ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 1 1990

НАУЧНАЯ ЖИЗНЬ

ХРОНИКАЛЬНЫЕ ЗАМЕТКИ

19—22 декабря 1988 г. в Таллинне состоялось рабочее совещание «Машинные фонды языков народов СССР», организованное Институтом языка и литературы АН ЭССР. В совещании приняли участие разработчики машинных фондов из Грузии, Казахстана, Киргизии, Коми АССР, Латвии, Литвы, с Украины, из Татарии, Эстонии, Ленинграда и Москвы.

По ряду языков были доложены результаты уже реализованных на ЭВМ разработок. По другим языкам заканчивается или продолжается проектирование баз данных, накопление первичных массивов данных.

В докладах были затронуты следующие основные проблемы: состав и структура машинных фондов различных языков, характеристика отдельных баз данных в составе машинных фондов, лингвистическое и программное обеспечение соответствующих разработок, машинные фонды и обучение языкам и некот. др.

Доклады показали, что в целом ряде республик имеются своеобразные концепции машинных фондов, которые активно разрабатываются и уже привели к получению существенных результатов. Так, в Грузинской ССР силами нескольких научных учреждений (Центральный институт научной информации по общественным наукам, ТГУ, Институт рукописей им. К. Кекелидзе и Институт истории АН ГССР и др.) Машинный фонд грузинского языка создается как иерархическая система взаимоувязанных баз данных и автоматизированных процессоров. Каждая из создаваемых баз данных включает определенный тип информации. Взаимоувязка баз данных делает возможным совместное использование нескольких баз и формирование новых баз на основе имеющихся данных. Создаваемые процессоры предполагают автоматизированное заполнение баз данных и аналитическую обработку данных (докладчик В. П. Гугушвили). В качестве одной из первых составляющих Машинного фонда грузинского языка создана база данных грузинских топонимов, включающая 300 тыс. географических названий. Создаются также полнотекстовая и словарная базы «Мученичества Шушаники», древнейшего грузинского письменного памятника, компьютерная версия грузинско-русского фразеологического словаря и другие лингвистические и фактографические базы данных.

Эстонские разработчики предлагают концепцию инструментального подхода к созданию машинных фондов языков СССР (докладчик М. Реммель). Данная, концепция предполагает создание программных средств высокого уровня, ориентированных не непосредственно на обработку языкового материала, а на генерацию программных средств более низкого уровня, предназначенных для решения конкретных лингвистических задач. Другой особенностью инструментального подхода является его принципиальная направленность на использование традиционных описаний (словарей и грамматик), что позволяет включить в оборот уже имеющиеся описания для многих языков, для которых формальных описаний нет. Третья особенность «эстонского варианта» — ориентация на хозрасчетные отношения между разработчиками машинных фондов. Эти принципы положены в основу разрабатываемого в Эстонии проекта PROFUNDA, отдельные части которого были доложены на совещании и показаны во время демонстрации в Таллиннском политехническом институте (докладчики Ю. Вике, К. Калдамяэ).

Лингвисты и программисты из Москвы и Ленинграда (О. А. Казакевич, Л. И. Колодяжная, Ж. Г. Мошкович, А. С. Асиновский) доложили о работах по созданию машинных фондов младописьменных языков народов СССР, в том числе о Машинном фонде селькупского языка. Последний разрабатывается в Лаборатории АЛС НИВЦ МГУ как совокупность текстовой и словарной баз данных, способных к постоянному пополнению. В качестве программного обеспечения Машинного фонда селькупского языка используется систе-

ма УНИЛЕКС-2. В настоящее время текстовая база данных содержит 28 текстов (свыше 10 тыс. словоупотреблений). Получены частотные и алфавитно-частотные словари словоформ по всему корпусу и по отдельным текстам, обратный словарь словоформ и конкордансы по говорам. Завершается автоматическая лемматизация корпуса. Основными компонентами словарной базы данных являются автоматический словарь селькупского языка и русско-селькупский словарь (около 2,5 тыс. словарных статей). На основе последнего автоматически получен словарь синонимов селькупского языка.

