На правах рукописи

Клочко Алексей Данилович

ОПТИМИЗАЦИЯ ИНДИВИДУАЛЬНЫХ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ СРЕДСТВАМИ СПЕЦИАЛИЗИРОВАННОЙ БАЗЫ ДАННЫХ

10.02.21. – прикладная и математическая лингвистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата филологических наук

Тверь – 2006

Работа выполнена на кафедре английской филологии в Армавирском лингвистическом университете

Научный руководитель: Сакиева Римма Сафраиловна**,** проф., д.ф.н.

Официальные оппоненты: Чагров Александр Васильевич, д.физ.мат.н.

Морозова Оксана Николаевна, д.ф.н.

Ведущая организация: Кубанский Государственный университет

Защита состоится «____» _____ 2006 г. в ___ час. на заседании диссертационного совета Д 212.263.06. при Тверском государственном университете по адресу: Россия, 170002, г. Тверь, проспект Чайковского, д. 70, корпус 4, филологический факультет, ауд. 47.

С диссертацией можно ознакомиться в научной библиотеке Тверского государственного университета (г. Тверь, ул. Володарского, 42).

Отзывы можно отправлять по адресу: Россия, 170013, г. Тверь, ул. Желябова, 33, Тверской государственный университет, ученому секретарю.

Автореферат диссертации разослан «____» _____ 2006 г.

Ученый секретарь
диссертационного совета
доктор филологических наук, профессор Л.Н. Скаковская

Общая характеристика исследования

Актуальность заявленной нами темы состоит в том, что индивидуальные лингвисты-исследователи, желающие оптимизировать сам процесс своей научной работы с помощью персонального компьютера, до сих пор не имеют обобщающих работ по данной проблеме. Правда, имеются многочисленные публикации, в т.ч. обобщающего характера, о применении обучающих программ на платформе ПК (персональных компьютеров), но при этом предмет исследований лежит в дидактической и общепедагогической плоскости. Рабочими группами исследователей в различных научных центрах (Киев, Москва – см. раздел Библиография) разработаны программы для ПК, напр., для разработки частотных словарей, т.е. для лексикологических исследований в области лингвостатистики. Однако, будучи раз созданы, они не позволяют без переписывания кода перенастраивать себя для новых исследовательских задач, и даже для модификации той же задачи.

С другой стороны, существует обширная литература по такой бурно развивающейся области лингвистики, как *прикладная и математическая лингвистика*, лабораторно-материальной базой которой служат большие ЭВМ, имеющие свой персонал ИТР и программистов, научные коллективы различных уровней и научных направлений. Их целью являются большие проекты: большие лингвистические базы данных, которые затем материализуются в многотомные словари (объемом до нескольких десятков томов), энциклопедические издания в области лингвистики (напр. «Языки мира»), электронные базы данных, постоянно пополняемые аспектно ориентированными лингвистической фактологией и применяемые для синхронических и диахронических исследований (см. далее в кратком историческом обзоре прикладной лингвистики).

Кроме того, мы можем найти достаточное количество литературы по разработке баз данных для применения в бизнесе или юридической практике. Но лингвистические исследования имеют свою ярко

выраженную специфику и почти необозримую широту предметов исследования.

Таким образом, мы констатируем, что существует своего рода *«серая зона»* между глобальными лингвистическими исследованиями, обеспеченные мощными материально-техническими и человеческими ресурсами (т.н. Большие Проекты) и исследованиями лингвистов-одиночек, ведущие исследования средствами 19 века.

Предлагаемая работа как раз и является попыткой обобщения и систематизации опыта создания и применения БД в указанных выше областях, а также изложением путей научно обоснованной оптимизации баз данных, специализированных для индивидуальных частнолингвистических исследований.

Объектом исследования является специализированные (лингвистические) базы данных).

Предметом исследования является оптимизация электронных средств хранения и обработки лингвистических данных в целом, т.е. и специализированных баз данных и (присоединенных) электронных таблиц.

Цель исследования — разработка и оптимизация специализированной электронной базы данных для хранения и обработки данных лингвистического исследования для индивидуального лингвиста-исследователя.

Задачи исследования. Из поставленной цели исследования следует необходимость решить несколько исследовательских задач теоретического и практического плана. В теоретическом отношении мы делаем попытку.

