимо иметь один массив формализованной информации, ориентированной на решение не одной, а нескольких задач. В связи с этим возникает проблема построения массива данных многоцелевого назначения, который можно считать базой лингвистических данных. При одноразовой содержательной обработке и вводе исходной информации база данных должна обеспечивать не только проведение различных видов информационного поиска, но также автоматическое построение различных указателей и решение других более частных задач. Для этого массив, создаваемый в виде базы данных, должен отвечать следующим требованиям:

В него должны входить данные, необходимые и достаточные для обеспечения всех задач, на которые ориентирована информационно-поисковая система. Данные должны быть представлены в виде системы разнопорядковых дискретных единиц, формально выделимых и содержательно значимых для решения поставленных задач.

Структура базы данных должна обеспечивать обращение к ней по многим «входам», для чего она не должна быть слишком жесткой. В процессе эксплуатации база данных должна допускать возможность локального переструктурирования и дополнения без нарушения ее общей схемы и потерь информации.

Центральным из этих требований является вопрос об основных единицах информации.

В рамках данной системы в общем виде под единицей информации следует понимать такие конструкции, которые могут быть вычленены как целостные образования на основе определенных формальных критериев. При этом такое вычленение должно производиться с учетом всех возможных задач, решаемых с помощью автоматизированной системы, т.е. эти единицы должны соответствовать поставленным задачам, быть необходимыми и достаточными для их решения.

Основной дискретной единицей энциклопедии «Языки мира» является отдельная статья, в которой описывается некоторый конкретный язык.

Каждая статья пишется по предварительно заданной типовой схеме, представляющей собой перечень наименований основных разделов или аспектов, которые должны быть обязательно раскрыты автором. В тексте статьи сохраняются цифровые индексы, соответствующие разделам (позициям) типовой схемы.

Языкознание, как и любая другая область науки, характеризуется тем, что, преследуя цель полного и адекватного описания своего объекта - языка, неизбежно распадается на ряд дисциплин, каждая из которых разрабатывает способы описания либо одной из сторон объекта, либо одного из аспектов его функционирования. Но в самом объекте все эти стороны и аспекты находятся в отношении взаимосвязи, взаимопересечения, взаимовлияния. Поэтому история и структура науки о языке демонстрируют постоянное стремление к интеграции достижений ее отдельных областей. Самое значительное проявление этого стремления - оформление в середине XIX века общего языкознания в отдельную дисциплину, определяющую и обосновывающую самые существенные свойства языка. Одной из реальных предпосылок создания общего языкознания была возникшая к тому времени лингвистическая типология как учение о языковых сходствах и различиях, независимых от родства языков. С тех пор общее языкознание и типология развиваются не столько параллельно, сколько в отношении интердепендентности.

Материалом для этих дисциплин должны служить описания языков, причем по возможности единообразные, и параметры этого единообразия должны задаваться лингвистической типологией на основании данных общего языкознания.

Энциклопедическое описание характеризуется следующими требованиями: 1) полнота охвата материала; 2) единообразие описания; 3) Отсутствие полемического аспекта в изложении, т.е. максимально возможная его объективированность.

Соблюдение этих требований позволяет фиксировать состояние языкознания в определенный момент; при этом энциклопедическое описание отнюдь не требует окончательного ответа на все вопросы, встающие перед языковедами; напротив, в нем ценен не только положительный материал, т.е. сведения о языках, но и материал отрицательный, т.е. отсутствие сведений. Энциклопедическое описание представляет необходимый для науки момент определения терминов и понятий, способствующий конкретизации последующих направлений исследования.

Требование единообразия описания повлекло за собой создание типовой схемы статьи, в которой в виде нумерованных позиций представлены те параметры, которые можно считать, во-первых, наличествующими во всех языках или в абсолютном большинстве их, во-вторых, характерными для конкретной языковой специфики, в-третьих, более или менее описанными для значительного числа языков. Наличие этих параметров во всех языках мира и достаточность их для описания «нетривиальных» по строю языков анализировались в процессе обсуждения типовой схемы.

Требование единообразия описания вводит нас в отдельную теоретическую проблему - сопоставимости языков, сравнения имеющихся в них категорий и явлений. Эту задачу можно сформулировать следующим образом: существуют ли реальные основания для единообразного описания всех языков мира независимо от их генетической принадлежности и типологической специфики, для описания, позволяющего сопоставить любой фрагмент системы конкретного языка с фрагментами систем других языков? Если на этот вопрос дается принципиально положительный ответ, то возникает тем не менее еще один: достигло ли языкознание второй половины XX века того уровня, при котором подобное описание осуществимо?

Философия общего языкознания и психолингвистика обосновывают и развивают тезис о единой ментальной основе человеческого языка, об отсутствии непреодолимых преград в общении различных представителей рода *homo sapiens,* о принципиальном единстве процессов вербализации мыслительной деятельности. Что касается типологии, то само ее существование показывает, что неисчерпаемое, на первый взгляд, разнообразие языков мира подчиняется некоторым законам, позволяющим классифицировать как формальные, так и содержательные аспекты языковых систем. Типологизация языковых свойств, начавшись с морфологической классификации языков, в наше время распространяется и на закономерности строения и развития фонологических систем, и на синтаксические и семантические явления, охватывает общие закономерности развития языков и социально-коммуникативные аспекты их функционирования. Эти общие положения вполне закономерно предопределяют соизмеримость языков и тем самым сопоставимость их описаний.

Сопоставимость описаний обеспечивает нам возможность систематизации исходного материала, что способствует извлечению из корпуса статей информации о межъязыковых сходствах и различиях и - в неяв-

ном виде - о наличии внутриязыковых структурных импликаций. Последний аспект требует дальнейших изысканий; здесь энциклопедическое описание может лишь навести специалиста на идею поиска. Впрочем, при достаточно корректной формулировке запроса специалист, оперируя грамматическим указателем, может получить ценный материал для самостоятельного исследования. Информативная ценность единичного описания (= статьи), построенного по принципу **соизмеримости,** возрастает, так как при этом наглядно выступает сопоставимость этих статей; принцип сопоставимости позволяет осуществить **систематизацию** материала на разных уровнях (работа с текстами - указатели - база данных). При этом, чем выше уровень единообразия, тем большие обобщения оно допускает.

Анализ, проведенный с целью определения основной единицы ввода и хранения энциклопедических знаний о языках в памяти компьютера, показал, что такой единицей должен быть сокращенный вариант статьи (ее реферат), содержащий все основные сведения об описываемом языке и представленный в виде специальной формализованной записи. Было установлено, что с формальной точки зрения условной структурной единицей реферата должна быть отдельная строка, соответствующая элементарной записи. В каждой строке записывается конкретный языковой факт, то есть явление, содержащееся в описании данного языка. Этот факт может быть описан одним или несколькими словами. Каждая элементарная запись связана с другими элементарными записями определенными отношениями, для чего применяются также специальные формальные средства.

