ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 3 2003

© 2003 г. В. А. ВИНОГРАДОВ, А. И. НОВИКОВ, Е. И. ЯРОСЛАВЦЕВА

БАЗА ДАННЫХ "ЯЗЫКИ МИРА" КАК ИНСТРУМЕНТ ЛИНГВИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

В Институте языкознания РАН давно ведется разработка базы данных (БД), включающей в себя свернутые, формализованные и определенным образом структурированные описания языков мира. БД "Языки мира" в настоящее время 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 статьи является краткое описание устройства этой БД и тех, полезных для лингвистики функций, которые она способна выполнять на данном этапе ее создания.

У истоков создания БД "Языки мира" стояла член-корреспондент РАН В.Н.Ярцева – автор и руководитель проекта «Энциклопедия "Языки мира"». Ей принадлежала идея создания с помощью компьютера справочного аппарата к этой энциклопедии в виде системы указателей. Затем идея трансформировалась в проект по разработке базы данных, создаваемой на основе энциклопедии, но позволяющей в дальнейшем извлекать сведения лингво-типологического характера без непосредственного обращения к самому энциклопедическому изданию.

1. КРАТКАЯ ИСТОРИЧЕСКАЯ СПРАВКА

Первоначально разработка базы данных осуществлялась в Институте языкознания РАН специально созданной для этой цели проблемной группой, куда входили сотрудники как отдела прикладного языкознания (А.К. Валентей, А.И. Новиков – руководитель группы, Н.К. Рябцева, Е.И. Ярославцева), так и сотрудники группы "Языки мира" (М.А. Журинская, В.П. Калыгин, А.А. Кибрик, Н. Рогова, Я. Тестелец). Большую помощь группе оказывали редакторы и авторы энциклопедии: М.Е. Алексеев, В.А. Виноградов, Г.А. Климов и др. В этот период был осуществлен теоретический анализ проблемы, подготовлен исходный материал и проведены экспериментальные исследования. Полученные результаты нашли отражение в ряде публикаций, в том числе, в монографии [Журинская, Новиков, Ярославцева 1986], представляющей собой своего рода эскизный проект создаваемой базы данных

В начале 90-х годов работы по наполнению базы данных информацией из-за плохого финансирования велись недостаточными силами Несмотря на это, старшим научным сотрудником отдела прикладного языкознания Ю.П. Скоканом был разработан комплекс программ, позволяющий осуществлять ввод, редактирование и корректировку информации. С помощью данного программного продукта было введено и отредактировано около 200 описаний языков как на русском, так и на английском языках усилиями главным образом старшего научного сотрудника Е.И.Ярославцевой.

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.

2 УСТРОЙСТВО БАЗЫ ДАННЫХ

Что представляет собой БД "Языки мира" в структурном отношении? Как и всякая другая БД, это – таблица, состоящая из строк и столбцов В каждой строке такой таблицы записан один определенный языковой факт, а столбцам поставлены в соответствие названия языков, по отношению к которым эти факты являются релевантными Наличие такой взаимной релевантности фиксируется в виде определенной пометки на пересечении соответствующих столбцов и строк Множество языковых фактов, относящихся к некоторому конкретному языку, составляет формализованный реферат описания данного языка Он может быть вычленен из общей структуры БД и представлен отдельно без пометок в виде перечня составляющих его языковых фактов

Содержимое всех строк, взятых без пометок, отдельно, как перечень языковых явлений и категорий, составляет так называемую модель реферата (МР) В отличие от рефератов, которые создаются в процессе функционирования БД, модель реферата, вернее, ее базовый компонент, создается предварительно на одном из первых этапов ее построения

Базовый компонент – это тот минимум языковых фактов, который является необходимым и достаточным для описания нескольких первых языков, попавших в экспериментальный массив, на котором отрабатывались основные принципы построения БД В дальнейшем МР постоянно расширяется в процессе эксплуатации БД за счет включения в нее тех языковых фактов, которые встречаются в описании очередного языка, но не содержатся в ней

В структурном отношении МР представляет собой классификационную схему в виде иерархического дерева Первому уровню такой иерархии соответствуют наименования классов языковых явлений Они практически идентичны тем классам, которые содержатся в схеме статьи, используемой в энциклопедии "Языки мира

Внутри классов выделяются возможные аспекты рассмотрения данного класса языковых явлений Совокупность аспектов – это дальнейшая градация содержания, осуществляемая на уровне каждого раздела.

