4
М-38

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: идеи и суждения

Любой прогресс в области построения моделей и алгоритмов останется 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.

ИЗДАТЕЛЬСТВО НАУКА

ИРЯ
институт русского языка
им. в. в. виноградова

www.ruslang.ru

80703

АКАДЕМИЯ НАУК СССР
ИНСТИТУТ РУССКОГО ЯЗЫКА
НАУЧНЫЙ СОВЕТ ПО ЛЕКСИКОЛОГИИ И ЛЕКСИКОГРАФИИ

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ИДЕИ И СУЖДЕНИЯ

Ответственный редактор
член-корреспондент АН СССР Ю. Н. КАРАУЛОВ



МОСКВА «НАУКА» 1986

4
П-38
БИБЛИОТЕКА
ИН-ТА РУ НОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР
80783

Сборник посвящен проблемам создания Машинного фонда русского языка — системы комплексной автоматизации теоретических исследований и прикладных разработок в русистике и лексикографии, направленной на расширение используемых в информатике данных о русском языке во всех его хронологических, пространственных и функциональных разновидностях, на развитие средств автоматической обработки данных и общения с ЭВМ на русском языке.

Редакционная коллегия:

академик А. П. ЕРШОВ,

член-корреспондент АН СССР Ю. Н. КАРАУЛОВ,

кандидат филологических наук В. М. АНДРЮЩЕНКО

Рецензенты:

В. П. ГРИГОРЬЕВ, Ю. Н. МАРЧУК

М $\frac{4602000000-336}{042(02)-86}$ 286—86—IV

© «Наука», 1986 г.

206—2

ПРЕДИСЛОВИЕ



Значимость русского языка как мирового языка и как языка межнационального общения народов СССР и стран социалистического содружества, качественные изменения в требованиях, которые предъявляет к лингвистике современное производство, основанное на широком использовании автоматизированных систем типа «человек-машина», фундаментальность тех областей духовной культуры, материальным выражением которых является русский язык,— все это определяет сегодня в качестве главного направления русистики разработку теории русского языка в его исторических, географических и функциональных разновидностях на основе прежде всего проблемной организации науки о русском языке и ее интеграции с другими гуманитарными и естественными науками.

Обеспечить эффективное решение этой задачи можно только в том случае, если резко, количественно и качественно, расширить фактическую базу многоаспектных системных исследований по русскому языку, обеспечив доступ к получаемым результатам в кратчайшее время потребителей самого широкого профиля — от работников просвещения, культуры и средств массовой информации до специалистов по автоматизированным системам обработки текстов на русском языке. Для фундаментальных исследований в области гуманитарных дисциплин такая база была бы гарантией достоверности и надежности получаемых выводов при решении канонических задач и мощным стимулом возникновения новых проблем, постановка которых на существующей фактографической базе в обозримом будущем является едва ли возможной. Такая база сведений о функционировании, строе и истории русского языка была бы наконец мощным инструментом научного управления теоретическими и экспериментальными исследованиями в сфере наук о человеке и человеческом обществе, его истории, его культуре и языке. Такая база данных о русском языке возможна только на основе использования вычислительной техники.

Использование ЭВМ как в сфере лингвистических, вообще филологических исследований, так и в практике использования результатов этих исследований отвечает тем требованиям широкой автоматизации научных исследований и современных производственных процессов, которые поставил перед нами XXVII съезд КПСС в связи с задачами всемерного ускорения научно-технического прогресса.

Осознание необходимости использования ЭВМ в лингвистических исследованиях и в ходе реализации получаемых результатов как лингвистами, так и теоретиками и практиками автоматизированной переработки информации на естественных языках привело к идее создания Машинного фонда русского языка. Термин «Машинный фонд русского языка» сегодня стал уже достаточно привычным, что, однако, совсем не означает существования у него единственного и общепринятого понимания. Обсуждению вопросов о том, каким должен быть Машинный фонд русского языка по составу и структуре, решение каких задач он должен обеспечить в первую очередь, каков минимальный и максимальный круг его лингвистических источников, какими могли бы быть организационные и технические формы его существования, и посвящен этот сборник, составленный на основе материалов состоявшейся в феврале 1983 г. Всесоюзной конференции по проблемам создания машинного фонда данных для автоматизированной системы лексикографических исследований.

Участники конференции поддержали предложение Научного совета по лексикологии и лексикографии и Секции лингвистических проблем обработки информации Научного совета по комплексной проблеме «Кибернетика» АН СССР разработать проект Машинного фонда русского языка как системы комплексной автоматизации лингвистических исследований и разработок, ставящей следующие задачи:

1. Перевести академические словари и грамматики в машинную форму, т. е. создать академические словарно-грамматические базы данных. Решение этой задачи позволит:

а) передавать академические словари и грамматики в формализованном виде для использования в прикладных разработках;

б) регулярно и оперативно вести пополнение этих источников и их уточнение;

в) ускорить публикации исправленных и дополненных переизданий словарей и грамматик.

2. Создать автоматически пополняемые словоуказатели и словари на базе текстов деловых и разговорных стилей, а также текстов научно-технической литературы и документации. Решение этой задачи позволит:

а) ввести в научный обиход до сих пор мало использованный языковой материал;

б) уточнить данные академических словарей и грамматик в той части, которая в наибольшей степени связана с использованием русского языка в автоматизации процессов переработки информации во всех сферах народного хозяйства и общественной жизни;

в) ввести в научный обиход на широком материале принципиально новые типы лингвистических данных о статистико-комбинаторных свойствах лингвистических явлений.

3. Соединить в единый языковый фонд данные об общеупотребительном русском языке и данные терминологических фондов. Решение этой задачи позволит:

а) улучшить состав, структуру и лексикографическую информацию терминологических словарей;

б) расширить представление о современном русском литературном языке;

в) уточнить концепцию подъязыка как проекцию общеупотребительного русского литературного языка на определенную предметную область или класс ситуаций общения.

4. Создать фонд лингвистических алгоритмов и программ, включая процессоры русского языка. Решение этой задачи позволит:

а) обеспечить типовыми программными средствами автоматизацию лингвистических исследований в академических и отраслевых институтах и вузах;

б) быстро настраивать общеязыковые процессоры на конкретные применения;

в) применять общелингвистическое программное обеспечение для решения специальных задач, в частности задач автоматического анализа и синтеза звучащей речи.

Подготовка и осуществление этого проекта является также одним из мероприятий по реализации решений июньского (1983 г.) Пленума ЦК КПСС и решения ЦК КПСС по вопросам преподавания русского языка в национальных республиках. Создание Машинного фонда русского языка будет, несомненно, иметь большое культурное и научно-техническое значение. Его

данные, алгоритмы и программы будут широко использоваться в автоматизированных системах обучения русскому языку и другим дисциплинам, в системах автоматизированного редактирования и набора, в подготовке лингвистического обеспечения ЭВМ новых поколений, информационных систем и интеллектуальных роботов. В конечном счете Машинный фонд русского языка будет работать на решение нашей ключевой экономической задачи — кардинальное повышение производительности труда. В этой огромной работе существенную роль сыграет и подготовка лингвистического обеспечения школьной и вузовской информатики, обеспечения эффективного взаимодействия людей и вычислительной техники в процессе труда. Созданию такого обеспечения на высоком научном уровне с учетом всех достижений языковедения и всех накопленных им знаний будет способствовать Машинный фонд русского языка.

Редколлегия

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ПОСТАНОВКА ЗАДАЧИ

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ВНЕШНЯЯ ПОСТАНОВКА

Л. П. ЕРШОВ

Осенью 1978 г. один специалист по программированию, имевший опыт создания диалоговых систем общения с ЭВМ, получил приглашение выступить с докладом перед сообществом физиков, биологов и программистов, собравшихся на первую (но ставшую впоследствии традиционной) конференцию Диалог-78 в Биологическом центре АН СССР в Пущине-на-Оке. Доклад содержал изложение общей гипотезы о продуктивности диалога человека с ЭВМ на естественном языке при условии ограничения этого диалога так называемой «деловой прозой». Надежда на продуктивность гипотезы базировалась на другой гипотезе о внутренней формализованности любого устойчивого процесса производства и возникающих в нем человеческих отношений. Эта тема сама по себе интересна, в том числе и для обсуждения на этом совещании, но дело не в основной линии доклада. Говоря об общенаучных предпосылках диалога с ЭВМ на естественном языке, докладчик, между прочим, сказал:

«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 очень большую научную, общекультурную и прикладную ценность. Не мне, конечно, составлять спецификацию такого фонда, но думается, что по крайней мере он должен содержать полный словарь и генератор словоформ, а также формализованный толковый словарь (тезаурус) русского языка. 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»[1].

[1] *Ершов А. П.* Методологические предпосылки продуктивного диалога с ЭВМ на естественном языке.— Вопр. философии, 1981, № 8, с. 115.

Это была первая высказанная вслух мысль о Машинном фонде русского языка. Тот факт, что я же и был докладчиком, не имеет особого значения; сейчас мне хочется обратить внимание на то, что эта постановка вопроса была сделана нелингвистом и адресована нелингвистической аудитории.

Надо сказать, что на этом же докладе мне был задан вопрос, заинтересовались ли лингвисты идеей Фонда. В тот момент вопрос был явно преждевременен, однако сегодня реакция превзошла всякие ожидания, и вообще это движение набирает силу столь стремительно, что полагать его реакцией на чей бы то ни было конкретный призыв было бы просто смешно.

И все же, оставаясь профессионально вне лингвистики, я как специалист по информатике беру на себя смелость продолжить этот внешний разговор о Машинном фонде русского языка как о важнейшей задаче, решение которой очень сильно повлияет как на саму лингвистику, так и на ее место в обществе. По существу должна произойти некоторая смена парадигмы этой науки и, собственно, это-то и придает мне смелости, так как известно, что предпосылки к смене парадигмы созревают зачастую вне рассматриваемой науки. Кроме того, я беру в союзники тридцатилетний опыт наблюдения за компьютеризацией самых разных областей деятельности, в том числе и научной, и, если угодно, мой собственный полувековой опыт применения русского языка: ведь человек, который пишет, читает, говорит, думает, вообще живет в своей родной языковой среде, относится к ней очень глубоко и активно. Кажется, у нас нет формально объявленного Общества любителей русской словесности, но на самом деле такое сообщество существует, объединяя сотни миллионов человек. От этого сообщества науке о языке передаются мощные импульсы потребностей и интереса, удовлетворению которых Машинный фонд может сильно помочь. В глазах этих людей русистика — это спящий Гулливер, которому пора проснуться и заговорить о себе полным голосом. Вычислительная техника может для этого оказаться подходящим будильником.

Каждый знает, какую роль сыграла в становлении В. И. Ленина как вождя русской революции написанная им в молодости работа «О развитии капитализма в России». Но сделаем еще один шаг. Эта работа не могла бы появиться без так называемой земской статистики. Я сошлюсь на свидетельство академика А. Г. Аганбегяна, который в одном из своих выступлений говорил об общей значимости земской статистики в России во второй половине XIX в., ознаменовавшей в экономической науке того времени переход от наблюдательного изучения к измерительному. Действительно, только такой переход позволил выдвинуть экономические методы управления, создать экономические модели — в целом приблизить экономику к разряду точных наук, сделать ее стратегическим оружием в развитии общества.

Между экономикой и языком как системами существует немало сходных черт. И та и другая являют как свою изменчивость,

так и свою устойчивость в миллиардах единичных актов человеческой деятельности. В каждой из этих систем их объективные законы получаются интегрированием индивидуальных актов сознания, которые при всей своей независимости, поддерживаемой волей участника событий, хотят тем не менее явно опираться на эти законы. Так вот, если поддерживать эту аналогию, то можно сказать, что русистике как науке тоже нужно в полной мере перейти от наблюдательного периода к измерительному. Конечно, этот переход давно наметился, но появление вычислительной техники дает этой проблеме совершенно новую постановку. И еще раз приведенная аналогия с экономикой показывает русистике как диапазон возможностей, так и величину дистанции, которую надо преодолеть, чтобы получать от языка прямые ответы на сто тысяч вопросов, которые ему хочется задать.

Итак, Машинный фонд русского языка может стать могущественным концентратором нашего знания о языке, знания полного, не отстающего от хода времени, детального и обобщенного вместе, подвижного и инерционного, накапливающего все предыдущие временны́е срезы.

Еще один аспект отношения «ЭВМ в языке» — это проблема овладения языком. Признано, что сейчас в культуре налицо две разнонаправленные тенденции. Одна из них — это явное увеличение роли печатного, услышанного и произнесенного слова. Круг словесного общения сейчас очень интенсивен. Средства массовой информации многократно повышают интенсивность такого общения, хотя, возможно, и делают его не столь прямым. Феноменальный рост личных книжных собраний дополняет эту интенсификацию. В то же время отмечается некоторое снижение уровня грамотности, не очень хорошо обстоит дело с изучением русского языка в национальных школах.

Можно, по-видимому, найти этому разумное объяснение путем более внимательного изучения обстановки в школе и в семье, пытаться улучшать положение традиционными методами. Есть, однако, ряд свидетельств тому, что широкое внедрение учебных ЭВМ в школе позволит радикально изменить ситуацию с изучением русского языка. Первые опыты применения ЭВМ в школе показывают, что машина может стать активным партнером в учебном процессе и не менее сильно стимулировать самого ребенка, создавая, где нужно, игровую обстановку, увеличивая элемент состязательности, обеспечивая индивидуальность и мгновенность реакции, побуждая к сознательному знанию и одновременно стимулируя и тренируя всяческую моторику, закрепляющую добытое знание.

Другой аспект интенсификации работы с русским языком я бы назвал проблемой «экспансии» языка, имея в виду его роль как языка межнационального общения и как мирового языка. И здесь я не вижу никаких причин тому, чтобы уклоняться от культурного соревнования с другими мировыми языками, и прежде всего с английским. И как раз тут у нас не все в порядке.

Зайдите в магазин иностранной литературы и посмотрите на обилие словарей английского языка: и один толковый словарь, и другой, и английский, и американский, словари синонимов, жаргонов, словари для студентов и школьников, с картинками и без, словари рифм, обратные словари и многое-многое другое. Жаль, что мы не обнаруживаем у себя такого обилия, такой кажущейся избыточности, такого языкового напора.

Конечно, мы делаем очень много. У нас есть замечательные словари. Например, у нас вышел великолепный однотомный «Энциклопедический словарь». Но ведь его нужно издавать каждый год, поддерживая конъюнктурные изменения. Праздником русской культуры стало появление Лермонтовской энциклопедии, но где энциклопедия Пушкина, Достоевского, Толстого, Чехова, Горького? Почему Академический словарь русского языка должен выходить один раз в пятнадцать — двадцать лет? Можно этому найти много причин и обнаружить их объективность. Но дело каждого преодолевать те причины, которые зависят от него.

Так вот, одна из причин — это та, что словарное и вообще лексикографическое дело для ведущих языков — английского, французского, немецкого — уже давно поставлено на машину и соединено с достижениями полиграфии, прежде всего с фото- и электронным набором. Можно и должно дорожить накопленными фондами в их теперешнем виде, уважать сложившиеся методы работы, но там, где идет речь об освоении пространства, старая карета и новая машина — не конкуренты.

Мне эта задача представляется очень важной. Она важна не только для повышения культурного уровня нашего общества и не только для усиления интегративных процессов в нашей стране, но и для реализации функции русского языка как языка международного общения. А для того чтобы понять, что тут все не так просто, достаточно просмотреть международные научные журналы, которые публикуют статьи как на русском, так и на английском языках, и посмотреть, как изменилось соотношение языков опубликованных текстов за последние двадцать лет.

Продолжая линию внешних постановок, я хотел бы обернуть отношение и поговорить о «языке в ЭВМ». Сначала небольшое философское отступление. За последние десятилетия начинает складываться представление о трех формах существования материи — вещества, отражающего постоянство материи, энергии, отражающей движение материи, и информации, отражающей структуру материи и сам процесс ее познания. Мало того, эти три ипостаси дают определенную периодизацию развития человеческой цивилизации: освоение вещества, создание материального производства, овладение принципом единства материи (от глубины тысячелетий до начала XX в.); затем освоение энергии, создание энергетического изобилия, овладение принципом единства энергии и ее связи с веществом (от XVIII до XXI в. в надежде на овладение термоядерной энергией) и, наконец, освоение информации, создание полной информационной доступ-

ности, овладение принципом единства процессов обработки информации в естественных и искусственных системах (от XX в. в будущее).

Здесь нам через развитие средств электросвязи, вычислительных средств, автоматики и робототехники предстоит где-то в середине XXI в. войти в период «сплошной информатизации», который в своем внешнем выражении будет характеризоваться следующими показателями:

грандиозная сеть электронной связи с полумиллиардом входов;

полная автоматизация техносферы с парой миллиардов встроенных микроЭВМ;

порядка двухсот-трехсот миллионов персональных ЭВМ и интеллектуальных терминалов, подключенных к сети связи;

десятимиллионная иерархия универсальных ЭВМ, поддерживающая управление обществом и сетью связи;

перенос на машинные носители практически всей информации, циркулирующей в обществе.

Все это должно не только существовать, но и воплощать полноту и доступность знания, хранить в себе достоверную информационную модель мира, быть в постоянном употреблении всеми людьми, реализуя на новой основе принцип единства человеческого рода. Возникает вопрос: как человек будет общаться с этой инфосферой, как он будет побуждать машины к действию, как он будет черпать из этого грандиозного фонда знания, как он будет относиться к новому жителю своего дома — компьютеру?

При первых же размышлениях над этим вопросом сразу возникает уже достаточно хорошо известная номенклатура конкретных технических проблем: организация диалога с базами данных; составление баз знаний и общение с ними (база знаний отличается от базы данных, грубо говоря, тем, что представляет теорию более высокого порядка: если в базе данных — конкретные факты, то в базе знаний — общие суждения и определенные дедуктивные свойства); извлечение смысла из текста в виде команд, фактов, энциклопедических знаний и т. п.; пополнение знаний (очень деликатная задача, так как речь идет не о механическом добавлении нового факта или суждения, а об установлении всех возможных связей с уже накопленной системой знаний); машинный перевод; синтез текста. На последней проблеме я хотел бы остановиться более подробно.

Каждый из нас сейчас читает много текстов, которые почти исключительно написаны людьми. Мы должны готовиться к тому, что это человеческое начало в текстовой информации начнет уступать заметное место машине. Хочу предупредить, что я совершенно не посягаю на фантастические сюжеты машинного стихосложения или любовных писем робота. Я говорю о реальных, прозаических вещах, связанных прежде всего с деловой прозой. Сейчас, например, автоматизация проектирования становится стержнем технического прогресса. Но неотъемлемой частью авто-

матического проектирования становится синтез технических описаний. Причем это не отдельные фразы или перечни слов в каких-то фиксированных таблицах. Это нормальный связный текст. Уклониться от рассмотрения такой проблемы нет никакой возможности. Во все большем количестве конструкторских бюро изготовление технической документации — узкое место. Профессия «технического писателя» самая дефицитная. Нет людей, которые имеют вкус к такой работе. Устранять человеческий язык из технической документации никак нельзя. Нам с этой техносферой надо сосуществовать веками и тем более нам нельзя терять наше человеческое. И если мы привыкли читать нормальный текст, мы предпочтем читать именно его, даже если он и синтезирован на машине.

Кроме технической документации, к этой синтетической прозе нужно будет отнести большое количество сводок, отчетов и других текстовых свидетельств. Здесь — мой особый призыв к лингвистам. Очень не хотелось бы обрекать деловую прозу на посредственность. На первый взгляд деловая проза бесконечно далека от интереснейших и красивейших задач внутреннего изучения русского языка. Есть ли у деловой прозы коммуникационная сверхфункция, есть ли у нее одухотворенность и чем она может быть красива — все это тоже, как мне кажется, достойнейшая задача большой науки о русском языке.

Мы хотим как можно глубже познать природу языка, и в частности русского. Одним из выражений этого познания должна стать модель русского языка. Это формальная система, которая должна быть адекватной и равнообъемной живому организму языка, но в то же время она должна быть анатомически отпрепарированной, разъятой, доступной для наблюдения, изучения и изменения. Я хочу сопоставить эту еще не существующую модель с другой моделью, созданной человечеством для изучения неживой природы,— математическим анализом и даже с одной лишь его ветвью — дифференциальными уравнениями. Нет никаких сомнений в том, что по сравнению с русским языком это куда более скромная область. И в то же время посмотрите, как огромен интеллектуальный багаж этой части науки. Сколько мы имеем разных руководств, теорий, сколько методов решений дифференциальных уравнений, как мощен поток литературы, как тщательно мы учим этому в университетах и технических вузах. Конечно, мы учим и русскому языку. Но если, действительно, сравнить размах области знаний, связанных с моделированием форм неживой материи, с нынешним статусом и объемом науки о языке, то станет ясно, что мы только в самом начале пути. Мне бы не хотелось этим предостережением сбивать ноту высокого оптимиза, на которой началась наша дискуссия, но, думается, что лишь сознание необходимости и неотвратимости качественного перелома в науке о языке, сознание ее общенаучной и государственной важности смогут создать тот запал и энтузиазм, без которого проблему Машинного фонда русского языка не поднять.

МЕТОДОЛОГИЯ ЛИНГВИСТИЧЕСКОГО ИССЛЕДОВАНИЯ И МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА

Ю. Н. КАРАУЛОВ

Машинный фонд русского языка мыслится как сложная, разветвленная и глубокая автоматизированная система, способная решать как информационно-поисковые, так и исследовательские языковедческие и шире — филологические задачи в области русистики. По отношению к известным нам зарубежным и отечественным лингвистическим и филологическим архивам (имеются в виду, например, текстовый архив «корпус Брауна» в США, лексико-семантический архив английского языка — там же, диалектный архив финского языка в Турку, архив итальянских литературных текстов от XIV в. до современности, терминологические банки разных языков и т. п.) Машинный фонд русского языка выступает как интегрированная система, учитывающая опыт и вбирающая в себя особенности названных, как правило, односторонних, однонаправленных, фондов и ориентированная по преимуществу на комплексные исследовательские цели. Интегрированный характер Фонда проявляется, в частности, и в том, что в его архитектуре предусматривается взаимодействие нескольких автономных подфондов, каждый из которых, будучи внутренне однородным, может расцениваться как аналог одного из уже существующих архивов. Так, иллюстративно-текстовой подфонд коррелирует с «корпусом Брауна», терминологический подфонд — с автоматическими терминологическими банками и т. д.

В качестве методологической основы построения Машинного фонда русского языка я выдвигаю два тезиса:

1. Любые данные о языке могут быть представлены в лексикографической форме. Если лексикографическое воплощение некоторых подсистем языка и даже целых уровней (например, синтаксиса) еще слабо разработано, это вовсе не означает, что оно принципиально недостижимо. В такой формулировке первый тезис является просто более сильной версией охарактеризованной мною ранее «тенденции к лексикографической параметризации языка», к «ословариванию» результатов лингвистических описаний, исследований и классификаций.

2. Любые лексикографированные данные о языке могут быть переведены в алгоритмизированную, машинную форму. Этот тезис возникает отчасти как обобщение уже реализованного на ЭВМ опыта разных стран, организаций и отдельных исследователей, а отчасти — и как развитие первого тезиса. Поскольку известно: то, что может быть представлено в словарной форме, доступно, следовательно, машинному представлению, а коль скоро в словарной форме представимы любые данные о языке, то и представлению в машине они, естественно, поддаются.

Таким образом, идеология Машинного фонда русского языка лексикографическая, словарная в своей основе. Именно исходя из этих соображений, но и не только из них, я считаю, что узловой единицей, стержнем, вокруг которого будет группироваться языковая информация в Фонде, должно стать слово. Не морфема, не концепт, не предложение, а слово как наиболее естественная единица языка, которая структурируется и в глубину — на морфемы, фонемы, слоги, буквы, значения, и во вне как член словоизменительных, словообразовательных, синтаксических, грамматических, ассоциативных и концептуальных рядов. Более того, в каждом слове амальгамирована его история, которая, в свою очередь, коррелирует с историей народа и его культуры, а следовательно, Фонд должен быть организован так, чтобы всю соответствующую информацию, сконцентрированную вокруг данного слова, в любой момент можно было извлечь, амальгаму расшифровать и все это дискурсивным образом представить. Значит, в основе способа хранения информации лежит своеобразное «досье» на каждое слово, представленное как подробно структурированная область базы данных, содержащая какое-то число потенциальных сегментов, которые отражают всю систему грамматических, стилистических, фонетических, контекстуальных и т. п. параметров, т. е. всех возможных сведений о слове. Очевидно, должна быть разработана лишь принципиальная схема такой области и соответствующим образом подготовлены источники для автоматического формирования «досье» или какой-то его части. Хранить такую зону в компактном виде и невыгодно, и нерационально, поскольку отдельные параметры для разных задач могут «склеиваться» в особые лексикографические модули, а резервные «пустые» клетки — заполняться информацией «на случай», некоторыми данными ad hoc. Степень изученности русского языка, степень готовности русистики к такому способу обобщения своих результатов я определяю как достаточную, адекватную сформулированной задаче.

По пути предлагаемого обобщения данных о каждом слове мы можем пойти и дальше: представим себе, что мы имеем возможность включить в базу данных не только стандартные лексикографические параметры в том виде, как они обозначены, например, в моей книге [1], но и более широкие сведения о варьировании слова, понимая под последним не обычные нормативные колебания вариантов, а своеобразное гиперварьирование. Такое понимание потребует отказа от методологического ограничения в трактовке этого понятия, навязанной теорией языкового варьирования, существенного расширения его амплитуды вплоть до границ, когда коммуникация и взаимопонимание еще возможно. Тогда в сферу гиперварьирования, а соответственно и в потенциальную структуру нашей области войдут:

[1] *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. М., 1981.

устно-разговорное (субнормативное) варьирование на разных уровнях. Например, на уровне грамматических характеристик — *черный/черное кофе*; на уровне семантизации — расширение синонимического ряда: *Я вот тут пурпурю, около храма...* (глагол *пурпурить* употреблен в значении 'наводить порядок, придавать блеск', а храмом названо здание пансионата);

поэтическое (наднормативное) варьирование, например, в поэтической лексической парадигме, где в качестве эквивалентов выступают не только общеязыковые синонимы: *туман—мгла—дым—пелена—завеса—кудри седые—облака*;

территориально-временно́е (диалектное) варьирование — анормативное;

историко-временно́е, или диахроническое (донормативное), варьирование, которое может быть представлено как в виде отдельных проявлений, модификаций слова (*стакан—достакан, любопытность—любопытство*), так и в виде набора его алломорфов и хрономорфов;

наконец, максимальной амплитуды достигает варьирование единиц русского языка при использовании его в межнациональном общении, когда влияние родного языка говорящего — даже без нарушения нормального протекания коммуникации — может доводить степень изменчивости единицы до уровня «ошибки». Например, во фразе из выступления: *Мы не должны делать розовые глаза, когда видим такие явления*,— произошло, как свидетельствует контекст (здесь опущенный), разрушение идиомы *смотреть сквозь розовые очки* и ее креолизация с идиомой *делать большие глаза*. Однако благодаря контексту и возможности семантизации по ключевому слову *розовые* понимание не нарушается. Каталог таких ошибок, связанных с употреблением данного слова или эквивалентной ему единицы в межнациональном общении, составит специальную область слова в базе данных и будет служить задачам обучения русскому языку как неродному, задачам повышения речевой культуры лиц нерусской национальности.

Итак, сфера гиперварьирования слова в файле должна состоять из зон нормативного, субнормативного, наднормативного, анормативного, донормативного и территориального варьирования, а каждый из названных параметров вариативности должен формироваться из специального подфонда текстов и словарей — разговорной речи, поэтического языка, диалектного, исторического, русского как межнационального и т. д.

В целом лексикографическая идеология Машинного фонда русского языка представляется перспективной, и методологические принципы его построения, основывающиеся на двух охарактеризованных тезисах, дополнительных ограничений на поставленную задачу, по-видимому, не накладывают.

Ограничения другого рода, которые можно назвать эмпирическими, связаны с выбором того или иного пути практического формирования Фонда и его ориентации на определенные типы работ. Здесь предстоит рассмотреть такие дилеммы:

1. Будет ли Фонд создаваться и функционировать только как архив — пусть и многоаспектный — как банк готовых сведений о русском языке или это будет исследовательская лаборатория, обеспечивающая производство новых знаний, добычу и извлечение новых данных о языке? Если мы принимаем первую альтернативу, то мы имеем дело с гигантской, во много раз увеличенной по сравнению с известными, хорошо параметризованной и систематизированной, но всего лишь с картотекой. В такой картотеке инвентаризованы, грубо говоря, вещи и свойства, но отсутствуют отношения. Ценность подобной гиперкартотеки безусловна: она дает возможность передать машине трудоемкие рутинные операции и получать по запросу исследователя — в нужном объеме и сразу — списки, перечни любых единиц и элементов системы, выбранных по заданным условиям. Хотя это выгодно и даже заманчиво, но всем нам хочется, конечно, разрешить рассматриваемую дилемму в пользу второй альтернативы, т. е. «лаборатории». Во-первых, потому, что такое решение богаче возможностями, так как вбирает в себя также и все достоинства первого решения; во-вторых же, потому, что интуитивно на первом пути мы ощущаем, что в какой-то момент нас ждет своеобразное перенасыщение, «отравление данными», с которыми мы не в состоянии будем справиться. Ведь человек был и остается самым важным звеном во всех системах обработки и передачи информации, и ее компрессия и манипулирование ею с помощью специальных эвристик становится насущной задачей. Значит, Фонд должен как бы обладать «пониманием» известных нам языковых законов, в нем должны быть смоделированы синхронные и диахронные языковые процессы, должны быть в динамике представлены, функционировать отношения единиц, уровней, подсистем. Следовательно, чтобы быть «лабораторией», проектируемая система должна быть не только банком данных самого объекта, но и банком научных знаний об объекте, т. е. Машинный фонд русского языка — это и банк данных, и «разумный», специализированный банк знаний, которым можно пользоваться как книгой, способной предложить решение или нацелить исследователя на определенное решение.

2. Следующую дилемму я сформулировал бы в виде оппозиции: культура или структура? Речь идет о том, следует ли в Машинный фонд включать лишь собственно лингвистические данные, относящиеся к разным уровням языковой структуры и поддающиеся достаточно последовательной параметризации, или в его составе с необходимостью следует учесть и лингво-культурный, историко-этнографический аспект русского языка? Я думаю, все мы, будучи минималистами по отношению к задачам сегодняшнего дня, дружно выступим как максималисты с позиций дня завтрашнего и поэтому выскажемся за то, чтобы в Машинном фонде русского языка были представлены обе альтернативы — и «культура» и «структура». Во-первых, потому, что эти два аспекта жизни языка органически взаимодействуют, взаимопро-

никают друг в друга. Во-вторых же, потому, что мы с самого начала высказываемся за всесторонность и междисциплинарность информации, связанной с русским языком и входящей в Фонд.

Взаимопроникновение двух названных аспектов может рассматриваться до глубины только национальной русской культуры, русского видения мира, но может рассматриваться и до глубины человеческой цивилизации в целом. Хороший пример последнего подхода — в творчестве индийской художницы Амины Ахуджи. В ее картине художественный образ, навеянный, например, двустишием А. Майкова:

Гармонии стиха божественные тайны
Не думай разгадать по книгам мудрецов,

предстает в виде изящно изогнувшего шею и застывшего с поднятым передним копытом благородного коня. Причем сам рисунок выполнен не обычными штриховыми линиями, не красками, а кириллическими буквами, повторяющими приведенное двустишие. Это уже не абстрактные арабески-украшения, а своего рода «орфопись», конкретно-образные «кириллески», если можно так выразиться. И когда такой рисунок видит и читает (иначе не скажешь) русский человек, то в его душе (а здесь уже без слова «душа» не обойтись) рождается адекватный отклик: ему на память приходят очень похожие, может быть, несколько условные, стеклянные фигурки игрушечных коней, изготовляемых народными умельцами в г. Гусь-Хрустальный, или вспоминаются стихи А. К. Толстого

Край ты мой, родимый край,
Конский бег на воле,
В небе крик орлиных стай,
Волчий голос в поле.

Гой ты, родина моя,
Гой ты, бор дремучий,
Свист полночный соловья,
Ветер, степь да тучи.

Основная эмпирическая сложность в реализации решения — «и культура, и структура» — связана с недостаточностью наших знаний о взаимодействии двух названных аспектов и с недостаточностью опыта в манипулировании такой синкретической, комплексной информацией о русском языке. «Душа» не параметризуется. Более того, у меня нет уверенности, что подобная информация, т. е. культурно-лингвистическая, может быть параметризована в обычном смысле, например, так, как это делается частично в наших словарях. В них, по сути дела, находит отражение (да и то далеко не полностью) лишь страноведческая и отчасти историко-этнографическая информация о слове. Правда, у нас нет пока словаря русского фольклора, нет ассоциативного тезауруса и словаря поэтических парадигм, но даже если бы та-

80783

www.ruslang.ru

кие словари были, они решали бы проблему параметризации культурно-лингвистической информации, вероятно, только отчасти.

Может показаться, что констатируемое таким образом положение противоречит двум методологическим тезисам, сформулированным в начале моего выступления. На деле это не так. Нам необходимо найти новые подходы, новые категории и приемы обращения с языково-культурной информацией. Принципиальное, на мой взгляд, отличие «структурной», условно говоря, информации от «культурной» заключается в том, что при манипулировании первой мы, как правило, легко обходимся линейными зависимостями, линейными функциями. Так, глагольное управление, например, однозначно задается соответствующими валентностными характеристиками. Если нам известно, что глагол *завершить* имеет одну обязательную (прямой падеж — *завершить что*) и две факультативные валентности (*завершить чем* и обстоятельственное управление, как: *он завершил свой доклад песней под аплодисменты зала*), то при пассивизации этого глагола реализуется линейная зависимость: у него остается лишь две валентности, причем бывшая факультативной становится обязательной (*завершиться чем*), а вторая факультативная сохраняется без изменений. Подобные линейные зависимости имеют место при разного рода грамматических преобразованиях на всех уровнях языковой структуры: таковы словообразовательные цепочки, звуковые изменения, словоизменительные парадигмы и т. п. Они предсказуемы, хорошо описаны и довольно просто параметризуются, во всяком случае — при анализе. Сложнее дело обстоит с синтезом тех же явлений. Здесь проявляется уже нелинейность, нелинейные закономерности, которые предполагают изменчивость зависимых величин от многих факторов, часто не учитываемых, не принимаемых во внимание при анализе. Так, при словообразовательном синтезе мы сталкиваемся с полимотивированностью дериватов, при семантическом синтезе — с множественностью поверхностных структур и т. п. Эта структурная нелинейность в общем преодолевается нами с использованием разных хитрых приемов линеаризации. Для «культурной» нелинейности эти приемы не годятся. Надо искать новые способы описания этих явлений, их лексикографирования, если мы хотим совместить в Фонде «культуру» со «структурой».

Определенные успехи лингвистики текста обнадеживают и дают основания думать, что часть задач в этой области мы сможем решить с ее помощью. А это означает, что тексты — исторические, литературно-художественные, публицистические, фольклорные — то, что называют иногда первичной языковой информацией, должны составить обязательную часть нашего Фонда. Однако и этого недостаточно. По сути дела, необходимо развернуть целую серию исследований по программе «Русский язык и русская культура», чтобы быть готовыми одолеть эту вторую эмпирическую трудность, или ограничение.

3. Наконец, третья дилемма — производственно-прикладная или филологическая ориентированность Фонда. Существующие в разных областях и для разных языков мини-фонды характеризуются главным образом производственной ориентированностью, нацелены на решение узких задач и немало страдают от этого. Из моих предыдущих рассуждений слушателям ясно, что я склоняюсь к другой альтернативе, прежде всего потому, что филологическая ориентированность требует универсальности. Но при этом мне хочется подчеркнуть, что производственный аспект тем не менее не должен упускаться из виду. И здесь становится очевидным следующее эмпирическое ограничение: мы практически очень мало знаем о русском языке в том его варианте, в каком он существует и функционирует в разного рода информационно-поисковых и документально-фактографических системах. По объему языковых данных и своих функций этот вариант теперь уже не уступает, видимо, тому, который находится в обращении между людьми, обслуживая наши коммуникативные и познавательные потребности. Значит, если раньше стояла задача приспособить естественный язык — в некотором его ограниченном объеме — для обучения ему машины и последующего общения с нею человека на этом языке, т. е. нужно было синтезировать человеко-машинный язык, то теперь стоит задача чисто лингвистическая: обобщить опыт его использования в этой роли, дать его полный анализ и описание, точно так же, как мы описываем региональный язык (говор) и особенности профессиональных или социальных диалектов.

Человеко-машинный русский язык (или стиль?) носит искусственный характер, хотя основа-то его вполне естественна. В этом отношении он сопоставим с искусственными языками типа эсперанто и как объект исследования мог бы быть включен в сферу интерлингвистики. Из разновидностей русского языка, функционирующих в человеческой среде, к нему ближе всего стоят деловой и научный стили. Задача лингвистического изучения человеко-машинного языка становится особенно актуальной в связи с развитием систем типа «искусственный интеллект». Для таких систем характерно использование адаптированного естественного языка на входе, на выходе и внутри, причем его функционирование внутри самой системы могло бы служить прообразом того языка, который выступает как посредник, промежуточный язык между естественным интеллектом и естественным языком и который Н. И. Жинкин назвал «универсальным предметным кодом». Таким образом, в каких-то его частях человеко-машинный язык мог бы рассматриваться и в качестве модели языка человеческого интеллекта, обеспечивая нам тем самым выход от прикладного аспекта Машинного фонда к фундаментальным проблемам науки о языке. Так появляется оппозиция прикладной — фундаментальный.

С одной стороны, противопоставление прикладного и фундаментального — это просто иной облик, более широкая формули-

ровка все той же рассматриваемой мною дилеммы — производственный характер Фонда или филологический. С другой стороны, за членами этого противопоставления более отчетливо, чем в первой формулировке, встает не только ориентация Фонда, но сам его статус и оценка как объекта либо фундаментального исследования, либо прикладной разработки, а та или иная оценка чревата очень сильными по своей значимости психологическими последствиями. Ведь одно дело — видеть в Машинном фонде такую конструктивную задачу, решение которой стимулирует выдвижение новых интересных идей и проблем в изучении русского языка, отдавать себе отчет в том, что любое инженерно-техническое, конструкторское решение всегда богаче по своим возможностям в порождении актуальных и потенциальных проблем и следствий, чем чисто теоретические разработки, и совсем иное дело — подходить к Фонду как к примитивному переносу готовых результатов в другую область, воплощению их в новой технологии, внедрению в практику. Оценка Машинного фонда русского языка как рутинного дела, как чисто технической реализации уже достигнутых и существующих решений, на мой взгляд, помимо того, что она не отвечает содержанию и характеру работы по созданию Фонда, с психологической точки зрения повредила бы нашему продвижению вперед. В глазах большинства людей, занимающихся научной деятельностью, фундаментальные, теоретические исследования имеют больший вес и престижность, чем прикладные. Но ведь границы-то между тем и другим зыбки и неопределенны. Нам только кажется, что мы всегда можем четко развести два эти типа. На деле же представление о том, что все, что неприкладное, что не имеет непосредственной практической приложимости, все это фундаментально и теоретично — такое представление ложно. Среди проблем, не имеющих приложимости, еще много псевдопроблем, игры в исследовательскую деятельность. Но однозначных и надежных критериев отделения подлинно фундаментальных проблем от лжепроблем у нас нет.

Перехожу к характеристике третьей серии ограничений, которые я назвал теоретическими и которые связаны с выполнением требования, сформулированного в самом начале: Фонд должен стать не просто банком данных, а творческой лабораторией. Это значит, что лингвист, пользующийся Фондом, не ограничивается наведением справок, а находится в собственно исследовательской позиции и сама организация Фонда осуществлена таким образом, что полностью соответствует, удовлетворяет этой позиции.

Что же означает *исследовательская позиция*? В самом общем виде она предполагает три момента: во-первых, рефлексию над материалом, его субъективное переживание; а чтобы рефлексия была возможна, необходимо, во-вторых, перекомбинирование, переконцентрация материала, получение новых классификаций; наконец, в-третьих, исследовательская позиция предполагает отчетливо сформулированную цель и определенную последовательность операций, шагов по ее достижению. Ясно, что

рефлексию осуществляет исследователь и заложить ее в Фонд мы не можем. Иное дело — переконцентрация материала. Именно такую возможность Машинный фонд и должен обеспечивать в самых широких масштабах на основании максимально разветвленной и дробной параметризации языковых данных и явлений. Сложнее обстоит дело с последней характеристикой исследовательской позиции — ее целевой ориентированностью и программой операций. При ближайшем рассмотрении оказывается, что ее нельзя брать недифференцированно: комбинации разных целей и техник оказываются настолько разнообразными, что создаются определенные типы исследовательских позиций, о которых мы практически очень мало пока знаем. А ведь Фонд должен обеспечить реализацию любой исследовательской позиции, и, значит, в его организацию должна быть заложена типология самих лингвистических подходов, лингвистических исследований. Теоретическую трудность здесь я вижу в том, что типы исследований в лингвистике никогда нами не дифференцировались, и задача теперь состоит в том, чтобы взглянуть на этот вопрос по-новому. Рассмотрим с этой точки зрения некоторые реальные или воображаемые темы.

Там, где речь идет о получении разного рода словников и выборок, особых проблем как будто не возникает: например, словари писателей, словоуказатели к текстам, словари новых слов, перечень глаголов с приставками *раз-*, *на-*, *пере-*, *за-*, имеющих значение интенсивности действия (*разобидеть*, *надымить*, *перестараться*, *запариться*) и т. п., выборки могут осуществляться автоматически — и из текстов, и из словарей. На таких списках могут развертываться некоторые типы главным образом классификационных исследований. Например, языкотворчество писателя — в этом случае выборка новообразований в его текстах будет получена автоматически путем сопоставления состава его словаря со стандартным генеральным словником Фонда. Дальше, работа исследователя — рефлективная и классификационная, причем последняя опять-таки может опираться на данные Фонда в области словообразования и синтаксической сочетаемости. Аналогичным образом может быть организована выборка фразеологизмов для изучения их индивидуального варьирования и использования в текстах, для изобразительных синтаксических конструкций; или может быть решена задача сведения к общему знаменателю этимологий одного и того же слова по разным источникам и снятия противоречий в его этимологизировании (при условии, конечно, что разные источники введены в ЭВМ и хранятся в Фонде).

В таком, например, традиционном типе исследования, как история слова, существенно облегчается и сокращается по времени этап сбора материала, т. е. подборки контекстов его употребления в разные эпохи и в источниках разных жанров (опять-таки при условии, что необходимое число источников введено в Фонд). Конечно, требует прояснения и это самое «необходимое

число»: сколько это — много или мало? Может быть, чем больше, тем лучше? В справедливости последнего во всяком случае я сомневаюсь. Но главную проблему в такого типа исследованиях составит идентификация самого слова, история которого воссоздается, поскольку орфографический его облик на протяжении времени менялся, причем менялся не только «по правилам», т. е. в соответствии с известными фонетическими и морфологическими законами, но иногда и помимо них. Здесь многое нам еще предстоит обобщить и уточнить, но тем не менее задача идентификации слова представляется вполне решаемой.

Иначе, как мне кажется, обстоит дело с таким типом исследования, как выявление поэтических парадигм, или построение словаря образов русской поэзии. В этом случае, даже имея в своем распоряжении все необходимые тексты, трудно предложить такие операции, чтобы автоматически удалось установить, что у Фета, например, о беде, невзгодах, неприятностях сказано так:

гроза:

Если тихо у нас и не веет *грозой*;

гроза житейская:

...хоть снесешь не одну ты *житейскую грозу*;

сумрак:

Твой чистый луч передо мной горел...
Но *сумрака* кругом не одолел;

ночь жизни:

Что *жизни ночь* над нами лишь сгустится,
Твоя душа... Передо мной умчавшись, загорится;

мрак жизни вседневной:

Еще теменее *мрак жизни вседневной*;

терний язвительные:

Можно ли, друг мой, томиться в тяжелой кручине?
Как не забыть, хоть на время, *язвительных терний*?

дебри и метели:

Куда безумство завлекло?
В какие *дебри и метели* я уносил твое тепло?[2]

Принципиальная особенность членов таких поэтических парадигм в том, что их нельзя семантизировать, т. е. идентифицировать их семантику по близкому контексту, как семантизируется, например, то или иное значение полисемичного слова. В последнем случае эта процедура поддается формализации, и по специальному «контекстологическому словарю» мы можем установить, скажем, что многозначный английский глагол engender при наличии справа от него слов fear (n.) или resentment пере-

[2] Примеры заимствованы из работы: *Григорьева А. Д., Иванова Н. Н.* Язык поэзии XIX века. М., 1985.

водится 'возбудить', при наличии справа от него слова confidence переводится 'породить', во всех остальных случаях — 'вызвать'[3]. Более того, контекст дает исследователю возможность восходить от уровня пословной семантизации к более высокому когнитивному уровню понимания и толкования. Так, используя некоторые формальным образом выделяемые контекстуальные индикаторы, исследователь оказывается в состоянии производить, т. е. почти автоматически получать полное семантическое описание соответствующего слова, иными словами,— строить его дефиницию по контекстам и из их фрагментов. Я имею в виду не обязательно лексикографическую дефиницию, а в широком смысле — семантическую, которая может находить разнообразные применения.

В отличие от подобных типов исследовательских позиций поэтическая парадигма не семантизируется дискурсивным образом, а постигается, открывается исследователю сразу, вдруг — на основе герменевтического проникновения в целый текст, филологического его понимания как целостного единства. И здесь уже решающую роль играет рефлектирующая позиция наблюдателя, которому никакие контекстуальные «индикаторы» не помогут установить поэтическую многозначность слова *огонь*, например, у того же Фета:

о любви:

Люби меня!... Окрылены неведомым стремленьем...
В каком огне, с каким самозабвеньем
Мы полетим.
Светил нам день, будя огонь в крови;

о молнии:

Под сенью ласковых ресниц
Огонь небесный мне не страшен;

о поэзии:

Кто плюет на алтарь, где твой огонь горит;

о разуме:

...огонь,
Что просиял над целым мирозданьем
И в ночь идет, и плачет, уходя;

о рассвете:

И до огней зари восточной
Под звездным пологом усни;

о силе возможности (сделать что-л.):

Где кистью трепетной я наберу огня,
Где я возьму небесной краски?

---

[3] *Марчук Ю. Н.* Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский. М., 1976. Ч. 1, 2.

о стихах:

Ты вынес, отрок дерзновенный,
свои алмазные огни...

Хотя некоторой подсказкой исследователю могли бы служить и определенные формальные моменты — тонкие различия в грамматическом оформлении слова, особенности его сочетаемости (например, притяжательное местоимение непервого лица — *твой, свои* — отождествляют огонь с поэзией, стихами) и т. п. Таким образом, чтобы Фонд был адекватен исследовательским задачам, взаимопереходы и взаимосвязи между его частями, или подфондами, а также между базами данных внутри одного подфонда (например, морфологической и словообразовательной внутри грамматического подфонда) должны строиться с учетом обслуживания всего многообразия исследовательских позиций. Ну, а типы исследовательских позиций в общем-то нам еще предстоит установить.

Итак, когда мы говорим о Машинном фонде русского языка, на первый план выдвигаются новые технические возможности, более перспективная технология, удовлетворяющая насущному внутреннему требованию современного исследователя — не ограничиваться образцами реализации того или иного правила, явления, иллюстрациями проявления какого-то признака, приблизительными обозначениями некоторой материальной области языковой структуры, а в каждом конкретном случае получать точное знание, доходить до исчерпывающего перечня всех элементов соответствующего множества. Но это качество Фонда лежит на поверхности. Более важны методологические последствия постановки такой задачи.

Для науки о русском языке Машинный фонд — это прежде всего возможность взглянуть на свой объект в новом ракурсе. Чем масштабнее будет Фонд, тем глубже в особенности строения языка он позволит проникнуть, тем шире раздвигает он рамки понимания самого объекта, отчетливее выявляет «белые пятна» в наших знаниях о нем. Даже беглое рассмотрение методологических, эмпирических и теоретических ограничений на решение поставленной задачи подводит к необходимости наметить некоторые направления, без проработки которых нельзя ожидать ни полноценного формирования Фонда, ни его эффективного использования. Это — и дальнейшее «ословаривание» результатов изучения разных уровней, распространение «лексикографической параметризации», которая в семантическом синтаксисе, например, выступает как прием линеаризации; это и разработка базы данных, обобщающей все реальные и потенциальные аспекты изучения слова; это необходимость освоения русистикой сферы «деловой прозы» и человеко-машинного языка; это и задача алгоритмизации правил идентификации слова в диахронии; это, наконец, осмысление феномена гиперварьирования. Перечисленные задачи служат и стимуляторами, и одновременно подцелями создания Машинного фонда русского языка.

Другое важное достоинство глобального взгляда на проблемы русистики с позиций Машинного фонда заключается в том, что у нас появляется возможность разделить их на решаемые и не решаемые с помощью Фонда. В сфере самого объекта — русского языка — принципы Фонда опираются на системные закономерности его организации, понимаемые в данном случае не как общие методологические установки, некоторые стратегические формулы, своего рода исследовательские горизонты, но воплощенные в конкретных алгоритмах структурных преобразований, в машинных процессах, вполне определенных технических зависимостях. И в этом отношении Машинный фонд — апофеоз системности, можно сказать, сама материализованная система языка. И вот здесь-то мы впервые начинаем осознавать его ограниченность как метода и как прибора одновременно, наталкиваемся на рамки, которые ставит нам именно система. Дело в том, что понимание системности, которое, главным образом, и реализуется в лингвистике и русистике за последние 70 лет, в качестве обязательного условия предполагает рассмотрение языка «в себе и для себя», говоря словами Соссюра. Однако выдвижение и постановка таких проблем, как «Русский язык и русская культура», «Естественный язык и язык интеллекта», «Структура и функционирование языковой личности», которые возникают, конечно, не в связи только с Машинным фондом, но которые можно уже разглядеть на указателях магистрального развития лингвистики и невооруженным глазом, требуют исследовательских действий по ту сторону традиционным образом понимаемой системности. А это в общем означает изменение и в традиционных представлениях об объекте лингвистики, подведение некоторой черты под соссюровским периодом ее развития. Языковая система становится лишь частью объекта науки о языке, а главное внимание переносится на носителя языка, «человек с языком» и «язык в человеке». Итак, Машинный фонд русского языка многократно усиливает нашу исследовательскую позицию, высвобождает творческую энергию исследователя, которая направляется на доработку системных свойств языка, достройку языковой системы как таковой. Программа по созданию Фонда должна быть комплексной в том смысле, что содержит равно необходимые и теоретические, и экспериментальные разработки. Программа по созданию Фонда должна быть целевой, так как в результате ее осуществления возникает некоторый новый объект, материальная система, новый суперприбор для исследования языка во всем многообразии его свойств и проявлений.

КОНЦЕПЦИЯ И АРХИТЕКТУРА МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА

В. М. АНДРЮЩЕНКО

1

Впервые задача создания Машинного фонда русского языка была сформулирована А. П. Ершовым в докладе «К методологии построения диалоговых систем: феномен деловой прозы»[1] в 1978 г. В связи с проблемой взаимодействия человека и ЭВМ на естественном языке, прежде всего в производственных и других 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 на 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 исполнения, подрывая все выгоды автоматизации. Мы не можем внедрить машины в повседневную жизнь, выделяя касту жрецов-посредников, но в таком случае мы должны быть готовы к тому, что число человеко-машинных интерфейсов в диалоговых системах возрастает с течением предстоящих 50 лет на несколько порядков. Не хватит никаких сил на то, чтобы снабдить эти миллионы интерфейсов специализированными языками и процессорами».

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 (разрядка наша.— *В. А.*). Это фундаментальная проблема, решение которой будет иметь очень большую научную, общекультурную и прикладную ценность... Очень хотелось бы видеть, что создание машинного фонда русского языка квалифицированными лингвистами опережало бы создание производственных лингвистических систем (разрядка наша.— *В. А.*), потому что это не только бы позволило избежать дублирования больших усилий, но и защитило бы здо-

[1] *Ершов А. П.* К методологии построения диалоговых систем: феномен деловой прозы.— Препр. № 156, ВЦ СО АН СССР. Новосибирск, 1979.

ровую ткань русского языка от самоуправства и неквалифицированного подхода».

Идея создания Машинного фонда русского языка как технологической основы для разработки систем общения с ЭВМ и обработки данных на естественном языке нашла широкий отклик и послужила предметом многочисленных обсуждений на конференциях и семинарах — как в лингвистической среде, так и в среде системных программистов и разработчиков ЭВМ. Актуальность этой задачи с каждым днем возрастает: сегодня она может быть рассмотрена также в связи с обсуждением проектов ЭВМ 5-го поколения, которые будут обладать встроенными языковыми процессорами и программно-аппаратными средствами обеспечения банков знаний.

Осенью 1982 г. Научный совет по лексикологии и лексикографии АН СССР совместно с Секцией лингвистических проблем обработки информации Совета по комплексной проблеме «Кибернетика» при Президиуме АН СССР распространил с целью собрать и систематизировать мнения специалистов вопросник, посвященный проблемам создания Машинного фонда русского языка. Полученные ответы на вопросы, записи бесед и обсуждений, замечания к текстам документов комиссии по машинным фондам языка, образованной Секцией лингвистических проблем обработки информации, послужили материалом для написания данной статьи [2].

2

Наиболее общее толкование понятию «Машинный фонд русского языка» можно дать, если представить его как систему комплексной автоматизации лингвистических исследований и разработок, состоящую из накопленных лингвистических данных, объективированных текстами, картотеками, словарями, грамматиками и другими лингвистическими источниками, и программного обеспечения использования этих данных и конструирования новых лингвистических объектов — словарей, грамматик, языко-

---

[2] Ответы на вопросник 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. Пиотровский, 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.

Материалы 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, мнение которых оказало 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.

вых процессоров, которые, в свою очередь, могут войти в Фонд в качестве единиц хранения и источников новых данных и средств для новых разработок.

Такая система может быть рассмотрена с нескольких точек зрения: организационной, пользовательской, информационной, кибернетической, технологической и вычислительно-лингвистической.

2.1. В организационном аспекте Машинный фонд русского языка — это учреждение, аналогичное фонду алгоритмов и программ, который содержит, как правило, справочные и информационные материалы, программные документы и средства, предназначенные для использования в вычислительных центрах. Однако в отличие от фонда алгоритмов и программ Машинный фонд русского языка должен получать, хранить и передавать не только программные документы и программные продукты, но и данные — тексты, словари и грамматики в машинной форме. Другим важным отличием Машинного фонда русского языка от фонда алгоритмов и программ является необходимость ведения в самом Фонде научно-исследовательской и опытно-конструкторской работы, связанной, во-первых, с переработкой поступающих данных в новые разработки, например, объединение различных словарей, текстов и программ в новые системы, пакеты программ и т. д., и, во-вторых, с созданием нового программного, лингвистического и информационного обеспечения Фонда русского языка. Пользователями Фонда могут быть любые лица и организации, заинтересованные в получении и использовании этих материалов.

2.2. В пользовательском аспекте Машинный фонд русского языка представляется как система автоматизированных (т. е. программно-управляемых) картотек, содержащих текстовые, словарные, грамматические, анкетные, программные и документальные источники данных о русском языке во всем объеме этого понятия. В понятие данных о русском языке мы включаем: данные о лексике и семантике в виде словарных статей различных словарей, сегментов текстов, содержащих употребление каждого учтенного слова, статистических данных об употребительности слова и различных комментариев к слову, содержащихся в грамматиках и лингвистических трудах; данные о грамматике русского языка в виде тезауруса, адресующего категориальный состав грамматических форм и конструкций, и свода правил, характеризованных условиями применимости в определенном структурном или лексико-семантическом контексте; данные о фонетике и фонологии, о морфемике и морфонологии, включаемые в состав словарных статей и грамматической информации; диалектологические данные в виде словарей и лингвогеографических анкет, социолингвистические и психолингвистические данные в виде источников, применяемых в соответствующих дисциплинах, и т. д. Любой факт, квалифицируемый как явление русского языка, должен найти в конечном счете

отражение в этих картотеках либо в виде объекта хранения, либо атрибута одного или нескольких объектов.

Программное обеспечение таких картотек должно содержать средства для выборки фактов и явлений, средства статистической оценки употребительности явлений, степени их связанности и взаимообусловленности, средства поиска и группировки фактов, обладающих теми или иными заданными характеристиками, и т. д.

Наряду с пользовательским, информационно-справочным обеспечением картотека должна быть оснащена средствами ввода и редактирования данных, лингвистического анализа текстов, оперирования словарями как целыми источниками, например средствами объединения словарей и их сравнения (пересечения и взятия симметрической разности).

Большое значение имеют также программные средства литературного и технического редактирования, автоматической корректуры и автоматического набора. Наличие таких средств позволит сосредоточить весь процесс подготовки издания в месте его создания и сделать компоненты Фонда, представляющие интерес для широких слоев читателей, доступными для них без дополнительных затрат труда издателей и полиграфистов, исключая, разумеется, работу по тиражированию.

Особое значение этот последний аспект имеет в связи с подготовкой новых изданий словарей: словари, содержащиеся в Машинном фонде, будут непрерывно пополняться новой лексикой из поступающих текстов, каждое слово может быть непосредственно проверено на текстах во всех своих аспектах, соответствующие словарные статьи могут быть отредактированы, так что машинный словарь в каждый момент времени будет представлять собой готовое к полиграфическому воспроизведению новое «исправленное и дополненное издание».