а) в историко-научном аспекте осветить проблемы компьютерной лингвистики и лингвистических баз данных различного типа, обсуждаемые в работах отечественных и зарубежных исследователей; б) изучить и обобщить опыт применения больших ЭВМ в различных сферах прикладной лингвистики, оценить возможность перенесения части этого опыта на платформу ПК; в) определить возможности и ограничения

персонального компьютера, в оптимизации и интенсификации труда индивидуальных лингвистов; г) на реальном примере разработанной нами базы данных для частнолингвистического исследования показать упомянутые возможности и ограничения ПК; д) предложить классификацию лингвистических баз данных, посильных для разработки индивидуальными лингвистами-исследователями.

Из общетеоретической цели исследования необходимо следуют частные практические задачи: а) формулирование принципов общий структуры и подсистем частнолингвистической (словообразовательной) базы данных для индивидуального исследователя; б) выделение критериев (частичный аналог зон или помет словарных статей) для запросов на выборку[1] в частнолингвистической БД; в) компьютерный поиск и отбор языковых единиц и их эксплицитных словоформ по заданным параметрам; д) разработка частнолингвистической базы данных в качестве примера.

Методология предлагаемой работы опирается на теоретическую базу прикладной и компьютерной лингвистики, созданную работами отечественных отечественных ученых в области прикладной лингвистики (А.Е.Кибрик, Р.К.Потапова, Ю.В.Рождественский, Б.Ю.Городецкий, Л.Н.Беляева, Р.Ю.Кобрин, С.Д.Шелов, Р.Г.Пиотровский, А.С.Герд, В.М.Лейчик, А.Н.Баранов, Г.В.Колшанский и др.). С учетом междисциплинарного характера данного исследования, мы обратились также к теории баз данных (БД), аспекты которой изложены в трудах основоположника реляционных баз данных, американского математика Э. Кодда (Edgar F. Codd), а также экспертов по СУБД MS Access – Дж. Вискас (John L. Viescas), и в работах отечественных экспертов по теории БД – Кагаловский М.Р., Бойко В.В., Каратыгин С.А..

Методы исследования. Нами применялись следующие общенаучные и частные методы**:**: общенаучные методы (методы

[1] Как пример возможностей частнолингвистической БД; это замечание относится и к пунктам «в, г, д».

эмпирического исследования — наблюдение, сравнение, измерение, эксперимент, моделирование; методы эмпирического и теоретического уровня – абстрагирование, анализ и синтез; методы теоретического уровня – метод восхождения от абстрактного к конкретному. Лингвистические методы – поскольку наше исследование носит междисциплинарный характер, то в нашем случае речь может идти о комплексе частных методов из частных дисциплин: сравнительно-сопоставительный метод, метод компонентного семантического анализа.

Новизна исследования. Созданная нами специализированная база данных для хранения и обработки результатов индивидуальных лингвистических исследований на примере коллоквиальных композитных существительных представляет собой первый опыт оптимизации и интенсификации НИР средствами СУБД для ПК в среде лингвистов-индивидуалов, т.е. не входящих в «команды Больших Проектов». Это особенность предлагаемой работы и определяет её новизну.

Теоретическая значимость данного исследования состоит в том, что оно вносит вклад в систематизацию и развитие компьютерных методов индивидуальных частнолингвистических исследований. Сформулированные принципы создания и оптимизации лингвистических баз данных, модифицируемых для индивидуальных исследований, могут послужить стимулом для дальнейшей компьютеризации и информатизации НИР индивидуальных лингвистов с чисто гуманитарным менталитетом. Это поможет им рационализировать, ускорить поиск релевантного для исследования языкового материала и обработку полученных результатов в целях повышения объективности формулируемых закономерностей.

Практическое применение результатов исследования заключается: а) в возможности пополнения индивидуальными лингвистами-практиками разработанной нами БД «Словообразовательные аспекты коллоквиализмов» из доступных источников разговорной лексики; б) полученный словник, снабженный индексацией по релевантным

частнолингвистическим (словообразовательным и семантическим) параметрам, может использоваться ими для лингводидактических задач; в) в возможности разработки оригинальных БД для других специфических направлений индивидуальных лингвистических исследований; г) в оптимизации разработки тематических учебных словарей по узким отраслям.

Научная гипотеза: индивидуальная электронная лингвистическая база данных (ИЭЛ БД) является особым видом БД, которая должна обладать специфической структурой, оптимизированной для индивидуальных лингвистических исследований и предусматривающей возможность модификации в случае последующего уточнения задач, что неизбежно в ходе НИР.