Совокупность таких элементарных записей, связанных между собой определенными отношениями, и составляет реферат. Таким образом, если реферат можно считать как бы максимальной единицей информации, (он соответствует описанию отдельного языка и является аналогом статьи), то языковой факт - минимальной единицей. Оба типа этих единиц соответствуют критерию формальной выделимости и содержательной значимости.

Очень важной была проблема стандартизации процесса реферирования. Как известно, реферирование представляет собой процесс свертывания, цель которого - минимальным количеством языковых знаков передать максимум информации, содержащейся в первичном документе. Поэтому может существовать несколько семантически адекватных рефератов для одного и того же первичного текста (в нашем случае - статьи энциклопедии «Языки мира»).

По аналогии с типовой схемой статьи в энциклопедии «Языки мира», являющейся некоторым шаблоном при описании языка, было принято решение создать сходный шаблон и для реферата в базе данных. В таком качестве стала выступать т.н. модель реферата. Она сначала создавалась априорно, на основе знаний и опыта лингвистов в разных областях языкознания и при опоре на имеющиеся в энциклопедии описания языков. Эта модель стала эффективным средством стандартизации процесса реферирования. Рефераты статей о языках представляют собой свернутые, формализованные и определенным образом структурированные описания языков, отличающиеся стандартизованностью и унифицированностью. Кроме того, референту предоставляется право дополнять список характеристик, задаваемый моделью. В этом случае предусмотрен режим добавления строки в модель. Новые строки добавляются и в модель реферата. Модель, таким образом, итеративно пополняется, что, на наш взгляд, является наиболее ценным ее качеством.

Реферирование статей энциклопедии "Языки мира" осуществляется специально проинструктированными референтами-лингвистами. Оно осуществляется по тем позициям и в той последовательности, в которой они даны в статье (и, соответственно, в модели), и для каждой позиции состоит в просмотре части модели, соответствующей данной позиции, и в пометке специально предусмотренным образом тех характеристик (а также аспектов и подаспектов), которые содержатся в статье, описывающей данный язык.

Реферирование отличается преобладанием номинативных конструкций, терминологической насыщенностью, привнесением слов обобщающего характера и клишированных оборотов, отсутствующих в тексте первоисточника, укрупнением содержательных единиц, использованием сжатых конструкций. Рефераты строятся с учетом принципов проблемной ориентации, алгоритмизации, простоты и универсальности, разработанными специалистами в области прикладной лингвистики.

Критерием семантической адекватности первичного и вторичного текстов мы будем считать, вслед за Н.М.Нестеровой (Нестерова Н.М. Реферативный перевод как смысловое преобразование текста. Дисс. …канд. филол. наук. М., 1984), факт тождественности их денотатных структур, структуры же эти эксплицируются с помощью денотатного графа, «иерархического построения, в котором можно выделить «главный предмет описания» (тему), «подтемы», «субподтемы» и «микротемы». (Новиков А.И. Семантика текста и способы ее формализации. М., 1983, с.83).

В диссертации приводится список языков, введенных в базу данных (это сейчас 330 языков Евразии) и полностью модель реферата, которая является также универсальным грамматиконом или обобщенной моделью грамматики языка.

Была составлена, отлажена и описана программа автоматического реферирования. В компьютерном варианте энциклопедии "Языки мира" каждый язык представлен наименованием языка и упорядоченной последовательностью параметров, присущих ему. Параметры делятся на лингвистические и индивидуальные.

С формальной точки зрения условной структурной единицей реферата является отдельная строка, соответствующая элементарной записи. В каждой строке записывается отдельный языковой факт, т.е. языковое явление, содержащееся в описании данного языка. Внутри строки эта запись может занимать различные позиции: крайнюю левую, с различным количеством сдвигов вправо, крайнюю правую. Сдвиг содержимого строки является средством выражения отношений между единицами информации внутри реферата. Для большей наглядности вместо сдвига употребляется соответствующее ему количество «ведущих точек», т.е. точек, предшествующих текстовой информации.

Каждая позиция (раздел) типовой схемы статьи о языке содержательно соответствует в модели классу языковых явлений.

В классах выделяются возможные аспекты их рассмотрения. Совокупность аспектов - это дальнейшая градация содержания, осуществляемая на уровне каждого раздела.

Классы и аспекты - это универсалии, априорно задаваемые в модели как наименования явлений, общих для большинства описываемых языков, либо для некоторой группы языков. Они соответствуют подтемам и субподтемам описания некоторого языка. Элементы, находящиеся на более низких уровнях иерархии, - подаспекты и характеристики - соответствуют более конкретным языковым явлениям, специфичным для одного или нескольких описываемых языков, т.е. микротемам статей энциклопедии.

Характеристика - это запись такого языкового факта, который не дробится на более мелкие факты и не имеет в модели подчиненных себе элементов. Группа однородных характеристик, подчиненных одному подаспекту, называется массивом характеристик. В записи цифровой индекс класса и его наименование всегда занимают крайнюю левую позицию. Название аспекта, подчиненного классу, записывается в следую-

щей строке со сдвигом вправо (с одной "ведущей точкой"). Если аспект содержит хотя бы одну характеристику (или подаспект), то их следует записывать с еще большим сдвигом вправо (с двумя и более "ведущими точками").

По своей роли в модели рассмотренные выше структурные элементы делятся на два основных типа: класифицирующие и фактографические. К первому типу относятся наименования классов, аспектов и подаспектов, то есть все такие, которые являются подчиняющими. Среди классифицирующих элементов можно выделить постоянные, а именно, такие, которые задаются в модели обязательно и с необходимой полнотой. Наибольшей степенью заданности, а следовательно, и устойчивости, обладают не только классы, но и аспекты. Поэтому возникает возможность присвоить каждому аспекту определенный код. Для нас кодом аспекта будет его порядковый номер внутри вышестоящего класса, присоединяемый с помощью точки к цифровому индексу этого класса.

Характеристики, в отличие от классифицирующих элементов, не могут быть заданы с исчерпывающей полнотой и точностью.

Процесс реферирования статей энциклопедии оттачивался сначала в ходе ручного реферирования, затем машинного с помощью программ, выполненных на ДВК-2, потом на отечественных компьютерах и в настоящий момент - на более совершенных импортных компьютерах

Реферирование должно производиться с учетом определенных принципов:

1. Принцип проблемной ориентации означает, что компоненты системы должны строиться с учетом тех информационных задач, которые стоят перед системой. В нашем случае такими задачами будут документальный и фактографический поиск информации, автоматический перевод рефератов, автоматическое формирование справочного аппарата энциклопедии «Языки мира», получение типологических обобщений, определение степени близости языков, выявление лакун (=пробелов) в описании языков, создание тезауруса лингвистических терминов.

2. Принцип алгоритмизации определяет не только возможность создания достаточно простых и надежных алгоритмов обработки информации, но и предусматривает требование функционирования системы при минимальном участии человека после ее создания и отладки.

