ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 2 1985

АНДРЮЩЕНКО В. М.

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ПОСТАНОВКА ЗАДАЧИ И ПРАКТИЧЕСКИЕ ШАГИ*

1. Впервые задача создания машинного фонда русского языка была сформулирована А. П. Ершовым в докладе «К методологии построения диалоговых систем: феномен деловой прозы» в 1978 г. [1]. В связи с проблемой взаимодействия человека и ЭВМ на естественном языке, прежде всего в производственных и иных регламентирующих отношениях, А. П. Ершов поставил задачу научить машину полностью воспринимать и понимать деловую прозу как языковой носитель модели производственных отношений. Социально-экономическая посылка необходимости решения этой задачи состоит, говоря словами автора, в следующем: «Мне кажется нереальным насытить производственные отношения вычислительными машинами, расставляя во всех стыках бесчисленные интерфейсы в виде формализованных предписаний, стандартных бланков и других средств предварительной подготовки информации для машины. Время на обработку или синтез документа будет почти всегда соизмеримо со временем его исполнения, подрывая тем самым все выгоды автоматизации. Мы не можем внедрить машины в повседневную жизнь, выделяя касту жрецов-посредников, но в таком случае должны быть готовы к тому, что число человеко-машинных интерфейсов в диалоговых системах возрастет с течением предстоящих 50 лет на несколько порядков. Не хватит никаких сил на то, чтобы снабдить эти миллионы интерфейсов специализированными языками и процессорами» [2, с. 6].

Рассмотрев конструктивные следствия, вытекающие из этой посылки, а также накопленные знания и технологические приемы системного программирования в области построения языковых процессоров, А. П. Ершов пишет далее: «Любой прогресс в области построения моделей и алгоритмов останется, однако, академическим упражнением, если не будет решена наиважнейшая задача создания машинного фонда русского языка. Это фундаментальная проблема, решение которой будет иметь большую научную, общекультурную и прикладную ценность... Очень хотелось бы видеть, что создание машинного фонда русского языка квалифицированными лингвистами опережало бы создание производственных лингвистических систем, потому что это не только бы позволило избежать дублирования больших усилий, но и защитило бы здоровую ткань русского языка от самоуправства и неквалифицированного подхода» (разрядка наша.— *А. В.*) [2, с. 11].

Идея создания Машинного фонда русского языка как технологической основы для разработки систем общения с ЭВМ и обработки данных на естественном языке нашла отклик в среде системных программистов и разработчиков ЭВМ, послужила предметом дискуссий на конференциях и семинарах, была поддержана лингвистической общественностью. Языковедческая мысль связывает с этой идеей возможности подлинной и комплексной автоматизации лингвистических исследований и разработок, прежде всего в области лексикографии, и на этой основе — возможности

* В основу данной статьи положен доклад автора «Концепция и архитектура машинного фонда русского языка», прочитанный им на Конференции по проблемам создания машинного фонда данных для автоматизированной системы лексикографических исследований (Москва, 21—23 февраля 1983 г.), а также материалы состоявшейся дискуссии.

более широкого участия лингвистов в решении задач, сформулированных в Постановлении ЦК КПСС и Совета Министров СССР «О мерах по дальнейшему ускорению научно-технического прогресса в народном хозяйстве» от 18 августа 1983 года (№ 814) (опубликовано в Правде 28 августа 1983 г.). Актуальность проблемы автоматизации лингвистических исследований с каждым днем возрастает: сегодня эта проблема может быть также рассмотрена в связи с обсуждением проектов ЭВМ пятого поколения, которые должны обладать встроенными языковыми процессорами и программно-аппаратными средствами обеспечения банков знаний [3].

Осенью 1982 г. Научный совет по лексикологии и лексикографии АН СССР совместно с Секцией лингвистических проблем обработки информации Научного совета по комплексной проблеме «Кибернетика» АН СССР распространил вопросник, посвященный проблемам создания Машинного фонда русского языка, с целью собрать и систематизировать мнения специалистов. Полученные ответы на вопросы, записи бесед и обсуждений, замечания к предложениям Комиссии по машинным фондам языка помогли сформулировать излагаемую ниже концепцию Машинного фонда русского языка [1].

Прошедшая в феврале 1983 г. Конференция по проблемам создания Машинного фонда русского языка и обсуждение результатов этой конференции и выработанных по ее решению предложений на заседании Бюро Отделения литературы и языка АН СССР показали глубокую заинтересованность академических и вузовских языковедов в реализации этой программы, их готовность к освоению технологических средств «безбумажной информатики» [см. 4] с целью повышения производительности своего труда, более широкого участия в решении прикладных задач современности.

