
Н.А. Мишанкина

БАЗЫ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Статья посвящена проблемам структурирования семантических областей, параметрического описания значения слова при представлении семантики лексических единиц в формате базы данных. Технологии баз данных активно применяются в лексикографической практике. Создание лексикографических баз данных, содержащих семантическое описание единиц определенных семантических областей, необходимо для разработки систем автоматического анализа естественного языка, систем машинного перевода.

Ключевые слова: прикладная лингвистика, лексикография, лингвистические базы данных, лексико-семантическое поле «Звучание».

Современная лингвистика характеризуется направленностью многих исследований на решение прикладных задач. Подобное положение дел связано с необходимостью ответа на вызовы внешней среды: информационные запросы общества, решение проблем коммуникативного взаимодействия, разработку информационных и телекоммуникационных новых технологий.

В настоящее время изменяются представления о роли информации в жизни как отдельного человека, так и общества в целом. Информация приобретает статус ресурса на первых этапах развития информационного общества, и эта ситуация неизбежно влечет за собой необходимость хранения больших объемов информации, что, в свою очередь, приводит к возникновению новой проблемы, связанной с ее организацией и систематизацией.

В науке, и в частности в лингвистике, потребность в создании информационных систем связана с тем, что широкий доступ к необходимой исследователю информации позволяет осуществлять более результативные действия, дает возможность эффективно управлять информационными потоками. Именно поэтому актуальной задачей становится разработка междисциплинарных методик, позволяющих осуществлять аналитическую обработку информации, и поиск способов ее структурирования для эффективного хранения и обработки. Оптимальными в этой связи представляются технологии **баз данных**

(БД), позволяющие создавать структурированные массивы данных, хранимые и управляемые с применением компьютерных технологий, используемые для создания и функционирования эффективных информационных систем [1–5].

Цель настоящей работы – провести обзор исследований в области лингвистических баз данных и рассмотреть возможности использования данной технологии в лексикографических проектах, а также представить вариант подобного проекта в виде лексикографической базы данных, отражающей лексику русского языка с семантикой звучания.

Технология БД используется в процессе создания как традиционных, так и электронных словарей. Активно разрабатываются словарные БД специальной и терминологической лексики [1, 3, 6]. Для разработки этой технологии в лингвистике и создания ресурсов подобного типа необходимо решить следующие задачи: 1) задачу структурирования и первичного анализа эмпирического материала, начиная от фиксации единиц языковых уровней (грамматиконы, словари, фонетические базы данных) до фиксации целостных текстов, что позволяет, с одной стороны, дополнить и уточнить структурную модель языковой системы, а с другой – сформировать функциональные модели дискурсивных областей и модель языковой системы в целом; 2) задачу поиска новых способов фиксации и сохранения языковых данных, а также организации доступа к этим материалам; 3) задачу поиска новых методов обработки материала для оптимизации исследования и получения новых результатов; 4) задачу верификации результатов исследования за счет обращения к материалу большего объема. Например, Е.И. Ярославцева отмечает, что компьютерная база данных «Языки мира», представляющая собой универсальный грамматикон, может быть использована в лингвистической типологии для верификации гипотез о типологической близости языков [7. С. 356].

В России разработка теоретических основ создания баз данных впервые была осуществлена в рамках создания информационно-поисковых систем для информационно-библиографического поиска. Работа над лингвистическими ресурсами была начата в связи с созданием Машинного фонда русского языка. В современной лингвистике компьютерное представление информации на основе технологии БД используется в самых разных областях: в типологических и сравни-

тельных, фонетических, лексических, грамматических исследованиях; лексикографии; семантических исследованиях, включая составление тезаурусов; в компьютерной лингвистике и для решения прикладных лингвистических задач (лингводидактических, автоматизированного перевода, автоматического распознавания и синтеза речи).

В настоящий период развития этой технологии в сфере лингвистики можно говорить о двух основных типах БД:

1) полнотекстовые базы представляют собой документальные БД, в которых представлены целостные тексты, относящиеся к какой-либо дискурсивной сфере. Примером такого ресурса может выступить БД «Среднеобский фольклор» (<http://mion.tsu.ru/song>), созданная в рамках развития информационной системы «Межрегиональные исследования в общественных науках» [8];

2) собственно БД являются информационными системами фактографического типа и содержат структурированную информацию о лингвистических единицах различного рода. Например, «Хронологический морфемно-словообразовательный словарь русского языка», БД по русским прилагательным «EDGE», база фонетических данных и др. [1]

Несмотря на активные и продуктивные исследования, работа в этой области далека от завершения, т.к. объект лингвистики – язык – принципиально безграничен и многомерен в силу целого ряда специфических свойств: элементы языковой системы *неоднородны* и количественно необозримы; языковые структуры *конструктивны*, т.е. множество порождаемых высказываний, адаптированных для выражения самых сложных ментальных построений, потенциально бесконечно; языковой знак *динамичен* – в процессе функционирования происходит трансформация его формальной и содержательной сторон, что влечет за собой порождение вариантов знака или новых элементов системы; языковая деятельность тесно связана с коммуникативной и приспособляется к выражению самых сложных коммуникативных смыслов; языковые структуры связаны с когнитивными процессами и в силу этого языковая семантика достаточно сложно формализуется.