3. Принцип простоты средств лингвистического обеспечения предполагает наличие в языке **и** логике системы только тех средств, которые являются обязательными и эффективными.

4. Принцип универсальности означает возможность использования одного информационного массива для решения различных исследовательских и практических задач.

Реферат, естественно, должен также удовлетворять всем требованиям, предъявляемым к машинным документам: он должен быть по возможности кратким, иметь четкую и единообразную структуру; унификация формы в свою очередь предполагает введение ограничений на использование средств естественного языка и введение дополнительных графических элементов в язык записи информации.

Прежде всего, мы стремились обеспечить семантическую адекватность реферата тексту статьи энциклопедии, т.е. сохранить все основные положения статьи, отразить концепцию ее автора и не «потерять» имеющуюся терминологию (в частности, путем создания словаря синонимов), старались исключить субъективные моменты, влияющие на процесс реферирования. Мы учитывали, что реферат, как и статья энциклопедии, — это потенциальный источник новых, уникальных в типологическом плане сведений, поэтому свобода реферирования не ограничивалась запретами на включение в реферат элементов, отсутствующих в абстрактной схеме.

Проблема синонимии терминов, употребляющихся в статьях энциклопедии, - одна из очень сложных и важных.

Было решено пойти по следующему пути: явные (полные) синонимы задаются априорно, в скобках после "базового" (выбранного в качестве предпочтительного) термина и со знаком "="; среди всех прочих синонимов отдается предпочтение тем, которые имеют греко-латинскую основу. Так, "увулярный" предпочтительнее, чем "язычковый", "фиксированность" предпочтительнее, чем "закрепленность" и т.д.).

Мы собирали отзывы на рефераты, сделанные авторами прореферированных статей, и они были в основном положительны. В них отмечалась информативность рефератов, их лингвистическая корректность и объективность, лаконичность, отражение существенных черт и специфики языков мира.

Мы вполне отдаем себе отчет, во-первых, в субъективности материалов, послуживших основным источником сведений о языках мира (это статьи энциклопедии "Языки мира", написанные хотя и по единой схеме, но разными авторами, находящимися под влиянием как своего родного языка, так и той группы языков, изучению которой посвящена их научная деятельность), во-вторых, в субъектиности референтов, преобра-

зующих данные сведения в рефераты, в-третьих, в субъективности выбранного метода присвоения строкам модели весовых коэффициентов. Именно поэтому проведенные исследования и были названы пилотажными - они призваны служить базой для уточнения проделанной работы и основой для проведения многих других возможных исследований, базирующихся на созданной базе данных.

Раздел 1.5 посвящен освещению проблемы **грамматикона** - (универсального, конкретно-языкового и частных), названных так по аналогии с лексиконом перечнями всех грамматических категорий и явлений. Термин был введен Ю.Н.Карауловым. Эти категории и явления представляют собой строки реферата, совокупность его элементарных записей.

На основании грамматических категорий и явлений, имеющихся в модели и в рефератах, был создан тезаурус грамматических категорий и явлений - он обогащает знания о грамматике, накопленные лингвистами разных школ **и** направлений к настоящему моменту, и является первой приблизительной моделью **«грамматикона».**

Наряду с универсальным набором грамматических явлений - **универсальным грамматиконом** - мы также можем получить для каждого конкретного языка проекцию на него универсального грамматикона, т.е. присущие именно этому языку грамматические категории. Такой грамматикой будем называть **конкретно-языковым грамматиконом.**

Внутри универсального и любого конкретно-языкового грамматикона, в свою очередь, можно выделить его части, составляющие компоненты - **частные грамматиконы** (или таксоны). Таковыми будут следующие: для категории "имя" - **ономастикон,** для категории "число" - **нумерикон,** для категории "падеж" — **птотикон,** для "глагольных категорий" - **вербатикон**, для "дейктических категорий" - дейктикон.

(Все названия частных таксонов условны, за основу взяты греколатинские корни как интернационализмы. Большую помощь в выборе этих названий оказала Н.В.Васильева). Итак, мы имеем грамматикон и его виды - универсальный, конкретно-языковой и частный.

Аналогично можно представить себе универсальное описание таких категорий, некоторые из которых можно считать грамматическими лишь условно, например, такие, как "просодические явления" (условно **просодикон),** "фонетически обусловленные процессы" **(фонотактикон),** "фонологическая структура" **(фонотипикон),** "слог" (силлабикон), "морфологический тип языка" **(морфотипикон),** "части речи" **(партикон),** "парадигмы" парадигматикон), "структура словоформы" (форматикон),

"словообразование" (дериватикон), "простое предложение" (сентенсикон), "сложное предложение" (комплексикон). Можно даже по аналогии сформировать перечень всех встретившихся видов письменности (алфавитов, почерков в их границах и т.п.), и назвать его, скажем графикон. Итак, мы имеем грамматикон и его виды - универсальный, конкретно-языковой и частный.

Универсальный графикон дает перечень всех разновидностей данного алфавита, типов направлений письма (по горизонтали, по вертикали, сверху вниз, снизу вверх и т.п.).

К сожалению, пока нет возможности сформировать универсальный фонематикон, поскольку в модели языка, действующей в рамках нашей компьютерной базы данных, не представилось возможным указать сведения о совместимости/сочетаемости отдельных артикуляционных признаков.

На основе универсального грамматикона был построен тезаурус грамматических категорий и явлений. Как и любой словарь, он будет источником информации (в данном случае, грамматической), а его нестандартная схема словарной статьи, разработанная автором, внесет определенный вклад в лингвистическую терминологию.

Тезаурус (от греч. thesaurus - сокровище, сокровищница) - 1) словарь, в котором максимально полно представлены все слова языка с исчерпывающим перечнем примеров их употребления в текстах; 2) идеографический словарь, в котором показаны семантические отношения (родо-видовые, синонимические и др.) между лексическими единицами. Тезаурус в первом значении в полном объеме осуществим лишь для мертвых языков, ср. "Thesaurus linguae latinae" (с 1900). К этому типу приближается, например, "Словарь польского языка XVI в." (с 1966).

Для живых языков требование исчерпывающей цитации примеров неосуществимо (ср. попытку изменения типа академического "Словаря русского языка" А.А.Шахматова, т.2, 1907, т.4,1916, и Л.В.Щербы, продолживших работу Я.К.Грота).

Структурной основой тезауруса во втором значении обычно служит иерархическая система понятий, обеспечивающая поиск от смыслов к лексическим единицам (т.е. поиск слов, исходя из понятия). Для поиска в обратном направлении (т.е. от слова к понятию) используется алфавитный указатель. Так, например, построен тезаурус П.М.Роже "Roget's thesaurus of English words and phrases" (1852), от названия которого в лексикологическую практику вошло второе значение термина "Тезаурус".

Наш словарь-тезаурус, конечно, относится ко второму типу, он обогащает знания о грамматике, накопленные лингвистами разных школ и направлений к настоящему моменту, и является первой приблизительной моделью «грамматикона». Как написано в статье «Словарь» Лингвистического Энциклопедического Словаря (ЛЭС, с.462), «в них (словарях) отражаются знания, которыми обладает данное общество в определенную эпоху».

