***M. Everaert, S. Musgrave, A. Dimitriadis* (eds.). The use of databases in cross-linguistic studies.** Berlin: Mouton de Gruyter, 2009. 409 p.

Рецензируемая книга является коллективной монографией, посвященной новому и быстро развивающемуся направлению лингвистики – созданию и использованию компьютерных баз данных для типологических исследований.

С развитием вычислительной техники появились новые возможности для представления и обработки результатов исследований – базы данных. К настоящему времени в мире созданы десятки больших и малых типологических баз данных (далее ТБД). Вероятно, впервые ТБД начала использовать Дж. Николс [Nichols 1992] еще в начале 90-х годов, но ее база данных существовала только в бумажной форме или в виде набора отдельных файлов. В предисловии к книге приводятся интересные сведения по истории развития ТБД.

Первые компьютерные лингвистические базы данных появились в середине 90-х годов [Liberman 1997; MacWhinney 1995]. Первая коллективная монография по данной теме – работа [Nerbonne 1998]. Первая самостоятельная конференция по лингвистическим базам данных прошла в Пенсильвании в 2001 г. [Buneman et al. (eds.) 2001]. Секция по ТБД была включена в программу XVII Международного лингвистического конгресса в 2003 г. в Праге. Крупный проект по интеграции разноструктурных ТБД выполнялся в начале нашего века при поддержке Евросоюза [Everaert 2003].

Как представляется, интерес к данному инструменту значительно возрос с появлением в 2005 г. «The world atlas of language structures» (WALS) [Haspelmath et al. (eds.) 2005] – первой большой общедоступной ТБД, содержащей описания более чем 2500 языков. В настоящее время в ведущих лингвистических журналах

мира практически ежемесячно появляются статьи по этой проблематике.

В какой-то мере можно считать, что рецензируемая книга подводит итоги первого этапа (10–15 лет) развития ТБД. Следует отметить крупный проект по созданию подобной базы данных, осуществляемый в России в Институте языкознания РАН, – «Языки мира» [Поляков, Соловьев 2006]. К сожалению, данная ТБД до сих пор мало известна на Западе и не нашла отражения в рецензируемой книге. Пожалуй, это наиболее крупный пробел в рецензируемой монографии. Впрочем, в последнее время в этом отношении ситуация меняется к лучшему: статья, сравнивающая проекты «Языки мира» и WALS, опубликована в журнале «Linguistic typology» (см. [Polyakov et al. 2009]), на ТБД «Языки мира» появляется все больше ссылок в западных публикациях.

Рассматриваемая книга состоит из введения и 10 статей, написанных разными авторами, но в совокупности отражающих все наиболее интересные аспекты создания и использования ТБД, и потому книга производит впечатление, скорее, целостной монографии, чем сборника статей.

Общие проблемы создания ТБД обсуждаются в статье А. Димитриадиса, С. Мусграве «Основы создания баз данных для лингвистов» («Designing linguistic databases: A primer for linguists»). К ним относятся: выбор инструментария, структуры данных и некоторые другие вопросы. Авторы делят все ТБД по степени программистской сложности на три категории: «все в одном компьютере» (all-in-one desktop database), «веб-базы данных» и сложные базы с профессиональной программистской поддержкой. ТБД первой и второй категории являются небольшими и могут быть созданы исследователями самостоятельно. Они предназначены для небольших групп пользователей или для размещения в Интернете. С помощью ТБД последней категории реализуются большие исследовательские проекты. В статье достаточно подробно обсуждаются особенности каждого из этих типов баз данных, средства их реализации. Например, для первой категории рекомендованы несложные программы Microsoft Access и FileMaker.

Важным этапом создания ТБД является выбор структуры представления данных (модели данных) – так называемое концептуальное проектирование. Предлагается выделять объекты и связи между ними. Для представления модели данных рекомендован специальный язык E-R диаграмм (Entity-Relationship diagram).

Основным форматом представления материала являются таблицы. Такие базы данных получили название реляционных. В статье описывается преобразование E-R диаграмм в таблицы реляционных баз данных. По нашему мнению, возможный следующий шаг в развитии ТБД будет состоять в обогащении E-R диаграмм и переходе к схемам объект-атрибут-значение (RDF-схемы, общепринятые в работах по созданию семантического Интернета).

Рассмотрим некоторые наиболее важные проблемы создания ТБД и ограничения, связанные с применением реляционной (табличной) схемы данных.

1. Представление последовательностей элементов. Допустим, что требуется хранить слова, разбитые на морфемы. В языках агглютинативного типа число морфем потенциально не ограничено, что делает затруднительным хранение всех морфем слова в одной строке в разных столбцах. Возможным решением, сохраняющим преимущества стандартных реляционных баз данных, является хранение каждой морфемы в отдельной строке.

2. Расширяемость. Чаще всего возможные значения того или иного параметра задаются фиксированным списком. Однако при описании новых языков, могут возникнуть новые значения, не предусмотренные заранее. Создатели ТБД должны позаботиться о легком способе расширения списков значений.

