

АНДРЮЩЕНКО В. М.

ВЫЧИСЛИТЕЛЬНАЯ ЛЕКСИКОГРАФИЯ, ЕЕ ВОЗМОЖНОСТИ И ПЕРСПЕКТИВЫ

1. В последнее десятилетие в применении вычислительной техники для автоматизации лексикографических работ наметились новые тенденции, позволяющие говорить о новом этапе развития вычислительной лексикографии. Эта научная дисциплина, развившаяся в рамках вычислительной лингвистики, первоначально имела своим объектом исследования и конструирования так называемые «машинные словари», т. е. словари, использующиеся в составе лингвистического обеспечения систем машинного перевода и автоматизированных информационных систем [1].

Вычислительная лексикография недавнего прошлого была прежде всего «отраслевой» лексикографией: она занималась компьютерными формами терминологических (большой частью информационно-поисковых) словарей, предназначенных для информационного обслуживания тех или иных отраслей знаний и сравнительно ограниченных в лексическом отношении систем автоматического перевода. Переход к массовой автоматической и автоматизированной (computer assisted) документалистике, переводческой и редакционно-издательской деятельности изменил ситуацию: требования к машинным словарям возросли как в количественном, так и в качественном отношении.

С первых шагов вычислительной лексикографии намечалась тенденция противопоставлять машинные словари обычным лексикографическим источникам по ряду признаков — по методу их использования, по методам отбора словников, по составу информации, комплекующей словарную статью [1, с. 29—47]. Однако сейчас многие особенности машинных словарей [1, с. 21—28; 2, с. 40] воспринимаются как случайные, характерные лишь для определенного этапа развития технологии обработки данных на естественных языках или для определенных направлений вычислительной лингвистики. Созданные в 70-х годах терминологические банки данных (ТБД) и автоматические словари (АС), ориентированные на человека, развитие средств автоматизации лексикографических работ существенно обогатили арсенал вычислительной лексикографии и сблизили ее проблематику с проблематикой общей лексикографии. В 80-е годы перед вычислительной лексикографией встали задачи овладения новой информационной технологией, ориентирующейся на автоматическую обработку данных, представленных на естественном языке, на интерактивное человеко-машинное решение задач, на «автоформализацию профессиональных знаний» [3] (в данном случае — знаний лексикографов), на новые технические средства — персональные компьютеры и ЭВМ пятого поколения, на использование вычислительных и информационных сетей. Это ставит и перед общей лексикографией ряд новых проблем.

На наших глазах создается новая «лексикографическая ситуация», в которой уже нельзя проводить резкие границы между академической, отраслевой, вычислительной и издательской лексикографией. Необходима интеграция этих направлений. Отраслевая лексикография сейчас нуждается в академических словарях, представленных или автоматически переводимых в форму и модели, используемые автоматизированными системами обработки данных на естественных языках. Издательская лексикография нуждается в средствах, развитых в вычислительной лексикографии, так как перед ней стоят задачи автоматизации редакционно-издатель-

ских процессов и поиска новых форм словарных изданий, соответствующих современным средствам передачи информации. Академическая лексикография не может не учитывать как требования издательской лексикографии, выдвигаемые задачами ускорения научно-технического прогресса в полиграфии и издательском деле, так и прикладные задачи, связанные с созданием систем обработки данных на естественном языке.

Начиная с 70-х годов все более определяются новые функции лексикографии, которые обещают стать ее ведущими функциями на ближайшие десятилетия, — лексикографическое обеспечение человеко-машинного общения и обработки данных на естественном языке, создание национальных лексикографических служб [4]. С ведущими функциями лексикографии, как было показано В. Г. Гаком [5], связаны и определенные типы словарей, каждый из которых символизирует свой период в развитии лексикографии. Предложенная В. Г. Гаком периодизация демонстрирует одну показательную закономерность — рост объемов словарей в каждом периоде на порядок. Так, в дословарном периоде объемы вокабуляриев и ономастиконов исчислялись в среднем тысячами единиц, в раннесловарном периоде лексиконы и двуязычные словари насчитывали десятки, а толковые словари периода развитой лексикографии — сотни тысяч словарных статей. Национальные лексикографические службы, оснащенные АС, ТБД и автоматизированными лексикографическими системами, надо полагать, выведут нас на объемы словарей — генеральных тезаурусов языка — порядка миллиона словарных статей, а технология «безбумажной информатики» потребует от лексикографии овладения десяти- и даже стомиллионными объемами словарных баз данных, или словарных карточек, если говорить современным лексикографическим языком. Возможности компьютерной лексикографии должны привести к тому, что в будущем различие между словарной карточкой и готовым словарем должно уменьшаться и в конечном счете исчезнуть: бесчисленное количество различных типов словарей должно программно порождаться из лексикографически отработанной автоматизированной словарной картотеки.

Экстраполируя периодизацию В. Г. Гака на конец XX — начало XXI в., мы можем говорить о периоде компьютерной¹ лексикографии, который сменится периодом лексикографии в вычислительных и информационных сетях.

Реализация словарей, содержащих свыше 300—500 тыс. статей, в книжной форме уже нерациональна. Современные и перспективные средства обработки данных и связи подсказывают иной путь: первичной формой словарей должна стать их компьютерная форма, а книжная форма и сложившиеся в ней типы словарей должны стать производными формами, проекциями лексикографических баз данных по составляющим, структуре и номенклатуре словарных статей.