Тезаурус строился с использованием всех лингвистических классов модели.

Следует отметить некоторую условность включения всех вышеперечисленных классов модели в тезаурус грамматических категорий и явлений. Мы вполне осознаем такую условность, но считаем ее оправданной, потому что хотим отразить в нашем тезаурусе всю словарную статью (а точнее, ее реферат) полностью. Всё, что касается морфологии и синтаксиса, вполне можно считать грамматикой. Довольно спорно, конечно, отнесение к этому классу всевозможных фонологических и просодических явлений (но это продиктовано уже упоминавшимся стремлением к полноте и всеохватности, а также емкостью и большим деривационным потенциалом предложенного Ю.Н.Карауловым термина **грамматикон).**

Все элементы - входы в тезаурус - расположены в алфавитном порядке. Отметим, что в качестве элементов (входов в тезаурус) мы НЕ брали такие параметры, которые, на наш взгляд, являются малоинформативными для изучения общей структуры языка и для целей определения общности/различия языков.

Для описания каждого элемента (строки модели или ее части в случае неоднословного элемента) мы использовали определенные «зоны» или единицы метаязыка (их названия в тезаурусе даны жирным шрифтом). Все зоны условно делятся на три группы: статусные, конкретизаторы, уточняющие роль признака в структуре тезауруса, и имплицированные - такие уточняющие зоны, которые необходимо помимо конкретизаторов добавить в словарную статью. Как следует из названия, введение этих зон продиктовано характером имеющегося материала. Если указан некий конкретизатор, например, «синкретическое выражение», то логично и даже необходимо указать перечень признаков, выражаемых синхронно с заглавным.

В диссертации перечислены названия зон словарной статьи тезауруса, которые понимаются не совсем традиционно, в отличие от такиx традиционно употребляемых как ***оппозиция, оппозит, синоним, антоним,*** *квазисиноним, квазиантоним.*

В качестве входов (заголовков словарных статей) выступают обычно термины (а точнее говоря - строки), применяющиеся в модели языка (базе данных). Их словоизменительные варианты (слова, отличающиеся от заглавного по роду, числу и пр.) считаются априорно эквивалентными заглавному слову (это относилось и к некоторым изменениям плана выражения, морфологически обусловленным чередованиям, и словообразовательным, и словоизменительным; лишь только в том случае, если такой словоизменительный или словообразовательный вариант имел другую семантическую (в плане репрезентирующего понятия) окраску, он приводился как синоним (или квазисиноним) заглавного слова.

В разделе 1.6 описывается система указателей, составленная по базе данных с помощью специальных алгоритмов и программ.

В данном разделе описывается разработанная нами форма представления всех указателей: географического, генетического, систематического предметного и алфавитного предметного.

Цель географического указателя - дать читателю возможность узнать, какие языки распространены в интересующей его стране или регионе.

Входом в указатель может быть название некоторой географической единицы (страны, области), отсылкой - перечень языков, распространенных в пределах этой географической единицы.

Целесообразно включить в географический указатель информацию социолингвистического характера, содержащуюся в индивидуальной части реферата в классе «Статус языка», а также снабдить каждый мертвый язык пометой (м.), а языки позднейших миграций пометой (я.п.м.). Предполагается, что входы в указатель должны иметь следующий вид: наименование географической единицы и со сдвигом - перечень распространенных в ее пределах языков, а в скобках - наименование тех функций (статусов), в которых выступают описанные в энциклопедии языки в пределах этой географической единицы.

В указателе при таком способе организации информации появляются два иерархических построения: иерархия географических и политико-административных единиц и двухступенчатая иерархия, в которой вышестоящим уровнем является какая-либо географическая единица, а нижестоящим - названия и статусы языков в пределах этой единицы (государственный, разговорно-обиходный, религиозно-культовый и т.д.). Сразу же возник вопрос: каковы должны быть географические единицы, служащие входами указателя? Ведь часто недостаточно иметь информацию о распространении языков по таким крупных единицам, как страны.

Для России, например, перечень распространенных в ней языков очень велик, и весьма желательно иметь более точные сведения о том, какие именно языки распространены в тех или иных республиках, областях, автономных округах и т.п. Кроме того, существуют языки, для которых в статьях энциклопедии в качестве области распространения указаны не единицы современного политико-административного деления, а исторические области (Крым, Кавказ, Средняя Азия и т.п.); особенно часто такие указания встречаются в статьях, описывающих мертвые языки.

Мы пришли к выводу, что географический указатель должен быть построен по смешанному алфавитно-систематическому принципу. Все названия географических единиц, которые встречаются в статьях энциклопедии (кроме, может быть, самых мелких, таких как селения или аулы), должны быть включены в указатель в качестве входов в едином алфавитном порядке. Часть их получит отсылки к другим входам в указатель типа «см.Х». Обычно это отсылки к более крупным географическим единицам (чаще всего к названию страны), приведенному в указателе. Крупные географические единицы должны быть представлены с указанием входящих в них более мелких географических единиц и их иерархии, т.е. фрагментов систематического указателя. Отсылки в виде перечня названий языков могут соответствовать географической единице любого уровня иерархии.

Приведем в качестве примера фрагмент географического указателя:

айвилингмиут гов. < иглулик диал.иннуитов Канады яз.

иглулик диал. < **инуитов Канады**

инуитов Канады < инуитские

инуитские < эскимосско-алеутские

эскимосско-алеутские < палеоазиатские

Полностью географический указатель приводится в приложении к диссертации.

Генетический указатель должен отражать группировку языков мира, т.е. вхождение их в семьи, группы, подгруппы, а также вхождение говоров, наречий, диалектов и диалектных групп в **те** или иные языки.

В каждом реферате приводится цепочка последовательных включений языка в подгруппу, группу и семью языков.

Проанализировав многочисленные варианты генетического указателя, мы решили, что наиболее удачным является сохранение того же основного принципа фиксации генетической информации, который применяется в рефератах, т.е. от меньшей генетической единицы к большей.

Необходимо, чтобы читатель смог найти по генетическому указателю сведения, в какой статье энциклопедии следует искать интересующий его диалект или другую меньшую, чем язык, единицу. Это приводит к необходимости выделения в указателе тех единиц, которые являются заголовками статей энциклопедии, что легко осуществимо с помощью полужирного шрифта.

Генетический указатель отражает информацию о включенности более мелких генетических единиц в более крупные и указание соотношения между названиями различных генетических единиц (в том числе синонимическими и историческими названиями) и заглавиями статей энциклопедии.

Отсылкой для указанных входов служит либо название более крупной генетической единицы, либо синонимичное название, употребляющееся в статьях энциклопедии (для вариантов названий и исторических названий языков и диалектов). Для названия семьи языков, выступающего в качестве входа, отсылка не указывается, поскольку это наиболее крупная генетическая единица. Возможна также отсылка типа "язык-изолят". Вход и отсылка связываются знаком "<" при наличии отношения включения и знаком "=" при наличии синонимического отношения.