Со временем такой фонд языковых источников мог бы дополнить многочисленные и объемные словарные и справочные картотеки данных по русскому языку Академии наук и вузов, а также многочисленных коллективов отраслевых научно-исследовательских и информационных институтов и служб, связать их воедино, обогатиться программным обеспечением, разработанным для автоматического анализа и синтеза русского текста и общения с ЭВМ на естественном языке и, таким образом, способствовать как повышению научного уровня инженерно-прикладных разработок, так и обогащению академических материалов о русском языке за счет включения в фонды Академии терминологических фондов отраслевых институтов, результатов и достижений вычислительной лингвистики. Это даст, в свою очередь, новый импульс для развития теории как общего языкознания, так и русистики, позволит с большей глубиной решать теоретические и прикладные проблемы, связанные с ролью русского языка в современном обществе, в эпоху научно-технической революции.

2.3. В информационном аспекте Машинный фонд русского языка мыслится как распределенный автоматический банк данных, содержащий несколько согласованных по интерфейсу систем управления базами данных (СУБД). Наиболее общим требованием к системам управления базами данных являются: возможности многоаспектного поиска данных, иерархическая организация объектов хранения (главным образом текстов и словарных статей), возможности оперирования целыми базами данных (например, возможности их слияния), простота прикладного программирования, возможности телекоммуникационного доступа. Конкретные базы данных могут создаваться и существовать в различных организациях, пополняться за счет кооперативного сотрудничества с другими организациями и из центрального Фонда. Пополнение, в частном случае создание баз данных, должно быть возможным в двух аспектах: по номенклатуре хранимых объектов, например словарных статей, и по их атрибутам, например путем передачи из одной базы данных в другую определенных частей соответствующих словарных статей.

Наиболее общим понятием единицы хранения в Машинном фонде русского языка является понятие хранимого объекта, характеризуемого типом объекта, его именем (не обязательно единственным) и иерархически организованными атрибутами, каждый из которых имеет определенное значение, возможно, в некоторой конкретной базе данных в конкретный момент времени — пустое.

Типичным примером объекта является тип «словарная статья». Каждая словарная статья имеет имя — ее заголовок и атрибуты: грамматические формы, сфера употребления (возможно, несколько), стилистические пометы, толкования, сочетаемостные свойства и т. д. Каждый такой атрибут состоит, возможно, из нескольких атрибутов меньшего ранга, но каждый терминальный атрибут в этой иерархии имеет определенное значение: атрибут «грамматические формы», например, имеет значение «неизменяемый» или «тип парадигмы» и «тип исключения»; атрибут «сфера употребления» может иметь либо «пустое значение», т. е. сфера употребления не определена, либо значение одной или нескольких сфер употребления по списку, например «общеупотребительное», «авиация и машиностроение» и т. д.

Тип хранимого объекта определяет тип базы данных. Все базы данных, содержащие словарные статьи, принадлежат к словарному типу; базы данных, содержащие тексты, принадлежат к текстовому типу; базы данных, содержащие программы, принадлежат к программному типу и т. д. Число типов баз данных, естественно, не ограниченно, но и не может быть слишком большим — их всего порядка десяти.

Базы данных одного типа, содержащие объекты с логически совместимыми типами атрибутов, образуют естественное семейство. Базы данных, принадлежащие одному семейству, мо-

гут находиться под управлением одной логической[3] или физической системы управления базами данных. Такие базы данных можно считать информационно однородными.

Взяв за основу принцип информационной однородности баз данных, мы теперь можем выделить в первом приближении естественные компоненты Машинного фонда русского языка.

2.3.1. Первым таким компонентом Фонда является Генеральный словник русских словарей, который может быть создан как семейство однородных баз данных на основе Сводного словника, созданного в Словарном секторе Института русского языка АН СССР, и других «словнико-подобных» словарей, таких, как «Грамматический словарь русского языка» А. А. Зализняка, Орфографический словарь и др.

Объектом хранения в этой базе данных является вокабула, т. е. слово, являющееся потенциальным заголовком какой-либо словарной статьи и именем определенного лексического значения, а его атрибутами — номера или идентификаторы значений, а значениями атрибутов — имена и входы баз данных, в которых данное слово или его формы зафиксированы в качестве значений атрибутов каких-либо объектов — словарных статей, текстов, грамматических правил, статистических сводок, научных статей и т. п.

Очевидно, что многоаспектный поиск по заданным значениям атрибутов с соответствующими логическими условиями может давать ответы на такие запросы, как: «Найти перечень словарных статей определенных словарей, для которых имеется информация в терминологической базе данных», «Найти значения вокабулы, стилистически маркированные в словарях», «Найти все глаголы с дефектными парадигмами и сопоставить структуры их словарных статей» и т. п.

2.3.2. Другим типом семейства словарных баз данных является Терминологический фонд, в котором могут быть выявлены совместимые подсемейства отраслевых тезаурусов, терминологических ГОСТов, многоязычных автоматических словарей и автоматических энциклопедических общих и отраслевых словарей и энциклопедий. Эти подсемейства объединяются такими атрибутами, как «многоязычные соответствия», «отрасль знаний» («сфера употребления»), «тезаурусные отношения», «на-

[3] Мы говорим об одной логической системе управления базами данных, если конкретные базы находятся на разных вычислительных установках, каждая из которых оснащена одной и той же системой управления базами данных. Если же базы данных размещаются на одной вычислительной установке, оснащенной одной определенной системой управления базами данных, и находятся под ее управлением, то мы говорим об одной физической системе управления базами данных. В случае если одна вычислительная установка оснащена несколькими разными СУБД и базы данных одного семейства находятся под управлением разных СУБД, они, как правило, непосредственно несовместимы, и даже несопоставимы, так же как базы данных, размещенные на разных вычислительных установках под управлением разных СУБД.

учно-техническая дефиниция». Работы в области автоматизации ведения терминологии достаточно далеко продвинуты. Сейчас реально находятся на машинных носителях сравнительно большие материалы по отраслевым тезаурусам, терминологическим ГОСТам и многоязычным словарям. Объединение этого материала в одной логической базе данных могло бы обеспечить непосредственное сопоставление этого материала как внутри семейства, так и с семейством баз данных Академического словарно-грамматического фонда.

2.3.3. Академический словарно-грамматический фонд только предстоит заложить. По нашему мнению, он должен состоять из нескольких подфондов, число которых не следует заранее ограничивать. Одним из таких подфондов станет свод академических словарей — наиважнейших источников зафиксированных на сегодняшний день знаний о русском языке. Такой свод должен образовать однородное семейство баз данных, в пределах которого каждый из словарей продолжал бы свое существование в качестве отдельной базы данных и в то же время был бы связан средствами СУБД и номенклатурой своих составляющих со всеми другими словарями. Этот объединенный словарь мог бы не только выполнять справочные функции, но и быть источником нового систематического изучения лексики.

В советской лексикографии утвердилось учение о системе словарей как о такой совокупности, которая позволяет описать лексическую систему языка (словарный состав языка) в ее полном объеме. Для русской лексикографии эту систему словарей должны составить: словарь современного русского литературного языка, исторический словарь русского языка XIX в., исторический словарь русского языка XVIII в., словарь русских народных говоров, словарь древнерусского языка XI—XVII вв. Эту систему должны дополнять: словарь синонимов, фразеологический словарь, словарь антонимов, словообразовательный словарь и другие типы специальных лексикографических изданий. В совокупности эти словари должны выполнить задачу, которую ставил перед своим словарем-тезаурусом А. А. Шахматов[4].

Понятие системы словарей возникло и разрабатывалось в связи со стремлением сохранить гуманитарную традицию русских словарей, ставящую естественный предел допустимым объемам словарных статей и их сложности, что приводит к многообразию типов словарей, каждый со своим составом лексикона, схемами словарных статей и рубриками лексикографического описания. В теории баз данных этим понятиям соответствуют понятия схемы базы и ее подсхем, поэтому в автоматической лексикографии, опирающейся на концепцию базы данных, понятие типа словаря преобразуется в понятие режима эксплуатации сводного словаря, в составе которого каждый его компонент

[4] *Сороколетов Ф. П.* Традиции русской советской лексикографии.— ВЯ, 1978, № 3, с. 40—41.

«виден» пользователю таким, каким нужно, но в то же время связан внутренними, «незримыми» связями со всеми другими словарями и компонентами соответствующих словарных статей. Любой словарь, размещенный соответствующим образом в базе данных, предстает перед пользователем то как полный словарь, то как сокращенный — путем задания ограничений на доступ к его лексике как по номенклатуре объектов, так и по номенклатуре их атрибутов и/или их значений. Любой словарь чисто автоматически может быть выдан из базы данных в алфавитной форме, как прямой, так и обратной, либо в порядке словообразовательных гнезд, либо в тематическом порядке. Аналогичным образом могут быть прослежены и объективированы в виде словарных статей синонимические и антонимические связи, извлечены из текстов словарных статей фразеологические обороты и т. д. В многоязычной и двуязычной лексикографии при надлежащем конструировании словарных статей многоязычный словарь может быть разложен на $n(n-1)$ двуязычных словарей, где n — число языков, и, наоборот,— несколько двуязычных словарей с входами из разных языков и выходом на один язык могут быть соединены в один автоматический многоязычный словарь и порождают таким образом $A_n^k - 2n$ новых k-язычных словарей ($k \geqslant 2$) [5]. Эта идея принадлежащая И. И. Убину [6], реализована в автоматическом многоязычном словаре МУЛЬТИЛЕКС, созданном в Лаборатории по применению вычислительных средств в гуманитарных науках НИВЦ МГУ по заказу ВЦП ГКНТ и АН СССР. Почти по каждому из лексикографических параметров возможно объединение, сопоставление или извлечение данных из одного или группы автоматических словарей, хранимых в одном семействе баз данных. Это позволяет перевести теорию лингвистического конструирования Ю. Н. Караулова [7] в алгоритмический план.

Естественным расширением словарного академического фонда явятся историко-этимологический, диалектологический, топонимический и другие фонды, скомплектованные из соответствующих источников. Общее для них то, что их данные могут быть сгруппированы в словарноподобные объекты, наделенные, разумеется, своими наборами атрибутов, лишь частично сопоставимые с атрибутами академических словарей.

Подбор таких атрибутов и их увязывание в обобщенную структуру словарной статьи должны соответствовать требованиям всестороннего описания лексики, т. е. содержать грамматические, лексикологические, семасиологические, стилистические,

---

[5] $A_n^k = n(n-1)\ldots(n-k+1)$.

[6] *Убин И. И.* Общие принципы построения и использования автоматического словаря MULTILEX.— В кн.: Вычислительная лингвистика. Теоретические аспекты. Вопросы автоматизации лексикографических работ/Под ред. В. З. Демьянкова. М., 1982, с. 120—131.

[7] *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. М., 1981, с. 75—77.

фонетические, диалектологические, исторические, энциклопедические, библиографические и другие параметры.

Поэтому представляется целесообразным непосредственно связать собственно лексикографические базы данных с базой данных Академической грамматики, рассматриваемой в лексикографическом аспекте как систематизированное по способу описания изложение типов морфологической, синтаксической и семантической информации словарных статей, соотнесенное с определенными лексическими группировками. Одновременно грамматический фонд должен сохранять и свой абстрактно-категориальный характер описания фактов языка.

Представляется, что на первом этапе целесообразно разместить в базе данных словарный и предметный указатели Грамматики так, чтобы каждый вход адресовал соответствующие извлечения из текста, которые могли бы рассматриваться как грамматические справки к соответствующим местам словарных статей наряду с историко-этимологическими, диалектологическими и другими, например энциклопедическими, справками.

2.3.4. Наряду со словарными данными Машинный фонд русского языка целесообразно снабдить фондом текстовых источников, создав аналог картотеки цитат. Такой фонд не только мог бы служить источником систематического пополнения словарных и грамматических данных, но и для получения необходимых статистических оценок употребительности и совместной встречаемости лексических и грамматических явлений, статистического изучения свойств связного текста и накопления иллюстративного материала для будущих словарей. Иллюстрационно-текстовой фонд должен состоять, по нашему мнению, из нескольких подфондов, среди которых должен быть подфонд полных авторских текстов крупных мастеров слова («образцовые» тексты), подфонд статистических выборок разных жанров, стилей, временны́х интервалов и т. д., подфонд цитат на заданные лингвистические явления, а также подфонды речений, пословиц, поговорок, заголовков, надписей и т. п. Уже сейчас на машинных носителях находится значительное количество текстов как современного русского языка, так и русского языка XV—XVI вв. Основные работы по подготовке текстов на машинных носителях были выполнены на кафедре структурной и прикладной лингвистики Минского государственного педагогического института иностранных языков и на кафедре математической лингвистики Ленинградского университета.

Размещение текстов в базе данных одновременно образует словоуказатель к ним, т. е. фактически создается готовый конкорданс и частотный словарь. С пользовательской точки зрения вид такого словоуказателя-конкорданса — это режим его сканирования: по алфавиту, по обратному алфавиту, по убыванию или по возрастанию частот, по убыванию или по возрастанию значения датировки текста и т. д.

2.3.5. Адресуемые Генеральным словником иллюстрационно-

текстовой, терминологический и академический словарно-грамматический фонды могут служить источниками для формирования лингвостатистической базы фонда, содержащей исходные статистические данные для вычисления употребительности лингвистических явлений, статистических связей между явлениями и статистических характеристик текстов. В общем случае такими данными являются лексические спектры полных текстов, суммы частот, а также суммы их квадратов, кубов и четвертых степеней в различных совокупностях текстов и другие численные оценки, из которых могут быть построены оценки употребительности и эмпирические распределения, необходимые для систематического получения статистических выводов.

2.3.6. В значительной степени работа по извлечению из текстов, словарей и грамматик необходимой информации может быть автоматизирована уже накопленным к настоящему времени и созданным в рамках Машинного фонда русского языка программным обеспечением. Это обеспечение может быть разделено на два класса: программные комплексы, реализующие автоматический морфологический, синтаксический и/или семантический анализ текстов, и программы различного назначения, свободно комбинируемые в целях формирования программных комплексов какого-либо специального назначения. Первый класс мы называем языковыми процессорами, или, точнее, процессорами русского языка, второй — прикладными лингвистическими программами. Четкую границу между ними провести трудно, но в общем случае можно утверждать, что «настоящие» языковые процессоры наряду с программами, реализующими собственно лингвистический анализ или синтез, включают в себя достаточно полные формальные грамматики и словари с встроенными в них или эксплицитно сформулированными и записанными на определенном формальном языке лингвистическими алгоритмами. Таким образом, языковые процессоры — это комплексы программ и данных, реализующие определенные модели лингвистического разбора и/или конструирования текстов, а в ряде случаев — и модели понимания и/или порождения. В рамках своего назначения и тематической области, охватываемой словарем, языковые процессоры являются машинными фондами русского языка в собственном и самом узком смысле слова, ибо «машинизируют» все основные структурные свойства языка, хотя оставляют в стороне исторические, социальные, психологические, культуроведческие, эстетические аспекты его изучения и рассмотрения. Однако именно языковые процессоры являются с точки зрения приложений и продуктом Машинного фонда, и его внутренними инструментами автоматического лингвистического анализа.

Сегодня в области конструирования языковых процессоров накоплен значительный опыт, а в ряде случаев инженерно реализованы наиболее глубокие теоретические модели автоматического анализа и синтеза слов, словосочетаний, предложений,

связного текста и диалога. Ряд работ, опубликованных в последние годы, показывает, насколько более высокие требования к описанию и интерпретации языковых фактов предъявляет автоматическая обработка по сравнению с обычными лексикографическими и грамматическими описаниями, ориентированными на человека. Суть отличия лингвистического обеспечения автоматизированных систем от «лингвистического обеспечения общечеловеческой деятельности» (если можно так выразиться) скорее состоит не в строго формальном построении первого и в неформальном, гуманитарно ориентированном построении второго, а в более тонкой дифференциации в системе фиксируемых грамматических и лексических знаний, средств связи слов и выражений на всех уровнях языковой структуры, в более тонкой согласованности параметров разных уровней. В ряде словарей, разработанных в качестве лингвистического обеспечения языковых процессоров, число параметров описания слова в каждом из разделов словарной статьи может в десятки раз превосходить число параметров обычных словарей. Эти разработки открывают новые перспективы и для общей лексикографии, но вряд ли могут быть реализованы в ней в условиях ручного труда лексикографов, без поддержки средств автоматизации и в традиционной полиграфической форме. В связи с развитием микрокомпьютерной техники, с появлением персональных ЭВМ словари будущего имеют возможность полностью перейти в компьютерную форму со скрытой от пользователя реальной сложностью описания лингвистического явления. Именно в фонде языковых процессоров мы видим прообраз лингвистических автоматов будущего. Это не только собственно словари, т. е. справочные базы данных, но это также и инструменты анализа и синтеза текстов, выполняющие разнообразные функции — от орфографического контроля текста и стенографирования речи до ее перевода в устную или письменную форму на том же самом или ином языке. Этим и определяется ядерная, центральная роль языковых процессоров в Машинном фонде русского языка.

2.3.7. Все программы более частного применения, такие, как формирователи словоуказателей, частотных словарей, конкордансов, сегментаторы текстов, формирователи карт, текстовые редакторы, корректоры и т. п., образуют Фонд лингвистических алгоритмов и программ, в котором включенные в него программы могут быть объединены в пакеты программ и программные комплексы. С их помощью могут быть автоматизированы самые различные процедуры обработки лингвистических данных.

2.3.8. Систематическое и технологически организованное применение таких процедур и языковых процессоров к текстовым и словарным источникам Фонда может быть оформлено в виде автоматизированной лексикографической системы, предназначенной для автоматизации самого массового и трудоемкого вида лингвистической деятельности — создания словарей. Такая си-

стема должна включать средства первичного накопления данных — текстов и словарей, их редактирования и воспроизведения в полиграфическом формате, средства оперирования словарными статьями — членения их на сегменты, размещения сегментов в базе данных, средства доступа к словарным статьям и текстам и сборки их из сегментов во внешнее представление, средства коррекции словарных статей и их создания на основе словарных фрагментов и текстовых данных; средства различных сортировок, объединений, подсчета статистик; средства лемматизации текстовых форм и другие средства лингвистического анализа. Минимальная версия такой системы разработана в Лаборатории по применению вычислительных средств в гуманитарных науках НИВЦ МГУ. Система УНИЛЕКС-ЕС[8] позволяет накапливать как тексты, так и словари, сегментирует их, размещает в базе данных, имеет средства генерации частотных и обратных словарей, связывает словарь и конкорданс, имеет морфологический анализатор, опирающийся на «Грамматический словарь» А. А. Зализняка. В настоящее время в системе размещен этот словарь, тексты русской разговорной речи объемом 400 тыс. словоупотреблений и текстовой материал по двуязычной лексикографии, частотным словарям и одно- и двуязычным конкордансам. В состав системы входит также новая версия многоязычного автоматического словаря МУЛЬТИЛЕКС, в котором хранится сейчас материал объемом более 60 тыс. словарных статей. На этом материале отрабатывается методика пополнения словаря и его привязки к текстовым данным и языковым процессорам. Процедуры системы функционируют как в пакетном, так и в диалоговом режиме.

На нынешнем этапе в автоматической лексикографии существует, на наш взгляд, три сложнейших проблемы, без решения которых немыслимо дальнейшее существенное продвижение в направлении создания крупных автоматических словарей. Первая проблема состоит в накоплении и поддержании в актуальном состоянии больших текстовых и словарных источников, насчитывающих миллионы и десятки миллионов словоупотреблений. Такие источники не могут создаваться и переноситься на машинные носители в одном месте, обычно они создаются различными коллективами в различных местах в течение длительного промежутка времени.

Это требует единой системы их кодифицирования, членения, разметки, отслеживания их прохождения через все этапы обработки, их учета и комплектации и т. д. В первом приближении эта проблема была решена в процессе создания системы УНИЛЕКС и закреплена в «Методике первичной обработки

---

[8] *Андрющенко В. М.* Автоматизированная лексикографическая система UNILEX: (Основные проектные решения).— В кн.: Вычислительная лингвистика: Теоретические аспекты. Вопросы автоматизации лексикографических работ/Под ред. В. З. Демьянкова. М., 1982, с. 104—119.

лингвистических данных в интерактивном режиме»[9]. В этой методике предлагается единая система абстрактного членения совокупностей на пакеты, тексты, модули, сегменты, поля и элементы с соответствующей их позиционной нумерацией и выносом содержательного членения в отдельные справочные таблицы; предлагается единый язык разметки на деревья составляющих, основанный на модифицированной польской инверсии записи; разработаны управляющие записи к модулям вводимой информации и программы их обработки. Последовательное применение этой методики позволяет единым программным обеспечением собирать разные по происхождению источники ввода в единый входной формат системы УНИЛЕКС.

Вторая проблема связана с разработкой процедур группового контроля соблюдения заданного формата словарных статей и непосредственного сопоставления форматов словарных статей для лексики разных групп. Эта проблема хорошо известна и в традиционной лексикографии. В одной из своих работ Н. Ю. Шведова писала: «Для того чтобы избежать разнобоя в словаре, лексикограф должен работать не с алфавитным списком слов, а с определенными их лексическими группировками внутри отдельных частей речи. Только на этом пути могут быть достигнуты единство в разграничении значений, однотипность толкований, непротиворечивые приемы подачи фразеологизмов и примеров, последовательность и единообразие стилистических характеристик. Иными словами, словарь станет не просто квалифицированно составленной коллекцией слов, за которой стоит лексикограф со своим знанием языка и со своим пониманием отдельного слова, а также научным произведением, которое опирается на специально для него разработанные эталоны описания слов, так или иначе объединенных формально и семантически»[10]. В плане автоматизации эта проблема частично решается введением формального языка для задания структуры словарной статьи. Наличие такого языка позволяет на входе и на выходе системы контролировать соблюдение формата с какой угодно степенью подробности, а использование системы управления базами данных, допускающей многоаспектный поиск, обеспечивает выполнение процедур, дающих возможность групповой подготовки словарных статей, связанных определенной группой признаков.

Третья проблема связана с построением специальных лексикографических процессоров. В идеале хотелось бы, чтобы такой процессор, получив на входе формализованный проект словаря и доступ к базам данных, содержащим источники для формиро-

[9] *Андрющенко В. М.* Методика подготовки и первичной обработки лингвистических данных в интерактивном режиме.— Препр. № 138, 139 Ин-та русского языка АН СССР. М., 1981.

[10] *Шведова Н. Ю.* Однотомный толковый словарь (специфика жанра и некоторые перспективы дальнейшей работы).— В кн.: Русский язык: Проблемы художественной речи. Лексикология и лексикография. М., 1981, с. 171.

вания словаря, формировал новую базу данных, содержащую в качестве объектов структурно связанные фрагменты словарных статей, предусматриваемых проектом. Конечно, сейчас нельзя ожидать от такого процессора, чтобы он писал толкования значений слов или квалифицированно подбирал иллюстрации. Достаточно потребовать, чтобы он в нужных местах словарной статьи приводил возможные варианты толкований и/или фрагменты конкорданса в качестве потенциальных иллюстраций или материала для формирования толкований и создавал бы нужные формальные признаки, такие, как грамматические пометы. Остальная работа могла бы быть проделана лексикографом вручную за терминалом после предварительного анализа распечатки подготавливаемой совокупности словарных статей. В качестве составной части такой лексикографический процессор должен включать в себя общеязыковой процессор и средства лексикографической интерпретации текстов и словарных статей. Работа по созданию такого процессора должна быть, по нашему мнению, выполнена в рамках Машинного фонда русского языка.

2.3.9. Проект Машинного фонда русского языка должен предусматривать также разработку Информационно-справочного фонда, который в свою очередь может состоять из нескольких информационно-поисковых систем. Одной из таких систем должна быть информационно-поисковая система литературы по языкознанию. В отличие от обычных информационно-поисковых систем дескрипторного типа, такая система может быть снабжена не только понятийно-рубрикационным тезаурусом, но и предметно-справочным аппаратом, позволяющим получать информацию об интерпретации в публикациях конкретных лингвистических явлений — конкретных слов, морфем, словосочетаний, конструкций, грамматических форм и связей и т. п. Близким по назначению к такой системе является и автоматизированный фонд лингвистических анкет, часто используемых в социолингвистических, психолингвистических и других исследованиях. Эта система может быть построена по тому же принципу, что и другие автоматизированные системы обработки анкетной информации, применяемые в социологических и исторических исследованиях, проводимых с помощью ЭВМ. Наконец, в рамках самого Машинного фонда русского языка могут быть созданы внутренние документы в виде программной документации, проектов словарей, проспектов работ и др. Для хранения и распространения такой документации должен быть создан автоматизированный документальный фонд. Такой фонд должен обладать также средствами автоматизации труда по составлению документов и подготовке научных трудов от подбора материалов для них до выдачи издательского оригинала.

Наконец, Машинный фонд русского языка, состоящий из многих систем и подсистем, должен иметь свой информатор фонда, т. е. информационно-справочную систему, позволяющую пользователю ориентироваться в этой информационной системе и снаб-

женную определенными обучающими функциями. С помощью информатора фонда пользователь должен иметь возможность получать оперативные справки об источниках фонда, о работах, проводимых в рамках фонда, о способах использования программного обеспечения и данных и др.

3

Как уже говорилось, Машинный фонд русского языка мыслится как система семейств баз данных, каждое из которых управляется одной или несколькими системами управления базами данных. Связь между различными базами данных должна осуществляться одной или несколькими мониторными системами автоматизации лингвистических исследований. Каждая из мониторных систем в зависимости от свойств операционной системы и свойств СУБД должна обладать по крайней мере следующими функциями: запуска пользовательских задач в пакетном режиме в качестве своих подзадач, задач операционной системы или задач СУБД; организации диалогового взаимодействия с пользователем, запуска его задач и обеспечения диалога пользователя со своей задачей; интерпретации языка пользователя, формирования и передачи в СУБД описаний пользовательских баз данных и запрошенных пользователем функций обработки; передачи результатов работы во внешние файлы или в определенные пользователем базы данных; анализа процесса диалога пользователя с мониторной системой и генерации недоопределенных пользователями функций обработки, например, выход на получение справок от информатора фонда.

Этот язык пользователя первого уровня должен быть по возможности прост, в функциональной части он должен содержать лишь имена обобщенных функций обработки, таких, как *ввод/вывод, сортировка/слияние, поиск, редактирование, печать, уничтожение, выполнение, определение, помощь* и т. п. В объектной части монитор должен позволять определить простыми средствами формат ввода и вывода информации, специфицировать конкретную программу, приспособленную к данным форматам, обработку различных типовых условий завершения задачи. Элементарной реконструкцией такого языка является директива. С помощью директив могут формироваться процедуры, которые, в свою очередь, могут запускаться функцией *выполнение.* Преобразования таких типовых процедур в конкретные запросы к базе данных должны осуществляться монитором путем передачи параметров в вызываемые программы и формирования необходимых управляющих блоков для СУБД.

Функции защиты информации от несанкционированного доступа, разграничения доступа во времени, ведения статистики обращений, диагностики ошибок обработки, управления памятью и другие сервисные функции остаются в ведении конкретных систем управления базами данных.

Пользовательские языки второго уровня, предназначенные для более тонкого программирования лингвистических работ, например для записи формальных грамматик, словарной информации в словарях языковых процессоров, лингвистических алгоритмов, должны разрабатываться в рамках конструирования языковых процессоров и использоваться в их диалоговых режимах.

Разделение пользовательского языка на два уровня позволяет выделить простейшие функции обработки, сделать их доступными массовому пользователю-лингвисту и создать простую систему автоматизации лексикографических работ путем применения типовых программных средств. Например, для подготовки словаря писателя в его минимальной форме (система статистических словоуказателей, конкорданс, если необходимо, справочник рифм) достаточно определить базу данных, содержащую источники ввода, базу данных, предназначенную для размещения результатов, и типовую процедуру переработки источника в результаты. Такая типовая процедура может содержать следующую последовательность шагов:

1. Получение из базы данных-источника упорядоченной последовательности текстов данного писателя и формирование промежуточного набора данных A.

2. Если набор данных A имеет стандартную разметку, то шаг 3, иначе произвести разметку и создать набор данных A_1.

3. Определить новую текстовую базу данных и разместить A или A_1 в этой базе данных, определив в качестве ключа каждое текстовое слово каждого контекста, заданного разграничителем определенного уровня.

4. Отсканировать базу данных и определить дополнительные ключи в форме словарных заголовков и, если нужно, рифмующихся окончаний.

5. Отсканировать базу данных столько раз, сколько разных упорядочений словника необходимо; в процессе упорядочения произвести подсчёт статистики и выдать на печать результаты или, если необходимо, образовать нужное число выходных наборов данных.

6. Если необходимо, разместить полученные выходные наборы данных в словарной базе данных писателя.

Таким же образом можно, получив словарь другого писателя, разместить его в уже созданной базе данных и слить полученные словари, сохранив при этом возможность их раздельного воспроизведения.

Аналогичным образом могут быть подготовлены словоуказатели к памятникам и простые частотные одноязычные или двуязычные словари и конкордансы.

Программное обеспечение, развитое для содержания в базе данных многоязычных словарей, может с небольшими модификациями и дополнениями быть использовано, например, для создания в пакетном или диалоговом режиме ассоциативных словарей. В этом случае можно считать, что исходный список A_1

образует словник одного языка, а словник реакций A_2 — словник второго языка; аналогично, словник реакций на A_2—A_3 образует словник третьего языка и т. д.

Словники связываются прямыми и обратными ссылками, так что становится возможным проследить не только все пути от первичного слова-стимула ко всем словам-реакциям, но и обратные пути от каждого слова конечного словника к первичному слову-стимулу (возможно, не единственному).

Такого рода работы, производимые пользователями, с помощью программного обеспечения и на материалах, накопленных в Фонде, могут включаться в Фонд в качестве его компонентов и, таким образом, обеспечивать рост Фонда.

4

С технологической точки зрения Машинный фонд русского языка на данном этапе можно рассматривать как совокупность вычислительных установок и средств подготовки и воспроизведения данных, находящихся в распоряжении коллективов, ведущих работы в интересах Фонда. Такими установками могут быть вычислительные комплексы ЕС и СМ ЭВМ, в некоторых случаях машина БЭСМ-6. Для каждой пары «операционная система — система управления базами данных» должна быть создана своя мониторная система, т. е. минимально три версии мониторной системы, каждая из которых должна иметь столько интерфейсов по форматам данных, сколько необходимо для передачи данных из одного монитора в любой другой. При согласованном конструировании мониторов возможно добиться единого интерфейса на базе КОИ-7, единого коммуникативного формата на магнитной ленте и соглашений о системе единой управляющей информации, сопровождающей каждый модуль передаваемых данных. Минимально каждая вычислительная установка должна иметь четыре дисковода типа ЕС-5061 (29 Мбайт), из которых один предназначается для операционной и мониторной систем, один — для входной, один — для выходной баз данных и один — для размещения рабочих наборов данных и два накопителя на магнитной ленте. На одном пакете дисков емкостью 29 Мбайтов может быть размещена текстовая база данных объемом 0,5—0,7 миллиона словоупотреблений или словарь объемом 60—80 тыс. словарных статей.

В зависимости от объема оперативной памяти процессора вычислительной установки, количества терминалов и качества программного обеспечения на базе данных такого объема может быть обеспечена одновременная работа от 1 до 16 пользователей-лингвистов.

Можно назвать несколько разновидностей работ, которые могли бы выполнять пользователи-лингвисты непосредственно на вычислительной установке: ввод, разметка и редактирование данных при поддержке тестирующих программ и автоматизированного редактирования; запуск стандартных программ обра-

ботки источников и получения словоуказателей, конкордансов, проектов словарных статей и т. п.; запуск стандартных программ сортировки и группировки данных по определяемым лингвистом признакам с попутным подсчетом статистики; конструирование формальных грамматик и словарей с попутной проверкой их качества путем применения их для анализа и синтеза текстов; получение различных справок из накопленных в базах данных лингвистических источников; исправление и дополнение источников в версии, находящейся в индивидуальном ведении пользователя, и др.

5

С точки зрения вычислительной лингвистики Машинный фонд русского языка является инструментальной системой для производства языковых процессоров и средств автоматизации лингвистических исследований, снабженной информационным фондом источников данных о русском языке во всем объеме этого понятия. Как инструментальный комплекс для производства программ и их лингвистического обеспечения Машинный фонд должен в дополнение к перечисленным выше средствам предоставлять в распоряжение лингвистов-специалистов в области вычислительной лингвистики средства для определения формальных языков, средства для конструирования интерпретирующих и транслирующих программ для этих языков и определения для них баз данных, средства для конструирования мониторов, организующих обработку данных. Разработка таких средств должна опираться на глубокие теоретические исследования, приводящие к построению адекватных моделей русского языка и его достаточно представительных подмножеств, к построению адекватных моделей лингвистической деятельности и моделей понимания текста.

* * *

Изложенный проект создания Машинного фонда русского языка направлен на решение следующих основных 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;

— создать систему комплексной автоматизации лингвистического труда: составления словарей, поиска и обработки научной информации, анализа текстов, проведения классификационных работ, подготовки аппарата изданий и т. п.;

— заложить понимаемую в современном смысле сокровищницу данных о русском языке во всем объеме этого понятия, во всех его временных и территориальных формах.

Мы считаем, что переход к новым методам сбора, хранения, анализа и сопоставления данных о языке, новые методы созда-

ния 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 лингвистической традиции и новых задач практики нужно осуществлять на путях новой информационной технологии, развиваемой системами обработки данных на естественном языке в интеллектуальной среде человеко-машинного общения.

Из изложенного следует, что достигнуть заметных успехов на пути создания Машинного фонда русского языка мы можем, лишь соединив усилия многих коллективов, специализирующихся сейчас в решении различных задач, каждая из которых может быть рассмотрена как задача формирования или задача обеспечения функционирования Фонда. Объединение тематики этих коллективов единой научно-технической программой создания Машинного фонда русского языка создаст условия для реализации изложенного проекта без значительных новых материальных затрат, только за счет усилий по координации и согласованию содержания, этапов и форм представления результатов отдельных работ. Однако фактически приступить к созданию Фонда мы можем, не ожидая утверждения такой программы государственными органами, путем простой коллегиальной координации работ — взаимного информирования, принятия совместных технологических решений, взаимопомощи во внедрении и эксплуатации современных систем управления базами данных, передачи конкретных данных и программ.

К СОЗДАНИЮ МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА
(определение, применения, актуальные проблемы)

Б. Ю. ГОРОДЕЦКИЙ

О понятии «Машинный фонд русского языка»

Мне хотелось бы всячески поддержать идею А. П. Ершова и Ю. Н. Караулова о методологической значимости Машинного фонда русского языка для лингвистических и — шире — гуманитарных исследований[1]. Эта мысль при ее более широкой трактовке далеко не нова: Машинный фонд лишь поднимает идею на более высокую (и в этом смысле принципиально новую) практическую ступень. Создание Машинного фонда легко вписывается

[1] *Ершов А. П.* К методологии построения диалоговых систем: феномен деловой прозы.— В кн.: Вопросы кибернетики: Общение с ЭВМ на естественном языке. М., 1982.

в контекст той деятельности, которую Ю. Н. Караулов называет «лексикографической параметризацией языка»[2]. По-иному ее можно назвать *лингвистической инвентаризацией языка*[3]. И эта научно-практическая деятельность такая же древняя, как и сама лингвистика. Достаточно вспомнить грамматику Панини с ее результатами системной инвентаризации языка. Именно такого типа работу мы и хотим поднять на качественно новый уровень с помощью вычислительной машины. Методологическая идея здесь состоит в том, что подлинно научное лингвистическое описание должно быть основано на исчерпывающем инвентаре лингвистических фактов определенного рода. ЭВМ позволяет строить инвентари в безграничных масштабах и тем самым в гораздо более обозримые сроки, чем раньше, создавать для лингвистики достоверную эмпирическую базу, которая до сих пор, по существу, отсутствовала.

Под *лингвистическим инвентарем* я имею в виду определенным образом упорядоченный перечень всех лингвистических объектов определенного класса, зафиксированных с помощью некоторого метаязыка с заданной точностью и эксплицитностью. Отсюда ясны и основные вопросы, связанные с построением таких инвентарей: это и разработка метаязыка, и методов исчерпывающего построения инвентаря, и принципов упорядочения объектов и т. д. (Заметим, что вместо термина «инвентарь» во многих случаях удобно употреблять термин «словарь».) На инвентарь надо наложить такие требования, как однородность его состава, полнота представления, равномерность описания, различительная сила метаязыка. Среди инвентарей мы различаем формальные (включающие элементы плана выражения) и семантические (включающие элементы плана содержания). Среди формальных — фонологические, морфонологические, лексемные и т. д. Среди семантических — лексико-семантические, морфосемантические, ориентированные на парадигматику или на синтагматику и т. д. Казалось бы, для русского языка в области фонологии и морфонологии с инвентарями дело обстоит неплохо. Но, как я понял из доклада Л. В. Бондарко, прочитанного недавно на симпозиуме в Ленинграде, и с ними для русского языка положение на сегодня весьма и весьма неудовлетворительное. Что же говорить о семантических инвентарях и моделях, которые должны быть на них основаны? По сей день мы имеем дело с таким научным описанием семантики русского языка (и с соответствующим отражением его в учебниках), которое зиждется во многих своих частях лишь на отдельных примерах.

---

[2] *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. М., 1981.

[3] *Городецкий Б. Ю.* О принципах инвентаризации семантики языка и подъязыка.— В кн.: Теоретические и экспериментальные исследования в области структурной и прикладной лингвистики. М., 1973.

Понятие «Машинный фонд русского языка» может иметь более узкое и более широкое толкование. В узком смысле его можно определить как динамическую (гибкую) и развивающуюся (пополняемую и углубляемую) систему инвентаризации русского языка, включающую машинное представление основных типов лингвистических данных и богатый набор программ (предусматривающих в перспективе диалоговое общение с банком данных). Именно такое понимание укладывается в русло упомянутой концепции лингвистической инвентаризации языка. И тогда функции подобной системы состоят в том, чтобы хранить лингвистические инвентари, автоматизировать процедуры составления новых инвентарей на базе первичных и автоматизировать процедуры использования инвентарей в тех или иных исследовательских или прикладных целях. Заметим, что инвентари могут создаваться не только в виде списков, но и в виде исчислений.

В широком смысле Машинный фонд, как указывалось, в частности, в докладе В. М. Андрющенко,— это уже более сильная система, связанная прежде всего с задачами прикладной лингвистики и включающая, кроме этой инвентарной части, еще принципиально иные компоненты — стандартные алгоритмы и программы, которые целенаправленно моделируют некую человеческую деятельность. Здесь нужно четко различать два направления этого моделирования: во-первых, моделирование тех или иных аспектов речевой деятельности носителей языка, прежде всего анализа (понимания) и синтеза (вербализации); во-вторых, моделирование деятельности лингвиста-ученого. К последнему направлению относятся, например, модели выделения тех или иных единиц языка на основе корпуса текстов, модели определения дистрибуции, модели классификации стилистических явлений, алгоритмы построения семантического пространства в лексике языка, в морфемике языка и т. д. Таким образом, алгоритмы и программы Машинного фонда в широком смысле воплощают в себе лингвистические модели речевой или научной деятельности, оперирующей языковыми данными.

Первичными источниками Машинного фонда в узком смысле могут быть три основных ресурса данных: а) существующие тексты (устные и письменные); б) существующие словари и грамматики; в) результаты психолингвистических и социолингвистических экспериментов. Эти три источника покрывают всю эмпирию, с которой имеет дело лингвист (если отвлечься от того обстоятельства, что в эту эмпирию входит весь материальный и психический мир постольку, поскольку он отражается в языке).

Машинный фонд целесообразно строить на концептуальной базе модели структурного типа в собственном смысле этого слова, т. е. модели, использующей базовые понятия единиц и отношений. В зависимости от исследовательской цели могут применяться и модели других типов (анализирующие, синтезирующие, порождающие). Организация и представление семантической

информации должны основываться на достаточно общей модели плана содержания. Главные типы семантических словарей — тезаурусы и «синтагмарии». Семантический метаязык будет, как правило, предусматривать компонентный анализ гибкого типа. В необходимых случаях полезны семантические сети, средства ситуативной семантики.

Если представить себе обобщенную словарную статью Машинного фонда, то в ней должны найти отражение парадигматические и синтагматические свойства единиц, их морфо-семантическая структура, денотативная, сигнификативная, экспрессивная, синтаксическая семантика, полисемия. При каждой единице необходимо давать спектр ее функционирования по подъязыкам (об этом см. ниже). Должны указываться продуктивность и другие количественные данные. В перспективе можно ввести сведения, отражающие логические свойства слов, их информативность в семантике текстов, их роль в процессе понимания текста.

О применениях Машинного фонда русского языка

Круг применений Машинного фонда русского языка определяется его методологическим статусом, очерченным выше: лингвистические инвентари — это обязательная основа построения как описательных, так и прикладных моделей языка и речевой деятельности. Именно такая основа во многом обеспечивает реальность, адекватность и эффективность моделей. Здесь уместно еще раз обратить внимание на такую черту проектируемого Машинного фонда, как его гибкость, динамичность: благодаря этому инвентаризация может непосредственно перерастать в обобщенное моделирование, так как богатый набор программ позволяет получать и новые знания на базе инвентарей языковых объектов. Общенаучную роль тщательно составленных инвентарей Л. В. Щерба предвидел еще в 1923 г., анализируя одно из пушкинских стихотворений: «Я чувствую, конечно, все несовершенство этих моих „опытов“; однако полагаю, что путь в них найден правильный — путь лингвистический, путь разыскания значений: слов, оборотов, ударений, ритмов и тому подобных языковых элементов, путь создания словаря, или, точнее, инвентаря, выразительных средств русского литературного языка... Я буду счастлив, если самые мои ошибки поведут к занятию этим вопросом и к разработке в широких размерах русского литературного языка как выразительного средства...»[4].

Внедрение ЭВМ в процедуры лингвистической инвентаризации языка целесообразно везде, где мы имеем дело с достаточно большими массивами текстов или с достаточно большими инвен-

[4] *Щерба Л. В.* Опыты лингвистического толкования стихотворений, 1. «Воспоминание» Пушкина.— Руская речь, 1. Пг., 1923, с. 15—16.

тарями языковых единиц. Машинная инвентаризация языка стимулирует разработку новых путей выявления лингвистической информации, дает новые способы представления информации и новые способы ее использования.

Существует два типа использования Машинного фонда русского языка: а) построение вторичных инвентарей на базе первичных; б) непосредственное решение научных и практических задач. Круг этих задач, естественно, широк и не может быть очерчен исчерпывающим образом. Укажем некоторые из них, отличающиеся, на наш взгляд, особой актуальностью.

К задачам фундаментального характера относится прежде всего упоминавшееся выше обоснование лексикографической деятельности: Машинный фонд позволит поставить на практические рельсы конкретное моделирование лексико-семантической, словообразовательной и прочих структур русского языка в действительно широком масштабе. Может быть начата планомерная работа по созданию тезаурусов и синтагматик русского языка. Фундаментальное значение имеет также привлечение Машинного фонда в типологических исследованиях языков мира, в частности в контрастивной лингвистике, занимающейся выявлением специфики русского языка на фоне того или другого из иностранных языков или языков народов СССР.

Для прикладных задач единый Машинный фонд русского языка важен прежде всего в плане обеспечения единообразия в разработке автоматизированных систем обработки текста (АСОТ) различных типов — в составе АСУ, АСНТИ, САПР, АСПР и др. Особенно существенно такое единообразие в практике создания информационных и других искусственных языков. На этом пути следует искать и решение проблем совместимости уже имеющихся автоматизированных прикладных систем.

Следует отметить, что в некоторой точке фундаментальные и прикладные задачи сходятся. И такой точкой являются терминологические банки данных для отдельных подъязыков русского языка. Такие банки должны рассматриваться как составные части Машинного фонда русского языка. Без них невозможно обеспечить адекватное представление о русском языке как системе подсистем. С другой стороны, динамическая система терминологической инвентаризации подъязыков — это необходимый инструмент автоматического и полуавтоматического построения рабочих компонентов информационных систем (как информационных тезаурусов, так и словарей других типов). В свете сказанного ясно, что создатели Машинного фонда русского языка должны учесть словари, уже имеющиеся на машинных носителях (например, терминологические словари в органах научно-технической информации). Кроме этого, необходимо определить круг словарей, существующих на традиционных носителях (книги, отчеты, картотеки) и отвечающих требованиям представительности (для соответствующих подъязыков) и научной качественности.

К конкретным прикладным задачам, которые выиграли бы от создания Машинного фонда русского языка, можно также причислить машинный перевод, составление учебных пособий по русскому языку, создание статистических описаний русской речи для систем передачи информации по каналам связи и многие другие (см. также ниже).

Покажем на одном примере, как даже простой инвентарь с простым программным обеспечением может давать пищу для важных выводов о семантической структуре языка. Возьмем в качестве такого примера ту часть Машинного фонда русского языка, которая уже существует и опубликована в книге Ю. Н. Караулова[5]. Если обратиться к учебникам по лексикологии и лексикографии, то в разделе о толковых словарях мы не найдем практически ничего о метаязыке толкований, о том, как он конкретно устроен в словарях русского языка, о том, как его надо совершенствовать. Инвентарь, содержащийся в книге Ю. Н. Караулова, будучи первым опытом анализа толкований с помощью ЭВМ, дает в этом плане уже многое.

С методологической точки зрения эта работа поучительна в двух отношениях. Во-первых, она показывает, что как только некий инвентарь (в данном случае исходный для Ю. Н. Караулова первичный инвентарь — 11 тыс. дефиниций) попадает в машину, появляется возможность быстро автоматически получить принципиально новый продукт — вторичный инвентарь (который и опубликован в упомянутой книге). Массив дефиниций С. И. Ожегова был в руках языковедов и до упомянутой работы, но получение о б о б щ е н и й по нему вручную было бы делом крайне трудоемким (тем более это относится к обработке более объемных толковых словарей). Следовательно, введение дефиниций в ЭВМ знаменует собой качественно новый этап в лексикографической работе. В результате некоторого огрубления дефиниций и обработки их на ЭВМ был получен инвентарь семантических множителей в количестве 8415 единиц — слов, которые используются в толкованиях (они представлены в виде квазиоснов). Каждый множитель снабжен индексом — показателем его дефиниционной активности (аналогом частоты). Этот вторичный инвентарь может быть полезен для десятков исследовательских задач. Он используется самим Ю. Н. Карауловым для измерения семантической связи между словами при построении тезауруса. Далее, он может быть использован при создании новых словарей, новых дефиниций, при отборе словарей-минимумов для иностранцев (из него легко порождается 30 тыс. слов). Мы знаем теперь, что наиболее частыми семантическими компонентами, используемыми в дефинициях русского языка, являются следующие: ‘двигать’, ‘часть’, ‘время’, ‘место’, ‘действие’, ‘растение’, ‘животное’, ‘общество’, ‘земля’, ‘направление’,

[5] *Караулов Ю. Н.* Частотный словарь семантических множителей русского языка. М., 1980.

‘работа’, ‘жить’, ‘человек’, ‘поверхность’, ‘сторона’, ‘производство’, ‘вещество’, ‘количество’ и другие[6]. Знаменательно, что полученное упорядочение понятий находит и интуитивное оправдание: здесь мы имеем названия основных типов действий, основных категорий времени и пространства, основных объектов внешнего мира и т. д.

Во-вторых, рассматриваемая работа удобна для иллюстрации того положения, что эмпирически обоснованный вторичный инвентарь позволяет легко получать из него инвентари третьего порядка, ориентированные на те или иные исследовательские задачи. Так, на основе имеющегося инвентаря мы можем получать (с помощью несложных арифметических операций) любые списки укрупненных (более абстрактных) семантических компонентов. Скажем, выясняется, что приставка *пере-* имеет суммарную дефиниционную активность, равную 624, а приставка *до-* — 342. Таким образом, мы узнаем, что один из этих компонентов в дефиниционном отношении почти в 2 раза активнее, чем другой (кстати, «квазипродуктивность» этих префиксов, т. е. число множителей, в состав которых они входят, равна соответственно 161 и 33). Далее, если взять некоторый синонимический ряд, например ‘желать’, ‘хотеть’, ‘стремиться’, ‘мечтать’, ‘жаждать’, ‘вожделеть’, ‘желание’, ‘воля’, ‘стремление’, ‘долгожданный’ и т. д., то легко узнать, что его вес в толкованиях составляет 152. Семантическое поле цветообозначений имеет суммарную дефиниционную активность 176; его члены по употребительности в дефинициях упорядочиваются следующим образом (по степени убывания дефиниционной активности): ‘белый’, ‘красный’, ‘серый’, ‘черный’, ‘зеленый’, ‘желтый’, ‘синий’, ‘голубой’. Оранжевый, коричневый и фиолетовый цвета вообще не встречаются в обследованных дефинициях. Подобным образом, имея простое программное обслуживание, мы можем строить модель семантического пространства русского языка. Этот метод позволяет сравнивать разные толковые словари, вырабатывать оптимальные (с той или иной точки зрения) списки семантических компонентов и т. д. Имеются пути повышения эффективности и точности данного метода.

Об актуальных проблемах, связанных с созданием Машинного фонда

Создание Машинного фонда русского языка — широкомасштабная задача, требующая основательной теоретической подготовки проекта. Укажем некоторые актуальные проблемы в этой области (одна из них будет рассмотрена подробно):

соотношение инвентаризации подъязыков и языка в целом;

основные направления составления лингвистических инвентарей, их виды. (Для одних задач нужны смысловые группы лек-

[6] *Городецкий Б. Ю.* [Рец:] Караулов Ю. Н. Частотный словарь семантических множителей русского языка.— ВЯ, 1983, № 6.

сем, для других — их сочетаемостные ограничения, для третьих — списки алломорфов, для четвертых — сочетаемость фонем и т. д. Предстоит разработать стройную систему направлений лингвистической инвентаризации русского языка. Применительно к инвентаризации плана содержания этот вопрос предварительно рассмотрен нами в статье «О принципах инвентаризации семантики языка и подъязыка», где в качестве основных направлений выделены лексико-семантическая и морфо-семантическая инвентаризация, каждая с подразделением на парадигматический и синтагматический аспекты);

объекты хранения в Машинном фонде. (Чаще всего говорят о словах, но это не единственная важная единица русского языка. Особенно остро вопрос о единицах стоит в области терминологии, это вопрос о слове и словосочетании, о морфемах и квазиморфемах, об основах и квазиосновах и т. д. Кроме языковых единиц, объектами хранения могут быть языковые отношения, а также комплексы единиц и целые тексты);

виды словарной информации о единицах Машинного фонда и оптимальная структура словарной статьи;

типы количественных показателей языковых единиц;

специфика фундаментальных и проблемно-ориентированных словарей. (Имеется в виду зависимость словаря от степени и характера привязки к практической задаче);

соотношение данных и процедур, взаимопереходы между ними. (При одном подходе некую информацию можно записать как данные — это базовая информация. При другом подходе ее можно получить с помощью процедур — это производная информация. Языковеды должны по-новому взглянуть на свои научные потребности и разработать новые решения указанной проблемы. Это позволит понять, какие виды сведений в различных лингвистических задачах могут быть оставлены в скрытом виде и при необходимости выводиться из банка с помощью программных средств);

диалоговое общение с Машинным фондом;

введение Машинного фонда русского языка. (Речь идет об изменениях состава и структуры Фонда, вызываемых, во-первых, расширением его диапазона и углублением его функций, а во-вторых, необходимостью оперативно отражать ход развития языка. Кстати, отсюда один шаг до управления этим развитием).

Остановимся подробнее на проблеме соотношения инвентаризации подъязыков и языка в целом. Значимость этой проблемы обусловлена, во-первых, тем принципиальным обстоятельством, что оптимальный путь описания русского языка (оптимальный как в методологическом, так и в методическом плане) должен мыслиться как инвентаризация языка через подъязыки. Последнее понятие в широком смысле охватывает и тематические подмножества русского языка, и функционально-стилистические подмножества, и подмножества, обслуживающие ту или иную конкретную сферу общения, и т. д. Еще Л. В. Щер-

ба отмечал, что «грамматически мы имеем единый язык в довольно широких группировках; в области же словаря для единства языка должно быть более полное единство материала, а потому мы видим, что с точки зрения словаря язык дробится на очень маленькие ячейки вплоть до семьи»[7]. Но, с другой стороны, Л. В. Щерба признавал и норму. Он говорил, что «хороший нормативный словарь не придумывает нормы, а описывает ту, которая существует в языке, и уж ни в коем случае не должен ломать эту последнюю. Может быть, норму трудно иногда подметить; но это уже несчастье исследователя (разрядка наша.— *Б. Г.*) и не имеет никакого отношения к принципиальной стороне дела»[8]. Действительно, «несчастье исследователя» и было, и есть велико, потому что возможности обозрения и анализа бескрайнего океана речевых произведений до сих пор были весьма и весьма ограниченны. Создание машинных банков языковых данных начинает несколько менять ситуацию. Теперь есть надежда, идя от отдельных подъязыков, систематически изучить специфику каждого из них и нащупать эту неуловимую норму или, точнее, нормы. Именно на этом пути разрешима, видимо, ситуация «Гулливер и лилипуты», упомянутая в докладе А. П. Ершова.

Значимость подъязыкового подхода к инвентаризации обусловлена, во-вторых, тем, что для большинства прикладных задач подъязык фиксируется в самом начале и лишь затем определяется диапазон и мощность создаваемого лингвистического обеспечения[9]. Впрочем, и для собственно описательной лингвистики этот принцип, думается, немаловажен: ведь особенности функционирования языка прежде всего проявляются в каких-то конкретных сферах.

Задать подъязык практически, сформировать объективную экспериментальную базу его изучения — задача отнюдь не простая. Сформулируем некоторые основные принципы исследования подъязыков, вытекающие как из теоретических соображений, так и из опыта нашей работы с несколькими подъязыками в пределах русской деловой прозы. Подъязык — это некоторая подструктура, способная функционировать как язык, т. е. полностью определять строение некоторой совокупности речевых произведений — текстов. Путь исследования подъязыка включает три важнейшие стадии: внешнее задание подъязыка, выявление его внутреннего состава как системы элементов, сопоставление с другими подъязыками (в ходе которого может обнаруживаться и полное совпадение элементов). С внешней стороны подъ-

---

[7] *Щерба Л. В.* О трояком аспекте языковых явлений и об эксперименте в языкознании.— В кн.: Щерба Л. В. Языковая система и речевая деятельность. Л., 1974, с. 28.

[8] *Щерба Л. В.* Опыт общей теории лексикографии.— Там же, с. 26.

[9] *Городецкий Б. Ю.* Семантические проблемы построения автоматизированных систем обработки текстовой информации.— В кн.: Вычислительная лингвистика. М., 1976.

язык естественно задавать через множество текстов, которые образуют представительный экспериментальный массив — эмпирическую базу, достаточную для познания закономерностей подъязыка (или, в общем случае,— комплекса подъязыков). Формирование экспериментального массива требует значительного времени для серьезной научной проработки и предварительного описания исходного множества текстов, доступных наблюдению, что видно хотя бы на примере составления частотного словаря американского варианта английского языка Г. Кучеры и Н. Фрэнсиса: в этом случае два года ушло на подготовку корпуса текстов, на продумывание его структуры и один год — собственно на получение словаря [10]. В группе прикладной семантики кафедры структурной и прикладной лингвистики филологического факультета МГУ было сформировано четыре представительных экспериментальных массива, отражающих комплексы подъязыков русского языка: комплекс подъязыков, обслуживающих высшие звенья руководства процессом коммунистического строительства в нашей стране; комплекс подъязыков, обслуживающих центральное отраслевое звено управления геологической отраслью; многоотраслевой комплекс подъязыков, определяющих строение заголовков неопубликуемых научно-технических отчетов по одиннадцати отраслям; комплекс подъязыков, участвующих в порождении политематических научно-технических рефератов по цветной металлургии. Эти работы показывают, что формирование массива на научной основе означает тщательное моделирование структуры текстовых информационных потоков в соответствующей области [11].

К важнейшим принципам формирования представительного массива относятся следующие: выборка речевых произведений не должна быть полностью случайной; в массив включаются не произвольные сегменты, а целые тексты; в массиве сохраняется набор существенных признаков текстов; обеспечивается количественная представленность признаков; учитываются прогнозы относительно тенденций развития подъязыка; массив строится как структурированное множество текстов, из которого могут автоматически получаться различные системы подмассивов с учетом тех или иных важных признаков.

---

[10] *Gorodeckij B.* Review of: Henry Kučera, W. Nelson Francis. Computational Analysis of Present-Day American English.— Linguistics, 1972, N 84; *Francis W. N.* Problems of Assembling and Computerizing Large Corpora.— In: Empirische Textwissenschaft. Aufbau und Auswerkung von Text-Corpora. Königstein, 1979.

[11] *Городецкий Б. Ю.* Лексико-статистическая инвентаризация комплекса подъязыков.— В кн.: Проблемы теоретической и экспериментальной лингвистики. М., 1977; *Городецкий Б. Ю., Зевахина Т. С.* Метод моделирования структуры текстовых информационных потоков.— В кн.: Всесоюз. совещ. по проблемам автоматизации информационных процессов в области общественных наук: (Метод. материалы). М., 1979.

Наши исследования показывают, что существует восемь основных групп признаков, необходимых и достаточных для адекватного формирования представительного массива текстов. Иначе говоря, подъязык (или, в общем случае, комплекс подъязыков) задается с внешней стороны допустимыми значениями признаков текстов. Эти группы признаков таковы:

1. Источники (авторы) текстов (например, министр, начальник главка, тот или иной научно-технический журнал, ведущий ученый, тот или иной научно-исследовательский институт и т. д.).

2. Хронологические рамки появления текста (например, десятая пятилетка, 1983 г., первый квартал 1984 г., довоенный период и т. д.).

3. Жанр текста, или вид документа (например, постановление Совета Министров СССР, приказ по министерству, директивное письмо, научно-технический отчет, научная статья, реферат научной статьи, ГОСТ, патент, заголовок монографии и т. д.).