Положения, выносимые на защиту:

1. Существующие аппаратные и программные средства для лингвистических исследований в подавляющем своем большинстве разрабатывались или оптимизировались для научных коллективов и т.н. Больших Проектов.
2. Опыт применения персональных компьютеров в прикладной лингвистике, особенно БД, также относится в основном к исследовательским или проектным группам.
3. Существует необходимость оптимизации и интенсификации индивидуальных лингвистических исследований, которые зачастую ведутся без применения ПК (если не учитывать набор текста), что
   а) затягивает накопление фактического материала и его обработку и
   б) при ручном методе некоторые закономерности трудно прослеживаются или допускают субъективную интерпретацию.
4. Оптимальным решением было бы создание электронного рабочего места индивидуального лингвиста-исследователя, которое состояло бы из: а) системы управления базами данных (СУБД) с набором специализированных баз данных (БД) с оптимизированной

структурой для задач исследования; б) приложений на основе электронных таблиц для автоматизации статистических вычислений; в) шаблонов MS Word для хранения макрокоманд, предназначенных для автоматической обработки больших текстовых корпусов.

5. Примером применения специализированной базы данных для индивидуальных лингвистических исследований может служить БД «Аспекты словообразования», в структуру которой входят: основная таблица, вспомогательные (подстановочные) таблицы со специализированными перечнями лингвистических критериев; запросы в соответствии с задачами исследования; вспомогательных электронных таблиц; меню для запуска: специализированных форм (для ввода ЛЕ) и запросов для извлечения и просмотра лингвистических данных; коллекцией ярлыков для быстрого запуска объектов БД; главного кнопочного меню для упрощения поиска и запуска лингвистически специализированных объектов БД.

Апробация и внедрение результатов исследования в практику. Содержание диссертации изложено в 8 публикациях общим объемом 16 п.л. Отдельные этапы исследования обсуждались на научных конференциях, статьи по темам выступлений опубликованы в материалах межвузовских научных конференций «Проблемы теории и практики преподавания научных конференции иностранных языков. Краснодар, КВАИ, 2002, 2003, 2004, 2005»; сборника научно-методических статей с материалами научно-методической конференции Армавирского лингвистического университета в феврале 2006 г. Получен отзыв о результатах тестирования базы данных «Словообразовательные аспекты коллоквиализмов» на факультете иностранных языков Армавирского Государственного педагогического университета.

Объем и структура исследования. Композиция диссертации соответствует целям и задачам исследования и состоит из введения, трех глав, заключения, библиографии (всего 120); в качестве приложений –

иллюстрированный перечень объектов базы данных и CD-ROM с дистрибутивом лингвистической базы данных «Словообразовательные аспекты коллоквиализмов» объемом около 7 МБ. Приводим перечень глав: Глава I. «Краткий историко-научный обзор развития компьютерной лингвистики»; Глава II. «Категории общего и особенного в лингвистической базе данных как частного случая реализации теории баз данных»; Глава III. «Оптимизация структуры БД для задач индивидуального лингвистического исследования».

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **Введении** обоснованы актуальность темы исследования, научная новизна, теоретическая и практическая значимость, определены предмет и объект исследования, методы, научная база, основная гипотеза, сформулированы цель, задачи и выносимые на защиту положения. Указаны данные об апробации результатов и структуре диссертации.

В **первой главе** «Краткий историко-научный обзор развития компьютерной лингвистики» освещены основные этапы развития и современное состояние субдисциплин, направлений и методов компьютерной лингвистики, обобщается опыт применения больших вычислительных систем в некоторых из указанных далее областей лингвистических исследований, делается предварительный вывод о влиянии соотношения «уровень сложности актуальной научно-лингвистической тематики / вычислительные ресурсы ПК (персональных компьютеров)», а также человеческого фактора лингвиста-исследователя» на выбор областей для лингвистических исследований с учетом указанных ресурсов и конкретных условий индивидуальной НИР.

Задачи прикладной лингвистики в гуманитарных науках на современном этапе, по мнению одного из ведущих прикладников Гердт А.С.[1], в самом общем виде следующие: «Автоматизация научно-

[1] http://www.phil.pu.ru/lib/ — статья «Предмет и основные направления прикладной лингвистики».