2. Предложения по созданию Машинного фонда русского языка определяют фонд как систему комплексной автоматизации научных исследований и прикладных разработок в области языкознания, опирающуюся на ряд лингвистических банков данных, реализуемых на средних и мини-ЭВМ в нескольких наиболее крупных академических и отраслевых институтах и вузах под единым координирующим руководством Института русского языка АН СССР. Предусматривается создание следующих фондов-составляющих:

Генеральный словник Машинного фонда русского языка;
Иллюстрационно-текстовой фонд русского языка;
Терминологический фонд русского языка;
Академический словарно-грамматический фонд русского языка;
Лексикографическая база Машинного фонда русского языка;
Лингвостатистическая база Машинного фонда русского языка;
Фонд процессоров русского языка;
Фонд лингвистических алгоритмов и программ;
Информационно-справочный фонд по русистике.

Каждый из этих фондов-составляющих может быть организован как подсистема единой системы, называемой Машинный фонд русского языка,

[1] Ответы на вопросы по проблемам создания Машинного фонда русского языка прислали А. Б. Антопольский, Ю. Д. Апресян, М. А. Балабан, К. Б. Бектаев, Л. Н. Беляева, И. Г. Бидер, И. А. Большаков, В. В. Бородин, А. С. Герд, Б. Ю. Городецкий, В. П. Григорьев, Т. А. Грязнухина, А. В. Зубов, Л. П. Крысин, Н. Н. Леонтьева, М. Г. Мальковский, Т. М. Николаева, Ю. К. Орлов, В. И. Перебейнос, Р. Г. Piотровский, Р. П. Рогожникова, Ю. В. Рождественский, В. Ю. Розенцвейг, А. А. Романовский, В. Ш. Рубашкин, Л. В. Сахарный, В. Н. Телия, Б. В. Сухотин, Г. С. Цейтин, А. Я. Шайкевич, З. М. Шаляпина, Д. Н. Шмелев, И. Б. Штерн. Предложения Комиссии автор обсуждал с А. П. Ершовым, Ю. Н. Карауловым, Ю. Д. Апресяном, И. А. Большаковым, А. С. Гердом, О. С. Кулагиной, М. Н. Реммелем, Г. С. Цейтиным, мнения которых оказали наибольшее влияние на формирование точки зрения автора, который стремился в той или иной мере учесть и обобщить все предложения и замечания. Автор искренне благодарен всем участникам обсуждения этой проблемы.

и состоять из частных баз лингвистических данных, таких, как собрание полных авторских текстов, статистических выборок из текстов, цитат и т. д. в составе Иллюстрационно-текстового фонда; собрание информационно-поисковых тезаурусов, стандартизованной лексики, отраслевых научно-технических номенклатур, лексики, зафиксированной наиболее крупными русскими словарями, такими, как БАС, МАС, словари В. И. Даля, Д. Н. Ушакова, С. И. Ожегова; справочный фонд, образованный рядом ортологических словарей и справочников, справочный фонд русских академических грамматик — в составе Академического словарно-грамматического фонда; собрания статистических данных о лексике, грамматике, текстообразовании в составе Лингвостатистической базы Машинного фонда русского языка.

В состав Академического словарно-грамматического фонда должны войти наряду с данными о современном языке данные о лексике и грамматике предшествующих периодов развития русского языка, соотнесенные с текстами памятников, образующих отдельные базы данных в составе Иллюстрационно-текстового фонда, а также данные диалектологии, социо- и психолингвистики, соотнесенные со средствами обработки анкетных форм в составе Информационно-справочного фонда.

Фонд процессоров русского языка может быть образован системами программ, формальных словарей и формальных грамматик русского языка, разработанных в наиболее развитых системах автоматического перевода и в системах общения с ЭВМ на русском языке. Языковые процессоры, т. е. программы автоматического анализа и синтеза русского текста, будут использоваться и как средство автоматизации лингвистических, прежде всего лексикографических, работ и будут развиваться в составе Машинного фонда русского языка как его собственные продукты, предназначенные для поставки заинтересованным организациям в качестве лингвистических подсистем автоматизированных систем управления, обработки информации и проектирования. Лингвистическое обеспечение процессоров русского языка — формальные словари и грамматики — будет служить также образцом для постепенного преобразования основных данных о современном русском языке в форму автоматических словарей и грамматик, используемых лингвистическими программами.

Фонд лингвистических алгоритмов и программ может быть образован находящимися в эксплуатации и созданными в рамках Машинного фонда русского языка программами и руководствами по их использованию, предназначенными для автоматизации основных видов работ,— текстовыми редакторами, автокорректорами, программами издательской подготовки, синтеза табличных форм, статистической обработки языкового материала, обучения русскому языку, анализа и синтеза русской речи [5].

Информационно-справочный фонд по русистике должен состоять из нескольких информационных систем: библиографической системы, документальной системы для учета документации самого фонда, информатора фонда, систем обработки анкетных форм и др.

Лексикографическая база Машинного фонда русского языка может быть образована типовой автоматизированной лексикографической системой, снабженной основными программными средствами автоматизации лексикографических работ: средствами формального проектирования словарей, образования словника, подбора иллюстрационного материала, процессорами русского языка и др.