4. Функциональная роль текста (например, управление научно-техническим прогрессом, определение организационной структуры, информирование о решениях вышестоящих органов, стандартизация документации, стандартизация терминологии, определение задач профсоюзов, согласование деятельности разных ведомств, пропаганда передового опыта и т. д.).

5. Тематика текста (сюда входят, во-первых, элементарные тематические индексы типа «сельское хозяйство», «культура», «транспорт», «геологоразведочные работы», «финансирование», «социальное обеспечение», а во-вторых, комбинированные тематические индексы типа «сельское хозяйство и наука», «геологоразведочные работы, финансирование и международное сотрудничество»; если документу приписан элементарный индекс, то он является монотематическим, а если комбинированный — то политематическим).

6. Характер информации, выражаемой в тексте (например, преобладание качественной или количественной информации, модальность подачи информации — «законодательное установление», «ориентационная установка», «выдвижение гипотез», «критика и осуждение» и т. д.).

7. Объем текста (может измеряться числом словоупотреблений, термоупотреблений, числом предложений, абзацев).

8. Внешняя структура текста, или композиция (например, документы с эксплицитной рубрикацией или без рубрикации, с приложениями или без приложений, с преобладанием списков или развернутого текста и т. д.).

Конкретный набор признаков, фиксируемых при исследовании того или иного подъязыка, уточняется с учетом специфики последнего: уточняется как список признаков, так и принимаемых ими значений. Объем экспериментального массива текстов может быть любым. Это зависит от статистических соображений, технических возможностей и ряда прагматических факторов.

Но в любом случае он должен адекватно отражать состав реальных моделируемых информационных потоков.

Структура этих потоков гибко моделируется посредством выделения различных систем подмассивов. Система подмассивов — это разбиение множества текстов, образующих экспериментальный массив, на подмассивы на основе одного или с учетом двух и более признаков. Основной инструмент при этом — система элементарных подмассивов, получаемая с учетом всех значений всех признаков (два текста попадают в один и тот же элементарный подмассив тогда, и только тогда, когда они одинаково характеризуются по всем признакам). Остальные системы подмассивов (или изолированные подмассивы) получаются автоматически путем теоретико-множественных операций над элементарными подмассивами. Например, для комплекса подъязыков отраслевого управления (директивные документы Мингео СССР) на основе предварительного обследования исходного информационного потока (888 документов, 250 тыс. словоупотреблений) нами был сформирован и структурирован экспериментальный массив из 328 документов объемом 100 тыс. словоупотреблений, в котором оказалось 189 элементарных подмассивов. Последнее число наглядно говорит о том, насколько нетривиальна задача выделения подъязыков русского языка через представительные массивы текстов. Разумеется, процедура формирования экспериментальных массивов может с самого начала огрубляться: можно учитывать не все группы признаков, а, скажем, только источники, хронологию, жанр и тематику. Но, чтобы обеспечить сопоставимость данных, должна применяться единая стратегия для разных сфер функционирования русского языка.

Структурированный представительный массив текстов позволяет создавать любые интересующие нас инвентари языковых (в частности, семантических) единиц с точным указанием места и роли каждой единицы в системе подъязыков русского языка. Мы можем получать универсальные инвентари (общие для всех основных подъязыков, представленных в экспериментальном массиве) и специфические инвентари (показывающие особенности лексического состава отдельных подъязыков). Так, в исследовавшемся нами массиве рефератов по математике, физике, горному делу и информатике в универсальный словарь вошли такие слова, как *анализ, вариант, величина, внимание, время, замена, краткий* и др. Для комплекса подъязыков, обслуживающих центральное звено руководства основными сторонами жизни нашей страны, универсальный словарь включает слова *изменение, женщина, клуб, лаборатория, месяц, наличие, необходимый, система, учреждение* и др. (всего 400 слов). Из специфических словарей подмассивов мы узнаем, например, что слово *композитор* встречается в документах, посвященных вопросам культуры, идеологии, повышения уровня материальной и духовной жизни народа, работе партийных, комсомольских и

профсоюзных органов, но не встречается (в вероятностном смысле) в документах, связанных с управлением промышленностью, транспортом, сельским хозяйством, планово-финансовой и законодательной сферами государственной жизни. Слово *металл*, с другой стороны, ни разу не встретилось в подмассиве документов, посвященных повышению уровня жизни народа.

В общем случае мы можем говорить об исчислении лингвостатистических инвентарей, которые можно получать по структурированному экспериментальному массиву текстов [12]. Каждый такой инвентарь можно характеризовать по четырем параметрам: 1 — подмассивы, отражаемые в инвентаре; 2 — вид языковых единиц, взятых в качестве лингвостатистических объектов; 3 — степень полноты инвентаря; 4 — принцип упорядочения единиц. Теоретически говоря, исчисление всех разновидностей инвентарей даже для одного ограниченного комплекса подъязыков будет охватывать сотни тысяч возможных инвентарей. При создании Машинного фонда целесообразно строить для сформированного массива текстов сводный инвентарь языковых единиц (прежде всего слов), дающий при каждой единице полный спектр ее качественных и количественных характеристик [13]. При условии хранения этого сводного инвентаря в ЭВМ мы сможем легко и быстро получать любые его подмножества, упорядоченные нужным образом. Мы сможем также наводить необходимые справки в диалоговом режиме.

Стратегия инвентаризации языка через подъязыки выдвигает целый ряд специальных проблем, главная среди которых — это выявление множества структурно (внутренне) различных подъязыков и определение их соотношений. Более частные проблемы включают, например, следующие вопросы: методы оценки языковой близости текстовых подмассивов; квантитативно-тезаурусная характеристика словарей подмассивов; типы взаимодействия подъязыков в тексте (отсюда — уточнение понятия «комплекс подъязыков»); способы экстраполяции данных, полу-

---

[12] *Городецкий Б. Ю.* Методы семантического исследования ограниченного подъязыка. М., 1971. Ч. II; *Он же.* Принципы оптимизации лингвистического обеспечения автоматизированных информационных систем в области общественных наук.— В кн.: Всесоюз. совещ. по проблемам автоматизации информационных процессов в области общественных наук: (Метод. материалы); *Городецкий Б. Ю., Еленевская Л. Е.* К методике исследования репрезентативных словарей для политематической выборки.— В кн.: Количественные методы в гуманитарных науках. М., 1981.

[13] *Городецкий Б. Ю., Зевахина Т. С., Еленевская Л. Е., Лебедева О. С., Пожарискый И. Ф.* Методика семантико-статистического анализа Базового отраслевого тезауруса по геологии.— В кн.: Научно-техническая информация в геологии: Экспресс-информация. М., 1981. Вып. 5; *Городецкий Б. Ю., Зевахина Т. С.* К построению Терминологического банка данных как компонента лингвистического обеспечения информационной системы.— В кн.: Всесоюз. конф. «Основные направления развития и совершенствования работ по стандартизации научно-технической терминологии в XI пятилетке»: (Тез. докл.). М., 1983.

ченных по экспериментальному массиву; представление общего лексического ядра для комплекса подъязыков; установление степени однородности (монолитности) каждого данного подъязыка.

Последовательное осуществление намеченной стратегии позволит определять эмпирически обоснованный функционально-тематический облик русского языка с любой степенью подробности. Каждая единица русского языка будет снабжена спектром ее употребления по различным подъязыкам.

С точки зрения прикладной лингвистики адекватное моделирование подъязыков необходимо прежде всего для разработки реализуемых алгоритмов анализа и синтеза текста в различных автоматизированных системах: на этапах проектирования, создания, корректировки и развития любых информационных систем разработчики опираются, в частности, на характеристики соответствующих информационных потоков. Точное знание этих характеристик повышает обоснованность и ручных работ по рационализации письменного делового общения (стандартизация документов, упорядочение процессов именования, семантическая систематизация лексики, оптимизация способов сжатия информации). Все эти задачи лингвистического обеспечения АСУ, ИПС или САПР должны решаться на базе надежных словарей слов, словосочетаний и других единиц. Эта надежность словарей гарантируется представительностью экспериментального массива как в количественном (статистика), так и в качественном (структура) отношении. Представительный массив текстов и основанные на нем адекватные словари необходимы для изучения информационных потоков не только в русле лингвистики, но и в русле документоведения, информатики, теории управления, прогнозирования, программирования.

Возвращаясь к проблемам фундаментальной науки, необходимо в заключение подчеркнуть, что создание Машинного фонда русского языка требует фронтального пересмотра задач лексикологии под особым углом зрения: какого рода закономерности мы ищем и как мы можем эффективно использовать ЭВМ? Должны быть усилены исследования по теории лексикографии, по лингвистической семантике, по методам машинного представления и обработки лингвистических данных. Само собой разумеется, на соответствующем уровне должна вестись и подготовка кадров в вузах.

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА И СЛОВАРНОЕ ДЕЛО

Р. П. РОГОЖНИКОВА

В соответствии с толкованием слова «фонд» в толковых словарях русского языка Машинный фонд русского языка можно понимать как ресурсы, запасы единиц русского языка, введенные в память ЭВМ и извлекаемые оттуда по мере необходимости.

Машинный фонд можно рассматривать и как собственно фонд, хранилище разного рода данных, и как исследовательскую лабораторию, в которой на основе введенных в ЭВМ данных ведется изучение русского языка в различных аспектах.

Наиболее актуальным, хотя и трудоемким, является создание машинного фонда лексико-фразеологических единиц. Актуально это потому, что именно на основе собрания лексико-фразеологических материалов возможна организация словарной работы, создание больших, фундаментальных словарей русского языка, а на их основе — кратких общих и учебных словарей.

Именно словарная работа, особенно для больших словарных предприятий, таких, как большой словарь национального языка, требует наиболее полного охвата лексического состава русского языка, привлечения большого количества текстов произведений самых различных жанров, использования словарей как современного языка, так и предшествующих эпох, материалов русской разговорной речи и т. п.

С другой стороны, собрание лексико-фразеологических материалов русского языка дает возможность проводить разнообразные исследования лексики, ее семантических, грамматических, стилистических, словообразовательных и других особенностей. Поскольку слово — это основная единица языка, собрание лексики может служить источником самых разнообразных работ, словарных и несловарных.

Словарная работа обычно начинается с создания картотек. Создание картотеки — это сбор и подготовка языковых фактов к дальнейшей работе. Языковой факт представляет собой употребление слова в ограниченном контексте, извлеченное из какого-либо текста, где выявляется его словарная форма, грамматические, стилистические особенности, словесное окружение, сочетаемость и т. д. Взятые в совокупности, они представляют собой лексико-фразеологический состав русского языка.

Можно сказать, что у нас уже существует фонд русской лексики по источникам XIX—XX вв., правда, составленный вручную. Это Большая картотека Словарного сектора в Ленинграде. То же можно сказать и о картотеках исторических: Словаря русского языка XI—XVII вв., XI—XIV вв., XVIII в., представляющих фонд русской лексики по этим временным периодам.

Каждая из этих картотек используется не только для непосредственной словарной работы, но и для многих других работ, словарных и несловарных. Широта охвата материала, большое и разнообразное количество материалов на каждый языковой факт позволяет сделать это.

Так, в Большой картотеке Словарного сектора на материалах языка XIX—XX вв. постоянно работают не только штатные лексикографы — сотрудницы Словарного сектора, но и другие исследователи, в том числе и иностранные ученые. Здесь ведется работа по сбору материалов для кандидатских и докторских диссертаций, монографических исследований, сбор и проверка материалов для учебных пособий, каталогов, учебных словарей, для плановых исследований Института русского языка, Института русского языка им. А. С. Пушкина, университетов страны и других научных и учебных учреждений. Картотекой пользуются литературоведы для комментирования текстов, изучения языка и стиля писателей и для других целей.

Хотелось бы обратить внимание на те сложности, которые возникают при подготовке словарных картотек и которые следует учесть при создании Машинного фонда русского языка.

В материалах картотеки необходимо по возможности максимально точно отразить лексико-фразеологическую систему русского языка, представленную через отдельные языковые факты. Сложность этой работы состоит в том, что она связана с охватом большого материала, огромного количества лексико-фразеологических единиц, которые нужно представить в возможно большей полноте, т. е. отразить лексико-фразеологический состав вширь. Это можно сделать, привлекая большое количество разнообразных источников различных жанров.

С другой стороны, каждый факт языкового употребления должен быть обеспечен достаточным количеством материалов, чтобы выводы, которые будут делаться в отношении семантики слова, его сочетаемости, грамматических и других особенностей слова, не были случайными, т. е. необходимо отразить состояние лексико-фразеологической системы вглубь.

И полнота отражения лексико-фразеологических единиц, и обеспеченность их достаточным количеством материала позволят объективировать словники словарей и отразить каждое языковое явление в словаре с наибольшей точностью.

Исходным материалом на этом этапе являются письменные тексты, которые отбираются в зависимости от целей и задач словаря [1]. Они представляют во всем жанровом и стилистическом многообразии русский литературный язык. Отбор текстов для картотеки — серьезный и важный вопрос, которому необходимо уделять особое внимание.

---

[1] *Рогожникова Р. П.* Некоторые вопросы организации словарной картотеки.— В кн.: Русский язык: Источники для его изучения. М., 1980, с. 247.

Что касается источников картотеки словаря современного языка, то они должны быть значимыми и актуальными, в них должен быть представлен русский литературный язык во всех важнейших сферах его функционирования. Необходимо, чтобы они были достаточными в количественном отношении, поскольку важно выявить реальный лексический состав русского литературного языка и обеспечить каждый языковой факт необходимым количеством материала (контекстов). При выборе источников учитывается принадлежность к определенному стилю и жанру, тематика и языковые качества источника.

Таким образом, только большое и разнообразное количество источников отразит состояние литературного языка в данный временно́й период.

В настоящее время в словарной картотеке, на основе которой подготовлены такие фундаментальные словари, как семнадцати- и четырехтомный, а также составлен и составляется ряд других словарей (синонимический, фразеологический, вариантов и трудностей словоупотреблений, новые слова и значения, словарь русских народных говоров и др.), насчитывается более 7 млн. карточек-цитат. Это огромнейшее собрание лексико-фразеологических материалов.

Казалось бы, можно было перевести на магнитные ленты картотеки, созданные вручную, и это создало бы машинный фонд лексики русского языка. Однако в картотеках, составленных вручную, имеются, как правило, уязвимые места, связанные с тем, что пополнение картотек происходит неравномерно в разное время, основным принципом выборки является частичная выборка, ориентированная прежде всего на качество цитаты, т. е. на непосредственное использование ее в словарях. Не учитывается весь реальный лексический состав хотя бы какого-то специально отобранного круга источников. Все эти недостатки картотек объяснимы. Они связаны с тем, что обработка источников — очень трудоемкое дело — происходила, да и сейчас еще происходит вручную. Это не дает возможности учесть весь реальный лексический состав источников, не всегда имеется достаточное количество материалов на отдельный языковой факт.

Сейчас появилась возможность использовать для этих целей ЭВМ. Хранение и обработка на ЭВМ полностью всех текстов, необходимых для полного толкового словаря национального языка, и сейчас еще представляет определенные трудности из-за недостаточной памяти ЭВМ, из-за трудоемкости подготовки текста для ввода в ЭВМ. Тем не менее идея создания машинного фонда лексики представляется очень важной. На первых порах кажется целесообразным сочетание ручных и автоматизированных методов создания картотек (т. е. пополнение картотек, созданных вручную, данными, полученными с ЭВМ) и одновременно с этим разработка программ и подготовка текстов для машинного фонда лексики.

В настоящее время значительную долю времени, связанного с обработкой текстов на ЭВМ, занимает подготовка текстов для ввода в ЭВМ, перфорирование текстов. Пока перфорирование текстов ведется вручную, неизбежны ошибки, а это влечет за собой проверку, правильно ли отперфорирован текст, что тоже делается вручную. Таким образом, доля ручного труда при подготовке текста для ввода в ЭВМ остается пока еще значительной. Ведутся поиски того, как автоматизировать этот этап. При наличии терминальных устройств ввод в ЭВМ может быть произведен непосредственно с терминала, но и это не исключает ручного труда.

Для повышения эффективности обработки на ЭВМ текстов целесообразно использовать перфоленты, которые готовятся во многих типографиях для фотонаборных машин. Это экономит время подготовки информации для ЭВМ и позволяет ввести в ЭВМ текст, адекватный тексту источника. Текст, записанный на перфоленту в закодированном виде, вводится в память ЭВМ и редактируется.

Возможности редактирования машинного текста в режиме непосредственного доступа дают лексикографу дополнительные удобства при работе с источником. Он может оперативно менять издание, выбирая первое или наиболее авторитетное, или любое другое в зависимости от целей обработки текста. Например, если обрабатывается текст произведения XIX в., изданный в XX в., введя его один раз, можно работать с ним дважды, по изданию XIX и XX в.

Редактирование машинного текста сводится к расстановке страниц в соответствии с выбранным изданием и сверке текстов. После этого текст обрабатывается с помощью программы получения прямого указателя словоформ. Просмотр указателя словоформ с терминала позволяет выявить слова с ошибками, пропущенные при первой проверке. Наличие номера страницы и строки при этих словах дает возможность легко их найти в исходном тексте и исправить.

Языковые явления и факты поступают с ЭВМ в виде алфавитно-частотных указателей словоформ на листах, в виде конкордансов, алфавитно-частотных указателей на карточках и т. п.[2] В алфавитно-частотном указателе словоформ приведены все словоформы текста, с их вариантными, морфологическими и другими особенностями, даются количественные указания использования их в тексте. Все это очень важно для лексикологов, лексикографов. Например, в статье В. Г. Белинского «Взгляд на русскую литературу 1847 года»[3] фиксируются такие словоформы: благороднее (1) и благородней (1), болезнию (1) и болезнью (1), большей (3) и большею (10), вежливостию (1),

[2] *Рогожникова Р. П., Чернышева Л. В.* Об автоматизации в лексикографии.— В кн.: Алгоритм и системы автоматизации исследований и проектирования. М., 1980.

[3] Словоуказатель получен Е. В. Вертелем на ЭВМ ЕС-1035.

вежливостью (2), верностию (8), внутренно (1), всею (14), всей (2), вымысл (1), вырабатывать (1), вырабатываться (1)[4].

При ручной выборке все эти словоформы могут быть не зафиксированы, попадут в картотеку лишь некоторые из них. В то же время фиксация этих словоформ с количественными показателями дает материалы для исследований функционирования их в языке XIX в., сопоставления их с более поздними материалами.

На основе прямого указателя словоформ составляется словоуказатель текста, в котором все словоформы приведены в основной форме слова. Он необходим для сопоставления данного текста со словарями, картотеками и т. п., где за основу берется исходная форма слова.

Приведение словоформ к основной (словарной) форме слова, или лемматизация, невозможно автоматическим путем из-за большого количества фоно-морфологических особенностей лексики, представляющих национальные отличительные черты русского языка, например чередование в корне слова, аффиксах (*носить—ношу*, *разбить—разобью*). Кроме того, приведение к основной форме слова выявляет большое количество омоформ и омографов, которые могут быть заданы не иначе, как только списком, например: родительный падеж ед. числа существительного *бег—бегу* и форма 1-го лица ед. числа глагола *бежать—бегу*. Поэтому после составления указателя словоформ каждой словоформе в интерактивном режиме приписывается его словарная форма, а затем производится пересортировка в алфавитный порядок по словам и словарной форме.

Для автоматизации этого процесса необходимо создание списка соответствий, в который войдут все словоформы, соотносящиеся с основной словарной формой слова и не имеющие каких-либо особенностей (*делаю—делать*), омоформы и омографы (*бегу* от *бег* и *бежать*) и формы с чередованием в корне или аффиксах (*разобью—разбить*). Использование такого списка облегчит лемматизацию и сортировку слов при автоматической обработке текста.

Возможно также приписывание к каждой словоформе основной формы слова в исходном тексте[5], но это слишком трудоемкая работа, и она целесообразна лишь на небольших по объему текстах.

Задача составления списка омоформ и омографов представляется важной не только в узкопрактическом (для машинной обработки текстов), но и в теоретическом отношении, а также для обучения русскому языку[6]. В настоящее время работа по

[4] В скобках указана частота словоформы в тексте.

[5] *Koubourlis B. J.* From word-form concordance to a dictionary-from concordance.— In: Computers in the Humanities. Minneapolis, 1975, p. 225.

[6] *Виноградов В. В.* Об омонимии в смежных явлениях.— В кн.: Виноградов В. В. Избр. труды: Исследования по русской грамматике. М., 1975; *Мельникова А. И.* К вопросу о русских омографах.— РЯШ, 1974. № 4, с. 60.

выявлению и составлению списка омографов ведется в Картотеке Словарного сектора.

Словоформы, приведенные к основной форме слова, могут быть выданы как списком, так и на карточку вместе с количественными данными употребления словоформ и всего слова в тексте. Появляется возможность организовать картотеку-словоуказатель, являющуюся необходимой частью словарной картотеки, но отсутствующую в современных картотеках, поскольку создание ее является очень трудоемким делом.

Это по существу словник с приведением всех словоформ к основной форме слова и количественными данными их употребления в полностью обработанных текстах. Она представляется весьма важной с разных точек зрения: она дает представление о словнике по полностью обработанным произведениям. Здесь воспроизведен весь набор словоформ текстов, даются сведения о количественных показателях употребления словоформ и слов. Картотека-словоуказатель позволяет пополнять картотеку цитат в необходимых случаях по частям, дает возможность быстро найти нужную цитату в тексте, а также использовать количественные данные об употреблении слов и словоформ в текстах. Гораздо легче проводить сверку словника со словарями по картотеке-словоуказателю, поскольку здесь все словоформы, как и в словаре, приведены к основной форме слова.

АЛТАРЬ 5
1 АЛТАРЕЙ . . . . . .
69 (6)
1 АЛТАРЕМ . . . . . .
42 (7)
2 АЛТАРЬ . . . . . .
50 (33) 62 (26)
1 АЛТАРЯ . . . . . .
51 (16)
П. А. ВЯЗЕМСКИЙ ЛИРИКА
М., 1979

Получение карточек-цитат из произведений, введенных в ЭВМ,— следующий этап автоматизации. Контекст выделяется формальным путем: по количеству словоупотреблений, строк от заданного слова. В стихотворных произведениях — это строфа, которая обычно представляет законченный отрезок в смысловом отношении. На карточке имеются библиографические указания. Таким образом, лексикограф, исследователь получает по своему требованию любые цитатные материалы тех произведений, которые обработаны с помощью ЭВМ.

Обработка текстов с помощью ЭВМ предполагает создание программы для получения обратных указателей словоформ и обратных словоуказателей. Они могут быть использованы для выявления вариантных форм и слов и сбора материалов для них. Обратный указатель словоформ дает полную картину сло-

воизменения, которую невозможно найти в обычном обратном словаре, в основу которого положены исходные формы слова. В то же время обратный словоуказатель, составленный по исходным формам слов, может послужить основой для выделения и детального обследования определенных групп лексики, исходя из их словообразовательных признаков.

Составлялись программы для выявления двух-, трех-, четырехсловных сочетаний. Работа по выявлению таких словосочетаний способствовала появлению небольшого учебного «Словаря сочетаний, эквивалентных слову»[7]. Могут быть составлены и другие программы, по которым будут получены необходимые для исследователей материалы.

Полная обработка текстов дает возможность значительно полнее отразить в толковых словарях лексический состав русского языка, что окажет влияние и на другие стороны исследований по русскому языку. Так, в толковые словари по разным соображениям, в том числе и из-за недостаточного количества материалов, не включаются целые группы лексики (например, наречия на *-о*, *-е*, наречия с *по-* и др., слова с аффиксами субъективной оценки и т. п.). Это характерно не только для русской лексикографии. Отсутствие подобных групп лексики в обратном словаре испанского языка, составленном на основе семантических словарей, как отмечают исследователи, искажает количественные соотношения между разными морфологическими типами слов[8].

Работа по созданию машинного фонда лексики безусловно является трудоемкой, но она открывает новые возможности для словарной работы, поскольку обрабатываются полностью произведения определенного временно́го периода. Для этой цели необходимо обрабатывать такие материалы, которые могут быть использованы многократно для разных работ. Так, если взять произведения классической литературы XIX в. не только художественной, то они могут быть использованы и в словарях современного русского языка, и для создания словаря языка XIX в., необходимость которого ощущается, и для различного рода лингвистических и других исследований.

Другая важная часть словарной работы — отбор лексики в толковый словарь. Сверка словника картотеки со словниками словарей — важный этап работы при формировании словников словарей. В настоящее время в распоряжении лексикографов имеются различные словари, позволяющие обобщенно характеризовать объект исследования — слово. Однако пока еще нет сводного справочника, в котором бы нашел отражение хотя бы один вопрос: в каких из существующих словарей зафиксировано то или иное слово. Поскольку словники различных, в том

[7] *Рогожникова Р. П.* Об эквивалентах слова в русском языке.— ВЯ, 1977, № 5; *Она же.* Словарь сочетаний, эквивалентных слову. М., 1983.

[8] *Арутюнова Н. Д., Максимова О. С.* Размышления о пользе обратных словарей.— НДВШ. Филол. науки, 1968, № 6, с. 25.

числе и толковых, словарей русского языка не совпадают, представляется целесообразным создание справочника, в котором бы все словники наиболее значительных словарей были объединены. Таким справочником является Сводный словник словарей русского языка, составленный коллективом сотрудников Словарного сектора. Он представляет собой полный перечень слов, включенных в одиннадцать наиболее полных по охвату лексического состава русского языка словарей нашего времени.

В основу сводного словника словарей положены следующие наиболее значительные по охвату лексического материала словари: Толковый словарь русского языка под ред. Д. Н. Ушакова — в словнике обозначен У; Словарь современного русского литературного языка (в 17-ти т.) — Б; Словарь русского языка (в 4-х т.) — М; Словарь иностранных слов. М., 1964 — И; Новые слова и значения. Словарь-справочник по материалам прессы и литературы 60-х годов — НС; Словарь русского языка С. И. Ожегова, 9—11-е изд.— Ож; Орфографический словарь русского языка (1974) —Орф; Большая советская энциклопедия. 3-е изд.— БСЭ; Политехнический словарь. М., 1976 — П; Словарь иностранных слов, 7-е изд. М., 1978 — И-7; Словарь русского языка (в 4-х т.), 2-е изд.— М-2. В сводном словнике сейчас более 163 тыс. слов. (В самом большом по объему семнадцатитомном «Словаре современного русского литературного языка» несколько более 120 тыс. слов.) При каждом слове словника указываются словари, в которых зафиксировано слово. Они располагаются в следующем порядке: У Б М И НС Ож Орф БСЭ П И-7 М-2. «Словарная статья» Сводного словника имеет следующий вид:

жанр	У Б М И Ож Орф БСЭ И-7 М-2
живой	У Б М Ож Орф БСЭ П М-2
жизнеобеспечение	НС Ож Орф БСЭ П М-2

Это позволяет видеть, в каком из современных словарей впервые было зафиксировано то или иное слово. Кроме того, в Сводном словнике отмечаются варианты слов, имеющиеся в словарях, омонимы с кратким раскрытием значения, некоторые грамматические особенности. При отборе лексики в словари Сводный словник наряду с цитатными материалами явится объективным справочником, с помощью которого можно формировать словники различных по объему словарей, как общих, так и учебных. Разнообразие задач, которые могут решаться с помощью Сводного словника словарей, обусловливает его многократное использование в различных целях. Сводный словник словарей русского языка может стать основой для Генерального словника в машинном фонде лексики, создаваемом на основе словарей. С помощью Сводного словника можно проверить, зарегистрировано ли какое-либо слово в современных словарях и в каких именно. Это важно для лексикографов, ведущих наблюдение за лексикой, не зафиксированной до сих пор в сло-

варях, за лексикой, появляющейся на страницах газет, книг, журналов.

Сводный словник может быть сопоставлен с любым текстом, представленным в виде словника или алфавитного словаря. Необходимость этой работы для лексикографов очевидна. Это важно для проверки и уточнения семантики слова, других особенностей лексикографической обработки слова. Так, сопоставление Сводного словника с «Частотным словарем русского языка» обнаружило довольно большое количество лексических единиц, не включенных в словари русского языка. На букву «б», например, таких слов 117, в том числе *байбачество, балочки, басовито, батин, бедноватый, безбилетность, безвылазно, бехтеревка, богоравный, боровок, бритвочка* и др.

Необходимость такой работы при подготовке словарей трудно переоценить. Лексика, не вошедшая в толковые словари,— это и резерв их пополнения, и индивидуально-авторские и другие образования, которые не должны входить в общие толковые словари, но тем не менее представляют интерес в разных отношениях.

На основе имеющихся полностью обработанных текстов и частотных словарей произведений авторов XIX в. предлагается создание словаря редких слов, т. е. слов, не включенных в современные толковые и другие наиболее значительные словари русского языка, но употреблявшиеся в произведениях русской классической литературы [9]. Этот словарь может быть составлен на основе полностью обработанных на ЭВМ текстах произведений авторов XIX в. В настоящее время по имеющимся словоуказателям и полным частотным словарям авторов XIX в. выявлено около 3 тыс. таких слов. Такая работа важна для лексикографов, литературоведов. Она помогает увидеть жизнь слов. Так, в ежегодниках «Новое в русской лексике. Словарные материалы 1977, 1978, 1979 гг.» фиксируются не только новые, появившиеся в эти годы лексические единицы, но туда попадают и такие слова, которые употреблялись у авторов XIX в., но не были зафиксированы словарями. Например, слова: *будитель, догробный, несходящий, нечеловечий*, отмеченные в этих выпусках, имеются у П. А. Вяземского (*будитель*), Н. А. Некрасова (*догробный*), Л. Толстого (*несходящий*), Н. В. Гоголя (*нечеловечий*). Отсутствие их в словарях свидетельствует, что эти слова не входили в активный словарный запас, не являлись словами частого употребления.

Сводный словник словарей имеет и самостоятельное научное значение. Он помогает проанализировать состав словников словарей, на основе которых он создан. По Сводному словнику можно определить, какой лексикой пополнялись словари, будут пополняться в дальнейшем, какие слова исключались из словарей.

[9] *Рогожникова Р. П.* Редкие слова в произведениях авторов XIX в.— ВЯ, 1982, № 1.

С помощью сводного словника легко выявить наиболее употребительную лексику современного русского языка и отобрать ее в словари различных объемов. Это важно при составлении толковых словарей разного объема как общего, так и учебного типа, а также для отбора лексики в русско-иноязычные словари. В настоящее время у нас отсутствуют толковые словари на 10, 20, 30 тыс. слов, необходимые и для русских, и для тех, кто изучает русский язык на продвинутых этапах. Не очень ясны пока критерии отбора лексики в такие словари. Думается, что Сводный словник мог бы оказаться полезным в этом отношении.

Поскольку в Сводном словнике отражены варианты слов, отмеченные в словарях, а также омонимы, некоторые грамматические особенности слов, он может оказать существенную помощь в научных исследованиях и учебных занятиях по лексикологии. Он может служить базой для отбора лексики по различным семантическим, грамматическим и другим группам, для изучения омонимов, вариантов слов.

По типу Сводного словника словарей современного языка могут быть созданы словники словарей предшествующих эпох, терминологических словарей и т. п.

Возможности использования ЭВМ в лингвистических исследованиях с каждым годом расширяются в связи с усложнением функций современных вычислительных машин, увеличением объема памяти и быстродействия, оснащенностью дополнительными устройствами, облегчающими ввод и обработку исходного материала (текстов или словарей). Применение интерактивных методов позволяет исследователю-лингвисту использовать ЭВМ даже в тех случаях, когда какие-то языковые факты или явления трудно или невозможно формализовать. Занимаясь автоматизацией в лингвистике, важно видеть не только возможности ЭВМ, но и целесообразность ее использования. Желательно передать ЭВМ наиболее трудоемкие и часто повторяющиеся задачи, что позволит значительно сократить сроки выполнения различных разработок, связанных с исследованием языка, и в то же время откроет новые возможности его изучения.

ТИПЫ РУССКИХ ТЕКСТОВ И ОРГАНИЗАЦИЯ МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА

А. С. ГЕРД

Обсуждая проблемы создания единого Машинного фонда русского языка, целесообразно, с одной стороны, разделить разные типы фондов, а с другой — различные уровни самого обсуждения.

Машинный фонд русского языка — это большая автоматизированная макросистема, включающая в себя ряд подсистем

и микросистем. Одна из таких подсистем — фонды, ориентированные на фиксацию фактов речи; это подсистемы текстов, слов, словоформ, словосочетаний и предложений. Такие фонды относятся к подсистемам регистрирующего типа в отличие от систем, ориентированных уже на ту или иную интерпретацию речевых данных (фонды словарных статей, грамматик), на их исследование.

Ниже обсуждаются преимущественно подсистемы регистрирующего типа. Каковы цели и задачи такой подсистемы? Хранить тексты и выдавать по запросу потребителя отрывки текстов, точные адреса и статистику словоформ, полные словоуказатели или их фрагменты, цитаты на слово, частотные и обратные словари, в идеале сводить разные формы слова к одной, исходной.

Реальные тексты порождаются в речевых ситуациях, в тех или иных языковых состояниях, в которых реализуется система языка. Типы текстов-источников определяют и характер лексики будущего Машинного фонда.

Приведем ниже некоторые примеры таких речевых ситуаций, языковых состояний: язык художественной литературы, язык газеты, язык науки и техники, язык организационно-распорядительных документов, язык суда, язык школы, язык собраний, язык обиходных бесед, реплики в транспорте, магазинах, аптеках, ателье, поликлиниках, язык личных дневников и частных записок. Даже этот неоднородный и неполный перечень показывает, что почти все типы ситуаций потенциально могут быть реализованы как в устной, так и в письменной форме. Однако некоторые из них функционируют преимущественно в письменной форме, а устная форма здесь вторична или необязательна (язык художественной литературы, науки, законов и постановлений, деловых документов, писем, дневников), другие — преимущественно в устной форме (диспут на собрании, военные приказы, профессионально-производственный диалог, телефонные разговоры, обиходные беседы, реплики дома, на улице, в транспорте, магазинах, аптеках, ателье, радиопереклички метеорологов, диспетчеров аэродромов, железнодорожных станций).

Вряд ли у кого-нибудь вызывает сомнение включение в Фонд письменных текстов художественной литературы, науки, техники, законов, постановлений, деловых документов.

Здесь можно сделать лишь небольшие комментарии.

В современных научных, технических и деловых текстах основное ядро — термины. Однако необходимость ввода в Машинный фонд русского языка полных научных и технических текстов не бесспорна. Контекстом для термина, в отличие от обычных слов общего языка, является не столько данное непосредственное научное сочинение, сколько та терминосистема, то терминополе, в котором существует данный термин. Каждый термин имеет значение только в конкретной терминосистеме,

вне которой он уже перестает функционировать как термин данной науки, отрасли знания.

На первый взгляд терминология других наук по своему объему может затопить и поглотить весь Фонд, занять львиную долю его машинной памяти. Здесь нужен детально продуманный и взвешенный подход, хорошее деловое обсуждение в кругу терминологов. Но, может быть, важнее как раз то, что никакой другой тип лексики по характеру своих значений не расположен заранее в такой степени к автоматизации, как терминология. Лексика языков науки и техники — это не только слова-термины, но прежде всего термины-словосочетания, символы, идентификаторы, число которых растет с каждым днем, это профессиональные глаголы, условная символика, за которой скрываются научные понятия, концепции.

Наиболее остро в этой области сегодня стоит вопрос о создании отраслевых и межотраслевых автоматизированных терминологических банков данных (АТБД). В идеале АТБД — это достаточно глубокая, семантическая сильная система документально-фактографического типа. АТБД может быть организован как хранилище целостных отраслевых текстов или терминосистем. Это АТБД, ориентированный на текст на естественном языке. Такой АТБД сможет ответить на такие запросы, как: 1) выдать весь текст автора; 2) выдать все тексты (фрагменты) по определенной тематике; 3) выдать термины a, b, c, их значения; 4) выдать словник текста; 5) выдать частотный словарь; 6) выдать сведения о наличии рисунков и схем.

Выдавая сведения о научном или техническом термине, о всех его текстовых употреблениях или о месте в терминополе, такая система тем самым выдает данные и о самом понятии, ибо описание термина в научной работе — это и есть описание самого понятия. Банк терминов во многих случаях ведет к автоматизации создания и самих терминологических словарей.

Более сложный вопрос о вводе в Машинный фонд текстов — записей устной речи, разговорной речи, отражающих такие ситуации, как диспуты, ход судопроизводства, собрания, обиходные беседы, реплики, споры, телефонные разговоры, радиопереклички и т. п. В целом современная разговорная речь — явление весьма неоднородное. В случае наличия фондов таких записей на магнитофонной ленте они могли бы быть вовлечены в Машинный фонд русского языка. Такие фонды имеются в ИРЯ АН СССР, в Саратовском университете, в Университете дружбы народов им. П. Лумумбы, в Ленинградском университете.

Таковы в целом те виды текстов, порождаемые в синхронических языковых состояниях, которые требовали комментариев.

Не менее важно хронологическое расслоение текстов. Древнейший период XI—XIV вв. в истории русского языка оставил нам такие памятники, созданные на Руси, как Остромирово евангелие, Чудовская псалтырь, Изборники 1073 и 1076 гг., Моление Даниила Заточника, Повесть временных лет и др.

От XV—XVII вв. мы имеем уже огромное, вряд ли исчислимое количество отдельных источников деловой, повествовательной, конфессионально-повествовательной и собственно конфессиональной письменности. С другой стороны, ни одна славянская страна не сохранила такого количества самых разнообразных деловых памятников, как Россия. Это и акты, и челобитные, и хозяйственные описи, и приходно-расходные и таможенные книги древних монастырей, и записи иностранцев. Часть их издана, большая часть хранится в архивах Москвы, Ленинграда, Вологды, Архангельска и других городов. Лексика деловых памятников XV—XVII вв. во многом представляет нам разговорную речь Московской Руси во всем многообразии ее словарного состава.

В XV—XVII вв. в Москве, Пскове, Новгороде продолжается русское летописание, расцветает жанр хождений и путешествий. XV—XVI вв.— это золотой век русской повествовательной литературы. Московские, новгородские, тверские повести, полемические сочинения и послания древнерусских писателей, замечательные по языку жития святых, созданные в самых различных местах Руси, заложили основы книжной лексики русского литературного языка, основы его семантических микросистем. Таким образом, как сами памятники, так и язык памятников письменности XI—XVI вв.— это уникальный в славянском культурном мире фонд слов, образов, выражений, фразеологии, это золотой фонд нашей культуры.

При автоматической обработке этих текстов есть свои, преимущественно чисто филологические трудности, но, как показывает опыт кафедры математической лингвистики Ленинградского университета по обработке текстов XV—XVI вв., трудности эти при соответствующей подготовке и постановке дела в кругу историков языка и специалистов по математической лингвистике вполне преодолимы.

XVIII век принес в русский язык лексику новой русской литературы, русского классицизма и сентиментализма. Это язык поэзии Кантемира, Ломоносова, трагедий Сумарокова, новой драмы и поэзии. XVIII век дал нам первую русскую газету, язык писем и бумаг Петра Великого, косвенно — разговорный язык Петербурга XVIII в.

Такие гигантские фигуры в истории русской национальной мысли и культуры, как Державин, Радищев и Карамзин, знаменуют уже прямой переход к литературе и языку XIX в.

XVIII век — это и век рождения Российской Академии наук, и, главное для нас, языка русской науки. Ломоносов, Георги, Гмелиан, Лепехин, Паллас — вот далеко не полный перечень тех, кто стоял у истоков рождения русского научного текста и стиля. История Российской Академии наук — это история русской науки, история ее языка. В течение последних десятилетий в Ленинграде создана фундаментальная картотека Словаря русского языка XVIII в. Нужно ли при наличии такой карто-

теки вводить в Фонд тексты XVIII в.? Здесь следует подчеркнуть, что создание и выполнение традиционных академических тем, таких, как, например, Словарь русского языка XVIII в. и реализация новых тем типа создания Машинного фонда русского языка, не должны пересекаться, а без необходимости и соприкасаться. Создание Машинного фонда, как и всякая новая большая тема, требует своей новой организации источников, материала, которая ни в коей мере не должна затрагивать фонды, материалы и методы работы, сложившиеся ранее. Опыт всех больших новых тем показывает, что каждая такая тема требует новой организации исходных данных, своей стратегии разработки.

В то же время принципиальная ориентация Машинного фонда на полноту адресных отсылок по каждой словоформе, контексту не снимает в перспективе вопроса о русских текстах XVIII в. как одном из его источников. Более того, в случае исключения из Фонда данных по XVIII в. он, выдавая, например, справки по XI—XVII вв. и, далее, по текстам XIX—XX вв., не сможет выдавать ответы по лексике XVIII в., и тем самым в большой фундаментальной системе окажется потерянным одно из важных звеньев в истории русского языка.

С XIX в. мы входим в новую русскую литературу, а с 30—40-х годов XIX в. и в эпоху современного русского языка.

В связи с рассмотрением хронологической стратификации русских текстов и лексики целесообразно остановиться на диалектных текстах как источнике Машинного фонда русского языка. Диалект функционирует в синхронии, но по сути своей он одновременно принадлежит и диахронии. Диалект древнее любого национального языка и тем более языка литературного. Лексика современных диалектов — явление неоднородное. С одной стороны, она таит в себе целые пласты древнейших славянских слов и выражений, восходящих к донациональной, дописьменной эпохе, а с другой стороны,— немало заимствований из других языков, поздних локальных новообразований, слов, связанных с взаимодействием со своим литературным языком. Русская диалектная лексика неоценима как исторический источник, как источник по истории славянских языков, по истории и этимологии слов, как источник по истории русского народа, по исторической географии и истории этноса. Народная речь — это поистине живая сокровищница представлений о духовном мире русского народа, о его верованиях, обрядах, о материальной культуре русских на протяжении ряда веков. И в этом отношении картотеки русской диалектной лексики — это своеобразные музеи материальной и духовной культуры русского народа.

Как исторические картотеки русской лексики, так и картотеки областной лексики — это наше культурно-историческое национальное достояние, это фонды, которые сродни сокровищам Эрмитажа, Музея антропологии и этнографии им. Петра Вели-

кого, рукописным отделам Библиотеки им. В. И. Ленина или Публичной библиотеки им. М. Е. Салтыкова-Щедрина.

Наконец, современная диалектная лексика — это часть лексики современной русской разговорной речи, это одна из синхронических форм ее реализации, и если мы думаем, что Машинный фонд русского языка должен как-то учитывать современную разговорную речь, то вопрос о соотношении с современной диалектной речью не может не встать. Вполне возможное в принципе формирование и пополнение Машинного фонда диалектными текстами осложняется традиционной проблемой дифференциального и недифференциального подхода к областной лексике. Проблема эта имеет уже свою традицию в региональной лексикографии. Как ни парадоксально, как раз Машинный фонд позволяет разрешить эту традиционную антиномию в сторону охвата диалектной народно-разговорной лексики во всем ее объеме, во всей ее полноте по отдельным диалектам, говорам, ибо здесь речь идет не о публикации, а только о накоплении данных и об их хранении. Новые типы носителей в виде дисков, лент даже сегодня позволяют уже чрезвычайно экономно хранить огромные объемы информации. Постепенное, неторопливое накопление диалектной лексики в текстах на магнитных лентах создаст с годами неоценимые фонды материалов.

Хотя мы располагаем рядом уникальных по охвату и полноте картотек областной лексики в Институте русского языка АН СССР, в Ленинграде, в Ленинградском и Московском университетах, в Педагогическом институте им. А. И. Герцена в Ленинграде, в вузах Вологды, Новгорода, Пскова, Ярославля, Петрозаводска, и здесь создание Машинного фонда должно идти своим путем, совершенно независимым от существующих картотек и тем. Перенос данных этих картотек в Машинный фонд русского языка нецелесообразен, да и невозможен практически. Картотеки создавались в течение десятилетий, по своим программам, с ориентацией на конкретные темы, они крайне неоднородны и сложны по своему характеру, структуре. Машинный фонд может ориентироваться только на полные магнитофонные записи диалектных текстов. Наиболее сложной проблемой здесь является многоступенчатость обработки таких данных для ввода в ЭВМ. Это — расшифровка магнитофонных лент, оформление черновых и чистовых записей, разрешение проблемы сохранения всех знаков фонетической транскрипции при вводе данных в ЭВМ и выводе их на печать. Те же проблемы встают и при строгой обработке записей по разговорной речи в целом.

Подобно диалекту, как синхронии, так и диахронии принадлежат также язык фольклора и язык земли — топонимика. Всем известно жанровое многообразие русского фольклора — былины, сказки, исторические и лирические песни, пословицы, обрядовая поэзия, загадки, частушки,— все это предопределяет разнообразие фольклорных текстов. Другой их отличительной чертой яв-

ляется множество разнородных и пестрых вариантных записей, сделанных разными собирателями в XIX—XX вв.

Ввод фольклорных текстов в Машинный фонд русского языка требует большой исследовательской текстологической и лингвистической подготовительной работы.

Среди проблем ономастики прежде всего следует остановиться на топонимике. Для топонимики совсем иначе стоит вопрос о текстах. Текстом для топонима является привязка его к объекту наименования в определенных границах (город, село, река, гора, озеро, море, урочище, лес, овраг, поле, кряж, холм). Топонимы по своей семантике близки к терминам, и именно поэтому перспективы организации машинных фондов по топонимике можно считать весьма обнадеживающими.

Особого обсуждения требует организация и других подфондов по ономастике — имен и фамилий людей, кличек животных и т. д. Разумеется, каждый тип текстов образует в рамках единого Машинного фонда русского языка свою подсистему, создание каждой из которых требует отдельного специализированного обсуждения. В связи с тем, что современные программы позволяют нам одновременно извлекать из памяти ЭВМ и отрывки текстов, и все словоформы с их точными адресами вплоть до строки на странице, и контексты на ту или иную словоформу, практически отпадает необходимость в отдельных подсистемах хранения лексики того или иного типа. Тщательное хранение целостных текстов и наличие соответствующих программ и программ сведения разных форм слов к одной исходной позволяет выдавать пользователю все основные справки по лексике.

Какого рода справки чаще всего нужны филологу, лингвисту? в каких источниках, текстах, подъязыках встречается то или иное слово? есть ли это слово в таких-то текстах? каковы все формы слова, включая его графические и фонетические формы? каковы все контексты на то или иное слово? было ли известно слово, форма в прошлом, в текстах, например, XIV, XV, XVI вв.? с какого времени оно появляется в словарях? какова суммарная позиция отдельных букв, морфем? кто, что, когда и где писал о том или ином слове, форме, явлении? найти все слова, обозначающие, например, одежду, обувь или отрицательные свойства и качества человека.

Можно предположить, что в конечном счете для лингвиста во многих случаях наиболее релевантными и пертинентными окажутся те ответы, которые помогут ему в выявлении новых тонких семантических оттенков, употреблений.

Различие между традиционными картотеками и регистрирующими подсистемами Машинного фонда отнюдь не чисто технологическое, как это может показаться на первый взгляд. Это фонды хранения качественно иного типа. Это фонды не слов и цитат, а прежде всего — целостных текстов на новых типах носителей со множеством параметров и входов по каждому из параметров.

Подсистемы машинных фондов регистрирующего типа не исключают фондов других типов, а именно интерпретационных подсистем исследовательского характера. Так, в рамках единого Машинного фонда русского языка как микросистемы различных макросистем крайне необходимы, например, подфонды словарных статей по отдельным типам лексики, фонд семантических представлений текстов, фонд грамматик, но это уже подсистемы и фонды качественно иного уровня, чем подсистемы регистрирующего типа. В то же время именно совокупность подсистем регистрирующего типа — основа различных интерпретационных фондов. В результате одноразового ввода в ЭВМ текста тех или иных источников потребитель впоследствии получает широкие возможности неоднократного обращения к данной автоматизированной системе с самыми различными запросами. Подобные системы в виде так называемых словарных банков данных не только обеспечивают предельную точность, надежность и проверяемость хранения всей информации, содержащейся в тексте, но и дают возможность получения ответов о всех употреблениях этого слова, о всех контекстах на каждое слово, на каждую словоформу, о всех типах вариантов как по отдельному источнику, так и по той или иной совокупности источников.

Первоначально в течение довольно длительного времени, пока тот или иной подфонд не набрал еще соответствующую глубину и полноту, число обращений к нему будет невелико. Со временем, постепенно, с ростом числа введенных текстов, данных количество запросов начнет неуклонно расти. Машинный фонд русского языка не заменяет существующих картотек и словарей и не отменяет новых словарей. Напротив, он впервые создает такую базу для всей нашей словарной работы, которой никогда ранее не было. Машинный фонд русского языка — залог появления новых типов словарей самого разного назначения и ориентации. Как и всякая большая новая тема, идея Машинного фонда русского языка ждет серьезного и детального проекта.

Среди неясных и требующих обсуждения проблем организации Фонда назовем: прогресс в технике, сменяемость марок машин, типов носителей, развитие языков программирования и сохранение стабильности Фонда; обеспечение сохранности фондов, предупреждение их от разрушения; совместимость различных подсистем между собой в рамках единой макросистемы; разрешение антиномии глубина банка — полнота банка данных; ориентация Фонда и его прагматика; системы универсального типа и системы проблемно ориентированные; соотношение ручного и автоматического труда при подготовке данных.

Отдельно следовало бы обсудить чисто психологические проблемы контактирования пользователя с будущим Машинным фондом русского языка.

Создание Машинного фонда русского языка как единой макросистемы в единстве и многообразии взаимно совмещенных

www.ruslang.ru

микросистем — дело сложное и трудоемкое, но реальное. Оно осуществимо путем неторопливого, поэтапного ввода одной подсистемы за другой, в зависимости от степени их филологической, концептуальной, программной, экспериментально-машинной и технологической подготовленности. Обсуждая далее проблемы создания единого Машинного фонда русского языка, целесообразно, с одной стороны, разделять разные типы фондов, а с другой — различные уровни самого обсуждения.

Машинный фонд русского языка — это большая автоматизированая макросистема, включающая в себя ряд подсистем и микросистем. Одна из таких подсистем — это подсистемы текстов, слов, словоформ, словосочетаний и предложений, фонды, ориентированные на фиксацию фактов речи.

ОПЫТ СОЗДАНИЯ НАЦИОНАЛЬНЫХ ЛЕКСИКОГРАФИЧЕСКИХ СЛУЖБ ЗА РУБЕЖОМ

В. Ю. РОЗЕНЦВЕЙГ

Перед тем как перейти к изложению относящегося к теме материала, сделаем несколько предварительных замечаний.

Под национальными лексикографическими службами мы будем понимать здесь службы, созданные не только государственными учреждениями (университетами, институтами), но и частными предприятиями. В промышленно развитых странах Запада словарное дело в последние годы вышло далеко за традиционные рамки и включено в сферу интересов крупных фирм.

Второе замечание касается понятия службы: мы не будем говорить лишь о таких лексикографических службах, которые используют идеи и методы современной информатики и оборудованы электронно-вычислительной техникой. Как мы увидим, отличие таких служб от традиционно сложившихся в словарном деле и использующих бумажные носители и механическое оборудование столь значительно, что можно говорить о стыковке, если не о слиянии лексикографии, информатики и современного полиграфического производства.

Отметим, наконец, что все, о чем будет здесь сообщено, основано лишь на изучении специальной литературы: никаких собственных наблюдений работы зарубежных лексикографических служб у автора не было. Если к этому добавить, что технологические сведения о работах таких служб и вообще сведения об использовании ЭВМ для создания и эксплуатации систем обработки информации являются, как известно, объектом торговли, а следовательно, держатся в секрете, то ясно станет, что это положение нуждается в дополнениях и, вероятно, в многочисленных уточнениях.

Рассмотрим вкратце типы современных лексикографических служб.

Научно-исследовательская лексикографическая служба (составление по тексту списков слов, индексов, конкордансов, частных и обратных словарей)

Использование ЭВМ в лексикографии началось с автоматизации весьма традиционной словарной работы, частично решаемой в прошлом при помощи счетно-перфорационных машин (СПМ),— выявления словарного состава текстов, как правило — древних, их частотных и грамматических характеристик (см.: Алфавитный, частотный и обратный словарь восточнославянских юридических текстов XI—XVI вв. (Сост. В. З. Санников) Под ред. В. И. Борковского. Ин-т рус. яз. Группа структурной и экспериментальной лингвистики. Предварительные публикации. М., 1974. Вып. 53—54). На предварительном этапе машины (СПМ, а затем — ЭВМ) использовались в основном для сортировки и подсчета частот слов и грамматических категорий, печати соответствующих списков, причем кодировка текстового материала производилась на перфокартах при помощи перфораторов. Преобразование текстовых единиц в словарные единицы — приведение словоформ текста к исходным заглавным формам, идентификация и дифференциация омографов, а тем более выявление синтаксических свойств слова — оставалось целиком делом исследователя-лингвиста.

Существенно важно отметить и полное отделение операций по такой обработке текста от операций по распространению и изданию полученных при этом результатов.

Использование методов, разработанных в автоматизированных системах поиска научной информации, и применение ЭВМ позволило в последние 10—15 лет объединить все операции по обработке текста с целью выявления лингвистических характеристик его словарного состава и издания полученных результатов в виде конкордансов и словарей в единую автоматизированную систему. Филологу (лингвисту, литературоведу) остается определить тексты, которые желательно обследовать, и вид публикации результатов, а сотрудникам лексикографической службы — обеспечить кодировку текстового материала на машинных носителях, главным образом на магнитной ленте для ввода в ЭВМ, а также корректуру. Подчеркнем, однако, что материал вводится в ЭВМ непосредственно с пульта (печатанием на клавиатуре) и также корректируется. Отметим еще, что автоматическое выявление контекста словарной единицы (строки или строк, в котором словоформа встречается в тексте, или же определенного числа слофором, предшествующих слову и/или следующих за ним — по типу KWIC) значительно увеличивает научную ценность результатов, позволяет изучать словосочетание, синтаксис, стилистику данного текста (группы текстов) как с лингвистической, так и с литературоведческой точки зрения. Много памятников культуры, в особенности средневековой

письменности, впервые включены таким образом в научный обиход.

Из многочисленных зарубежных программ такого рода словарной обработки текстов можно выделить многоцелевую программу Оксфордского университета. Ее достоинства: она не зависит от типа и модели ЭВМ, ее язык команд легко и быстро усваивается и не требует никаких специальных знаний. После того как текст введен с пульта (фрагментами или полностью, причем возможны пропуски), система выдает по запросу полные списки слов текста, индексированные списки (каждому слову приписан его адрес в тексте), конкорданс с указанием частотных характеристик словарной единицы и ее контекста, списки морфем (аффиксов) слов, словосочетаний, сгруппированных по частотным характеристикам, по количеству букв, а также списки, сочетающие разные характеристики (например, возможен такой запрос: выдать все слова длиной больше n букв с частотной характеристикой, с окончанием на -ed и т. п.). Списки могут быть получены в любом порядке (алфавитном, обратном, по длине, частоте, по контексту — левому и/или правому); формат распечатки (количество строк, интервал между строками, размер левого и правого поля, количество колонок и интервал между ними и т. п.) также может быть задан. Вот пример работы рассматриваемой службы: в ЭВМ введены (разными шрифтами!) тексты пьес Софокла (с указанием названия пьесы, ролей, адресов). Требуется: конкорданс всех хоровых текстов в «Антигоне» и «Эдипе», выбрав только слова, в которых предпоследняя буква «а», с указанием их адресов — пьесы, роли, номера строки контекста (6 слов слева и 6 слов справа от каждого выдаваемого слова), с частотной характеристикой, для каждого контекста указать 2 предшествующие и 2 следующие строки; выдачу печатать двумя колонками, в каждой колонке 55 строк, ширина левого поля — 5 букв, интервал между колонками — 8 букв, оставить интервал в одну строчку между ключевыми словами и т. п.

Отметим, что уже при составлении конкордансов выявилась необходимость использования грамматических сведений; без такого рода сведений программа не может автоматически сводить словоформы текста к единой, исходной форме, различать омонимичные формы, не говоря уже о синтаксическом анализе.

В работе научно-исследовательских лексикографических служб встала также проблема сочетания удовлетворения индивидуальных запросов ученых с публикацией полных результатов, полученных при обработке текстов. Хорошо знакомый лексикографам конфликт между стремлением к полноте охвата материала и требованием его обозримости и практической доступности проявляется и при составлении конкордансов. Так, примерно одну треть объема конкордансов составляют служебные слова. Очевидно, между тем, что список, скажем, всех контекстов союза and в сочинениях Шекспира требуется немногим

специалистам, а занимает такой список в печатном конкордансе сочинений Шекспира десятки страниц, что не только не экономно, но и неудобно для пользователя. Понятна поэтому наметившаяся тенденция к выделению двух разных функций зарубежных университетских лексикографических служб: справочной и издательской. Первая отвечает на индивидуальные запросы специалистов, вторая обеспечивает публикациями широкий круг научной общественности. В этой связи усиливается развитие служб рассматриваемого типа в направлении национальных, многоаспектных лексикографических информационных центров, располагающих обширным банком данных на магнитных носителях и современным полиграфическим оборудованием. Такие центры, в виде университетских лабораторий или вычислительных центров, имеются сейчас в Англии, Франции, ФРГ, в Скандинавских странах. Некоторые из них, например вычислительная лаборатория Гетеборгского университета, ставят задачу создания фонда современного шведского языка.

Подчеркнем, что уже автоматизированная лексикографическая служба научно-исследовательского типа требует не просто замены традиционной словарной картотеки «автоматизированной», а такого словарного описания, которое обеспечивает автоматизацию всех этапов перехода от текста к данным и от данных к тексту и объединения их в единый процесс поиска, хранения и выдачи (включая печатание) словарной информации. Сказанное относится еще в большей мере к лексикографической службе терминологического типа.

