

ТЕХНОЛОГИЯ БАЗ ДАННЫХ В СОЦИОГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ

Н.А. Мишанкина

Обсуждается проблемы создания структурированных массивов лингвистической информации. Приводится типология лингвистических БД. Проводится анализ перспектив их применения.

DATABASE TECHNOLOGY IN SOCIOGUMANITARNYH RESEARCH

N.A. Mischankina

Discusses the problems of creating structured arrays of linguistic information. Provides a typology of linguistic databases. Analyses the prospects of their application.

Современный мир может быть назван гиперинформационным. Период общественного развития со второй половины XX – начала XXI в. тесно связан с изменением представлений о роли информации в жизни как отдельного человека, так и общества в целом. Появление и активное развитие новых технологий обработки информации повлекло за собой кардинальное изменение ценностных ориентаций общества – информация как ресурс стала приобретать все большую значимость. Это, в свою очередь, привело к резкому увеличению информационных потоков и формированию банков информации. Необходимость хранения больших объемов информации повлекла за собой такую актуальную проблему как ее организация и систематизация.

Создание информационных систем связано, в первую очередь, с тем, что широкий доступ к необходимой информации позволяет осуществлять более результативные действия в самых разных областях, эффективно управлять информационными потоками. В этой связи актуальной задачей становится разработка междисциплинарных методик, позволяющих осуществлять аналитическую обработку информации и поиск способов ее структурирования для эффективного хранения и обработки.

Наиболее адекватными в этом аспекте представляются комплексные методики, позволяющие создавать **базы данных** – структурированные

информационные массивы, используемые для создания эффективных информационных систем.

Итак, предпосылками разработки технологии и создания баз данных можно считать насущную необходимость в решении следующих задач:

- **Создание информационных систем** (процедуры первичного анализа информации → технологии структурирования и организации информационных массивов).
- **Исследование когнитивной деятельности** (методики исследования концептуализации действительности основываются на принципах представления информации в базах данных и ведут к формированию «Моделей – онтологий» – «комплексных разносторонне формализованных представлений предметной области и знаний... синтез лингвистических, тезаурусных, понятийных, энциклопедических и процедурно-декларативных специальных знаний») [8, с. 53].
- **Повышение достоверности и эффективности** лингвистических (и в целом, гуманитарных) исследований (систематизация и верификация материала), которое непосредственно зависит от доступности научной информации (как первичной, в виде материала для анализа, так и вторичной – результата теоретического осмысления). Основа научного исследования – накопленная эмпирическая информация об изучаемом объекте реального мира. При этом, чем масштабнее исследовательский проект, тем более объемным должен быть подлежащий обработке материал. Поэтому эффективность реализации многих научных лингвистических программ в значительной мере определяется степенью автоматизации обработки данных и организации доступа к ним.

В частности, для фиксации различных проявлений языковой системы возникла необходимость в создании лингвистических ресурсов, направленных на решение задач структурирования лингвистической информации; структурирования и первичного анализа эмпирического материала; поиска новых способов его фиксации и хранения; верификации результатов исследования; поиска новых методов обработки материала для оптимизации исследования и получения новых результатов. Это привело к созданию информационных систем для аккумуляции, хранения и первичной обработки лингвистической информации, таких как словари, лингвистические корпусы и лингвистические базы данных.

60-е гг. XX в. – время, когда начинается разработка различных концепций компьютерного представления и обработки данных. К этому

времени уже существовали развитые и разнообразные информационные системы: библиотеки, архивы. Поэтому автоматизация обслуживания информационных систем была в первую очередь востребована именно в этих областях. Позднее компьютерные базы и архивы стали создаваться и для конкретных исследовательских задач в разных областях филологии.