Источником для образования словников новых словарей должен служить Генеральный словник Машинного фонда русского языка, в основу которого может быть положен Сводный словник словарей русского языка, составленный коллективом сотрудников Словарного сектора ЛО Института языкознания АН СССР [6]. В дальнейшем состав сводного словника может пополняться путем учета в нем всех слов, адресующих словарные модули всех компонентов Машин-

ного фонда русского языка. Генеральный словник должен содержать наряду со сведениями о вхождении слов в те или иные словари и частные фонды также сведения, необходимые для порождения формальных частей словарных статей, зависящих от характера соответствующей вокабулы — часть речи, лексико-грамматический класс, состав форм, если необходимо, исключения и нерегулярности форм, общенормативные сведения и др.

Особенностью излагаемого проекта создания Машинного фонда русского языка является то, что он должен реализовываться как распределенный банк лингвистических данных, т. е. его компоненты могут физически находиться в разных местах и дублироваться в соответствии с потребностями их использования. Это относится также и к пользовательским базам данных, которые будут создаваться в рамках фонда, в качестве продукта исследователей, работающих в фонде. Это означает, что вновь созданные в среде фонда словари и грамматики как бы автоматически включаются в фонд, расширяя его возможности. Единство фонда будет поддерживаться единством проекта, регламентом обмена информацией, однородностью систем управления базами данных, соглашениями о применении технических и программных средств. Предложения включают также задачи по проектированию машинных фондов национальных языков народов СССР. Проект создания Машинного фонда русского и национальных языков в полном объеме рассчитан на 15 лет, в течение которых во всех институтах языкознания и крупных вузах должны быть внедрены типовые системы комплексной автоматизации лингвистических работ и информационно-справочного обслуживания языковедов, средства автоматизированной подготовки изданий и автоматизации редакционно-издательского процесса на базе мини- и микро-ЭВМ.

Изложенная концепция и архитектура Машинного фонда русского языка является обобщением многочисленных работ, проводившихся как в нашей стране, так и за рубежом в области автоматизации лингвистических исследований, автоматического перевода, автоматической лексикографии, автоматизированного терминологического обслуживания, автоматической обработки текстов и общения с ЭВМ на естественном языке. Основной предпосылкой создания Машинного фонда является дальнейшее согласованное ведение этих работ в рамках единой программы, возможность обмена данными и программами, возможность комплектации этих материалов в пакеты программ, в автоматизированные системы обработки лингвистических данных, в лингвистические базы и банки данных. В настоящее время в различных организациях накоплены на машинных носителях значительные текстовые и словарные фонды на русском языке, имеются разнообразные программы обработки лингвистических данных. Эти программы и данные при известных организационных и технологических условиях могут быть объединены в фонды, аналогичные фондам алгоритмов и программ, и, таким образом, могут стать основой для дальнейшего развития перечисленных выше компонентов Машинного фонда.

3. Изложенный проект создания Машинного фонда русского языка направлен на решение следующих основных и взаимосвязанных задач:

Создать возможность эффективной централизованной разработки и поставки промышленности и НИИ лингвистического обеспечения для разрабатываемых систем общения с ЭВМ на русском языке и систем обработки документов в естественной языковой форме [2];

---

[2] Одной из основных характерных черт проекта ЭВМ пятого поколения является ориентация на взаимодействие пользователя и ЭВМ на естественном языке. В число базовых прикладных систем этих машин планируется включить многоязычные системы автоматического перевода со словарем объемом порядка 100 тыс. словарных статей, запросно-ответные системы с естественно-ограниченным языком общения, системы понимания письменной и устной речи. Это означает, что в ближайшие 10—15 лет потребуется формализованный словарь русского языка, сравнимый по своему объему с БАС, и формальная грамматика, сравнимая по полноте с Академической грамматикой русского языка. Такие труды не могут быть подготовлены лишь силами отраслевых подразделений, необходима надотраслевая лингвистическая организация, ответственная за разработку лингвистического обеспечения ЭВМ пятого поколения.

составления словарей, поиска и обработки научной информации, анализа текстов, проведения классификационных работ, подготовки аппарата изданий и т. п.;

заложить на машинных носителях сокровищницу данных о русском языке во всем объеме этого понятия, во всех его временных и территориальных формах.

Переход к новым методам сбора, хранения, анализа и сопоставления данных о языке, новые методы создания и новые формы лингвистических источников, таких, как автоматические словари и грамматики, могут быть жизнеспособными и эффективными, если они опираются на общую филологическую традицию и культуру, на глубокое изучение языка и учет информации о нем во всех формах его существования. Однако соединение традиции и новых задач практики нужно искать на путях новой информационной технологии, развиваемой системами обработки данных на естественном языке в интеллектуальной среде человеко-машинного общения.