Классы и аспекты – это универсалии, априорно задаваемые в модели как наименования явлений, общих либо 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 Различие между подаспектами и характеристиками носит скорее формальный, чем содержательный характер подаспект определяется как обобщенное название какой-либо однородной группы характеристик, являющееся в модели подчиненным некоторому аспекту 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 Количество подаспектов внутри некоторого аспекта может варьироваться в достаточно широких пределах аспекту могут быть непосредственно подчинены характеристики (т е подаспекты отсутствуют), и внутри аспекта может содержаться до пяти (в принципе неограниченное число) уровней иерархии, последний из которых (самый низкий) представлен характеристиками, а все промежуточные являются по определению подаспектами На одном и том же уровне иерархии могут находиться элементы, имеющие статус подаспекта и статус характеристики первые подчиняют себе некоторые другие элементы, вторые не характеризуются подчиненными элементами

Например, внутри аспекта "тон" (класс 2 1 2 "Просодические явления") выделяются подаспект "число тонов" (подчиняет характеристики "два", "три", 'более трех), подаспект регистровые признаки" (подчиняет характеристики "высокий/невысокий , высокий/средний/низкий" и т д), подаспект "контурные признаки" (подчиняет ха-

рактеристики 'восходящий/нисходящий" и пр), характеристика фонологически значимый тон", подаспект "носитель тона" (подчиняет характеристики "слог", 'группа слогов", "слово" и т д.)

Для отражения иерархии элементов, составляющих модель, используется специальная формализованная запись. Каждый следующий уровень иерархии записывается со сдвигом вправо Каждый сдвиг обозначается точкой В результате каждый последующий уровень иерархии имеет на одну "ведущую точку" больше, чем предшествующий Такая форма записи достаточно наглядна и легко позволяет автоматически определять уровень иерархии

По своей роли в модели рассмотренные выше структурные элементы делятся на два основных типа классифицирующие и фактографические К первому типу относятся наименования классов, аспектов и подаспектов, т е все те элементы, которые являются подчиняющими Среди классифицирующих элементов можно выделить постоянные элементы, т е такие, которые задаются в модели обязательно и с необходимой полнотой

Ко второму типу структурных элементов (фактографическому) относятся характеристики Характеристики, в отличие от классифицирующих элементов, не могут быть заданы с исчерпывающей полнотой и точностью Они соответствуют микротемам описания языков мира и должны задавать способ языкового выражения этих микротем Поэтому характеристики, включенные в модель, следует рассматривать как примеры возможного раскрытия подаспектов, языковое оформление которых служит своего рода образцом для записи характеристик конкретных языков

В модели обычно приводятся наиболее типичные, часто встречающиеся характеристики Если они совпадают с теми языковыми явлениями, которые описаны в статье некоторого языка, то их можно механически переносить в реферат В противном случае необходимо формулировать микротему таким образом, чтобы получилась новая характеристика, аналогичная по своим лексико-синтаксическим параметрам тем характеристикам, которые приведены в модели

По отношению к характеристикам, отражающим конкретные языковые данные по соответствующим языкам, классифицирующие элементы играют вспомогательную роль Они служат своего рода координатами, задающими место той или иной характеристики в модели и в конкретных ("реальных") рефератах

Для иллюстрации соотношения структурных элементов в модели приведем в качестве примера фрагмент класса 2 1 2 "Просодические явления"

2 1 2 Просодические явления	*– класс*
Ударение	*– аспект*
тип ударения	*– подаспект*
нефонологическое	*– характеристики*
фонологическое	
вид ударения	*– подаспект*
музыкальное	*– характеристики*
динамическое	
количественное	
смешанное	
Тон	*– аспект*
число тонов	*– подаспект*
два	*– характеристики*
три	
более трех	
и т д	

Реферат имеет идентичную структуру, поскольку является результатом включения в него из модели 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. Можно считать, что если реферат является основной единицей ввода, хранения и обработки в БД, 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 БД.