В России разработка теоретических основ создания баз данных впервые была осуществлена в рамках программы ГАСНИТИ (разработка ИПС) и в рамках работы по Машинному фонду РЯ. Подобные системы были, как правило, закрытыми, т.е. их дополнение и реорганизация требовали избыточного копирования данных и переделки программ обработки. Еще одна сложность, связанная с отсутствием универсальной концепции, – различие в моделях баз данных, разрабатываемых в разных научных центрах. Эти различия вели к тому, что информационные системы не имели уникальных программ обработки, которые могли быть использованы в других системах. Однако значимым свойством для каждой информационной системы является возможность ее укрупнения за счет объединения с другими. Эта проблема стала еще более острой в 70-е гг. с появлением ЭВМ, позволяющих создание банков данных, аккумулирующих данные разных научных центров. Попытки решения привели к созданию новых способов организации информационных систем, реализующих принцип независимости данных от прикладных программ. Она и получила название базы данных. «Представление информации в виде баз данных позволило разработчикам реализовать ряд существенных преимуществ по сравнению с используемыми ранее информационными системами, имеющими файловую организацию данных. Так, в частности, появилась возможность использования одних и тех же данных одновременно для решения многих прикладных задач, а вследствие того что отпала необходимость в их дублировании, была существенно сокращена избыточность в представлении информации» [1, с. 38].

В настоящее время компьютерное представление информации в виде баз данных используется в самых разных областях языкознания и прикладной лингвистики: в типологической и сравнительной лингвистике, 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 гуманитарной сфере можно говорить о следующей типологии.

Среди гуманитарных баз данных можно выделить **общегуманитарные базы**, содержащие информацию, актуальную для многих наук гуманитарной сферы, и **специализированные базы**, ориентированные на определенную гуманитарную область. Предварительно следует оговорить, что разделение на общегуманитарные и специализированные представляется достаточно условным, так как информация, представленная в специализированных базах, вполне может быть использована для исследования в другой гуманитарной области. Но все же следует помнить о том, что создается специализированная БД специалистами одной области знания, при создании решаются узкоспециальные задачи, а способы представления также ориентированы на достижение узкоспециальных целей.

Среди специализированных БД мы можем назвать *исторические, лингвистические, культурологические* базы данных и т.п. *Исторические базы данных* связаны с представлением массовых источников (например, БД архивных материалов по массовым репрессиям), исторических документов, фактов материальной культуры, интересных для историков (экспозиции электронных музеев, исторические архивы) и др.

В лингвистике можно говорить о существовании следующих видов лингвистических баз данных:

Полнотекстовые базы, к которым могут быть отнесены как собственно базы, так и корпуса текстов, например самый крупный в России «Национальный корпус русского языка» или «Компьютерный корпус текстов русских газет конца XX века». А также базы данных, в которых представлены целостные тексты, относящиеся к какой-либо дискурсивной сфере. Они могут представлять собой библиотеки электронных текстов, включающие как произведения художественной литературы, так и критические, исследовательские работы по их изучению. Подобные базы размещены на различных филологических сайтах (портал «Кирилл и Мефодий», Интернет-версия журнала «Philologica», «Ruthenia», «Русская виртуальная библиотека»), а также могут представлять собой отдельный самостоятельный ресурс («Фундаментальная электронная библиотека: Русская литература и фольклор»).

Примером такой базы данных может выступить База данных «Среднеобский фольклор» (<http://mion.tsu.ru/song/>), созданная в рамках развития информационной системы «Межрегиональные исследования в общественных науках».

Материалом для создания этой БД послужили записи текстов народной культуры жителей сибирского и алтайского регионов, в том числе Среднего Приобья, которые интенсивно велись в Томском государствен-

БАЗА ДАННЫХ "СРЕДНЕОБСКИЙ ФОЛЬКЛОР"	
АННОТАЦИЯ ФОТОАРХИВ ПОИСК АВТОРЫ ССЫЛКИ	
Жанр	---- →
Территория	
Год фиксации	-- ↓
Исполнитель	
Пословный поиск	
Поиск	
Расширенный поиск	
Всего записей: 1922	

ном университете с 60-х гг. XX века. Записи проводились как в рамках научных экспедиций, так и во время студенческих фольклорных практик. Архив записей хранится на филологическом факультете ТГУ в виде рукописных тетрадей.