Представляется необходимым уже на ранних этапах осуществления этого проекта зафиксировать те исходные теоретические установки, которые будут определять в дальнейшем его лицо и сущность. В основе проекта Машинного фонда русского языка лежат идеи «безбумажной информатики», соединенные с принципами концепции «лингвистического конструирования», как они изложены в книге [7, с. 16—17]: «Лингвистическое конструирование — это совокупность обобщенных способов и приемов компиляции и комбинирования „образцов решений проблем", экстраполяции уже имеющихся, готовых теоретических и практических результатов, полученных в разных областях лингвистики, и их прямого или эвристического использования для преодоления трудностей и решения проблем, возникающих в тех же или других областях при построении новых лингвистических объектов. Создать, построить какую-то „вещь", значит не только уметь объяснить те свойства языка, которые в ней использованы и на которых она основана, не только объяснить определенные закономерности языковой структуры, но и выяснить новые свойства построенного объекта, так или иначе характеризующие исследуемый язык, а значит — расширить наши знания о человеческом языке вообще. Таким образом, новым лингвистическим объектом будем называть такое представление фактов, языковых данных, которое генерирует новую информацию о языке. Как правило, такие объекты получаются не в результате описания некоторого материала, а возникают как результат эксперимента, причем эксперимента, понимаемого широко. В этом смысле и нестратическую теорию, и порождающую грамматику, имеющие целью каждая создание нового объекта, следует трактовать как лингвистический эксперимент».

Концепция «лингвистического конструирования» связывает воедино две цели: удовлетворение насущных потребностей практики использования языка в компьютерных системах и развитие самой лингвистики путем преодоления ее собственных трудностей и противоречий. О наличии таких трудностей свидетельствует тот факт, что наиболее фундаментальные лингвистические труды, каковыми являются академические словари и грамматики, оказываются в наименьшей степени пригодными в качестве источников для словарей и грамматик прикладных систем обработки данных на естественном языке. Это приводит к тому, что в прикладных областях возникает собственная «кибернетическая» лингвистика, собственная «машинная» лексикография, собственное отраслевое терминоведение и т. д. Причины этого явления следует искать в сложившемся противоречии между реальным атомизмом исследовательской практики и уже хорошо осознанной целостностью и системностью лингвистического объекта. Не раз отмечалось в литературе, что ни в одном словаре не удается последовательно провести принцип системности описания лексики даже в пределах уже известных приемов ее отражения — это, несомненно, связано с атомистическим подходом к конструированию словарных статей [см. 7,

с. 24; 8]. Несмотря на то, что давно осознана и теоретически изучена динамичность языковой синхронии, в рамках теоретической лингвистики нет ни одной реальной динамической грамматики языка — это, несомненно, связано с «маломерностью» лингвистического восприятия: чтобы понять конструктивную суть генеративной лингвистики и развить далее этот аппарат, необходимо было выйти в вычислительный эксперимент, размерности которого превышают обычную в исследовательской практике двумерность бумажного листа [3].

Переход к использованию ЭВМ, т. е. к «безбумажной информатике», снимает противоречие двумерности бумаги как традиционного носителя лингвистической информации и многомерности лингвистического объекта; противоречие статичности описания и динамичности существования объекта и результатов конструирования; реальности традиционного результата как непременного условия существования научной истины и виртуальности существования объектов, описываемых порождающими механизмами, алгоритмами и реализующими их программами; исторически обусловленной кумулятивности лингвистических данных и избирательности их использования.

Последнее противоречие характерно для лексикографии. Ф. П. Сороколетов писал в [10, с. 40], что понятие системы словарей в советском языкознании возникло и разрабатывалось в связи со стремлением сохранить гуманитарную традицию, ставящую естественный предел допустимым объемам словарных статей и их сложности. Это приводит к физическому многообразию типов словарей, каждый со своим составом лексикона, схемами словарных статей и рубриками лексикографического описания. В свое время лексикографическая концепция системы словарей, которые «в совокупности... должны выполнить задачу, которую ставил перед своим словарем-тезаурусом А. А. Шахматов» [10, с. 41], стала значительным достижением лексикографической мысли. Действительно, собрать в одном словаре, в одной словарной статье все мыслимые данные о каком-либо слове и при этом согласовать между собой структуры словарных статей по всем вокабулам не представляется ни возможным, ни практичным с точки зрения использования словаря. С другой стороны, использовать в практике множество словарей, рассогласующихся между собой в составе словников и в методах подачи, не менее непрактично, а создать серии согласующихся между собой словарей по всем параметрам еще менее возможно, чем один словарь-thesaurus в смысле А. А. Шахматова. Ярким примером тому является наличие многих ортологических словарей, изобилие которых вряд ли намного лучше, чем их недостаток [11].