В настоящее время в БД содержится около 300 рефератов, что соответствует количеству статей, содержащихся во всех вышедших из печати к настоящему времени томах энциклопедии "Языки мира". Модель реферата, пополненная лингвистическими терминами в процессе ввода этих рефератов, включает в себя в настоящее время около 4 тысяч единиц, образующих своего рода "грамматикон", описание которого содержится в [Ярославцева 2002].

3. ОСНОВНЫЕ ЗАДАЧИ, РЕШАЕМЫЕ С ПОМОЩЬЮ БД

3.1. Формирование и ведение базы данных. Основные функциональные возможности данной системы реализованы в ее программном обеспечении. Как уже отмечалось, комплекс программ для данной БД был разработан Ю.П. Скоканом. Они были написаны на языке Clipper и позволяли осуществлять ввод, хранение, инспекцию, редактирование и преобразования рефератов, а также ввод новых строк в модель реферата. Кроме того, с помощью этого программного продукта можно осуществлять автоматизированный перевод рефератов на английский язык. Программная реализация этих функций позволяет рассматривать процесс формирования БД как процесс ее постоянного расширения как по горизонтали (ввод новых языков), так и по вертикали (ввод новых строк в модель реферата).

В настоящее время данное программное обеспечение было перепрограммировано на языке Delphi и адаптировано под Windows[1]. Во второй версии программного обеспечения полностью сохранена идеология первой версии. Дополнительно были реализованы функции БД, связанные с информационным поиском.

3.2. Формальное сопоставление языков для установления степени близости между ними. Принятая форма представления информации в БД дает возможность проводить построчное сопоставление рефератов между собой и вычислять количественные показатели, характеризующие степень близости языков на структурном (грамматическом) уровне. В принципе специально разработанная для этой цели программа[2] позволяет осуществлять сопоставление каждого языка с каждым и получить 90 тысяч результатов такого попарного сопоставления. Но очевидно, что такой тотальный сопоставительный анализ не требуется, поскольку во многих случаях и без этого соотношение языков является очевидным. Поэтому наиболее целесообразным является такой режим эксплуатации данной программы, когда пользователь системы сам выбирает в качестве исходных и сравниваемых языков то множество языков, которые он считает наиболее интересным с точки зрения решаемых с помощью данной системы задач.

Результатом каждого попарного сравнения является некоторый количественный показатель в интервале от 1 до 0. Он характеризует степень близости сравниваемых языков в целом. Кроме того, количественные показатели близости определяются и на уровне каждого класса внутри сравниваемых языков. Все это фиксируется в виде протокола, который записывается в отдельный файл, и к нему впоследствии можно обращаться для анализа полученных данных. При необходимости он может быть скопирован, напечатан и т.д.

---

[1] Модернизация программного обеспечения осуществлена В.Н. Поляковым и В.В. Логуновым.

[2] Алгоритм и программа составлены А.И. Новиковым.

В верхних строках протокола фиксируется название сравниваемых языков, а также интегральная характеристика близости этих языков. Затем фиксируется наименование очередного класса и содержимое тех строк, которые совпали в результате сравнения данных рефератов. В каждой такой строке справа приводится "вес", который присваивается ей автоматически в процессе сравнения. Запись данных, относящихся к отдельному классу, завершается строкой, где записывается показатель близости сравниваемых языков на уровне этих классов. В конце протокола приводится список индексов классов и количественных показателей близости, упорядоченный по убыванию их величины. Ниже приводятся фрагменты такого протокола в качестве примера.