исследовательских работ в гуманитарных науках должна оптимизировать: а) поиск литературы предмета на разных языках; б) подбор источников, фондов материалов; в) оперирование данными источников; г) многоаспектную, глубоко эшелонированную классификацию материала; д) создание сводных описаний, реестров, каталогов по заранее заданным параметрам; е) применение методов статистики, картографии, теория классификации и системного анализа; ж) графическое представление данных в виде схем, рисунков, карт». К разработанной нами БД относится п.п. б) и г).

Судя по библиографии, наиболее востребованными областями хотя бы частично автоматизированных лингвистических исследований являются компьютерная лексикология и компьютерная лексикография и тесно связанная с ними корпусная лингвистика, машинный (автоматический) перевод.

История компьютерной лингвистики началась непосредственно после создания первых экземпляров электронно-вычислительной машины (компьютера). Сразу же они нашла применение в такая специфической области, как шифрование и дешифрование данных, машинный перевод (автоматический перевод), автоматизация лексикографических работ и др.

Пионером в области автоматизированной лексикографии стоял итальянский ученый Р. Буза. С помощью ЭВМ были составлены словоуказатели к древним рукописям. В 1959 г. в г. Безансоне (Франция) был основан Центр лингвистических исследований. На ЭВМ были созданы картотеки (читай: базы данных) для существующих больших словарей. В 1960 в Нанси (Франция) был создан специальный Исследовательский центр для создания с помощью ЭВМ т.н. Сокровищницы французского языка. В 1964 г. в Академии делла Круско (Флоренция). Другие подобные научные и университетские центры Лейден (Нидерланды), Гётеборг (Швеция), Будапешт (Венгрия), Вашингтон (США). Из них наиболее впечатляющих проектов – машинная картотека французского языка

(вместе с терминами – 800 000 слов). В начале 80-х г.г. в Мангейме (Германия) подобные работы (Lexicographical Data Base for German) развернулись на уже более современном аппаратном и программном обеспечении. Другие проекты: индексы (словоуказатели) к древним (напр. Библии – г. Осло.) и современным литературным произведениям, конкордансы (Флоренция, Турин) и частотные словари (Гётеборг, Копенгаген), Этимологический словарь итальянского языка (Саарбрюккен - Германия). Словарные базы данных в европейских научно-лингвистических центрах создавались в качестве электронной версии бумажных словарей, напр.: Машинный словарь итальянского языка. Проект BONNLEX (машинный словарь немецкого языка) интересен как словарный банк лексической базы данных немецкого языка LEDA. Шведская логотека – г. Гётеборг. Система словарных данных японского языка и Система для англо-японского словаря – г. Киото. Электронная версия Оксфордский словарь английского языка создана в университете Ватерлоо (Канада). По аналогии созданы электронные версии словарей английского языка Хорнби, Коллинз, Оксфордского словаря идиом, Оксфордского словаря цитат, фонетического словаря Джоунз.

Наряду с общими национальными словарями велась разработка терминологических баз данных. Получены первые результаты (Документ ISO 12616.2 ресурс Интернет http://www.iso.ch/). Некоторые национальные проекты: Лексикографическая информационная система LEXIS (Федеральное языковое бюро ФРГ). EURIDICAUTOM (банк терминологических данных Бюро терминологии Комиссии европейских сообществ в Люксембурге).

Текстовые БД (корпусная лингвистика) образуют особую категорию лингвистических БД. Корпус текстов есть множество текстов естественного языка, организованное для изучения конкретных языковых аспектов или прикладных задач (составление отраслевых частотных словарей, например). Примеры: Боннский корпус газетных текстов.

Фрейбургский корпус текстов. Проекты для универсальных целей: Брауновский корпус английских текстов (Брауновский университет, США, 1962-63). Текстовой корпус Ланкастер-Осло-Берген. Текстовой корпус LIMAS (Институт исследований проблем коммуникации и фонетики при Боннском университете). Корпус Хауза (250 тыс. словоупотреблений). Корпус текстов для американского словаря Хартвига Даля по разговорной речи американского английского языка. Он получен в результате расшифровки магнитозаписей. Лондонско-Лундский корпус текстов (1979 г.) представляет собой комбинацию текстов из письменной и устной речи. Банк английского языка в Бирмингемском университете, Великобритания (начало 1980-х г.г.) – его несомненным плюсом является выход за рамки 1 млн словоупотреблений. Источником для корпуса TEPL. являются школьные учебники. Объем – 1 млн словоупотреблений.