В докладе М. М. Пещак (Киев) было охарактеризовано состояние работ по Машинному фонду украинского языка. К концу 1988 г. в нем имелось пять подфондов (словарный, социолингвистический, иллюстрационно-текстовый, диалектологический и только зарождающийся фонд письменных памятников). Основным видом работы до последнего времени было накопление материалов в составе первичных баз данных, создаваемых на основе печатных изданий и ответов на анкеты, и вторичных баз данных, являющихся результатом автоматизированной обработки текстов и анкет. Начата также разработка программных средств для обслуживания имеющихся баз данных.

В нескольких республиках ведется работа преимущественно над словарными компонентами машинных фондов. Об этом было доложено представителями Казахстана (С. А. Ускомбаев), Коми АССР (С. В. Лесников), Литвы (В. Жилинскене и др.), Татарии (К. Р. Галиуллин).

Совещание прошло в деловой обстановке, при хорошем уровне взаимопонимания между его участниками, что позволяет надеяться на дальнейшее активное развертывание работ по машинным фондам национальных языков .

*

С 22 по 27 мая 1989 г. в Москве проходила третья Всесоюзная конференция по созданию Машинного фонда русского языка (МФ РЯ), организованная Институтом русского языка АН СССР и МГУ им. М. В. Ломоносова [1]. В конференции приняли участие 170 исследователей и разработчиков в области компьютерной лингвистики. Было проведено 2 пленарных и 14 секционных заседаний. Работали секции: лингвистических процессоров, словарных фондов, машинных фондов языков народов СССР, диалектных фондов, текстовых фондов, фонетических фондов и терминологических фондов. В Институте русского языка АН СССР впервые были проведены демонстрации программно-источниковых продуктов на ЭВМ ЕС-1036, СМ-4 и персональных компьютерах.

В. М. Андрющенко (Москва) в докладе «О рабочем проекте МФ РЯ» охарактеризовал состояние работ в головной организации — ИРЯ АН СССР. К настоящему моменту в основном создана вычислительная база и активно ведутся работы по главным компонентам МФ РЯ. Центром МФ РЯ является Генеральный словник русского языка, идея которого принадлежит акад. А. П. Ершову. С Генеральным словником должны быть согласованы все остальные базы данных. Уже имеются машинные версии Грамматического, Словообразовательного, Орфографического словарей и словаря «Слитно или раздельно?». Следующим этапом работы является создание сводного машинного словника русского языка, включая программно-технологические средства автоматизированного формирования словарной статьи и пополнения словника за счет обработки текстовых массивов. На конец 1990 г. запланировано иметь в машинной форме словник объемом около 200 тыс. словарных статей с морфологической и частично синтаксической информацией. По Иллюстрационно-текстовому фонду активно ведется формирование пяти корпусов текстов (разговорной речи, фольклорных, общественно-политических, поэтических и прозаических художественных текстов).' В составе Академического словарно-грамматического фонда вступила в эксплуатацию первая версия Автоматического варианта ДАРЯ, первая версия Автоматического синтаксического словаря Г. А. Золотовой. Активно ведутся работы по созданию принципиально нового типа словаря — Автоматического фразеологического словаря русского языка, а также по Ассоциативному тезаурусу русского языка.

В докладе Ю.Н. Караулова (Москва) «Всесоюзная программа "Русский язык,, и задачи Машинного фонда» последний был охарактеризован как средство поддержки программы «Русский язык». В разработках, ведущихся в рамках МФ РЯ, отражаются три шага информатизации общества. Первым шагом является создание разнообразных ИПС. Применительно к МФ — это конкордансы и программно-источниковые пакеты по разным типам текстов. Второй шаг информатизации предполагает создание экспертных систем. В качестве таковых могут рассматриваться все процессоры и автоматизированные словари в составе МФ РЯ. Третьим шагом информатизации является создание систем типа «гипертекст», которые способны порождать но-

вые знания и синтезировать новые тексты. К системам такого типа приближаются два компонента МФ РЯ: Автоматический вариант ДАРЯ и Ассоциативный тезаурус русского языка.