Терминологическая лексикографическая служба

Тенденция к созданию национальных и даже международных служб в области научно-технической терминологии проявилась, как известно, давно, но быстрое продвижение в этом направлении началось лишь с применением ЭВМ. Значительным стимулом здесь была потребность в системах автоматического поиска научно-технической информации, предполагающих наличие сверхбольших словарей-тезаурусов, в которых о каждом термине сообщаются его стандартное наименование, возможные варианты наименования, включая сокращения, отрасль, к которой он относится, некоторые грамматические (морфологические) сведения, а также, весьма часто, эквиваленты на другом языке.

В нашей стране таких словарей имеется великое множество (вопросами стандартизации научно-технической терминологии ведает ВНИИКИ Госстандарта); ведутся также работы по созданию двуязычных (многоязычных) словарей-тезаурусов.

Применение ЭВМ позволило и здесь создать банки терминологических данных, в которых ввод, хранение, пополнение, обработка, выдача и распространение, включая издание терминологической лексики, объединены в единый процесс, причем

все функции, за исключением опять-таки кодирования и ввода информации, автоматизированы. Функции этих банков данных и соответствующих терминологических служб разные. Основные из них: поиск научно-технической информации, составление и перевод на другой язык научно-технической документации, нормализация терминологического узуса. По объему охваченного терминологического материала, разнообразию услуг и оперативности (срокам ввода, переработки и выдачи информации) терминологические банки не сопоставимы с составленными вручную терминологическими картотеками и словарями.

Проиллюстрируем сказанное на примере терминологического банка западногерманской фирмы «Сименс» (TEAM) [1]. Банк не только обслуживает потребности этой фирмы в изготовлении на немецком, английском, испанском, русском и других языках научной документации к своей продукции, простирающейся от домашних электроприборов до крупных вычислительных машин, но и ведет документацию крупнейших западногерманских банков, промышленных фирм, многочисленных государственных и провинциальных учреждений, а также терминологическую службу Западногерманского института стандартизации терминологии (DIN). Это по существу национальная терминолгическая служба ФРГ (с ней сопоставима разве что западногерманская государственная лексикографическая информационная служба Lexis). Объем банка (к началу 1981 г.) — 1 млн. немецких терминологических единиц и, соответственно, 2—3 млн. терминов на других языках. Структура информации терминологической единицы (ТЕ) следующая: максимальный объем данных — 3000 букв, число категорий 99, а именно 00—09 — название ТЕ (вхождения), буквенно-числовой код, позволяющий «вызвать» ТЕ для добавления, исправления или стирания, дата ввода и/или последнего просмотра, обозначение лица, ответственного за ввод, обозначение блока, к которому принадлежит ТЕ, предметная область ТЕ и др.; 10—19: немецкая ТЕ (слово или словосочетание) — название, грамматическая информация, источник сведений, индекс или поисковый образ (для ТЕ-словосочетаний), толкование, примеры контекстов, синонимы и т. п. Остальные категории отводятся другим языкам: 20—29: английская ТЕ; 30—39: французская ТЕ; 40—49: испанская ТЕ; 50—59: русская ТЕ; 60—69: итальянская ТЕ; 70—79: португальская ТЕ; 80—89: датская ТЕ. Оставшиеся категории предназначены для более редких языков и для терминологических кодов УДК. Строится добавочный банк для арабского языка.

Все данные ТЕ соответствуют международным стандартам и совместимы с другими банками данных. Всего заданы 17 стандартных характеристик ТЕ, из которых обязательны 7: код, обеспечивающий однозначную идентификацию, оценка «качества» ТЕ, дата ввода и просмотра, ответственный за ТЕ, область

[1] *Brinkmann K.-H.* Machine Aids to Translator.— META, XXVI (1981), p. 67—75.

применения, слово или словосочетание — название ТЕ, источник. Необязательные: толкование, примеры контекста, название исходного языка, ограничения на применение ТЕ, предметная область, местные ограничения на применение термина, грамматические сведения, сокращения, рекомендации, касающиеся употребления ТЕ, библиографическая информация. Банк, естественно, регулярно пополняется. Функции терминологической службы: первая и главная функция — терминологическое обслуживание переводчиков (их от 300 до 400): выдача по запросу списков терминов по предметной области, применительно к оборудованию, в порядке их встречаемости в тексте и/или в алфавитном порядке. Термины выдаются на экран терминала и/или в распечатанном виде. О ТЕ сообщаются частотные характеристики, сведения о «возрасте» и источнике и т. п.

Вторая функция — унификация терминологии. Этой цели служат списки-глоссарии, бюллетени новых терминов-эквивалентов — на бумаге, микрофильмах, микролентах и т. п.

Третья функция — терминологическое обслуживание обучающихся на курсах фирмы «техническому» языку — издание словарей-минимумов по данной специальности на данном языке.

На четвертой функции остановимся несколько подробнее. Речь идет о составлении и издании технических словарей. Объединение всех функций терминологической службы (поиска терминологической лексики по тексту, хранения, обновления, выдачи как по индивидуальному запросу, так и в виде глоссариев и т. п.) в единый информатизированный процесс естественно смыкается с изданием словарей. Показателен и поучителен в этом отношении опыт совместного издания терминологических словарей, осуществленного в 1972 г. терминологической службой TEAM и западногерманским издательством, специализированном в выпуске технических словарей. При этом ставились следующие задачи: поставить англо-немецкий политехнический словарь (около 150 тыс. вхождений) на машину, дополнить его, обратить его в немецко-английский. Осуществление проекта потребовало прежде всего определения структуры (формата) словарной статьи. «Тщательное определение задаваемого формата,— пишет сотрудник терминологической службы, участвовавший в реализации проекта,— имеет решающее значение для экономической реализации словарного банка данных и для успеха подобного проекта в целом»[2]. Суть этой работы — такое уточнение и стандартное представление информации о словарной единице в «обычном» словаре, которое позволило бы задать ЭВМ программу операций, необходимых для ввода, обновления, неоднократного извлечения и всякого рода преобразования части или всей этой информации. За основу был взят формат, принятый в терминологической службе TEAM, включающий следующие основные зоны: имя терминологической единицы

[2] *Vollnhals O.* Die datengesteuerte Herstellung einer Wörterbuchreihe.— BABEL, XXIV (1978), 2, p. 65.

(ТЕ), грамматические категории, синонимичные названия (включая сокращения), предметная область, перевод. Отношение количества знаков в «обычном» исходном формате к принятому формату — приблизительно 1 : 2. Перекодировка формата и ввод дополненного словаря (152 тыс. вхождений) в машину в новом формате заняли около 15 месяцев. Двойная корректура, неоднократные проверки на совместимость разного рода информации, добавление нового материала, выявление омографов (их оказалось около 65 тыс.), устранение омографов автором (ему была выдана распечатка полного списка омографов) заняли чуть больше четырех месяцев. Введенные окончательно в машину 140 676 ТЕ были рассортированы по нескольким признакам (по алфавиту, номеру ТЕ, области применения и др.), на что потребовался примерно один час работы машины. Фотоскопически набранный (в заданном формате) словарь объемом около 1000 стр. был в виде ксерокопии передан автору для просмотра и внесения дополнений. Предельный срок авторской правки — 8 недель до выпуска словаря в свет. В силу того что материал был подвергнут двойной корректуре и разного рода проверкам, авторский просмотр (1000 страниц!) длился не боле двух с половиной месяцев (потребовалось исправление около 5% вхождений). Окончательный фотонабор, корректура и печатание словаря заняли 3 месяца. Всего, таким образом, от начала работы над терминологическим словарем (начало 1972 г.) до выхода в свет этого словаря в новом формате (сентябрь 1974 г.) ушло около двух с половиной лет. Работа по обращению англо-немецкого словаря в немецко-английский (включая пополнение английской части и конвертирования в новый формат), начатая в октябре 1974 г., была закончена к концу октября 1975 г.

Представляют интерес данные о стоимости указанной работы. Изготовление банка данных (перекодировка, пополнение, ввод в машину и т. д.) и первое издание словаря обошлось несколько дороже перенабора. С учетом обратимости словаря (с англо-немецкого в немецко-английский) стоимость работы равна половине стоимости обычного издания; с учетом переиздания (при десятипроцентном изменении материала) стоимость снижается до одной трети обычной; с учетом обращения словаря на третий язык (немецко-французский) стоимость работы снижается еще на 50%. При этом следует принять в соображение, что наличие словарного банка данных обеспечивает многократность издания, причем в разных объемах и форматах.

Многоцелевые общесловарные лексикографические службы

Сближение информатики и лексикографии идет не только со стороны терминологических банков данных, но и со стороны лексикографии, точнее — со стороны лингвистов, стремящихся

использовать ЭВМ и современные возможности полиграфической технологии для издания больших, национальных толковых словарей. Имеются в виду словари, содержащие не 100—150 тыс. статей (ССРЛЯ), 400 тыс. статей (Оксфордский словарь) и даже не 600 тыс. статей (словарь Вебстера), а еще более крупные, причем построенные на основании обследования колоссального объема текстов. Ясно, что сбор и обработка таких массивов в сжатые сроки немыслима без современных методов и техники информатики.

Эта тенденция проявляется в разных странах, полнее и реальнее всего — во Франции. Уже с 1956 г. издательство Larousse перешло на машинную картотеку, а с 1963—1964 гг. использует ЭВМ для автоматического составления и набора небольших словарей [3].

Еще более значительным предприятием является создание словаря Trésor de la langue francaise, работа над которым была начата в 60-е годы по инициативе французских лексикографов (P. Imbs, B. Quemada) в специально для этого созданном Научно-исследовательском центре (ныне — Институте французского языка) Национального центра научных исследований Франции. Концепция издаваемого словаря менялась: задуманный первоначально как полное словарное описание французского слова от его истоков до наших дней, чисто лингвистическое (а не энциклопедическое), предназначенное не столько для широкой публики, сколько для специалистов, Тезаурус должен был состоять из двух частей — старофранцузского словаря (до XVI в.) и новофранцузского словаря (от XVI в. до нашего времени). Ставилась задача создать картотеку объемом около 250 млн. цитат (25 млн. карточек старофранцузского, 40 млн.— среднефранцузского, 40 млн.— XVI в., 60 млн.— XVII и XVIII вв., 85 млн.— XIX—XX вв.), извлеченных из художественно-литературных (80%) и прочих (20%) текстов [4]. Для начала издается Тезаурус французского языка XIX—XX вв. (ТФЯ) в 15 томах (объемом 1300 стр. каждый). Издание будет завершено в 1987 г.

Здесь нет возможности лексикографически охарактеризовать этот словарь. Отметим лишь, что по количеству цитат, иллюстрирующих употребление лексических единиц, он значительно превосходит существующие французские, да и не только французские словари. Следует вместе с этим признать, что осуществ-

[3] В последнее время во французской печати обсуждается идея постановки на машину известного общеупотребительного словаря Petit Larousse и создания автоматизированного словаря, к которому можно было бы обратиться по телефону, набрав соответствующим образом закодированный запрос, и который сообщил бы ответ на домашний телеэкран. Предположительная стоимость системы — 10 млн. франков. Рассматриваются и альтернативные возможности автоматизации словарного поиска, связанные с широким распространением микрокомпьютеров и видеодисков. См.: Le Monde, 30.IV.1982.

[4] Напомним для сравнения, что картотека, служащая основой БАС и МАС, насчитывает около 7 млн. цитат. См.: *Сороколетов Ф. П.* Традиция русской советской лексикографии.— ВЯ, 1978, № 3, с. 34.

ление первоначального замысла натолкнулось на экономические трудности, несмотря на то что ТФЯ не коммерческое, а научное издание, финансируемое государством. Этим объясняется некоторое изменение «лица» ТФЯ: задуманный, как уже сказано, как полное теоретическое лингвистическое описание современного литературного языка, предназначенное для филологов, он становится полным, нормативным словарем широкого пользования. Достаточно сказать, что из 100 млн. цитат современного языка, имеющихся в распоряжении составителей, в ТФЯ используется всего лишь 1%.

Лексикографическая служба ТФЯ — колоссальный банк словарных данных, создание и эксплуатация которого возможны лишь благодаря использованию методов и технологии современной информатики. Обработка текстов на ЭВМ позволила получить для каждого слова (словник ТФЯ насчитывает 800 тыс. единиц) его контекст из трех строк и соответствующий фрагмент текста объемом в 15 строк. Преобразование словоформ текста в заглавную форму и дифференциация омографов, статистика встречаемости и т. п. осуществляются алгоритмически. Кроме текстов, были обработаны все французские словари, собрана документация об описаниях французских слов в лингвистических работах.

Лексикографическое использование материала такого объема и составление словарных статей — дело штата сотрудников, образующих Институт французского языка[5] с отделами документации, этимологии, фонетики, грамматики, семантики, статистики, редакционно-издательской работы. Особое место отведено отделу лексикографической консультации: любой филолог может здесь получить непосредственно или же письменно справку по словарному составу французского языка. Предусмотрена разработка диалоговой системы, что открыло бы непосредственный доступ к банку данных. И здесь, таким образом, заметна тенденция к созданию многоцелевой национальной лексикографической службы.

Из сказанного намечаются следующие выводы.

1. Создание лексикографических систем, объединяющих процессы сбора, обработки, хранения и выдачи (включая издание) словарных данных, — вполне реальное, экономически рентабельное и научно перспективное дело.

2. Создание единых автоматизированных служб особенно эффективно при сочленении лексикографической службы со словарным изданием, на базе современных методов и техники информатики и полиграфии. Современная лексикография находится на стыке науки и промышленности.

---

[5] Институт пользуется услугами многочисленных французских лингвистов, присылающих ему карточки-цитаты выражений, привлекших их внимание. В обмен на подобные услуги Институт высылает своим членам-корреспондентам издаваемые им лексикографические бюллетени.

3. Развитие автоматизированной лексикографической службы требует не только формализации лексикографии, но и стыковки лексического описания с грамматическим описанием в рамках интегральной лингвистической модели. Лексикография становится современной экспериментальной базой лингвистики.

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА И АВТОМАТИЗАЦИЯ СЛОВАРНО-ТЕРМИНОЛОГИЧЕСКОГО ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ СЛУЖБЫ

Р. Г. КОТОВ

Обсуждение путей построения Машинного фонда русского языка свидетельствует об осознании широкими кругами лингвистов необходимости создания своего рода машинно-информационной базы, которая позволила бы перейти на качественно новый уровень теоретических и прикладных исследований языка. Понятие Машинного фонда русского языка появилось в среде лингвистов давно, но обсуждение вопросов, связанных с его конкретным содержанием и возможным использованием, фактически началось в 1983 г. Оказалось, что в создании Машинного фонда русского языка заинтересованы представители самых разных языковедческих дисциплин и направлений, занятые решением различных не связанных между собой проблем (история языка и литературы, языковое нормирование, структурные исследования разных уровней языковой системы, описание отдельных категорий языка и формальные модели языка в целом, автоматизация обработки языка в различных прикладных целях и т. п.). Естественно, что при этом выявились различные взгляды на то, что же должен представлять собой такой Фонд.

Если обобщить различные представления о Машинном фонде, то получается, что он должен обладать по крайней мере свойствами некоторой коллекции результатов речевой деятельности (тексты разных жанров и стилей), представлять собой хранилище лингвистических данных (словари и грамматики), быть инструментом исследований (программные средства обработки текстов) и одновременно программой исследований в соответствии с интересами представителей различных языковедческих дисциплин. В общем случае Фонд рассматривается как средство, обеспечивающее решение ближайших и перспективных проблем теоретического и прикладного языкознания, а также языковых проблем информатики. Отсюда следует, что речь идет о крупномасштабном проекте использования вычислительной техники в области лингвистики и смежных дисциплин.

Машинный фонд русского языка в такой постановке — это выражение потребностей внутреннего развития языкознания как

науки и потребностей технологической сферы общества в решении лингвистических задач, выдвигаемых быстро развивающейся информационной службой страны.

При любом толковании содержания и назначения Машинного фонда русского языка несомненно, что это должна быть большая система, основными компонентами которой являются сложная информационная база (база данных) с соответствующими программными средствами ее ведения и постоянно развивающаяся система программных средств обработки лингвистических данных — элементов текста или полных текстов (программы задач, преследующих исследовательские или прикладные цели — от выявления статистических и структурных закономерностей до анализа и синтеза текста, включая и задачи смысловой обработки заключенной в нем информации). Машинный фонд русского языка представляется в данном случае как уникальная информационная система многоцелевого назначения и коллективного пользования. Но это означает, что при проектировании и создании такой системы вряд ли будет достаточно тех отработанных решений, которые используются при построении типовых систем обработки информации в нашей стране и за рубежом. Своеобразие Машинного фонда русского языка как системы потребует новых решений для информационного, лингвистического, программного, технического и организационного обеспечения на всех этапах ее проектирования и реализации.

Следует учесть и особенности круга пользователей будущего Фонда: это не однородная среда абонентов системы, это лингвисты и филологи различного профиля с практически неограниченным перечнем своих задач. Они не только пользователи системы, но и ее создатели, по крайней мере в части сбора и подготовки информационных фондов. Они не образуют единый коллектив, они разобщены профессионально, организационно и территориально. И в этих условиях значительный объем работ по подготовке информационных фондов, очевидно, должен выполняться на общественных началах. Это весьма трудная задача. Еще большие трудности представит проектирование и разработка минимального начального варианта программного обеспечения для ведения информационного фонда (можно предположить, что стандартные системы управления базами данных окажутся мало пригодными для этого). Особый вопрос — это организация начальной технической базы Машинного фонда русского языка. В этих условиях на первый план выходят организационные проблемы, проблемы рационального использования возможностей лингвистов, филологов и специалистов других дисциплин, находящихся в различных учреждениях, НИИ и вузах. Насколько известно, опыта организации и проведения таких крупномасштабных работ в области языкознания нет.

С учетом сказанного очевидно, что проект Машинного фонда русского языка в такой широкой постановке можно рассматри-

вать лишь как «голубую мечту». Требуется более узкая формулировка цели и задач Фонда с учетом весьма ограниченных возможностей реализации и отсутствия опыта таких работ.

Основной вопрос — это определение ориентации, т. е. определение круга пользователей и тех задач, которые могут реально решаться на базе Машинного фонда. Четкая ориентация позволит определить как содержание информационных фондов, так и состав программных средств.

Видимо, удовлетворить потребности всех направлений языкознания и потребности технологической сферы сразу невозможно. Действительно, задачи литературоведения и филологии, задачи исследования структуры языка, задачи нормирования языка, нормализации и унификации терминологии, задачи прикладного характера, основу которых составляют анализ и синтез, понимание текстов, требуют информационных фондов разной структуры и содержания, а также и разных программных средств. Вряд ли все это можно сразу объединить в одной системе. Однако при любом ограничении ориентации Машинного фонда русского языка необходимо, на наш взгляд, обеспечить интересы и лингвистики, и технологической сферы. Для этого Фонд должен иметь такие информационные базы и программные средства, которые позволяли бы ему решать и практические задачи (хотя бы ограниченные) и быть инструментом исследования.

С этой точки зрения интересы языкознания (общее, русское и прикладное) и технологической сферы (информационная служба) наиболее совпадают в проблемах нормирования языка, особенно языка деловой документации, унификации и стандартизации научной, технической, экономической терминологии (в более общей формулировке — специальная лексика) и, наконец, в проблемах автоматической обработки документов (индексирование, реферирование, форматирование, перевод) и использования естественного языка для общения с ЭВМ (доступ к базам данных, управление вычислительным процессом, программирование).

Это два взаимосвязанных направления ориентации Машинного фонда русского языка: нормирование языка и терминологии и обработка текста. Однако по наличию задела в части информационных фондов и программного обеспечения, который может быть использован в Фонде, они неравнозначны.

Информационный фонд первого направления — это словари. Языкознание в рамках лексикографических и прикладных работ накопило большой словарный фонд — словари разного типа и назначения, среди которых имеются и словари русского языка и терминологические словари по отдельным отраслям науки и техники, не считая многочисленных частотных словарей. Информационная служба совместно с инженерной лингвистикой также создали большой словарно-терминологический фонд, реализованный в общесоюзных и ведомственных классификаторах,

рубрикаторах и тезаурусах, машинных словарях разного назначения. Имеются и достаточно отработанные основные программные средства составления и ведения информационных словарей, имеют место работы по созданию программных средств ведения больших терминологических баз.

Второе направление, связанное с проблематикой обработки текста, не вышло пока еще из стадии теоретических и экспериментальных исследований (если не считать немногочисленных систем промышленного машинного перевода научно-технической документации). Это направление еще не имеет массивов лингвистических данных с достаточно отработанной структурой, которые можно было бы включать в Машинный фонд русского языка. Нет до сих пор полностью законченных формальных моделей языка, а перспективы их завершения становятся все более проблематичными. Неизвестно, когда могут быть созданы словари с «энциклопедической» информацией, которые рассматриваются как составная часть формальных описаний языка. Программные средства обработки текста в этом направлении также не вышли еще из экспериментального статуса.

С учетом этих обстоятельств представляется целесообразным и вполне оправданным решение ориентировать Машинный фонд русского языка на словарно-терминологическое обеспечение потребностей языкознания и информационной службы страны. При этом обеспечивается реальная основа для создания информационных фондов и определенной части программного обеспечения. Использование всего этого, а также всего опыта словарно-терминологической работы, накопленного языкознанием, лексикографией и информатикой, ставит проект Машинного фонда русского языка на реальную почву, так как в этом случае он создается не на пустом месте. Опасение того, что Машинный фонд в этом варианте возьмет на себя задачи терминологического обеспечения информационной службы, вряд ли состоятельно. В этом случае он выступает как необходимое дополнение и продолжение тех работ, которые ведутся информационной службой в своих целях. Информационные фонды и исследования, которые будут проводиться на их основе, не дублируют работы в области информатики. Однако и сами фонды, и результаты проводимых на их основе исследований могут оказать большую пользу работам по терминологическому обеспечению информационной службы.

Даже при таком серьезном ограничении рамок проекта Машинного фонда русского языка потребуется детальная проработка всех его компонентов уже на этапе проектирования. Особое внимание следовало бы обратить на следующие вопросы: а) выбор типа словарей и информационного формата словарной статьи с учетом ее заполнения данными по мере расширения круга задач; 2) выбор программных средств ведения словарей, включая широкие возможности поиска и извлечения информации с использованием разных входов в словарную статью;

в) выбор средств, обеспечивающих функционирование Фонда на распределенных технических средствах и базах данных в режиме коллективного пользования и др.

Решение этих вопросов может быть облегчено, если проект Машинного фонда русского языка будет увязан с проблемой автоматизации словарно-терминологического обеспечения одной из больших информационных систем, разрабатываемых в рамках информационной службы страны. Словарно-терминологический фонд таких систем, как ОГАС, ГАСНТИ, АИСОН, представляет широкую социальную информацию всех видов: технико-экономическую, научно-техническую, информацию по общественным наукам. Меньшие по масштабам системы машинного перевода и автоматических словарей для перевода работают с терминологией достаточно большого объема по отдельным областям науки и техники.

Словарно-терминологический фонд таких систем по своему объему и содержанию является сам по себе достаточно сложным и интересным объектом исследований как самой лингвистики, так и для прикладных целей. Накопление и исследование на базе Машинного фонда русского языка тех слоев общеупотребительной лексики и терминологии, которые используются в документации, но не учитываются в терминологическом обеспечении информационных систем, дает возможность комплексного изучения функционирования делового языка в условиях современного общества. Результаты таких исследований могут дать много полезного для языкового нормирования не только в части терминологии, но и языка в целом.

К настоящему времени информационная служба обладает огромными массивами терминологии, сформулирована проблематика ее нормирования, унификации и стандартизации, накоплен опыт создания и использования терминологии, создается задел в проектировании больших терминологических баз и банков.

Привязка проекта Машинного фонда русского языка к проектам информационной службы может означать не только облегчение всех вопросов проектирования, но и реальную поддержку со стороны заинтересованных организаций и ведомств. Выбор того или иного проекта, к которому будет привязан Машинный фонд русского языка, не является принципиальным вопросом. Важно другое, в этом случае уже на первых этапах его реализации (например, как словарь русского языка и общенаучной лексики) он может использоваться не только в интересах языкознания, но и для информационной службы страны как дополнительный компонент (со своими специфическими задачами) общего словарно-терминологического обеспечения.

Машинный фонд русского языка в таком варианте может рассматриваться как база для проведения исследований, связанных с проблемами лексикологии и лексикографии, с проблемами нормирования языка деловой документации и т. п., что соответствует интересам языкознания. С другой стороны, на его основе

могут проводиться исследования и поиск методов работы с терминологическими единицами, с информацией, содержащейся в словарных статьях. Опыт больших терминологических словарей на базе ЭВМ за рубежом показывает, что в разработке и эксплуатации таких словарей (какой бы простой на первый взгляд ни казалась эта задача) существуют нерешенные вопросы. Достаточно указать на необходимость обеспечения гибкого входа в массивы многокомпонентных терминов, в словарные статьи, возможности преобразования информационных форматов и т. д. Без этого пользователи всегда будут испытывать трудности при извлечении всей релевантной информации из словарей. А это — вопрос эффективности автоматических словарей в практической эксплуатации.

И в «простом» варианте Машинного фонда остаются и требуют тщательной проработки все системные стороны проекта (помимо выше упоминавшихся), в том числе: технические средства (ЭВМ и периферия), реально имеющиеся в вузах и учреждениях и доступные для данной работы, а также средства, на которые можно рассчитывать в ближайшем будущем; вопросы их совместного использования; технологический процесс подготовки различных видов лингвистических данных с учетом того, что они могут готовиться разными коллективами; вопросы программирования — языки, форматы, оформление документации и др.; вопросы использования систем редактирования, которые в условиях подготовки больших объемов текстовых данных приобретают большое значение (полезно сразу же подумать о методах автоматического обнаружения и исправления ошибок при вводе), и др.

В заключение можно сказать, что обсуждение проекта Машинного фонда русского языка — это столкновение мнений сторонников программы-максимума и программы-минимума.

Программа-максимум предполагает построение Фонда в виде некоторой сложной системы, включающей словари, тексты и программные средства для их содержательной обработки. Перспективы такого проекта заманчивы. Однако отсутствие апробированных программных средств содержательной обработки текста, отсутствие соответствующего лингвистического и программного обеспечения системы в целом, а главное — ясности в разрешимости многих связанных с ними проблем, к которым добавляются проблемы технического и организационного порядка, ставят под сомнение реальность такого проекта.

Программа-минимум, предполагающая создание автоматического словаря русского языка многоцелевого назначения в качестве базы для лексикологических и лексикографических исследований и работ, может рассматриваться как реальный вариант. Реализация его и опыт эксплуатации позволят более обоснованно подойти к проектам его дальнейшего расширения и по информации, и по программному обеспечению. Однако и эта задача при всей своей реальности остается еще достаточно сложной.

Разработка проекта такого варианта Машинного фонда русского языка также требует привлечения лингвистов различного профиля. Видимо, учитывая реальное положение в языкознании, с этого проекта и нужно начинать дело. Выполнение более простой и реальной задачи даст возможность проверить силы лингвистической общественности в решении языковедческих и системных вопросов и приобрести опыт, необходимый для решения более сложных задач.

НЕКОТОРЫЕ СЛЕДСТВИЯ ИЗ ЕДИНОЙ МОДЕЛИ ЯЗЫКА ДЛЯ МАШИННЫХ ФОНДОВ *

Ю. Д. АПРЕСЯН

Начавшееся в литературе обсуждение проекта создания машинных фондов (МФ) русского языка [1] позволяет говорить о двух типах МФ, которые можно условно назвать культурными и прикладными.

Была предложена следующая концепция культурных МФ, определяющая их объем, аспекты описания и функции. По объему — это самые разнообразные в стилистическом и жанровом отношении тексты на русском литературном языке, записи разговорной и диалектной речи, технические и научные терминосистемы, словарные картотеки по современному языку и по различным его историческим срезам, а также всевозможные готовые словари — от толковых до энциклопедических. Тем самым определяется и основной аспект описания языка в рамках культурных МФ — лексикографический. Наконец, по функциям культурные МФ мыслятся как сокровищница русской лексики, имеющая ценность сама по себе, или как архив для справочной службы по русскому языку, или как картотека материалов, которые могли бы служить основой разного рода лексикографических и лексикологических исследований.

Трудно отрицать полезность понимаемых таким образом МФ. Однако они не могут покрыть всех существующих в настоящее время информационных нужд. Необходим еще один род МФ, которые в противовес культурным МФ были названы выше прикладными. Прикладные фонды нужны для разного рода систем автоматической переработки текстов на естественных языках. В общих чертах прикладные МФ могут быть противопоставлены культурным следующим образом.

---

* Статья представляет собой расширенный и переработанный текст доклада, прочитанного 24 февр. 1983 г. в МГУ на Первой конференции по машинным фондам русского языка.

[1] *Андрющенко В. М.* Машинный фонд русского языка: постановка задачи и практические шаги.— ВЯ, 1985, № 2.— Некоторые из приводимых здесь примеров заимствованы из опубликованных работ автора.

По объему они должны быть ориентированы не на весь русский язык, а на тот язык, который представлен в научной, научно-технической и деловой прозе, технической и деловой документации и т. п. Следовательно, с этой точки зрения прикладные МФ значительно уступают культурным. По аспектам описания, наоборот, они должны превосходить культурные МФ. Действительно, для многих информационных систем, например систем АП или систем общения с базами знаний на естественном языке, недостаточно дать только лексикографическое описание русского языка. Необходимо по меньшей мере описать его грамматику — морфологию и синтаксис. Поэтому прикладные МФ должны быть в равной мере ориентированы и на словарь, и на грамматику. Наконец, по функциям прикладные МФ должны представлять собою лингвистическое обеспечение систем переработки текстовой информации на естественном языке. Рассматриваемые с этой точки зрения, они перестают быть просто сокровищницей, архивом или складом интересных сведений о языке. Это скорее фабрика, которая получает сырье и перерабатывает его в какой-то новый продукт.

Во всех существенных отношениях наиболее полными и надежными прикладными МФ русского языка являются соответствующие лингвистические компоненты систем АП. Их полнота объясняется самой природой переводческой деятельности, которая требует прежде всего владения словарем в достаточно большом объеме и владения практически полной грамматикой языка — морфологией и синтаксисом. Их надежность объясняется тем, что АП представляет собой своеобразный экспериментальный полигон для проверки «на прочность» различных моделей языка. Он дает лингвистике естественнонаучную методологию поиска научной истины. С одной стороны, АП является, так сказать, абонентом лингвистики: чем совершеннее модель языка, тем выше качество перевода. С другой стороны, лингвистика становится абонентом АП. Функционирование систем АП дает непосредственный импульс для формулировки новых правил, для возникновения новых теоретических понятий и целых концепций, оказывая тем самым мощное обратное воздействие на лингвистику.

В Советском Союзе было создано несколько доведенных до реализации на ЭВМ систем АП. Я буду говорить лишь об одной из них, в разработке которой принимал непосредственное участие. Речь идет об экспериментальной системе АП электротехнических текстов с французского языка на русский ЭТАП-I (Электротехнический автоматический перевод, первая очередь).

ЭТАП-I — система второго поколения, т. е. система, в которой переход от одного языка к другому совершается на уровне синтаксических структур, явным образом выделяемых алгоритмом в процессе перевода. Она работала в экспериментальном режиме около полутора лет и дала переводы, которые с рядом оговорок можно считать сравнимыми по качеству с человечески-

ми [2]. Тем самым входящая в ее состав модель русского языка получила некоторое естественнонаучное подтверждение, а вместе с ним — и право на то, чтобы рассматриваться как возможный компонент прикладных МФ.

Модель русского языка, входящая в состав ЭТАПа-I, представляет собой редуцированный вариант модели «Смысл — Текст» и состоит из трех основных компонентов — автоматического комбинаторного словаря, формальной морфологии и формального синтаксиса. Ниже я коротко опишу лишь два компонента модели — комбинаторный словарь и синтаксис.

Автоматический комбинаторный словарь в системе ЭТАП-I включает шесть типов информации о лексеме. Из них я коротко упомяну лишь три, представляющие не только технический, но и содержательный интерес: синтаксические признаки лексемы, факт ее цитирования в правилах и ее модель управления.

Синтаксическими признаками называются сокращенные и обобщенные обозначения тех синтаксических свойств лексемы, которые позволяют ей участвовать в одних конструкциях и не допускают ее участия в других. В традиционной грамматике хорошо известны по крайней мере имплицитно такие синтаксические признаки, как три рода, одушевленность и неодушевленность существительных, переходность, непереходность, связочность и безличность глаголов, качественность прилагательных и наречий, первообразность и непервообразность предлогов и т. п. Признак качественности прилагательных и наречий, например, необходим для описания атрибутивных конструкций, в которых такие прилагательные и наречия подчиняют наречия степени, ср.: *очень красивый, очень красиво.* Никакие другие прилагательные и наречия русского языка этим свойством, как известно, не обладают. В нашей формальной модели синтаксиса русского языка использовалось свыше ста синтаксических признаков. Понятно, что в дополнение к уже известным синтаксическим признакам пришлось выделить десятки новых.

Цитирование — это способ пометить определенную лексему как синтаксически релевантную. Лексема считается синтаксически релевантной, если она упоминается в синтаксических правилах. Один пример. В правиле формирования элективной конструкции типа *самый* [*наиболее, наименее*] *интересный из докладов* необходима ссылка на конкретные лексемы САМЫЙ, НАИБОЛЕЕ, НАИМЕНЕЕ. Только в непосредственном контексте этих лексем (или, в меньшей степени, лексем БОЛЕЕ, МЕНЕЕ) качественные прилагательные в положительной степени приобретают способность быть вершиной элективной конструк-

[2] Подробнее см.: *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-1.— В кн.: Разработка формальной модели естественного языка. Новосибирск, 1981 (в работе содержатся и библиографические данные, необходимые для понимания лингвистических истоков нашей работы).

ции; ср. неправильность: **интересный из докладов*. В словарных статьях лексем САМЫЙ, НАИБОЛЕЕ, НАИМЕНЕЕ должно быть помещено указание о том, что они цитируются в правиле. Помета о цитировании является, таким образом, частным случаем синтаксического признака. Она используется в ситуации, когда класс лексем, обладающих определенным синтаксическим свойством, численно мал и разнороден. Тогда введение синтаксического признака не дает никакого полезного обобщения и, следовательно, не имеет никаких содержательных преимуществ перед прямой ссылкой на конкретные лексемы, тем более что в предельном случае синтаксически релевантные лексемы образуют одноэлементные классы. Чтобы не выходить за пределы элективной конструкции, обратимся к лексеме ПОСЛЕДНИЙ. Наряду с порядковыми прилагательными (*первый* [*третий*] *из посетителей*) и прилагательными в превосходной степени (*лучший* [*худший*] *из людей*) она сама, без какого-либо специального контекста, обладает свойством быть вершиной элективной конструкции; ср.: *последний из них*.

Модели управления — это сложные теоретические конструкты, с помощью которых описываются управляющие свойства предикатных лексем. Они приписываются таким предикатным лексемам, семантические валентности которых могут быть реализованы синтаксически зависящими от них словами. Модели управления содержат до пяти типов информации о семантических валентностях слова, иллюстрируемых ниже на примере лексемы ПЛАТИТЬ: 1) Число возможных валентностей (мест, семантических актантов) данной лексемы; для ПЛАТИТЬ оно равно четырем: «кто платит» (субъект), «кому платит» (получатель), «сколько платит» (количество) и «за что платит» (объект); 2) способы поверхностно-синтаксической реализации валентностей; для ПЛАТИТЬ — это именительный падеж субъекта (*А платить вы собираетесь*?), дательный падеж получателя или В+винительный падеж (*Платить можно прямо продавцу; Платить следует в кассу*), винительный падеж именной группы со значением количества или количественное наречие (*Придется платить тысячу рублей*; *Сколько надо платить*?), ЗА+винительный падеж объекта (*Платите за проезд*); 3) семантические признаки лексем, претендующих на заполнение данной валентности предикатного слова; для первой и второй валентностей глагола ПЛАТИТЬ — это «человек» или «учреждение», для последней — «предмет» (*платить за книгу*) или «действие» (*платить за вход*); 4) синтаксические признаки слов, претендующих на заполнение данной валентности предикатного слова; для третьей валентности ПЛАТИТЬ — это синтаксический признак «колич», приписываемый существительным ТЫСЯЧА, МИЛЛИОН, МИЛЛИАРД и т. п. и наречиям типа МНОГО, МАЛО, СКОЛЬКО, или синтаксический признак «измер», приписываемый существительным со значением единицы измерения

(в частности, денежной единицы); 5) имена лексем, способных заполнять данную валентность предикатного слова, если их можно задать только «поименно», т. е. если у них нет никаких общих семантических или синтаксических признаков; для третьей валентности ПЛАТИТЬ — это лексемы СУММА и ДЕНЬГИ (ср.: *Придется платить крупные суммы* [*большие деньги*]).

Формальная модель синтаксиса русского языка является ядром модели языка в целом и поэтому будет изложена подробнее. Она включает в себя в качестве основной части правила, которые называются синтагмами. Каждая синтагма устанавливает соответствие между двумя элементами морфологической структуры предложения[3] и фрагментом его синтаксической структуры — бинарным деревом зависимостей, где эти два элемента связаны каким-то подчинительным отношением. Всего используется около 40 подчинительных отношений, таких, как предикативное (от сказуемого к подлежащему), комплетивные (от управляющего слова к его первому, второму и т. д. дополнению), определительное (от существительного или прилагательного к синтаксически согласованному с ним прилагательному) и т. п. Набор подчинительных отношений национально своеобразен (меняется от языка к языку) и семантически достаточен (обеспечивает сохранность смысла обрабатываемого предложения).

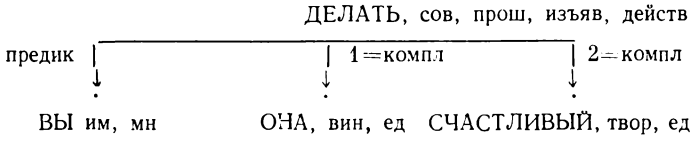
Как ясно из сказанного, каждое отношение подчинения задает большой класс однотипных синтаксических конструкций и обслуживается, в зависимости от объема этого класса, большим или меньшим количеством синтагм. Общее количество синтагм в модели русского синтаксиса — около 200. Каждая синтагма обслуживает один подкласс синтаксических конструкций. Так, основная комплетивная синтагма обрабатывает конструкции с каноническим управляемым падежом — тем, который записан в словаре в модели управления главного слова. Отдельная комплетивная синтагма обрабатывает конструкции с родительным падежом прямого дополнения в случае мены винительного на родительный в контексте отрицания (*За лето он не прочитал ни одной книги*). Другие комплетивные синтагмы обрабатывают конструкции, в которых вместо ожидаемого по модели управления винительного или родительного падежа представлена аппроксимативно-количественная группа (*За лето девочка прочитала около десяти книг; Командование бросило в прорыв до трех батальонов*), группа с распределительным ПО (*Каждому дали по рублю*) и т. п.

В ЭТАПе-I формальная модель русского синтаксиса использовалась в режиме синтеза. Однако ниже мы рассмотрим синтагмы только как правила синтаксического анализа текста, по-

[3] Под морфологической структурой предложения понимается линейная последовательность содержащихся в нем словоформ с приписанными каждой словоформе грамматическими характеристиками, т. е. значениями категорий падежа, числа, времени, вида и т. п.

тому что это дает лучшую возможность показать, как различные компоненты грамматики взаимодействуют друг с другом и со словарем, превращая лингвистическое описание в действующую модель языка.

В синтагмах используются результаты морфологического анализа предложения, словарная информация, текстовая информация (порядок слов, знаки препинания, наличие у определенных слов определенных зависимых), грамматическая информация (например, требование согласования или какого-то другого соответствия двух словоформ) и информация об универсальных условиях правильности синтаксических структур. Рассмотрим в качестве примера предложения типа *Вы сделали ее счастливой* с фактитивной связкой ДЕЛАТЬ. Они имеют следующую поверхностно-синтаксическую структуру:



Точно таким же образом могут употребляться и многие другие фактитивные связки (ср.: ДЕРЖАТЬ, ИЗБИРАТЬ, НАЗНАЧАТЬ, ОСТАВЛЯТЬ), а также путативные, декларативные, дескриптивные и перцептивные связки (ср.: НАХОДИТЬ, ПРИЗНАВАТЬ, СЧИТАТЬ; НАЗЫВАТЬ, ПРОВОЗГЛАШАТЬ, ОБЪЯВЛЯТЬ; ИЗОБРАЖАТЬ, ОПИСЫВАТЬ, ПОКАЗЫВАТЬ, ПРЕДСТАВЛЯТЬ, РИСОВАТЬ; ЧУВСТВОВАТЬ, ОЩУЩАТЬ).

Не будем перечислять всех правил, необходимых для установления этой структуры: они достаточно просты. Упомянем лишь одно из них. Оно находит второе комплетивное зависимое глагола в случае, когда последнее выражено прилагательным в положительной или превосходной степени и требует проверки согласования прилагательного в роде и числе с именной группой, называющей объект действия; ср.: *сделать ее счастливой* [*его несчастным, это невозможным, их послушными*].

Обозначим через X вершину рассматриваемой комплетивной конструкции, т. е. глагол; через Y — ее зависимый член, т. е. прилагательное; через Z — тот актант глагола (прямое дополнение или подлежащее), с которым согласуется Y. Тогда, с некоторыми непринципиальными упрощениями, формальное правило установления второго комплетивного зависимого глагола примет следующий вид:

Условия.

1. =(X, СОГЛАКТ-3-2) **и** VAL(3, X, ТВОР) **и** =(Y, A, ТВОР)

2. =(X, ДЕЙСТВ) **и** DOM (X, Z, 1-КОМПЛ)/=(X, СТРАД) **и** DOM (X, Z, ПРЕДИК)

3. CONMB (Y, Z) и COGEN (Y, Z)/=(Z, ЕД) и DOM (Z, W, СОЧИН)

Действие.

1. СВЯЗАТЬ: (X, Y, 2-КОМПЛ)

Комментарии к условиям 1—3.

В условии 1 говорится, что 1) X должен обладать синтаксическим признаком согласования третьего актанта со вторым (СОГЛАКТ-3-2); ср. глагол ДЕЛАТЬ, который им обладает, и квазисинонимичный ему глагол ПРЕВРАЩАТЬ, у которого его нет (ср.: * *Головокружительная карьера превратила его в самоуверенного*); 2) третья валентность этого глагола выражается именной группой в творительном падеже; 3) в предложении должно быть прилагательное в творительном падеже.

Условие 2 имеет целью найти ту именную группу *Z*, с которой согласуется прилагательное *Y*. Если глагол *X* употреблен в действительном залоге, то *Z* является его первым комплетивным зависимым (в данном случае — прямым дополнением). Если глагол употреблен в страдательном залоге, то *Z* является его предикативным зависимым, т. е. подлежащим (ср.: *Чаадаев был высочайше объявлен сумасшедшим*).

В условии 3 проверяется согласование прилагательного *Y* с именной группой *Z*. Здесь требуется совпадение числа и рода *Y-а* и *Z-а*. Отсутствие такого совпадения допускается только в том случае, когда *Z* в форме единственного числа открывает собой сочиненную группу. В этом случае по весьма общему правилу русской грамматики прилагательное *Y* имеет форму множественного числа, ср.: *Вы сделали свою жену и сына несчастными.*

Если все перечисленные условия удовлетворены, то выполняется единственное действие правила — *X* связывается с *Y*-м второй комплетивной зависимостью.

Как легко видеть, в этом правиле явно используются следующие три типа информации: 1) текстовая (ср. характеристику падежа у прилагательного, характеристику залога у глагола, характеристику числа у существительного, требование первого комплетивного или предикативного зависимого у глагола); 2) словарная (ср. синтаксический признак СОГЛАКТ-3-2 и способ выражения третьей валентности глагола; и то и другое приписывается глаголу непосредственно в его словарной статье); 3) грамматическая (ср. требование согласования прилагательного *Y* с именной группой *Z* в числе и роде). Кроме того, в нем имеется имплицитная ссылка на лингвистическую универсалию — закон синтаксической проективности большинства правильных предложений. Если определенная синтаксическая конструкция допускает нарушение проективности, этот факт эксплицитно оговаривается в соответствующей синтагме.

Рассмотренное правило достаточно рельефно демонстрирует меру взаимодействия синтаксического компонента модели с ее

морфологическим компонентом, со словарем, с другими правилами грамматики и с лингвистическими универсалиями.

Такова в самых общих чертах формальная модель русского языка, входящая в лингвистическое обеспечение системы ЭТАП-I.

Я хотел бы подчеркнуть, что считаю преждевременной канонизацию какой бы то ни было лингвистической модели в качестве единственно возможной основы прикладных МФ. Нельзя, разумеется, канонизировать и только что рассмотренную модель. Однако одно ее свойство — а оно является неотъемлемой принадлежностью всякой действующей, т. е. допускающей реализацию на ЭВМ модели языка, — представляет общий интерес и должно учитываться во всех работах по МФ. Это свойство состоит в том, что комбинаторный словарь, морфология и синтаксис образуют в совокупности единое описание русского языка, т. е. настроены друг на друга и способны к информационному взаимодействию.

Концепция единого описания языка, сложившаяся в рамках идеологии лингвистического моделирования и на основе опыта эксплуатации формальных моделей, представляется естественной теоретической основой прикладных МФ.

Эта концепция открывает новые перспективы и перед культурными МФ. Действительно, нельзя сомневаться в том, что рано или поздно между прикладными и культурными МФ, если им суждено реализоваться, начнется нормальный информационный обмен. Прикладные МФ будут обогащаться за счет более обширных лексикографических материалов культурных МФ. Культурные МФ, в свою очередь, должны будут постепенно утратить свою преимущественно лексикографическую ориентацию и начать заимствовать из прикладных МФ идеологию и форматы описания грамматической информации. Это, несомненно, окажет глубокое обратное воздействие на принципы обработки и способы хранения информации в культурных МФ, сделает их более гибкими и расширит их функциональный диапазон. В связи со сказанным полезно уже сейчас уяснить некоторые следствия из идеологии единого описания языка для культурных МФ. Рассматривая этот вопрос, мы, естественно, должны будем использовать более широкий круг языковых фактов и лингвистических понятий, чем те, которые фигурируют в рамках формальной модели.

Будем называть описание языка единым, или интегральным, если грамматика (морфология, синтаксис) и словарь способны взаимодействовать друг с другом, т. е. согласованы 1) по типам помещаемой в них лингвистической информации, 2) по способу, или формальным языкам, ее записи.

На первый взгляд требование интегральности описания языка может показаться тривиальным. В самом деле, каким же еще оно может быть? Лишь обращение к существующим лингвистическим описаниям позволяет понять, насколько это требование,

при всей своей самоочевидности, трудно осуществимо на практике и, следовательно, нетривиально. Ведь до самого последнего времени единых описаний языка в лингвистике не существовало. Более того, даже теоретически необходимость такого описания не была ясно осознана.

Выработать концепцию единого описания языка и создать практические описания, которые ее реализуют, оказалось непросто по двум причинам. Одна из них — психологическая. Дело в том, что грамматисты и лексикографы работают раздельно. Одни пишут грамматики и не пишут словари, а другие пишут словари и не пишут грамматики. Между тем, для того чтобы создать единое описание языка, надо работать на всем его пространстве. Вторая причина — информационная: до последнего времени отсутствовали многие типы сведений, необходимые для единого описания языка, и совсем не существовало формальных языков для их записи.

Рассмотрим сначала вопрос о типах сведений, по которым грамматика и словарь должны быть согласованы друг с другом, а затем значительно более коротко — вопрос о формальных языках их записи.

Даже на основании относительно несложного правила анализа синтаксических конструкций типа *Вы сделали ее счастливой* можно было видеть, что грамматические правила опираются на самые разнообразные типы языковой информации, извлекаемой из самых различных источников, — словаря, текстов, грамматики, лингвистических универсалий. К этому следовало бы добавить внеязыковую информацию, в частности, знания о мире и о текущей ситуации. Однако в контексте единого описания языка особый интерес представляет только один тип информации — собственно языковые сведения о лексических единицах— и только один источник информации — словарь.

Собственно языковая информация, подлежащая включению в словарь, тоже, конечно, достаточно разнообразна. Лингвистические правила могут требовать обращения к коммуникативным, прагматическим, стилистическим, референционным, семантическим, лексическим, синтаксическим, сочетаемостным, деривационным, морфологическим и просодическим свойствам лексем.

Так, ДАЖЕ, ТОЛЬКО и многие другие выделительные и усилительные частицы всегда маркируют рему предложения — ту синтаксическую группу, слева от которой они стоят. В предложениях *Даже мои дети едят кашу, Кашу едят даже мои дети,* независимо от порядка слов и интонирования, рему образует группа подлежащего. Точно такой же вклад в коммуникативную структуру предложения вносят усилительные прилагательные и наречия типа КАКОЙ, КАК и т. п.: *Какой тихий был вечер*!, *Как тихо было в лесу*! Быть рематизаторами — словарное, т. е. лексикографически фиксируемое свойство таких лексем.

Оно должно учитываться правилами установления коммуникативной организации предложения.

Правило формирования (или анализа) обстоятельственной конструкции типа *Грузчики молча переносили ящики на баржу*; *Ящики немедленно переносились на баржу* должно учитывать тот факт, что наречия типа МОЛЧА, ВЕСЕЛО, УГРЮМО ведут себя в этой синтаксической функции более избирательно, чем наречия типа НЕМЕДЛЕННО, БЫСТРО, ПОСТЕПЕННО. Наречия первой группы подчиняются следующему сочетаемостному референционному ограничению: их семантический субъект должен быть кореферентен с субъектом глагола, от которого они зависят. При этом, если у глагола есть эксплицитное подлежащее, оно должно обозначать именно субъект действия. Данный запрет объясняет неправильность предложений типа **Ящики молча* [*весело, угрюмо*] *переносились грузчиками на баржу*. В этом отношении рассматриваемые наречия похожи на деепричастия, но только для наречий данное свойство является словарным, т. е. должно регистрироваться в их лексикографическом описании, а для деепричастий оно является чисто грамматическим. Соответствующая информация помещается в зоне сочетаемостных референционных ограничений в виде фильтра, т. е. условия правильности синтаксических структур, в которых эти наречия выполняют функцию обстоятельства.

Аналогичный лексикографический фильтр регулирует употребление лексемы СОБСТВЕННЫЙ в определительных и присвязочных конструкциях: владелец, или субъект собственности, должен быть упомянут или подразумеваться в данном высказывании. Ср.: *Где же ваши собственные вещи?*; *У ребенка есть собственная авторучка; Эта машина — моя собственная*; *Пользоваться в больнице собственными лекарствами воспрещается* (подразумеваются пациенты больницы). Нарушение данного запрета порождает грамматическую неправильность: * *На столе лежала собственная сумка*; * *В камере хранения находились собственные вещи.*

Правила формирования условных и временных предложений предъявляют особые требования к лексикографической обработке глагола СТОИТЬ в том значении, в котором он квазисинонимичен союзам ЕСЛИ и КОГДА; ср.: *Стоит подняться на горку — и будет виден наш дом*; *Стоило ему выйти из дому, как пошел дождь*. В словарную статью соответствующей лексемы надо включить одно морфологическое и одно сочетаемостное ограничение. Морфологическое ограничение состоит в том, что данная лексема имеет неполную парадигму. Она употребляется только в настоящем или прошедшем времени. Сочетаемостное ограничение состоит в том, что она управляет инфинитивом только совершенного вида; ср. неправильность: * *Стоило подниматься на горку — и становился виден наш дом.*

Дать здесь сколько-нибудь представительный материал по всем типам лексикографически релевантной, но не отражаемой

в толковых словарях информации невозможно. Однако в интересах дела — как можно нагляднее продемонстрировать существующий разрыв между грамматической и лексикографической обработкой языковых фактов. С этой целью мы вернемся к уже затронутой теме синтаксических признаков лексем и приведем еще несколько иллюстраций. Начнем с очевидных примеров, а затем перейдем к менее очевидным и, следовательно, лексикографически более интересным примерам.

1. Чтобы правильно строить или анализировать определительные словосочетания *высокий юнга, красивое дитя, моя зазноба, сущий тать, большая выпь, крепкий кофе, жесткое ложе,* надо, в частности, знать род соответствующих существительных. Если говорящий, например, иностранец, не знает рода существительных, ничто не помешает ему построить неправильные словосочетания * *высокая юнга*, * *красивая* [*красивый*] *дитя*, * *мой зазноба*, * *сущая тать*, * *большой выпь*, * *крепкое кофе*, * *жесткий ложе*. Равным образом ничто не помешает системе автоматического синтаксического анализа сформировать эти неправильные гипотезы, если в нее не будет введена информация о родовой принадлежности существительных. Как известно, общих правил, позволяющих стопроцентно вычислять род существительного по его внешней форме или значению, нет. Поэтому целесообразнее всего приписывать существительному информацию о роде непосредственно в словаре. Именно так и решается вопрос в толковых словарях русского языка. Тем самым по синтаксическому признаку рода существующие словари удовлетворяют требованиям единого описания языка.

2. Известно, что при установлении определительных связей приходится учитывать не только род, но и одушевленность существительного. Форма винительного падежа множественного числа прилагательного совпадает с формой родительного падежа, когда оно зависит от одушевленного существительного (*вижу белых слонов*), и с формой именительного падежа, когда оно зависит от неодушевленного существительного (*вижу белые камни*).

В русском синтаксисе имеется еще несколько конструкций, описание которых в виде эффективно действующих правил требует обращения к информации об одушевленности — неодушевленности существительных. Одна из них — конструкция с так называемым дательным посессивным, ср.: *смотреть тигру в глаза, гладить девочке волосы, стричь ребенку ногти* и т. п. Если не считать фразеологических единиц типа *смотреть смерти в глаза, заглянуть опасности в глаза*, позицию дательного падежа в таких конструкциях может занимать лишь одушевленное существительное; ср. неправильность словосочетаний типа * *заглянуть дому в окна* (надо: *заглянуть в окна дома*).

Третья синтаксическая конструкция русского языка, реагирующая на свойство одушевленности — неодушевленности существительных, — это конструкция с родительным падежом

подлежащего в отрицательных предложениях. Большинство экзистенциальных (бытийных) глаголов одинаково свободно сочетается и с неодушевленными, и с одушевленными подлежащими в родительном падеже; ср.: *Таких книг не существует* [*не бывает*] и *Таких людей не существует* [*не бывает*]. Между тем глаголы со значением положения в пространстве свободнее сочетаются с неодушевленными, чем с одушевленными, подлежащими в родительном падеже; ср.: *На деревьях уже не висело ни одного яблока* при невозможности: * *На заборе уже не висело ни одного мальчишки* (ср.: *Когда в саду сторож, мальчишки на заборе не висят*).

Итак, в русском синтаксисе имеется несколько участков, описание которых в виде строгих, т. е. буквальным образом применимых, правил требует обращения к информации об одушевленности—неодушевленности существительных. Однако удивительным образом ни один толковый словарь русского языка этой информации не сообщает. Она дается лишь в словаре А. А. Зализняка [4]. Между тем информация об одушевленности—неодушевленности существительного гораздо менее тривиальна, чем например, информация о его роде.

Действительно, эта информация совершенно невыводима из внешней формы существительного и гораздо менее выводима из смыслов «живое» — «неживое», чем информация о роде — из значения биологического пола. Хорошо известно, что имеются целые классы существительных, обозначающих неживые существа, которые синтаксически трактуются как одушевленные. Это — названия покойников (МЕРТВЕЦ, ПОКОЙНЫЙ, УДАВЛЕННИК, УТОПЛЕННИК и т. п.), тушек животных, употребляемых в пищу, кукол, роботов, шахматных фигур [5], карточных мастей. С другой стороны, имеются существительные, которые обозначают живых людей и тем не менее синтаксически трактуются как неодушевленные. Мне известен, по крайней мере, один бесспорный разряд существительных такого рода. Это — метонимическое название человека по его части, органу или свойству: *верная рука, важное лицо, беспокойные сердца, великие умы.* Мы говорим: *Я знаю такие беспокойные сердца — вечно они ищут романтики.* Нельзя сказать: * *Я знаю таких беспокойных сердец.* Однако даже в этой узкой группе есть исключения; ср. слово ЯЗЫК в словосочетании *брать языка.*

Таким образом, по синтаксическим признакам одушевленности и неодушевленности существительных ни один толковый словарь русского языка не удовлетворяет принципам единого описания, т. е. не может служить лексикографической основой эффективно действующих грамматических правил.

---

[4] *Зализняк А. А.* Грамматический словарь русского языка: Словоизменение. М., 1977.

[5] Любопытно, что одушевленными являются и субстантивированные прилагательные БЕЛЫЕ и ЧЕРНЫЕ, ср.: *лучший ход за белых* ⟨*за черных*⟩.

3. В предыдущем пункте мы уже упоминали хорошо известный в русской грамматике феномен преобразования именительного падежа подлежащего в родительный в контексте отрицательного предложения, ср.: *Такие сведения к нам уже поступали — Таких сведений к нам еще не поступало.* Не будем сейчас задерживаться на том, почему имя в родительном падеже считается подлежащим; можно было бы в соответствии с господствующей традицией считать такие предложения безличными — на ход дальнейших рассуждений это никак не влияет.

Очевидно, что не все глаголы допускают такую мену. Глагол ЧИТАТЬ, например, ее не допускает: нельзя сказать: * *Мальчика не читало*, можно только *Мальчик не читал.*

Итак, все глаголы русского языка разбиваются относительно данного явления на два класса. Одни допускают мену именительного падежа подлежащего на родительный в контексте отрицания, другие — нет. Система синтаксических правил, будь то автоматические правила или правила, предназначенные для человека, не может быть эффективной, если она не располагает информацией о том, к которому из двух классов принадлежит данный глагол. Эта информация должна быть помещена либо в грамматике, либо в словаре.