3. Иерархические структуры. Древовидные структуры, возникающие, например, при синтаксическом разборе предложений, не удобны для размещения в реляционных базах данных. Это же касается и больших корпусов текстов. В этих случаях, вероятно, придется отказаться от реляционных баз данных.

4. Число значений атрибутов. Часто в ТБД используются бинарные (т. е. принимающие только два значения: 1 и 0) признаки. Многозначные признаки могут быть представлены как композиции бинарных. Такой подход применялся еще Хомским [Chomsky 1968]. Однако следует иметь в виду и определенные минусы такого подхода. В [Wichmann, Kamholz 2008] приводится следующий пример. В WALS на карте № 106 (Reciprocal constructions) выделены следующие значения: отсутствие реципроков, реципроки совпадают с рефлексивами, реципроки отличаются от рефлексивов и смешанный вариант. Частотность этих значений сопоставима. Если же выделить бинарный признак 'присутствие / отсутствие реципрока', то получится следующая картина: присутствие реципроков отмечено в 159 языках, отсутствие – только в 16 (для остальных языков в WALS нет данных). Карта в этом случае получится неинтересной. Часто в такой ситуации авторы стремятся разбить значение признака с высокой частотностью на несколько подзначений.

Возможным способом сведения многозначных признаков к бинарным является представление каждого значения многозначного признака в качестве самостоятельного признака. Например, признак 'приоритетный порядок слов' со значениями 'SVO', 'SOV' и т. д. можно заменить набором бинарных признаков: 'порядок слов SVO', 'порядок слов SOV' и т. д. При этом надо иметь в виду, что эти бинарные признаки взаимосвязаны в том смысле, что любой язык может иметь только один из них. По мнению авторов статьи, это должно быть явно указано. При статистическом анализе также лучше иметь дело с многозначными признаками. Следует иметь в виду, что подобного рода выводы и рекомендации носят предварительный характер. Опыт использования ТБД к настоящему времени недостаточен, чтобы прийти к окончательным заключениям.

Изложенное до сих пор относилось к созданию отдельной базы данных. Однако после того как было создано несколько ТБД, естественным образом возникло желание их интегрировать и обеспечить возможность комбинированного поиска релевантной информации во множестве ТБД. Появляющиеся при этом проблемы многократно усложняются. Важная (и единственная на сегодняшний день) попытка решения данной проблемы предпринята в проекте TDS (Typological Database System), выполненном в Нидерландах. Ему посвящена статья А. Димитриадиса и его соавторов «Как объединить базы данных и не начать типологическую войну» («How to integrate databases without starting a typological war»).

В проекте TDS не создавалась оригинальная коллекция данных, а только обеспечивался единый веб-интерфейс к независимо созданным ТБД. Основные трудности при реализации проекта порождали не различия в структурировании данных или в программном обеспечении, а различия в терминологии и теоретических предпосылках создателей баз данных, а также огромное число параметров, используемых в разных ТБД при слабой документированности.

TDS (http://languagelink.let.uu.nl/tds/) объединяет следующие базы данных:

– по типологии анафоры (Anaphora Typology);

– по личному согласованию (Person Agreement Database);

– по фонемному инвентарю (Smith's Phoneme Inventory);

– по типологии ударения (Stress Typology Database);

– по типологии слога (Syllable Typology Database);

– по порядку слов и составляющих (Typological Database Amsterdam);

– по общей информации об устройстве языка (Typological Database Nijmegen);

– по интенсификаторам и рефлексивам (Typological Database of Intensifiers and Reflexives);

– по сегментному инвентарю фонетических систем (UCLA Phonological Segment Inventory Database);

– по редупликации (Graz Database on Reduplication);

– по цветонаименованиям (World Color Survey).

Перечисленные базы данных различаются в следующих основных аспектах.

1. Различные типы контента. Выделяются два типа ТБД – аналитические и коллекции предложений. Первые содержат переменные, описывающие язык как целое. Вторые – примеры предложений с подробным описанием. Разумеется, возможна и комбинация обоих типов информации в одной базе данных.

2. Различная теоретическая основа. Поскольку не существует общепринятой и исчерпывающей лингвистической теории, то создатели каждой из ТБД используют по своему усмотрению ту или иную теорию или версию теории. Например, в некоторых ТБД используются традиционные понятия 'субъект' и 'объект', в то время как в других – S/A/P/R [Haspelmath 2005].

3. Различия в целях создания. Даже при общей теоретической платформе могут, естественно, различаться цели создания ТБД: образовательные, исследовательские и т. д.

4. Различия в обозначениях. Так, множественное число может обозначаться и как pl, и как Plural.

5. Различия в структуре данных. Ясно, что даже одни и те же данные могут быть организованы и представлены различными способами.