В связи с этими перспективами актуальным представляется новое осмысление роли и места вычислительной лексикографии как инструмента «лингвистического конструирования» [2].

2. Вычислительная лексикография является одним из направлений вычислительной лингвистики. В 1986 г. вычислительная лингвистика отмечает свое двадцатилетие. Правда, этот термин существовал и до 1966 г., однако именно в этом году вычислительная лингвистика впервые была названа научной дисциплиной со своим объектом и кругом задач [6]. Сейчас к ее компетенции относят автоматический лингвистический анализ, синтез и преобразование текста и речи, проблемы представления и использования знаний в системах искусственного интеллекта, обработку статистических данных о языке и текстах, автоматизацию лингвистических исследований, в том числе автоматизированную подготовку словарей, разработку лингвистического обеспечения информационных систем, систем машинного перевода и искусственного интеллекта, редакционно-изда-

¹ Компьютерной мы называем лексикографию как совокупность словарей, создаваемых и хранимых в этой форме; вычислительной — лексикографию как научное направление, занимающееся изучением и созданием средств для разработки и использования компьютерных словарей.

тельских систем и многие другие применения ЭВМ в целях обработки данных на естественном языке, будь то в интересах филологии или других дисциплин и приложений. К ведению вычислительной лингвистики относят также преподавание вычислительной техники и программирования для филологов. Однако принадлежность этой дисциплины и ее «ядерное» содержание до сих пор остаются спорными (см. историю и обсуждение вопроса в [7]).

Наше определение вычислительной лингвистики основано на следующих аналогиях. Если рассматривать языкознание как дисциплину, предметом изучения которой являются лингвистические структуры, образуемые единицами и конструкциями языка, аналогично тому, как в качестве предмета классической математики рассматриваются математические структуры (произвольные множества с заданными на них отношениями), а предметом информатики (computer science) соответственно — вычислительные структуры (структуры данных, алгоритмов, аппаратуры и реализованных в ней схем), мы выделяем класс соотносительных дисциплин, предметом изучения которых являются соотношения между структурами, изучаемыми смежными дисциплинами. Такой дисциплиной является, например, математическая лингвистика, занимающаяся изучением соотношений математических и лингвистических структур; такой же дисциплиной является и программирование, занимающееся соотношениями математических и вычислительных структур [8]². По аналогии мы и определяем вычислительную лингвистику как дисциплину, занимающуюся соотношениями лингвистических и вычислительных структур. Между этими соотносительными дисциплинами существует тесное родство. Теория формальных грамматик математической лингвистики нашла свое приложение в программировании, получила здесь дальнейшее развитие и создала основной аппарат теоретического программирования. Этот аппарат воспринял вычислительной лингвистикой, которая становится наукой о «лингвистическом» программировании, наукой об алгоритмическом и автоматизированном конструировании лингвистических объектов.

В этой схеме вычислительная лексикография занимает свое место, которое и определяет арсенал ее средств, — понятийный аппарат, методы, приемы и технологию решения задач. Именно этим арсеналом современная вычислительная лексикография существенно отличается от того направления в прикладной лингвистике, которое мы привыкли называть автоматизацией лексикографических работ, связывая его прежде всего с подготовкой словоуказателей, частотных словарей и конкордансов. Отличие это, в частности, состоит в том, что вычислительная лексикография заимствует из современного программирования понятия баз и банков данных, информационных систем, преобразуя эти понятия в своей области в соответствующие конструкции текстовых и словарных баз и банков, автоматизированных лексикографических систем. Вычислительная лексикография использует существующие и создает свои собственные базы данных, системы терминального доступа к ним, языки описания и манипулирования лингвистическими и лексикографическими данными, рассматривая понятие структуры словаря и схемы словарной статьи по аналогии с понятием схемы базы данных, разработанным в теоретическом программировании [9].

3. Вычислительная лексикография сейчас является динамично развивающимся направлением. Это объясняется не только успехами теоретического, системного и практического программирования прошедшего десятилетия, но и успехами, достигнутыми в области конструирования языковых процессоров, а также появлением новых перспективных технических средств, использование которых способно в ближайшие десятиле-

² Термину «структура» мы не придаем какого-либо специального смысла, считая структуры произвольными абстракциями изучаемых объектов и отношений между ними при условии, что дисциплина, предметом изучения которой являются структуры, рассматривает их в соответствии с природой и свойствами тех объектов и отношений, абстракциями которых они являются.

тия преобразовать облик современной научной и издательской лексикографии.