Приведем несколько примеров пар типа "вход - отсылка" из генетического указателя:

македонский < южнославянские
южнославянские < славянские
славянские < индоевропейские
хевсурский диал. < грузинский
моркинско-сернурский гов.< луговой марийский диал.
луговой марийский диал. < марийский

Во **второй главе** «Применение базы данных в лингвистических исследованиях» рассматривается: использование базы данных для совершенствования труда педагогов, переводчиков, студентов и лингвистов, автоматизированный перевод базы данных на английский (в принципе на любой другой) язык, а в соответствии с этим выпуск рефератов и указателей на английском (или других) языках, а также другие задачи, которые могут быть решены при помощи того способа представления информации, который применяется в разрабатываемой системе.

Это также автоматизированный поиск информации, создание формальной (основанной на сравнении имеющихся в языках категорий и яв-

лений) типологии, которую можно использовать для верификации традиционной типологии.

Автоматизация перевода базы данных (модели и рефератов) базируется на том, что можно предварительно перевести на другой язык модель (ее первоначальный, исходный) вариант, а впоследствии вручную переводить те элементы, которые были добавлены в этот исходный вариант модели при обработке статей конкретных языков. Как показывает практика, количество вновь вводимых элементов на каждый реферат незначительно (приблизительно 10% от количества уже имеющихся в исходном варианте модели элементов).

Релевантность при поиске означает не что иное, как «соответствие», свойство смысловой близости между текстами и/или их фрагментами (Жданова и др., 1971, с. 152).

Критерием поиска должен служить признак качества сопоставления, т.е. признак, по которому можно отделить релевантные фрагменты базы данных от нерелевантных.

Применительно к энциклопедии таким аппаратом является алфавитное расположение описаний языков (статей), такой поисковый аппарат как бы "встроен" в саму энциклопедию, а потому является ее внутренним справочным аппаратом. Он обеспечивает ответы на те запросы, где объектом поиска является конкретный язык. По такому "входу", как уже отмечалось, можно найти и языковые явления, характеризующие данный язык. Методом последовательного перебора языков можно найти и другие характеристики, некоторые из которых могут оказаться общими для данного множества языков. Но существуют запросы, где абонента интересует не конкретный язык, а некоторые языковые явления, и при этом неизвестно, в каких языках они встречаются. Ответ на такой запрос можно получить путем сплошного просмотра всех статей, что крайне трудоемко, а потому практически невыполнимо.

В разделе 2.5.3 «поисковые запросы и поисковые предписания» рассматривается поиск по информационным запросам потребителей с использованием специальных алгоритмов поиска.

База данных может служить и удобным инструментом для выявления лакун в описании языков, для получения сведений об импликациях различных характериртик.

В целях создания формальной типологии языков мы опирались на то, что принятая форма представления информации в БД позволяет осуществлять построчное сопоставление рефератов между собой и вычис-

лять количественные показатели, характеризующие степень близости языков на структурном (грамматическом) уровне. Специально разработанная для этой цели программа позволяет осуществлять сопоставление каждого языка с каждым и получить количественные результаты такого попарного сопоставления.

Для этого прежде всего предлагается провести сравнение (построчное) языков разных групп, подгрупп и семей на основе приписывания каждой строке модели некоторого веса, что должно отразить значимость данной строки, т.е. данной категории или явления для описываемого языка.

Составлены, отлажены и описаны программы автоматического определения степени формальной близости языков, использующие разные методы.

В качестве примера в диссертации приводятся результаты сравнения трех языков (айнского, арабского классического и бенгальского) со всеми остальными языками, входящими в базу данных. Полученные данные показывают, какие языки более близки друг к другу, а какие формально более далеки друг от друга.

В качестве одного из возможных методов присвоения весов предлагается следующий: приписывание весов в соответствии со степенью универсальности той или иной категории.

При этом мы считаем, что степень универсальности должна иметь различные градации. На наш взгляд, было бы разумно установить четыре уровня универсальности разной степени: универсалии, фреквенталии, раритарии и уникалии. Следует заметить, что термины *универсалии* и *уникалии* употребляются нами не совсем традиционно: *универсалии* не означают наличия категории абсолютно во всех единицах анализируемой группы, а лишь в достаточно большом их количестве, (эти универсалии можно назвать вслед за Дж.Гринбергом, Ч.Осгудом и Дж.Дженкинсом, статистическими), *уникалии* - не в одной такой единице, как это обычно принято считать, а в малом их количестве.

Мы решили проанализировать, в каком количестве языков рассматриваемой семьи, группы и подгруппы встречается та или иная категория, и в зависимости от этого присваивать ей определенный вес. Если категория встретилась в большинстве (от 75 до 100%) языков, то ей будет присвоен вес 1, если в количестве, превышающем половину (от 50 до 75%), то 2, если меньше, чем в половине (от 25 до 50%), то 3, и, наконец, если в малом количестве (от 0 до 25 %), то 4. Нам представляется, что чем

чаще встречается в анализируемой группе некоторая категория, тем меньше должен быть ее вес, т.к. она является «универсалией» для данной группы, а более редкие совпадающие категории свидетельствуют о ее «уникальности», и их совпадение в языках данной группы говорит об их близости.

Программа LangWorld (самый последний вариант программы, работающий с базой данных «Языки мира»), предусматривает поиск строк модели в рефератах, для чего нужно просто ввести искомые строки с логическими отношениями между ними («не», «и», «или») и нажать кнопку «поиск». Через некоторое время на экране появится список языков, в рефератах которых содержатся искомые строки, а также цифра, указывающая на общее количество таких языков. Это очень удобно для определения процента языков, имеющих данную категорию. При этом можно также устанавливать фильтр для поиска только в пределах заданной группы языков.

Категории, получившие вес 1, будем называть универсалиями, вес 2 - фреквенталиями, вес 3 - раритариями, а вес 4 - уникалиями.

Такое присвоение весов в зависимости от степени универсальности категории или явления позволит нам решить две задачи: 1) определение кластеров («пучков» характерных признаков) для крупных единиц языковой систематики (подгрупп, групп и семей) и 2) установление формальной близости языков для верификации традиционной типологии.

Определение кластеров в некотором роде аналогично элементам компаративистики на формальном уровне, так как они выявляют общие для подгрупп, групп и семей языков характеристики. Эти параметры наиболее общего плана - универсалии и фреквенталии - дают представление о квази-универсальном строении единицы, что аналогично описанию семьи или группы языков в каждом томе энциклопедии «Языки мира» и некоторым образом верифицирует такое описание, а менее общие характеристики - о ее специфичности по сравнению с другими аналогичными единицами.

Совпадение специфичных категорий для некоторых языков в рамках исследуемой группы дает возможность говорить об их типологическом сходстве.