6. Различия в программном обеспечении. Базы данных, включенные в TDS, созданы с использованием Microsoft SQL Server, MySQL, Microsoft Access, Exsel, 4th Dimention, а также с помощью программ, разработанных самими создателями ТБД. К счастью, существуют интерфейсные модули или форматы обмена данными для всех этих систем. Дополнительные сложности создают различия в операционных системах, шрифтах, форматах данных и др.

Принятый в TDS подход состоит в том, чтобы по возможности нивелировать расхождения в кодировании и представлении данных (объ-

ектная модель) и в то же время сохранить различия в семантике и теоретических подходах и сделать их более ясными за счет тщательного документирования. В TDS не предпринимается попытка унификации типологической информации на семантической основе в связи с практической неосуществимостью этого, по крайней мере, на данном этапе развития лингвистики. Кроме того, это позволяет избежать «типологической войны».

Для решения проблемы интероперабельности выбрана двухуровневая модель [Stuckenschmidt, Harmelen 2005], включающая глобальную онтологию общих лингвистических концептов (TDS-GO) и локальные онтологии компонент отдельных баз данных. Для преодоления различий в структуре данных и обозначениях разработан декларативный язык Data Transformation Language (DTL).

Возможно, наиболее интересной частью проекта TDS является глобальная онтология TDS-GO, организованная в форме иерархии классов. Пример:

1. Linguistic property → phonetic or phonological property → syllable structure property → onset feature → obligatory onset.
2. Linguistic property → phonetic or phonological property → suprasegmental property → stress placement → main stress placement → variable stress placement systems → non-lexical stress placement → edge placement → right word edge stress placement → antepenultimate if heavy, else penultimate if heavy, else antepenultimate.
3. Linguistic property → phonetic or phonological property → marker function → agreement marker function → agreement marker for core argument → subject agreement marker.

Глобальная онтология строится с использованием языка OWL, являющегося стандартом в области формализации семантики в сети Интернет. Авторы разработки подчеркивают, что глобальная онтология создавалась не как каноническая система лингвистических понятий, а скорее для выражения набора всех точек зрения, представленных в базах данных TDS. TDS-GO создавалась независимо от ранее разработанной онтологии GOLD (http://www.linguistics-ontology.org) и в дальнейшем может быть объединена с ней.

TDS-GO содержит следующие типы концептов: лингвистические объекты ('морфема', 'предложение' и др.), лингвистические свойства ('базисный порядок слов' и др.) и лингвистические отношения ('согласование' и др.).

Поиск в TDS реализуется в виде двухшагового процесса. На первом этапе выбираются релевантные поля базы данных, которые помещаются в корзину запроса. Затем пользователь обращается к корзине и уточняет параметры поиска.

Авторы TDS дают общие рекомендации разработчикам ТБД. Прежде всего, следует тщательно продумать структуру базы данных. Далее, если она рассчитана на длительное использование и / или широкий доступ, то ее необходимо детально документировать и снабдить библиографическими источниками информации. Наконец, необходимо позаботиться об использовании стандартных обозначений. Например, для наименования языков это стандарт ISO 639-3. В статье также обсуждаются проблемы комментариев, нулевых и неопределенных значений, хотя рекомендации по этим вопросам носят менее ясный характер.

По широте охвата языков и признаков ТБД можно разделить на универсальные и специализированные. К первым относятся базы данных, имеющие целью описать грамматики языков более или менее полно хотя бы в ключевых аспектах. Как следствие, такие ТБД содержат информацию по многим свойствам (и для многих языков). В таком случае информацию уместно структурировать в виде одной таблицы, в которой строки соответствуют свойствам, а столбцы – языкам (можно и наоборот).

Единственной, кроме базы данных «Языки мира», универсальной ТБД является уже упоминавшийся выше WALS – крупный международный проект, выполненный под руководством М. Хаспельмата, Б. Комри и др. и включающий компьютерную базу данных и бумажное издание. Ему посвящена статья М. Хаспельмата «Типологическая база данных "Мирового атласа языковых структур"» («The typological database of the *World Atlas of Language Structures*»). WALS оперирует со значительно большим числом языков и признаков по сравнению со всеми другими ТБД (кроме «Языков мира»): 2560 языков и 142 признака, каждый из которых может принимать несколько значений – от 2 до 9, в среднем, примерно 5.

Две отличительные особенности придают проекту большую значимость. Во-первых, для каждого признака построена карта Земного шара, на которой кружочками разного цвета обозначены языки с различными значениями выбранного признака. Хотя идея графического изображения географического распределения признаков предлагалась и ранее (в том числе, в работе Дж. Николс), но впервые она была реализована столь масштабно. Во-вторых, база данных снабжена чрезвычайно удобным интерфейсом, позволяющим легко ориентироваться в огромном массиве информации, и включает

огромное количество справочных статей по различным признакам и языкам. Признаки классифицируются по тематике и по алфавиту, языки – по семьям и регионам. Поисковые средства позволяют находить нужную информацию по комбинации признаков, генерировать новые карты с заданными свойствами.