Поясним сказанное примерами. Всем хорошо известны такие простейшие формы компьютерных словарей, как словоуказатели, частотные словари и конкордансы. Известны также многочисленные претензии, которые способны предъявить к этим словарям лексикографы. Однако значительная часть таких претензий объясняется не природой этих словарей, а применявшейся до сих пор технологией обработки лингвистических данных (отсутствием лемматизации, неразличением омонимии, фиксированным контекстом, неудобной текстовой адресацией, засорением лишним материалом, ненадежностью статистики и т. п.). Применение современной технологии во многом снимает эти претензии. Благодаря использованию систем управления текстовыми базами данных и терминального доступа к ним эти словари приобретают форму автоматизированных лексикостатистических справочников. Автоматизированный лексикостатистический справочник — это одновременно и словоуказатель, и частотный словарь и конкорданс к текстам, которые постоянно хранятся в машине, образуя в ней текстовую базу данных. Доступ к текстам осуществляется через словоуказатель (индекс), в котором ведется и статистика. Задавая на терминале (дисплее) слово или группу слов, можно получить все контексты употребления заданных слов в нужном интервале текста. Задавать слова можно как в виде словоформ, так и в словарной форме. При просмотре словаря пользователь может устранить омонимию, ввести необходимые пометы к словам, а затем задавать уточненные запросы, вводя вместе со словами их признаки. Пользователь может запретить выдачу некоторых слов, например, служебных, может заказать или написать сам и включить в систему новую программу обработки, может полностью сменить программное обеспечение и обрабатывать данную текстовую базу новым программным обеспечением, но может сменить и текстовую базу, оставив прежнее программное обеспечение. Пользователь может «практически бесконечно» дополнять и расширять как текстовую базу, так и программное обеспечение, он может по мере необходимости делать выборочные распечатки материала или заказать полную печать конкорданса с любым упорядочением (в том числе и в виде обратного словаря) заглавных слов, получая сверстанную распечатку заданного формата.

Пользователь может, применяя системные средства редактирования и коррекции данных, вносить любые исправления во входной, промежуточный или выходной текст, он может и сам сгенерировать систему из готовых программных и текстовых модулей.

Это — результат реализации в вычислительной лексикографии принципа отделения системы хранения и доступа к данным от функциональной системы обработки, а в последней — отделения алгоритмов от лингвистических описаний (в данном случае — от схем источников). Всех этих возможностей были лишены старые системы обработки лексикографических данных, технология которых побуждала автоматизированным составлением словарей заниматься не лексикографов, а самих вычислителей. Современная вычислительная лексикография создает не словари, а программные средства для составления словарей. Словари же, используя эти средства, могут делать те, кто и должны их делать, т. е. профессиональные лексикографы. В этом состоит первая особенность современной вычислительной лексикографии.

Вторая особенность состоит в том, что в вычислительной лексикографии используются достижения в области конструирования языковых процессоров, т. е. программ автоматического анализа и синтеза текстов. До сих пор разработкой лингвистических процессоров занимались создатели систем машинного перевода и общения с базами данных на естественном языке. К сожалению, непосредственно перенести в АЛС разработанные ранее языковые процессоры невозможно по многим причинам, из них главными являются две: 1) в лексикографии нужны процессоры, дающие лингвистически точный результат (что не всегда имеет место в инженерных разработках), 2) в лексикографии нужны процессоры, оперирующие словарями

объемом в сотни тысяч словарных статей (в инженерных разработках объемы словарей языковых процессоров достигают лишь двух-трех десятков тысяч словарных статей). Поэтому в вычислительной лексикографии ставится и решается задача создания языковых процессоров, использующих данные обычных словарей. Для морфологии русского языка такая работа проделана: используя «Грамматический словарь русского языка» А. А. Зализняка, коллектив под руководством М. Г. Мальковского создал программы морфологического анализа и синтеза для практически полного русского языка [10]. Аналогичную работу этот коллектив ведет сейчас над «Словарем сочетаемости слов русского языка» (к сожалению, объем этого словаря слишком мал, чтобы надеяться на достаточно эффективное его использование в лексикографии).

Третья особенность современной вычислительной лексикографии состоит в том, что она выдвигает задачу преобразования обычных словарей в форму АС, а также обратную задачу — преобразования АС в книжную форму. Автоматическим словарем мы называем обычный словарь, хранимый в ЭВМ и сохраняющий в ней (для пользователя) внешнюю форму своих статей. АС можно пользоваться как обычным словарем. Его можно также непрерывно и оперативно корректировать, пополнять и обновлять, готовя очередное «исправленное и дополненное издание» (т. е. АС — это картотека изданного словаря). Можно, например, осуществлять групповой поиск словарных статей по признакам, приписанным заглавным словам или любым фрагментам статьи; образовывать различные «проекции» словаря, т. е. рассматривать только словник, или только толкования, или только иллюстрации, или только фразеологические единицы, или только группы производных слов, помещенные в словарные статьи, и т. д.; можно образовывать подсловари различных объемов, отбирая в них лексику по заданным признакам; можно соединять и сопоставлять данные различных словарей, образуя пересечения, объединения и разности как по макро-, так и по микроструктуре словарей-источников.

Соединение в одной системе текстовой и словарной баз данных мы называем автоматизированной лексикографической системой (АЛС). В АЛС реализуемы многие процедуры, связанные с подготовкой новых словарей, такие, как проверка полноты словарей на текстах (с выработкой оценок реальной употребительности слов и выражений), автоматический отбор слов в словники, формирование единых словников для групп двуязычных словарей заданного объема, контроль лексики толкований и иллюстраций, выявление логических кругов в дефинициях, отбор иллюстраций, контроль соответствия стилистических помет заданной классификации текстов и их фрагментов, учет вариативности форм и употреблений, связывание вариативности с определенными сферами употребления и многое другое. В АЛС могут одновременно храниться несколько ранее изданных словарей. Используя эти словари, можно согласовывать дефиниции в них (в особенности это касается терминологической лексики), автоматически (а значит без ошибок и пропусков) формировать справочные отделы, проводить любые работы, связанные с выверкой данных формируемого словаря по источникам.