Машинный (автоматический) перевод. Наибольший интерес к машинному переводу с максималистскими, идеалистичекими ожиданиями характерен для периода 1955 – 1965 г.г. (в СССР – Бельская И.К.; Нелюбин Л.Л., Рябцева Н.К., Марчук Ю.Н., Котов Р.Г., Пиотровский Р.Г.; в Германии – Bruderer H., США – Hutchins W.J.). Естественный язык оказался более сложным явлением, чем казалось энтузиастам, что ясно показали аналитические возможности любых ЭВМ (энциклопедия Britannica, статья Computational linguistics). Поддержка исследований в этой области несколько сократилась, но наработки применяются для автоматизированного лингвистического анализа (определение авторства).

Квантитативная типология и конфронтативная лингвистика. Как известно, во всякой дисциплине столько науки, сколько в ней математики. Поэтому количественный (квантитативный) анализ позволяет объективизировать закономерности типологии языка, полученные. в частности, при конфронтативном методе лингвистических исследований (Арапов М. В. Алексеев П. М., Арапов М. В., Херц М. М.)., Джозеф Гринберг (США).

Лингводидактика. Другое употребительное наименование этой прикладной дисциплины – Computer Assisted Language Learning (CALL). Ее статистический компонент занимается квантитативным анализом процесса обучения. Первые опыты обучающего лингвистического автомата — ОЛА (термин Р. Г. Пиотровского), относятся еще к периоду до появления персональных компьютеров (60-х гг. в США в Стэндфордском университете (Russian-Program) и Нью-йоркском университете (Das deutsche Programm). Среди подходов лингводидактики сущесвтуют:

Бихевиористский подход (упражнения подстановочного типа с заранее жестко заданной структурой): анализ и оценка ответов обучаемого со стороны ОЛА (А. Мензель (Академия наук Берлин); развитие баз данных и баз знаний, позволяющих повысить интеллектуальные возможности систем ОЛА (Д. Миндт (Университет Западного Берлина); разработка компьютерного тестирования (П. Дункель (Пенсильванский университет, США).

Когнитивно-интеллектуальный подход (создание универсального программного обеспечения для CALL (CALL-Software) на базе опыта, полученного при разработке различных форм автоматической переработки текста). В СНГ центрами разработки CALL в 1990 г.г. являлись и продолжают вести лингводиактические исследования: Казань (КГУ), Минск (БГЛУ), Москва (РГПУ). Когнитивно-интеллектуальный подход в разработке автоматических учебных словарей России представлен : Г. В. Дроздецкой и др. (ИРЯ им. А. С. Пушкина); Н. А. Обносовой и К. Р. Галиулиным (Казанский университет), К. Р. Пиотровской (РГПУ) и др.: Х. Б. Масляева (Казанский университет) – программный комплекс коррекции произношения с обратной связью на дисплее. Лексико-фонетические курсы с аудиоподдержкой разрабатывались П. А. Скрелин (ЛГУ), Л. В. Златоустовая и др. (МГУ) Р К. Потаповой (Московский ЛУ). Продолжаются усилии в области распознавания речи (Зиновьева, Н.В., Кривнова О.Ф. Кейтер Дж, Кузнецов В.И., Скрелин П.А. один из пионеров

Фант Г.).

В результате проведенного автором историко-научного обзора сформулирован вывод о том, что к наиболее реальным, перспективным и посильным направлениям компьютерной лингвистики с точки зрения материальных, временных и человеческих ресурсов можно отнести следующие: лингвистическая лексикография и терминография; компьютерная лексикология и лексикография; терминологическая и статистическая лексикография и терминография; вероятностные и статистические модели языка и речи; статистическая семантика и статистическая стилистика; квантитативная типология; компьютерные отраслевые словари-минимумы; специализированные лингвистические базы данных; статистическая и машинная лингводидактика; компьютерная лингводидактика. Некоторые другие подобласти компьютерной лингвистики, по нашему мнению, малопригодны для научно-исследовательской деятельности индивидуальных лингвистов в силу высокого «ресурсного порога», о котором мы говорили выше. Второй вывод по выбору одной из подобластей компьютерной лингвистики: тема индивидуальной НИР в области компьютерной лингвистики должна быть возможно более прикладной, «узконаправленной».

Во **второй главе** «Категории общего и особенного в лингвистической базе данных как частного случая реализации теории баз данных» рассматриваются тенденции, сложившиеся в современной теории и практике баз данных, описываются специфика различных моделей БД.