Отметим, что WALS заведомо не претендует на полноту описания языка, выделяя только 142 (причем из них лишь 126 относятся собственно к грамматике) наиболее интересных типологам свойства. Кроме того, многие клетки таблицы WALS оказались незаполненными как из-за отсутствия данных, так и из-за нерелевантности некоторых признаков.

WALS является одной из немногих ТБД, активно используемых в научных исследованиях. Некоторые примеры приведены в обзорной работе [Соловьев 2010].

Создание универсальных ТБД сопряжено с огромными затратами ресурсов. Значительно проще создать ТБД, посвященную какому-либо небольшому разделу грамматики или даже одному признаку. В статье Ф. Гаста «К двухуровневому языковому описанию: типологическая база данных по интенсификаторам и рефлексивам» («A contribution to 'two-dimensional' language description: the typological database of intensifiers and reflexives») такие ТБД названы межъязыковыми базами данных по отдельным областям грамматики («domain-specific cross-linguistic database»). Эта работа, хотя и посвящена специализированной базе данных, начинается с интересного обсуждения общих методологических вопросов.

Отмечается, что грамматика традиционно описывается либо семасиологически (от формы к функции), либо ономасиологически (от функции к форме). Разумеется, возможны и различные комбинации этих подходов, одна из которых приводит к интересному формализму семантических карт [Auwera, Plungian 1998], весьма удобному для сравнительных исследований.

Другой возможностью описания грамматик, специально ориентированной на сопоставительные исследования, является разработка некоего универсального шаблона, представляющего собой набор признаков. Независимо от серии монографий «Языки мира», где применен именно этот подход, и чуть раньше попытка единообразного описания языков в заранее фиксированных терминах была предпринята в серии Lingua Descriptive Studies в конце 70-х годов XX века. В монографиях этой серии языки описывались на основе опросника, разработанного Б. Комри и Н. Смитом [Comrie, Smith 1977]. Позже идея была продолжена в модифицированной форме в серии Routledge Descriptive Grammars (например [Hewitt, Khiba 1989]).

Описания языков на основе этой анкеты, к сожалению, не были переведены в форму компьютерной базы данных. Сам проект часто критиковался и, видимо, не имел большого успеха. Объектом критики была заведомая неполнота описания: во-первых, не все категории релевантны для всех языков, т. е. некоторые клетки двумерной таблицы будут пустовать, и, во-вторых, многие детали грамматик языков не будут представлены в этом описании, т. к. учесть все в одном опроснике представлялось невозможным.

Однако, по мнению автора статьи, все же именно такой вид описания наиболее удобен для межъязыковых исследований, чем и мотивирован выбор формата для описываемой в статье ТБД по интенсификаторам и рефлексивам (http://www.tdir.org).

Это довольно маленькая база данных, содержащая всего 689 (на момент написания статьи, т. е. на 30.04.2007) примеров конструкций. Отмечается, что, хотя примеры охватывают более 100 языков, выборка не сбалансирована и поэтому ее нельзя использовать для статистических исследований. Проект создавался, скорее, с описательными целями. ТБД по интенсификаторам и рефлексивам включает релевантные ссылки на литературу и может рассматриваться как отправная точка для исследователей, интересующихся этой областью. Автор статьи указывает на ограниченность финансовой поддержки проекта и выражает надежду, что эта работа послужит созданию в будущем более объемных и усовершенствованных ТБД.

В статье Х. Блисс и Э. Риттер «Типологическая база данных по личным и указательным местоимениям» («A typological database of personal and demonstrative pronouns») обсуждается ТБД (http://136.159.142.10:591/), которая создавалась с целью изучения морфосинтаксических свойств, присущих системам местоимений в языках мира. Например, хорошо известна универсалия Гринберга № 42 [Greenberg 1963]: «Все языки имеют, по меньшей мере, три лица и два числа». Естественно, возникают вопросы, сколько лиц и чисел может быть максимум? Как эти параметры комбинируются между собой? Какие виды синкретизма и лакун невозможны, редки или, напротив, часто встречаются? Какие другие морфосинтаксические свойства характерны для местоименных систем языков мира?

ТБД личных и указательных местоимений создана с использованием FileMarker Pro 5.5. Она включает местоимения из 109 типологически различных языков 52 семей. Каждому местоимению выделяется отдельная запись, в

которой оно описывается значениями из некоторого набора параметров.

Данные, собранные в этой ТБД, подтвердили ранее отмечавшиеся закономерности: категории лица и числа варьируются в незначительной степени, в то время как падеж, род, вежливость могут иметь широкий спектр значений, образующих открытый класс.

В языках базы данных представлены следующие значения: 1-е, инклюзивное, 2-е, 3-е, 4-е (для категории лица) и единственное, двойственное, тройственное, паукальное, множественное, общее (для категории числа). Следует отметить, что такая интерпретация языковых данных не является бесспорной. Например, И. Мельчук [Мельчук 1998: 204] настаивает на том, что лиц бывает только 3, а инклюзивность является особой категорией (близкой к числу). В связи со сложностью интерпретации инклюзивности авторы ТБД сознательно уклоняются от принципиальной теоретической дискуссии, стремясь просто предоставить данные для исследователей. Детальный анализ инклюзивности у Мельчука демонстрирует неполноту рассматриваемой ТБД. Четвертое лицо также может являться просто проявлением дополнительной категории (в языке юпик это 'упоминавшееся ранее лицо').