Посмотрим, можно ли ее поместить в грамматике. Очевидно, что перечислить в грамматике все русские глаголы, обладающие рассматриваемым свойством, невозможно, потому что их слишком много. Но, может быть, в этом и нет необходимости. Может быть, глаголы, допускающие мену именительного падежа подлежащего на родительный в контексте отрицания, обладают общим семантическим признаком. Тогда данное синтаксическое свойство становится вычислимым непосредственно по словарному толкованию глагольной лексемы.

Традиционно считается, что у всех таких глаголов действительно есть семантический инвариант — значение существования. Легко, однако, назвать два класса глаголов, которые нарушают это идеальное соответствие «существование»=«мена именительного падежа на родительный в контексте отрицания».

С одной стороны, имеются глаголы, которые явным образом не являются экзистенциальными и тем не менее допускают такую мену. Это, например, модальные глаголы типа ПОЛАГАТЬСЯ, ПРИЧИТАТЬСЯ, ТРЕБОВАТЬСЯ (ср.: *Ребенку отдельного места не полагается* [*не требуется*]) и глаголы со значением количества истекшего времени типа ИСТЕЧЬ, МИНОВАТЬ, МИНУТЬ, ПРОЙТИ (ср.: *Года не минуло* ([*не прошло*], *как он снова заболел*).

С другой стороны, имеются различные экзистенциальные глаголы, особенно фазовые глаголы со значением начала существования чего-либо, которые, вопреки ожиданиям, не допускают мены именительного падежа подлежащего на родительный в контексте отрицания. При этом удивительным образом для всякого такого глагола можно подыскать достаточно близкий си-

нонимом, который эту мену допускает. Возьмем, например, пару НАЧАТЬСЯ — НАСТУПИТЬ. Нельзя сказать: * *Перестановок в руководстве еще не началось* (надо: *Перестановки в руководстве еще не начались*). Между тем с глаголом НАСТУПИТЬ в том же инцептивно-экзистенциальном значении оба варианта оформления подлежащего грамматически равно правильны (хотя семантически и коммуникативно могут различаться). Ср.: *Улучшения в состоянии больного все еще не наступило — Улучшение в состоянии больного все еще не наступило.* Интересно, что финитно-экзистенциальные и континуативно-экзистенциальные глаголы вообще, по-видимому, не допускают мены именительного на родительный в контексте отрицания. Предложения типа * *Симптомов заболевания не проходило* [*не исчезало*], * *Дождя не кончалось* [*не прекращалось*], * *Наступления не продолжалось* неправильны.

Неверны и все другие предположения о семантическом инварианте глаголов, обладающих рассматриваемым синтаксическим свойством. Оно лишь частично мотивировано семантически, а в значительной мере идиосинкратично. Рациональное место его фиксации не грамматика, а словарь: оно должно отмечаться индивидуально для каждой глагольной лексемы в качестве ее синтаксического признака.

Излишне говорить, что не существует такого словаря русского языка, не только толкового, но и грамматического, который приписывал бы глагольным лексемам этот синтаксический признак. Следовательно, в рассматриваемой точке системы русского языка ни один словарь не может эффективно взаимодействовать с грамматикой.

4. Рассмотрим, наконец, лексикографические предпосылки для полного и строгого синтаксического описания атрибутивной конструкции типа *человек широких взглядов, женщина редкой красоты, ребенок удивительных способностей.* Отмечаемое во всех руководствах отличие данной атрибутивной конструкции от канонической конструкции с атрибутом в родительном падеже (*дом отца*) заключается в том, что в ней атрибут обязательно имеет собственное зависимое, и притом в форме согласованного определения. Нельзя сказать * *человек взглядов*, где атрибут не имеет своего зависимого, и * *ребенок способностей к музыке*, где зависимое налицо, но не в форме согласованного определения.

Более интересно и менее известно второе свойство этой конструкции, непосредственно связанное с понятием синтаксического признака: в роли атрибута в ней могут выступать лишь некоторые существительные — названия физических или духовных характеристик личности. Не обладает этим свойством большинство существительных, обозначающих отрицательно оцениваемые обществом характеристики личности. Словосочетания * *человек исключительной бедности*, * *женщина редкого безобразия*, * *мужчина большой трусости*, * *мальчик исключительной грубо-*

сти сомнительны или неправильны. Более того, это свойство присуще далеко не всем существительным, обозначающим положительно оцениваемые характеристики личности. Его нет, например, у существительных АВТОРИТЕТ, ЛЮБЕЗНОСТЬ, НЕЗАВИСИМОСТЬ и ряда других; ср.: сомнительность или невозможность * *человек большого авторитета* [* *большой любезности*], * *девушка исключительной независимости.* Следовательно, способность формировать указанную синтаксическую конструкцию — типичный синтаксический признак, который следует приписывать лексемам в словаре.

Третье интересное свойство этой конструкции состоит в том, что ее вершиной в неспециальных условиях может быть лишь слово с достаточно общим значением лица (ЧЕЛОВЕК, МУЖЧИНА, ЖЕНЩИНА, ДЕВУШКА и т. п.), не являющееся к тому же именем отношения. Нельзя сказать: * *чернорабочий большой физической силы,* * *профессор большого ума,* * *отец широкой культуры,* * *родственник европейского образования.* Словосочетания такого типа становятся правильными только при описании ситуации, когда лицо, обозначенное существительным со специализированным значением (но все-таки не со значением отношения), действует в привычной для себя профессиональной обстановке; ср.: *В пассажирский салон вошла стюардесса исключительной красоты* при невозможности: * *По берегу моря гуляла стюардесса исключительной красоты* (последним наблюдением я обязан Е. М. Вольф). По-видимому, указанный запрет объясняется тенденцией избегать высказываний с двумя ремами одного порядка [6].

Особый статус таких существительных в русском синтаксисе, обнаруживающий себя по крайней мере в двух независимых точках, должен быть отмечен специальным синтаксическим признаком.

Если словам в словаре не приписаны явным образом сведения об их роде, одушевленности, способности менять именительный падеж подлежащего на родительный в контексте отрицания и т. п., они не могут эффективно участвовать в грамматических правилах. В эффективно действующем описании языка каждой лексеме должны быть эксплицитно приписаны все типы сведений, необходимые для того, чтобы она могла стать объектом грамматических правил.

---

[6] Любопытно, что рассматриваемая атрибутивная конструкция — не единственная, для которой оказывается необходимым выделять существительные с общим значением лица. Такова же и обособленная конструкция типа *Человек, никогда ничего не боявшийся, он и сейчас готов был рискнуть.* Вершиной этой конструкции могут быть лишь достаточно неспециализированные названия человека, не являющиеся к тому же именами отношений; ср неправильные фразы: * *Долина* [X] *богатая теплом и влагой, она давно славилась своими виноградниками* (X — не название человека) и **Мать* [X] *жившая только интересами детей, она работала не покладая рук, чтобы дать им образование* (X — название отношения).

Чтобы представить сказанное в правильной перспективе, необходимо принять во внимание следующие три факта.

Факт первый. Мы рассмотрели всего семь синтаксических признаков русских лексем и показали, что эти признаки должны приписываться лексемам в словаре в рамках любой эффективно действующей системы синтаксических правил. Но мы используем в формальной модели русского синтаксиса, о которой я говорил выше, больше ста таких признаков. Подчеркнем, что это число не более чем показатель порядка синтаксических признаков, которые необходимо фиксировать в словаре, чтобы он мог взаимодействовать с грамматикой, а отнюдь не окончательная истина о том, сколько их на самом деле. При более широком исследовании синтаксических свойств лексем на материале русского языка вообще, а не только научно-технической прозы или делового стиля, может возникнуть нужда в новых синтаксических признаках.

Факт второй. Синтаксические признаки — лишь один из типов забытой, т. е. не отражаемой в словарях лексикографической информации о синтаксических свойствах лексем. Другие типы забытой синтаксической лексикографической информации мы уже называли. Это — модели управления, синтаксически релевантные лексемы, сочетаемостные синтаксические ограничения и некоторые другие.

Факт третий. Синтаксическая информация — лишь один из забытых типов лексикографической информации о свойствах лексем, которая может оказаться существенной для успешного функционирования грамматических правил. Без труда можно было бы назвать не менее пяти других типов забытой, но лексикографически весьма важной информации (см. ранее приведенные примеры и в особенности даваемый ниже ориентировочный перечень зон стандартного формата описания лексемы в толковом словаре). Особенно существенна в этом отношении информация о сочетаемостных ограничениях — прагматических, коммуникативных, референционных, семантических, лексических, синтаксических, морфологических, стилистических, просодических, которые накладываются на контекст данной лексемы, т. е. на другие лексемы или конструкции, с которыми она сочетается.

Итак, чтобы эффективно взаимодействовать с грамматикой в рамках единого описания языка, толковый словарь должен включать многие типы информации, традиционно остававшиеся за его пределами. В их число входят и совершенно новые типы информации, открытые и введенные в оборот в самое последнее время.

Перейдем ко второму вопросу, связанному с концепцией единого описания языка, — о способах записи этой информации. Никакая формальная система, тем более такая сложная, как МФ русского языка, не может работать в условиях разнобоя. Принцип единого описания языка требует, чтобы однотипная

информация записывалась в одних и тех же местах и всякий раз одинаково, независимо от того, в каком компоненте описания— морфологическом, синтаксическом, словарном или каком-либо ином — она фигурирует. В более технических терминах это сводится к двум условиям. Одно накладывается на форму представления лингвистической информации в словарях и правилах МФ, другое — на язык ее записи.

Формальная система требует унификации внешней формы представления лингвистической информации, т. е. навязывает лингвистам определенный жесткий формат ее описания и в словаре, и в грамматике.

Из этого вытекает прежде всего зонный принцип организации лексикографической информации. Пространство словарной статьи разбивается на зоны, причем каждая зона вводится своей меткой. Приблизительный состав лингвистических зон для словаря как части единого описания языка с учетом известных в настоящее время фактов и нужд мог бы выглядеть следующим образом [7].

1. Фонетические и орфографические сведения.
   1.1. Имя лексемы в орфографической записи.
   1.2. Фонетическая транскрипция.
   1.3. Просодические сведения.
2. Морфологические сведения.
   2.1. Тип словоизменения.
   2.2. Неполнота парадигмы и другие ограничения.
   2.3. Часть речи.
   2.4. Сведения о морфологической и деривационной структуре.
3. Синтактика.
   3.1. Модель управления.
   3.2. Синтаксические признаки.
   3.3. Сочетаемостные ограничения (семантические, референционные, коммуникативные, прагматические, стилистические, лексические, синтаксические, морфологические, просодические).
   3.4. Стандартные лексические функции-параметры.
   3.5. Перечень фразем, содержащих данную лексему.
   3.6. Порядок слов.
4. Семантика.
   4.1. Семантические стилистические пометы.
   4.2. Нетривиальные семантические признаки.
   4.3. Толкование.
   4.4. Другие семантические правила (зачеркивание, модификация, области действия).

---

[7] Приводимая здесь схема учитывает следующие аспекты языкового знака: 1) означающее, 2) морфологическую структуру, 3) синтактику, 4) семантику, 5) прагматику, 6) коммуникативные свойства. Более подробно она изложена в работе автора «Интегральное описание языка и толковый словарь» (ВЯ, 1986, № 2).

4.5. Референционные и анафорические сведения.
4.6. Энциклопедические сведения.
4.7. Контекстная (конситуативная) информация.
4.8. Сведения о семантической производности.
4.9. Парадигматические семантические отношения (синонимы, конверсивы, антонимы, гиперонимы, гипонимы, аналоги, семантические производные).

5. Прагматика.
5.1. Прагматические стилистические пометы.
5.2. Коннотации, культурный и образный мир лексемы.
5.3. Статусы говорящего и адресата.
5.4. Иллокутивные функции.
5.5. Прагматические признаки.

6. Коммуникативные сведения.
6.1. Функция тематизации.
6.2. Функция рематизации.

7. Иллюстрации (в дополнение к примерам по каждой зоне и подзоне словарной статьи).

Логически различаются два основных типа лексикографической информации. Содержанием одних зон словарной статьи являются объекты языка, или модели,— части речи, синтаксические признаки, стилистические пометы. Содержанием других зон являются правила, т. е. предписания о том, в какой объект должен быть переработан данный объект при той или иной совокупности условий. Казалось бы, правила должны помещаться в грамматике или другой общей части лингвистического описания. Однако в языке есть много правил, касающихся не бо́льших классов слов, а отдельных лексем, и естественное место фиксации таких частных правил — словарь. Типичным примером словарного правила является толкование.

Идея форматизации, или зонной организации лингвистических сведений, важна не только для словаря. Следует выработать определенный формат и для записи правил. Каждое правило должно включать минимум следующие три зоны: зону проверки условий; зону действий; неформальную зону комментариев и иллюстраций (см. приведенное выше правило синтаксического анализа конструкций типа *Вы сделали ее счастливой*).

Итак, вся лингвистическая информация в словаре и в правилах записывается в стандартных форматах, имеющих фиксированное количество зон. Каждая зона открывается своей меткой, по которой она опознается алгоритмами.

Второй и, видимо, более важный уровень формализации лингвистических описаний (вообще и в рамках МФ) — это формализация языков описания. На них накладывается следующее общее требование: язык записи однотипной лингвистической информации должен быть единым для всех компонентов лингвистического описания. Это, в частности, значит, что каждому языковому объекту, фигурирующему в массивах МФ, должно

соответствовать ровно одно имя, а каждому имени — ровно один объект.

Рассмотрим простой пример из области морфологии и морфологической семантики. Конечными объектами морфологии являются морфологические категории и граммемы, т. е. значения морфологических категорий. Если какая-то категория в морфологическом компоненте названа «видом», то во всех остальных частях грамматики и в словаре она должна фигурировать под этим именем и не имеет права называться «аспектом». Если в каком-то компоненте единого списания языка два значения этой категории обозначаются именами СОВ и НЕСОВ, то именно эти сокращения, а не СВ и НВ, должны фигурировать и во всех остальных компонентах описания. Если основное значение граммемы СОВ описывается как «результативное» (ср., с другой стороны, потенциальное, узуальное, наглядно-примерное и другие частно-видовые значения СОВ), то надо отказаться от всех синонимичных или квазисинонимичных терминов — перфективное, предельное, точечное, целостное, тотальное. По справедливому замечанию Э. Кошмидера, они всего лишь переназывают одно и то же семантически недостаточно эксплицированное понятие.

Если этого не сделать, формальная передача морфологической информации из одних компонентов в другие (скажем, из морфологического или словарного в синтаксический) станет невозможной (если, конечно, не использовать никаких дополнительных средств трансляции): разные компоненты описания утратят способность взаимодействовать друг с другом.

Уже здесь возникает серьезный, отнюдь не только технический вопрос об унификации лингвистической терминологии. Этот вопрос перерастает в труднейшую проблему при переходе от морфологии к синтаксису, где терминологический разнобой возрастает на порядок, а тем более к семантике, где вообще не существует никакого хотя бы отдаленно общепринятого семантического метаязыка и, следовательно, никакой устоявшейся терминологической традиции.

Терминологический разнобой в науке всегда является свидетельством отсутствия проверенной системы теоретических понятий. В этом отношении машинные фонды, требующие безусловной унификации не только терминологии, но и всех лингвистических метаязыков в целом, могли бы оказать колоссальное воздействие на упорядочение языка лингвистики.

ОБ ИНФОРМАЦИОННОЙ СИСТЕМЕ СЛОВАРЕЙ МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА

Н. Н. ЛЕОНТЬЕВА

Фонд русского языка как информационная система

Машинный фонд русского языка в целом — это большая, сложная и неоднородная информационная система (ИС), состоящая из словарей и текстов. Неоднородной ее делает массив документов, различающихся типом, стилем, жанром, функциональным использованием и т. д. каждого документа. Из видов документов, которые предполагается помещать в Фонд русского языка (РЯ), достаточно назвать основные типы: Словари, Грамматики, Тексты, причем тексты делятся на лингвистические тексты (являющиеся продолжением грамматик) и художественные тексты. Последнюю категорию можно отнести к «культурному» фонду, а все остальные — считать входящими в «научный» фонд русского языка, при этом документы «культурного» фонда являются иллюстрацией тех явлений, которые описаны источниками «научного» фонда.

Свободную ИС как объединение лингвистических источников, в первую очередь словарей, можно строить уже сейчас и не в модельном, а в реальном масштабе: начальный информационный массив образуют материалы, которые перенесены или переносятся на машинные носители. Словари, существующие только в традиционной форме, могут быть представлены в ЭВМ в «рекламном» виде; при этом они должны сопровождаться описанием основных параметров: назначение словаря, его объем, тип входных единиц, тип информации к словам, форма существования словаря (на магнитной ленте, в традиционной картотеке и т. д.), местонахождение словаря (в секторе структурной поэтики ИРЯ, в МГУ, в ВЦП и т. д.), примеры словарных вокабул каждого словаря.

При управлении такой системой важнее всего смоделировать динамический характер информационного массива: многие рекламные сведения о «живых» разработках быстро изменяются (словарь пополняется, передается для использования в разных системах, переносится на машинные носители и т. д.). Словари, имеющие наибольший спрос (например, «Грамматический словарь русского языка» А. А. Зализняка), желательно ввести в ЭВМ как можно скорее, другие могут вводиться постепенно; словари, имеющие узко специальный, «ведомственный» характер, вообще не нужно вводить в общий Фонд, они могут храниться «на местах». К этой категории мы бы отнесли разные словари отраслевых ИПС, тезаурусы и терминологические словари, двуязычные словари систем машинного перевода и другие словари

специального назначения. Но сведения о наличии таких словарей и их основных параметрах чрезвычайно нужны разным разработчикам: учет всех словарных работ и хотя бы слабый контроль служб РЯ поможет, во-первых, снизить дублирование в трудоемкой области создания словарей, во-вторых, поднять профессиональный уровень специальных словарей, составляемых подчас нелингвистами. Учет реальной информационной потребности в разных словарях поможет планировать создание или компиляцию новых словарей на основе имеющихся.

Введенный таким образом фонд словарей будет рассматриваться в одном ряду с другими источниками информации (например, текстами Академической грамматики и других грамматик РЯ, а также естественными художественными текстами) как необходимый компонент информационно-поисковой системы (ИПС) Фонда.

Функционирующая ИПС применительно к Фонду в целом должна осуществлять хранение, извлечение, обработку и выдачу информации. Основными «актантами» этого процесса являются:

A — источники информации (документы, массивы);
B — состав, виды, содержание информации (в ответ на запрос);
C — инструмент (средство) извлечения информации;
D — адресат информации (или цель, адресация информации).

Запросы к ИПС Фонда могут быть разного свойства: справочного характера (например: «Сколько словарей и какие конкретно могут быть использованы для автоматического перевода?»); документографического характера (например: «В каких источниках Фонда есть информация об инфинитивных конструкциях РЯ?») и фактографического (например: «Что говорится в Академической грамматике РЯ об инфинитивных конструкциях с модальными глаголами?»). В Фонде РЯ как ИПС можно выделить три стандартные ситуации, различающиеся комбинациями актантов ИПС.

В первой ситуации источником информации (актант A) будут специальные лингвистические тексты (грамматики, словари разных видов, научные статьи по лингвистике, справочники, монографии и т. д.). Выдаваемой информацией (актант B) может быть фактографическая информация в виде изолированных фактов (например, в ответ на запрос «Выдать парадигмы или всю словарную статью слов, оканчивающихся на *-ция*») или в виде систематизированных фактов («Выдать информацию о значениях инфинитивов в русском языке, по всем источникам, в неизбыточном виде»); информационная цитата («Выдать весь словарь или всю грамматику определенного автора или определенной системы»); документальная информация («Перечислить все источники, в которых говорится о простом предложении или о семантике личных местоимений»). Адресатом информации (D) в первой ситуации могут быть специалисты-лингвисты, использующие информацию для построения новых лингвистических текстов, и спе-

циалисты-нелингвисты, использующие полученную информацию для включения в ИПС по своей специальности.

Во второй ситуации:

A — специальные «нелингвистические» тексты («деловая» проза);
B — информация таких же типов, как и в первой ситуации;
D — специалисты разных научно-технических областей.

В третьей ситуации:

A — художественные тексты;
B — аналогично в 1-й и 2-й ситуациям (например, фактографическая информация: «Собрать из текстов былин все ситуации, где Илья Муромец сидит на коне или печи»; документальная информация: «Выдать все русские былины, в которых упоминается Соловей-разбойник» или «Перечислить все сказки, где участвует баба-Яга»; цитатная информация не требует пояснений);
D — специалисты-литературоведы, историографы, текстологи и др.

Актант C (инструмент анализа) во всех трех ситуациях — словари, с помощью которых происходит перевод сведений, содержащихся в документах на естественном языке (ЕЯ), в такую форму, в которой выдается информация: сначала в единицы принятого в системе информационного языка (ИЯ), затем в форму текста информации. В системах, где ИЯ задается лексически, списком, как множество информативных единиц (дескрипторов, терминов), т. е. понятия ИЯ и словаря совпадают, ИЯ является и средством поиска, и средством фиксации найденной информации. Но поскольку они могут и расходиться, будем считать словарь типичным средством поиска, обнаружения информации (его можно также назвать в этой функции поисковым языком — ПЯ), а ИЯ — типичным средством фиксации найденной информации.

О форме представления источников информации. Словари как источник информации

Можно предложить два способа хранения информации в Фонде; они соответствуют двум путям моделирования понимания текста.

Первый путь состоит в признании ценности естественного объекта — текста. Задача состоит в создании системы средств, обеспечивающих построение результата «понимания» для произвольного естественного текста (ЕТ), без его «настройки» на средства анализа или на желаемый результат.

Второй путь требует препарирования исходного материала либо отбора узкого подмножества документов, которые заведомо укладываются в схему результата. Это путь превращения естественного объекта в неестественный. Таким путем создается большинство систем искусственного интеллекта, принимающих на

входе тексты, состоящие только из грамматически правильных предложений, без местоимений, без эллипсисов, без придаточных и т. п., либо тексты на так называемом ограниченном естественном языке (ОЕЯ), который оказывается ограниченным не структурно, а тематически: входные тексты допускают лишь небольшие отклонения от заданного сюжета, на который настроена система.

Источники «культурного» фонда — художественные тексты во всем их многообразии — обладают безусловной самостоятельной ценностью. Если формальная модель «не берет» этот материал, не объясняет в нем ничего из того, что составляет суть и специфику данного источника по сравнению с другими подобными объектами, то «виноватой» следует признать модель, а не анализируемый материал. Пока модели анализа художественных текстов могут объяснить не очень много, но даже небольшие, объективно добытые знания здесь очень важны (таковы, например, «естественные» классы А. Я. Шайкевича, полученные методом дистрибутивно-статистического анализа, или разработанная И. П. Севбо стилистическая диагностика художественных текстов на основе «прочтения» рисунков деревьев зависимостей).

В случае с художественными произведениями очевидно, что нужно идти по первому пути, т. е. без изменения анализируемого объекта совершенствовать средства его анализа (грамматики, словари, сам ИЯ как средство фиксации информации, добытой из ЕТ) и автоматические процедуры, определяющие порядок обращения к имеющимся в системе средствам, чтобы строить все более информативный результат анализа.

Рассмотрим ситуацию, когда источником информации являются материалы научной части Фонда. Выделим из всех материалов Фонда словари как источник информации. Это специфически лингвистический материал, подобный специальным таблицам или особым, хорошо структурированным текстам некоторых предметных областей (математика, химия). Казалось бы, извлечение информации из этого источника должно моделироваться по второму пути, т. е. на вход в ЭВМ должны допускаться словарные статьи в каком-то одном, заранее разработанном шаблоне. Однако разработка единого формата описания словарных данных сама по себе является грандиозной задачей, эквивалентной построению единого метаязыка словарных описаний, который составит часть общего информационного языка для лингвистики. Это легко понять, обратившись к автоматическим словарям (АС), создаваемым в разных системах автоматической обработки текстовой информации. По самой своей сути АС должны быть готовым организованным материалом, имеющим четкую структуру: имена разделов словаря являются теми вопросами, которые могут быть заданы АС как источнику информации, а содержимое разделов должно быть готовыми ответами. Но в этой функции АС представляют еще большую трудность для извлечения из них отдельных сведений, чем обычные словари. Так, даже если взять

несколько АС из систем, имеющих одно назначение и заведомо содержащих сведения о синтаксической модели управления (МУ) слов, поиск этой информации сопряжен с преодолением многих трудностей, таких, как:

а) разные именования искомого понятия (морфолого-синтаксическое управление; синтаксическая валентность; семантическая валентность; морфологический и синтаксический способ реализации семантической валентности; фрейм; обойма управлений; МУ и другие аббревиатуры; сборная, совмещенная МУ и др.);

б) трудность вычленения нужного значения слова, поскольку МУ является функцией от значения слова. Эта трудность представляется пока непреодолимой: деление на значения в разных словарях существенно разное, поскольку это самая творческая процедура в составлении словаря;

в) трудность вычленения запрашиваемой информации из состава словарных сведений: ведь сведения о МУ могут быть «размыты» по нескольким словарям системы, дающим описание той же единицы с разной степенью глубины (как узел синтаксической структуры, как единица семантической структуры, как составная часть термина, как ключевое слово словосочетания);

г) возможность распределения однородных сведений по разным компонентам системы: так, МУ некоторого слова может быть найдено не в словарном, а в грамматическом (трансформационном, переводном) компоненте системы автоматического перевода.

Большое количество разных, несводимых друг к другу типов словарных описаний, создаваемых в словарях разных систем, свидетельствует о том, что еще не найдено теоретическое решение проблемы АС. Не найден уровень квантования параметров словарных описаний такой, который удовлетворил бы и авторов словарей, и требования внутрисистемного описания, и внешние, поисковые требования к словарям. Поиск нужной информации в массиве разных АС как источнике информации подобен поиску фактографической информации в массиве текстов на разных языках и жаргонах, неравномерно структурированных, с большой степенью фразеологичности, с большой эллиптичностью и с большим дублированием.

Множество АС образует плохой информационный массив из-за «разноязычия» составляющих его источников. Чтобы сделать его хорошим информационным массивом, нужно привести все составляющие его источники к виду, допускающему сравнение с другими подобными документами по всем параметрам. Лишь тогда можно ставить задачу поиска какой-либо информации по всему массиву источников как целому. Такая экспликация содержимого АС средствами единого информационного эквивалента едва ли выполнима в обозримые сроки. Видимо, АС нужно пока рассматривать как неотъемлемую часть конкретной системы автоматического анализа текста и оценивать их по тем устойчивым результатам, которые обеспечивает включающая их система.

Если стандартизация АС вызывает такие трудности, то о стандартизации информации, содержащейся в традиционных словарях, тем более говорить пока не приходится, точно так же, как о стандартизации других источников научной части Фонда, помещаемых в ЭВМ (например, текста Академической грамматики). Это означает, что и все научные источники Фонда нужно рассматривать как исходные «тексты». Таким образом, и в ситуации с грамматиками и словарями мы вынуждены выбрать второй путь — помещение их в ЭВМ в своем «естественном виде» (т. е. признание неизменности, самоценности их как объектов анализа) и создание гибких инструментов анализа, эксплицирующего содержание этих документов.

Все сказанное вынуждает рассматривать информационную систему Фонда РЯ в категориях наиболее сложных систем автоматического понимания текстов (АПТ). Проблемы искусственного интеллекта (ИИ) встают здесь в натуральную величину. Реальный, а не модельный масштаб материалов Фонда не позволяет прибегать к таким стандартным упрощениям проблемы ИИ, как введение ОЕЯ для представления источников информации и узко специализированных словарей в качестве средств анализа. Язык во всем его богатстве — видов источников, жанров, стилей — это данность, из которой должна исходить любая автоматическая система (справочная, информационная, интеллектуальная), моделирующая работу с материалами Фонда РЯ.

Рассмотрение параметров проектируемого Фонда в категориях систем АПТ помогает понять сложность задачи и подсказывает возможные решения. Предлагается начинать эту работу в жанре исследования, имеющего целью постепенное превращение системы Фонда в кибернетическую информационную систему. Главным свойством кибернетической системы мы считаем ее способность «работать» с реальным естественным объектом, учитывая те изменения, которые претерпевает объект в процессе обработки. В нашем случае этими объектами являются и естественные тексты, и такой специфический источник информации, как словари.

Основные компоненты системы АПТ

В стандартных системах понимания текста (собственно, правильнее было бы говорить об экспериментальных моделях таких систем) на вход поступает текст на ограниченном естественном языке (ОЕЯ). В качестве средств его анализа (понимания) привлекаются, с одной стороны, лингвистические знания (ЛингвЗн) в виде словарей и грамматик, с другой стороны, специальные знания в выбранной предметной области (СпецЗн) в виде структуры семантической сети, в которой фиксированы шаблонные ситуации, основная схема той предметной области, к которой принадлежит и текст на ОЕЯ. Предметная область, как правило, ограничивается одним сюжетом (например, посещение зубного врача,

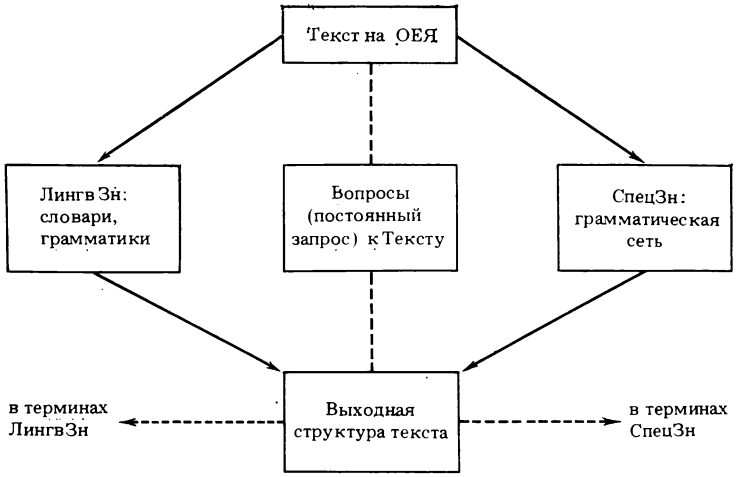


Рис. 1. Компоненты системы АПТ

посещение ресторана, крушение поезда, некоторая политическая ситуация и т. п., ср. многочисленные системы Р. Шенка). В ответ на конкретный вопрос к тексту или на постоянный информационный запрос (заданный той же схемой предметной области) строится выходная структура конкретного текста, которая и фиксирует результат «понимания» данного текста. Упрощенная схема компонентов системы АПТ приводится на рис. 1.

Характер системы, т. е. используемые инструменты анализа и тип выходных структур, часто определяется тем, кто был разработчиком — лингвист или специалист в выбранной предметной области. Основой лингвистического подхода является какая-либо синтаксическая модель, а схема предметной области используется стихийно, набегами, в виде фильтров, уменьшающих неоднозначность синтаксического анализа. Выходная структура в терминах ЛингвЗн представляет собой квазиреферат исходного текста, состоящий из предложений или частей предложений, содержащих нужные слова схемы предметной области. Для «специалиста» основным средством анализа является семантическая сеть, а морфолого-синтаксические признаки и структуры привлекаются в несистематичном виде: как фильтры, как правила локального контекстного разбора, задачей которого является заполнение оставленных в схеме предметной области пустых мест, т. е. «слотов» подходящих фреймов. Выходная структура в терминах СпецЗн — это фрагмент семантической сети с заполненными слотами, развертывание которого в текст может достигаться иными средствами, чем использованные в исходном тексте.

По сути дела в системах АПТ моделируется не столько понимание, сколько узнавание заданной заранее схемы сюжета, а понимание сводится к поиску в тексте значений некоторых пере-

менных (*поезд шел из Вены в Париж, в результате крушения погибло 100 человек, ранено 12* и т. д.). Понимание произвольного текста системы АПТ не моделируют.

Словари как средство анализа в реальных системах АПТ многоаспектные, сведения в них разносортные; существенной частью словарного описания оказываются эмпирические правила построения нужного фрагмента сети СпецЗн. Элементы СпецЗн образуют большую гамму переходов — от атомов смысла (есть словесные атомы, текстовые атомы, у Р. Шенка — спецатомы для каждой узкой проблематики) до самых крупных концептуальных единиц (гиперфреймы, планы, сценарии). Самой популярной единицей словаря в системах АПТ оказался «фрейм», так как это средняя единица и по величине, и по глубине, и по разнородности своих дифференциальных признаков. Слово в пределах одной словарной статьи выступает как представитель и своих морфологических вариантов, и синтаксических вариантов, и представитель соответствующего понятия, и элемент крупной ситуации.

Стандартным решением проблемы адекватного понимания в системах АПТ является «ранняя специализация» словарей и грамматик. Это значит, что слова с самого начала вводятся в систему только в тех значениях, которые ожидается встретить в данной узкой области, а грамматики настроены на узнавание лишь тех ситуаций, которые описаны в СпецЗн. Такое решение может привести к тому, что при выходе за пределы уже обработанных текстов система перестанет узнавать и свое, и чужое, т. е. и специальное, и общезначимое. Вывод грустный: чтобы угнаться за реальными текстами, нужны все новые и новые словари и грамматики, так что средств анализа окажется больше, чем текстов, которые «берутся» этими инструментами.

Фонд русского языка как система АПТ

Лингвистика в задаче построения фонда РЯ и последующего поиска в нем оказывается в ситуации тех специальных дисциплин (систем класса ИИ, АПТ, ИПС), которые должны создавать свой тезаурус, свою базу спецзаданий, процедуры анализа материала и словари как средство анализа.

Рассмотрим ситуацию, когда лингвисту (как адресату информации) нужно получить ответы на определенные вопросы к «тексту» грамматики (как источника информации). Например, ему нужно узнать, какие типы сложных предложений выделяются в Академической грамматике русского языка. В ответе системы должна быть выдана информация относительно содержания (элемент СпецЗн) некоторого раздела грамматики, безотносительно к тому, каким лингвистическим способом это содержание оформлено (элемент ЛингвЗн). Но лингвиста как специалиста может также интересовать эта последняя информация о чисто лингвистической организации данного материала (причем содержание



Рис. 2. Компоненты понимания «текста» Грамматики

этого материала — описание сложных предложений — является в данном случае лишь адресом, по которому можно найти объект анализа. На рис. 2 показана ситуация, когда анализируемым «текстом» (актантом А, т. е. источником информации) является фрагмент грамматики.

Разделение выдаваемого результата на информацию о форме и информацию о содержании источника имеет целью не только удовлетворение информационных потребностей специалиста-лингвиста. Оно нужно и для специалиста-нелингвиста как способ устранения «лишней» информации (которую иногда называют «лингвистическим шумом») из выдаваемых системой структур. Так, специалисту-фольклористу, подсчитывающему в текстах былин все упоминания о коне, принадлежащем Илье Муромцу, нужно убрать из структур, сопоставленных разным поверхностным реализациям (*конь под Ильей Муромцем, конь Ильи Муромца, Илья седлает своего коня* и др.), собственно лингвистическую информацию (род. пад., посессивное отношение, притяжательное местоимение), оставив интересующее адресата смысловое отношение

принадлежность (конь, Илья Муромец).

Как и в случае чтения специалистом технических текстов, лингвист может в лингвистическом тексте понять все слова и не понять «по-крупному», о чем говорит автор. Такой же эффект непонимания проявит и система, если в нее вложить сведения только о «мелких» единицах (ЛингвЗн) и не поместить «крупных» концептуальных знаний (СпецЗн). Если же поступить наоборот

и поместить только «крупные» единицы, система будет проявлять «жесткость», например, узнавать одну и ту же концепцию во всех разнящихся по существу утверждениях. Это противоречие непреодолимо при установившейся на данный момент методике составления словарей для автоматических систем.

И все-таки именно лингвист может грамотно выполнить эту задачу, может грамотно произвести вычитание ЛингвЗн из общего содержания (состава) текста, отделить то, что сообщается о каких-то лингвистических явлениях, от того, как об этом сообщается. Именно лингвист может определить такие процедуры автоматического анализа произвольного текстового материала, которые приведут к построению информативных для адресата информации единиц.

Можно ли создать сейчас единый формат описания слов, который был бы логическим объединением всех видов информации, необходимой в разных задачах АПТ? Ответ очевиден: по состоянию на данный момент сделать это практически невозможно. Можно только, идя навстречу такому пожеланию, рассмотреть множество словарей как стихийную базу словарных данных и попробовать применить к нему процедуры автоматической обработки, применяемые к обычным текстам.

Задачу построения информационной системы словарей Машинного фонда РЯ можно уточнить, назвав два аспекта систематизации:

1) создание системы словарей как источников информации, или создание фонда словарных сведений, упорядоченной системы содержимого словарных статей;

2) формирование комплекса словарей как системы средств автоматического анализа разных источников информации.

В следующем разделе формулируется некоторая общая идея создания системы словарей в первом смысле как части кибернетической системы Фонда. Один из вариантов системы словарей во втором смысле иллюстрируется на примере рассмотрения организации словарей в одной системе автоматического перевода (ФРАП, ВЦП). Комплекс словарей ФРАП интересен в обсуждаемом аспекте тем, что ориентирован на осуществление полного цикла автоматического понимания естественного текста.

Контуры Фонда РЯ как кибернетической системы

Чтобы быть не просто складом (или даже кладбищем) разных словарей и словарных комплексов, а действительно полезным источником лингвистической информации, с одной стороны, и инструментом извлечения информации из разных специальных и общезначимых текстов — с другой, словарный фонд РЯ должен стать системой в кибернетическом смысле. А это требует, чтобы «внутренний» язык системы, на котором происходит обмен информацией между разными словарями и всеми частями системы,

совпадал или хорошо стыковался с ее «внешним» языком, на котором возможно общение с адресатами информации. Выше говорилось о том, что системы АПТ принимают один язык: это или язык лингвистических структур (язык ЛингвЗи) — и тогда он плохо воспринимается потребителем информации, или язык знаний — и тогда он плохо соотносится с реальными текстами, весь интеллект системы уходит на преодоление искусственного «языкового» барьера. В обоих случаях анализ жесткий, отсутствует обратная связь между частями системы и с внешним потребителем, не строится то, что можно назвать Информацией.

Чтобы построить кибернетическую систему словарей как источника информации (это система словарей в первом смысле), необходимо ввести единый информационный эквивалент, на который и с которого переводимы любые высказывания на ЕЯ как общезначимого, неспециального характера, так и высказывания профессионального языка, в частности содержание словарных вокабул или словарных статей автоматического словаря. В качестве такого смыслового эквивалента можно использовать язык элементарных смысловых отношений (СО), двухместных предикатов типа

P (A, B)

Высказывания такого ИЯ имеют одинаковую длину — это триады, являющиеся структурной формулой минимальной порции информации. В отличие от отраслевых ИПЯ, единицы которого задаются лексически (т. е. понятия информационного языка и словаря совпадают), а также от универсальных языков «семантических множителей», предлагаемый ИЯ двухместных СО задается структурно:

A есть P по отношению к B, или

A находится в отношении P к B,

где A, P и B — переменные, которые могут принимать разные лексические значения в зависимости от типа анализируемого материала. В этом виде представимы и сколь угодно большие порции информации, т. е. A и B могут быть сколь угодно большими, сложными единицами.

Переводом словарей (вернее, состава словарных вокабул) на ИЯ двухместных отношений мы делаем семантическое пространство словарей структурно однородным, снимаем специализации, разукрупняем, «рассыпаем» сложные, неделимые единицы (типа «фрейм»), каковой является словарная статья. Данный ИЯ позволяет задавать любые степени огрубления материала. Переводом на него могут быть устранены многие «диалектные» различия, созданные разными философиями авторов словарных статей или обусловленных разными способами решения задач анализа. Создается основа для сравнения разных словарных параметров, для формирования метаязыка словарных описаний. Тем самым

будет подготовлена основа для создания, компиляции системы словарей во втором смысле — как ряда, последовательности словарей, служащих инструментом анализа в конкретных системах АПТ.

Ко всякой начинающейся работе естественно предъявлять максимальные требования. Соответствующая таким максимальным требованиям идеальная ИС Фонда должна, на наш взгляд, обладать свойствами, которые мы перечисляем ниже.

Кибернетическая ИС имеет то большое преимущество перед традиционными системами АПТ, что может начинать работу на любом исходном материале, как по количеству, так и по степени его совершенства. Если средства анализа хорошо выбраны, различие в исходном качестве проявится в результирующем продукте.

Кибернетическая ИС не просто накапливает и хранит данные — она наводит «внутренний порядок» в множестве своих объектов: постоянно обобщает их, суммирует, сравнивает, устанавливает сходства на любом стихийном множестве объектов, т. е. строит «естественные классы» в терминологии А. Я. Шайкевича. Назовем этот процесс «идеализацией». Результатом такой идеализации является создание потенциальной информации. Если нет внешних требований на эту внутреннюю информацию, система строит производную, вторичную информацию (статистические сведения о количестве полученных классов, о количестве единичных явлений, не вошедших в классы, и т. д.), выводы и даже «поступки»: часто повторяющиеся сведения переводит из словаря в грамматику, из долговременной памяти в оперативную и т. д. Если для каких-то выводов не хватает материала, она формулирует, какого рода сведений не хватает, или ставит вопросы исследователю. Если на систематизированную «изнутри» информацию нет спроса, она требует ответа, что с этим делать дальше, или предлагает некоторое стандартное использование полученной информации и т. д.

Кибернетическая система осуществляет постоянный обмен информацией с другими кибернетическими устройствами. Она принимает на входе информацию в виде вопроса или в виде новой порции знаний и т. д., перерабатывает его указанным выше способом «идеализации» и вырабатывает ответ. Как и любому разумному существу, ей нужна эта постоянная связь с внешней средой. Знание о том, как использована выработанная ею информация системой-адресатом, должна ею учитываться в виде перестройки ее единиц, выводов, оценок и т. д. Названные свойства выполняются для всех ее частей, в том числе для такой большой подсистемы, какой является подсистема словарей, и для каждого отдельного словаря. Идеализация массива словарных вокабул состоит в том, что отмечаются все регулярности, вокабулы разбиваются на типы, повторы группируются в отдельную рубрику, неполные фрагменты выравниваются по более полному образцу и т. п. Обмен информацией осуществляется и между всеми ча-

стями, и между любой частью и целой системой. Если в каком-то компоненте системы вырабатывается новая информация, она должна «отзываться» во всех других компонентах, связанных с данным.

Возможность построения системы словарей Фонда автоматической процедурой «идеализации» исходного множества словарных описаний является, конечно, некоторой смелой гипотезой, но ее позволительно ставить перед Фондом как творческой лабораторией.

Если объектом систематизации словарей в первом смысле является состав словарных информаций, то объектом систематизации во втором смысле является множество единиц, образующих сами словарные входы. Создаваемые для систем АПТ словари имеют в качестве входов не только отдельные слова, но и более крупные единицы: обороты, термины, фраземы и т. д., поскольку некоторые виды информации можно приписать только термину или только понятию, но не слову. При анализе текста возникает необходимость на промежуточных стадиях анализа фиксировать единицы, которые нельзя было задать априорно,— это единицы текста, единицы информационного анализа, единицы общения двух (нескольких) интеллектуальных систем. Дело в том, что словарь не остается неизменным в ходе анализа (понимания) текста — он обогащается, изменяется, в общем обучается. В процессе анализа текста исходная словарная единица, взятая как болванка, как заготовка, обрастает разными текстовыми характеристиками, которые нельзя было предсказать заранее: это текстовые термины, информационные термины. Создание окончательного словаря, настроенного на заданный корпус текстов, проходит через построение таких текстовых единиц. Поэтому систему словарей во втором смысле можно назвать «каскадом» словарей: это ряд словарей, последовательно работающих как инструмент анализа текстов и позволяющих строить все более крупные единицы, принадлежащие уже данному корпусу текстов. Для текстов, представляющих художественную ценность, эти промежуточные продукты семантического анализа можно фиксировать — будем получать таким образом авторские словари. Фиксируя таким же образом текстовые единицы для научно-технических текстов, можно строить «изнутри» термины информационного тезауруса.

Организация словарей в одной системе машинного перевода

Остановлюсь коротко на организации словарей в системе французско-русского автоматического перевода (ФРАП, ВЦП). Словари как средство анализа образуют систему в смысле «каскада» словарей, где единицы каждого словаря описаны в категориях, подготавливающих построение единиц следующего уровня. Система ФРАП строится как информационно-переводческая систе-

ма, целью которой является перевод одновременно с переработкой информации исходного текста (сжатие, ответы на вопросы и др.). Словарная информация распределена по 10 словарям, которые включаются или выключаются в разных по сложности режимах работы системы. Позволю себе перечислить и кратко аттестовать эти словари:

ФМС — французский морфолого-синтаксический словарь;
СФО — словарь французских фиксированных оборотов;
СОБ — словарь гибких оборотов и аналитических комплексов;
ФСЧ — словарь французских словосочетаний;
СОТ — словарь отношений (предлогов, союзов, беспредложных связей);
ФСС — французский семантический словарь;
РСС — русский семантический словарь;
РСЧ — русский словарь словосочетаний;
РГС — русский грамматический словарь;
РМС — русский морфологический словарь.

Намечено ввести еще два словаря:

ТЕЗ — словарь-тезаурус терминов определенной области знаний, с задаваемыми на нем энциклопедическими отношениями;
СИТ — словарь стандартных научно-технических ситуаций, схем предметной области.

СФО привлекается еще на стадии графематического (доморфологического) анализа, он позволяет собирать единицы, описываемые жесткой последовательностью текстовых элементов, включая окказиональные образования, цифровые комплексы и аббревиатуры, действительные только на ограниченном корпусе текстов.

Входами в ФМС являются слова с нерасчлененными значениями, различие в значениях фиксируется только при наличии грамматической омонимии (типа *печь* сущ.— *печь* глаг.). У слов, имеющих разные модели управления (МУ), в этом словаре записывается одна обобщенная МУ. Словарная статья ФМС содержит только формально-грамматические категории и не содержит никаких семантических характеристик. Словарь привлекается при морфологическом и синтаксическом анализе предложений.

СОБ содержит регулярные для языка составные единицы, задаваемые лексемами, грамматическими классами лексем, порядком их следования и возможными вставками на месте разрывов. Этот словарь позволяет также собирать сложные времена и другие аналитические комплексы (возвратные, вопросительные и отрицательные формы и др.). Результатом работы этого словаря является построение синтаксического узла и частичное предсказание связей, проходящих через него.

Входом в ФСЧ являются уже синтаксические поддеревья, которые при опознании в тексте формируют более крупный синтак-

сический узел с возможными новыми синтаксическими свойствами. Все упомянутые словари не имеют дела с семантической информацией, они только строят более крупные синтаксические узлы, которые могут быть входами в интерпретирующие словари СОТ и ФСС.

СОТ дает перевод единиц ЕЯ или уже построенных системой единиц в одной или несколько альтернативных смысловых отношений, т. е. единиц будущей семантической и затем информационной структуры текста.

ФСС описывает смысловое поведение слов, взятых в определенных значениях. Разбиение на значения оказывается не дробным (так, прямое и переносное значения слова могут описываться в пределах одной словарной статьи), поскольку ФСС предсказывает текстовое поведение слова в достаточно общих смысловых валентностях, а также возможную реализацию этих валентностей в пределах предложения (в терминах синтаксических отношений). ФСС предсказывает семантические характеристики своих актантов, вернее, актантов называемой этим словом ситуации, которые могут оказаться и в других предложениях. По сути дела, СОТ и ФСС описывают в терминах ИЯ связи входящих в них единиц с их смысловыми составляющими, подготавливая тем самым собирание единиц следующего, текстово-семантического уровня и следующего за ним — информационного.

Организация словаря в виде каскада разгружает словарную систему в целом, обеспечивает «монотонность» сведений в каждом отдельном словаре, что позволяет строить (автоматически) шаблоны, используемые для вылавливания ошибок в словарных статьях, а также для полуавтоматического составления следующих порций словаря. В системе ФРАП успешно работают эти два вида словарной службы. Кроме того, работает программа автоматической сверки входов — выходов разных словарей, позволяющая выявлять расхождения в составе словарных единиц.

Реализованные в системе ФРАП 10 словарей дают, так сказать, общеобразовательные сведения о словах, в категориях приведенных на рис. 1 и 2 схем они относятся к компоненту ЛингвЗн. Это значит, что словари не только синтаксического, но и семантического компонентов системы не настроены жестко на одну предметную область, они могут использоваться для анализа естественных текстов произвольной тематики. Настройку на предметную область должны задавать два последних словаря — ТЕЗ и СИТ, в категориях упомянутых схем они относятся к СпецЗн. Вообще говоря, уже имеющаяся система словарей имеет формальные средства описания терминов. Последние рассредоточены по разным словарям, в зависимости от степени связанности составных частей термина: одноэлементные термины описываются в ФМС и ФСС, состоящие из двух и более слов могут входить в СФО (типа ВЦ АН СССР) или в СОБ (типа двигатель внутреннего сгорания); достаточно свободные термины, которые трудно описать в категориях лексем и порядка их следования, но

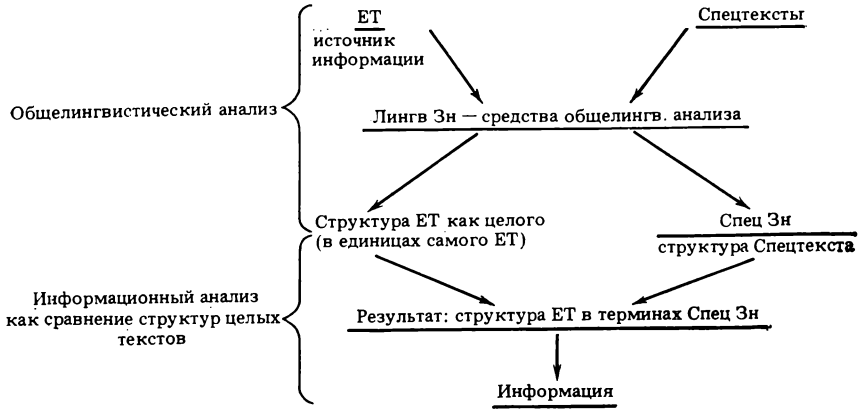


Рис. 3. Уточненная схема компонентов АПТ

можно описать лексемами с определенными на них синтаксическими связями, входят в ФСЧ; совсем свободные термины, которые задаются ключевым словом и смысловыми связями с другими элементами, описываются в ФСС. Это дает возможность уже имеющимися общеязыковыми средствами описывать общетехнические, общеязыковые или случайные термины. Однако терминосистема в целом такими средствами не может быть охарактеризована. Эту задачу должны выполнять последние два специальных словаря. Они, как уже было сказано, могут вводиться в систему в своем естественном виде: списком терминов стандартного информационно-поискового тезауруса (ИПТ), энциклопедическим словарем и т. п., наконец, в виде канонического текста (для тех областей, в которых еще не систематизирована терминология). Наличие в системе ФРАП единого информационного эквивалента, в форму которого переводимо содержание любого источника, позволяет задавать компонент СпецЗн в свободном виде «текста». При анализе ЕТ другой тематики меняется только компонент СпецЗн. Такие «тексты» знаний (включая ИПТ) должны проходить все те же этапы анализа, что и анализируемый источник информации. Последний компонент системы — информационный — должен осуществлять сравнение между собой структур, построенных лингвистическими компонентами для анализируемого текста и текста знаний. Предлагаемое (и реализуемое в системе ФРАП) изменение в распределении компонентов системы АПТ показано схемой на рис. 3.

Итак, в составе Фонда РЯ словари могут выступать, во-первых, как специфическая разновидность источников информации, во-вторых, как инструмент извлечения информации из разных источников. Построение машинного фонда словарей РЯ предлагается осуществлять в жанре лингвистического исследования, которое направляется «категорическим императивом» самого материала.

Построению словаря, являющегося хорошим инструментом поиска информации в разных массивах документов, должно предшествовать превращение стихийной базы словарных данных в систему словарных сведений. Первый шаг систематизации в этом смысле состоит в декомпозиции, приведении разнородных словарных сведений к единому информационному эквиваленту. Результат такого перевода может быть объектом формальных процедур анализа (применимых и к обычным текстам), имеющих целью построение типов словарных единиц и словарных информаций.

Систему словарей во втором смысле образует множество словарей, работающих последовательно при анализе (понимании) конкретного текстового материала. Это множество составляет как бы поисковый язык, настраивающийся на анализируемый материал: анализ начинается с самого общего словаря, безразличного к тематике текстов, и доходит до самого специфичного, справедливого только для данного типа текстов.

Разукрупнение и стандартизация, затем укрупнение и специализация — вот путь системного исследования разных лингвистических материалов, который предлагается реализовать на источниках Машинного фонда русского языка. Моделирование этих процессов может быть начато на любом небольшом фрагменте материалов Фонда. Что может дать такой эксперимент специалистам, разрабатывающим разные системы класса АПТ?— а) «средний», неспециализированный семантический словарь русского языка; б) механизмы «настройки» словаря на корпус текстов заданной тематики; в) правила автоматического анализа текста, включающие методику освоения новых текстов, лексическое и семантическое обновление словарных массивов.

СОСТАВЛЯЮЩИЕ И ПРИНЦИПЫ ФОРМИРОВАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА

И. А. БОЛЬШАКОВ

Введение

Важной проблемой, возникающей при создании Машинного фонда русского языка, является построение адекватной алгоритмико-программной базы, т. е. фонда лингвистических процессоров и фонда утилитных лингвистических алгоритмов и программ (совместно эти лингвистические программные фонды обозначаются далее через ФЛАП).

На наш взгляд, ФЛАП должен являться неотъемлемой частью Машинного фонда русского языка, поскольку без программного обеспечения невозможно ни сформировать сам фонд, ни вести на его базе исследования теоретического характера, ни

вести исследования прикладного характера с передачей части возникающих лингвистических и программных продуктов заинтересованным организациям.

В части последнего отметим, что число промышленных организаций, вникающих в задачи лингвистического характера и начинающих самостоятельные исследования в этой области, порой на низком уровне, постоянно растет и шефство над такими работами со стороны организаций — держателей Машинного фонда русского языка представляется неизбежным.

В данной работе, носящей сугубо предварительный характер, описывается предположительный состав ФЛАП и перечисляется ряд принципов, которые разумно положить в основу его формирования и использования.

Роль отдельных перечисляемых компонентов при формировании Машинного фонда русского языка и проведении исследований на его базе будет заведомо неодинаковой. Мы настаиваем лишь на наличии всех перечисляемых ниже потребностей, которые по мере развития вычислительной техники и представлений о структуре Фонда будут трансформироваться и уточняться.

Ниже не рассматриваются требования к системам управления базами данных (СУБД), на основе которых следует строить Машинный фонд в целом. Только мощная современная СУБД сделает ненужным для пользователя-лингвиста знание внутренней структуры Фонда на машинном («физическом») уровне, упростит проблемы пополнения, обновления, рационального дублирования и обеспечения целостности столь ценных материалов. Такую СУБД можно будет заимствовать из числа тех, которые будут иметься в СССР на момент технического проектирования Фонда. Рассматриваемые же ниже компоненты ФЛАП являются пакетами прикладных программ (ППП), составляющими как бы окружение для выбранной СУБД.

Фонд лингвистических процессоров

Лингвистические процессоры — это программы, являющиеся функциональными моделями языковой деятельности человека и осуществляющие преобразование естественно-языковых высказываний (ныне — в их письменной, а в перспективе и в речевой форме) в некоторые их смысловые представления (=анализ) и, обратно, смысловых представлений высказываний — в соответствующую им письменную (в перспективе речевую) форму (=синтез).

По мере развития вычислительной техники и совершенствования алгоритмов лингвистической обработки возможно появление лингвистических процессоров, имеющих вид специализированной вычислительной аппаратуры, но это не повлияет на существо рассматриваемых проблем.

Фонд лингвистических процессоров должен содержать по меньшей мере:

1. Математический аппарат описания функциональных моделей языка. Поскольку практика последних 20 лет показала, что единого математического аппарата для описания всех лингвистических преобразований и фактов предложить не удается, необходимо привлечь все средства, показавшие свою практическую ценность ранее, как то:

— формальные грамматики, укладывающиеся в классификацию Н. Хомского или не укладывающиеся в нее (например, грамматики с управляемым выводом) [1];

— аппарат описания восходящего и нисходящего грамматического разбора, заимствованный из богатой практики трансляции языкового программирования [2];

— древесные преобразователи, развиваемые Гренобльской школой [3];

— вариант узкого исчисления предикатов, предложенный для описания синтаксических признаков и правил [4];

— модифицированный аппарат шкал и масок [5] и многое другое.

В данном случае идет речь о чисто формальных приемах описания, которые могут быть представлены в машинной форме в виде литературных ссылок, рецептуры и образцов применения.

2. Совокупности формальных грамматических правил, описывающих конкретные уровни (фрагменты) естественного языка, представленных в исходной (т. е. ориентированной на прикладного лингвиста) форме и удовлетворяющих требованиям выбранного математического аппарата (см. п. 1).

3. Формализованные словари в их исходной форме, согласованные по структуре и содержанию с грамматическими правилами п. 2. Структура отдельных статей этих словарей должна быть предметом особого внимания исследователей, чтобы удовлетворять требованиям разных школ и методов описания языковых фактов. Некоторый опыт, возможно, удастся заимствовать из практики обмена научно-технической информацией с помощью библиографических описаний.

---

[1] *Гладкий А. В., Диковский А. Я.* Теория формальных грамматик.— В кн.: Итоги науки: Теория вероятностей. Математическая статистика. Теоретическая кибернетика. М., 1972, Т. 10.

[2] *Льюис Ф., Розенкранц Д., Стирнз Р.* Теоретические основы проектирования компиляторов/Пер. с англ. М.: Мир, 1975.

[3] *Буате К., Недобежкин Н.* Русско-французский перевод в группе исследований по автоматическому переводу: общая методика и разбор примера.— В кн.: Разработка формальной модели естественного языка. Новосибирск, 1981, с. 29—52.