Материал в БД систематизируется по жанровой принадлежности, территории бытования, дате фиксации, имени исполнителя.

Уже первый параметр – жанровый состав – позволяет говорить о специфике сибирского фольклора, так как на исследованной территории полностью отсутствуют былины и исторические песни (по крайней мере, это можно сказать в отношении рассматриваемого периода – с 60-х гг. XX в. по настоящее время). Здесь представлен песенный фольклор (баллады, хороводные, протяжные и др. песни, романсы, песни литературного происхождения и частушки) и загадки. Устная проза представлена жанровыми формами сказок и быличек. Среди сказок в основном встречаются бытовые (реже – сказки о животных). Наблюдается единичная фиксация текстов народной драмы, особой составляющей материала являются тексты наивной литературы. Представлен бытовой обрядовый фольклор (описания свадебного обряда), тексты заговоров [6, с. 110].

Объектом БД «Среднеобский фольклор» выступает отдельный текст, снабженный описанием по указанным параметрам.

БАЗА ДАННЫХ "СРЕДНЕОБСКИЙ ФОЛЬКЛОР"	
Жанр:	Наречная литература
Исполнитель:	Хромов Алексей Игнатьевич, г.р. 1917.
Название:	Утро
Территория:	Парабель, 1963г.
Текст:	Полумрак на дворе, Чуть влечет восток, А вон там в стороне Все горит огонек. Тихо, тихо вокруг В этот упрямый час. Только внизу я вдруг Огонек тот погас. Ближе, ближе рассвет Побелеет небосвод.



8.jpg



Августина
Васильевна
Крохина



Арина Алексеевна
Токарева



Гуси



деревня
Чернокурья



Дом № 1



Кроме текстов, БД содержит фотоархив – уникальные фотоматериалы, сделанные во время записи фольклорных материалов.

Вторая группа – **собственно базы данных**, содержат информацию о лингвистических единицах различного рода. Например, «Хронологический морфемно-словообразовательный словарь русского языка», база данных по русским прилагательным «EDGE», база фонетических данных и др. Технология баз данных используется как в процессе создания традиционных словарей, так и для создания электронных словарей. Активно разрабатываются словарные базы данных специальной и терминологической лексики. Также можно назвать Международный Компьютерный Архив Современного Английского Языка (International Computer Archive of Modern English – ICAME); Летний Институт Лингвистики США (SIL), где формируются базы данных по 850 языкам; Фонетический фонд русского языка, разработанный и пополняемый на кафедре фонетики и Лаборатория экспериментальной фонетики СПбГУ [1].

При всем том исследовательская работа в этой области далека от завершения, так как объект лингвистики – язык – безграничен и многомерен, в силу своих специфических свойств:

- элементы языковой системы *неоднородны* и количественно *необозримы*;
- языковой знак *динамичен* – в процессе функционирования происходит трансформация формальной и содержательной сторон,

что влечет за собой порождение вариантов знака или новых элементов системы;

- языковые структуры *конструктивны*, т.е. множество порождаемых высказываний, адаптированных для выражения самых сложных ментальных построений, потенциально бесконечно;
- 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 формализации лингвистической информации для решения прикладных задач ведет к актуализации исследований в этом направлении.

Можно говорить о следующих перспективах использования лингвистических баз данных.

Теоретические перспективы видятся в использовании технологии БД в лингвистических и, в частности, лингвокогнитивных исследованиях, так как решают проблему поиска, фиксации, хранения и верификации эмпирического материала.

Практическое значение БД видится в решении задач:

автоматизированной обработки и синтеза устной речи (фонетические БД), письменной речи (морфологические БД);

информационного поиска (лексикографические БД). В частности, базы данных, ориентированные на описание лексической семантики, позволяя 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 БД [2, 3, 4, 5, 7 и др.] говорится о перспективах использования естественной языковой категоризации в логико-лингвистическом концептуальном моделировании. Р.Ю. Кобрин указывает на то, что в настоящее время существуют 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 [4]. Варианты решения данных задач представлены и в работах А.Н. Баранова [2], С.Е. Никитиной [7].