Это позволяет решить вопрос о более точном отнесении некоторого языка к определенной семье или группе, если до сих пор такая отнесенность допускала разные (альтернативные) варианты, а также формальное отнесение языка к группе языков-изолятов.

Таким образом, мы построим так называемую формальную типологию, базирующуюся на системной организации языка.

В хорошо организованных (жестко структурированных) системах, каковой и является предложенная нами запись информации о языках в базе данных «Языки мира», изменение одного ее элемента влечет за собой изменения в других точках системы. Различные подсистемы языка развиваются с неодинаковой скоростью. Именно поэтому мы выбрали фонологию, грамматику и синтаксис как основу для описания структуры языка и отказались от описания лексики и грамматики как плохо формализуемых или с трудом поддающихся формализации, стандартизации и унификации систем языка.

Когда мы сравнивали рефераты языков с проставленными для каждой строки весами по уже имеющейся программе сравнения, то в результате получили искомую численную величину степени близости языков.

Наиболее важным из проделанного анализа является вывод о том, что во многих случаях автоматизированный поиск рационально проводить не по массиву рефератов, составляющих базу данных, а по одному из указателей, содержащихся в памяти компьютера, и которые можно рассматривать как самостоятельные поисковые массивы. Существуют такие запросы, ответы на которые требуют комбинированного поиска, т.е. поиска не только по рефератам или отдельному указателю, а по обоим видам поисковых массивов.

Как показывает проведенный статистический анализ, из общего количества собранных мной в 1985 г. многоаспектных запросов потребителей требуют поиска по одному из указателей 28, комбинированного поиска - 17, что в сумме составляет 43% или почти половину анализируемых запросов. Остальные запросы требуют отдельного рассмотрения.

Все это является веским основанием для того, чтобы считать, что, для осуществления эффективного автоматизированного поиска, база данных, представленная в виде множества рефератов, должна быть определенным образом адаптирована. Такая адаптация может быть осуществлена за счет включения в базу данных указателей, полученных на предшествующем этапе работы системы.

Систематический указатель становится в системе основным массивом, а все остальные выступают в качестве вспомогательных компонентов. Систематический указатель синтезирует в себе одновременно и модель реферата, и массив рефератов. Реальные рефераты находят в нем

именований языков, поставленных в соответствие языковым явлениям, содержащимся в его левой части.

Модель реферата в определенной степени дублирует в базе данных массив реальных рефератов, а поэтому в принципе может являться его заместителем. Если это так, то можно предположить, что наличие реальных рефератов в базе данных не является обязательным. Развивая это предположение, можно допустить, что в том случае, если для удовлетворения определенного вида запросов в качестве ответа на них потребуются не фактографические данные, а сам реферат или какие-либо его фрагменты, то такой реферат может быть выведен из информационной модели. Это может оказаться вполне возможным при наличии специального алгоритма и программ, являющихся «обратными» по отношению к программам составления указателя.

Реализация такой программы позволила бы исключить если не полностью, то хотя бы частично массив рефератов из базы данных, что способствовало бы значительной экономии машинной памяти и времени обработки информации.

Если же провести сравнение частных грамматиконов и их фрагментов, в которых, по нашему мнению, отражены особенности восприятия реальной действительности разными этносами, что определяется условиями его жизнедеятельности (к таким фрагментам частных грамматиконов можно, например, отнести конкретные наборы падежей (птотикон), набор местоименных форм и средств пространственной ориентации (дейктикон), такие глагольные категории, как набор различных форм наклонений, категорий модальности, способа действия (вербатикон) и некоторые другие, то подобные исследования могут помочь в описании картины мира у разных этносов.

Раздел 2.6. посвящен банку данных. Термин "банк данных" появился в середине 60-х годов и первоначально применялся для обозначения совокупности взаимосвязанных массивов, находящихся под общим управлением.

В дальнейшем для обозначения такого рода массивов информации чаще стал применяться термин "база данных".

Одновременно изменились и требования к этим массивам. Основным из них, очевидно, можно считать требование, заключающееся в том, что в базе данных элементы должны быть связаны между собой, причем преимущественно не на основе каких-либо внешних формальных при-

знаков, а на основе отношений, отражающих закономерности функционирования моделируемых объектов.

В настоящее время под базой данных понимают "именованную совокупность данных, отображающую состояние объектов и их отношений в рассматриваемой предметной области." (Кокарева Л.В., Милошилин И.И. Проектирование банков данных. М, 1984, с.8). Понятие "банк данных", применяемое в настоящее время, является более широким, чем "база данных". Оно определяется как система "программных, языковых, организационных и технических средств, предназначенных для централизованного накопления и коллективного использования данных". (Там же).

Следует подчеркнуть, что банк данных представляет собой систему, включающую в себя все необходимые средства для решения стоящих перед ним задач. Основным компонентом является база данных, в которой сосредоточена вся фактографическая информация, ее свойством является централизованный принцип формирования. База данных может постоянно изменяться, дополняться, подвергаться декомпозиции без изменения общих принципов ее организации. Другим не менее важным свойством базы данных является то, что она рассчитана на коллективное использование содержащейся в ней информации. В этой связи база данных должна быть максимально приближена к потребителю, который должен иметь возможность непосредственного доступа к ней. Такая возможность наиболее полно реализуется в так называемом диалоговом режиме, что и определяет ее особенности как поискового массива.

Для реализации этого свойства в качестве одного из основных компонентов они должны содержать так называемую базу знаний, которая может быть включена в базу данных или выделяется в самостоятельный информационный массив. База знаний содержит в себе следующие основные компоненты:

1) Сведения, которые отражают закономерности, существующие в предметной области, и позволяют, как выводить новые факты, имеющие место в данном состоянии проблемной среды, но не зафиксированные в базе данных, так и прогнозировать потенциально возможные состояния;

2) Сведения о структуре и содержании базы данных;

3) Сведения, обеспечивающие понимание входного языка, т.е. перевод исходных вопросов и утверждений на внутренний язык.

Основной функцией базы знаний является обеспечение эффективного управления базой данных.

Важным компонентом базы данных является так называемое лингвистическое обеспечение, предназначенное для перевода запросов с входного естественного языка на язык внутреннего представления информации в банке данных. Лингвистическое обеспечение может как входить в состав базы знаний, так и быть самостоятельным массивом. Необходимым компонентом банка данных является комплекс программ, обеспечивающих внутримашинное формирование и ведение базы данных, а также обращение к ней. Совокупность указанных компонентов в их связи и взаимодействии составляет общую структуру банка данных.

Существует необходимость в проведении анализа запросов, требующих автоматизированного поиска с точки зрения определения того, какие требуются преобразования и дополнения базы данных для адаптации ее к задаче автоматизированного многоаспектного поиска информации.