Возможные комбинации лиц и чисел ограничены. Обнаружено, что нет языков, которые имели бы и тройственное, и паукальное число, что подтверждает гипотезу Корбетта [Corbett 2000].

Вместе с традиционной категорией рода в базе данных рассматриваются различия по степени одушевленности: живые, неживые, люди, сверхъестественные существа, растения и нейтральный класс. Падежи, естественно, образуют большой открытый класс. Однако их можно разбить на три группы: грамматические (номинатив, эргатив, генитив, аккузатив), локативные и остальные. Статистические подсчеты показывают, что из этих трех групп каждая следующая встречается реже предыдущей, что подтверждает предположения Блейка [Blake 1994].

На момент создания этой ТБД вежливость была изучена меньше всего по сравнению с другими категориями. Поэтому обнаруженные в этой части закономерности представляются наиболее интересными. Помимо хорошо известных форм выражения вежливости в местоименной системе индоевропейских языков (только для второго лица) были обнаружены языки, где это различие проявляется в 1-м и / или 3-м лицах, а также показано, что нет языков с различением вежливости в инклюзивном лице. Некоторые языки (тайский, вьетнамский) имеют более чем два значения этой категории. На выбор формы влияют возраст, род, должность и иные факторы.

Следует подчеркнуть, что, в отличие от работ Корбетта и Блейка, в которых гипотезы формулировались на основе интуиции авторов и несистематического обзора языков, результаты, полученные в этой работе с использованием ТБД, носят более доказательный характер. Конечно, они не являются универсалиями, т. к. эта ТБД не содержит всех языков, однако широкий спектр представленных в ней типологически различных языков из большого числа семей (сбалансированность выборки) делает полученные результаты более обоснованными.

Авторы обращают внимание на два типологически интересных случая редких местоименных систем. Инклюзивное лицо по своей семантике является множественным ('говорящий и собеседник(и)'). Однако некоторые языки, например, нганди (Австралия), имеют две формы этого лица – единственного и множественного числа. Единственное число обозначает 'я и ты', а множественное – 'я и несколько адресатов речевого акта'.

Еще один интересный случай – использование указательных местоимений в качестве личных местоимений 3-го лица, что имеет место, например, в баскском.

Таким образом, рассматриваемая ТБД оказалась эффективным инструментом исследования типологии местоименных систем, выявляя как статистически значимые тенденции, так и необычные феномены.

Другая база данных по этой проблематике описывается в статье Г. Сежерера «База данных по личным местоимениям в африканских языках» («A database on personal pronouns in African languages», http://sumale.vjf.cnrs.fr/pronoms/).

Местоименные системы являются идеальным кандидатом для типологических исследований: они универсальны, образуют замкнутое множество с жесткой организацией, но широким спектром внутренних свойств. Это позволяет использовать местоименные системы и в генеалогических исследованиях. В то же время многие данные рассредоточены по многочисленным публикациям, и не всегда легко доступным. Все это делает желательным создание базы данных, которая бы объединяла описания местоименных систем.

ТБД личных местоимений в африканских языках была создана в рамках трехлетнего проекта, длившегося с 2001 по 2004 г., группой лингвистов из Германии и Франции с целью собрать и представить в унифицированном виде всю доступную информацию по местоименным системам африканских языков. В настоящее время ТБД содержит материал бо-

лее 500 языков и является одной из немногих (наряду с «Языками мира»), так сказать, региональных баз данных: включенные в нее языки относятся к одной из макрообластей Земного шара. Это придает ей дополнительный интерес, т. к. позволяет использовать для изучения динамики языковых систем.

При создании БД были приняты следующие решения: существует только три лица и три числа: единственное, множественное и двойственное. Другие варианты в африканских языках не найдены. Далее, каждая форма описывается следующими параметрами: одушевленность, род, эксклюзивность / инклюзивность, определенность, логофоричность. Выделено 5 укрупненных синтаксических позиций: субъект, объект (прямой, косвенный), посессор, рефлексив, независимое употребление.

Общая таблица личных местоимений какого-либо языка содержит все возможные комбинации параметров, даже если они в данном языке не представлены. Это, по мнению авторов, облегчает процедуру сравнения систем разных языков. Поисковая система позволяет осуществлять поиск по разнообразным сочетаниям форм, признаков и по языковым семьям с выдачей статистики и представлением результатов в виде карт (в духе WALS).

Г. Сежерер отмечает, что эта база данных создавалась как инструмент для сопоставительных и типологических исследований, но за прошедшее время опубликована всего одна статья, использующая материал этой ТБД (причем самим разработчиком!). Автор выражает надежду, что результаты его проекта еще будут востребованы типологами.