Существует несколько таких моделей: графовые модели (или иерархические), семантические сети, модель "сущность-связь". Сначала стали использовать иерархические даталогические модели БД Что касается лингвистических БД, то именно на этом принципе целесообразно выстраивать иерархию пользовательских субменю в интерфейсе различных программ, а также для тематического упорядочения источников и разработанных материалов по НИР по виртуальным папкам (темы, их разделы, назначение и т.п.).

Пример иерархической модели в системе субменю лингвистической (морфологической) БД

Морфология

↓

1. Части речи (перечень) → Глагол → Перечень категорий (1. Залог, 2. Наклонение, 3. Время, 4. Аспект, 5. Лицо, 6. Число, 7. Неличные формы) → Рубрикация внутри каждой категории.

2. Морфологические средства (перечень)

…

С точки зрения философских категорий общего и особенного, общим для любых иерархических БД и для лингвистических БД, основанных на этой модели, является принцип «вложенности» подчиненных объектов БД в объект на один уровень выше. Особенным для лингвистических БД являются; а) нецелесообразность распространения этой модели на всю структуру БД в целом, ввиду неприемлемой для лингвиста-пользователя сложности управления и настройки; б) иерархичность языковых категорий весьма относительна, поскольку сама сущность естественного языка разнопланова. Отчасти такая негибкость иерархической модели БД может быть компенсирована элементами сетевой модели (гиперссылки), применяемых в лингвистической БД.

В теории БД отмечается крайняя сложность разработки и высокая вероятность логических ошибок разработчика сетевой модели БД.

Реляционная модель БД является ныне наиболее распространенной и соответствующей современным возможностям аппаратно-программного обеспечения. Теорию реляционных баз данных (← relation – отношение, связь) разработал американский математик Е. Кодд. Она зиждется на нескольких ключевых понятиях: информационный объект, реквизиты (= атрибуты) информационного объекта, нормализация отношений, тип связи, инфологическая (информационно-логическая) модель. Информационный объект – есть совокупность имени и реквизитов некоторой сущности разной степени абстрагированности (предмет,

явление, процесс, событие)[1]. В прикладной лингвистике такими объектами могут быть лексемы в начальной форме, словоформы, словосочетания, синтаксические конструкции, предложения, высказывания, микродиалог, текст. Реквизиты (= атрибуты) информационного объекта являются элементами описания информационного объекта. С точки зрения лингвистической семантики (в широком смысле), указанные реквизиты могут соответствовать или семам в компонентном анализе, или принадлежностью ЛЕ семантическому классу, или даже лексической теме (в лингводидактическом смысле). В грамматической базе данных реквизитами (= атрибутами) являются грамматические категории разного уровня. В таблицах БД им соответствуют поля (они же столбцы). Информационные объекты с одинаковым набором реквизитов объединяются в класс, которому присваивается имя, напр. Части речи. Класс информационных объектов физически представлен специализированной таблицей БД, обычно с тем же именем. Экземпляры информационных объектов соответствуют записи базы данных. Запись БД представляет собой строку основной таблицы.

Особенным в реляционной БД для лингвистических исследований является необходимость дополнять вспомогательные таблицы по мере изучения объекта исследования: списки семантических классов лексем, список словообразовательных моделей и др.

Типов связи таблиц БД – три: 1) один к одному, 2) один ко многим и 3) многие ко многим. Первый тип у нас практически не представлен. Но им может быть: «один фонетический признак – одно место образования фонемы». Пример типа «один ко многим» в нашей БД: Одна часть речи – много лексических единиц. «Многие ко многим»: каждый из множества примеров лексем может иллюстрировать много аспектов словообразования, каждый из множества аспектов словообразования представлен многими примерами лексем, в т.ч. в контексте.

[1] Определение наше (– автор).

Вывод: частнолингвистическая база данных однопользовательского типа имеет много общего с обычными реляционными БД. Особенное: а) необходимость внесения модификаций в ходе исследования б) наличие элементов иерархической БД (меню, субменю, иерархия виртуальных папок) и сетевой БД (гиперссылки к нужным данным за пределами основной БД).

В **третьей главе** в конкретном плане рассматриваются способы оптимизации при разработке структуры лингвистической БД (стадиально).