С этой целью специалистам, представляющим различные разделы языкознания, после кратких пояснений о составе энциклопедии и форме представления в ней информации было предложено сформулировать запросы, являющиеся актуальными для них с точки зрения решаемых ими проблем. В результате мной было собрано 200 запросов, характеризующихся большим многообразием, в частности, с точки зрения поиска необходимой информации. В этом плане наиболее простыми являются запросы, в которых необходимо установить наличие или отсутствие какой-либо определенной характеристики в описании конкретного языка.

Все собранные запросы были проанализированы с целью их классификации, а также определения того, в какой степени они могут быть удовлетворены при обращении непосредственно к энциклопедии и какой дополнительный справочный аппарат для этого требуется. В основу методики анализа были положены следующие положения.

Априорно можно считать, что не существует никаких принципиальных причин, чтобы ответы на все анализируемые запросы нельзя было получить только на основе описания языков, содержащихся в энциклопедии. Но ответы на некоторые запросы могут быть получены только в результате аналитико-синтетической обработки информации, локализованной в разных фрагментах энциклопедии. Такая обработка связана с большими трудностями. Во многих случаях ее, очевидно, нельзя считать информационным поиском, поскольку она скорее представляет собой своего рода самостоятельное исследование.

Информационный поиск - это такой процесс, выполнение которого в определенной степени формализовано. Такая формализация может быть

осуществлена только в том случае, если имеется какой-либо специальный аппарат для этой цели.

Были выделены группы запросов. Для адекватного ответа на запросы будем моделировать процесс многоаспектного поиска, взяв за основу существующую базу данных, с одной стороны, и запросы рассматриваемого вида - с другой.

Всё многообразие видов информационного поиска может быть расклассифицировано по нескольким основаниям. Такими основаниями являются: тип запроса, тип поискового массива, являющийся наиболее оптимальным для удовлетворения конкретного запроса, критерий выдачи ответа, тип единиц, составляющих ответ, тип поисковых процедур, применяемых для получения ответа и т.д.

Каждый из данных видов поиска в свою очередь может быть разбит на подвиды по другим основаниям.

Такая классификация видов поиска позволяет осуществить дифференцированный подход к различным видам запросов, применяющих наиболее оптимальную стратегию поиска.

Наиболее важным является вывод о том, что во многих случаях автоматизированный поиск рационально проводить не по массиву рефератов, составляющих базу данных, а по одному из указателей, содержащихся в памяти компьютера, и которые можно рассматривать как самостоятельные поисковые массивы.

Существуют такие запросы, ответы на которые требуют комбинированного поиска, т.е. поиска не только по рефератам или отдельному указателю, а и по обоим видам поисковых массивов.

Таким образом, на этапе работы системы в режиме информационного обслуживания специалистов исходное состояние базы данных преобразуется в интегрированную совокупность нескольких определенным образом связанных между собой поисковых массивов, главным из которых является систематический указатель. При этом левая часть систематического указателя уже не является аналогом модели реферата как средства стандартизации процесса реферирования, а представляет собой полную модель энциклопедического описания. Такую модель можно назвать информационной.

Первым этапом содержательной обработки запроса является составление поискового предписания. Оно должно иметь структуру, которая позволяла бы автоматически распознавать тип запроса и осуществлять

выбор соответствующего алгоритма поиска информации, ее обработки и формирования ответов.

Критерии релевантности запроса и выдаваемых в ответ на него сведений в развитых информационно-поисковых системах зависят от способа эксплицитной фиксации смысла искомого документа.

Мы выделяем следующие виды запросов: простые, сложные и распространенные.

Простые запросы сформулированы в тех терминах, которые совпадают с терминами представления информации в базе данных; ответом на них обычно служит перечень языков, в которых содержится то или иное языковое явление.

Сложные и распространенные запросы записываются в виде булевой формулы.

Поисковые предписания составляются на специальном формализованном языке, единицами которого служат наименования данных, встречающихся в запросах в форме слов и словосочетаний. Такие единицы делятся на несколько семантических классов: 1 - языки (от language), k - языковое явление (от kind), g - генетические единицы (от genttic), p - географические названия (от place). Кроме того, выделяются виды данных, которые служат для уточнения запроса, определения его цели. Это различного рода качественные параметры (f- formants), количественные параметры (r - results), а также вид конкретных операций, необходимых для получения ответа (n - number). По своей функции обозначения вида данных 1, k, g, p являются указателями роли, а f, r, n - определителями.

Приведем примеры запросов и их символическое представление - поисковое предписание.

Простые запросы:

В каких языках имеется категория "род"?

$z_1=k/l:n$

В каких языках отсутствует залог?

$z_2=k/l:n$

Какие залоги есть в грузинском языке?

$z_3=l_0/k:n$

Какие языки входят в уральские?

$z_4=g/l:n$

Какие языки распространены в Италии?

$z_5=p_0/l:n$

В каких языках число падежей меньше 5?

$z_6=k/l:r<5$

Распространенные запросы:

Какие признаки гласных есть в романских языках?

z_7=(l<g)/k:n

Сколько существует кавказских языков?

z_8= 1(l<g)/k:n)

Чем отличаются тюркские языки от монгольских?

z_9=(l<g)/:r

Сложные запросы:

В каких языках есть изафет?

z_{10} = k/l:n →g:n

Где распространены языки с иероглифической письменностью?

z_{11}=k/l:n → l/p:n

К каким семьям принадлежат языки Ближнего Востока; СССР?

z_{12}= p/l:n →l/g:n

Дальнейший сбор потенциально возможных запросов, составление соответствующих поисковых предписаний и разработка программы автоматического получения ответов на вопросы, набираемые на клавиатуре компьютера, сможет в будущем позволить создать такой диалог с компьютером, который будет незаменимым подспорьем для лингвистов и всех, кого интересует энциклопедия «Языки мира».

Информационный поиск в отличие от поиска по поисковым предписаниям является более простой процедурой, при которой происходит непосредственное обращение к имеющимся в базе данных массивам информации Поиск по запросам пользователей и составление по ним поисковых предписаний - более сложен и до конца не отлажен как в плане составления соответствующих алгоритмов, так и в плане написания и отладки программ. Сбор потенциальных запросов и составление по ним поисковых предписаний - дело будущего.

В выводах отмечается, что поставленные в диссертации цели достигнуты.

1) Доказано, что созданная в отделе прикладной лингвистики Института языкознания РАН компьютерная база данных «Языки мира» является электронным аналогом издания «Языки мира», его детализацией и конкретизацией типовой схемы статьи о языке, применяемой в этом издании.

2) В работе описано, как можно использовать базу данных для педагогов, переводчиков, студентов и лингвистов; предложен метод автоматизированного перевода базы данных на английский (уже осуществле-

но), а в принципе на любой другой иностранный язык; предложено несколько вариантов сопоставления языков по формальным критериям.

3) Созданная общая схема универсального реферата, называемая «моделью реферата», выступает в роли шаблона при составлении рефератов.

4) Имеющиеся программы сопоставления языков, основанные на разных критериях соответствия, позволяют создать так называемую формальную типологию языков - материал для верификации традиционной типологии.