[4] *Цинман Л. Л.* Язык задания лингвистической информации для автоматического анализа и синтеза текстов.— В кн.: Формальное описание структуры естественного языка. Новосибирск, 1980, с. 105—123.

[5] *Бидер И. Г.* Синтез словоформ с помощью преобразовательных грамматик.— НТИ. Сер. 2, 1979, № 6, с. 22—30.

4. Трансляторы исходных грамматических правил и словарей и их рабочее (т. е. сжатое, ориентированное на обработку текста) представление. Сжатие происходит в 2—4 раза, обеспечивает резкое повышение быстродействия обработки, но ценой лишения исходных объектов их понятной текстовой формы. Такие трансляторы могут быть сопоставимы по сложности с самими программами лингвистической обработки (см. п. 6).

5. Трансляторы исходных словарных материалов иных компонентов Фонда в исходные словарные статьи лингвистических процессоров. Такие трансляторы позволят получить лишь полуфабрикаты (полузаполненные статьи), которые подлежат пополнению и коррекции в интересах конкретной функциональной модели языка, но это позволит сэкономить средства и время на ввод множества малоинтересных в грамматическом отношении словарных статей.

6. Собственно рабочие программы анализа и синтеза текста. Это ядерная часть фонда лингвистических процессоров, расчленяемая на морфологические, синтаксические и другие уровни. Достойное место здесь должны занять уже существующая система автоматического перевода, разработанная под руководством Ю. Д. Апресяна и Л. Л. Цинмана [6], и система общения с ЭВМ на естественном языке, разработанная под руководством Э. В. Попова [7].

7. Программы взаимодействия лингвистических процессоров с базами фактографических и документальных данных, а также с приводными и сенсорными устройствами роботов и т. п.

ППП, упомянутые в п. 2, 3, 4 и 6, желательно иметь по всем уровням языка, но полезными как в исследовательском, так и в прикладном планах окажутся и парциальные программы, ориентированные, например, только на предморфологический («графематический»), морфологический или предсинтаксический («словосочетательный») уровни.

Все программные средства должны уметь обрабатывать русские тексты как можно более широкого круга жанров и проблематик, но особую ценность приобретут ППП, рассчитанные на обработку целой группы естественных языков.

Фонд утилитных лингвистических алгоритмов и программ

Особо важную роль среди утилитных средств сыграют программы для ввода, редактирования, коррекции и издательской подготовки текстов произвольного характера. Единого русского

---

[6] *Апресян Ю. Д.* и др. Лингвистическое обеспечение в системе автоматического перевода ЭТАП-1.— В кн.: Разработка формальной модели естественного языка, с. 3—28.

[7] *Попов Э. В.* Общение с ЭВМ на естественном языке. М., 1982.

термина для этого круга программ пока нет (ср. англ. word processing). Здесь особо отметим:

Текстовые редакторы — программы интерактивного (за дисплеем) ввода новых текстов в память ЭВМ, располагающие не только средствами замены, перестановки и дублирования строк, но и редактирования внутри строки путем стирания, замены и вставки новых литер, замены в выделенной части текста одной подстроки на другую, расщепления и слияния строк и др. Это очень мощное средство, заменяющее морально устаревшие специализированные системы подготовки данных (в том числе — многопультовые). В наиболее развитых случаях они позволяют хранить несколько вариантов текста в виде одного канонического варианта и файлов изменений, вносимых иными вариантами в выбранный канон.

Текстовые автокорректоры («спеллеры»), т. е. программы автоматического обнаружения и исправления ошибок произвольного происхождения. В простейшем случае они работают в пакетном режиме и производят обнаружение буквенных последовательностей, отсутствующих в приданном им словаре (словоформ и/или основ). В более сложных случаях автокорректоры работают интерактивно и, останавливаясь около каждого «подозрительного» слова, выдвигают альтернативы на его замену из состава приданного словаря. В последнем случае можно либо допустить данное слово без его внесения в рабочий словарь, либо принять и вставить его в словарь, либо локально заменить на новое слово, либо ввести замену данному слову (например, аббревиатуре) по всему последующему тексту, либо, наконец, отредактировать текущую строку в целом (см. текстовые редакторы).

Со временем автокорректоры будут сочетаться с лингвистическими процессорами, что позволит обнаруживать ошибки, замаскированные под существующие, но не сочетающиеся по смыслу с окружающим контекстом слова, а также ошибки синтаксического и стилистического характера.

Программы издательской подготовки, осуществляющие форматизацию текста в широком смысле этого слова, т. е. придающие тексту полиграфически законченную форму, вне зависимости от конкретного устройства, которое выдает этот текст на печать. Для такой форматизации, как правило, необходима специальная редакторская разметка текста, которая заблаговременно вносится в диалоге. Осуществляемые операции форматизации включают выравнивание строк слева и справа, формирование правильных переносов слов там, где без переносов не обойтись, подчиняющееся определенным правилам распределение текста по строкам и страницам (полосам) и др. Весьма важно также управление шрифтовыми гарнитурами.

Программы синтеза готовых форм и заполнение таблиц, являющиеся узкоспециализированным, но очень

полезным средством подготовки анкет, отчетов, словарных статей и т. п.

В качестве следующей крупной группы утилитных ППП отметим программы сортировки/слияния. Они осуществляют лексикографическое упорядочение заданных файлов (например, словарных) по выделенным полям отдельных записей в этих файлах (например, по буквенной оболочке основ), слияние нескольких упорядоченных файлов в единый, а также вычеркивание составляющих одного файла из другого. В качестве одного из примеров применения таких программ служит подготовка прямых и обратных словников.

Очень важными для Машинного фонда русского языка окажутся примыкающие к предыдущей группе ППП для статистической обработки: составители частотных словников, составители конкордансов и индексов, средства набора статистики по парциальным лингвистическим фактам.

Для того чтобы последние средства могли работать применительно к фактам, относящимся не к орфографическому (поверхностному) уровню языка, они должны сочетаться с элементами лингвистических процессоров.

Чтобы переводить тексты из одного кодового представления в другое, будут также необходимы конверторы. В простейшем случае они переводят текст из кода КОИ-7 в ДКОИ и обратно, но возможны также случаи расширенных и многостраничных кодовых таблиц для представления нескольких алфавитов сразу (например, кириллического и латинского), отображения множества литер с диакритическими знаками, биграммного и триграммного изображения специальных литер и лигатур и т. п. Сюда же относятся программы архивации — перевода текстов в каноническую форму хранения на магнитных лентах или иных архивных средствах, которые возникнут в будущем. Отдельную, но очень важную группу утилитных средств составят программы обучения русскому языку в его письменной, а в перспективе и устной форме. Естественно, они будут различаться ориентацией на детей — учащихся национальных школ или иностранцев.

Наконец, для всех упоминавшихся ранее средств, включая лингвистические процессоры, необходимо обеспечить диалоговый монитор, оперирующий доступным для лингвиста языком управления заданиями. Именно с помощью команд этого монитора лингвист должен иметь возможность, не программируя сам, ввести, отредактировать и проверить текст; отформатизовав, вывести его на печать; составить по тексту (текстам) словник, провести статистическую обработку материала; получить нужную библиографическую справку и т. д. и т. п.

Основные принципы формирования ФЛАП

Уже без глубокой проработки проблемы ясны следующие принципы формирования и использования лингвистических программных фондов.

1. Полное покрытие нужд лингвистов-русистов, включая не только нужды лексикографии в узком смысле.

2. Опора на ограниченный набор универсализированных структур данных, показавших свою особую полезность применительно к лингвистическим проблемам. Спектр возможных структур современной информатики настолько широк, что условия для взвешенного, рационального выбора таких структур достаточно подготовлены.

3. Опора по возможности на ограниченный инвентарь лингвистических теорий и методов, со стремлением в перспективе перейти к единому принципу описания языка. Возможности для этого пока ограничены, но программное обеспечение как инструмент должно по мере сил способствовать такой тенденции.

4. Преимущественно одноразовый ввод как словарных материалов, так и текстов произвольного характера.

5. Преимущественно однокопийное хранение редко используемых материалов с возможностью быстрого дублирования их для нужд произвольного пользователя.

6. Иерархичность размещения и использования Машинного фонда русского языка в целом, предполагающая банк лингвистических сведений, распределенный по разным организациям и даже городам, оживленный обмен материалами по каналам связи между организациями — держателями компонентов Фонда и мощную сеть пользователей, которым доступны в электронной форме все словарные, текстовые и программные богатства. Отдельный пользователь должен сам решать, какие материалы ему продублировать из Фонда, а какие запрашивать эпизодически из общего хранилища.

7. Адекватное документирование программных средств. Существуют советские стандарты, требующие, чтобы всякий программный продукт, сдаваемый в Государственный фонд алгоритмов и программ, был снабжен как минимум тремя документами: текстом программы, описанием программы и описанием ее применения. Для удобного пользования лингвистическими ППП этого, на наш взгляд, мало. Необходим развернутый набор руководств для пользователя (с описанием примеров), для прикладного программиста, для прикладного лингвиста, для системного программиста (если ППП требует больших ресурсов и сложной наладки). Опыт выпуска таких руководств (кроме руководств для прикладного лингвиста) есть, например, у Международного центра научной и технической информации. Итак, создание фондов лингвистических алгоритмов и программ потребует значительных усилий, включая привлечение

готовых, адаптацию принципиально применимых и разработку ряда новых ППП. На наш взгляд, уровень развития вычислительной техники и программных средств уже достаточен, чтобы решить задачу комплектации программных фондов в 5—7 лет. Проблема остается в необходимых организационно-технических усилиях для запуска работ на современной технике и методологии.

ИЗУЧЕНИЕ ДРЕВНИХ РУКОПИСЕЙ И МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА

Л. П. ЖУКОВСКАЯ

Насущные задачи исторического языкознания вообще, а не только истории русского языка требуют от филологов с помощью специалистов по другим наукам и вычислительной технике найти способы и пути использования сокровищ древнерусской письменности таким образом, чтобы можно было получать основанные на всех сохранившихся материалах объективные выводы по истории русского, украинского и белорусского языков. Заключения эти должны в максимально возможной степени отражать действительное состояние литературно-письменного и народно-разговорного языка в разные периоды их исторического состояния и в применении к разным территориальным диалектам древности. Эти проблемы применения ЭВМ стали занимать меня с конца 1964 г. и нашли письменное выражение в книге 1968 г. Но тогда я писала достаточно пессимистично: «Ускорить необходимый формальный анализ всех сохранившихся до нашего времени списков, хотя бы от древнейших до середины XV в., могло бы лишь использование машинной вычислительной техники... В ближайшее время проведение такой работы не представляется реальным» [1].

При изучении языка по древним памятникам и их спискам нам нужно выйти из состояния ремесла и знаточества на путь электроники и электронно-вычислительной техники. Дошедшие до нашего времени списки памятников древней и средневековой письменности хранят множество фактов письма и языка, изучение которых является важным компонентом в понимании нашей письменной и языковой культуры прошлого. Известно, что, основываясь на данных Б. В. Сапунова о том, что на Руси до Батыева нашествия существовало 10 тыс. церквей и 200 монастырей, и учитывая, что христианство является религией с необхо-

[1] *Жуковская Л. П.* Типология рукописей древнерусского полного апракоса XI—XIV вв. в связи с лингвистическим изучением их.— В кн.: Памятники древнерусской письменности: Язык и текстология. М., 1968, с. 293.

димыми для богослужения крупными рукописными книгами (их число для древнейшего периода истории устанавливается довольно точно), акад. В. В. Виноградов насчитывал примерно 85 тыс. книг до середины XIII в. К тому же он добавлял: «Сюда же следует отнести и развивавшиеся на базе книжно-славянского типа древнерусского литературного языка произведений житийной, апокрифической, проповеднической и другой религиозно-дидактической литературы»[2]. Аналогичные числа были получены мною по процентному отношению сохранившегося наследия для каждого века[3]. Теперь предлагаю еще и третий способ подсчета. Так, если предположить, что в древности в год готовили всего лишь 10 грамотных мальчиков (число явно занижено) и они переписывали книги всего лишь в течение 20 лет своей жизни, переписывая в год две книги (вспомним, что Остромирово евангелие было переписано за полгода), то к середине XIII в. было переписано примерно 100 тыс. книг. Между тем до наших хранилищ от XI — XIV вв. дошло 1500 книг (считая привезенные южнославянские)[4], от XV в. их сохранилось примерно 2500. Следовательно, все это наследие вполне обозримо и учтено.

Сохранившиеся крохи древнерусской письменной культуры отражают множество разнообразных явлений языка. Если учесть, сколько отклонений при переписке книги сделает машинистка, то естественно, что при переписывании стотысячного фонда от руки древние писцы внесли множество написаний, восходящих к их живой речи. Кроме того, нужно еще учесть, что писцы подчас редактировали тексты в соответствии со своими собственными задачами или со своим пониманием переписываемого. Выявить все это и передать в руки лингвистов и текстологов без помощи вычислительной техники невозможно.

Как можно представить себе изучение древнего и средневекового рукописного фонда, который должен войти в Фонд русского языка? Естественно, нужно, чтобы вся информация об этих памятниках письменной культуры была введена в память ЭВМ и могла быть доставлена в любой научный центр с электронными скоростями. Но следует отдать себе отчет, что существует кардинальная разница между теми задачами, которые стоят при вводе в память машины современного текста и вводе рукописного древнего текста. Современный текст, если он введен оператором в память машины с пропуском или неверной заменой буквы, в дальнейшем может быть откорректирован по законам современной орфографии. Древний же текст должен быть введен буква в букву, строка в строку. Пропущен-

---

[2] *Виноградов В. В.* Основные проблемы изучения образования и развития древнерусского литературного языка. М., 1958, с. 39.

[3] *Жуковская Л. П.* Сколько книг было в Древней Руси? — Русская речь, 1971, № 1, с. 73—80.

[4] См.: Сводный каталог славянорусских рукописных книг, хранящихся в СССР (XI—XVIII вв.). М., 1984.

ная в нем буква может породить каскад ошибочных лингвистических заключений самого разнообразного характера. Поэтому ввод древнего текста должен осуществляться только автоматически. Эта задача сейчас еще неразрешима, но это не значит, что она не будет разрешена в обозримое время.

Главное требование состоит в том, чтобы ни на каком этапе ввода информации не было вмешательства в закодированный текст со стороны не историка языка. Однако устройства, которое читало бы микропленки древних текстов (да, кажется, и современных тоже), пока не существует. Ведь оно должно было бы различать и идентифицировать изменявшиеся с течением веков, а подчас и в индивидуальной практике начертания одних и тех же букв! Как известно, машина может хранить в памяти и картинку листа текста... Но пока это и сложно, и дорого, во всяком случае значительно дороже, чем хранить знаковую информацию. Следовательно, нужно, чтобы у машины появились возможности распознавать буквы кириллицы (1), и нужна программа для этого (2). Иными словами (может быть, не вполне профессиональными), нужны приборы, которые непосредственно с пленки будут вводить в память машины буквенный текст. В принципе, кажется, уже и теперь возможны такие устройства, которые способны воспроизводить любые буквы (т. е. не только современный русский, латинский и цифровой материал) в памяти машины.

Второй шаг работы составит уже труд квалифицированного *лингвиста-древника* (именно так!), который будет работать на дисплее с преобразованными машиной в строки текстами рукописей. Прежде всего перед этим специалистом стоит подчас крайне трудная задача разбить строки буквенного текста на слова, т. е. сделать древнее беспробельное письмо пробельным: вставить между слитно написанными в древней рукописи словами знак пробела. С точки зрения машинной техники это задача легкая. Сейчас строки на дисплеях содержат до 80 знаков (символов), а на алфавитно-цифровых печатающих устройствах — до 160 символов. Для древнего текста достаточна плотность в 80 символов, поскольку даже рукописные книги большого формата, написанные в один столбец (практически с XV в.), имеют значительно меньше букв на строке. Тем более это относится к двухстолбцовым рукописям (имевшиеся в византийской практике трехстолбцовые в славянской письменности не представлены). Таким образом, средствами дисплея можно и без программы увеличить в необходимой степени число символов на строке.

Кроме символа пробела, квалифицированный лингвист-древник проставит (до или после пробела) разные символы, несущие собственно грамматическую информацию (например, символ времени для глаголов — аорист, имперфект, перфект и т. д., символы рода, числа, падежа и т. п.). Естественно, что конкретные лингвистические вопросы будут обсуждены в кругах исто-

риков языка, даже и не имеющих отношения к машинной технике. Одновременно лингвист устранит ошибки распознавания текста (букв), которые машина может допустить на первом этапе работы. Следовательно, нужен не только дисплей с введенным в машину текстом, но и рядом увеличенный кадр негатива или позитива с оригинала, находящегося в древлехранилище.

На дисплее пригодится существующая уже теперь клавиша сдвига строки. Нужны также клавиши для включения в строку символов пробелов и другой научной символики; нужен также световой карандаш, наподобие имевшегося на машине МИР-2, для исправления букв, неверно распознанных машиной.

Нетрудно заметить, что в изложенной «голубой мечте» почти все реально, кроме чтения текста машиной. Но наука развивается крайне быстро, и это вселяет определенные надежды...

Есть еще важная археографическая задача, о которой не следует забывать. В последние годы в лучших хранилищах страны начато планомерное микрофильмирование всех древних рукописей, но особенно ценнейших, которые нежелательно выдавать специалистам из-за охранных мер. Микрофильмирование делает доступными сокровища древней письменности для исследователя, живущего в любом краю, для лица, не имеющего возможности работать в самом хранилище. Ясно, что если все микрофильмированные тексты рано или поздно будут введены в память ЭВМ, то будет создана та часть Фонда русского языка, которая составляет его прошлое и без которого невозможно не только историческое языкознание, но и сама история письменной культуры.

Хочется надеяться, что пока будет осуществляться планомерное микрофильмирование сокровищ древней письменности, с одной стороны, пока будут разрабатываться способы «чтения» (распознавания) машиной букв кириллицы без помощи оператора — с другой, за это время будут созданы и практически вечные носители информации для ЭВМ (каким был пергамен, сохраняющийся тысячелетия). Эти носители не должны требовать систематической перезаписи материала, как современные магнитные ленты и диски (непригодность перфокарт тем более вне сомнения).

Нетрудно заметить, что приведенное выше высказано с верой в будущее. Хочется надеяться, что оно не будет чересчур отдаленным!

ОРЕОЛЬНАЯ СЕМАНТИКА В МАШИННОМ ФОНДЕ РУССКОГО ЯЗЫКА

Л. П. ЖУРАВЛЕВ

Наиболее важной и наиболее сложной информацией, которой должен располагать Машинный фонд русского языка, будет являться семантическая информация, а в ней прежде всего должна быть отражена понятийно-логическая семантика. Однако понятийно-логическое значение (обычно отражаемое толковыми словарями) не единственный аспект семантики. Очень часто в живом функционировании слова на первый план выходят другие аспекты значения, и если в конкретной языковой ситуации их не учитывать, то можно потерять существенную часть семантической информации, а иногда и основную семантическую информацию в данном случае.

Речь идет прежде всего о коннотативно-признаковом аспекте значения слова, который операционно можно было бы определить как тот аспект значения, который может быть охарактеризован путем перечисления оценочных признаков типа «хороший — плохой», «большой — маленький», «нежный — грубый», «быстрый — медленный» и т. д. Коннотативно-признаковым значением обладают все слова. В одних случаях этот аспект едва намечен и не играет существенной роли в функционировании слов: *уровень, вторник, середина, подоконник* и т. д. С другой стороны, есть слова, понятийное ядро значения которых неопределенно, и почти весь объем значения выражается признаковыми характеристиками: *прелесть, гадость, добро, зло, очарование* и т. п. Большинство же слов расположено между этими двумя полюсами, и для их значения важны как логико-понятийное ядро, так и коннотативно-признаковый аспект.

Во всяком случае привлечение к семантическому анализу коннотативно-признаковых характеристик всегда обогащает анализ, а иногда без этого привлечения значение слова вообще остается нераскрытым. В дополнение к приведенным выше примерам доминирования признаковых характеристик в семантике слова можно указать на те случаи, когда семантические различия слов проявляются в основном в признаковых аспектах: *здание—дворец, мать—мама, ветер—ураган, глаза—очи, лицо—лик* и т. п.

Кроме того, следует, видимо, учитывать и другой аспект, совсем уж редко попадающий в поле зрения исследователей и, может быть, считающийся третьестепенным. Тем не менее иногда он проявляет себя весьма активно. Здесь имеется в виду аспект фонетико-содержательный, содержательность самой фонетической формы слова. Нельзя думать, что эта содержательность оживает только в поэзии или в каких-то редких случаях эстетического порядка.

Как показывают имеющиеся в нашем распоряжении материалы, фонетической содержательностью обладают все слова, и эта содержательность находится в различных отношениях с коннотативно-признаковыми значениями слов, что оказывает существенное влияние на восприятие слов и текстов, а также их функционирование.

Почти любое семантическое явление так или иначе связано с диалектикой взаимодействий между значением и содержательностью звуковой формы языковых единиц. Так, синонимические ряды часто строятся не только с помощью варьирования признаковых, но и фонетико-содержательных характеристик. Например, в ряду *ветер—буря—ураган—шторм—тайфун* изменяющимся признаковым характеристикам соответствуют сходные изменения в содержательности звучания слов.

А что касается экспрессивной лексики, то в ее функционировании фонетическая содержательность играет особую роль, иногда выходя даже на первый план, затмевая и коннотативно-признаковое и понятийное значение. Так, семантика слов *хмырь, карга, хрыч* и т. д. почти целиком охватывается содержательностью их звучания.

Признаковые коннотаты и фонетическая содержательность создают как бы семантические ореолы вокруг понятийных семантических ядер слов. Эти ореольные аспекты значения входят в общую лексическую и текстовую семантику, игнорировать их нельзя, иначе семантическая характеристика слова и текста окажется обедненной, неполной, а иногда и искаженной.

Для характеристики этих двух аспектов семантики в настоящее время уже существуют программы, реализованные на ЭВМ, причем оказалось очень удачным, что и тот и другой аспекты описываются сходным образом — в терминах одних и тех же измерительных признаковых шкал. Для машины это удобно, потому что она может не только описывать эти аспекты значения, но и автоматически сравнивать их и по требованию выдавать результаты такого сравнения.

Программы для анализа фонетической зависимости слов и текстов уже работают в Московском институте народного хозяйства им. Г. В. Плеханова, и с их помощью мы формируем специальный словарь фонетической значимости слов. Но для Машинного фонда нет необходимости заранее формировать словарь. Достаточно снабдить машину Фонда имеющимися программами и ввести в ее память необходимые исходные данные (их немного), по которым она по запросу пользователей сразу же рассчитает этот аспект семантики для слова, словосочетания или текста любого типа. В практическом плане немаловажно, что обращение пользователя к программам и получение результатов не требуют никакой специальной подготовки. Слово или текст можно вводить обычной печатью с телетайпа и получать результаты как в числовой, так и в содержательно интерпретированной форме.

Что касается коннотативно-признакового аспекта, то здесь дело обстоит сложнее. Некоторые материалы по описанию этого аспекта значения уже накоплены. Разработана и реализована программа автоматической классификации слов по их признаковым коннотатам. Так что на начальном этапе формирования Фонда можно было бы этими материалами воспользоваться. Но для основательного отражения коннотативно-признакового аспекта семантики в Фонде необходимо провести большую работу по экспериментальному измерению признаковых коннотатов русских слов. В принципе хорошо было бы иметь специальную лингвистическую (лингвометрологическую?) службу измерения этого аспекта русской семантики. Ведь в признаковых коннотатах отражаются не только лингвистические, но и социальные условия функционирования языка.

Анализ взаимодействия ореольных аспектов значения, основанный на материалах Фонда, дал бы возможность получить семантическую информацию, которой в настоящее время лингвистика не располагает и которую никаким другим способом получить нельзя. Опираясь на такие данные, можно было бы гораздо глубже рассматривать целый ряд общесемантических проблем. Скажем, проблему образования, функционирования и распада синонимических и антонимических рядов, которые часто выстраиваются и функционируют как бы в силовом поле коннотативно-признакового и фонетико-содержательного аспектов значения.

Явления полисемии тоже находят интересное преломление, если учесть взаимодействие этих ореольных аспектов семантики. Нередко переносы значений имеют в своей основе сходство семантических ореолов тех или иных понятий: *невеста—цветок*, *злодей—паук*, *тип—фрукт*, *тупица—дуб* и т. п. Видимо, возможно даже прогнозирование метафор, основанное на анализе ореольной семантики. Многие проблемы, связанные с историческими изменениями в значениях или в звучаниях слов, не могут получить всестороннего объяснения без учета взаимодействия содержательности звуковой формы и признакового значения слова.

Особые возможности данные Фонда об ореольной семантике открыли бы перед исследователями восприятия и воздействия речи, автоматического анализа и синтеза семантики речевых сообщений, а также проблем автоматической переработки речевой информации.

И если сейчас изучение семантических ореолов идет не столь активно, как они того заслуживают, то одной из причин этого можно считать «машинную» специфику их описания, чуждую большинству филологов. Материалы Фонда, избавив заинтересованных от математико-технических проблем, помогали бы ввести анализ этих аспектов семантики в лингвистический обиход. Думается, что такое большое дело, как создание Машин-

ного фонда русского языка, нужно делать так, чтобы предвидеть развитие семантических интересов и требований тех, кто будет этим Фондом пользоваться.

СОВРЕМЕННАЯ РАЗГОВОРНАЯ РЕЧЬ КАК СОСТАВНАЯ ЧАСТЬ МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА

Ю. Г. ОВСИЕНКО

Поскольку материал машинного фонда должен отражать свойства совокупности (русского языка в целом), его выборка должна быть достаточно представительна, охватывая по возможности все сферы речевого общения. В соответствии с этим при накоплении материала необходимо включать в него образцы разных функционально-речевых разновидностей общелитературного языка.

Начало этому уже положено. В лаборатории по применению вычислительных средств в гуманитарных науках МГУ закончена обработка материала современной русской разговорной речи, собранного на кафедре русского языка Университета дружбы народов и частично полученного из сектора Е. А. Земской (Институт русского языка АН СССР) и с кафедры О. Б. Сиротининой (Саратовский государственный университет).

На опыте уже проделанной работы (записи образцов живой разговорной речи и обработка собранного материала на машинах) можно сделать определенные выводы.

1. Необходима консолидация сил на этапе сбора материала, поскольку требуется единый подход к определению самого понятия «разговорная речь», а также единые критерии, предъявляемые к условиям сбора материала. Консолидация сил на уровне предварительной обработки материала необходима, так как живая разговорная речь нестандартна, некодифицирована. Материал ее очень труден для лингвистической интерпретации и зачастую даже отнесение слова к той или иной грамматической категории не находит однозначного решения.

2. Представительность выборки по отношению к разговорной речи может быть обеспечена при условии предварительного проведения серьезных социолингвистических исследований, которые приблизили бы нас к созданию модели речевого поведения.

3. Разговорная речь динамична — это живой, непрерывно меняющийся материал, на протяжении жизни одного поколения претерпевающий существенные изменения (на всех уровнях). Поэтому нужна большая мобильность в сборе и обработке такого материала с тем, чтобы каждые 20 лет можно было

бы получать сведения относительно тех явлений, которые произошли в разговорной речи. Это тем более важно, что существенная часть инноваций приходит в общелитературный язык из разговорной речи. Учитывая, что имеющиеся сейчас записи разговорной речи (в Москве, Ленинграде, Саратове) сделаны в основном с начала — середины 60-х годов по начало 70-х годов, очередные записи разумно начать в середине — конце 80-х годов, в противном случае многие интересные явления ускользнут от внимания исследователей, оставшись зафиксированными лишь в литературных источниках, а это материал неадекватный. Единственным источником изучения разговорной речи может служить лишь живая разговорная речь.

4. Сбор материала — процесс трудоемкий и требующий предварительных материальных затрат, поэтому необходимо уже сейчас подумать о создании соответствующей материально-технической базы.

5. По мере дальнейшего лингвистического исследования разговорной речи могут возрастать и меняться требования к системе записи (интонационные знаки, фонетическая транскрипция и т. д.), поэтому необходимо такое хранение фонетики, которое давало бы возможность в случае нужды обращаться к ней.

Без выполнения этих минимальных условий разговорная речь — живой памятник своего времени — не найдет должного представления в Машинном фонде русского языка. И потеря эта явится столь же невосполнимой, как утрата для нас живой речи предшествующих поколений.

Сохранить для потомков образцы непринужденной речи наших современников — одна из задач, вторая — чисто практическая. Полученные в результате машинной обработки словари разговорной речи, лингвостатистические справочники и т. д. имеют широкую область практического применения, и прежде всего при обучении русскому языку нерусских — изучение русского языка в национальных республиках, обучение русскому языку иностранцев.

Роль русского языка в современном мире велика, растет интерес к нему, увеличивается число изучающих русский язык. Материалом для изучения языка служат в основном письменные источники (художественная литература, публицистика, научные журналы и т. д.). Разговорная речь, не будучи зафиксированной и соответствующим образом обработанной, останется за рамками процесса обучения. Таким образом, изучающие русский язык не только не получат полного представления о нем, но и будут лишены надежного инструмента для практического овладения языком, овладения языком как средством общения.

Таков основной, далеко не исчерпывающий перечень задач, связанных с необходимостью включения в Машинный фонд русского языка живой современной разговорной речи.

О МАШИННЫХ ФОНДАХ РУССКОГО ЯЗЫКА ЗА РУБЕЖОМ

Ю. А. ТУЛДАВА

В связи с обсуждением проблемы создания Машинного фонда русского языка в Советском Союзе было бы интересно рассмотреть также вопрос о том, имеются ли какие-нибудь машинные фонды русского языка за рубежом. Наша исследовательская группа прикладной лингвистики при Тартуском университете имеет научные контакты с некоторыми центрами прикладной и вычислительной лингвистики за рубежом, в частности у наших ближайших соседей — в скандинавских странах. Они посылают нам свои материалы — препринты, отчеты и т. д. На этом основании я хотел бы сообщить о той работе, которая ведется в Скандинавии.

В г. Упсале (Швеция) работает Центр вычислительной лингвистики, руководителем которого является проф. А. Согвалл Хейн. В этом центре созданы машинные фонды целого ряда языков, в том числе русского языка. К содержанию фонда русского языка относятся словари и тексты. Например, введен в ЭВМ орфографический словарь русского языка (Ин-т русского языка АН СССР, 1963 г. издания) объемом 104 тыс. слов. Далее, в ЭВМ хранятся частотные словари Йоссельсона и Штейнфельдт, а также словарь основ русского языка, который применяется в специальных программах для исследования русских текстов, русской грамматики и лексики. Кроме словарей, в машинном фонде упсальской группы имеются русские тексты, по последним данным — около 60 текстов. Эти тексты относятся к жанрам художественной литературы (45%), научно-технической литературы (40%) и учебной литературы (15%). К текстам созданы частотные словники и даются различные лингвостатистические характеристики. Среди текстов художественной литературы имеются произведения русской классики и советской литературы. Из научно-технических текстов большинство относится к научно-популярному жанру. В машинном фонде имеются также такие тексты, как Программа КПСС, Устав КПСС и Конституция СССР.

В программной части упсальского машинного фонда имеется целая система программ, среди которых можно особо отметить программу ATEX для автоматизированного морфологического анализа русского текста.

Машинный фонд словарей и текстов русского языка упсальского центра применяется, судя по данным имеющихся отчетов и препринтов, при решении педагогических задач (преподавание русского языка, создание учебных пособий и учебников) и в лингвистических исследованиях (имеются сведения об исследо-

вании синтаксиса русского языка, о контрастивном изучении русского и шведского языков и т. д.).

Кроме Упсалы, в Скандинавии имеются еще другие центры, которые занимаются изучением русского языка и располагают машинными фондами русских словарей и текстов. Так, можно отметить Славяно-балтийский институт при университете г. Осло (Норвегия), где под руководством профессора Гейра Хетсо создан машинный словарь-конкорданс к поэзии Тютчева. Профессор Г. Хетсо и его группа занимаются также исследованиями по выявлению авторства некоторых русских текстов. Недавно было опубликовано предварительное сообщение о проблеме авторства отдельных публицистических текстов конца прошлого века (где ставился вопрос о принадлежности некоторых текстов Достоевскому). Широко известны работы проф. Г. Хетсо по изучению поэтического языка эпохи Пушкина, в которых используются данные машинного фонда текстов начала XIX в. (одна статья на эту тему опубликована в «Ученых записках Тартуского ун-та», 1981, вып. 585).

Центры прикладной и вычислительной лингвистики в Скандинавии тесно сотрудничают не только между собой, но и с подобными центрами в других странах. Например, обратный словарь чешского языка был предварительно создан в Австралии, доработан в Швеции (в Упсале) и напечатан в США.

Машинные фонды русского языка имеются не только в скандинавских странах. Можно полагать, что число и объемы таких фондов увеличатся в связи с развитием технологии вычислительной лингвистики и повышением интереса к изучению русского языка и литературы за рубежом. Представляется, что было бы полезно и целесообразно установить контакт с зарубежными центрами, где проводятся работы по созданию машинных фондов русского языка, и обмениваться опытом с теми, которые на это согласны. Это может дать импульс к более успешному решению хотя бы некоторых частных проблем создания Машинного фонда русского языка у нас в стране.

ТЕРМИНОЛОГИЧЕСКАЯ БАЗА ДАННЫХ АВТОМАТИЗИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ПО ОБЩЕСТВЕННЫМ НАУКАМ ИНИОН АН СССР

С. В. БАЗАРНОВА, Р. Р. МДИВАНИ

В Институте научной информации по общественным наукам АН СССР создается Автоматизированная информационная система по общественным наукам (АИСОН), предназначенная для обслуживания библиографической информацией в области общественных наук. В настоящее время такое обслуживание осущест-

вляется по трем областям знания: научному коммунизму, экономике и демографии, философии и социологии, в перспективе — по истории, государству и праву, литературоведению, языкознанию, науковедению. В поисковую базу данных включается комплексное описание содержания документа, состоящее из трех компонентов: библиографического описания, дополненного в большинстве случаев аннотацией; совокупности классификационных рубрик, определяющих принадлежность документа к определенной тематической области; совокупности дескрипторов и ключевых слов, описывающих содержание документа.

Для описания содержания вводимых в систему документов используется комплекс лингвистических средств, состоящий из словарей системы: отраслевых списков нормализованной лексики (в дальнейшем развитии — информационно-поисковых тезаурусов), отраслевых рубрикаторов, грамматики (указателей роли и связи) и методических инструкций.

Отраслевые списки нормализованной лексики — промежуточный этап создания отраслевых информационно-поисковых тезаурусов. Они созданы на основе результатов индексирования ключевыми словами реального потока советской и зарубежной научной литературы, поступающей в ИНИОН, и содержат весь набор лексических единиц, выделенных в сфере функционирования в качестве релевантных для задач информационного поиска. Лексические единицы в нем нормализованы в лексикографическом отношении и разделены на две категории: разрешенные для использования в системе (дескрипторы) и запрещенные (недескрипторы). При каждом дескрипторе списка указаны все лексические единицы, вместо которых он используется в системе (т. е. приведены все синонимы и квазисинонимы), а также для ряда дескрипторов приведены семантические пометы и краткие дефиниции. Таким образом, список нормализованной лексики отличается от информационно-поискового тезауруса только отсутствием двух видов парадигматических отношений — отношений выше — ниже и ассоциативных отношений.

Список нормализованной лексики содержит разные категории лексических единиц отраслевого языка науки, в первую очередь — терминологию, взятых из сферы функционирования. Сфера фиксации, т. е. различные терминологические словари, справочники, энциклопедии, привлекались главным образом при редактировании и нормализации лексических единиц и структурировании всего массива в целом. В этом состоит специфическая особенность списка нормализованной лексики как явления языка науки, предназначенного для информационно-поисковых задач.

Из сказанного выше следует, что список нормализованной лексики является ограниченным представлением языка науки, созданным в соответствии с прагматическими задачами системы, в первую очередь в связи с задачей описания содержания вводимых в систему документов. В каждой отрасли общественных

наук при этом должна быть решена проблема представления в соответствующем отраслевом списке нормализованной лексики разных терминосистем данной отрасли. Существование разных терминосистем внутри отраслей отмечается как широко распространенное явление в общественных науках. Их возникновение и использование обусловливается существованием многочисленных научных школ и направлений, использованием различной методики научного исследования, а также разным идеологическим подходом к рассмотрению научных проблем. Эти терминосистемы отличаются в количественном и, что более важно, в качественном отношении: план содержания разных терминосистем отличается набором представленных в нем понятий, семантикой терминов и соответственно системными отношениями между терминами. План выражения одних и тех же понятий в разных терминосистемах отрасли в целом ряде случаев различается, что приводит к синонимии. С другой стороны, план выражения разных понятий может совпадать, что приводит к омонимии.

Список нормализованной лексики для индексирования документов, созданных в рамках разных терминосистем одной отрасли, должен обладать достаточной семантической силой для описания этих терминосистем, т. е. служить в конечном счете метаязыком для соответствующих терминосистем. Специфика общественных наук состоит в том, что такой метаязык не может быть создан путем механического объединения существующих терминосистем в силу их принципиальных (в том числе и идеологических) расхождений. Вследствие этого такой список может быть создан лишь путем анализа разных терминосистем, в ходе которой выявляются расхождения между терминосистемами как в плане выражения, так и в плане содержания и затем синтеза лексических единиц в метаязыке. В результате образуется новый язык, который полностью не совпадает ни с одной из исходных терминосистем.

Процесс нормализации плана содержания и плана выражения лексических единиц, взятых из терминосистем отрасли, позволяет выделить все понятия, релевантные задачам информационного поиска, и выявить совокупность лексических единиц, соотнесенных с данным понятием (полные синонимы и условные синонимы, т. е. синонимы только в рамках системы). Из этой совокупности для каждого понятия по определенным критериям выбирается репрезентирующая его лексическая единица — дескриптор. Полисемичные лексические единицы, выбранные в качестве дескрипторов, снабжаются разного рода семантическими пометами и дефинициями, обеспечивающими точность толкования семантики дескриптора. Очевидно, что в процессе такой нормализации исходные термины, функционирующие в одной или нескольких отраслевых терминосистемах, и дескрипторы, совпадающие с ними в плане выражения, могут расходиться в плане содержания. С другой стороны, значения нескольких исходных терминов из разных терминосистем могут быть синтезированы

в одно семантическое поле метаязыка, т. е. представлены одним дескриптором списка нормализованной лексики. Таким образом, план содержания дескриптора эксплицитно представлен в виде собственно лексической единицы в сочетании с семантическими пометами и дефинициями, набором синонимов и квазисинонимов (в дальнейшем еще и набором парадигматических отношений выше — ниже и ассоциативных отношений). Устранение синонимии и омонимии терминов, выступающих в качестве дескрипторов в рамках системы, обеспечивает точность семантики и тем самым одно-однозначное соответствие между планом содержания и планом выражения дескрипторов, т. е. в дескрипторах в отличие от обычных лексических единиц в системе нивелирована асимметрия языкового знака.

Эксплицитное представление содержания дескрипторов обеспечивает их однозначное понимание, а также позволяет установить их соотношение с соответствующими терминами других терминосистем. Такое представление лексических единиц метаязыка позволяет эффективно пользоваться им как для описания содержания научных работ, созданных в рамках разных терминосистем, так и для описания содержания запросов потребителями, пользующимися разными терминосистемами, т. е. использовать его в качестве информационно-поискового языка.

Отраслевые рубрикаторы созданы на основе принятых в советской науке классификаций путем создания их модификаций в соответствии с практическими задачами системы. Каждая единица (рубрика) отраслевого рубрикатора представляет собой наименование определенной тематической области/группы, которое может быть различной длины — от однословного термина до нескольких назывных предложений. План содержания рубрики, ее место в структуре рубрикатора, ее парадигматические отношения, преимущественно—иерархические, определяют семантику каждой рубрики.

Отраслевые списки нормализованной лексики и отраслевые рубрикаторы являются взаимодополняющими средствами. Рубрикатор является средством упорядочения списка нормализованной лексики по крупным тематическим полям. Каждой рубрике рубрикатора соответствует определенное множество дескрипторов.

Методические инструкции по содержательной обработке документов включают в себя набор правил, позволяющих отразить синтагматические отношения между дескрипторами или рубриками, т. е. отношения, существующие только в сфере функционирования для конкретного документа. Содержание документа определяет то множество дескрипторов и/или рубрик, которые выделяются в качестве его репрезентации. Внутри этого множества они объединяются в семантические группы посредством использования указателей роли и связи и позволяют выделить связь между ними на уровне частей документа.

Таким образом, поисковая база данных в принципе служит

источником для выявления групп лексических единиц, с которой в сфере функционирования используется каждая нормализованная единица сферы фиксации (списка нормализованной лексики, рубрикатора), и тем самым сфера функционирования и сфера фиксации оказываются неразрывно связанными.

Списки нормализованной лексики и рубрикаторы образуют терминологическую базу данных АИСОН, созданную на основе сферы функционирования отраслевых языков науки. Как говорилось выше, эта база данных является метаязыком, отображающим отраслевые терминосистемы, которые используются в языке общественных наук. В то же время в поисковой базе данных АИСОН содержатся контексты для каждой лексической единицы в виде заглавий, аннотаций, дескрипторных описаний содержания документов (ПОДов), которые позволяют выявить синтагматические отношения для каждой терминологической единицы. Несмотря на различие требований, предъявляемых к дескриптору и термину, такая база данных позволяет вести широкие терминологические исследования в области общественных наук, а также создавать различного типа словники, в том числе и словники, содержащие термины с контекстами.

ПСИХОЛОГИЧЕСКИЕ, ТЕОРЕТИЧЕСКИЕ, ОРГАНИЗАЦИОННЫЕ И ТЕХНИЧЕСКИЕ АСПЕКТЫ СОЗДАНИЯ МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА

Р. Г. ПИОТРОВСКИЙ

Приветствуя и поддерживая идею создания Машинного фонда данных для автоматизированной системы лексикографических исследований, я хотел бы подчеркнуть, что успех этого грандиозного предприятия национального масштаба зависит от того, насколько будут учтены, правильно поняты и скоординированы психологический, теоретический, организационный и технический аспекты этого предприятия.

Сложность психологического климата, в котором придется работать создателям Машинного фонда, состоит в том, что здесь встретятся представители разного стиля научного и научно-технического мышления. Опыт разработки действующих систем машинного перевода — задачи, очень близкой по своему духу и существу к задаче создания Машинного фонда,— показал, что сотрудничество языковедов, привыкших к интуитивно-эмпирической и индивидуально-эстетической оценке объектов языка и речи, с логико-дедуктивно мыслящими математиками, которые часто отодвигают на задний план прагматическую сторону дела, а также с инженерами-программистами, которые, сохраняя дедуктивно-логический стиль мышления математиков-теоретиков,

подчиняют свои действия прагматике и реальным возможностям решения задачи, сразу же сталкивается с недостаточным взаимным пониманием и разночтением основной задачи. И дело здесь не в недостатке доброй воли со стороны участников таких инженерно-лингвистических предприятий.

Чтобы понять, почему десятки научных коллективов, пытавшихся в конце 50—60-х годов заняться решением проблемы машинного перевода, вынуждены были через несколько лет прекратить свои исследования и разработки, оставив потомкам наивно-романтические концепции языка-посредника, перевода через трансформации, теории типа «смысл — текст» и т. п., необходимо вспомнить о тех парадоксах, которые создают барьер отторжения, преграждающий путь автоматизации рече-мыслительной деятельности человека. Речь идет об антиномии «нечеткость языка человека и четкость языка ЭВМ», парадоксе языка и идиолекта, парадоксе актуализации, антиномии стрелы и других парадоксах, достаточно хорошо описанных в кибернетической литературе[1]. Эти парадоксы и являются причиной разночтения инженерно-лингвистических задач со стороны языковедов, с одной стороны, и математиков-теоретиков и программистов — с другой. Первые внутренне убеждены, что автомат также легко воспримет нечеткую, текучую материю языка, как это делает человек, вторые предполагают, что объекты естественного языка также дискретны и прозрачно-системны, как и объекты искусственных языков.

Чтобы преодолеть разночтение задачи создания Машинного фонда и недостаток взаимного понимания между языковедами, математиками-теоретиками и программистами, необходимы не столько организационные или дидактические меры, сколько выработка правильной теоретической стратегии и технологической тактики. В частности, помня о неудачах тех коллективов МП, которые пытались создать всеобъемлющие, закрытые системы МП, разработчики Машинного фонда должны, как мне кажется, идти по пути создания открытой модульной системы, способной к дообучению и адаптации к новым задачам, системы, учитывающей вероятностный характер лингвистических явлений и работающей не автономно, но в интерактивном человеко-машинном режиме. При таком подходе на начальном этапе разработки Фонда возникают две тесно связанные между собой задачи, без решения которых нельзя приступить к созданию Фонда. Первая состоит в определении характера и объема лексико-энциклопедического, грамматического, историко-лингвистического и сти-

[1] *Piotrowski R.* The Antinomies of Linguistics and Automatic Interpretation of the Text.— In: 3-rd International Congress of Applied Linguistics. Copenhagen, 1972; *Брушлинский А. В.* Почему невозможен искусственный интеллект.— В кн.: Кибернетика: Неограниченные возможности и возможные ограничения. Перспективы развития. М., 1981, с. 89—92; *Пиотровский Р. Г.* Лингвистические аспекты искусственного разума.— ВЯ, 1981, № 3, с. 30—35; *Дрейфус Х.* Чего не могут вычислительные машины: Критика искусственного разума/Пер. с англ. М., 1978.

левого материала, который должен быть включен в Фонд. Вторая заключается в определении структуры лингвистико-информационной базы Фонда (эта структура может иметь тезаурусный, линейносписочный или сетевой вид), а также в организации схемы словарной статьи [2].

Переходя к организационной стороне дела, я хочу поддержать идею, выдвинутую В. М. Андрющенко, о том, что организация работ по созданию Машинного фонда русского языка должна иметь строго иерархический характер. Во главе предприятия следует поставить небольшое число компетентных языковедов, математиков и программистов, наделенных реальными полномочиями и отвечающих за успех дела. Драматический опыт машинного перевода показывает, что решение инженерно-лингвистических задач нельзя пускать по пути «партизанской» самодеятельности, а во главе координирующих советов ставить менеджеров-любителей. Вместе с тем нельзя забывать о том, что построение Машинного фонда связано с обработкой колоссальных объемов лингвистического материала. В ходе подготовки нового толково-энциклопедического словаря французского языка, включающего 70 тыс. словарных статей, потребовалось обработать с помощью ЭВМ тексты объемом в 70 млн. словоупотреблений (т. е. около 200 тыс. страниц), а также ряд словарей [3]. Другим примером машинной обработки больших массивов лексической информации являются издаваемые в Англии под редакцией проф. Р. А. Висби словари-указатели и конкордансы [4]. К сожалению, у нас в стране опыт обработки с помощью ЭВМ больших массивов текста и лексикографических материалов невелик. Самое крупное лингвостатистическое исследование, выполненное с помощью ЭВМ,— «Частотный словарь русского языка» (около 40 тыс. слов. Под ред. Л. Н. Засориной. М.: Русский язык, 1977) — составлен на выборке в 1 млн словоупотреблений (т. е. около 3 тыс. стр.). В общесоюзной группе «Статистика речи» за 25 лет работы на ЭВМ переработано не более 8 млн. словоупотреблений (т. е. 24 тыс. стр.). Поэтому, мне кажется, следует отнестись самым внимательным образом к предложениям профессоров К. Б. Бектаева, Г. Г. Белоногова, А. Я. Шайкевича и др., указывающих на необходимость использовать имеющийся в стране готовый задел в области машинной лексикографии [5].

Благополучнее всего обстоит дело с программистско-технической стороной дела. Для создания, хранения и функциониро-

[2] *Чижаковский В. А., Беляева Л. Н.* Тезаурус в системах автоматической переработки текста. Кишинев, 1983.

[3] Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789—1960), publié sous la direction de P. Imbs. Tome premier P., 1971, p. XX—XLII.

[4] Compendia. Computer-generated aids to literary and linguistic research/Ed. R. A. Wisbey. Leeds, 1968—1979. Vol. 1—9.

[5] *Алексеев П. М.* Методика квантитативной типологии текста: Учеб. пособие/ /ЛГПИ им. А. И. Герцена. Л., 1983; *Нелюбин Л. Л.* Перевод и прикладная лингвистика. М., 1983.

вания Машинного фонда лексикографических данных имеется достаточно мощная и доступная вычислительная техника, созданы пакеты программ, позволяющих осуществлять лексико-статистический, дистрибутивный, синтаксико-семантический анализ текста, а также осуществлять выборку и разного типа упорядочение морфем, словоформ и словосочетаний[6]. Тем не менее и здесь предстоит еще много работы. Имеющиеся программы нужно свести в единый комплекс, а также дополнить этот комплекс недостающими компонентами, в частности программами, обеспечивающими интерактивный режим работы Фонда (в первую очередь человеко-машинный диалог через дисплей).

Сложны задачи и велик объем работ по созданию Машинного фонда русского языка. Однако можно быть уверенным, что при разумном сочетании психологического, теоретического, организационного и технического аспектов это грандиозное предприятие будет успешно завершено к началу следующего века.

О СОЗДАНИИ БАНКА ДАННЫХ ДЛЯ СИНТАКСИСА РУССКОГО ЯЗЫКА

И. П. БЕЛЕЦКАЯ

Наряду с созданием Машинного фонда русского языка применительно к исследованиям лексикографического плана было бы полезно и вполне своевременно приступить к созданию машинного фонда, включающего данные о типах поверхностно-синтаксических структур. Подобные данные должны храниться в ЭВМ не в форме словесных описаний (типа: «сложное предложение с придаточным времени» или «простое предложение с однородностью»), а в виде формальных структур. Удобный и хорошо разработанный для этого аппарат — деревья зависимостей (ДЗ). Коды, описывающие многочисленные разновидности ДЗ, без труда могут быть введены в ЭВМ. Таким образом, предлагаемый банк данных (БД) будет содержать тексты фраз и параллельно сведения об их синтаксической структуре, представленные в терминах деревьев зависимостей. Лингвисты без труда улавливают специфические особенности деревьев зависимостей, изображенных в виде рисунков (последние легко восстанавливаются из машинных кодов). Техника «чтения рисунка» напоминает навыки врача, рассматривающего рентгенограмму, или фонетиста, который интерпретирует осциллограмму. Так, например, из двух рисунков (см. рис. 1, 2) видно, насколько различными могут быть ДЗ даже для коротких фраз: развертывание рисунка дере-

[6] *Пиотровский Р. Г.* Текст, машина, человек. Л., 1975, с. 58—92; *Зубов А. В.* Обработка на ЕС ЭВМ текстов естественных языков. Минск, 1977; *Он же.* Программирование лингвистических задач на языке ПЛ/1. Минск, 1978.

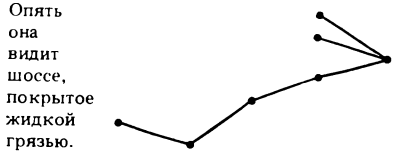


Рис. 1



Рис. 2

ва преимущественно вправо у Чехова и обилие левых дуг, а также перемены направления в пути графа (зигзаг на рисунке ДЗ) — у Шолохова (последнее связано с инвертированным порядком слов).

Библиотека ДЗ могла бы включать широкий набор сведений, полезных для лингвистических исследований по теории синтаксиса, по стилистике, а также для решения задач прикладного плана, связанных с автоматическим синтаксическим анализом. Перечислим некоторые из возможных задач данного плана:

сопоставление рисунков деревьев для синтаксических структур двух авторов;

неоднородность ДЗ как метод стилизации в рамках одного произведения;

особенности деревьев зависимостей стихотворных текстов;

закономерности чередования структур, следующих друг за другом в связном тексте;

поведение рисунков деревьев в аспекте диахронии;

психологические аспекты рисунков деревьев;

использование ДЗ как наглядного способа изображения синтаксических структур при обучении русскому языку иностранцев.

Кроме того, на материале ДЗ решается практически безграничное число задач статистического плана. Их решению предшествует установление характеристик дерева, которые служат объектами статистических подсчетов, например: количество уровней в дереве; ширина ветвления у корня; максимальное количество перемен направления пути в дереве; максимальная протяженность дуг в ДЗ и др. Вариации значений перечисленных примеров дают возможность различать синтаксические особенности стилей отдельных авторов, функциональных стилей и т. п. Существующие пакеты программ для стандартных статистических подсчетов вполне пригодны для исследований структур деревьев зависимостей. Здесь открывается широкий простор для автоматизации работ в области синтаксиса; полученные данные, хранимые в БД, служат сырым материалом, требующим лингвистических интерпретаций и комментариев к рисункам деревьев.

О НЕКОТОРЫХ ЛИНГВИСТИЧЕСКИХ АСПЕКТАХ ПРОБЛЕМЫ СОЗДАНИЯ МАШИННОГО ФОНДА РУССКОГО ЯЗЫКА

Ю. В. ВАННИКОВ

Сложность проблемы создания Машинного фонда русского языка на этапе постановки и предварительного обсуждения проявляется прежде всего в изобилии соображений и предложений, различных по масштабу, по значимости и нередко противоречивых. В этих условиях представляется принципиально важным разделить все предложения на первоочередные, т. е. предложения первого потока, и на предложения непервоочередные, т. е. предложения второго потока, более подробное рассмотрение и реализация которых может быть осуществлена во вторую очередь.

На какой основе должно быть проведено такое разделение? По-видимому, не на основе оценки теоретической актуальности тех или иных лингвистических проблем и не на основе оценки прагматической актуальности описания тех или иных аспектов и категорий языкового материала. Вообще понятие актуальности, как нетрудно убедиться, весьма относительно. То, что актуально сегодня, скорее всего не будет актуально завтра (если бы это было не так, понятие актуальности не имело бы смысла), но ведь Машинный фонд русского языка — дитя завтрашнего дня. Однако главное заключается, пожалуй, в другом. Обычно считается, что повышенный интерес, проявляемый к какой-либо лингвистической проблеме, вызван ее актуальностью. Но более внимательный анализ ситуации нередко приводит к противоположному выводу: именно тот факт, что данный вопрос привлекает к себе большее внимание, чем другие, и служит основанием, чтобы назвать его актуальным. Поскольку же, как известно, лингвистические исследования не организованы — да и не могут быть организованы — по принципу равномерного распределения внимания по всем возможным объектам, и выбор объектов для анализа зависит от многих факторов, в том числе и случайных, постольку и оказывается, что понятие актуальности в значительной степени имеет характер количественной оценки. Поэтому, думается, определять содержание и структуру будущего Фонда, исходя из сегодняшнего представления об актуальности теоретических проблем и важности описания определенных аспектов языкового материала, не является наилучшим решением вопроса.

Выделение первого и второго потоков подлежащих обсуждению идей и подлежащих реализации предложений должно осуществляться на основе выявления той роли, которую они могут сыграть в разработке и воплощении Машинного фонда. К проблемам первого потока по понятным соображениям следует отне-

сти тщательно обоснованное и достаточно детальное определение целей и функций Фонда. Сюда же должно быть отнесено и создание общей программы работ с выделением этапов и установлением очередности их выполнения, а также определение организационной структуры и материальной базы Фонда. Среди собственно лингвистических проблем к первому потоку надо отнести разработку содержания и структуры Фонда. Лингвистическое содержание и структура Машинного фонда, с одной стороны, очевидным образом связаны с его целевым назначением и функциями, а с другой — они зависят от единиц представления лингвистической информации. А вопрос о единицах представления, в свою очередь, связан с проблемой объема Фонда. Разумеется, желание ввести в Фонд как можно больше разной информации в виде самых разных единиц, взятых из самых разных источников — от энциклопедий и технических словарей до диалектных текстов, психологически вполне понятно. Но не следует забывать, что, чем большим будет объем Фонда, тем сложнее им будет пользоваться, и что неоднородность и неограниченность единиц представления является фактором, не благоприятствующим структурированию содержания Фонда и тем самым препятствующим системному подходу к объекту, оптимальному решению вопроса обработки, хранения, поиска и выдачи лингвистической информации. Таким образом, выбор единиц представления и системы их характеристик является центральным лингвистическим вопросом, который должен решаться уже на этапе постановки проблемы.

Единицы представления в Фонде не должны быть сведены к лексическим единицам языка. Фонд должен включать не только единицы плана выражения всех структурных уровней языка, но и единицы плана содержания. Введение в Фонд единиц разных аспектов языка и разных уровней языковой системы ставит вопрос об их отношениях. Принципиальные возможности сводимы здесь к двум решениям: (1) можно выбрать одну единицу в качестве базовой и все остальные сводить к ней или соотносить через нее; (2) можно включить все единицы в фонд на равных правах.

В первом случае следует разработать правила перехода к базовой единице «сверху» (от единиц более высоких структурных уровней) и «снизу» (от единиц более низких структурных уровней) с сохранением и интеграцией соответствующей лингвистической информации. Такой принцип построения был применен в некоторых информационных лингвистических системах, например в лингвостатистическом справочнике русской разговорной речи[1]. К аналогичному решению склоняются, по-видимому, и

[1] См.: работы: *Андрющенко В. М.* Некоторые проблемы автоматизации лексико-статистических работ.— В кн.: Вопросы лингвостатистического анализа русской разговорной речи. М., 1976; *Ванников Ю. В.* Лингвостатистический справочник русской разговорной речи и получение лингвистической информации высших уровней.— В кн.: Вопросы лингвостатистического анализа рус-

авторы книги «Анализ метаязыка словаря с помощью ЭВМ»[2].

Во втором случае следует разработать единую систему описания единиц различных аспектов и уровней и обеспечить возможность перехода с одного уровня структуры на другой (т. е. анализ и синтез различных единиц представления) с сохранением необходимой информации.