Исх. =.ИСПАНСКИЙ
.галисийский
S = .7374066987752987
2.1.1.ФОНЕМНЫЙ СОСТАВ
гласные 1
подъем 2
верхний/средний/нижний 3
ряд 2
передний/средний/задний 3
открытые/закрытые 2
общая схема монофтонгов 2
различия по ряду 3
в нижнем подъеме 4
отсутствуют 5
в верхнем подъеме 4
передний/задний 5
число монофтонгов 2
.....
вибранты 5
переднеязычные 6
глайды 5
губные/среднеязычные 6
дополнительные признаки 4
ретрофлексные/неретрофлексные 5
вибранты 6
Skl= .8155563472271702

2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ
ударение 1
вид 2
динамическое 3
носитель ударения 2
слог 3
фиксированность 2
связанное 3
Skl= .6555183946488294
.....
2.5.4.СЛОЖНОЕ ПРЕДЛОЖЕНИЕ
особенности подчиненного компонента 1
оформление сказуемого 2
нефинитные формы 3
тип построения 1
сочинение/подчинение 2
тип связи 1
союзная/бессоюзная 2
союзы 2
самостоятельные служебные элементы 3
Skl= .8795454545454545

2 3 1 = 1
2 3 7 = 1
2 5 4 = 8795454545454545
2 5 2 = 8376623376623377
2 3 4 = 8232998885172798
2 1 1 = .8155563472271702
2 3 6 = 7682539682539682
2 4 0 = 762987012987013
2 3 3 = 7617021276595746
2 1 4 = 7222222222222222
2 1 3 = 7086681974741675
2 5 1 = 6862745098039216
2 1 2 = 6555183946488294
2 2 3 = 6529411764705882
2 3 5 = 6286518435125246
2 5 3 = .6076555023923444
2 3 2 = 589247311827957
2 3 0 = 3434022257551669
2 2 1 = 3333333333333334
2 2 2 = 2916666666666666

После сопоставления всех отобранных пользователем языков также производится упорядочение интегральных показателей их близости. Список таких упорядоченных показателей вместе с наименованиями соответствующих языков фиксируется в специальном файле, что позволяет хранить результаты сопоставления для последующего обращения к ним. Приведем в качестве примера такой упорядоченный спи сок для группы романских и некоторых других языков

Исходный = ИСПАНСКИЙ
еврейско-испанский 7864494160209003
галисийский 7374066987752987
итальянский 7142775508188599
старофранцузский 6977204222698277
португальский 6928346624175639
английский 690846066048237
провансальский 6872384720814565
ретороманский 6766647294743898
румынский 6681295708664254
латинский (м) 658930540489609
французский 6539475905969674
арагонский 6439591580251719
гасконский 6255930668137184
франкопровансальский 5963383739859804

Эти протоколы сами по себе могут представлять интерес в качестве материала для анализа при решении тех или иных вопросов теоретического или практического характера Но преимущественно интерес они могут представлять как большой целостный массив информации, к которому могут быть применены определенные автоматизированные средства обработки.

Получаемые в процессе описанной выше процедуры количественные данные имеют не абсолютную, а относительную значимость, что предполагает их определенную интерпретацию Для этого необходимо знать те принципы, которые были положены в основу данного метода, и те конкретные операции, которые используются в данной процедуре И ниже приводятся основные положения, на которых базируется данный метод

1) Структура реферата позволяет отразить описание того или иного языка, хотя и свернуто, но с той степенью полноты и точностью, с которой оно представлено в энциклопедии

2) Стандартность, унифицированность, формализованность процедуры составления рефератов делает их в структурном отношении подобными друг другу, а, следовательно, и сопоставимыми между собой.

3) Если это так, то возникает возможность прямого (построчного) сопоставления рефератов между собой

4) При этом можно считать, что чем большее количество строк сравниваемых рефератов совпало, тем ближе между собой эти рефераты, а следовательно, и языки, описываемые ими.

5) Такой чисто количественный подход является слишком упрощенным, в связи с тем, что в структурном и содержательном плане сравниваемые строки рефератов не являются равнозначными по отношению друг к другу. Поэтому здесь необходимо учитывать также и качественный аспект такого сопоставления.

6) Качественный аспект в данной процедуре может быть задействован, например, через приписывание строкам реферата "весов", отражающих их значимость в описании данного языка. При этом желательно, чтобы "весомость" строк определялась на основании формальных критериев.