Ряд докладов был посвящен лингвистическому, математическому и программному обеспечению процессоров русского языка. Доклады показали, что работы в этой области весьма продвинуты и продолжают активно развиваться. Разработки НИИ систем автоматизации НПО «Каскад» (Москва, - докладчики Э.И.Королев и др.) характеризует комплексный подход к лингвистическим процессорам как системе средств, предназначенных для многоцелевой и многоаспектной обработки русских научно-технических текстов, включая средства создания и ведения комплекса машинных словарей, системы анализа, коррекции и синтеза текста, средства создания баз знаний автоматизированных систем. В докладах ученых МГУ (М. Г. Мальковский и др.) лингвистический процессор рассматривается как часть более общей системы понимания естественного языка в рамках работ по искусственному интеллекту. Принципом системы искусственного интеллекта TULIPS-2 и ее лингвистического процессора «АДАМАНТ» является открытость как лингвистических, так и предметно-ориентированных знаний. Представители ЛГУ (Г. С. Цеитин и др.) разрабатывают систему машинного понимания текстов в ограниченной предметной области на основе распределенного представления языковых и предметных знаний. В данной системе в качестве промежуточной структуры данных используются ассоциативные сети, в которых отдельными узлами представлены как языковые объекты, так и соответствующие им внеязыковые объекты. В ИППИ АН СССР под руководством Ю. Д. Апресяна разрабатывается лингвистический процессор для перевода запросов на естественном (русском) языке на искусственный язык SQL, предназначенный для обращения к базам данных реляционного типа (докладчики И.М.Богуславский, Л. Л. Цинман).

На секции Словарных фондов обсуждались три основных круга проблем: разработка машинных словарей, программное обеспечение словарей МФ РЯ, конкретные лингвистические задачи, которые могут быть поставлены на машинных словарях. Совместный доклад представителей ЛО ИЯЗ АН СССР и НИВЦ МГУ (Р. П. Рогожникова, Л. И. Колодяжная и др.) был посвящен работе над сводным словником русских словарей, который включает 14 наиболее авторитетных и массовых словарей русского языка. В настоящее время сводный словник содержит более 170 тыс. слов и находится в печати. В выступлении по докладу В. М. Андрющенко подчеркнул, что сводный словник не следует смешивать с машинным Генеральным словником русского языка. Последний предполагает особую структуру словарной статьи и будет, в частности, содержать такую информацию, которой нет в традиционных словарях. Источниками Генерального словника будут сводный словник русских словарей, нетрадиционные словари (например, словари лингвистических процессоров), а также автоматически обрабатываемые тексты. В докладе Н.Н. Леонтьевой (Москва) был охарактеризован Русский общесемантический словарь (РОСС). Словарь не связан жестко с какой-либо предметной областью и включает большое количество эксплицитно поданной информации. Словарная статья имеет около ста полей, объединенных в десять зон, четыре из которых содержат семантические и энциклопедические сведения. Словарь предназначен для лингвистической экспертизы текста в режиме «человек — машина». Результаты экспертизы в форме семантического представления текста будут накапливаться в базе текстовых знаний, которая станет основным компонентом информационно-справочной системы по общественно-политической тематике. На секции была также освещена работа над Автоматическим словарем русских неологизмов (планируемый объем — 60 тыс. словарных статей), которая ведется в ЛО ИЯЗ АН СССР (докладчики Т. Н. Буцева, СИ. Алаторцева), и другие проекты.

Большое внимание <ловарным разработкам было уделено также на секции Машинных фондов языков народов СССР. Преимущественное развитие словарного подфонда характерно для ряда МФ: эстонского (доклады Ю. Вике, И. Хейн, К. Калдамяэ), татарского (доклад К. Р. Галиуллина и др.). В МФ украинского языка создан Морфемно-словообразовательный фонд, представляющий собой словарную базу данных со словником свыше 165 тыс. единиц, снабженную разнообразным программным обеспечением, с информационно-справочной и исследовательской функциями (доклад Н. Ф. Клименко и др.). В МФ латышского языка получили наибольшее развитие текстовые базы данных (общий объем — более 350 тыс. словоупотреблений). По части массива (250 тыс. словоупотреблений) получен автоматический частотный словарь словоформ (37 тыс. единиц) (докладчики В. А. Дризуле и др.). Активно ведутся работы над МФ младописьменных

и бесписьменных языков. В Л О ИЯЗ АН СССР создается информационная система, включающая словарные и текстовые данные на машинных носителях по чукотскому, керекскому, корякскому, ительменскому, эскимосскому, ненецкому, юкагирскому, нивхскому, нганасанскому, адыгейскому и гагаузскому языкам (докладчики А. С. Асинов ский и др.). В НИВЦ МГУ продолжается работа по МФ селькупского языка (докладчик О. А. Казакевич). К настоящему времени проведена автоматическая лемматизация корпуса текстов, получены частотные, алфавитно-частотные и обратные словари словоформ и лексем, словоуказатели и конкордансы. Результаты использованы при подготовке учебника селькупского языка для педучилищ.