Переход к методам «безбумажной информатики», к средствам современной автоматической лексикографии позволяет в е р н у т ь с я к ш а х м а т о в с к о й т р а д и ц и и, спрятать внутрь базы данных всю действительную сложность и объемность описания, сделав для пользователя «видимой» каждый раз ту часть thesaurus'а и в том представлении, которое ему необходимо и соответствует его лексикографическому восприятию. В автоматической лексикографии понятию типа словаря соответствует понятие режима обращения к нему: путем ограничений на выдачу словарных статей и их компонентов словарь может быть во внешней форме представлен в нужном объеме (большой, средний, малый) и в нужном аспекте (толковый, перево́дный, семантический, словообразовательный и т. д., синонимов, антонимов, конверсивов, фразеологизмов и т. д.).

Дальнейшую конкретизацию соединение идей «безбумажной информатики» и «лингвистического конструирования» находит в систематическом применении концепции реляционных баз данных [см. 12] и тезиса о лексикографируемости любого языкового факта [13]. Этот тезис обоснован тем, что любой факт может быть представлен в форме его имени, соединенном с набором значений его атрибутов. Следовательно, понятие факта обра-

[3] Результатом является то, что все современные эффективные методы полного автоматического грамматического разбора оказались основанными на идеях порождающей и трансформационной грамматики и являются дальнейшим развитием этих идей [см. 9].

зует отношение, т. е. множество кортежей, состоящих из значений атрибутов данного факта, так что каждый кортеж является определенной реализацией соответствующего факта [4]. В этом легко усматривается аналогия со словарем, для которого фактом является слово, а атрибутами факта — конкретные реализации слова как вокабулы с определенными дефинициями, в определенных контекстах и т. д. Словарная статья есть кортеж из отношения СЛОВО, доменами которого (отношения) являются лексикографические параметры, описанные в книге [7, с. 75—77]. Если каждый факт представим как отношение, то он представим также и в лексикографической форме. Должно быть верно и обратное: любая лексикографическая форма представима в виде реляционной базы данных. Следовательно, и Машинный фонд русского языка может быть в своей лексикографической части сконструирован в виде системы реляционных баз данных, согласованных между собой именами отношений и доменов, а также значениями доменов. Более того, операции над отношениями, используемыми в реляционных базах данных, хорошо моделируют лексикографические работы, такие, как отбор словника (операции ВЫБОР или ДЕЛЕНИЕ), компоновка словарной статьи из имеющихся источников (операции СОЕДИНЕНИЕ и ПРОЕКТИРОВАНИЕ), добавление и исключение словарных статей (операции ОБЪЕДИНЕНИЕ и РАЗНОСТЬ) [см. 12, с. 104—109]. Задавая определенные модификации этих операций, можно получить реальные лексикографические операции, такие, как ПЕРЕУПОРЯДОЧЕНИЕ, КОРРЕКЦИЯ, ПЕРЕСТАНОВКА и ИСКЛЮЧЕНИЕ компонентов словарных статей и т. д.

Однако реализация реляционных лексикографических баз данных требует от лексикографии четкой разработки не только полного перечня параметров (доменов) лексикографических отношений и полного набора значений каждого домена, но и полного перечня и описания видов лексикографических работ (операций).

Реляционный подход к автоматической лексикографии позволяет также разработать методы автоматизации проектирования словарей и самого Машинного фонда путем использования формальных описаний словарей в программах, реализующих операции над словарями: их размещение в базах данных, поиск словарных статей, их коррекцию, слияние и переупорядочение и т. д.

В отличие от традиционных реляционных систем управления базами данных, в лексикографических системах требуются и необычные операции. Речь идет о специфических для лингвистики проверках условий и специфических операциях преобразования лингвистических объектов. К ним относится прежде всего лемматизация, т. е. перевод словоформы в вокабульное представление, проверка условий реализации синтаксической связи и семантических ограничений и т. д. С наибольшей полнотой эти методы в алгоритмическом плане сейчас разработаны в некоторых системах автоматического перевода и общения с ЭВМ на естественном языке [см. 14, 15]. Поэтому представляется необходимым расширить систему реляционных операций процедурами, заимствованными из этих реализаций.

Проектирование реляционных лексикографических баз данных для Машинного фонда русского языка ставит перед вычислительной лексикографией задачу конструирования автоматизированных лексикографических систем (АЛС) нового типа.

До сих пор практически освоенными являются лишь тексто-ориентированные и информационно-справочные АЛС [см. 16], реализующие автома-

[4] В сторогом определении отношением R_n называется декартово произведение множеств $D_1 \times D_2 \times \ldots \times D_n$, не обязательно различных. Сами множества $D_1, D_2 \ldots, D_n$ называются доменами отношения R_n. Отличие домена от атрибута состоит в том, что атрибут представляет использование домена внутри отношения. Атрибуты, используемые в отношении, уникальны, но двум или более атрибутам может соответствовать один домен. Так, слово может иметь несколько значений, каждое из которых принадлежит своему атрибуту отношения СЛОВО, но все значения слов входят в один домен ЗНАЧЕНИЕ. Понятие домена почти соответствует понятию лексикографического параметра, сформулированному в книге [7].