7) В качестве такого критерия может служить степень конкретности и фактуальности содержимого строк реферата. Как уже отмечалось, наибольшей конкретностью обладают те языковые явления, которые в реферате выступают в роли характеристик и выражают фактографическую информацию, в отличие от других, выполняющих классифицирующую роль. Например, если при сравнении двух рефератов совпали такие категории, как "гласные", "подъем" и "ряд", но не совпали строки, в которых содержится информация о самих этих гласных (верхний/средний/нижний, передний/средний/задний и др.), то значимость такого совпадения является очень низкой Но здесь существует и обратная зависимость: если совпали конкретные характеристики, то должны совпасть одновременно и все вышестоящие подчиняющие подаспекты и аспекты. Таким образом, в сопоставлении участвует по сути не отдельная строка, а вся ветвь иерархического дерева, образующая иерархический контекст данной характеристики. При этом наибольшую значимость имеет именно характеристика, занимающая конечную позицию в этой цепи, а вышестоящие звенья этой цепи имеют значимость тем меньшую, чем более высокую ступень они занимают в этой иерархии.

8) В связи с этим в качестве формального критерия приписывания "весов" можно использовать количество сдвигов в каждой строке. На этом основании был реализован принцип, в соответствии с которым "вес" каждой строки равен количеству сдвигов, с которым записано ее содержимое: чем больше сдвигов, тем большим будет вес, приписывемый данной строке.

9) Этот критерий является формальным, а потому "веса" приписываются автоматически

10) Степень близости сравниваемых языков определяется по следующей формуле:

$$S = \frac{\sum_{i=1}^{m} Skl(i)}{m}, \qquad (1)$$

где M – количество классов, а $Skl(i)$ – результат сравнения по классу, полученный следующим образом:

$Skl(i) = (F + P)/2$, где·

$$F = \frac{\frac{Ws1}{Ws} + \frac{Ws1}{Ws2}}{2}, \qquad (2)$$

$$P = \frac{\frac{Kl1}{Kl} + \frac{Kl1}{Kl2}}{2}, \qquad (3)$$

где:
Ws = вес всех строк очередного класса исходного реферата;
*Ws*1 = вес всех совпавших строк сравниваемых классов;
*Ws*2 = вес всех строк очередного класса сравниваемого реферата;
Kl = количество всех строк очередного класса исходного реферата;
*Kl*1 = количество совпавших строк сравниваемых классов;
*Kl*2 = количество строк очередного класса сравнивамого реферата;

Возможны альтернативные варианты присвоения "весов". Например, для этой цели можно использовать знания экспертов, которые присваивают "веса" определенным строкам модели реферата на основе своих знаний о их типологической значимости. Затем эти веса могут присваиваться соответствующим строкам реферата уже в автоматическом режиме. Но здесь имеются свои недостатки. Во-первых, в этом случае неизбежна определенная субъективность в оценке значимости строк. Во-вторых, используя знания экспертов, можно получить в результате то, что уже известно (по принципу: "что ввели, то и получили"). В случае же, когда при использовании полностью автоматизированной процедуры приписывания весов будут получены уже известные данные, значимость этого факта не будет тривиальной, поскольку получены они другим методом. Если же будут иметь место расхождения с общеизвестными данными, то можно предполагать, что мы имеем дело возможно с новым знанием. Но такое расхождение может быть также и следствием погрешностей в работе системы. Они вполне возможны как по объективным, так и по субъективным причинам. К объективным причинам относится разная степень полноты и точности описания языков вследствие недостаточной их изученности или отсутствия сведений вообще по какому-либо классу того или иного языка, что фиксируется в реферате знаком ". О ". Такие лакуны имеются, например, в описании гасконского и франко-провансальского языков, вследствие чего они занимают нижние позиции в приведенной выше таблице. В связи с этим одной из функций БД может служить поиск лакун в описании языков, что, однако, связано с другой задачей – задачей информационного поиска.

3.3. Поиск информации. Основной функцией БД, как и всякой информационной системы, является автоматизированный поиск информации, необходимой для ответа на запросы, поступающие в систему, а также обработка этой информации для формирования ответа в зависимости от типа запроса и его цели.

Как показывает анализ запросов, которые могут поступить в БД, они являются простыми, распространенными и сложными. Приведем примеры.