Работа секции Диалектологического фонда (ДФ) показала, высокий уровень взаимопонимания и координации усилий основных разработчиков: представителей Отдела диалектологии и лингвогеографии АН СССР, Саратовского и Сыктывкарского ГУ. Общая концепция Диалектологического фонда была изложена в докладе Н. Н. П ш е н и ч н о в о й. Основными составляющими ДФ должны быть: словарный, текстовый, справочно-грамматический и лингвогеографический подфонды. В рамках последнего сотрудниками Отдела диалектологии и лингвогеографии разработано лингвистическое обеспечение Автоматического варианта ДАРЯ (программное обеспечение разработано Г. А. Черкасовой). Детальный проект текстового подфонда, центрального для ДФ в целом, был представлен в докладе В. Е. Гольдина (Саратов). Доклады представителей Сыктывкарского ГУ (О. В. Загоровская, С. В. Лесников и др.) по Автоматизированному словарю русских народных гороров могут служить основой проекта словарного подфонда ДФ.

На секции Фонетических фондов (ФФ) продолжалось обсуждение концепций ФФ в составе МФ РЯ. Одна из них представлена разработками кафедры фонетики ЛГУ (руководитель — Л. В. Бондарко). В соответствии с этой концепцией ФФ должен включать фонетическую базу данных, т. е. зафиксированные на магнитной ленте звуковые реализации, а также сведения о фонемном составе значимых единиц — морфем и словоформ, извлекаемые из машинных версий словарей морфем и словоформ, и, наконец, автоматический транскриптор, позволяющий представить орфографический текст в виде звуковых последовательностей (с разной степенью подробности). Более широкое понимание структуры ФФ как открытой системы, которая может пополняться различными компонентами в соответствии с возникающими исследовательскими задачами, было предложено в докладе О. Ф. Кривновой и Н. В. Зиновьевой (МГУ). ФФ, по мнению докладчиков, должен включать базу фонетических знаний, базу фонетических данных, инструментально-техническое обеспечение и корпус образцов звучащей речи (фонотеку). Представители Лаборатории экспериментальной фонетики АН СССР (Р. Ф. Касаткина и др.) в своих докладах подчеркнули необходимость включения в состав ФФ большого количества записей спонтанной речи, полученных от носителей региональных вариантов русского языка и диалектов.

В докладах секции Терминологических фондов нашли отражение разработки в области создания баз знаний на основе крупных терминологических банков данных. Эти системы не являются компонентами МФРЯ, но имеют для его развития большое значение. Так, Автоматизированный банк данных по стандартизованной терминологии (СОВТЕРМ) является развитием системы АСИТО, работающей во ВНИИКИ. СОВТЕРМ должен обеспечивать одновременную обработку не менее 300—400 тыс. терминологических статей и ежегодное пополнение базы данных не менее чем на 25—30 тыс. статей (докладчики И. Н, Волкова, В. А. Гарбарчик). Большой интерес вызвал комплекс терминологических банков данных по картографии, разрабатываемый в Горьковском ГУ (докладчики Р. Ю. Кобрин и др.). Система принимает запросы по терминологии картографирования и по «языку карты» (совокупности терминов и условных знаков, описывающих объекты действительности и отношения между ними, а также пространственное положение объектов) и обладает чертами разработки по искусственному интеллекту.

На секции Текстовых фондов большое внимание было уделено вопросам программного и информационного обеспечения соответствующих разработок. Общие перспективы развития Текстового фонда МФ РЯ были определены в пленарном докладе А. С. Герда (Ленинград) «Автоматизированные лексические базы данных ИРЯ как единое целое». По мнению докладчика, основой МФ РЯ должны быть базы данных полных текстов, вводимых без предварительной обработки. Докладчик предложил список из 10 разножанровых баз, которые должны быть созданы в течение пяти лет. Центральное место среди них занимает сводный иллюстрационно-текстовый фонд источников нового академического словаря (Словарь второй половины XX в.).