Каждое из названных свойств говорит, с одной стороны, о неограниченном информационном потенциале языковой системы, а с другой – о том, что эта информация необычайно сложно структурируется даже на уровне отдельных языковых элементов. Однако необходимость формального представления лингвистической инфор-

мации для решения прикладных задач ведет к увеличению исследований в этом направлении.

Можно говорить о теоретических и прикладных перспективах использования лингвистических баз данных. Теоретические перспективы видятся в использовании технологии БД в лингвистических, и в частности лингвокогнитивных, исследованиях. Применение данной технологии в исследовании когнитивной деятельности связано с изучением концептуализации действительности. В данном случае на первый план выступают принципы представления информации в базах данных, принципы, лежащие в основе когнитивного моделирования данных, формирования «Моделей-онтологий» – «комплексных разносторонне формализованных представлений предметной области...», синтезирующих «...лингвистические, тезаурусные, понятийные, энциклопедические и процедурно-декларативные специальные знания» [9. С. 53].

БД, ориентированные на описание лексической семантики, позволяют получать данные о способах концептуализации действительности в том или ином языке, используемые в системах текстового анализа и машинного перевода, в лингводидактике и т.п. С другой стороны, в современных работах по моделированию лингвистических тезаурусов и БД [1, 3, 6, 10, 11, 12 и др.] говорится о перспективах использования естественно-языковой категоризации в логико-лингвистическом концептуальном моделировании. Р.Ю. Кобрин указывает на то, что в настоящее время существует два подхода в создании банков данных: а) построение языковых моделей предметных областей; б) построение алгебро-логических баз данных. Однако в решении прикладных задач доминирует второй подход. При всем том автор указывает, что лингвистическое моделирование может выступать в качестве основы концептуального. В частности, такую задачу он решает на материале лингвистического анализа терминологии, устанавливая систему семантических отношений в рамках заданной области [10]. Попытки решения данной проблемы представлены и в работах А.Н. Баранова [6], А.С. Герда [3], С.Е. Никитиной [12].

Подобный подход представляется перспективным в свете современных теорий концептуализации, в основе которых лежит антропный принцип, находящий последовательное отражение в естественном языке. Например, БД, отражающая лексико-семантическое поле

«Звучание», позволяет представить в структурированном виде данную концептуальную область, специфичную для носителей русского языка, т.к. для представления материала в такой БД необходимо провести тезаурусное структурирование названного лексико-семантического поля и параметризацию семантики лексических единиц, формирующих эту область.

Таким образом, создание лингвистических информационных систем в виде баз данных – задача насущная и актуальная. Но вместе с тем и весьма сложная. Рассмотрим проблемы, возникающие в процессе концептуального проектирования лексикографических баз данных, на примере БД «Звучание».

Первый этап создания любой базы данных связан с формированием системы информационных задач, на решение которых направлено создание информационного ресурса. От этой системы напрямую зависит структурирование предметной области, получающей отображение в БД, и организация ее инфологической схемы.

В случае с лингвистическими ресурсами кроме информационных потребностей следует учитывать и то, что в качестве предметной области здесь выступает фрагмент языковой системы [13], поэтому необходимо учитывать при формировании структуры уже разработанные в лингвистике описания и классификации. В частности, при обращении к лексике с семантикой звучания необходимо, во-первых, установить границы лексико-семантического поля и, во-вторых, исследовать его структуру с точки зрения уже существующих в этой области работ и оценить возможности ее представления в БД.

Границы материала в этом случае могут быть установлены как минимум по двум основаниям: с одной стороны, в БД могут быть отражены единицы, непосредственно маркирующие типы звучания, пусть даже в синкретичном виде (например, глагол *бахнуть* одновременно указывает на совершение физического и звукового действия). С другой стороны, в БД могут получить отображение все единицы, содержащие семантический компонент «звучание», и в этом случае в выборку попадают и единицы, маркирующие музыкальные инструменты, и обозначающие человека по звучанию (*свистун*).

Еще одна проблема в отборе материала связана с активной полисемией в этой области: включать ли в БД единицы с переносным значением? От ответов на все эти вопросы непосредственно зависит структурная и функциональная специфика БД, ее информационный

потенциал. Полагаем, что для реализации максимально полного информационного потенциала необходимо включение в проектируемый ресурс широкого спектра единиц, но с учетом их неравноценного положения. Лексические единицы для представления в БД были исследованы в диссертационной работе [14] – это единицы литературного языка, содержащие в семантике компонент «звук».