I. Определение тем и подтем (классификация и рубрикация)

Тема: Словообразование коллоквиализмов. Внутри темы определяем подтемы, т.е. перечень способов словообразования. В отличие от обычного оглавления, в каждом пункте будем указывать краткое наименование (как иногда выражаются в классификациях) «материнской рубрики», т.е. рубрики на один уровень выше. В табл. 1 демонстрируется часть такой рубрикации.

Табл. 1: Классификация коллоквиальных существительных по грамматическим и лексико-семантическим аспектам словообразования

1) Способы словообразования – Аффиксация – Этимология: Исконнные/Заимствованные – Исконные
2) Способы словообразования – Аффиксация – Этимология: Исконнные/Заимствованные – Заимствованные
3) Способы словообразования – Аффиксация – Наличие видов аффиксов – Префиксы только
4) Способы словообразования – Аффиксация – Наличие видов аффиксов – Суффиксы только
5) Способы словообразования – Аффиксация – Наличие видов аффиксов – Префиксы+Суффиксы
6) Способы словообразования – Аффиксация – Продуктивность – Продуктивные
7) Способы словообразования – Аффиксация – Продуктивность – Непродуктивные
8) Лексико-тематический аспект (позиция в тезаурусе) – Результирующая семантика компонентов – (по аналогии).

II. Трансформация классификационных рубрик в структурные компоненты БД.

1) Главная и вспомогательная таблицы. Встроенные списки подстановки.

Определив перечень таких классификационных рубрик, создаем главную таблицу tblMain, записи которых представляют собой коллоквиальную лексическую единицу (ЛЕ разговорного стиля) вместе с лингвистическим описанием, степень углубленности которой зависит от задач исследования. Рубрикации можно подразделить на две категории: конечный (= нижний) уровень и прочие, более высокие уровни (принцип «матрешки»). Имена полей, как отражение рубрик, могут быть слегка перефразированы для сокращения длины имени. Для нижнего уровня рубрикации следует создать вспомогательные таблицы подстановки.

Таблица 1: Вспомогательная таблица со списком подстановки

tblThesaurPositionMainOrDetermComponent
Semantics
0_НЕ_СЛОВОСЛОЖЕНИЕ_0
Абстр_Время_16
Абстр_Здоровье_36
Абстр_ИскусствоЛитер_39
Абстр_Качество_34
Абстр_Количество_35
Абстр_МистикаФантазия_19
Абстр_ПриродаМетео_21
Абстр_ПространствОтнош_33
Абстр_ПроцессРезультатМероприятие_17
Абстр_Религия_20
Абстр_ЧувствоМысльХарактеристика_18
ИмяСобирательное_Неодуш_41
ИмяСобирательное_Одуш_40
КонкрНеодуш_Валюта_38

Если список критериев (= рубрики нижнего уровня) невелик, то целесообразно создать встроенный список подстановки. Приводим часть таблицы подстановки (далее по аналогии):

Табл. 2: Встроенный (в главную таблицу) список подстановки

"НеСловосложение_0";"АгглютСловосложБезСвязЭлем_1";
"АгглютСловосложАфф_2";"АгглютСловосложАббрев_3";
"МорфологСловослож_4";"СинтактСловослож_5".

2) Запросы на выборку

Далее создаваем запросы на выборку (их много). Маркировка _ знак подчеркивания с номером дает возможность в последующем легко создавать запросы с этим критерием. Напр. запрос для выборки ЛЕ с определяемым компонентом на тему ВРЕМЯ имел бы критерий _16. Запрос для выборки всех абстрактных ЛЕ имел бы критерий Абстр_*. Вот пример **имен запросов**:

qryThesaurResultSemantConcretPersonRelationOtherPerson

Имена достаточно красноречивы, легко определить назначение запросов.

3) Формы для ввода данных

Формы обеспечивают удобство ввода данных и просмотра ЛЕ (по одной ЛЕ с одним или атрибутами лингвистического описания, в зависимости от конструкции формы). Пример имен форм: frmCompositionModel.

4) Тематически группированные ярлыки с русскими псевдонимами запросов и форм

В левой части окна имеется раздел Группы с системным значком папки Избранное. Разработчик может создать и пользовательские тематические папки, напр. Аспект категориального моделирования композитов. Разработчик может дать русский внятный псевдоним назначения соответствующего запроса). Пример: Форма Модель словосложения по частям речи.

5) Главная кнопочная форма для запуска объектов БД

Для облегчения доступа к объектами БД разработчик может создать Главную кнопочную форму с набором кнопок, открывающих другие

узкоспециализированные страницы той же кнопочной формы. Напр.: Продуктивность аффикса).