Пленарный доклад С Е. Никитиной (Москва) «Фольклорные тексты в Машинном фонде русского языка» содержал характеристику одного из текстовых фондов. В его составе — произведения двух фольклорных жанров: духовные стихи и свадебные причитания, общим объемом около 80 тыс. словоупотреблений. Предложен ряд задач, которые могут быть поставлены на этом материале. Центральная из них — составление семантического словаря фольклора тезаурусного типа. Работа над ним активно ведется с использованием тексто-ориентированной компоненты автоматизированной системы «УНИЛЕКС». . Данной системе был посвящен отдельный доклад. Разработчик, Ж. Г. Мошкович (НИВЦ МГУ), познакомила собравшихся с основными возможностями пакета. Он обеспечивает полный цикл машинной обработки текстов: составление частотных словарей, словоуказателей и конкордансов, а также формирование базы данных, позволяющей работать с текстами в режиме «запрос — ответ». В качестве запроса может быть задано слово, словосочетание или группа слов; пользователь-лексикограф может варьировать такие параметры, как размер контекста, число и область поиска контекстов и др.

Конференция закончилась принятием решения, вобравшего в себя итоги многочисленных дискуссий и обсуждений. В нем отмечено существенное продвижение в разработке многих компонент МФРЯ (прежде всего словарных) и выделено главное направление дальнейшей работы: «создание систем автоматизации лингвистических исследований в институтах Академий наук и вузов, основой которого является накопление источников для изучения языков — текстов, словарей, грамматик, диалектного и фольклорного материала, памятников письменности».

СПИСОК ЛИТЕРАТУРЫ

1. Третья Всесоюзная конф. по созданию Машинного фонда русского языка: Тез. докл. Ч. 1—2. М., 1989.

Кукушкина Е. Ю. (Москва)

22 ноября 1988 г. в Словарном отделе Института языкознания АН СССР в Ленинграде состоялось совещание по вопросам русской исторической региональной лексикографии, организованное Комиссией лингвоэтногеографии Географического общества СССР совместно со Словарным отделом и Межкафедральным Словарным кабинетом им. Б. А. Ларина (Ленинградский ун-т). В совещании приняли участие лексикологи из Москвы, Ленинграда, Петрозаводска, Вологды, Арзамаса, Перми, Красноярска, Хабаровска.

Совещание открыл заведующий Словарным отделом А. С. Герд. Актуальность обсуждаемой проблематики, как подчеркнул председательствующий С. С. Волков (Ленинград), определяется в первую очередь тем, что исторические словари являются той базой, без которой невозможно создание исторической лексикологии русского языка.

В ходе совещания было заслушано семь докладов. О теоретических проблемах построения регионального словаря говорилось в докладе Г. В. Судакова (Вологда), посвященном изложению принципов создания Словаря Северной Руси **XIV—XVII** вв. и проспекта Вологодского исторического словаря как его составной части. На необходимость создания региональных словарей для воссоздания исторического ландшафта русского языка и в качестве базы для будущего общерусского регионального словаря XI—XVII вв. указала в докладе «О словаре воронежской деловой письменности» В. И. Дьякова (Москва), изложившая принципы построения Воронежского регионального словаря. В ряде докладов прослеживалось стремление провести границу между диалектными и региональными словарями (Г. В. Судаков, О. С. Мжельская, выступление З. М.Петровой). Различие между диалектизмами и регионализмами четко определила в докладе «Региональная лексика как предмет исторической лексикографии» О. С. Мжельская (Ленинград): если диалектизмы составляют принадлежность современных говоров, то регионализмы созданы в определенный исторический период и актуальны до тех пор, пока существуют обозначаемые ими реалии.