В статье Д. Брауна, К. Тибериус, М. Чумакиной, Г. Корбетта и А. Красовицки «Базы данных для изучения конкретных явлений» («Databases designed for investigating specific phenomena») представлены работы Морфологической группы университета Суррея (Великобритания) по созданию целой серии ТБД (доступны по адресу http://www.smg.surrey.ac.uk).

Эта исследовательская группа значительное внимание уделяет русскому языку, в частности, ей принадлежит диахроническая база данных по русской морфологии. В качестве источника данных использовался корпус русского языка, созданный А. Барентсеном в Амстердаме и содержащий 30 млн. слов из различных текстов с 1801 по 2000 г. Статистический анализ проводился для 10 двадцатилетних промежутков.

Интерес представляют нестабильные конструкции, допускающие множественную морфологическую маркировку в одной и той же синтаксической позиции. В базе данных представлены диахронические изменения в шести синтаксических конструкциях русского языка: глаголы с отрицанием (родительный / винительный падеж прямого дополнения), существительные в предикативной позиции (именительный / творительный падеж), предикативные прилагательные (краткая / полная форма / творительный падеж), предикаты в квантифицированных выражениях (единственное / множественное число), сочинительные конструкции (единственное / множественное число предикатов), именные группы с числительными 2–4 (именительный / родительный падеж прилагательных).

Для каждой альтернации в базе данных учтены различные параметры (как описанные ранее в литературе, так и найденные в оригинальных корпусных исследованиях), которые могут влиять на выбор конструкции. Это позволяет в деталях изучать диахронические процессы, их направленность и факторы, влияющие на языковые изменения. В частности, показано, что при выборе маркировки падежа у предикативных существительных в XX веке происходит сдвиг от семантически-ориентированного выбора (в первой половине века) к синтаксически детерминированному.

Другая разработка Морфологической группы Суррейского университета – база данных по согласованию (http://www.smg.surrey.ac.uk), которая содержит детальное описание данного явления в 15 языках. Вообще говоря, такое число языков может показаться очень малым, тем более что ранее Анной Северской [Siewierska 1999] создана ТБД по согласованию в 272 языках. Однако последняя содержит значительно меньше информации для каждого отдельного языка. Это общая дилемма – собирать данные по возможно большему числу языков или для относительно небольшого числа языков представить максимально детализированную информацию о рассматриваемом явлении. В данном случае авторы пошли по второму пути.

Согласование является достаточно сложным явлением и не имеет строгого определения. Перед исследователями обычно встает вопрос, что относится к объекту изучения, а что нет. В данном случае авторы использовали понятие канонического согласования, предложенное Г. Корбеттом [Corbett 2006]. Согласование описывается следующим набором структурных элементов: контролер, цель, синтаксическое окружение, согласовательные категории, условия согласования. Записи в базе данных соответствует согласовательная конструкция, характеризующаяся набором этих элементов.

Часто ТБД создаются для проведения конкретных сопоставительных исследований. В данном случае было тщательно изучено согласование в русском языке на предмет его каноничности.

Г. Корбетт предложил три принципа канонического согласования [Corbett 2006]:

Принцип 1. Каноническое согласование скорее избыточно, чем информативно.

Принцип 2. Каноническое согласование синтаксически просто.

Принцип 3. Согласование тем ближе к каноническому, чем оно ближе к аффиксальной словоизменительной морфологии.

На основе этих принципов Корбеттом выдвинуто 19 критериев. Например, критерий 1, поддерживающий принцип 1, гласит, что присутствие контролера более канонично, чем его отсутствие. В русском языке из 11 типов синтаксических конструкций контролер присутствует в 9. Авторы этой ТБД (и статьи) оценивают степень каноничности согласования в русском языке по этому критерию в $9/11 \approx 82$ %.

Вычисленная таким образом степень каноничности русского согласования составляет в среднем 86% для первого принципа и 90% для третьего.

В статье дано также описание и других проектов этой исследовательской группы: баз данных по синкретизму и супплетивизму. Любопытно, что и при описании базы данных по синкретизму также используется материал русского языка – окончания существительных в разных падежах.

Статья Р. Гудсманса и Г. Хульста «StressTyp: база данных по акцентным структурам в языках мира» («StressTyp: A database for word accentual patterns in the world's languages») посвящена базе данных по ударению StressTyp (http://stresstyp.leidenuniv.nl).

Этот аспект языкового устройства оказался достаточно хорошо представлен в типологических базах данных, см. также [Bailey 1995; Gordon 2002]. В статье детально сравниваются эти базы данных, обращено внимание на высокую степень корреляции между ними при наличии небольшого числа расхождений.

Приведем здесь краткие сведения по истории развития ТБД StressTyp, представляющиеся поучительными. Работа над ней началась в 1991 г. по инициативе ван дер Хульста в рамках EUROTYP [Hulst (ed.) 1999], более крупного проекта по типологии европейских языков, финансировавшегося Европейским научным фондом. На этой стадии в БД было введено доступное из типологической литературы описание ударения в 154 языках и были проведены дополнительные исследования. После завершения в 1994 г. EUROTYP состав разработчиков ТБД StressTyp в значительной степени сменился, но темпы разработки оставались высокими: за следующие 3 года было добавлено описание ударения в 116 новых языках.