6) Пользовательское меню

В строке меню можно поместить пользовательское меню с подменю, которые будут иметь столь же профессиональный вид, как привычные встроенные меню Файл, Правка, Вид и др. В нашем случае – Словообразование коллоквиализмов. Пример подменю:

Запросы по словосложению – Категориальная модель композита – (конечные команды запуска с именами всех запросов на все модели словосложения)

Вывод

1. Электронная база данных обеспечивает лингвиста-исследователя достаточно обширным инструментарием ввода данных, а также их сортировки, группировки (= систематизации), промежуточных и итоговых вычислений, в сумме неизмеримо превышающих возможности традиционных картотек.

2. Структура БД должна соответствовать поставленной задачам лингвистического исследования. Структура БД состоит из основной и вспомогательных таблиц и встроенных списков подстановки; запросов выборки данных по всем критериям, соответствующим применяемой классификации исследуемого материала; форм ввода данных для обеспечения удобства исследователя-пользователя при вводе и выводе данных; главной кнопочной формы для удобства запуска специализированных форм (согласно классификации материала); Пользовательского меню запуска всех объектов лингвистической БД.

3. Пользовательский интерфейс должен быть рассчитан на исследователя-гуманитария, быть интуитивно понятным и самоочевидным. Разработчик должен снимать прогнозируемые затруднения исследователя-пользователя с помощью всех системных средств подсказки.

Заключение**.** В заключительном разделе резюмируются итоги исследования и намечаются дальнейшие пути частнолингвистических исследований с применением методов компьютерной лингвистики, напр. проблемы разработки баз данных словообразовательных моделей частей речи по узкоотраслевым подъязыкам.

Основные идеи данного исследования в области компьютерной лингвистики, в частности, лингвистических баз данных, изложены в следующих публикациях общим 3 п.ч.

1. Клочко А.Д. Опыт разработки электронной лексико-грамматической базы базы данных для повышения качества знаний по иностранным языкам. В сб. Материалы межрегион. науч. конф. «Развитие внутривузовских систем обеспечения качества образования», Армавир: АГПУ, 2004.
2. Клочко А.Д. Проблемы оптимизации структуры учебника по иностранным языками для военных авиационных институтов к условиям обучения. В сб. Материалы межвуз. науч.-практ. конф. «Теория и практика обучения иностранным языкам». Краснодар: КВАИ, 2002.
3. Клочко А.Д. Место мультимедийных презентаций на занятиях по иностранному языку. В сб. Материалы межвуз. науч.-практ. конф. «Теория и практика обучения иностранным языкам». Краснодар: КВАИ, 2005.
4. Клочко А.Д. Профилактика логических ошибок при разработке мультимедийных обучающих презентаций по иностранным языкам В сб. Материалы межвуз. науч.-практ. конф. «Теория и практика обучения иностранным языкам». Краснодар: КВАИ, 2005.
5. Клочко А.Д. Маркировка английских многокомпонентных терминов в текстовом массиве как очередной этап разработки учебного отраслевого словаря с применением ПК. В сб. Материалы межвуз. науч.-практ. конф. «Теория и практика обучения иностранным языкам». Краснодар: КВАИ, 2003.
6. Клочко А.Д. Базы данных. Уч. пособие для преподавателей гуманитарных специальностей. Армавир: АЛУ, 2005
7. Авдеева Т.Б, Клочко А.Д. Сложные коллоквиальные существительные в английском языке. Практикум по лексикологии и словообразованию. Армавир: АЛУ, 2006. – 180 с.
8. Фетисов О.В. Клочко А.Д. Средства выражения причинно-следственных отношений в английском языке. Практикум по теоретической и практической грамматике. Армавир: АЛУ, 2005. – 60 с.

КЛОЧКО Алексей Данилович

ОПТИМИЗАЦИЯ ИНДИВИДУАЛЬНЫХ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ СРЕДСТВАМИ СПЕЦИАЛИЗИРОВАННОЙ БАЗЫ ДАННЫХ

10.02.21 – прикладная лингвистика

Отпечатано в, подписано в печать 00.09.2006
Формат 84x108 1/32. Объем 0,0 п/л. Тираж 100 экз.
Заказ № 00
Армавирский лингвистический университет,
170013, г. Тверь, ул. Желябова 33.