Преимущество первого решения заключается в возможности полного «вертикального» структурирования Фонда, что обеспечивает переход от единиц низших уровней к текстам и от текстов к единицам низших уровней. Недостаток первого решения (который, впрочем, можно рассматривать и как достоинство) состоит в том, что оно требует глобального подхода к объекту при введении его в Фонд: каждый текст или совокупность текстов должны анализироваться в полном объеме и на полную глубину, что, вероятно, не всегда удобно. Преимущество второго решения заключается в том, что относительная автономность представления единиц отдельных структурных уровней языка дает возможность парциального подхода к объекту, возможность построения Фонда отдельными ярусами.

Первое решение ориентировано на получение информации главным образом о синтагматических свойствах единиц языка, хотя оно и не исключает получения информации об их парадигматических свойствах. Второе решение ориентировано на получение информации главным образом о парадигматических свойствах единиц языка, хотя оно и не исключает получение информации об их синтагматических свойствах.

Вопрос о характеристиках единиц представления в лингвистических справочно-информационных системах уже имеет свою историю. Наибольший интерес представляют многомерные системы характеристик, призванные дать максимально полное представление о лингвистических и отобранных экстралингвистических свойствах языковых единиц. Укажем здесь в качестве примера систему регистров и параметров Эссекского «Проекта

ской разговорной речи. М., 1976; *Он же.* Грамматические связи и синтаксические функции: анализ и представление в лингвостатистическом справочнике.— В кн.: Исследования в области вычислительной лингвистики и лингвостатистики. М., 1978; *Он же.* Единая справочная лингвостатистическая система как база учебной лексикографии.— В кн.: Актуальные проблемы учебной лексикографии. М., 1977; *Он же.* Единая справочная лингвостатистическая система как база учебной лексикографии. Ст. 2.— В кн.: Переводная и учебная лексикография. М., 1979; *Городецкий Б. Ю.* Семантические проблемы построения автоматизированных систем обработки текстовой информации.— В кн.: Вычислительная лингвистика. М., 1976; *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. М., 1981; *Овсиенко Ю. Г.* Задачи описания разговорной речи и вопросы построения лингвостатистического справочника.— В кн.: Вопросы лингвостатистического анализа русской разговорной речи. М., 1976; *Она же.* Частотный словарь-справочник русской разговорной речи.— В кн.: Актуальные проблемы учебной лексикографии. М., 1977.

[2] *Караулов Ю. Н., Молчанов В. И., Афанасьев В. А., Михалев Н. В.* Анализ метаязыка словаря с помощью ЭВМ. М., 1982.

исследования современного русского языка»[3], систему детерминирующих факторов и признаков в разработке лингвостатистического справочника русской разговорной речи, систему 68 параметров в лексикографической концепции Ю. Н. Караулова. Поскольку вопрос о характеристиках единиц представления в Машинном фонде русского языка требует более детального рассмотрения, ограничимся здесь самыми общими замечаниями.

Система характеристик должна давать информацию не только о поверхностных единицах языка, но и о глубинной структуре, должна обеспечить возможность извлечения из Фонда информации, релевантной для различных (и даже несовместимых) теоретических моделей описания языка, должна обеспечить возможность эвристического комбинирования отдельных характеристик (т. е. возможность создания новых лингвистических объектов в смысле Ю. Н. Караулова[4]), что позволит извлекать из Фонда нетривиальную лингвистическую информацию.

Лишь достаточно полный набор единиц представления и их характеристик позволит построить Машинный фонд русского языка как большую систему, знаменующую новый этап в развитии лингвистической философии, методологии и технологии.

---

[3] The Contemporary Russian language analysis project. University of Essex Language Centre, 1965; *Хаден Р.* К установлению лексических минимумов с учетом разных «регистров» русского языка.— В кн.: Актуальные проблемы учебной лексикографии. М., 1977.

[4] *Караулов Ю. Н.* Лингвистическое конструирование: (О путях расширения практических приложений науки о языке).— Изв. АН СССР. Сер. лит. и яз., 1979, № 2.

ДИСКУССИЯ О МАШИННОМ ФОНДЕ РУССКОГО ЯЗЫКА

От редакторов

Всесоюзная конференция по проблемам создания машинного фонда данных для автоматизированной системы лексикографических исследований, состоявшаяся в феврале 1983 г., завершилась широкой дискуссией. Материалы дискуссии внесли существенный вклад в разработку программы создания Машинного фонда русского языка и в подготовку соответствующих проектных документов. Ниже помещаются выступления ее участников.

А. Б. АНТОПОЛЬСКИЙ

Можно считать, что в дискуссии обсуждаются две проблемы: из чего создавать Фонд русского языка и как его создавать. По первой проблеме хотелось бы отметить следующее. Отдавая должное направлению создания Машинного фонда собственно для лингвистических исследований, нельзя не признать, что более актуальной задачей, которая может оправдать затраты на создание такого Фонда, является задача обслуживания этим Фондом лингвистической промышленности, т. е. большое количество прикладных лингвистических задач, программ и фондов, создаваемых в различных организациях, затраты на которые весьма велики и могут, видимо, значительно сократиться при создании Машинного фонда. Институт межотраслевой информации, так же как и большинство информационных органов, имеет свои машинные словари, в наших условиях сейчас это десятки тысяч терминов в словарях различных типов, но эти словари реально могут обслужить только весьма узкий круг задач — документальный поиск (ориентированный на ручную обработку текстов). Переход к автоматическому индексированию, с одной стороны, переход к созданию и функционированию систем фактографических и далее — задач искусственного интеллекта, с другой стороны, интеграция в рамках информационных сетей — с третьей, требует резкого увеличения объема и сложности используемых словарей в соответственно имеющейся в этих словарях лингвистической информации. Эта задача стоит не только перед нами, но и перед большинством информационных систем по мере их развития и интеграции. Мы уже, как правило, не можем увеличивать ресурсы на создание, поддержку, ведение в автономном режиме больших словарей — делать так, как было до сих пор, просто невозможно.

Много внимания было уделено инвентаризации лексики русского языка. Остановимся на лексике, которая используется в научно-технической литературе. Как можно оценить массив терминов? Никто не знает, сколько в русском языке терминов, но некоторой аппроксимацией этого объема может слу-

жить существующий на сегодня фонд в Австрии, где собраны и зарегистрированы тезаурусы,— фонд, который составляет примерно 0,5 млн. лексических единиц. Это далеко не все, но какую-то оценку дать можно. Этот фонд составляет очень небольшую часть лексики русского языка, если иметь в виду не только терминологию как таковую, но и то, что называют обычно номенклатурой с включением разного рода собственных имен. Это и есть основной в количественном отношении пласт русского языка, к машинной инвентаризации которого практически мы еще не приступали, а если приступали, то настолько разрозненно, что об этом всерьез никто и не думал в системном плане.

Приведу несколько цифр, чтобы было понятно, о каких объемах надо говорить. Скажем, химических соединений — порядка 5 млн. названий, геологических видов — порядка 3 млн., номенклатура изделий, только выпускаемых советской промышленностью (имена), — порядка 10 млн., географических наименований, зарегистрированных у военных картографов,— порядка 5 млн., советских юридических лиц — порядка 1 млн., зарубежных — соответственно столько же и т. д. Количество лексики такого рода невероятно велико, можно оценивать его в десятках миллионов лексических единиц. Причем к ним нельзя подходить как к словам.

Конечно, создание Машинного фонда не может опираться только на сбор и анализ пословной информации. Необходимо иметь информацию о лексической единице, как она устойчиво употребляется в тексте, особенно номенклатура, термины и т. д. Почему? Во-первых, для задач перевода, включая машинный; семантическую парадигматику отобразить на отдельных словах невозможно (в задачах не только информационного поиска, но и искусственного интеллекта). Парадигматика может быть задана только на термине — словосочетании, обозначающем понятие научно-технического языка.

Поэтому без инвентаризации терминов — словосочетаний, представляющих научные понятия, видимо, обойтись в Машинном фонде нельзя. В этом смысле я согласен с В. М. Андрющенко, что разумно Фонд делить на некоторые части: фонд словарный и фонд терминологический. Информация там должна собираться разная и должна взаимно дополнять друг друга.

В связи с этим хотелось бы перейти к организационным практическим проблемам, которые затрагивались весьма недостаточно. Во-первых, нужно опираться на имеющиеся фонды лексики, которых в стране насчитывается уже, видимо, несколько десятков, если не сотен, но которые информацию хранят по-разному, описывают по-разному. Отсутствует какая-либо координация в сборе, вводе, выводе и т. д. Так что первой задачей является координация действий нынешних фондов. Для этого необходима разработка даже до перехода к фонду как распределенному банку данных, создание единого языка описания лингвистических данных, включая описания и семантические, и морфологические и т. д., а также некоторый язык лингвистических алгоритмов обработки и доступа к этим данным. В качестве приближения или первой пробы пера в этом отношении можно считать имеющийся формат описания лексической терминологической информации, разработанный в рамках ИНФОРТЕРМа, и формат МАТЕР и его отечественную интерпретацию — формат ФОЛИЯ. Такой формат описания лексико-терминологической информации может послужить некоторой основой, хотя он, конечно, беден для тех грандиозных задач, которые здесь обсуждаются, но тем не менее надо с чего-то начинать. Далее, необходимо перейти к практической координации работы по

части распределения банка данных на содержательном уровне. Какие собственно лексические пласты, массивы? Кто собирает, кто хранит, как обеспечивают к ним доступ со стороны партнеров по фонду? В каком виде представляют эти данные и т. д.? Я считаю, что поставленная В. М. Андрющенко задача распределенного банка данных в узком смысле этого слова, задача очень приятная, но в ближайшие 5—8 лет, на мой взгляд, нереальная. Мы еще не имеем распределенных банков даже для более зрелых в техническом отношении систем (имеются в виду распределенные банки, находящиеся под сетевым мониторингом). Это, видимо, задача 90-х годов. А в 80-х годах реально сделать набор локальных банков, обменивающихся информацией на магнитных лентах. Но в такой постановке необходимо решить колоссальное число сложных технических проблем, возникающих в этой связи. Те, кто занимается разработкой автоматизированных систем, хорошо представляют себе, что такое сделать конкретную информационную систему, тем более распределенную, сложную и т. д. Нужна и Целевая программа, нужен заказчик для этой системы, нужно финансирование, материально-техническое снабжение, нужен координационный центр в достаточной степени снабженный, который мог бы осуществлять эту прямую координацию и в научном, и, главное, в организационно-техническом плане, кадровое обеспечение и т. д. Без этого о существенных сдвигах по сравнению с нынешним положением говорить преждевременно.

А. С. АСИНОВСКИЙ

Определенный интерес представляет опыт ввода в ЭВМ русского словообразовательного словаря Уорта. Этот словарь представляет собой 110 тыс. слов русского языка, разделенных пробелами на морфемы и собранных в 10 тыс. гнезд, единство которых обусловлено единством корня. При вводе в память ЭВМ каждое слово получило индекс части речи в соответствии с традиционным представлением о частях речи. Этот опыт был предпринят в связи с разработкой модели автоматического распознавания связного устного текста на русском языке. Такая модель разрабатывается на кафедре фонетики ЛГУ совместно с Лабораторией речи Института математики Сибирского Отделения Академии наук. Выбор единицы словаря распознавания обусловлен следующими свойствами морфемы. Морфемное описание слова дает экономию памяти как минимум в 10 раз по сравнению со словарем слов и в 100 раз по сравнению со словарем словоформ. Второе, морфемы — это наиболее приближенный участок слова к так называемым опорным участкам, получаемым на выходе первичной обработки сигнала. Эта приближенность к опорным фонетическим участкам дает возможность использовать морфемный словарь распознавания, содержащий дистрибуцию каждой морфемы, для адекватной интерпретации диффузных, слабых, фонетически не интерпретированных участков. Так, если мы получим надежно распознанный сегмент словоформы, то отождествление его с элементом морфемного словаря распознавания позволяет рассматривать только те гипотезы о внутрисловных контекстах этого сегмента, которые зафиксированы в словаре. Третьим обстоятельством, побудившим нас обратиться к морфологическому представлению слова в словаре распознавания, является такое свойство морфемы, как ее наиболее однозначная по сравнению со словом соотнесенность с лингвистическими категориями, используемыми при синтаксическом и семантическом анализе текста. Так, если

для словаря морфема является средством экономного описания его единиц, то для текста морфема представляется универсальным показателем согласования, и тип согласования — классификатор типов морфем. В связи с этим существенным оказывается такой момент, как возможность автоматической синтаксической индексации транспонирующих или реляционных суффиксов. Эта индексация может быть предпринята на базе морфемного словаря в его машинном виде. Поскольку каждое слово морфемного словаря заиндексировано, возникает возможность рассмотреть все гнезда, в которые входят слова с данным суффиксом, и соотнести частеречное значение слова с этим суффиксом с частеречным значением слова, отличающимся в каждом данном гнезде отсутствием описываемого суффикса и, таким образом, получить так называемый реляционный индекс этого суффикса, состоящий из двух знаков, левого — это частеречное значение слова без этого суффикса, и правого — частеречное значение слова в том же самом гнезде с данным суффиксом в сильной позиции. Под сильной позицией имеется в виду финальная позиция суффикса в основе. Такая индексация позволяет обойтись без хранения информации о частеречной принадлежности каждого слова, использовать реляционные индексы для определения частеречных значений слов, содержащих суффиксы, таким образом заиндексированные. Использование хотя бы минимальной информации такой, как частеречное значение слова, определение по правому знаку реляционного индекса суффикса в сильной позиции элементарных значений падежей дает возможность автоматически получить словарь словосочетаний по тексту и переорганизовать этот словарь словосочетаний в виде семантического поля текста любого объема. В связи с задачами, о которых говорилось, предполагаются следующие этапы работы по переорганизации морфемного словаря. Первое — это получение трех списков элементов словаря, трех списков морфем — префиксы, корни, суффиксы, с адресами всех слов, в которых эти морфемы содержатся в определенной позиции, и получение частотных характеристик каждого элемента. Второе — это синтаксическая индексация суффиксов по описанному способу. Третье — это словоизменительная индексация. Но поскольку словарь содержит только начальные формы, а в реальном тексте мы работаем со всем объемом, доходящим до 200 словоформ на каждое слово, в словарь вводятся словоизменительные парадигмы. Четвертое — транскрибирование словаря, проводимое также в алгоритмическом виде и позволяющее подробнее рассмотреть фонетические структуры русских внутрисловных границ, приблизить, смоделировать работу со звучащим текстом. В заключение хотелось бы особенно отметить, что, говоря о морфемном словаре, мы не предлагаем альтернативного решения по поводу выбора единицы словаря для Машинного фонда русского языка. Речь идет только о морфологически описанном слове, об удобном способе записи основных единиц словаря.

К. Б. БЕКТАЕВ

Все проблемы, изложенные Ю. Н. Карауловым, А. П. Ершовым и В. М. Андрющенко, упираются в создание лингвистического банка данных. В связи с этим хочу поделиться нашим опытом работы в этом направлении. Во-первых, всем хорошо известно, что большу́ю трудность представляет подготовка лингвистической информации для ввода в память ЭВМ. Поэтому необходимо за-

ранее продумать систему индексации тех лексико-грамматических и синтактико-семантических характеристик, которые надо извлечь из текста. Так, например, при лингвостатистическом исследовании тюркских языков по данным текстов художественных произведений (казахский), газетно-публицистических текстов (казахский, узбекский, киргизский, каракалпакский), научно-технических (узбекский, казахский), устно-разговорных (казахский), детской литературы (казахский, узбекский), учебников начальных классов (казахский) и произведений отдельных известных писателей и поэтов (казахский, узбекский) каждое словоупотребление было помечено заранее разработанным лексико-грамматическим индексом [1]. Таким образом, по этим текстам с помощью ЭВМ были получены алфавитно-частотные, обратно-частотные и частотные словари словоформ, по которым был проведен и проводится лингвостатистический анализ. С помощью этих материалов защищено несколько кандидатских и одна докторская диссертация. Как известно, сам частотный словарь без лексико-грамматических индексов (особенно для тюркских языков) с лингво-теоретической точки зрения особого интереса не представляет, а словарь с лексико-грамматическими пометами имеет весьма большое теоретическое и прикладное значение. Эта методика была использована при составлении частотных словарей по тексту М. Ауэзова «Путь Абая», 460 тыс. словоупотреблений [2]. Общий объем казахских текстов, введенных в память ЭВМ, в настоящий момент составляет 20 млн. байтов. Каждое слово мы измеряем 8 байтами. Во-вторых, для создания Машинного фонда русского языка большое значение имеет проблема ввода в память ЭВМ текстов отдельных произведений крупных писателей или текстов их полного собрания сочинений. Например, в Институте языкознания АН КазССР с помощью ЭВМ начали осуществлять поистине грандиозную работу по составлению словаря по текстам полного собрания сочинений Мухтара Ауэзова. В грандиозности этой работы можно убедиться сопоставлением времени, затраченного на выполнение подобной работы при составлении словарей по произведениям Абая и Пушкина. Известно, что на составление словаря языка Абая десять человек затратило 5 лет, а словарь языка Пушкина составлялся почти в течение 20 лет. Основное время составителей ушло на выполнение первого этапа работы, т. е. на составление картотек. Подсчитано, что текст произведений М. Ауэзова в 20 книгах (30 млн. байтов) превышает обработанный текст произведений Абая в 60 раз, а Пушкина — в 7—8 раз. Отсюда нетрудно подсчитать, сколько лет потратили бы люди на составление словаря М. Ауэзова. Следовательно, без использования ЭВМ эта работа не может быть выполнена, для этого потребовалось бы по крайней мере 30 лет, а институт планирует закончить и выпустить словарь в течение 10 лет. Тексты произведений М. Ауэзова, введенные в память ЭВМ, явятся особо ценным материалом для исследований не только по казахскому, но и по другим тюркским языкам. Это говорит о том, что при создании Машинного фонда русского языка необходимо вводить в память ЭВМ и тексты произведений крупных русских писателей и поэтов.

---

[1] См.: *Бектаев К. Б.* Статистико-информационная типология тюркского текста. Алма-Ата, 1978.

[2] См.: *Бектаев К. Б.* и др. Частотный словарь по роману М. Ауэзова «Абай жолы». Алма-Ата, 1979.

В каждой союзной республике большое значение придается терминологическим вопросам, особенно русско-национальным терминологическим работам. В Казахстане выпущено в настоящее время более 50 русско-казахских терминологических словарей по различным отраслям науки и техники. Но тем не менее при составлении словарей обнаруживается ряд недостатков. Поэтому Терминком Казахской ССР поставил перед лингвистами задачу устранения этих недочетов. Для решения этой задачи Институт языкознания АН КазССР решил объединить все изданные русско-казахские терминологические словари так, чтобы при каждом слове был указан его адрес, т. е. источник, из которого было извлечено то или иное слово (термин). Этот вопрос решается с помощью ЭВМ. Мы сейчас ввели в память ЭВМ около 50% терминологических словарей. Здесь имеется образец объединенного словаря по гуманитарным наукам. На его основе был получен казахско-русский словарь путем простого переворачивания. При внимательном анализе этих словарей можно убедиться, какую неоценимую пользу может дать ЭВМ для создания русских, национальных, русско-национальных, а также национально-русских терминологических словарей.

С 1976 г. в Чимкенте осуществляется промышленный машинный перевод текстов по химии полимеров с английского языка на русский. При этом создаются большие автоматические словари (словоформ и оборотов), которые также имеют большое значение в деле создания Машинного фунда русского языка. Недавно нам заказчик предложил текст по химии полимеров, который необходимо было перевести с русского языка на английский. Но для этого пришлось «переворачивать» наш англо-русский автоматический словарь, т. е. из него составить русско-английский автоматический словарь. С помощью машины мы это сделали за одну ночь и получили алфавитный русско-английский автоматический словарь. Правда, реализовать машинный перевод было почти невозможно, потому что семантическая покрываемость оказалась очень низкой. Новый словарь имеет большое самостоятельное значение.

При работе с частотными словарями разных языков по разным тематикам надо было выявить информационно-статистические характеристики этих словарей. Однако работа над этими словарями показала, что эти характеристики несопоставимы друг с другом. Поэтому мы были вынуждены обработать их заново на ЭВМ и сделать эти статистические характеристики сопоставимыми. Мне кажется, что здесь нужно обратить особое внимание на то, чтобы при создании Машинного фонда русского языка статистико-информационные характеристики, извлекаемые из этих словарей, были сопоставимы, поскольку материалы будут готовиться не только в Москве, но и по всему Советскому Союзу. В противном случае мы не сможем извлечь информацию, необходимую для лингвостатистического анализа.

Очень радует энтузиазм в деле создания Машинного фонда русского языка. Научный энтузиазм — это первое и идеальное условие успеха в разработке той или иной проблемы. Второе условие — это техника организации работы. Поэтому, естественно, большое внимание вызывает вопрос о том, из какого строительного материала будет состоять Машинный фонд русского языка и какие теоретические и прикладные выводы можно сделать. Как выявилось, в стране уже имеются машинные фонды словарей. Некоторые товарищи прямо заявили о том вкладе, который они могут сделать в создание Машинного фонда русского языка. Например, Л. Н. Беляева из Ленинграда, А. В. Зубов из

Минска безвозмездно передают результаты десятилетней работы целого коллектива. Это надо приветствовать. О своем взносе сказали также доктора филологических наук А. Я. Шайкевич и М. Н. Реммель.

Чимкентский коллектив может выделить для создания Машинного фонда русского языка одну перфораторщицу, которая в течение года сможет заперфорировать текст объемом 300—400 тыс. словоупотреблений, чтобы передать этот материал фонду безвозмездно.

В создание Машинного фонда активно могут включиться кафедры русского языка ряда институтов Казахстана. Но для этого нужно конкретно определить, какой фонд будем мы готовить. Думается, что для решения этого и ряда других вопросов надо создать рабочую группу под руководством Ю. Н. Караулова, А. П. Ершова, Ю. В. Рождественского и Ю. Н. Марчука. Эта рабочая группа должна координировать всю работу по созданию Машинного фонда русского языка: определить темы исследований и распределить темы, разработать инструкции по подготовке исходных данных, а также учесть все реальные взносы рабочих коллективов во избежание дублирования проделанной работы.

Р. Г. Котовым и Ю. В. Рождественским были затронуты некоторые вопросы, касающиеся использования ЭВМ в деле составления терминологических словарей. Наш опыт работы над русско-казахскими терминологическими словарями показал, что необходимо создать единую основу в русской части отраслевых терминологических словарей. Эта работа имеет большое значение, особенно для национальных республик, так как нет даже единообразия в записи и подборе терминов в терминологические словари. Например, если взять словосочетание «абсолютная величина», то в одном словаре пишут «абсолютная величина», в другом «величина, абсолютная», в третьем — «величина (без запятой) абсолютная» и т. п.

На одном из заседаний Координационного совета по машинному переводу выступил представитель кафедры, возглавляемой Ю. В. Рождественским (Московский университет), который в своем докладе говорил о грандиозном плане создания отраслевых терминологических словарей. Должен сказать, что вручную создать их сейчас невозможно. А создание их с помощью ЭВМ сопряжено с большой трудностью. После этого доклада нами была продумана методика создания машинного фонда русских терминологических словарей. Она заключается в следующем: совместно со специалистами отбираются учебники и монографии. Читая эти книги, специалисты отмечают терминологические слова и словосочетания. После повторной проверки подбираются переводные казахские эквиваленты. Собранный материал перфорируется с книжки и далее записывается в память ЭВМ. В конечном счете по каждой отрасли науки мы получаем термины (русский, русско-казахский) с указанием адреса, т. е. отрасли науки, названия (номера) книги и страницы. Когда данные занесены в память ЭВМ, то из них можно извлечь всю информацию, которая нас интересует.

Успех этой работы зависит от дружбы и взаимопонимания между членами уже сформировавшихся и имеющих большой опыт коллективов. Хочется сказать несколько слов по поводу мнения профессора В. Ю. Розенцвейга. Многие программистские и лингвистические задачи, изложенные им, уже реализуются в коллективе группы «Статистика речи». Например, В. Ю. Розенцвейг отметил, что в Западной Германии 150 тыс. немецких терминов зало-

жены в память ЭВМ; в нашей стране только в Чимкенте записано (я не говорю о других отделениях группы «Статистика речи») 100 тыс. терминов из русско-казахских терминологических словарей и около 50 тыс. терминов из переводных эквивалентов англо-русского словаря по химии. Поэтому прежде всего нам надо изучить то, что делается у нас, в Советском Союзе, по автоматической обработке текста.

А. С. Герд отметил, что многие энтузиасты уже давно отошли от машинного перевода. Это действительно так. Но отошли эти лингвисты и математики от машинного перевода по той причине, что их подход оказался тупиковым. Эти ученые не до конца учли лингвистические, инженерно-математические парадоксы и трудности машинного перевода. Несмотря на эти трудности и неудачи, экспериментальный машинный перевод в нашей стране существует и не только в традиционном англо-русском варианте, но и как МП с восточных языков. В данное время Чимкентский коллектив работает над повышением качества машинного перевода. Для этой цели используется дисплей: машина высвечивает на экран телевизора перевод текста, в котором можно произвести любой вид корректировки. Наш подход к реализации машинного перевода отличается от всех существующих в нашей стране. Он прежде всего ориентирован на перевод текстов определенной отрасли науки и узкого класса документов.

Итак, главный успех в работе по созданию Машинного фонда русского языка — это дружба всех энтузиастов.

Г. Г. БЕЛОНОГОВ

Когда А. П. Ершовым и Ю. Н. Карауловым была провозглашена программа создания Машинного фонда русского языка, аудитория, насколько я имел возможность наблюдать, возликовала. Почему? Потому что к этому давно стремились, и очень знаменательно, что первая конференция по Машинному фонду русского языка возглавляется одним из лидеров в области применения электронной вычислительной техники и директором Института русского языка. Это достаточно символично, поскольку союз именно этих двух сил является залогом успеха.

Мне всегда импонировало у лингвистов особое уважение к материи, к естеству языка. Оно и здесь прозвучало в ряде выступлений, например, о том, что древнерусские тексты нужно вводить в ЭВМ очень тщательным образом, без всяких искажений. Или другой прозвучавший здесь тезис о примате текстов или, в крайнем случае, словоформ по отношению к другим единицам языка, получаемым путем обработки текстов, в соответствии с той или иной лингвистической теорией. Я симпатизирую такой точке зрения.

Мне кажется, что Машинный фонд русского языка должен, как уже здесь было сказано, иметь два аспекта: первый — исследовательский, языковедческий; второй — сугубо прикладной, так сказать, народнохозяйственный. К реализации первого аспекта лингвисты давно стремятся, и в этой связи очень жаль, что наш ведущий академический Институт русского языка еще недостаточно опирается на электронно-вычислительную технику в лексикографической работе и в исследовании текстов. Второе направление связано с информатизацией общества. В настоящее время дело обстоит таким образом, что тот, кто отстал в области автоматизации обработки информации, тем самым

отстал и в наиболее существенном направлении развития научно-технического прогресса. Поэтому задача создания и изучения Фонда русского языка — это не задача первой трети будущего века, как здесь кто-то сказал, а актуальнейшая задача современности.

Чтобы поднять научный и практический уровень автоматической обработки научно-технической информации и информации в автоматизированных системах управления, надо вести интенсивную работу сегодня. Но, к сожалению, некоторые академические ученые-лингвисты и университетские преподаватели не понимают этого и скептически, я бы даже сказал враждебно, настроены к методам исследования языка, связанным с нуждами практики автоматической обработки информации. Кое-кого почему-то смущает термин «структурная лингвистика», хотя он обычно воспринимается как обозначение структурного, алгоритмического подхода, т. е. подхода, связанного с формализацией информации. А чтобы решать задачи с помощью ЭВМ, необходимо, как известно, уметь формализовать информацию. Пора бы прекратить безответственные разговоры о якобы идеологической несостоятельности структурной лингвистики, считать их реакционными и вредными.

Создание Машинного фонда русского языка — это не кратковременная кампания, а явление перманентное. И началось оно не сегодня. Здесь уже говорили о том, что в нашей стране создана Государственная автоматизированная система научно-технической информации, которая включает десять всесоюзных центров, такие, как ВИНИТИ, ИНИОН, ВНТИЦ и др., порядка 70 отраслевых центров (по числу министерств), 15 республиканских центров, региональные центры и т. д. И в этой системе создана так называемая магнито-ленточная служба. В частности, ВИНИТИ выпускает в год на машинно-читаемых носителях примерно 800 тыс. описаний документов. Это — безусловно машинный фонд русского языка, поскольку информация представляется на русском языке. Эта информация тиражируется и распространяется во многие центры. Следовательно, в системе научно-технической информации Машинный фонд русского языка уже реальность.

В. Н. БЕЛООЗЕРОВ

Развитие форм машинного существования языка является коренным преобразованием не только в самом языке, но и заметным этапом развития человеческого общества в целом, который сравним с такими этапами, как возникновение письменности и возникновение самого языка. Наряду с устной речью и письмом теперь становится реальностью существование языка в машинном виде, которое несводимо к его письменной форме, отличается от нее на всех уровнях, равным образом и в синтактике, и в семантике, и в прагматике. Становление машинной формы существования языка знаменует очередной шаг в отстранении языка от человека. Если устная речь непосредственно физически связана как с говорящим, так и со слушающим, то в актах письменной речи человек присутствует уже только односторонне: либо как пишущий в акте писания, либо как читающий в акте чтения; при этом взаимодействие читающего с пишущим значительно редуцируется, а языковое произведение приобретает характер самодостаточного переносчика информации. В актах машинной речи мы теперь можем наблюдать видимость полного отсутствия человеческого фактора. Такое максимальное отстранение речи от человека

происходит, например, при информационном взаимодействии в сетях автоматизированных центров научно-технической информации: документы, подлежащие обработке, автоматически описываются какими-то ключевыми словами в соответствии с какими-то правилами автоматического индексирования, потом фонды этих документов распределяются на основе автоматического сравнения этих ключевых слов с автоматически полученными поисковыми образами запросов других информационных центров, а человек остается вне этого процесса обработки информации, выраженной языковыми (лексическими) средствами, участвуя лишь в начальном этапе построения всей системы и в конечном звене — при оценке ее эффективности.

По мере все большего отстранения речи от человека увеличивается потребность в нормализации использования языка. Отсутствие человека в речевых процессах лишает возможности вносить в преобразуемые сообщения коррекцию в зависимости от результатов его преобразования. Поэтому требуется бо́льшая точность, меньшая вариативность в соотнесении речи с ее содержанием, которая при контактном речевом взаимодействии людей компенсируется внеязыковыми каналами передачи информации. Повышенные требования к точности речи ведут к разработке специальных правил использования языка, которые причудливым образом взаимодействуют с правилами естественной грамматики, трансформируют их и при том настолько, что заставляют говорить об искусственной грамматике, которая заменяет естественную грамматику языка в информационно-поисковых системах. Правда, в настоящее время на облик грамматик информационно-поисковых систем в большой степени оказывает влияние необходимость применяться к небогатым возможностям вычислительной техники, но было бы ошибкой считать, что с ходом развития машинной формы речи она будет приближаться к письменной форме. Напротив, в некоторых отношениях она будет даже отдаляться от нее, как отдалялась письменная речь от устной с развитием практики письма.

А развитие машинной формы речевых процессов, по-видимому, можно отнести к тем, по выражению А. П. Ершова, безальтернативным инновациям в обществе, которые не могут не произойти и которые будут определять развитие общества в дальнейшем. Поэтому, видимо, был прав Ю. В. Рождественский, утверждая, что судьбы национальных языков в настоящее время решаются на фронте информатики, как в свое время они решались на фронте письменности. Насколько язык сумеет войти в автоматизированные системы, настолько ему будет обеспечено развитие и будущее в обозримой перспективе.

Эти обстоятельства, а именно необходимость, с одной стороны, способствовать использованию русского языка в автоматизированных системах, а с другой — нормировать это использование, были, видимо, впервые осознаны практическими работниками информатики. Это осознание около 20 лет тому назад вылилось в постановку проблемы создания всесоюзного центра, ведущего инвентаризацию, систематизацию и управляющего развитием информационно-поисковых языков. Этот центр создан в настоящее время во Всесоюзном научно-исследовательском институте технической информации, классификации и кодирования (ВНИИКИ) под именем Автоматизированная система ведения информационных языков (АСВИЯ). По мере создания этого центра от него отпочковались еще две упоминавшиеся здесь системы — Автоматизированная система информационно-терминологического обслуживания (АСИТО) и Автоматизированная система ведения общесоюзных классификаторов

(АСВОК). Каждая из этих систем имеет богатые банки лексических данных, по-разному отражающие лексику и грамматику языка русской научной и технической литературы, преломленные под углом зрения их использования в автоматизированных системах.

Таким образом, можно считать, что первый шаг в создании автоматизированного фонда русского языка уже сделан — создан фонд русской научно-технической терминологии. В общей структуре будущего единого фонда русского языка, которую так четко обрисовал В. М. Андрющенко, та часть, которая направлена на деловую научно-техническую прозу, уже можно считать, вчерне готова. Однако, кроме научно-технической части, кроме лексики информационно-поисковых языков, в Фонде русского языка просто необходимо существование других его частей, включающих традиционные языковые источники, традиционные филологические словари. Дело в том, что словари информационно-поисковых языков, собираемые в систематизируемые фондом сейчас, являются весьма своеобразным языковым феноменом, резко отклоняющимся от традиционной лексикографической практики, и возникает опасность разрыва преемственности языковой традиции в развитии машинной формы существования русского языка. А обеспечить эту преемственность можно только тем, что включить достаточно представительный корпус традиционных филологических словарей в автоматизированный фонд и наладить взаимодействие между частями фонда. Поскольку различные части фонда по необходимости будут создаваться в географически и административно различных учреждениях, необходимо создать Центральную службу фонда (например, в Институте русского языка АН СССР), которая координировала бы взаимодействие частей фонда и его развитие. Взаимодействие предполагает налаженный обмен массивами данных, а для этого необходимы средства обмена: общие форматы описания данных, коммуникативные форматы представления данных на машинных носителях. Разработка таких форматов ведется уже сейчас для целей обмена научно-технической информацией без учета потребностей предполагаемого Фонда русского языка. Языковеды могли бы много сделать уже сейчас для успешного функционирования фонда, если бы более активно включились в эти работы — разработку ГОСТов на коммуникативные форматы для лексикографической и библиографической информации, аналогичные форматы СЭВ и международной терминологической сети ИНФОРТЕРМ. Разработка коммуникативных форматов, наряду с работами по созданию различных автоматизированных лексикографических фондов, является уже сейчас практической работой по созданию единого автоматизированного Фонда русского языка. На мой взгляд, вопрос не в том, следует или не следует организовывать Фонд — он будет организован просто ходом развития коммуникативных процессов в нашем обществе. Вопрос в том, какие организационные и материальные формы примет этот Фонд, будет ли он достойно служить целям нашего общества, каким процессам в развитии русского языка будет способствовать его наличие.

Л. Н. БЕЛЯЕВА

Все, кто имел дело с лингвистическими информационными базами (ЛИБ), используемыми в автоматизированных системах переработки текста, знают, как тесно связаны структура такой базы, организация словаря и словарной

информации с задачами, стоящими перед конкретной системой. В зависимости от поставленных задач лингвистическая информация описывается каждый раз особым образом, определяясь алгоритмами, на реализацию которых она ориентирована.

ЛИБ Машинного фонда, напротив, должна быть организована как система, хранящая самую разнородную информацию, необходимую для решения различных задач — задач универсальных и, что самое главное, задач, часть из которых мы сейчас просто не можем предусмотреть.

Такая постановка проблемы создания Машинного фонда предъявляет особые требования к его структуре. Ю. Н. Караулов и В. М. Андрющенко отмечают, что такой Фонд может быть организован двояко. Он может быть организован как набор отдельных баз, средств выбора информации из этих баз и средств работы с этими данными, либо этот Фонд может быть реализован как сложная иерархическая система, ориентированная на решение комплекса задач. При подобном подходе должна быть создана единая база и система лексикографических процессоров. Такие процессоры в зависимости от конкретных прикладных или фундаментальных лингвистических задач должны из этой общей базы выбирать ту информацию, которая необходима для решения поставленной в данный момент задачи. Такая ЛИБ строится как открытая модульная структура, в которой предусматриваются совместимость отдельных модулей, их корректировка и пополнение. Совместимость позволит на основе одного базового комплекса компоновать специализированные фонды для решения задач одного типа. Второй вариант реализации Машинного фонда, выступающего в виде некоторой единой системы, более перспективен, кроме всего прочего, потому, что первая версия Фонда, создаваемого как набор отдельных баз данных, противоречит основному принципу разработки этих баз, требующему отсутствия дублирования.

Если согласиться с тем, что Машинный фонд будет организован в виде единой базы, к нему можно предъявить следующие требования:

1. Фонд должен строиться как универсальная лингвистическая база, ориентированная как на решение задач автоматической переработки информации, так и более широко — на решение исследовательских, филологических задач, на создание систем искусственного интеллекта и т. п.

2. Фонд должен быть организован оптимально относительно задач, которые могут перед ним возникнуть. Такая организация предусматривает, как уже говорилось, модульное построение, которое заключается в том, что ЛИБ реализуется в виде набора нежестко связанных блоков. Эта модульность позволяет компоновать ЛИБ по мере готовности отдельных модулей и устранять дублирование информации, обеспечивая при этом возможность последовательного, поэтапного решения задач АПТ.

Задача разработки Фонда как единой базы представляется очень интересной, в этом плане я вполне согласна с Ю. Н. Карауловым в том, что в такой базе на каждое слово должно быть создано нечто вроде «досье». Но в то же время мы не знаем, как организовать структуру этого досье таким образом, чтобы охватить всю информацию, которая должна быть приписана некоторой лексической единице в базе. Это связано с тем, что и тип, и набор этой информации в достаточной мере зависят от тех задач, для решения которых база будет использоваться. Следовательно, при создании единой базы данных нужно будет в самой этой базе иметь иерархическую структуру. В основе

такой иерархии должен быть словник, который В. М. Андрющенко назвал Генеральным словником, т. е. ядерный словник, содержащий информацию о слове, которую можно приписать ему уже сейчас и которую достаточно подробно можно описать.

В большинстве реализованных ЛИБ ядерным блоком-модулем является автоматический словарь (АС) слов, основ, словоформ и словосочетаний. АС организуется как система словарных статей, каждая из которых содержит саму ЛЕ и ее описание.

При определении структуры и состава АС в архитектуре универсальной ЛИБ должны быть решены три основные задачи: 1) определение формы словарной единицы — заглавной единицы в словарной статье; 2) определение набора информации, составляющей описание ЛЕ; 3) определение оптимального способа организации словарной статьи в базе и, следовательно, самой ЛИБ.

Решение этих задач, в свою очередь, должно соответствовать следующим требованиям: информация, вводимая в ЛИБ, не должна дублировать уже имеющиеся в ней сведения; единицы ЛИБ должны обладать внутренней семантической цельностью, т. е. должны соответствовать некоторым смысловым единицам языка; словоупотребления и словосочетания перерабатываемого текста должны достаточно просто отождествляться с единицами ЛИБ.

Эффективность функционирования Фонда зависит как от общего объема сведений, включаемого в него, так и от организации каждой словарной статьи Фонда и характера помещенных в ней лингвистических сведений.

Информация, приписываемая в ЛИБ каждой ЛЕ, может состоять из описаний или характеристик четырех типов: морфологического, синтаксического, семантического и функционального.

В различных системах переработки данных реализованы ЛИБ, содержащие морфологическое описание единиц. В частности, в лаборатории инженерной лингвистики ЛГПИ им. А. И. Герцена создана ЛИБ, которая является единой для решения задач анализа и синтеза русского текста. С помощью этой базы мы можем по заданным лексико-грамматическим характеристикам получить нужную нам русскую словоформу, и наоборот — от русской словоформы получать все ее лексико-грамматические характеристики и, соответственно, приводить ее к канонической форме. Все остальные типы описания, включающие и синтаксические, и более глубинные семантические характеристики лексических единиц, вероятно, должны быть вынесены из самой словарной статьи и организованы отдельно. Связь между этой информацией и самой словарной статьей должна быть реализована на уровне отсылок, т. е. на уровне связей между отдельными единицами базы. Таким образом, в Машинном фонде должна быть реализована иерархия уровней описания лексических единиц, идущая от уровня эксплицитных поверхностных характеристик (морфологических, лексико-грамматических и т. д.) к семантическим глубинным уровням.

Машинный фонд должен быть открытой системой не только в том смысле, что его можно будет пополнять сколь угодно много с точки зрения количества словарных статей и количества единиц информации, но и можно будет достаточно оперативно модифицировать саму информацию и расширять информацию, которую можно приписать к конкретным единицам базы. Если говорить о конкретных взносах в Машинный фонд русского языка, то мы со своей стороны можем предложить систему морфологического и лексико-граммати-

ческого анализа и синтеза, которая у нас разработана на достаточно большом массиве, опробована и действует уже в течение нескольких лет. Эта система является частью Многоцелевого автоматического русского словаря (МАРСа), который в своей лексической части реализован как словарь основ.

Базовое лексико-морфологическое описание единиц МАРСа составляют модули лексико-грамматических характеристик и типовых парадигм. Собственно семантическое описание и описание тезаурусных связей, формирующие десигнативное и денотативное значение единиц словаря, реализуются в виде отдельных модулей. В словарной статье русского АС формируются отсылки к соответствующим записям в структуре модулей.

Следовательно, Фонд целесообразно организовать как иерархическую систему блоков-модулей, на верхнем уровне которой находится лексический словарь, содержащий минимально необходимую информацию. Нижние уровни Фонда должны составлять модули семантических описаний, связанные адресными отсылками с лексическими АС и между собой.

Такая структура ЛИБ Машинного фонда русского языка позволяет последовательно наращивать полноту и детальность семантического описания по мере разработки алгоритмов его использования.

Л. В. БОНДАРКО

Идея создания Машинного фонда русского языка действительно заслуживает самого пристального внимания. Создание такого Фонда позволило бы решить многие задачи, которые сегодня лишь усиленно обсуждаются.

Хотелось бы обратить внимание на два обстоятельства. Во-первых, Машинный фонд должен быть фондом, обеспечивающим автоматизированное исследование всей системы русского языка, а не только основой для лексикографических исследований. Машинное представление речевого материала — слишком дорогое дело, чтобы мы могли позволить себе поэтапно вводить лингвистическую информацию о слове — сначала только лексикографическую, затем морфологическую, синтаксическую и т. д. Нужно, видимо, предусмотреть максимально полную запись информации о слове, которая позволила бы не только создать своеобразное «досье» на каждое слово, но и обеспечила бы переход от одного уровня анализа к другому.

Второе, на чем необходимо остановиться, это то, какая форма существования языка должна быть представлена в Машинном фонде. В течение двух с половиной дней мы слышали выступления, касающиеся лишь письменной формы; можно подумать, что русский язык — это «немой» язык, не имеющий звуковой формы существования. Необходимость ввода в Машинный фонд именно фонетической информации определяется не только тем, что Машинный фонд должен обеспечивать исчерпывающий анализ разных уровней языковой системы, но и принципиальной невозможностью игнорировать материальную форму реализации языковых явлений.

Обращение к звучащей речи могло бы помочь лингвистам в решении целого ряда задач, связанных с вводом информации в память ЭВМ. Уже сейчас существуют устройства, преобразующие речевые сообщения в цифровую форму; можно было бы подумать об использовании таких распознающих систем для ввода лингвистической информации в машинную память. Что касается собственно фонетических характеристик, то их машинное представление

и хранение совершенно необходимо по многим причинам. Не секрет, что многие фонологические концепции базируются на довольно примитивных и часто неверных фонетических представлениях, так что получение унифицированного по форме и статистически достоверного фонетического материала смогло бы «переорганизовать» и дисциплинировать полеты фонологических фантазий.

Роль фонетических характеристик очень велика и при различных прикладных применениях результатов лингвистического анализа. Преподавание русского языка как иностранного, например, невозможно без учета довольно сложных отношений между сегментными и суперсегментными свойствами речи; работы по орфоэпии, повышение «фонетической культуры» речи, звучащей сейчас в устах тех, кто должен был бы представлять литературную норму, но чья речь несет явные следы диалектных или иных влияний; исправление дефектов речи и помощь людям, страдающим нарушениями слуха; диагностика заболеваний по речевым характеристикам — вот далеко не полный перечень проблем, где фонетические данные совершенно необходимы.

Прикладное значение работ в области фонетики особенно важно в связи с теми чисто организационными проблемами, которые стоят перед Машинным фондом. Эта программа, если она будет принята, должна выполняться не за 3—4 года, а за гораздо более долгий срок. Для ее выполнения нужна реальная материальная база, и выполнение прикладных исследований, которые, с одной стороны, вносили бы свой вклад в Машинный фонд, а с другой— финансировались бы заинтересованными министерствами и ведомствами, — один из способов реализации программы.

Последнее, на что хотелось бы обратить внимание. Необходима специальная подготовка — как профессиональная, так и психологическая — для того, чтобы Машинный фонд стал действительно орудием современного лингвистического анализа. Сейчас среди некоторых лингвистов старшего поколения царит какой-то страх: если мы введем в память машины тот или иной материал (словарь, тексты и т. д.), то мы как бы потеряем его, закроем доступ к нему для других. Что это неверно, кажется совершенно очевидным, однако даже на нашем совещании такие ноты звучали очень отчетливо. Для эффективного использования Машинного фонда нужно уже сейчас готовить кадры из числа сегодняшних студентов. В связи с этим возникает, как мне кажется, необходимость большего внимания к существующим в стране отделениям структурной, прикладной и математической лингвистики; вместе с тем необходимо и студентов-русистов подготавливать к новой технике лингвистической работы.

В. В. БОРОДИН

В Горьковском НИИ прикладной математики и кибернетики уже давно ведется работа по автоматизации лексикографических исследований. Были получены Лермонтовские словари, Частотный словарь русского языка, тезаурус философских терминов по книге В. И. Ленина «Материализм и эмпириокритицизм» и др. И хотя эта работа носила, как удачно выразились, партизанский характер, но кое-какой опыт у нас все-таки накопился. Мне бы хотелось остановиться на двух моментах.

Во-первых, это вопрос о достоверности текстов, которые должны приниматься в Машинный фонд русского языка. Этот вопрос уже поднимался от-

носительно исторических текстов. Хотелось бы его распространить на все печатные издания. Мне кажется, что Фонд должен с самого начала выдвинуть самые высокие требования к той информации, которой он собирается распоряжаться. А это в свою очередь выдвигает вопрос об ответственности, скажем, на уровне ответственности врача за запись о больных. Ведь в конце концов одна из иностранных фирм подала в суд на терминологический словарь за использование названия ее продукции.

Единственной силой, которая способна в полном объеме удовлетворить таким требованиям и нести полную ответственность за тексты, является издательство. И следует сделать издательство действующим лицом Машинного фонда и службу текста возложить на него, сделать его ответственным за хранение текстов на машинных носителях, которые получаются при фотонаборе. Такая служба может быть взаимовыгодной и Фонду и издательству. Это первый момент.

Второй момент связан с вопросом совместимости различных видов данных в Фонде. Судя по выступлениям, ситуация выглядит таким образом, что, с одной стороны, Фонду, который еще не начал функционировать, уже грозит стандартизация, а с другой стороны, если мы попытаемся избежать ее, то необходимо затратить 2/3 усилий на создание интерфейса между различными базами данных. Я склоняюсь к тому мнению, что все-таки следует затратить эти 2/3 усилий, но не смотреть на эти усилия как на необходимое зло, а использовать их в качестве средства и метода для изучения различных пар баз данных. Рассмотрим для примера отношение между парой словарей, статьи которых содержат различные наборы признаков (например, различные наборы лексических значений слов). Если при переходе от одного словаря к другому окажется, что некоторые признаки могут быть вычислены через признаки другого словаря, то очевидно, что признаки этого другого словаря необходимо исключить из Фонда, а в Фонд вносить только программы преобразований. Таким образом, создание самого интерфейса будет способствовать выявлению инварианта лексического описания и преобразований над ним.

На мой взгляд, Фонд может состоять из содержательно несовместимых баз данных. Что же касается доступа к разноформатным представлениям данных, то здесь было бы полезным уже сейчас собрать сведения об имеющихся форматах словарей в машинах и начать работу по созданию такого интерфейса. Я думаю, что на основе формальных грамматик здесь можно сделать много интересных вещей, и они будут не столь жестки, как требования стандартизации.

Е. М. ВЕРЕЩАГИН

Хотел бы как сотрудник ИРЯ им. А. С. Пушкина выразить всевозможную поддержку идее создания Машинного фонда русского языка. Очевидно, что возможные записи на машинах некоторых явлений русского языка, даже включая грамматику, смогут сыграть свою роль при определении минимумов учебных единиц для целей преподавания русского языка как носителям языка, так, пожалуй еще в большей мере и людям, для которых русский язык является неродным: и для советской национальной школы, и для преподавания русского языка в качестве иностранного. В этом смысле большая армия преподавателей русского языка, а это все филологи-русисты, имеющие нема-

лый опыт,— в распоряжении Оргкомитета, так что есть возможность привлечь людей, которые заинтересованы в конечном результате. Вспоминается клятва, которую дали аргонавты при отправлении в поход: «Будем полагаться на сотоварищей, как на самих себя!». Это выражение при всей своей образности хорошо отражает существо дела, которое я сейчас имею в виду. Это первое.

Теперь второе. Для меня, человека, далекого от всевозможных машинных интересов, было очень знаменательным в нашем совещании то, что круг проблематики оказался для меня неожиданным. Конечно, мы привыкли говорить об обратных словарях, скажем, о словарях терминологии, о частотных словарях. Они все очень важны для нашей проблематики, в том числе для преподавания русского языка как иностранного. Однако прозвучала мысль о необходимости создания Машинного фонда также и русской культуры, и вот на этом вопросе хотелось бы задержать внимание присутствующих.

Мне кажется, что мысль о том, что слово, вообще язык, является амальгамой культуры, и то, что эта мысль перемещается сейчас в центр внимания также и «машинников», заслуживает самого пристального внимания. Можно оспорить мысль о том, что язык создавался также как средство коммуникации. Во всяком случае Гумбольдт считал, что язык существует как средство познания мира, а коммуникативная функция является вторичной. Не будем вдаваться в подробности и принимать участие в этой контроверзе. Но то, что функция аккумуляции знаний является такой же важной, как функция коммуникативная, мне кажется, мысль, заслуживающая все-таки хотя бы взвешивания. Именно это отметил Ю. Н. Караулов, который призвал нас подойти с достаточной серьезностью к проблеме слова как носителя знания. Здесь мне хочется сделать несколько замечаний. Во-первых, вопросы, которые кроются за этой постановкой, чрезвычайно сложны, трудны, всеобъемлющи — этого сказать мало, потому что нет даже подступа к решению этой проблемы. Мы не знаем, какова, скажем, связь внешней формы слова с мыслью, со знанием. Ведь это большая загадка. И здесь проблема ассоциаций только прикрывает проблему, т. е. название ассоциаций только прикрывает вопрос, не открывает его никак. Я думаю, что если у нас имеются попытки описания группового знания в энциклопедиях, потому что там описывается знание ученых, то до сих пор не предпринималось попыток описать знания обиходного, или знания языкового, того, что стоит за простейшими русскими словами типа, скажем стол, стул, вода, кровь, голова и т. п., на уровне бытовом, всем известном. Такие описания не предпринимались. Мне кажется, что если мы действительно будем ставить задачу описания культуры, стоящей за языком, то это такое поле деятельности, к которому пора приступать, очень трудно, ответственно, но очень нужно. И этот вопрос, при всей его сложности и, может быть, даже запутанности, отягощенности всевозможными ассоциациями ненаучного характера, в принципе разрешим. Если различать задачи, которые неразрешимы, и просто нерешаемые задачи, то этот вопрос относится к числу не решаемых пока задач, хотя он разрешим. Призыв Ю. Н. Караулова развернуть исследования по теме «Русский язык и русская культура» в этом смысле является весьма актуальным и заслуживающим всевозможной поддержки. Но и трудности, которые здесь стоят перед нами, нельзя преуменьшать. При всей неразрешимости этой проблемы, очень важно обратить внимание на те вещи, где эта проблема неразрешима, выяснить те

ситуации, где она, может быть, принципиально решаться не может. Самое главное, мне кажется, здесь при выделимости семантического ядра, стоящего за словом, периферия все-таки остается текучей, зыбкой и не поддающейся никакому описанию. В этом смысле я хочу сослаться на кредо П. А. Флоренского, который выразил его, для того чтобы свою мысль представить более элегантно, в форме, предложенной поэтом Бальмонтом. Он писал так:

Слова — хамелеоны,
Они живут спеша.
У них свои законы,
Особая душа.
Они спешат меняться,
Являя все цвета,
Поблекнут, обновятся —
И снова красота.

Там, где слова «блекнут» или «обновляются», для машинного подхода нет места. И надо просто сказать, что эти вещи надо оставить за пределами нашего внимания, и только изучать ту семантику, которая является общеязыковой, выявляется как-то с помощью анкет или другими объективным способами. Много интересного было сказано о поэтическом языке, где, мне думается, приложимость машинного подхода также ограничена. И самое последнее. За разговорами вокруг применения машин, за этими большими разговорами нередко скрывается (я не хочу никого обидеть) отсутствие какого-то результата. Вот я человек поколения людей, которые начинали машинный перевод, и наблюдаю, как это развивалось на протяжении времени. Откровенно говоря, может быть, и есть основания для оптимизма, но для сдержанного оптимизма.

Мне бы хотелось закончить тоже цитатой из Одиссеи: ведь когда аргонавты отправлялись в странствие, они ставили перед собой цель «увидеть венец разговоров наших». Я призываю как раз к тому, чтобы наше дело обрело достойный венец, стефанос, которого оно заслуживает.

И. Н. ВОЛКОВА

Проблема создания Машинного фонда русского языка имеет важное общегосударственное значение, так как фонд русского языка представляет интерес не только для лингвистов, для которых он является объектом научных исследований, но и для специалистов практически всех сфер народного хозяйства, для которых он имеет прикладное значение. На конференции уже отмечалась роль лингвистического обеспечения при создании автоматизированных управляющих и информационных систем. Необходимо также учитывать важное значение языка в создании различного вида документации. Производство любого изделия, создание любого технологического процесса связано с подготовкой большого количества документов: проектных, конструкторских, технологических и т. д. Среди этих документов особое место занимают нормативно-технические — стандарты всех категорий и видов и технические условия. От того, насколько четко, точно изложены эти документы, зависит правильное, однозначное понимание их специалистами. Поэтому к этим документам предъявляются особенно жесткие требования по точности формулировок, однознач-

ности используемых в них терминов. Именно поэтому в процедуре подготовки нормативно-технических документов имеется специальный этап — терминологическая экспертиза. Для качественного проведения терминологической экспертизы нормативно-технических документов специалисты, осуществляющие эту экспертизу, должны располагать достоверным справочно-поисковым аппаратом, который позволял бы им легко и просто получать необходимую информацию о том, какая терминология стандартизована, какие термины следует употреблять в том или ином значении. Это было одной из причин создания банка стандартизованных терминов (Справочного банка терминов ВНИИКИ) и разработки на его основе Автоматизированной системы информационно-терминологического обслуживания (АСИТО). Эта система находится в промышленной эксплуатации, ее коллективными абонентами являются свыше 400 организаций (научно-исследовательских и проектно-конструкторских институтов, издательств, органов НТИ, промышленных предприятий).

База данных АСИТО содержит около 70 тыс. терминологических статей из государственных стандартов, стандартов СЭВ и стандартов ИСО на термины и определения. В дальнейшем в базу данных планируется ввести также терминологические статьи из отраслевых стандартов на термины и определения, стандартов МЭК и ряда других документов.

В каждой статье представлены следующие элементы данных (при условии их наличия в источнике): заглавный (стандартизованный) термин, код документа по Классификатору государственных стандартов, дата утверждения (издания) документа, номер термина в документе, краткая форма заглавного термина, синонимы заглавного термина (недопустимые или допустимые к применению), эквивалентные термины на английском, немецком и французском языках. В настоящее время в базу данных вводятся также и определения. Система обеспечивает возможность поиска в двух режимах: диалоговом с использованием дисплея и пакетном с выдачей ответа в виде распечатки на АЦПУ. Поиск осуществляется по любому элементу данных терминологической статьи и их сочетаниям, реализована возможность поиска с учетом словоизменительных вариантов слов и синтаксических вариантов словосочетаний. Ответ на запрос выдается и при неполном лексическом совпадении запроса с термином в базе данных АСИТО.

Промышленная эксплуатация АСИТО, контакты с потребителями терминологической информации показывают, что не на все вопросы терминологического характера, с которыми обращаются в систему, может быть дан ответ с помощью базы данных АСИТО.

Целый ряд вопросов касается правильности словоупотребления, правильности выбора той или иной модели термина. Например, очень часто задают вопрос — в каких случаях нужно использовать терминоэлемент стандартный, а в каких — стандартизованный, стандартизированный? Нередко поступают запросы с просьбой дать значения тех терминов, которые не стандартизованы и поэтому отсутствуют в базе данных АСИТО.

В этих случаях рекомендуем нашим обонентам обращаться к соответствующим терминологическим словарям. Однако библиографии таких словарей нет, по многим темам отсутствуют сами словари, а из имеющихся многие являются библиографической редкостью. В результате специалисты, которые не являются ни лексикографами, ни специалистами-терминологами, начинают сами создавать для себя словари. И такие словари создаются на промыш-

ленных предприятиях. Известен опыт разработки словаря на КАМАЗе. Попытки готовить для себя словари делают и другие организации. Многие эти вопросы могли бы быть сняты при наличии для специалистов легкого доступа в ту базу данных, которая представлена в виде коллекции терминологических словарей. Эта проблема может быть решена реализацией машинного словаря, имеющего, таким образом, большое практическое значение для специалистов как источник ненормализованной терминологической информации.