Программа создания Машинного фонда русского языка ставит задачу разработки словарных баз данных для всех наиболее авторитетных толковых и наиболее распространенных специальных словарей русского языка. После создания автоматического варианта «Грамматического словаря русского языка» А. А. Зализняка Институтом русского языка АН СССР совместно с Лабораторией по применению вычислительных средств в гуманитарных науках НИВЦ МГУ ведутся работы над автоматическими вариантами «Словообразовательного словаря русского языка» А. Н. Тихонова, «Орфографического словаря», «Толкового словаря живого великорусского языка» В. И. Даля, «Словаря русского языка» в 4-х томах, а также над рядом других источников.

Конструкция АС предполагает возможность использования формальных описаний словарных статей [11]. При наличии таких описаний программно реализуемой становится методика написания словарных статей под управлением ЭВМ: выбор однородной группы заглавных слов, гене-

рация типовой структуры словарной статьи для данной группы, выдача очередного задания лексикографу на формирование фрагмента словарной статьи или на проверку определенных условий, выбор из источников или написание фрагмента за терминалом и т. д. Такая контролируемая ЭВМ подготовка словарных статей за терминалом, в интерактивном режиме или off-line, с последующим вводом данных, создает определенную дисциплину, контролируемый порядок и единообразие выполнения основных лексикографических операций. Уже на этапе написания словарных статей могут быть автоматически проверены или сформированы стандартные пометы и лексикографические знаки, шрифтовые выделения, проверены или созданы ссылки, размещена фразеология, проверена комплектность всех сведений, которыми должна быть снабжена словарная статья в соответствии с формализованным проектом словаря. [Таким образом, методика словарной работы, развиваемая вычислительной лексикографией, объективно направлена на устранение многих хорошо известных недостатков современной словарной работы, таких, как несистематичность картотек, алфавитный подход к порядку написания словарных статей, отставание в регистрации лексических инноваций, разноречивость помет в словарях (и даже в одном словаре!) и т. п.].

При наличии встроенных в систему языковых процессоров могут быть осуществлены такие операции, как автоматическое формирование диагностических окончаний, генерация речений, выявление сочетаемости и моделей управления, подбор образцов употребления для заданных конструкций и др. АЛС может быть соединена с информационно-справочной системой, содержащей необходимые публикации, сведения из которых могут быть автоматически привлечены в момент написания словарной статьи. Сопряжение АЛС с программами редакционно-издательского сервиса делает возможным непосредственный вывод словаря на носители, используемые полиграфическим оборудованием, причем контроль орфографии и другие операции «технического» редактирования могут быть осуществлены автоматически.

Современная вычислительная лексикография по-новому ставит вопрос о типологии словарей. Попытки построения типологии и классификации словарей, как известно, делались неоднократно (см. краткий обзор классификаций в [12]). Каждая такая попытка была связана с решением определенной проблемы, интересовавшей автора. Так, Л. В. Щерба [13] стремился создать общую теорию лексикографии, связывая ее с теорией языка; его установка — установка автора словаря, его классификация — очерк руководства к написанию словарей. Феноменологическая классификация русских словарей М. Л. Апажева [14] — руководство для пользователя по выбору нужного ему словаря. Классификация, предложенная А. М. Цывиным [15], сделана как бы с позиций внешнего наблюдателя. Ю. Н. Караулов [2], хотя и не дал в явном виде классификации словарей, но ввел и исследовал понятие лексикографического параметра, указав типы фрагментов словарной статьи, совокупность которых и идентифицирует однородную группу словарей. Это — установка на конструирование; вычислительная лексикография добавляет к ней лишь понятие структурной организации лексикографических параметров и создает свою типологию АС, характеризующую единством структур хранения и доступа к лексикографическим данным. В компьютерной лексикографии многие «типы» словарей становятся несущественными (невидимыми) в их автоматическом исполнении:

1. Все словари, характеризующиеся порядком расположения словарных статей (алфавитные прямые и обратные, гнездовые, тематические и дескрипторные, статистические алфавитно-частотные и частотно-алфавитные) — пользователь, да и создатель системы не знают, как реально расположены словарные статьи в базе данных, получать же их можно в любом заданном порядке следования словарных статей. Указанные выше признаки — это не параметры словаря, а параметры его внешней формы.

2. Все так называемые односторонние словари (орфографический, ударений, фонетический, орфоэпический, грамматический, обратный):

все эти словари могут быть получены (в принципе) в результате задания на вывод из базы данных словника двустороннего словаря (при условии его достаточной полноты, стандартности оформления и комплектности).

3. Часть словообразовательных и фразеологических словарей, словарей синонимов, антонимов, омонимов, неологизмов и архаизмов может быть получена путем вывода соответствующих фрагментов «хорошего» толкового словаря («хорошим» мы называем такой словарь, который все эти факты в себе содержит).

4. Все справочники сокращений, собственных имен и любых номенклатур могут быть получены из энциклопедических словарей, хранимых в словарных базах данных.