Простые запросы:
1. В каких языках имеется категория "род"?
2. В каких языках отсутствует "залог"?
3. Какие языки распространены в Италии?
Распространенные запросы:
4 Какие признаки гласных имеются в романских языках?
5.Сколько существует кавказских языков?
6. Чем отличаются тюркские языки от монгольских?
Сложные запросы:
7. К каким семьям принадлежат языки, в которых есть "изафет"?
8. Где распространены языки с иероглифической письменностью?
9. К каким семьям принадлежат языки Ближнего Востока?
и др.

Как видно из приведенных примеров, все эти запросы имеют различную структуру, предполагают различные поисковые процедуры, а также различную форму выдачи ответа.

Простые запросы характеризуются тем, что здесь в качестве ответа должен выдаваться перечень языков, в которых содержится то или иное языковое явление, ин-

тересующее абонента системы. Кроме того, их особенностью является то, что они сформулированы в терминах, совпадающих с терминами представления информации в БД. Поиск информации здесь совершается как бы за один "проход" при непосредственном обращении к рефератам, а потому является одноступенчатым.

Распространенные и сложные запросы требуют проведения многоступенчатого поиска. Например, в ответ на 4-й запрос необходимо сначала определить, какие языки входят в понятие "Романские языки", а затем составить полный перечень характеристик, входящих в класс "2.1.1.Фонетические явления", для данной семьи языков. Кроме того, они могут формулироваться в терминах, не содержащихся в БД в качестве ее основных единиц, что предполагает процедуру перевода запроса с "естественного лингвистического языка" на язык представления информации в БД. Формой выдачи здесь не обязательно является перечень языков. Это может быть и перечень языковых явлений, присущих той или иной группе языков, и некоторые количественные данные и др.

В настоящее время в базе данных "Языки мира" реализован автоматический информационный поиск только для простых запросов. Он может осуществляться в двух основных режимах: поиск по отдельным языковым явлениям и по их комбинациям, где реализованы такие логические операции, как конъюнкция, дизъюнкция и отрицание.

Поиск по отдельным языковым фактам представляет собой фактически поиск выбранной строки во всем множестве рефератов. Этот выбор может осуществляться двумя способами. Если пользователь БД хорошо знаком со структурой модели реферата и легко ориентируется в ней, то для выбора строки ему предоставляется непосредственно сама модель, см. рис. 1.