5) Составлен тезаурус грамматических категорий и явлений по 330 языкам базы данных. Он написан по принципиально новой схеме словарной статьи, содержащей 3 группы элементов: статусные, конкретизирующие и имплицированные.

6) Составлены географический и генетический указатели к базе данных. Географический указатель позволяет без обращения к энциклопедии определять, какие языки распространены в какой-либо географической единице, и в каком статусе она там употребляется. Генетический указатель дает отсылки от любых генетических единиц к подчиняющим ее более крупным единицам.

7) Разработана система автоматизированного перевода базы данных на английский (в принципе на любой другой) язык.

8) Модель языка и составленные по ней рефераты дают возможность выявить лакуны в описании языков и определить существующие импликации языковых категорий и явлений.

9) Разработанная система составления поисковых предписаний по поисковым запросам потребителей позволяет организовать многоаспектный поиск в базе данных.

10) Описана структура банка данных, обеспечивающая возможность создания диалоговой вопросно-ответной системы по материалам базы данных.

В приложениях представлены следующие материалы: Приложение 1 - Генетический указатель. Приложение 2 - Географический указатель. Приложение 3 - Фрагмент тезауруса, касающаяся видо-временных категорий (часть вербатикона).

Содержание диссертации отражено в следующих публикациях:

Ярославцева Е.И. Методы определения семантической близости текстов // Семантика языковых единиц и текста (лингв. и психолингв. исследования). Сборник статей ИЯ АН СССР. М., 1979. 0,5 а.л. (в соавторстве с Б.В.Якушиным). Мой объем - 0,25 а.л.

Ярославцева Е.И. Критерий близости текстов по содержанию (ситуативный критерий) // Известия АНСССР, Сер. лит. и языка, т. 39, № 6, М., 1980. 0,2 а.л. (в соавторстве с Б.В.Якушиным). Мой объем - 0,1 а.л.

Ярославцева Е.И. Семантические сферы информационного анализа и поиска // Некоторые вопросы анализа поиска текста и терминосистем (деп.) // ИНИОН АН СССР, № 6252. М., 1980.0,5 ал.

Ярославцева Е.И. Критерий близости текстов по содержанию // Материалы семинара «Статистическая оптимизация преподавания языков и инж. лингвистика». Чимкент, 1980. 0,2 ал. (в соавторстве с Б.В.Якушиным). Мой объем - 0,1 ал.

Ярославцева Е.И. Исследование смысловой близости текстов // Дисс. ... канд. филол. наук. М., 1981.

Ярославцева Е.И. Реконструкция умственных ситуаций как условие установления релевантности текстов // Колл. монография «Лингвистические вопросы сообщений». М., Наука, 1983.0,5 ал.

Ярославцева Е.И. Проблемы семантического означивания при автоматическом установлении содержательной близости текстов // Вопросы семантики в процессах коммуникации. Ульяновск, 1981.0,5 ал.

Ярославцева Е.И. Экспликация смыслового содержания текста и установление релевантности текстов при АПТ // Всес. конф. «Переработка текста методами инженерной лингвистики». Минск, 1982.0,1 ал.

Ярославцева Е.И. Критерий адекватности перевода для текстов различных типов. 0,6 а. л. (в соавторстве с А.И.Новиковым). Мой объем - 0,3 ал.

Ярославцева Е.И. Критерий близости текстов и его экспериментальная проверка // Конф. «Семантика и синтаксис в языках народов СССР, народов мира и прикладных информационных системах». М., 1985. 0,2 ал.

Ярославцева Е.И. Влияние лингвистических параметров текста на эффективность речевого воздействия // VIII Всес. симпозиум по психолингвистике и теории коммуникации. М., 1985. 0,2 ал. (в соавторстве с А.В.Михеевым). Мой объем - 0,1 ал.

Ярославцева Е.И. Речевое воздействие и лингвистическое оформление текстов // Респ. научн.-техн. конф. «Психолого-педагогические и лингвистические проблемы исследования текста». Пермь, 1984.0,2 ал.

Ярославцева Е.И. Лингвистическое оформление пропагандистского текста // Сборник «Семантика текста и проблемы перевода». М., Ин-т языкознания, 1984.0,5 ал.

Ярославцева Е.И. База лингвотипологических данных и принципы ее функционирования // Вестник АН СССР, 1985, М, №2 0,8 ал. (в соавторстве с А.И.Новиковым). Мой объем - 0,4 ал.

Ярославцева Е.И. Энциклопедическое описание языков мира (Теоретические и прикладные аспекты). М., 1986. 10 ал. (в соавторстве с М.А.Журинской и А.И Новиковым). Мой объем - 5 ал.

Ярославцева Е.И. Принципы автоматической обработки информации о языках мира // Матер. III Всес. конф. по теор. вопросам языкознания. М., 1984. 0,8 ал. (в соавторстве с МАЖуринской и А.И.Новиковым). Мой объем - 0,3 ал.

Ярославцева Е.И. Формализованное представление системы знаний о языках мира // Конф. «Оптимизация преподавания языков и инженерная лингвистика». Ульяновск, 1985. 0,1 ал., (в соавторстве с А.И.Новиковым). Мой объем - 0,05 ал.

Ярославцева Е.И. Семантические расстояния в языке и тексте. М., 1990. 15 ал. (в соавторстве с А.И.Новиковым). Мой объем - 7,5 ал.

Ярославцева Е.И. Размышления о языкознании и языке. Эссе // RES Linguistika. Сборник статей к 60-летию В.П. Нерознака. М., Academia, 1999. 0,5 ал.

Ярославцева Е.И. Грамматикой, его виды и аналоги // Язык, сознание, коммуникация. Вып. 10. МГУ, М., 1999.1 ал.

Ярославцева Е.И. Грамматикой и база данных "Языки мира" // Проблемы прикладной лингвистики 2001. М., 2002,1ал.

Yaroslavtseva E.I. Linguotypological Data Bank // Social Sciences.USSR Academy of Sciences. Vol. XVII, No. 3, 1986, (Novikov). 1 ал. Мой объем - 0,5 ал.

Ярославцева Е.И. База данных «Языки мира» как инструмент лингвистического исследования // Вопросы языкознания. 2003, № 3. 0,8 ал. (в соавторстве с В.А.Виноградовым, А.И.Новиковым). Мой объем - 0,3 ал.

Ярославцева Е.И. Географический и генетический указатели к базе данных «Языки мира» // Проблемы прикладной лингвистики. Выпуск 2. М., 2004.1 ал.

Ярославцева Е.И. Компьютерная база данных «Языки мира» и формальная типология // Сборник памяти А.И.Новикова. Уфа, 2004, (в печ.), 0,8 ал.

Сдано в печать 25 февраля 2005г.
Объем печати 1 п.л. Заказ № 528. Тираж 000 экз.
Отпечатано: ООО «Спринт-Принт»
г. Москва, ул. Краснобогатырская, 92
тел.: 963-41-11,964-31-39