тизированное составление словоуказателей, частотных словарей и конкордансов и использование, в основном терминологических, словарей в автоматическом режиме [17—20]. В настоящее время в практике осваиваются инструментальные АЛС, в которых реализуется интерактивная подготовка и коррекция словарных статей [21]. В стадии проектирования находятся исследовательские АЛС, снабженные языковыми процессорами, которые позволят автоматически формировать образы словарных статей на основе автоматического анализа и синтеза текста. Для двуязычной терминологической лексикографии эта задача представляется достаточно реальной. Схема ее решения такова: берется представительная совокупность пар текстов, один из которых является переводом другого, и осуществляется машинный перевод «наоборот»: анализируются параллельно оба текста, результаты анализа сопоставляются, выявляются «переводные соответствия», из которых, а также из наличного цитатного материала, синтезируется образ словарной статьи согласно заданному формальному описанию ее структуры, причем для проверки может привлекаться какой-либо имеющийся словарь, по отношению к которому входной материал рассматривается в качестве дополнения.

Реализация таких систем позволит перейти к постоянному «ведению» словаря в машинном представлении, в котором он в каждый момент времени находится в готовом для полиграфического воспроизведения виде. Соединение АЛС с системами автоматического набора даст возможность резкого увеличения периодичности и снижения стоимости переизданий словарей, а переход в микрокомпьютерную (безбумажную!) форму существования словарей приведет к их массовому, индустриальному тиражированию в «карманном» исполнении.

Новым типом АЛС, который можно было бы назвать генерирующим, является тип, подсказываемый идеями «лингвистического конструирования» и реляционных баз данных. Действительно, если каждый языковой факт лексикографируем и каждый факт выступает как отношение, то число потенциальных типов словарей не меньше числа различных отношений, которые мы в состоянии сформировать на лингвистических данных. Систематически используя операции СОЕДИНЕНИЯ и ПРОЕКТИРОВАНИЯ над уже зафиксированными отношениями, можно получать разнообразные типы словарей, не меняя базы данных. Например, если в базе данных хранятся отношения словообразования, то из них выводимы словари морфов, основ, гнезд слов и т. д. Если хранится информация о синтаксических отношениях, то возможны словари словосочетаний, грамматических конструкций, управления, моделей предложения и т. д. Практически по любому лексикографическому параметру возможно формирование словарного входа и определение нового типа словаря.

4. На пути создания Машинного фонда русского языка лежит, конечно, и много трудностей и препятствий, которые нужно преодолевать.

Парадоксальным образом многих языковедов, для которых, очевидно, пишущая машинка является нормальным рабочим инструментом, смущает необходимость ручного ввода большого количества данных. Процедура ввода данных этим языковедам представляется чем-то в высшей степени не соответствующим их основной специальности. Большие надежды возлагаются на читающие автоматы будущего и даже на голографию (ср.: «Вероятно, большие успехи по механизации лексикографических работ будут достигнуты с привлечением голографии и созданием подлинно читающих автоматов, т. е. таких автоматов, которые смогут распознавать и обрабатывать любой книжный шрифт, любой почерк» [22, с. 24]. В связи с этим заметим, что реальные успехи не только механизации, но и подлинной автоматизации лексикографических работ были достигнуты без читающих автоматов. За 20 лет во Франции был создан машинный фонд французского языка, намного превышающий по своему объему картотеку цитат Словарного сектора. Это те самые 20 лет, в течение которых мы вполне удовлетворялись так называемой «малой механизацией» [22, с. 18].

Сегодня автоматические картотеки и словари стали реальностью [5], но, к сожалению, вне стен лингвистических институтов. Думается, что использование терминальных устройств непосредственно на рабочих местах, создание автоматизированных рабочих мест лингвиста (АРМЛ) [6], тесное сотрудничество с издательствами и типографиями — вот реальный путь накопления больших массивов данных о языке. Опыт показывает, что во многих случаях подготовка данных непосредственно исследователями — лучший способ обеспечить их высокое качество.

Высказываются также суждения о недостаточности дисплея как устройства для отображения картотечной информации. Действительно, типовые алфавитно-цифровые видиотерминалы сегодняшнего дня способны отображать на экране 8, 12, 16 или 24 строки по 80 символов. Однако не следует забывать о возможности выдачи синоптической распечатки любого объема, содержащей нужным образом упорядоченный цитатный материал, о возможности использования графопостроителей для компоновки словарного материала сложной структуры на листах большой площади.

Технология автоматизации в лексикографии не может быть создана умозрительным путем: необходимо интенсивное внедрение всех подходящих средств обработки информации и реальная лабораторная отработка наиболее удобных технологических приемов.

Проект Машинного фонда русского языка не предусматривает перевода в память ЭВМ существующих картотек, однако новые картотеки предлагается вести машинным способом, ибо ЭВМ — это и есть в сущности автоматизированная картотека огромной емкости. В машинной среде теряет смысл цитатная карточка как физическая единица: она может быть сгенерирована в любой момент на основе имеющегося в памяти текстового материала, выдана на экран или на бумажный носитель при любом объеме цитируемого материала.