С вычислительной точки зрения это не самостоятельные типы словарей, а различные режимы обращения к словарным базам данных. Какие же классификационные признаки остаются существенными и в вычислительной лексикографии? Прежде всего это признаки объекта лексикографирования (слово, морфема, словосочетание, выражение, понятие, классификационная рубрика и т. п.). Список возможных значений этого признака открыт: любые объекты, имеющие однородное по своей структуре описание, могут быть лексикографированы в одном типе словаря (например, морфемы как заголовки словарных статей в толковых и переводных словарях, словосочетания — в терминологических). Далее это типы атрибутов (лексикографических параметров, переменных лексикографического метаязыка), которыми характеризуются лексикографируемые объекты [2, 75], и состав значений этих атрибутов (термы лексикографического метаязыка, например, списки частей речи, списки помет разного рода, лексикографические знаки, типы фрагментов словарной статьи). Число типов атрибутов, видимо, бесконечно, хотя и не очень велико. Затем это типы отношений объектов и атрибутов к языковой системе или системам (в случае переводных и исторических словарей). Типов таких отношений, думается, также бесконечно много, если учесть все мыслимые способы членения языка на составляющие его сферы употребления в функциональном, жанровом, синхронном, диахронном, географическом, хронологическом и других планах вплоть до языка индивида. Следовательно, феноменологически число типов словарей бесконечно, а структурно мы имеем один тип универсального словаря (ср. [2, с. 36; 16]), т. е. словарную базу данных, в типовой схеме которой переменными заданы типы объектов, типы атрибутов и типы отношений объектов и атрибутов к языковой системе (системам).

Такой подход к классификации словарей не учитывает группу признаков, не поддающихся (показ) формализации. Эту группу признаков можно в целом охарактеризовать как «стиль» (или способ) описания объекта лексикографирования. Толковый, энциклопедический, терминологический (одноязычный справочник), разные виды исторических, учебный, толково-комбинаторный и все виды «словарей для МП» отличаются друг от друга не столько объектами лексикографирования и не составом их атрибутов (хотя в каждом имеются свои, специфические для него атрибуты), сколько способом подачи значений атрибутов (ср. дефиниции толковых, энциклопедических и учебных словарей, различия в толковании значений в больших, средних и малых толковых словарях, формально-грамматические пометы обычных и «машинных» словарей, имплицитное описание сочетаемости в иллюстрациях обычных словарей и эксплицитное — в толково-комбинаторном). Пока АС могут осуществлять преобразование «стиля» лишь в области морфологических помет. В остальных случаях они могут выдавать только то, что в них «заложено». Но по мере развития языковых процессоров, по-видимому, многие такие преобразования станут возможны. Думается, что именно эта группа признаков и является основной точкой развития как теории общей, так и зависимой от нее практики вычислительной лексикографии. Сюда нужно отнести также проблему разработки такой лексико-семантической классификации, позиция лексико-семантического варианта лексемы в которой являлась бы аргументом для выбора структуры или хотя бы некоторых термов метаязыка

ее толкования [17, 18]. Решение этой проблемы имело бы ключевое значение для создания нового вида поиска в словаре — поиска слов по приблизительно описанному значению, способствовало бы улучшению методик автоматического формирования семантических словарей. 1/8

Наряду с вопросом о типологии словарей вычислительная лексикография ставит также вопрос и о типологии пользователей и способов их обращений к словарю. Этот вопрос в теории лексикографии, кажется, совершенно не изучен.

Имея дело с «Орфографическим словарем» и не зная действительного написания слова, пользователь может довольно долго искать его правильное написание (например, в этом словаре нет слов *шестнадцатеричный* и *девятеричный*, широко употребляемых в наименованиях соответствующих систем счисления; узнать их правильное написание можно только, сравнив их со словами *четверичный*, *пятеричный*, *шестеричный*, *семеричный*, *восьмеричный*, *десятеричный* и *двенадцатеричный*). Имея дело с толковым словарем и не найдя в нем нужного слова, читатель далеко не всегда может быть уверен, имеется ли это слово в данном языке, или оно не является нормативным в каком-либо аспекте и поэтому не включено в словарь, или его там нет по случайным причинам. Так, в «Словаре русского языка» в 4-х томах нет слов *двоичный*, *пятеричный*, *шестеричный*, *семеричный*, *девятеричный*, *шестнадцатеричный*, а для слов *троичный*, *четверичный*, *восьмеричный* и *десятеричный* не указано значение «... в наименовании системы счисления с основанием 3 (4, 8 и 10 соответственно)». Слово *останов* имеется в «Орфографическом словаре», но отсутствует в «Словаре русского языка» в 4-х томах — только из этого факта можно заключить, что оно, видимо, имеет специальное, но все-таки достаточно широко употребительное значение (*останов машины*, *программы*, *тяжелый останов* и т. п.). Имея дело с ортологическим словарем и идя от неправильного (ненормативного) слова или выражения, пользователь не всегда может найти правильный ответ на свой запрос, даже если ответ в словаре имеется [например, в словаре «Трудности русского языка» (под ред. Л. И. Рахмановой, 2-е изд., изд-во МГУ, 1984) выражение *обречь на успех* нельзя отыскать по слову *успех*]. Пользуясь не только словарями, но и грамматиками русского языка, нельзя с уверенностью ответить на такие вопросы: соответствуют ли норме слова *записевый* и *массивовый* в наименованиях типов данных? Как определить выражения, имеющиеся в языках программирования — *алгольные* или *алгольские*, *фортранные* или *фортранские*, *паскальные* или *паскалевские* и т. д.?

Все способы поиска в словаре вполне программируемы, и программирование различных стратегий использования АС способно сделать его намного более эффективным справочником, чем словарь в книжной форме. Но для этого нам пока недостает систематических знаний о типах возможных запросов к словарю.