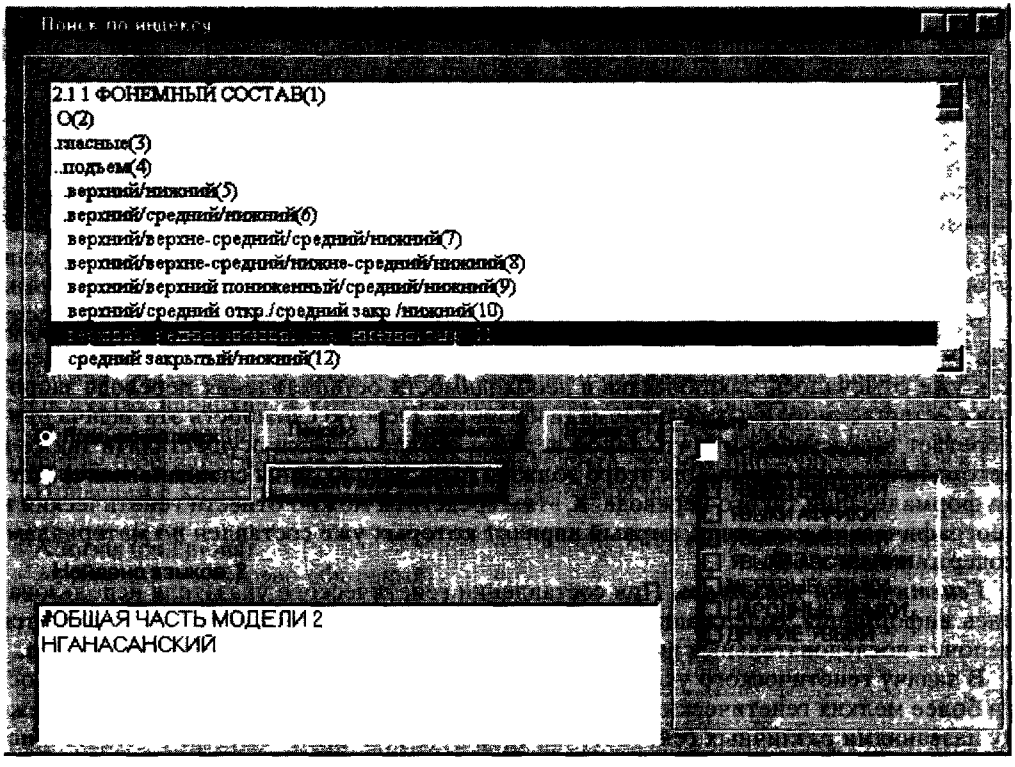

Рис. 1. Поиск информации по строке из МР

В том случае, если пользователь не знаком с внутренней структурой БД, ему предоставляется перечень наименований языковых явлений, идентичных тем, что содержатся в модели реферата, но расположенных по алфавиту. Кроме того, эти языковые явления представлены здесь в иерархическом контексте, т.е. в виде цепочек иерархического дерева, что устраняет возможность неоднозначного их восприятия компьютером, см. рис. 2.

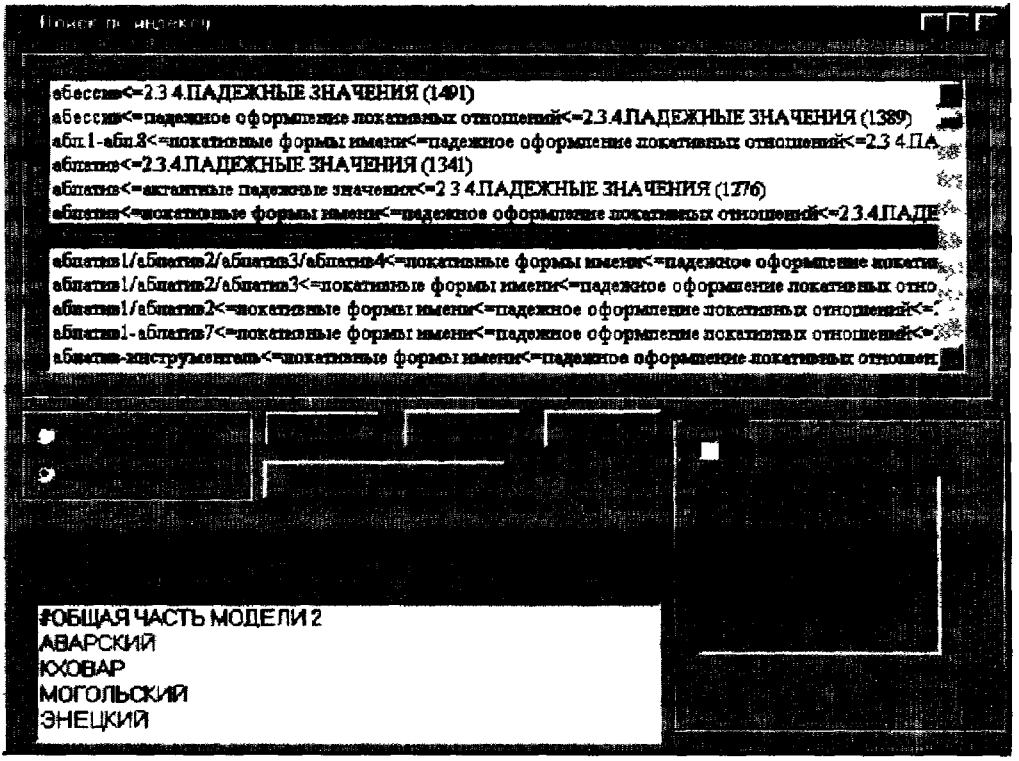

Рис. 2 Поиск информации по строке из алфавитно-предметного указателя

Разработка процедуры автоматического поиска по более сложным видам запросов – это задача, которая решается в настоящее время. Основная проблема здесь, как уже отмечалось, заключается в необходимости осуществления перевода запросов с естественного языка на формализованный. Особую сложность эта задача приобретает вследствие того, что такой перевод предполагается осуществлять также в автоматическом режиме. Для этого должны быть разработаны специальные средства формализации такого перевода. К этим средствам можно отнести генетический и географический указатели, первый вариант которых уже составлен по материалам, содержащимся в БД.