В центре внимания участников совещания оказались такие важные вопросы, как отбор и характер источников региональных словарей, состав их словников, пути создания картотек и т. д. Региональная лексикография, как отметила З. М. Петрова (Ленинград), значительно распшряет круг источников. Как свидетельствовали доклады, сегодня это и деловая письменность Сибири со време-

ни ее основания [об этих источниках, хранящихся в различных архивах страны, сообщила в своем докладе Л. М. Городилова (Хабаровск), выдвинувшая ряд требований, необходимых «при подборе рукописных материалов XVII в.], и памятники письменности Карелии XV—XVII вв. и т. д. Л. П. Михайлова (Петрозаводск) обратила внимание присутствующих на лексические данные не изучавшихся до сих пор документов Олонецкой приказной избы, Соловецкого и Палеостровского монастырей. Лексика этих памятников далеко не всегда находит отражение в издаваемых Институтом русского языка АН СССР Словаре XI—XVII вв. и Словаре русских народных говоров. Мнения по поводу источников свелись в целом к общему знаменателю: круг источников должен быть максимально широк и разнообразен как в жанровом, так и в тематическом отношениях.

Наиболее оживленную дискуссию вызвал, пожалуй, вопрос о том, каким быть региональному словарю: полным или дифференциальным. Так, Г. В. Судаков, В. И. Климкова, Л. А. Дьякова (как и З. М. Петрова в выступлении по докладам) отстаивали идею дифференциального словаря, тогда как Л. М. Городилова и выступавший в прениях С. С. Волков считали полный словарь предпочтительнее, хотя бы уже потому, что именно полный словарь способен отразить варианты, не зафиксированные словарем XI—XVII вв.

На необходимость этимологических исследований применительно к региональному словарю указал С. С. Волков в связи с интересным докладом Л. А. Климковой (Арзамас), посвященным проблеме составления микротопонимического словаря южных районов Горьковской области. Проблематика этого доклада в известной мере пересекалась с темой сообщения И. А. Кюршуновой (Петрозаводск), в котором затрагивался вопрос о номенклатурной топонимии старорусского языка по памятникам письменности Карелии. Как отметил Г. В. Судаков, обе докладчицы обратились к материалу, крайне редко привлекаемому составителями региональных словарей, что определило несомненную актуальность обоих выступлений.

Обсуждение докладов выявило общее стремление к координации усилий лексикографов и конкретизации дальнейшей работы.

Г. А. Богатова (Москва) вновь поддержала обсуждавшееся ранее на совещаниях в Вологде и Днепропетровске предложение о создании сборников по материалам совещаний, что дает широкие возможности в плане обмена мнениями. В связи с этим А. С. Герд выдвинул встречное предложение: включать в подобные сборники не авторские статьи, а суммарные обзоры по проблемам региональной и диалектной лексикографии. Идею создания сборников поддержал и С. С. Волков, отметив, что они представляют интерес для преподавателей вузов.

Чрезвычайно наболевший вопрос о статусе лексикографа затронула в своем выступлении Г. А. Богатова. Помочь его решению может недавно созданная Ассоциация лексикографов. Изменение статуса лексикографа безусловно скажется на интенсивности развития теории лексикографии. В выступлении Г. А. Богатовой прозвучала также мысль о необходимости технического совершенствования работ по созданию словарных картотек.

Если проблематика докладов была связана с исторической региональной лексикографией, то свободный обмен мнениями за Круглым столом касался одной из актуальнейших задач современной науки о русском языке: создания академической исторической лексикологии. Круг вопросов был намечен заранее в анкете, опубликованной к совещанию в Вологде. Уже в Вологде и Днепропетровске в первом приближении намечался облик будущего исследования. Собравшихся за Круглым столом в Ленинграде занимали в первую очередь такие проблемы, как основной объект исторической лексикологии, хронологические границы исследования, проблемы взаимоотношения общерусской и местной лексики, междисциплинарные связи, вопрос о группах слов, подлежащих изучению, и т. д. Понятно, что работа над таким фундаментальным трудом требует уже на первом этапе решения весьма непростых организационных вопросов. Участники Круглого стола сочли целесообразным принять предложение А. С. Герда о создании неформальной межведомственной комиссии со своим секретариатом в Москве и в Ленинграде, которая могла бы развернуть работу по созданию проекта и макета будущего исследования, координируя работу на местах. В организационном отношении подобная комиссия мобильнее, чем кафедры вузов или отделы академических институтов. При чрезвычайно сжатых сроках, отсутствии кадров, в условиях хозрасчета и самоокупаемости создание такой комиссии является единственно возможным путем решения поставленной задачи. Созданную по решению совещания комиссию возглавила Г. А. Богатова.

Этерлей Е. Н. (Ленинград).