В 1997–2001 гг. StressTyp вошел в состав другого проекта – «Просодия в индонезийских языках», координировавшегося Лейденским университетом. В этот период число языков в базе данных увеличилось до 510. В этот же период StressTyp был включен и в состав WALS, для которого было подготовлено 4 карты, № 14–17 [Haspelmath et al. (eds.) 2005]. С появлением упоминавшегося выше интегрирующего проекта TDS StressTyp вошел и в него.

Авторы прилагали большие усилия для того, чтобы эта разработка могла использоваться в научных исследованиях. StressTyp доступна через Интернет как напрямую, так и через TDS. Свободно распространяемая версия базы данных в формате Access может быть получена у авторов. В 1996 г. был издан сборник статей [Goedemans, Hulst 1996] по StressTyp и ее использованию, в настоящее время в издательстве John Benjamins готовится к изданию следующий сборник.

Создатели базы данных отмечают, что с самого начала проект имел очень небольшое финансирование. Они планируют и дальнейшее расширение базы данных, для чего необходимы новые гранты. Один из возможных путей разработка вопросника и заполнение его специалистами по отдельным языкам. Однако до сих пор эта идея не реализована. В статье авторы выражают заинтересованность в сотрудничестве с любыми обладателями ресурсов в этой области, даже если они имеются только в бумажной форме.

Другие их планы состоят в создании целой системы баз данных, посвященных ударению: аннотированной библиографии (StressBib, в стадии разработки), терминологической базы данных (StressTer, в стадии разработки), списка адресов лингвистов, работающих в этой области (StressRes). Объединение этих баз данных в единую систему приведет к появлению сети Stress Expert System.

Кроме того, Р. Гудеманс и Г. Хульст работают над ТБД фонотактической информации (SylTyp), которая совместно с ТБД по тонам (http://xtone.linguistics.berkeley.edu/) входит в состав базы данных по просодии (Word Prosody Database). Следующим шагом может быть интеграция с базой данных сегментного инвентаря Я. Меддисона. Наконец, возможно и объединение всех вышеупомянутых баз данных, описывающих существующие типы фонетических единиц, с ТБД, посвященными фонетическим процессам: NasDat (http://acvu.nl.staf/wlm.wetzels/ pwp_en.htm), ATR/Vowel Harmony и др.

Разработчики StressTyp дают хороший пример международного сотрудничества и интеграции усилий.

Статья Я. Мэтраса, К. Вайта и В. Элшика «База данных по морфосинтаксису романи» «The Romani Morpho-Syntax (RMS) database» посвящена проекту по языку европейских цыган (http://romani.humanities.manchester.ac.uk/). Этот язык, генетически индоарийский, возник в центральной части Индии несколько тысяч лет назад. Его важной особенностью является то, что у него нет статуса официального языка ни в одном государстве. На диалектах романи говорят в десятках стран. Проект RMS выполнялся при поддержке трех организаций с общим бюджетом 840 тысяч евро. Это один из немногих случаев, когда называется точная сумма, что дает возможность оценить уровень финансирования лингвистических проектов в Евросоюзе. В нем было задействовано три исследователя на условиях полной занятости и около 60 на условиях частичной занятости.

Цели создания ТБД RMS:

1. Исторические: изучение инноваций с фокусом внимания развития от функции к форме и от формы к функции.

2. Типологические: изучение структурных репрезентаций функций в различных диалектах, отношений функций и форм, кластеров функций.

3. Ареальные: изучение контактных влияний.

4. Диалектологические: изучение связей между инновациями и их географическим распределением, генеалогической классификации диалектов.

База данных по морфосинтаксису романи является, видимо, единственной базой данных, содержащей описание многих диалектов одного языка и ориентированной на изучение ареальных влияний. RMS не просто ТБД, это тщательно продуманный и спланированный проект, в котором особое внимание уделено стратегии сбора данных, их обработке и оценке.

ТБД по редупликации (http://reduplication.uni-graz.at/), созданная в Граце, описана в статье Б. Хурха и В. Матс.

Авторы отмечают, что редупликация имеет специфический статус, не позволяющий отнести ее ни к лексическому уровню языка, ни к грамматическому. Первая и весьма содержательная часть статьи посвящена рассмотрению феномена в общетеоретическом аспекте. Сама база данных по редупликации описана чрезвычайно подробно с приложением всех таблиц. Ко времени издания рецензируемой книги проект еще не был завершен.

Весьма интересной особенностью интерфейса является его иерархическая (древовидная) организация. Например, сначала в диалоговом режиме выбирается язык, затем функция редупликации в этом языке, затем словарные формы и т. д., что является удобным, но несколько идет вразрез с общей тенденцией к табличной организации данных в большинстве ТБД.