В отличие от словарей в полиграфическом исполнении АС практически не имеют предела ни в степени сложности словарной статьи [12, с. 39; 19] (пользователь может видеть в словарной статье только то, что его интересует), ни в количестве входов в словарь [2, с. 71]. Этим снимается основное, на наш взгляд, противоречие лексикографической теории (требование многоаспектности описания лексики) и практики подготовки словарей в книжной форме (линейность расположения словарных статей с неизбежными ограничениями на сложность их структуры и количество входов в словарь). Например, в n -язычном автоматическом словаре может быть осуществлен автоматический переход с каждого из n языков на ($n - 1$) других, т. е. для n языков автоматически создаются ($n - 1$) входов. Если различать в каждой паре языков родной и иностранный [13, с. 90; 12, с. 39] и соответственно варьировать выход (описание на родном или иностранном языке), то число входов (типов словарей) увеличивается еще в два раза, притом что в систему загружаются ($n - 1$) двуязычных словарей достаточно простой структуры [20].

Основной технологический прием лексикографирования — «координатное» фрагментирование словарной статьи и ссылочный аппарат. Дан-

ный прием в вычислительной лексикографии программируется в максимально обобщенном виде. Математическая структура словаря — это ориентированный граф с размеченными узлами (фрагментами словарной статьи) и дугами (ссылками). Соответствующие вычислительные структуры — списки и программно реализованные механизмы создания списков, поиска в них и их преобразования. (Лексикографические структуры обладают рядом особенностей по сравнению с соответствующими математическими и вычислительными структурами, одна из них — многоаспектная разметка узлов и дуг представляющего словарь графа. Поскольку такие структуры имеют лингвистическую природу, они обладают означающим, означаемым, типом означаемого и его категориальной соотнесенностью. Это создает свои особенности в размещении и в доступе к данным, что заставляет вычислительную лексикографию строить свои системы управления базами данных, а не использовать разработанные для других приложений.) Систематическое использование этого приема дает возможность решать самые разнообразные лексикографические задачи. Например, один и тот же алгоритм отыскивает в АС словосочетания всех типов — терминологические, фразеологические, речения; одна и та же техника компоновки словарных статей может быть применена для словообразовательных, дескрипторных и идеографических (тематических) словарей. Одна и та же структура хранения и доступа получается для частотных словарей, словоуказателей и конкордансов, с одной стороны, и для морфемных словарей, с другой (словник для морфемного словаря — то же, что тексты для словоуказателя). Структура данных и программное обеспечение многоязычного АС применимы для реализации ассоциативного словаря.

Таким образом, в вычислительной лексикографии однородные типологические группы образуют: 1) словоуказатели, частотные словари, конкордансы, морфемные словари; 2) орфографический и орфоэпический словари, словарь ударений, словарь «Слитно или раздельно?» [21]; 3) «Русский семантический словарь» [22], «Словообразовательный словарь русского языка» А. Н. Тихонова, все дескрипторные словари; 4) все переводные и ассоциативные словари; 5) все толковые, энциклопедические, этимологические, моновязычные терминологические словари и толково-комбинаторный словарь; 6) все ортологические словари, словари паронимов, синонимов и антонимов. Каждая из этих групп характеризуется единой обобщенной структурой словарной статьи и единой стратегией размещения и доступа к данным.

4. Для первых пяти типологических групп все основные вычислительные структуры и в части хранения и доступа, и в алгоритмической части к настоящему времени разработаны и опробованы на более чем достаточном материале, на нескольких типах машин.

Наиболее значительные разработки предприняты в терминологической лексикографии. Терминологические банки ТЕАМ, принадлежащий фирме Сименс АГ (ФРГ), LEXIS (Федеральное бюро языков, ФРГ), EURODICAUTOM (Комиссия европейских сообществ в Люксембурге), TERMIUM (Канада) содержат каждый более миллиона описаний терминов (обзор см. в [23, с. 65—193]). Еще в начале 70-х годов в СССР начала создаваться Автоматизированная система терминологического обслуживания (АСИТО, ВНИИКИ Госстандарта СССР) [23, с. 145—165], содержащая сейчас в своей базе более 70 тыс. описаний стандартизованных терминов. В 1980 г. в Лаборатории по применению вычислительных средств в гуманитарных науках НИВЦ МГУ по заказу Всесоюзного Центра Переводов был разработан многоязычный АС МУЛЬТИЛЕКС [24], который в версии 1985 г. содержал 80 тыс. словарных статей по вычислительной технике и ядерной физике на русском, английском, французском, немецком и (частично) венгерском языках. (Сообщалось о новой реализации этого словаря («вторая очередь») в ВЦП [20]). Распиренная версия этого словаря (ТЕРМИН) включена в состав АЛС УНИЛЕКС [25] в качестве ее словарной базы; текстовую базу этой системы образуют автоматический конкорданс и пакет программ для генерации частотных словарей. В си-

стеме предусмотрен специальный язык для задания схем источников — структур текстов и словарных статей. В настоящее время ведутся работы по наполнению системы источниками для Машинного фонда русского языка.