Что касается существующих картотек, то нам представляется целесообразным постепенный перевод их, с целью сохранения и лучшего использования, на фотоносители, хранилище которых может управляться ЭВМ, что обеспечивает автоматизированный или автоматический доступ к информации непосредственно с рабочего места. Такая же схема, видимо, наиболее целесообразна и для хранения памятников, и для автоматизации работ по составлению словарей.

В обсуждениях проблем создания Машинного фонда русского языка большое место занимали также проблемы шрифтов и шрифтовых выделений. Действительно, современные устройства ввода и вывода информации в большинстве случаев используют лишь один шрифт со сравнительно небольшим числом знаков и только заглавные литеры [7]. Эти трудности со временем, конечно, будут преодолены. Для вывода информации они преодолимы уже и сейчас, даже тремя способами: выводом на фотонабор, или на программируемые знакосинтезирующие или растровые устройства печати, или на графопостроители. Для ввода информации пока единственным методом является кодирование отсутствующих знаков и управляющих символов для знаковыделения с помощью комбинаций знаков, имеющихся в клавиатуре, или путем использования функциональных клавиш. Поэтому для лингвистических работ, налагающих жесткие ограничения на знаковый состав текста или его расположение, представляется целесообразным преимущественное использование терминального оборудования ЭВМ, а не устройств подготовки данных

[5] В мире сейчас насчитывается более 50 автоматизированных терминологических банков данных и автоматических словарей [см. 23], из них 10 содержат около 100 тыс. словарных статей, а три — более 1 млн. описаний терминов; существует несколько издательских систем подготовки словарей, начинается выпуск обычных словарей на машинных носителях.

[6] Устройство чтения и копирования микрофиш ISI-4000, управляемое ЭВМ, способно хранить в своей памяти 75 тыс. страниц. АРМЛ целесообразно комплектовать таким устройством, терминалом для связи с ЭВМ и персональным микрокомпьютером для ведения индивидуальных баз данных исследователя.

[7] Впрочем, уже начался массовый выпуск дисплеев, имеющих наборы заглавных и строчных литер; очевидно, что этим принципиально решена также задача смены наборов литер, что дает возможность отображения более 200 различных алфавитов.

Создание Машинного фонда русского языка будет ускорено, если ведущие лингвистические институты Академии наук уже сейчас будут оснащаться вычислительными комплексами, создаваемыми на базе мини-ЭВМ, расширенных внешней памятью на магнитных дисках большой емкости и сопряженных с фотонаборными автоматами [8]. В связи с этим возникают естественно проблемы стоимости и финансирования затрат на оборудование и его эксплуатацию. И здесь надо признать, что академические лингвисты, в отличие от вузовских, мало используют возможности хоздоговорных работ, возможности хозяйственного расчета. Думается, что заказы отраслевых НИИ и издательств на выполнение хоздоговорных работ на научные исследования, на подготовку словарей и грамматик уже сейчас способны дать значительные средства для оснащения академических институтов вычислительной техникой.

Такой союз — в интересах и промышленности, и языкознания. В связи с этим уместно процитировать слова В. М. Глушкова: «Одним из наиболее важных результатов научно-технической революции является бурная „компьютеризация" практически всех областей человеческой деятельности. Развитие сетей ЭВМ и систем терминального доступа к ним приводит к тому, что все большая часть информации, прежде всего научно-технической, экономической и социально-политической, перемещается в память ЭВМ.

Большинство выполненных к настоящему времени прогнозов сходится на том, что к началу следующего столетия в технически развитых странах основная масса информации будет храниться в безбумажном виде — в памяти ЭВМ. Тем самым человек, который в начале XXI века не будет уметь пользоваться этой информацией, уподобится человеку начала XX века, не умевшему ни читать, ни писать» [4, с. 7]. Это означает, с одной стороны, что основная масса материала для исследований по современному языку, независимо от воли и желания языковедов, будет находиться в компьютерной форме, а большая ее часть никогда не увидит бумаги. Языкознание может попросту потерять основной источник знаний о современном языке, если вовремя не перейдет к компьютерным формам обработки источников. Более того, основные продукты лингвистического труда — словари и грамматики — потеряют ценность, если они не станут доступными для использования в компьютерной форме, даже в издательской практике, не говоря уже о системах автоматизированного обучения и обработки данных на естественном языке. С другой стороны, в участии языковедов в создании систем обработки информации на естественном языке и систем человеко-машинного общения должны быть заинтересованы разработчики таких систем, ибо чем больший языковой материал охватывают такие системы, тем большее место в них начинают занимать чисто лингвистические проблемы. Одной из самых серьезных проблем уже сейчас является оперативное пополнение словарей таких систем в процессе эксплуатации. Здесь автоматизация лексикографической работы приобретает форму производственной службы по сопровождению лингвистического обеспечения. Элементарный анализ этой проблемы показывает, что для ее решения необходимы локальные машинные фонды языка, ресурсы которых должны оперативно использоваться для пополнения словарей в ситуации, когда система встречается с незнакомой лексикой.