Генетический указатель. При составлении генетического указателя использовалась информация, содержащаяся в классе 1.1.2. каждого реферата, где приводится цепочка последовательных включений языка в подгруппу, группу и семью языков.

В задачу генетического указателя входит отражение информации о включенности более мелких генетических единиц в более крупные и задание соотношения между названиями различных генетических единиц (в том числе синонимическими и историческими названиями) и заглавиями статей энциклопедии.

Входом в генетический указатель служит название любой единицы диалектной системы, название языка, подгруппы, группы и семьи языков, варианты перечислен-

ных названий, в том числе и исторические. Отсылкой для указанных входов служит либо название более крупной генетической единицы (для говора – диалект, для диалекта – язык, для языка – подгруппа языков, для подгруппы – группа языков, для группы – семья языков), 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 отсылка вида "язык-изолят". Особым образом выделены названия языков, подгрупп, групп и семей языков, для которых в энциклопедии имеются самостоятельные статьи, независимо от того, в каком качестве – входа или отсылки – употреблено выделяемое название. Вход и отсылка связываются знаком "<" при наличии отношения включения и знаком "=" при наличии синонимического отношения.

Все названия в генетическом указателе даны в полном (несокращенном) виде и расположены в алфавитном порядке входов. Приведем в качестве примера фрагмент генетического указателя.

абага-сунутский гов. < халхаский диал.
абадзехский диал. < адыгейский
абазинский < абхазо-адыгские
абазский (ист.) = абазинский
абаканских татар (ист.) = хакасский
абаканских тюрок (ист.) = хакасский
абдуллинский гов < средний диал. татарского яз.
абжуйский диал. < абхазский
абхазо-адыгские < иберийско-кавказские
абхазский < абхазо-адыгские
абхазо-адыгские < кавказские
.. ..

Географический указатель. Цель географического указателя – дать возможность узнать, какие языки распространены в той или иной стране или регионе.

Входом в указатель служит название некоторой географической единицы (топонима), отсылкой – перечень языков, распространенных в пределах этой географической единицы.

Входы в указатель имеют следующий вид: наименование географической единицы и со сдвигом – перечень распространенных в ее пределах языков, а в скобках после наименования каждого языка – наименование тех функций (статусов), в которых выступают описанные в энциклопедии языки в пределах этой географической единицы.

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.).

Приведем в качестве примера фрагмент географического указателя:

Абхазия (см.Грузия)
.абазинский (разговорно-обиходный)
.абхазский (официальный)
.грузинский (официальный)
.лазский (разговорно-обиходный)
.русский (разговорно-обиходный)
сванский (разговорно-обиходный)
Австрия
венгерский (разговорно-обиходный)
немецкий (государственный)
... ...

Данные указатели имеют и самостоятельную значимость как особые поисковые устройства, раскрывающие содержимое БД в генетическом и географических аспектах.

Другим, как предполагается, достаточно эффективным средством автоматизации процесса перевода запросов на формализованный язык будет служить тезаурус, в котором, кроме терминологии, содержащейся в модели реферата, будет содержаться синонимичная лексика, а также необходимые парадигматические связи и отношения между используемой в БД лингвистической терминологией. Работа над таким тезаурусом в настоящее время ведется.

СПИСОК ЛИТЕРАТУРЫ

Журинская, Новиков, Ярославцева 1986 – *М.А. Журинская, А.И. Новиков, Е.И. Ярославцева.* Энциклопедическое описание языков. М., 1986.

ПОЯМ 1985 – Принципы описания языков мира. М., 1985.

Ярославцева 2002 – *Е И. Ярославцева.* Грамматикон и база данных "Языки мира" // Проблемы прикладной лингвистики 2001. М., 2002.

Novikov, Yaroslavtseva 1986 – *A. Novikov, E.Yaroslavtseva.* Linguotypological Data Bank // Social Sciences. USSR Academy of Sciences. V. XVII. № 3. 1986.