Кроме упомянутых, в СССР имеется еще несколько реализаций АС и АЛС. Наиболее ранняя из них — Многоцелевой Автоматический Русский Словарь (МАРС) [26], созданный в Лаборатории инженерной лингвистики ЛГПИ им. А. И. Герцена. Ведутся работы над созданием автоматического варианта «Словаря украинского языка» в 11-ти томах [27]. Наиболее развитой является автоматизированная система генерации словарей эстонского языка, созданная в ИЯЛ АН ЭССР [28].

Из зарубежных систем — наиболее крупная АЛС LEDA, разработанная в Институте немецкого языка в Маннгейме (ФРГ) [29, 30]. Ее текстовый банк содержит (по данным 1984 г.) 20 млн. словоупотреблений, из которых 13 млн. составляют корпус письменной речи, корпус устной речи и корпус газет; 7 млн. словоупотреблений получены на машинных носителях из различных издательств. Планируется довести текстовую базу до 50 млн. словоупотреблений — корпус текстов, подготавливаемый для нового словаря немецкого языка объемом 200 тыс. словарных статей. В системе имеется словарная база данных, сформированная из различных словарей немецкого языка; ее обобщенная словарная статья включает транскрипцию заглавного слова, его лемму, класс слова, диагностические окончания, сведения о словообразовании и синтаксическом окружении, фрейм глубинных падежей, формальное и содержательное описание значений. Аналогичная АЛС создается в связи с подготовкой нового словаря голландского языка [31].

Системы нового поколения снабжаются средствами вывода словарей на носители, используемые в полиграфии (в СССР эта технология освоена в Институте языка и литературы АН ЭССР), а также на фотоносители (микрофильмы и микрофиши). Словари выводятся не только в книжной, но и в компьютерной форме (на магнитных лентах), для их дальнейшего использования в системах обработки данных. Академическая и издательская лексикография получают таким образом возможности новых приложений в сфере информатики и автоматической документалистики, приобретаая взамен практически неисчерпаемые источники данных по современному языку и необходимые средства для автоматизации исследований.

5. Уже сейчас на машинных носителях имеется значительное количество лексикографически релевантной информации. Ближайшее будущее готовит нам еще большие и принципиально новые возможности. Эти возможности связаны, во-первых, с использованием новых информационных служб в вычислительных сетях [32, 33]. Новые информационные службы — это цифровая передача средствами связи текстов, речи, рукописных и графических изображений в вещательном и интерактивном режимах. Через службу типа ВИДЕОТЕКС пользователь может получить доступ по телефону к удаленному текстовому банку и принять информацию от него на вычислительную машину или даже на бытовой телевизор. Таким банком может быть АС какого угодно объема, и, таким образом, в перспективе это — один из видов массового не книжного издания словарей. Служба ВИДЕОТЕКС может обеспечить также доступ к словарным картотекам, текстовым базам, хранящим сформированные лингвистами корпуса текстов, к фондам памятников письменности или к фондам диалектологических данных. В системах заочного обучения через службы ВИДЕОТЕКС можно получать контрольные работы и всю необходимую учебную и справочную литературу. Используя эту службу, можно отправить свои данные в издательство, в фонд на хранение, сдать в учебное заведение выполненные контрольные работы и т. д. Информационная служба ТЕЛЕТЕКСТ передает по телевидению текстовые и графические изображения только в вещательном режиме. Так может передаваться текущая информация (новый вид газеты), реклама, художественные и развлекательные тексты. Прием сообщений службы ТЕЛЕТЕКСТ на вычислительную машину обеспечивает непрерывное отслеживание языковых инноваций, создает воз-

возможности их оперативного учета в текстовых базах, а затем и в словарях.

Во-вторых, принципиально новые возможности как для подготовки, так и для использования словарей предоставляют персональные компьютеры — «первый массовый инструмент активной формализации профессиональных знаний» [3, с. 107]. Именно это свойство персональных компьютеров подводит заключительную черту под все еще распространенным тезисом о том, что роль лингвиста — содержательная постановка задачи, а ее решение — дело программиста. В применении к лексикографии персональный компьютер — это и оперативная картотека словарника, т. е. та часть словарной картотеки, над которой он в данный момент работает, это устройство связи с центральной картотекой, это и прибор для наблюдения (если компьютер подключен к сети ТЕЛЕТЕКСТ), это инструмент для письма и редактирования, обладающий гораздо большими возможностями, чем пишущая машинка; это и автоматический проект словаря, обеспечивающий автоматизированное написание словарных статей, их компоновку из готовых фрагментов, проверку ссылок, цитат, правильный выбор и расстановку формальных элементов словарной статьи. Для пользователя словарем — это АС объемом 30—60 тыс. словарных статей. Для издательской лексикографии — это новый вид издания словаря и учебника: издание на гибком магнитном диске.

Еще один, уже практикуемый [34], вид издания словарей, пока учебных и туристских — это «словарные компьютеры» (точнее управляющие файлы для изготовления электронных схем для карманных микрокомпьютеров — *handheld computers*), специализированные на выдачу словарной информации, запоминание новых слов и на функции лексического тренажера для учащегося.

Освоение новых форм словарей, новых форм словарной работы, новых приложений, открываемых современной информационной технологией, — насущная задача вычислительной лексикографии, один из путей продвижения научно-технического прогресса в языкознание в целом.