Наконец, организационные трудности. О них больше всего говорилось на февральской конференции 1983 г. В связи с этим уместно вспомнить слова вице-президента АН СССР академика П. Н. Федосеева: «Учитывая тенденции развития современного научного знания, нужно, по-видимому, шире использовать практику организации научных коллективов и подразделений, создаваемых из представителей различных отраслей естественных, технических и общественных наук для решения той или иной четко определенной комплексной научно-технической проблемы. Такие функциональные гибкие подразделения были бы свободны

---

[8] По примеру того, как это сделано в Институте языка и литературы АН ЭССР.

от недостатков жесткой, иногда десятилетиями неизменной структуры научных учреждений отраслевого типа» [24].

Нам представляется, что одной из современных форм организации таких работ является временный коллектив по решению научно-технической задачи, который мог бы быть создан из представителей заинтересованных ведомств и организаций при Институте русского языка АН СССР как базовой организации. Главная задача такого коллектива по созданию Машинного фонда русского языка состояла бы в разработке принципиальных положений, касающихся обмена данными и программами среди участников решения этой проблемы, комплектования аппаратных и программных средств фондов-составляющих, установления технологических режимов эксплуатации фондов, изготовления проектной, программной и общесистемной документации, выполнения заказов и поручений коллективами-соисполнителями. Соблюдение этих требований, соблюдение единой технологии и нормативов, определенный уровень исполнительской дисциплины — залог сохранения единства фонда в условиях физической разобщенности его компонентов.

ЛИТЕРАТУРА

1. *Ершов А. П.* К методологии построения диалоговых систем: феномен деловой прозы. Новосибирск, 1979.
2. *Ершов А. П.* К методологии построения диалоговых систем: феномен деловой прозы.— В кн.: Вопросы кибернетики. Общение с ЭВМ на естественном языке. М., 1982.
3. ЭВМ пятого поколения Концепции, проблемы, перспективы. Под ред. Т. Мотооока. М., 1984.
4. *Глушков В. М.* Основы безбумажной информатики. М., 1982.
5. *Большаков И. А.* Составляющие и принципы формирования программного обеспечения для машинного фонда русского языка (в печати).
6. *Рогожникова Р. П.* Машинный фонд русского языка и словарное дело (в печати).
7. *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. М , 1981.
8. Словарь древнерусского языка XI—XIV вв. Введение, инструкция, список источников, пробные статьи. М., 1966, с. 27.
9. Parsing natural languages. Ed. by King M. London, 1983.
10. *Сороколетов Ф. П.* Традиции русской советской лексикографии.— ВЯ, 1978, № 3.
11. *Гак В. Г.* О типах ортологических словарей.— В кн.: Переводная и учебная лексикография. М., 1979.
12. *Дейт К.* Введение в системы баз данных. М., 1980.
13. *Караулов Ю. Н.* Методология лингвистических исследований и машинный фонд русского языка (в печати).
14. *Попов Э. В.* Общение с ЭВМ на естественном языке. М., 1982.
15. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Крысин Л. П., Лазурский А. В., Перцов Н. В., Санников В. З.* Лингвистическое обеспечение в системе автоматического перевода ЭТАП-I.— В кн.: Разработка формальной модели естественного языка. Новосибирск, 1981.
16. *Андрющенко В. М.* Автоматизированные лексикографические системы.— В кн.: Теоретические и прикладные аспекты вычислительной лингвистики. М., 1981.
17. *Guckler G., Müller E., Wahrig G.* Wörterbuch als Datenbank. Ein komplexes linguistisches, logisches und technisches Problem.— Sprache und Datenverarbeitung, 1977, № 2.
18. *Hitzenberger L., Konrad W., Krause J., Schneider C.* Das Regensburger Textverarbeitungssystem COBAPH.— Sprache und Datenverarbeitung, 1977, № 1.
19. *Krollmann F.* Linguistic data banks and the technical translator.— Meta, 1971, v. 16, № 1—2.
20. *Schulz J.* A terminology data bank for translators (TEAM).— Meta, 1980, v. 25, № 2.
21. *Андрющенко В. М.* Автоматизированная лексикографическая система UNILEX (Основные проектные решения).— В кн.: Вычислительная лингвистика. М., 1982.
22. *Петушков В. П.* О возможных пределах механизации лексикографических работ.— ВЯ, 1981, № 5.
23. Preliminary feasibility study of a U. J. T. A. terminology data bank.— Union of international technical Associations. UNESCO, December, 1982.
24. *Федосеев П. Н.* В. И. Ленин и проблемы интеграции естественных, общественных и технических наук.— Вестник АН СССР, 1978, № 9, с. 27.