Результаты проведенных ранее исследований [14–18] позволяют представить следующую модель лексико-семантического поля «Звучание»: а) ЛСГ звукоподражаний (*бум, бабах, ку-ку, мяу*); б) ЛСГ глаголов звучания (*греметь, пищать, кричать*); в) ЛСГ имен звучания (*треск, вой, голос*); г) ЛСГ имен действия, результатом которого является звучание (*пение, гоготание, верещание*); д) ЛСГ имен лица по звучанию (*ворчун, стрекотуха, трещотка*); е) ЛСГ имен артефактов по звучанию (*свисток, пищалка*); ж) ЛСГ имен животных (*квакушка, мурлыка*); з) ЛСГ звуковых признаков (*громкий, звонкий, тихо, гулко*). Ядром БД, как и ядром лексико-семантического поля, выступает глагольная лексика, отражающая динамический аспект феномена звучания. Именно глагольная лексика представляет собой базовый объект описания, все остальные объекты (ЛСГ) являются в данном поле в той или иной степени производными от глагольной лексики.

Следующим этапом создания БД является проектирование дата-логической схемы БД, представляющей собой систему таблиц, поля которых отображают атрибуты описываемых объектов. В данном случае в качестве атрибутов выступают семантические, грамматические, стилистические параметры отдельных лексических единиц.

Сложность параметризации семантики естественных языковых единиц отмечают практически все исследователи. Отечественные и зарубежные работы в области семантики предлагают широкий спектр методик и приемов, но полное описание плана содержания языковых единиц остается проблемой, которая не решена по сей день.

Традиционная лексикографическая практика позволяет опираться на разработанные и апробированные образцы. Как и в традиционной словарной статье, при формировании параметров атрибуции единицы в БД выделяются следующие: грамматическая информация – принадлежность к части речи, характерные грамматические формы; стилистические пометы – информация о стилистической маркированности единицы; приводятся контексты функционирования единицы. Но зона толкования или представление собственно семантической информации в

отличие от традиционной формы нуждается в дополнительном структурировании, и ключевым здесь является прием компонентного анализа, позволяющий выявить ядерные семантические компоненты. Периферийные семы определяются на основе анализа переносных ЛСВ, анализа дистрибуции. Структурирование семантики для представления в БД тем более не является исчерпывающим, т.к. в данном случае предполагается унифицированное описание отдельных групп единиц, но при этом необходимо все же ориентироваться на учет максимального количества параметров.

Полагаем, что для единиц ядерной части лексико-семантического поля – глаголов звучания – будут релевантными следующие параметры:

Тип значения (прямое – метафорическое – метонимическое); Грамматическая информация; Форма национального языка (литературный язык – диалект – просторечие – жаргон); Тип звучания; Акустические характеристики звучания; Ситуативные характеристики звучания; Субъект; Характеристики субъекта; Тип действия; Характеристики действия; Коннотативная оценка звучания; Коннотативная оценка субъекта; Коннотативная оценка действия; Эмотивные смыслы; Сочетаемость; Контекст.

Сразу необходимо оговорить, что коннотативная оценка является факультативным компонентом семантики описываемых лексических единиц, поэтому в случае проявленной оценочности фиксируется ее наличие, а в случае неявной – ставится маркер «нейтральная».

Рассмотрим, например, параметрическое описание глагола *жужжать*: Лексема: жужжать; Тип значения: прямое; Грамматическая информация: глагол; Форма национального языка: литературный язык; Тип звучания: звучание насекомых; Акустические характеристики звучания: неголосовое, низкого тона, негромкое, шумное; Ситуативные характеристики звучания: однообразное; Субъект: насекомое; Характеристики субъекта: летающее; Тип действия: движение крыльями при полете; Характеристики действия: однообразное; Коннотативная оценка звучания: негативная; Коннотативная оценка субъекта: нейтральная; Коннотативная оценка действия: назойливое, надоедливое; Эмотивные смыслы: нет; Сочетаемость: жужжит пчела / муха / жук; Контекст: *Вот шмель жужжит около цветка.* (Гончаров).

Фиксация лексико-семантических вариантов возможна за счет введения ключевых идентификаторов в числовом формате, что позволяет описывать их как отдельные записи в таблице, но при этом форма слова, а также указание на тип значения дадут возможность установить их идентичность.