ЛИТЕРАТУРА

1. Марчук Ю. П. Вычислительная лексикография. М., 1976.
2. Карацлов Ю. П. Лингвистическое конструирование и тезаурус литературного языка. М., 1981.
3. Грозов Г. Р. Национальные информационные ресурсы: проблемы промышленной эксплуатации. М., 1984, с. 96.
4. Imbs P. Les travaux pour le Trésor de la langue française. — In: Tavola rotonda sui grandi lessici storici. Firenze, 3—5 maggio, 1971. Firenze, 1973.
5. Гак В. Г. О некоторых закономерностях развития лексикографии. (Учебная и общая лексикография в историческом аспекте). — В кн.: Актуальные проблемы учебной лексикографии. М., 1977.
6. Languages and machines — Computer in translation and linguistics. A report by the Automatic language processing advisory committee. Washington, 1966.
7. Batori I. S. Linguistische Datenverarbeitung. — Sprache und Datenverarbeitung, 1977/1.
8. Холл П. Вычислительные структуры. Введение в нечисленное программирование. М., 1978.
9. Мартин Дж. Организация баз данных в вычислительных системах. М., 1980, с. 68—89.
10. Мальковский М. Г. Диалог с системой искусственного интеллекта. М., 1985.
11. Kolodyazhnaya L. I. Linguistic software for an automatic dictionary of common language. — In: Symposium on automatic compilation of dictionaries (Tallinn, November 25—27, 1985). Summaries. Tallinn, 1985.
12. Денисов П. Н. Типология учебных словарей. — В кн.: Проблемы учебной лексикографии. М., 1977.
13. Щерба Л. В. Опыт общей теории лексикографии. — В кн.: Академик Л. В. Щерба. Избранные работы по языкознанию и фонетике. Л., 1958.
14. Апажеев М. Л. Лексикография и классификация словарей русского языка. Нальчик, 1971.
15. Цывик А. М. К вопросу о классификации русских словарей. — ВЯ, 1978, № 1.
16. Денисов П. Н. Практика, история и теория лексикографии в их единстве и взаимобусловленности. — В кн.: Проблемы учебной лексикографии и обучения лексике. М., 1978, с. 27—28.
17. Шведова Н. Ю. Однотомный толковый словарь (специфика жанра и некоторые аспекты дальнейшей работы). — В кн.: Русский язык. Проблемы художественной речи. Лексикология и лексикография. М., 1981, с. 171.

18. *Шведова Н. Ю.* Лексическая классификация русского глагола (на фоне чешской семантико-компонентной классификации).— В кн.: *Славянское языкознание. IX Международный съезд славистов. Киев, сентябрь 1983 г.: Доклады советской делегации.* М., 1983, с. 310.
19. *Сороколетов Ф. П.* Традиции русской и советской лексикографии.— ВЯ, 1978, № 3, с. 28.
20. *Убин И. И.* Место автоматических словарей в системе переводческого обслуживания и их основные свойства.— В кн.: *Международный семинар по машинному переводу.* Москва, 1983 г.: Тезисы докладов. М., 1983.
21. *Бужина Б. З., Калакуцкая Л. П.* Слитно или раздельно? (Опыт словаря-справочника). 3-е изд. М., 1982.
22. *Караулов Ю. Н., Молчанов В. И., Афанасьев В. А., Мизалев Н. В.* Русский семантический словарь. Опыт автоматического построения тезауруса: от понятия к слову. М., 1982.
23. *Волкова И. Н.* Стандартизация научно-технической терминологии. М., 1984.
24. *Убин И. И.* Общие принципы построения и использования автоматического словаря MULTILEX.— В кн.: *Вычислительная лингвистика. М., 1982.*
25. *Андрющенко В. М.* Автоматизированная лексикографическая система UNILEX. (Основные проектные решения).— В кн.: *Вычислительная лингвистика. М., 1982.*
26. *Лукьянова Е. М.* Информационная база автоматических словарей.— В кн.: *Статистика речи и автоматический анализ текста. Л., 1980.*
27. *Пецак М. М.* Машинный формат представления лексикографических текстов у «Фонді української мови». Київ, 1985.
28. *Viks Ü.* Applications of a databank of dictionaries.— In: *Academy of Sciences of the Estonian S. S. R., Preprint KKI-32.* Tallinn, 1985.
29. *Brustkern J., Hess K. D.* The BONNLEX lexicon system.— In: *Lexicography in the electronic age. Proceedings of a symposium held in Luxembourg, 7—9 July 1981.* Ed. by Goetschacks J. and Rolling R. Amsterdam — New York — Oxford, 1982.
30. *Teubert W.* Application of a Lexicographical data base of German.— In: *10th International conference on computational linguistics. 22th Annual meeting of the association for computational linguistics. Proceedings of Colig 84. 2—6 July 1984.* California, 1984.
31. *Sterkenburg P. van, Martin W., Al B.* A new van Dale project: Bilingual dictionaries on one and the same monolingual basis.— In: *Lexicography in the electronic Age...*
32. *Яновский Г. Г.* Новые информационные службы и вычислительные сети общего пользования.— В кн.: *Вопросы кибернетики. Проблемы теории вычислительных сетей.* М., 1983, с. 126—130.
33. *Ефимов А. Н.* Информационный взрыв: проблемы реальные и мнимые. М., 1985.
34. *Voigt W.* Der Wörterbuchverlag und das «optimale» Wörterbuch.— In: *Theoretische und praktische Probleme der Lexikographie. 1. Augsburger Kolloquium. München, 1984.*