Подобный подход представляется перспективным в свете современных теорий концептуализации, в основе которых лежит антропный принцип, находящий последовательное отражение в естественном языке. Например, база данных русских зоонимов позволяет представить в структурированном виде концептуальную область «Зоология», специфичную для носителей русского языка.

Проблемы терминологической эквивалентности (терминологические БД и тезаурусы) связаны с тем, что качество научной деятельности и научной коммуникации во многом определяется спецификой языковых единиц, фиксирующих эпистемологические модели знания.

Базовой единицей концептуализации и структурирования научной области выступает термин, концептуальная система отражена в терминосистеме научной области.

Одним из главных требований к терминологической единице является требование семантической стабильности. Но термины – это не только концептуальные, но и лингвистические единицы и вне зависимости от предъявляемых к ним требований попадают под действие общеязыковых законов. Семантика термина, как и семантика лексических единиц, подвержена действию закона асимметричного дуализма языкового знака, что проявляется как в образовании термина на основе лексико-семантического варьирования, так и в процессах метафоризации. За счет этого терминосистемы обладают определенной динамикой, кроме того, терминосистемы различных языков не являются симметричными, что влечет за собой проблему информационной однозначности в интерпретации термина.

Терминологические базы данных и тезаурусы позволяют проводить эффективный сопоставительный анализ эпистемологических моделей, функционирующих в национальных культурах, находить адекватные эквиваленты и вырабатывать универсальные эпистемологические системы.

Таким образом, решение проблемы и концептуализации и структурирования специализированных предметных областей – актуальная задача когнитивной лингвистики, особенно в части разработки прикладных проектов.

До настоящего времени не получили достаточного освещения проблемы метафоризации терминосистем, в частности, соотношения функцио-

нирования метафорических и неметафорических терминов, взаимодействия и взаимовлияния терминосистем различных научных дисциплин, внутридисциплинарной полисемии термина, а также принципы отбора лексико-семантических единиц при метафорическом терминообразовании.

Применение технологии БД позволит выявить принципы формирования современных терминосистем в разных языках, тенденции их динамики и на этом основании выработать принципы корректирования семантики терминов и терминологического соотнесения при научном переводе.

ЛИТЕРАТУРА

1. Асиновский А.С., Архипова Е.А., Богданова Н.В. и др. Полевая лингвистическая практика: Учебно-метод. комплекс сложной структуры. Часть 1: Теоретические основы и методика сбора лингвистических данных для представления их в речевом корпусе русского языка. – СПб., 2006.
2. Баранов А.Н. Введение в прикладную лингвистику. – М., 2001.
3. Герд А.С. Прикладная лингвистика. – СПб.: Изд-во СПб. ун-та, 2005.
4. Кобрин Р.Ю. Лингвистическое описание терминологии как база концептуального моделирования в информационных системах: Автореф. дис. ... д-ра филол. наук. Л., 1989.
5. Марчук Ю.Н. Компьютерная лингвистика. – М.: АСТ: Восток-Запад, 2007.
6. Мишанкина Н.А., Тубалова И.В., Эмер Ю.А. Филология и информатика: специфика электронного представления региональных фольклорных текстов // Гуманитарная информатика: Сб. статей / Под ред. Г.В. Можяевой. – Томск: Изд-во Том.ун-та, 2004. Вып. I. – С. 102–114.
7. Никитина С.Е. Семантический анализ языка науки: на материале лингвистики. – Изд. 2-е. – М.: Книжный дом «ЛИБРОКОМ», 2010. – 146 с.
8. Рябцева Н.К. Язык и естественный интеллект: РАН. Ин-т языкознания. – М.: Academia, 2005. – 640 с.