При описании лексико-семантических вариантов данная структура позволяет проследить трансформацию семантики. Рассмотрим описание лексико-семантического варианта уже описанного глагола: Лексема: жужжать; Тип значения: метафорическое; Грамматическая информация: глагол; Форма национального языка: литературный язык; Тип звучания: звучание неодушевленного; Акустические характеристики звучания: неголосовое, негромкое, низкого тона, шумное; Ситуативные характеристики звучания: повторяющееся; Субъект: механизм; Характеристики субъекта; Тип действия: в процессе функционирования; Характеристики действия: однообразное; Коннотативная оценка звучания: нейтральная; Коннотативная оценка субъекта: нейтральная; Коннотативная оценка действия: нейтральная; Эмотивные смыслы: нет; Сочетаемость: жужжит мотор; Контекст: *Жужжа, вентилятор хватает горячий воздух США металлической жаброй* (Бродский).

Кроме того, современные системы управления базами данных дополнены возможностями интеграции мультимедийных объектов. Это позволяет включить в БД данные другого типа: аудио- или видеоматериалы, демонстрирующие тип звучания, его субъект и действие, с ним связанное.

Таким образом, решение проблемы концептуализации и структурирования специализированных семантических областей может рассматриваться как актуальная задача, с одной стороны, прикладной, а с другой – теоретической лингвистики. Создание подобных ресурсов позволит эффективно и быстро получать данные о составе лексических единиц, формирующих определенную семантическую область, об их стилистической и грамматической специфике, о структуре значения, о функционирующих в данной сфере метафорических и метонимических моделях. Полагаем, что ресурс подобного рода будет востребован как в исследовательской работе, так и в учебном процессе. Подобная база данных может быть использована в образовательном процессе как лексикографический источник в преподавании курсов «Лексикография», «Лексикология», «Лингво-

культурология», «Языковая картина мира», «Русский язык как иностранный» и др.

Литература

1. Асиновский А.С., Архипова Е.А., Богданова Н.В. и др. Полевая лингвистическая практика: учеб.-метод. комплекс сложной структуры. Ч. 1: Теоретические основы и методика сбора лингвистических данных для представления их в речевом корпусе русского языка. СПб., 2006.
2. Гайдамакин Н.А. Автоматизированные информационные системы, базы и банки данных. Вводный курс: учеб. пособие. М.: Гелиос АРВ, 2002.
3. Герд А.С. Прикладная лингвистика. СПб.: Изд-во С.-Петерб. ун-та, 2005.
4. Советов Б.Я. Базы данных. М.: Высш. шк., 2005.
5. Хомоненко А.Д. Базы данных. СПб.: Корона-Принт, 2004.
6. Баранов А.Н. Введение в прикладную лингвистику. М.: Корона Эдиториал УРСС, 2001.
7. Ярославцева Е.И. Грамматикон и база данных «Языки мира» // Scripta linguisticae applicatae. Проблемы прикладной лингвистики 2001: сб. ст. М., 2001. С. 339–357.
8. Мишанкина Н.А., Тубалова И.В., Эмер Ю.А. Филология и информатика: специфика электронного представления региональных фольклорных текстов // Гуманитарная информатика: сб. ст. / под ред. Г.В. Можяевой. Томск, 2004. Вып. 1. С. 102–114.
9. Рябцева Н.К. Язык и естественный интеллект / РАН. Ин-т языкознания. М.: Academia, 2005. 640 с.: библи. (Монографические исследования: лингвистика).
10. Кобрин Р.Ю. Лингвистическое описание терминологии как база концептуального моделирования в информационных системах: автореф. дис. ... д-ра филол. наук. Л., 1989.
11. Марчук Ю.Н. Компьютерная лингвистика. М.: АСТ: Восток – Запад, 2007.
12. Никитина С.Е. Семантический анализ языка науки: на материале лингвистики. 2-е изд. М.: Книжный дом «ЛИБРОКОМ», 2010.
13. Мицкевич О.С. Лингвистическая база данных (ЛБД) специальной лексики белорусского языка с точки зрения потенциальных пользователей // Прикладная лингвистика в науке и образовании: сб. тр. VI Междуна. науч. конф., 5–7 апреля 2012 г., Санкт-Петербург. СПб., 2012. С. 203–206.
14. Мишанкина Н.А. Феномен звучания в интерпретации русской языковой метафоры: дис. ... канд. филол. наук. Томский гос. ун-т. Томск, 2002.
15. Васильев Л.М. Семантика глаголов звучания в современном русском языке // Системные отношения в лексике и методы их изучения. Уфа, 1977. С. 3–20.
16. Васильев Л.М. Семантика русского глагола (глаголы речи, звучания и поведения): учеб. пособие. Уфа: Изд-во Башк. ун-та, 1981.
17. Голубева. Е.Л. О семантических особенностях глаголов звучания // Вопр. теории и истории русского языка. Ташкент, 1967. Вып. 294. С. 15–22.
18. Рузин И.Г. Природные звуки в семантике языка: (Когнитивные отражения именованья) // Вопр. языкознания. 1993. № 6. С. 17–28.