Как и в ряде других случаев, отмечается, что эта ТБД позволяет лучше понять рассматриваемый феномен, обеспечивая релевантные данные и ссылки, а также предлагает средства для тестирования различных гипотез о природе редупликации.

Таким образом, рецензируемая работа демонстрирует, что к настоящему времени созданы десятки типологических баз данных, накоплен значительный опыт их разработки. В результате появились принципиально новые возможности квантитативных типологических исследований с применением математических и компьютерных методов. Проблемой является то, что методология использования типологических баз данных в научных исследованиях пока недостаточно разработана. В итоге они используются в настоящее время прежде всего в учебно-справочных целях.

В заключение следует подчеркнуть, что рецензируемая монография написана квалифицированными учеными, работающими на переднем крае исследований. Она полно отражает положение дел в области типологических баз данных и может быть рекомендована всем, кто хочет поподробнее познакомиться с этим быстро развивающимся перспективным направлением лингвистики.

СПИСОК ЛИТЕРАТУРЫ

Мельчук 1998 – *И.А. Мельчук*. Курс общей морфологии. Т. 2. Ч. 2. М.; Вена, 1998.

Поляков, Соловьев 2006 – *В.Н. Поляков, В.Д. Соловьев*. Компьютерные модели и методы в типологии и компаративистике. Казань, 2006.

Соловьев 2010 – *В.Д. Соловьев*. Типологические базы данных: перспективы использования // ВЯ. 2010. № 1.

Auwera, Plungian 1998 – *J. Auwera, V. Plungian*. Modality's semantic map // Linguistic typology. 1998. V. 2.

Bailey 1995 – *M. Bailey*. Nonmetrical constraints on stress. Ph. D. dissertation. University of Minnesota. 1995.

Blake 1994 – *B. Blake*. Case. Cambridge, 1994.

Buneman et al. (eds.) 2001 – *P. Buneman, S. Bird, M. Liberman* (eds.). IRCS Workshop on linguistic databases. Philadelphia (Pennsylvania), 2001.

Chomsky, Halle 1968 – *N. Chomsky, M. Halle*. The sound pattern of English. New York, 1968.

Comrie, Smith 1977 – *B. Comrie, N. Smith*. Lingua descriptive series: Questionnaire // Lingua. 42. 1977.

Corbett 2000 - *G. Corbett*. Number. Cambridge, 2000.
Corbett 2006 – *G. Corbett*. Agreement. Cambridge, 2006.
Everaert 2003 – *M. Everaert*. The use of databases in linguistic theorizing // Abstracts of the XVII International congress of linguists. Prague, 2003.
Goedemans, Hulst 1996 – *R. Goedemans, H. Hulst*. StressTyp manual. Leiden, 1996.
Gordon 2002 – *M. Gordon*. A factorial typology of quantity insensitive stress // Natural language and linguistic theory. 2002. 20.
Greenberg 1963 – *J. Greenberg*. Some universals of grammar with particular reference to the order of meaningful elements // J. Greenberg (ed.). Universals of language. Cambridge, 1963.
Haspelmath 2005 – *M. Haspelmath*. Argument marking in ditransitive alignment types // Linguistic discovery. 2005. 3(1).
Haspelmath et al. (eds.) 2005 – *M. Haspelmath, M. Dryer, D. Gil, B. Comrie* (eds.). The world atlas of language structures. Oxford, 2005.
Hewitt, Khiba 1989 – *G. Hewitt, Z. Khiba*. Abkhaz. London, 1989.
Hulst (ed.) 1999 – *H. Hulst* (ed.). Word prosodic systems in the languages of Europe. Berlin, 1999.
Liberman 1997 – *M. Liberman*. Introduction to the Linguistic Data Consortium. http://www.ldc.upenn.edu/ About/ldc_intro.shtml. 1997.
MacWhinney 1995 – *B. MacWhinney*. The CHILDES project: Tools for analysing talk. Hillsdale (NJ), 1995.
Nerbonne (ed.) 1998 - *J. Nerbonne* (ed.). Linguistic databases. Stanford, 1998.
Nichols 1992 – *J. Nichols*. Linguistic diversity in space and time. Chicago; London, 1992.
Polyakov et al. 2009 - *V. Polyakov, V. Solovyev, S. Wichmann, O. Belyaev*. Using WALS and Jazyki Mira // Linguistic typology. 2009. V. 13.
Siewierska 1999 – *A. Siewierska*. From anaphoric pronoun to grammatical agreement marker: Why objects don't make it // Folia Linguistica. 1999. V. 33. № 2.
Stuckenschmidt, Harmelen 2005 - *H. Stuckenschmidt, F. Harmelen*. Information sharing on the semantic web. Berlin, 2005.
Wichmann, Kamholz 2008 – *S. Wichmann. D. Kamholz*. A stability metric for typological features // Sprachtypologie und Universalienforschung. 2008. 61.3.

В.Д. Соловьев