При создании этого словаря очень важен вопрос о том, что должно составлять корпус данных, базу данных этого машинного словаря. Вопрос этот непростой, поскольку, с одной стороны — естественное желание ввести туда как можно больше данных, а с другой — трудности ввода. Практика создания АСИТО — сравнительно небольшой системы — показывает, что это очень трудоемкий процесс. Этот вопрос актуален для всех специалистов, работающих с банками терминологических данных.

Учитывая это, представляется необходимым при создании Машинного фонда русского языка очень тщательно выявить, собрать и проанализировать всю ту лексику, которая уже представлена на машинных носителях. В числе таких источников уже указывались информационно-поисковые тезаурусы и АСИТО. Однако только Р. Г. Котовым были упомянуты общесоюзные классификаторы технико-экономической информации (ОК ТЭИ). ОК ТЭИ — очень важный и очень обширный по своему составу источник лексической информации, который представлен на машинных носителях. Это десятки миллионов единиц лексической информации: наименования продукции, которая выпускается в нашей стране; наименования природных ресурсов; наименования профессий рабочих и служащих и их должностей. Важно, что это все уже представлено на машинных носителях. Поэтому при создании Машинного фонда русского языка обязательно нужно очень тщательно выявить всю лексику, уже записанную на машинных носителях. Видимо, еще не все источники лексической информации на машинных носителях известны, поэтому эта работа должна быть выполнена. Важен также вопрос обмена терминологической информацией. Уже ставился вопрос о необходимости разработки коммуникативного формата для обмена терминологической информацией на машинных носителях. ИСО/ТК 37/РГ 4 завершается разработка международного стандарта 6156 на коммуникативный формат для обмена терминологической и лексикографической информацией. Он может быть использован при разработке коммуникативного формата для обмена терминологической информацией между теми базами данных, которые уже имеются на машинных носителях и будут созданы в процессе реализации программы создания Машинного фонда русского языка.

Мне хотелось бы остановиться на необходимости глубокого изучения зарубежной практики создания больших массивов терминологических данных, таких, как система TEAM фирмы СИМЕНС, система ТЕРМДОК Шведского центра технической терминологии, ЕВРОДИКАУТОМ (Люксембург) и др. Необходимо учитывать не только их опыт накопления терминологической базы данных, но и опыт его использования, потому что они имеют уже сложившуюся практику издания словарей с помощью ЭВМ и располагают массивами русской терминологии, записанной на машинных носителях. Видимо, следует рассмотреть вопрос о возможности и целесообразности использования при создании Машинного фонда русского языка и этих терминологиче-

ских массивов, представленных на машинных носителях. Обмен терминологической информацией на международном уровне может осуществляться в рамках программы ТЕРМНЕТ, разрабатываемой по инициативе и под руководством Международного информационного центра по терминологии (Инфотерма).

При такой крупномасштабной целевой комплексной программе, какой является создание Машинного фонда русского языка и которая для своей реализации потребует значительных затрат, необходимо дать ее технико-экономическое обоснование и подсчитать ориентировочный экономический эффект от ее реализации.

А. С. ГЕРД

Все дела и темы требуют четкой научной организации и должны быть реалистичны. В. М. Андрющенко много внимания уделил организации Машинного фонда. Я бы поставил вопрос иначе — надо говорить об организации работ по Организации Машинного фонда. Во всяком случае тот небольшой опыт, который мы имеем в Ленинграде, в ЛГУ по автоматической обработке разных текстов, показывает, что в конечном счете все сводится к организации.

Приходилось слышать мнение, что создание Машинного фонда как бы снимает проблемы создания картотек, словарей. Идеальный словарь — это и есть машинный, но думаю, что это не так. Машинный фонд не только не отменяет актуальности обычных новых словарей, а наоборот, создает такую базу для всей словарной работы, которой ранее никогда и не было. Бережно сохраняемый Машинный фонд — это надежная основа множества разных словарей, которые должны быть и которых у нас нет, как очень верно отметил А. П. Ершов. Как человек, занимавшийся вплотную такими вещами, вижу очень много конкретных, частных трудностей, но сегодня я не хочу их обсуждать. Перечислю некоторые. Мне кажется, что как-то совсем забывают о том, что все эти проблемы выдвинуты научно-технической революцией, но она ведь продолжается, меняется техника, изменяются машины. Реально Машинный фонд будет осваиваться и вводиться на каких-то машинах, о которых мы сегодня ничего не знаем. Как известно, даже отладка минимальной фактографической системы перед ее промышленной реализацией требует нескольких лет работы с машиной. Как это все будет с Машинным фондом русского языка, сколько понадобится лет и какие тогда будут машины — это совсем не ясно. Я хочу вспомнить только об одном: мы помним и знаем, что лет 10—15 тому назад была масса прекрасных идей, реализованных на разных машинах. А многие ли из этих идей перенесены на новые типы машин? Многие из них остались там именно из-за сложнейшей традиционной проблемы перехода с одного вида машины на другую. Я не буду больше об этом говорить, по-видимому, нужны какие-то рабочие группы, которые будут заниматься проблемами совместимости, интерфейсов. Эти проблемы очень серьезные; мы знаем, многие благородные начинания в области разработки больших и малых систем, к сожалению, погибли, потому что они были ориентированы на те машины, которые были потом сняты, которые были ориентированы на те языки, которые были оставлены. Я не считаю, что проблема создания фундаментального Машинного фонда нереальна. Она реальна, осуществима, но решать ее надо неторопливо, серьезно, постепенно, конкретно обсуждая каждое звено отдельно. 25 лет развития отечественной математической линг-

вистики и зарубежный опыт ярко демонстрируют одну и ту же тенденцию, одну и ту же кривую. При появлении новых идей первоначально — медленное тление, затем — резкое возгорание и взлет, статистический рост сторонников этих идей в ожидании быстрых и скорых результатов, множество новых теорий, поспешных экспериментов, а затем наступает пора разочарования, пора депрессий, отходничества в другие области и переквалификации. И лишь постепенно, со временем, из уже затухающих углей медленно возгорается пламя новых идей и фундаментальных разработок. Так было с машинным переводом, так было со многими системами. Никто не гарантирован от ошибок в большом деле, но их число должно быть сведено к минимуму. К минимуму должна быть сведена и потеря кадров и союзников. И достичь этого можно только на почве трезвого реализма, продуманности, неторопливости и взаимной благожелательности.

С. И. ГИНДИН

Главный вопрос, к которому я хотел бы привлечь внимание, — это вопрос о соотношении словарей и текстов в составе Машинного фонда русского языка. Тот подход к его решению, который будет предложен, ведет к постановке второго, производного вопроса — об эвристико-методологической роли, которую могли бы сыграть при создании Фонда исследования по теории и типологии текста. Обе проблемы могут показаться на первый взгляд отвлеченно-теоретическими, между тем адекватная их трактовка имеет принципиальное значение для успеха обсуждаемого нами дела. Правильное же решение первого вопроса немаловажно и для такой жгуче практической задачи, как обеспечение самоокупаемости и доходности Фонда.

Вполне четко определился подход к Машинному фонду прежде всего как к собранию словарей (включая и словарно заданные грамматики). Ю. Н. Караулов, правда, отметил, что, кроме словарей, в Фонд должны входить и тексты. Однако за этим признанием тут же последовала оговорка о том, что текстов «не должно быть слишком много». Основное же внимание Ю. Н. Караулов уделил организации исключительно словарно-грамматической части Фонда. Что же касается анализа В. М. Андрющенко структуры Машинного фонда, то в нем текстам и вовсе была отведена только роль иллюстраций к положениям словарей. Не удивительно, что при таком понимании роли текстов предусматриваемая В. М. Андрющенко в структуре Фонда «иллюстративно-текстовая картотека» в основном должна состоять (как то бывает с иллюстрациями и в обычных словарях) совсем не из реально бытующих в обществе и культуре текстов, а лишь из их фрагментов, достаточных для целей иллюстрации.

Но словари, равно как и грамматические описания,— это не сам язык, а лишь фиксация наших знаний о нем. Непосредственной действительностью языка является речевая деятельность, создаваемые и циркулирующие в обществе тексты. В какой мере мы можем, создавая Машинный фонд, обойтись в нем без текстов, ограничившись словарями и грамматическими описаниями, зависит от того, какие функции мы предполагаем на этот Фонд возложить. Ю. Н. Карауловым были выделены две основные функции Машинного фонда русского языка: быть гигантских размеров памятью, позволяя оперативно находить и использовать уже имеющиеся сведения о языке, и быть лабора-

торией, инструментом для получения новых, ранее неизвестных данных о языке.

Пока мы рассматриваем Фонд только как память, стремление изгнать «сырой материал» текстов более или менее оправданно: материал в памяти естественно хранить в удобной для использования форме, а словари как раз и представляют собой такую форму, устоявшуюся и апробированную в многовековом употреблении. Но даже и тут полное изгнание текстов может оказаться недальновидным. Любой словарь дает нам языковые факты такими, как их выделил и систематизировал его составитель, а иллюстративная картотека по способу своего создания приспособлена для подтверждения данных тех словарей, вместе с которыми возникла. Если же появится необходимость в проверке словарных дефиниций и утверждений, нам не обойтись без новых примеров, а дать их могут только тексты.

Для использования же Машинного фонда в качестве исследовательской лаборатории в его составе и подавно должны иметься тексты в таком количестве и в сочетании с такой системой их автоматизированной обработки, чтобы можно было получать на ЭВМ достоверную новую информацию о языке и, в частности, создавать новые словари (в том числе, возможно, и словари каких-то еще неизвестных нам сегодня типов). Наличие в Фонде массивов текстов становится необходимым условием самой возможности его функционирования, и вопрос, нужны ли тексты, заменяется более практическим, но и более трудным: *сколько и каких именно* текстов надо включить в состав Фонда?

Для ответа на этот вопрос необходимо четко различать два типа текстов и, соответственно, два типа их массивов в составе Фонда. Массивы первого типа, которые я назвал бы коллекциями текстов, составляются из текстов, представляющих интерес не своей индивидуальностью, а именно как *типичные* представители определенного вида или жанра речевой деятельности. Например, нужна коллекция управленческих документов, коллекция научно-технических публикаций по определенной отрасли (на первых порах можно будет, наверное, воспользоваться выборками из массивов рефератов, введенных в действующие ИПС), коллекции публицистики определенных периодов и т. п. Поскольку любой текст в коллекции может быть заменен другим, не уступающим ему в типичности, в коллекции не нужно стремиться к полноте охвата соответствующего класса текстов, но лишь к репрезентативности выборки. Тексты в коллекции заменимы, но сама коллекция всегда необходима как отражение определенной сферы существования и фунционирования русского языка.

Наряду с коллекциями в Фонде должны быть представлены массивы другого рода, включающие тексты, за которыми в нашем обществе признается *индивидуальная культурная ценность* и которые и в языковом отношении рассматриваются как в определенном смысле образцовые или как яркий пример выявления выразительных возможностей русского языка. В случае большой величины подобные тексты могут составлять самостоятельные подмассивы, малые тексты будут группироваться в *корпусы* (чаще всего — корпус текстов одного автора), но и те и другие в отличие от коллекций должны характеризоваться *полнотой представления*, отсутствием изъятий и сокращений. Самое количество примеров, часто малоизвестных, из Фета, которыми иллюстрировал свои положения Ю. Н. Карау-

лов, свидетельствует о невозможности заранее, раз и навсегда решить, какие именно из индивидуально ценных текстов и какие из содержавшихся в них языковых фактов понадобятся лингвистам для их штудий.

Вместе с тем, поскольку количество индивидуально ценных текстов в столь развитой культуре, как русская, не может не быть крайне велико, разработчикам Фонда не обойтись без введения каких-то, хотя бы и внешних, но жестких и четких критериев определения очередности перевода таких текстов в машинно-читаемую форму. Например, можно начать с включения в Фонд тех текстов, для которых уже имеются текстологически авторитетные издания. Еще больше сожмется круг текстов, если мы возьмем только изучаемые по программе средней школы [3].

Включение в состав Машинного фонда русского языка текстов индивидуальной ценности будет способствовать не только реализации Фондом самой «высокой» из его функций — исследовательской, но и выполнению такого элементарного экономического требования, как самоокупаемость. Во-первых, Фонд сможет осуществлять автоматизированную подготовку включенных в него индивидуально ценных текстов к изданию, подобно тому, как это делается в Машинном фонде эстонского языка при подготовке академических словарей. Во-вторых, наличие в составе Фонда различных программ автоматической обработки текстов позволяет организовать на его базе службу лексикографического обеспечения индивидуально ценных текстов и покончить с тем ненормальным положением, при котором подавляющая часть конкордансов, частотных словарей, словарей рифм и т. п. пособий к памятникам русской культуры выходит за рубежом и совершенно недоступна нашим читателям. Думаю, что если Фонд сможет обеспечить хотя бы школьных учителей простейшими словарными пособиями ко всем произведениям, входящим в школьную программу, о его экономических показателях можно будет не беспокоиться.

В предварительной программе нашего совещания был пункт «Перспективные лингвистические исследования в интересах Фонда». Предложенная трактовка роли текстов в составе Машинного фонда русского языка позволяет наметить два рода таких исследований, находящихся в компетенции той ветви лингвистики, что не вполне удачно называют «лингвистикой текста». Во-первых, если будет решено не ограничиваться в Фонде лишь «иллюстративно-текстовой картотекой», а вводить в него реальные тексты, неотложным становится исследование типологии русского текста в разные периоды его истории. Без такого исследования мы не сможем дать достаточно обоснованного ответа на вопрос, какие коллекции современных текстов надо предусмотреть в составе Фонда. Нужно изучить и опыт зарубежных работ по типологии текстов (особенно интенсивно ведущихся в ФРГ).

Во-вторых, лингвистика текста может сыграть определенную эвристико-методологическую роль в перспективном продумывании задач Фонда и его взаимоотношений с лингвистами-исследователями. Ю. Н. Караулов, приводя

[3] Последнее не означает, что мы понимаем под текстами индивидуальной ценности только художественные. Со временем в Фонд безусловно должны будут войти и этапные произведения общественной мысли, и образцовые научные сочинения, прежде всего те из них, которые по самому замыслу своему были призваны, как «Грамматика» и «Риторики» Ломоносова, влиять на развитие русского языка.

примеры употребления Фетом слова «огонь», отмечает, что фактически оно используется поэтом «в разных семантических планах», так сказать, переключается. Но как мы узнаем, к какому семантическому полю принадлежит это слово в каждом конкретном случае, каково его значение? Очевидно, само устройство текста, соседние слова, связи между частями позволяют нам делать по тексту необходимые заключения. Поэтому если мы сумеем понять, что, собственно, структура текста позволяет нам делать в процессе его восприятия, какие виды выводов о значениях слов строить, то лучше представим себе и типологию возможных словарей, и ту классификацию исследовательских задач и позиций, о которых шла речь. Исследовательских позиций может быть много, но поняв, какие виды операций позволяет производить с текстом и какие типы информации из него получать сама организация текста, мы, думается, лучше сумеем очертить и тот круг задач, с которыми станут обращаться к Фонду будущие исследователи.

В. П. ГРИГОРЬЕВ

Одним из самых важных, ключевых слов нашей конференции является слово *начало*. Оно понимается по-разному, по-разному понимается этап, на котором мы находимся сейчас. Тем не менее очень многое упирается в вопрос: «С чего начать?» В практическом плане этот вопрос оказывается достаточно трудным, во всяком случае не менее трудным, чем вопрос о «началах» в евклидовском или лобачевском смысле.

В самом деле, нам предстоит, кроме всего прочего, во-первых, соотнести целевую программу Фонда с существующими перспективными планами различных ведомств, и прежде всего с перспективными планами академических учреждений, т. е. с некоторыми конкретными темами, разработка которых предстоит в этой и следующей пятилетках. Во-вторых, необходимо предусмотреть определенное обратное воздействие Фонда на эти планы, на будущие планы и их реализацию. Я имею в виду не только уже более или менее ясные вопросы от нас к Фонду (многие из них уже были заданы и еще будут задаваться), но и те вопросы (лучше, может быть, сказать, не вопросы, а творческие импульсы), которые последуют от Фонда к нам, смотря по тому, как он будет складываться.

Эти вопросы не могут не касаться и задач собственно семантического характера, в частности вопросов соотношения общеязыкового и поэтического смыслов. Поэтому я считаю принципиально важным постулат, который прозвучал в первом из заслушанных нами докладов,— постулат относительно целей Фонда — как лингвистических, так и, шире, филологических целей. Дальше я предложу конкретную тему общефилологического ранга, но сначала на минуту отвлекусь на беглую полемику с Е. М. Верещагиным.

Опираясь на Флоренского и Бальмонта, Е. М. Верещагин призывал в основном ограничиваться общеязыковой семантикой, однако в таком случае саму культуру мы стали бы понимать несколько обуженно. Сейчас, кстати, подтверждаются контакты Флоренского с Хлебниковым. Здесь придется ограничиться ссылкой на давно высказанное в полемическом плане мнение о том, что «слово, каким его впервые показал Хлебников, требовало для себя формул высшего порядка». По-видимому, формулы такого рода — и цель, и условие полноценного функционирования нашего Фонда.

Очень продуктивным оказался оппозитивный метод обсуждения нашей проблематики. Так, Ю. Н. Караулову этот метод позволил в явной форме представить проблемы идеологии Фонда и предложить их достаточно гибкие решения. Правда, имели место попытки выдвинуть на первый план одну из альтернатив там, где специфической является, на мой взгляд, как раз другая. Я имею в виду, например, некоторое заострение консервативной функции Фонда как хранилища или, скажем, призывы к сохранению давно действующих капитальных картотек. Понятно, что Фонд не грозит пустить их на поток и разграбление. «Сжигать Рафаэля» едва ли кто-нибудь будет предлагать.

Были выдвинуты и новые оппозиции. Ю. Н. Караулов определяет слово как центральную единицу Фонда, вводит понятие «досье на слово». Обращалось внимание на центральную роль словосочетания для терминологического компонента в структуре Фонда. Еще более принципиальными для Фонда оказываются тоже упоминавшиеся оппозиции слово/текст и словари/тексты. Со своей стороны, я хочу предложить такую оппозицию: некоторый относительно произвольный набор текстов/художественный идиолект как некоторая целостность. Нам нужны не только представительный корпус разнородных текстов, не только некоторое множество текстов индивидуальной ценности, по выражению С. И. Гиндина, но и целостные художественные идиолекты, которые могли бы быть интерпретированы как факты художественных идиостилей.

Тем самым перед нами возникают проблемы приоритетов, известная конкуренция среди задач, которые претендуют на помощь Фонда, на место в нем. Кажется очевидным, что беспристрастным Фонд быть не сможет. Ю. В. Рождественский полагает, что основой Фонда должна стать деловая проза. Сегодня Ю. Д. Апресян, разграничивая культурные и прикладные фонды, внес некоторые ограничения в этот тезис. Вероятно, было бы поучительно обсудить столь же сильное предложение строить Фонд преимущественно на технических, политических или художественных текстах, или — еще более сильный вариант — даже исключительно на стихотворных текстах как на манифестациях высшей формы культуры речи. Но уже ясно, что формирование текстового компонента Фонда будет многоканальным, хотя задачи обуздания стихийной конкуренции между отдельными каналами и отдельными порциями текстов нам, по-видимому, не обойти.

А. С. Герд остановился, в частности, на микротопонимике. Я хочу поддержать его мысль о значимости этих данных для Фонда именно с позиций эстетики русского языка. Такие исчезающие названия, как название пожни *Дрюменник* в районе Селигера и т. п. факты (диалектологи хорошо знают круг такого рода лексики), представляют собой в известном смысле не меньшую ценность, чем иной деловой текст. Но в этом же ряду задач еще существенней может оказаться задача подготовки словаря русских пословиц или, скажем, словарного свода лексики фольклора. Очень важной, а в методическом плане даже первоочередной представляется задача подготовки метрико-ритмического корпуса русской поэзии.

И все же в моих глазах особым приоритетом среди иных идей обладает конкретная тема перспективного плана Института русского языка АН СССР. Эта тема формулируется сейчас как «Очерки языка русской поэзии XX в.». Содержание и принципы осуществления этой работы (а она задумана как совокупность сжатых монографических описаний языка нескольких десятков

поэтов) будут еще обсуждаться. Но понятно, что при любом варианте реализации «Очерков» необходимы конкордансы к изданиям большой серии «Библиотеки поэта» и другим авторитетным изданиям.

Эта тема могла бы, оставаясь культурным достоянием машинных фондов, стать в то же время и прикладной, правда, несколько в ином смысле, чем тот, в каком это определение употребляет Ю. Д. Апресян. Конкордансная форма хранения соответствующей информации в Фонде должна учитывать недостатки большинства известных печатных конкордансов к поэтам, где словоупотребление в жестком контексте одной-единственной строки не допускает, как правило, перехода к более широкому контексту и восстановления целостного текста всего стихотворения.

М. П. ДАНИЛОВ

На конференции была четко и полно обрисована общая задача: ввести в машину достаточно представительный массив текстов на русском языке с целью получения с помощью компьютера тех результатов, которые принципиально или практически невозможно получить без него. Меньше внимания было уделено организационным путям и этапности работ. Предлагалось вести их на общественных началах. Это нереально. Необходима координация, возможно, межведомственная, общий план, распределение задач, соответствующее финансирование работ.

Первым этапом должен быть учет тех массивов, которые либо уже введены в машину, либо имеются на машинно-читаемых носителях. При этом не стоит переоценивать значение всякого рода информационных тезаурусов и словников. Это не цельные тексты, а лишь слова и словосочетания в канонической форме. С другой стороны, здесь есть и резервы. Например, ТАСС ежедневно передает по телетайпным каналам поток материалов для местных газет. Это представительный массив современной газетной прозы на машинно-читаемых носителях.

Другим резервом может быть ввод энциклопедических справочников (однотомного энциклопедического словаря, МСЭ, БСЭ, возможно, отраслевых энциклопедий). Эта работа может быть поддержана с нескольких сторон: издательством (за рубежом основные энциклопедические издания готовятся, ведутся и переиздаются с помощью компьютера), исследователями русского языка (энциклопедии — это представительный фонд научной прозы) и, наконец, информационными учреждениями (такие справочники, по существу, слабоформализованные тезаурусы и отличный материал для информационно-поисковых тезаурусов).

В организационном плане было бы разумно создание центрального широко доступного Машинного фонда русского языка — своего рода библиотеки с богатым фондом и необходимым минимумом услуг (система управления базами данных, компиляторы с основных языков программирования, программы сортировки, печати и т. п.). Такой «библиотекой» могли бы пользоваться все заинтересованные учреждения и исследователи на договорных, возможно, коммерческих началах.

В ряде случаев в этом центре могла бы проводиться лишь предварительная выборка и организация материала, а, учитывая, что все современные

ЭВМ совместимы на уровне магнитных лент, основная работа завершалась бы на ЭВМ заинтересованного исследователя.

В заключение мне хочется призвать лингвистов, имеющих доступ к ЭВМ, овладевать началами программирования, хотя бы на простых языках типа Бейсик или на подмножестве ПЛ/1 системы ДИСП. Программы для решения многих содержательных лингвистических задач могут включать всего несколько десятков операторов и составляются и отлаживаются за считанные часы, а это зачастую быстрее, чем требуется на то, чтобы поставить задачу профессиональному программисту.

П. Н. ДЕНИСОВ

Мы обсуждаем очень широкий круг вопросов, однако я бы не сказал, что каждый из поднимавшихся вопросов одинаково интересен. По-видимому, смысл данных обсуждений состоит в таком первоначальном обмене мнениями, при котором общей темой, общей идеей является создание Машинного фонда русского языка.

Первое, что мне хотелось бы отметить, поскольку я буду опираться на свой собственный, в основном лексикографический опыт,— это то, что, на мой взгляд, Машинный фонд русского языка позволит нам достигнуть принципиально другого количественного уровня охвата материала, а из другого количественного уровня, конечно, последует и качественное изменение в характере нашего видения предмета, например лексики, закономерностей ее состава, функционирования, распределения значений и т. д. Акад. А. Н. Колмогоров говорил о том, что математики иногда очень вольно обращаются с понятием бесконечности, недооценивая значение таких данных, которые имеются в достаточно большом количестве. Оперируя одним порядком величин, имеем одни возможности проникновения в словарный состав языка, если же количественный охват материала повышается на один-два порядка, открываются принципиально новые возможности проникновения в функционирование лексико-фразеологического уровня языка.

В Словарной картотеке ИРЯ АН СССР в Ленинграде имеется 8 млн. карточек, накопление которых шло более ста лет (с 1881 г.). Эти 8 млн. карточек должны отвечать за весь русский литературный язык XIX—XX вв. На этих 8 млн. карточек составлен после войны «Словарь современного русского литературного языка» (в 17-ти т., 1950—1965 гг.), на них же готовится переиздание этого самого фундаментального толкового академического словаря русского языка, созданного в советское время. Когда начиналась работа над «Словарем русских народных говоров», в его картотеке насчитывалось 165 тыс. карточек. Не мало ли для всех говоров русского языка на всей территории его распространения?

Сплошное расписывание Полного (пятого) собрания сочинений В. И. Ленина дало картотеку объемом 2,5 млн. карточек. 2,5 млн. для языка одного автора и всего 8 млн., извлеченных из произведений двухсот русских писателей, публицистов, ученых, деятелей искусства, для всего русского литературного языка двух последних веков! [4]

[4] Подсчеты по перечням авторов и их сочинений, содержащимся в первых томах большого академического словаря (17 т., 1950—1965 гг.), показывают, что русскую нацию в академическом словаре представляют примерно 200

Сопоставление цифр картотеки и словаря показывает следующее: большой академический словарь: картотека 8 млн. — число слов в словаре 120 тыс., словарь языка В. И. Ленина: картотека 2,5 млн. — число слов в словаре свыше 37 тыс. Общим делителем является тройка. Картотека и словник большого академического словаря всего в три раза больше соответственно картотеки и словника ленинского словаря.

Вывод может быть только один, и он будет не в пользу академического словаря. Для составления полного академического словаря русского языка XIX—XX вв. объем картотеки следует повысить хотя бы на один порядок, т. е. иметь 80 млн. карточек вместо 8 млн., что даст и больший словник — 300—400 тыс. слов. Заниженный объем академической картотеки объясняется не только ручным трудом, но и установкой на частичную выборку из источников как основной путь построения картотеки [5].

Чтобы поднять объем общей академической картотеки на один порядок, необходима помощь ЭВМ. Р. П. Рогожникова говорила о механизации картотеки. Фонды большой академической картотеки должны стать сердцевиной Машинного фонда русского языка. Должен быть акцентирован полный (сплошной) тип расписывания образцовых текстов.

Словник произведений В. И. Ленина намного превышает имеющиеся словники других авторов. А слов, не зафиксированных словарями русского языка конца прошлого века, начала и середины XX в., в его составе оказалось больше четырех тысяч. Р. Г. Рогожникова говорила о необходимости словарей редких слов. Они, конечно, необходимы, потому что у В. И. Ленина по общим статистическим закономерностям имеется очень длинный шлейф редких слов. Я говорю о том, что традиционные способы накопления материалов для исследования лексики все-таки не дают полной уверенности в том, что мы работаем на требуемом количественном уровне для обследования такого специфического объекта, как лексический состав и его семантическое устройство. Поэтому, с моей точки зрения, если выбирать, что чему предпочесть (a giant brain or a swift hand), то я бы не стал противопоставлять одно другому, т. е. требуется ли нам гигантский электронный мозг, который бы, так сказать, все это объединил, или просто проворная рука, то для лексикографической практики, конечно, потребовалась бы и проворная рука. Прежде всего нам необходима семантическая сводка всех имеющихся словарей или семантическая подборка текстов по заранее разработанной методике.

Выяснение значений, я думаю, для многих слов русского языка — еще не вполне законченный процесс. Для меня значение слова русского языка, если это слово устойчиво находится в системе языка в течение, скажем, двух-трех веков,— это совершенно такой же непреложный факт, как наличие или отсутствие какой-либо фонемы в фонологической системе русского языка. И практика показывает, что одним словам повезло в словарной обработке,

---

человек. Это все выдающиеся люди, цвет нации, но все-таки 200 человек из 200 млн. говорящих на русском языке как на родном — не мало ли?

[5] Академическая лексикография в отношении источников базируется на общей и специальной выборке. Та и другая может быть полной и частичной. Все виды дают подбор слов по указанным признакам. Частичная выборка является основным типом. См.: Разработка лексики и фразеологии современного русского литературного языка: Пособие по выборкам. Л., 1972, с. 7—8.

а другим нет. Семантическая структура многих слов, даже при сопоставлении всех имеющихся семантических разработок на это слово, все-таки оставляет известную неудовлетворенность и кажется, что семантическая структура данного конкретного слова еще окончательно не установлена, поэтому такого рода перечни, удобные для пользования, имели бы большое значение. Если составить полный алфавитный словоуказатель ленинских произведений, то это был бы большой вклад в Машинный фонд русского языка.

Я позволю себе еще несколько общих замечаний, уже не касающихся или выходящих за пределы собственно лексикографии. Мне кажется, что ретроспективные исследования, т. е. исследования источников прошлых веков, и исследования, допустим, нашего текущего языкового состояния — это вещи все-таки трудносопоставимые в том смысле, что ситуации по векам не полностью воспроизводят друг друга, каждая ситуация уникальна. И в этом смысле ретроспективные материалы не увеличиваются в объеме и не уменьшаются, как они сохранились, так они и отражают то состояние, которое было. Наша же задача состоит в том, чтобы языковое существование XX в. было отражено гораздо полнее, чем это было сделано для XIX или для XVIII в. На эту тему на конференции говорилось очень много, в частности, много говорилось о деловой прозе, о терминологии, о научно-технической литературе, о тех особенностях, которые составляют языковое существование XX в. Меньше говорилось о языке массовой коммуникации, о языке текущей периодики. Много говорилось о лингвистическом поведении в целом, т. е. о записях устной разговорной речи. Ретроспективные исследования имеют свои особенности, свои трудности, но это вещи фиксированные и определенные. Состав памятников известен, их количество известно, а вот сосредоточиться на современном этапе — это особая задача.

А. В. ЗУБОВ

Машинный фонд русского языка можно представить состоящим из двух основных частей: 1) большого постоянно пополняющегося машинного алфавитного словаря словоформ русского языка с максимальным набором информации (структурной, лексико-грамматической, семантической, текстовой и др.) и 2) комплекса программ по автоматическому оперированию единицами и совокупностью единиц этого словаря.

Единицей такого словаря должна быть словоформа. Объясняется это следующими причинами: а) главным источником пополнения словаря будут письменные тексты, основными единицами которых является словоформа; б) легкость автоматического пополнения такого машинного словаря новыми единицами из текстов; в) словоформы несут максимум грамматической информации; г) существуют алгоритмы приведения русских словоформ к канонической форме и к феноматической записи.

Комплекс программ по оперированию единицами такого словаря может включать следующие программы:

1. Программы оперирования с отдельной единицей словаря — преобразования ее к канонической форме; преобразования ее к фонематической форме; преобразования ее к фонетической записи; преобразования ее к специальным формам для диалектологических исследований; преобразования ее к различ-

ного рода квазиосновам; морфологического анализа словоформы; морфологического синтеза словоформы; деления словоформы на слоги; определения ритмических характеристик словоформы и др.

2. Программы оперирования совокупностью единиц словаря — постоянного учета употребительности каждой словоформы; внесения новых словоформ в словарь; выбора словоформы или группы словоформ из словаря по самым различным критериям (фонетическим, фонематическим, морфологическим, слоговым, ритмическим, синтаксическим, семантическим и т. п.) и др.

3. Программы оперирования словоформами словаря при его наложении на предложение или текст — определения синтаксической функции словоформы в предложении; построения структурных формул предложений и абзацев на уровне классов слов и членов предложений; определения ритмической формулы строки (предложения); определения семантических ролей словоформы в предложении; определения семантических формул предложений и абзацев и др.

4. Программы различных статистических подсчетов.

5. Программы перевода словоформ машинного русского словаря на различные иностранные языки и языки народов СССР (в перспективе).

Такая большая работа не может быть выполнена в обозримые сроки одной какой-либо организацией (даже если она будет иметь специальное финансирование). Поэтому решать ее должны 20—30 организаций страны на основе договора о научно-техническом сотрудничестве, по единому плану, с единым руководящим центром. На этом пути возникнет большое число различных проблем. Хотелось бы отметить две из них.

Первая — это проблема взаимопонимания различных специалистов, которые будут участвовать в создании Машинного фонда. Необходимо организовать работу таким образом, чтобы все задачи создания словаря, его пополнения и пользования им ставились с самого начала в тесном сотрудничестве лингвистов и программистов.

Вторая проблема — материальная база Фонда. Понятие «машинный фонд» предполагает, что основные усилия по извлечению информации из Фонда, различным преобразованиям исходной информации Фонда должны лечь на электронную машину. Чтобы создаваемый Фонд действительно был «машинным», необходимо добиться выделения для целей организации Фонда большой современной ЭВМ типа ЕС-1060, ЕС-1050 или ЕС-1045, обладающей большой памятью и высоким быстродействием. Такая машина может быть установлена в помещении какого-либо вуза центральной части нашей страны, что снимет вопрос о выделении фондов на строительство специального вычислительного центра. Работа создаваемого Машинного фонда языка на арендуемом машинном времени может не оправдать тех надежд, которые на него возлагаются.

И, наконец, последнее. Можно много и долго спорить о структуре Фонда, типе его единиц, тех задачах, которые он должен решать, и т. п. Главное сейчас — сделать первые практические шаги в создании Фонда, выяснить, какой конкретный вклад могут внести в его создание организации, заинтересованные в нем. Например, Минский госпединститут иностранных языков готов уже сейчас передать в Фонд следующие материалы:

1. Различные русские тексты на различных машинных носителях:

а) на перфокартах:

тексты всех поэтических произведений С. Есенина (общая длина 57 тыс. словоупотреблений);

тексты некоторых работ А. Р. Лурия (85 тыс. словоупотреблений);

тексты некоторых военных очерков К. Симонова (76 тыс. словоупотреблений);

тексты некоторых математических работ А. Маркова (60 тыс. словоупотреблений);

б) на пятипозиционной перфоленте:

тексты некоторых произведений В. Катаева (75 тыс. словоупотреблений);

тексты русских рассказов (50 тыс. словоупотреблений);

в) на перфоленте для наборных машин — около 100 тыс. словоупотреблений различных текстов (научных, публицистических, художественных).

2. Программы обработки текстов:

программу построения частотных списков двух-, трех- и четырехбуквенных сочетаний;

программу построения частотных списков словоформ;

программу построения частотных списков двух-, трех-, четырех- и пятисловных сочетаний;

программу выделения из текстов предложений по опорным словам;

программу полного статистического описания поэтического текста.

Т. Ф. ИВАНОВА

Остановлюсь на двух вопросах: 1) возможности создания в рамках Машинного фонда русского языка справочной лингвистической службы языка В. И. Ленина; 2) перспективах исследования языка В. И. Ленина с помощью ЭВМ.

1. В Институте русского языка АН СССР с 1971 г. ведется работа над «Словарем языка В. И. Ленина». Картотека создавалась вручную по текстам Полного собрания сочинений В. И. Ленина (5-е изд., далее — ПСС). Она насчитывает 2,5 млн. словоупотреблений. Картотека выборочная, т. е. нет полной выборки служебных слов, не вошли слова в латинской транскрипции, имена собственные.

На базе этой картотеки составляется словарь филологического типа, написано уже свыше двух тысяч экспериментальных словарных статей. Картотека представляет собой самостоятельную исследовательскую ценность, как и картотека Словарного сектора в Ленинграде. Однако наряду с задачей быть базой для составления словаря картотека во все возрастающем объеме выполняет справочную функцию. Это связано с тем, что ленинские тексты занимают особое место в нашем обществе, в идеологии, в научной и культурной жизни.

Как правило, запросы в картотеку поступают из Института марксизма-ленинизма при ЦК КПСС. В Справочном отделе ИМЛ поиск адреса цитаты ведется с помощью тематического указателя по приблизительно содержащейся в запросе теме. Поэтому сотрудники Справочного отдела не всегда располагают возможностями для точного ответа и запрос переадресуют к нам. Картотека «Словаря языка В. И. Ленина» позволяет указать точный адрес слова, если даже оно употреблено всего 1—2 раза в 55-ти томах ПСС.

Виды запросов примерно таковы:

1. Указать адрес цитаты (по названным ключевым словам).

2. Дать все контексты употребления какого-либо слова (например, *художественный*) или адреса сочетания слов (например, *внутренний враг*).

3. Есть ли в текстах В. И. Ленина цитата с такими-то словами (например, со словами *живучесть националистических предрассудков*)?

4. Говорил ли В. И. Ленин о том-то (например, о Киевской Руси), употреблял ли такое-то сочетание слов (например, *унитарное государство*)?

5. Нельзя ли найти точное высказывание, смысл которого заключается в следующем: ... (или: выражен приблизительно такими-то словами) (например, *общее может понять каждый, а главное — вникнуть в детали*).

В основном эти запросы предполагают два вида процедур: 1) поиск адреса (адресов) слова или отрезка текста по указанным словам; 2) идентификация, проверка гипотетического текста на принадлежность В. И. Ленину. Во втором случае мы часто имеем дело с неточным употреблением цитат из текстов В. И. Ленина и так называемыми «глухими ссылками», что иногда имеет место в зарубежной печати. Есть необходимость в идентификации обратного перевода, когда, скажем, автор английской книги дает «глухую ссылку», а при переводе книги на русский язык издательство (например, «Прогресс») обязано уточнить правильность цитаты, привести ее точный адрес.

Постепенно увеличивается роль картотеки в переводе произведений В. И. Ленина на другие языки. Так, издательство «Свобода» (ЧССР), осуществляя перевод Полного собрания сочинений В. И. Ленина на чешский язык, постоянно обращается в картотеку за помощью (например, что означает слово *серый*, каков смысл сочетания *черная Дума*, и пр.).

Поиск ответов на эти запросы осуществляется вручную, по картотечным шкафам, перебором карточек-цитат в ящиках. Если цитата гипотетическая, то текст считается ненайденным только после проверки его по трем-четырем словам, а не с порога. Ручной поиск занимает много времени, сложен и громоздок. Если поступает сразу несколько запросов, для ответа на них от основной работы отвлекается весь коллектив сотрудников сектора.

Составители «Словаря» считают, что объективно существует общественная потребность в лингвистической информационной службе языка В. И. Ленина, так как запросы касаются значений слов или адресов их употребления. Возникает вопрос: отвечает ли поставленной задаче — быть банком лингвистических данных по языку В. И. Ленина картотека в ее настоящем виде — в единственном экземпляре на бумажных носителях. Да, пока она вынуждена отвечать этой задаче, поскольку ничего другого не имеется. Для составления словаря филологического типа ее материалов достаточно, для ручного поиска адресов двух-трех слов в месяц она надежна.

Но речь идет о сотнях запросов в год. И постепенно информационная функция, становясь одной из основных функций картотеки, станет и ее узким местом при традиционном способе организации. Поэтому в создании Машинного фонда — наш прямой интерес. Мы ставим вопрос о создании на базе текстов ПСС автоматизированной лингвистической информационной службы языка В. И. Ленина. Думается, эта служба может стать одной из активно работающих частей Машинного фонда русского языка.

2. Как уже говорилось, к описанию языка В. И. Ленина лексикографическими средствами Академия наук приступила более 10 лет назад. Осуществляемый сейчас лексикографический проект (автор его — Ф. П. Филин) предусматривает составление словаря филологического типа. Толковый словарь языка В. И. Ленина нужен. Очевидно, надо четко представлять себе ограниченные возможности ЭВМ в такой области, как семантический анализ слова. Машина может легко перебрать и вывести на печать 2—4 тыс. адресов слова *давать*, но определить, в каком значении употреблен этот глагол в сочетании «*давали Леонида Андреева*» (продавать (книгу) или показывать пьесу) — дело квалифицированного лексикографа.

Для дальнейшего всестороннего изучения языка В. И. Ленина с помощью ЭВМ необходимо ввести в Машинный фонд русского языка тексты ПСС (около 9 тыс. наименований документов). Это отвечает и высказанным на настоящей конференции пожеланиям: 1) чтобы корпус текстов «индивидуальной ценности» вводился в Машинный фонд в полном объеме (С. И. Гиндин); 2) чтобы Фонд, поскольку он создается «в интересах нации», содержал тексты не только художественной, но и деловой прозы (Ю. В. Рождественский); 3) что кандидатом на такой полный ввод в Машинный фонд должны быть прежде всего тексты В. И. Ленина (А. Я. Шайкевич).

Проанализировав все высказывания и предложения о структуре Фонда, о его задачах, можно заключить, что лингвистическая часть Фонда вообще может быть тесно связана с текстами В. И. Ленина. Это будет не корпус препарированных иллюстраций, а именно корпус текстов (с разметкой). На этом материале можно развернуть исследовательские программы по синонимике, семантическим полям, словообразованию, ономасиологии, стилистике, фразеологии, исторической лексикологии, терминологии. Короче, в языке В. И. Ленина не найдут материала для исследования разве что фонологи и специалисты по стиховедению.

Какими материалами для «вклада» в Машинный фонд располагает ленинская тема? 1. Сами тексты ПСС. 2. Созданный вручную в секторе «Словаря языка В. И. Ленина» словоуказатель (свыше 37 тыс. слов ленинского лексикона) с индексом частотности. 3. Исследования по языку и стилю В. И. Ленина, «филологическая Ленинианa». 4. Опыт ввода в ЭВМ ленинских текстов и их обработки. Все это позволяет непосредственно на текстах ПСС развернуть и ряд лексикографических программ: составление конкордансов, словоуказателей, частотных, обратных и прочих словарей.

Ввод текстов В. И. Ленина в Машинный фонд окажет действенную помощь в составлении самых разных методических пособий, например для кафедр общественных дисциплин в вузах. При многочисленных переизданиях трудов В. И. Ленина можно доверить машине процедуру сверки (идентификации) текстов. Можно совместить лингвистическую обработку текстов с автоматическим фотонабором, чтобы исключить многие трудоемкие этапы при переиздании произведений В. И. Ленина.

Наконец, благодаря вводу ленинских текстов в Машинный фонд мы, лингвисты, будем иметь уникальную возможность на закрытом корпусе текстов одного автора проверить на «гармонию текста и словаря» весь реестр словарных предприятий, т. е. частотных, обратных и других словарей, в их числе — толкового, созданного вручную (составление его, очевидно, должно закончиться к концу XX в.).

Н. П. КОЗЛЕНКО

Я охарактеризую проект Машинного фонда русского языка с позиции своей специальности, сущность которой — моделирование на ЭВМ класса антропоморфных функций. Проектируемый столь мощный Машинный фонд будет правильно сконструирован и реализован, если он априори мыслится оптимально антропоморфным. Это ясно из того, что эффективность Машинного фонда существенно определяется его согласованием со свойствами пользователя, т. е. с характеристиками человека как источника и приемника информации. Если суммировать высказанные здесь авторитетные мнения в свете трудоемкости задачи, ее нацеленности на будущее (я здесь имею в виду не менее чем десятилетний срок реализации Фонда), то на фоне очень быстрого развития средств вычислительной техники очевидна необходимость потенциальной универсальности Машинного фонда. Его структура не должна ставить ограничений расширению класса допустимых операций, включая, конечно, открытость для проведения устного диалога. Из изложенного вытекает необходимость так называемого системного подхода к проектированию Фонда. Системный подход предполагает привлечение смежных дисциплин, в связи с чем Московский университет (и университеты вообще), наверное, мог бы служить хорошей организационной единицей, поскольку он объединяет авторитетных представителей этих дисциплин. Лаборатория прикладной лингвистики филологического факультета, например, уже имеет определенный опыт в проведении подобных работ. Кстати, критерием отбора смежных дисциплин не может служить неосведомленность лингвиста — иначе нам пришлось бы, например, немедленно отказаться от применения современных методов и средств вычислительной техники с примыкающими специальными дисциплинами. Однако действительность показывает нам, что совместная работа прикладного лингвиста и специалиста по вычислительной технике является естественной и при правильной организации — эффективной. Это же касается других смежных дисциплин, например социологии, психологии и т. д. Важно отметить, что оперативный устный диалог на базе динамической памяти ЭВМ (т. е. памяти, содержимое которой определенным образом изменяется в результате диалога) может оказаться неальтернативным изобретением, о котором здесь весьма уместно говорилось. Это обеспечит качественный сдвиг в процессе развития и эксплуатации Фонда, что соответствует трудоемкости его создания.

Н. З. КОТЕЛОВА

К обсуждаемой теме советские лингвисты приступают не впервые, в том числе и на уровне академических учреждений. В 60-е годы в Ленинградском отделении Института языкознания были организованы две поисковых группы, которым было поручено осуществить эксперименты, сходные с теми, которые обсуждаются на этой конференции. Мне кажется, что было бы неплохо осмыслить и проанализировать итоги и уроки прежних начинаний. Тогда была бы соблюдена преемственность в разработке проблемы.

Хотелось бы высказать некоторые соображения по поводу создания Машинного фонда русского языка, имея в виду в основном письменный вариант доклада В. М. Андрющенко, которым мы располагали в Ленинграде. Многие

положения прояснились в ходе дискуссии. Но все-таки кажется, что нужно более четко знать, что имеется в виду: лингвистика помогает технике или техника работает на лингвистику. И если техника работает на лингвистику, то что именно она делает? Это машинное исполнение того же, что делают «немашинные» лингвисты, или машина дает новые, нетривиальные знания о языке, или машина хранит полученные языковедами знания, или машина воплощает, модифицирует эти знания, или (об этом больше всего говорилось) имеется в виду подача пользователям готовых знаний, введение их в обиход, экономичное использование? Все это нужно разграничивать. Ю. Н. Карауловым было высказано много существенных соображений в отношении теоретических предпосылок автоматизации в лингвистике. Развивая идеи своей книги, Ю. Н. Караулов отметил, что можно руководствоваться возможностью лексикографической параметризации языка и весь инвентарь и строй языка представить в форме словарей, используя лексикографические способы описания. Я всю жизнь занимаюсь лексикографией, очень ее люблю, и меня радует, что уже перестали относиться к лексикографии как к простой инвентаризации слов, а к работе лексикографа — как к ремесленной работе. Лексикография, действительно, описывает значительную часть языка. Но думаю все-таки, что не все языковое можно представить лексикографически. В лексикографическом виде нельзя полностью осуществить инвентаризацию языковых знаний, потому что представить нужно разные языковые сущности, разнообразные лингвистические знания — о моделях, структурах, отношениях позиций, о законах и закономерностях. В структуре языка есть такие уровни, закономерности которых лишь вероятностно связаны со свойствами отдельных слов или совсем с ними не связаны. Сейчас, правда, наука идет к тому, чтобы представить устройство языка через слово. Можно привести в качестве примера популярную актантную теорию, стремящуюся объяснить структуру предложения семантическими свойствами слова, интерпретацию словосочетания и предложения в Грамматике-80, в соответствии с которой слово предсказывает, диктует, обусловливает состав и структуру словосочетания и предложения. Ю. Н. Караулов отмечает возможность описывать язык с помощью слова и его «шлейфа». Но, по-видимому, если бы человек располагал словами и их «шлейфами», знал хорошо парадигмы слова, его интенции и актанты, он не смог бы все же говорить, так как ему нужно еще то, что предоставляет говорящему человеку, например, синтаксис с его единицами, выражающими синтаксические отношения, которые статистически (вероятностно) согласованы со свойствами отдельного слова, но могут и «перешагивать» через них. Именно поэтому человек может говорить и о новооткрытых вещах, выражать средствами языка фантазии и мифы, т. е. выражать нетривиальное содержание, не нарушая языковую норму.

Приступая к решению задач, связанных с автоматизацией в лингвистике, неплохо было бы подумать, какие именно результаты доавтоматической лингвистики приводят к мысли о создании Машинного фонда. Нужно установить, для изучения каких свойств языка показана машинная техника. Наверное, это прежде всего количество. Очень многое в языке характеризуется по этому параметру. Это не только совершенно очевидно количественно оцениваемые явления, которые могут быть объектом статистического описания, но и оценки узуса, воспроизводимости. В словарях новых слов мы маркируем окказиональные слова, при этом последовательному разграничению узуальных и

окказиональных слов мешает часто то, что мы не знаем, один раз встретилось то или иное слово в речи или не один. В работе с инновациями постоянно нужна оценка по параметру «количество».

Показана машинная техника для обработки такого объекта, как язык, в связи со значимостью для языка соотношения формы и содержания. Известен в лингвистике как основной закон языка тезис о том, что в языке существует асимметрия формы и содержания. В то же время определяющим признаком языка является соответствие означающего означаемому. Как же «основной закон» может быть несовместимым с определяющим признаком языка, противоречить ему? По-видимому, лингвисты не очень правильно понимают языковую форму и языковое содержание, не везде видят проявления формы, где они есть, по-видимому, недостаточно изучено в языкознании соотношение формы и содержания. Создание Машинного фонда явится стимулом к тому, чтобы заниматься вопросами формы и содержания в языке, а высокая степень симметрии формы и содержания в языке является предпосылкой эффективного использования машинной техники в лингвистике.

Для меня всегда был привлекателен и всегда применялся в лаборатории метод использования образа машины. Образ прохождения через машину, с ее расчетом, последовательностью, объективностью, железной логикой дисциплинирует исследование и позволяет приходить к новым результатам. В Свердловске Э. В. Кузнецова занимается «вручную» тем, что сделал с помощью машины Ю. Н. Караулов. В принципе семантический словарь можно сделать без машины: взять толкования словарей, систематизировать входящие в них слова и т. д. Но преимущество машины в том, что она не знает, что «так не может быть». Человек знает, например, что слова «пенсне» и «пловец» или слова «единица», «конец», «выпустить» не составят одного семантического поля, а машина не знает. Она последовательно, как это было ей задано, выдала ряды слов, объединяемых хотя бы одним дифференциальным признаком. В результате мы получили нетривиальные знания о языке. Оказались выявленными не только поля, как в идеографическом словаре (при этом педантизм машины способствовал уточнению понятия поля), а и такие семантические связи слов, о которых мы ничего до сих пор не знали. Не знали, например, того, что слова «пенсне» и «пловец» связаны за счет глагола «держаться», и т. д. Использование машинного подхода имеет массу интереснейших последствий для науки. Не только запуск машины, но и подготовка его, и просто образ машины помогают работе лингвиста. Кое-где Ю. Н. Караулов ограничил результаты, даваемые машиной. Например, он убрал множители с частотой выше семи. Сведения о связях слов по множителю с частотой встречаемости в толкованиях выше семи в Словаре не приводятся. Но ведь необыкновенно интересно посмотреть, что находится в поле частотных, стереотипных множителей, каков их состав, как они соотносятся с другими множителями, какие именно результаты даст их учет как классификаторов.

В. М. Андрющенко не очень четко определил значимость Машинного фонда. Перечисляются десять специальностей, которым он нужен, но не сделано попытки подумать о том, кому именно, что именно и с какой целью нужно. Нужно оценить и зарубежный опыт. Сообщается, что во Франции, Швеции, ФРГ ведутся аналогичные работы. Это составление указателей, словников, сведение текстов словарей и т. д. Но не сказано, что же с их помощью до-

было нового, какие получены новые знания о языке, как эти результаты использовались.

Обсуждаемые проблемы поворачиваются иногда такими гранями, которые вызывают тревогу. Так, Словарная картотека в Ленинграде насчитывает всего 6 млн. единиц. В ближайшем будущем, по мнению В. М. Андрющенко, эту картотеку следует заменить подготовленной с помощью машины. Видимо, не надо ничего заменять, машинная картотека может дополнить существующую. 100 лет люди, тщательно обрабатывая тексты, создавали эту картотеку. Участвовали в работе академики Обнорский, Чернышев, чл.-корр. Истрина и многие другие известные лингвисты, лексикографы. Сравнивать материалы нужно не только по количеству, но и по качеству, по содержанию. Что находится в этих французских сверхкартотеках? Можно ли пользоваться содержащимися в них материалами? Ю. Н. Караулов справедливо отмечал, что возможно «отравление» избыточной информацией машинных фондов. Думаю, что это очень существенное соображение. Глагол «идти» представлен в Словарной картотеке 4 тыс. извлечений из текстов — это один ящик. Пропорционально количеству материалов французского языка на такой глагол должно быть 50 ящиков, 200 тыс. извлечений. Чтобы сделать словарную статью на слово «идти», я лучше возьму материалы «ручной» картотеки. 200 тыс. цитат только для прочтения потребуют огромного количества времени.

Отношение к существующим в науке результатам внушает иногда опасение. В. М. Андрющенко сообщает, что в толково-комбинаторном словаре будет дана информация, 70% которой не содержится в толковых словарях. Нужны ли и возможны ли такие противопоставления? Мне представляется, что успех проектируемых мероприятий прямо пропорционален степени учета той базы, которая уже создана лингвистикой. Говорилось сегодня о том, что возможен контакт человека с машиной с помощью миллиона кнопок. Эффективность такого контакта, содержательность информации, поставляемой машиной, зависит от того, насколько полно будет использоваться накопленное человеком знание о языке. Не надо ничего зачеркивать, говорить, что и словари, и грамматики плохие, что все надо делать заново. Это очень опасно. А если учесть, что планы по автоматизации должны получить организационное воплощение, с которым связан новый расклад сил и возможностей, то отрицание того, что делается в лингвистике, звучит угрожающе.

Автоматизация нужна лингвистике, и лингвистика нужна автоматизациям. Словарь А. А. Зализняка, например, при создании которого использовалась машина,— прекрасная вещь. Словарь Ю. Н. Караулова я использовала с первого дня его появления. В Ленинграде защищалась докторская диссертация на тему «Выражение отношения в русском языке»; в ней утверждалось, что частью речи, изосемичной семе «отношение», является глагол. Интуитивно усомнившись в этом, я искала доказательства. Помог Семантический словарь. Пользуясь семантическим множителем «*отно*» (отношение), даже при отсутствии дескриптора отношение, я смогла доказать, что не является глагол изосемичной частью речи по отношению к семе отношение. Я — за контакты лингвистики с машинной техникой. Но при условии, что будет отсутствовать противопоставление: машина — хорошо, все остальное — плохо. На пустом месте никогда ничего не создавалось. Во Франции, кстати, машинные разработки — это сведение воедино имеющихся француз-

ских словарей, получение словников, т. е. как раз обобщение той базы, которая уже создана совокупными усилиями науки, а не что-то такое, что воздвигнуто на пустом месте.

Не нужно забывать и о таком важном аспекте машинизации, как философское осмысление результатов. Когда лингвисты пользуются машиной, возникает опасность истолкования метода в духе операционализма. Операционализм, как известно,— это ядро современного позитивизма, абсолютизирующего метод по отношению к действительным свойствам объекта. Когда автоматизация войдет в научный обиход и все будут «нажимать кнопки», пользоваться машинами, понадобится адекватная интерпретация ситуации человек — машина. Что дают машинные методы в гносеологическом смысле, к какому этапу познания следует их отнести, индуктивный это этап или дедуктивный, нуждаются ли полученные результаты в проверке, могут ли они непосредственно использоваться как теоретические результаты и т. д.— все это вещи очень серьезные.

Уже имеющийся в лингвистике опыт использования машин свидетельствует о том, что не следует выходить за пределы трезвых оценок, адекватно отражающих то, что именно делает машина. Например, представители «машинных» направлений утверждали, что машина осуществляет рецензирование, реферирование научных работ. В действительности же по заданным ключевым словам машиной фиксировались отрывки текста — до рецензирования здесь еще далеко. Ставится вопрос и о составлении словарей с помощью машины, об анализе текста машиной. Разве в действительности это будет делаться? Выполнение каких-то отдельных вспомогательных процессов при аннотировании работ, составлении словарей, анализ текста будут обеспечиваться машиной, но, конечно, это не рецензирование, не составление словарей, не анализ текста!

Высказанные соображения вызваны желанием увидеть благополучный исход очередного приступа к автоматизации в лингвистике. Осмотрительность и чувство реальности никогда не будут лишними. Тем более, что речь идет о важном деле. Лингвистам нужен и дисциплинирующий их действия «образ машины», нужна и реальная машинная обработка данных, представляющих объект лингвистики, и обработка полученного знания о языках. Очередная задача лингвистов — выявить и исчислить все формальные параметры описания языка и все формальные проявления языкового содержания, установить все связи формы и содержания, привести в известность все возможности формализованного представления языковых явлений. Очень важно в экстралингвистических аспектах, что язык формализует интеллектуальные сущности, например единицы мышления. Открывается возможность построения с помощью машины матрицы формализованных языковых содержаний, репрезентирующих содержания внеязыковые, что позволит увидеть, как термины (кусочки знания разных наук) или слова литературного языка членят действительность, получить инвентарь и пересечения добытых человечеством в практике познания и общения содержаний — значений. Этот инвентарь отразит в формализованном виде значительную часть познанного человеком. Кроме того, при пересечении огромного количества «кусочков знания» построятся и искусственные объекты, которые нужно будет интерпретировать, используя существующий в науке арсенал проверки гипотез, и которые аналогично искусственным объектам в таблице Менделеева могут иметь корре-

ляты в действительности, которые ждут своей интерпретации. Лингвистические матрицы знания, упорядоченные с помощью современных информационно-вычислительных средств, будут способствовать интенсификации творческого процесса в разных науках. Такая проблема может стать темой лаборатории пока еще гипотетического Института прикладной лингвистики. Во всяком случае организация в системе учреждений Академии наук подразделения, которое учитывало бы роль языка в кругу проблем, связанных с построением информационно-поисковых систем, с использованием языка в ЭВМ и применением ЭВМ в лингвистике и в котором разрабатывались бы соответствующие проблемы, представляется делом прогрессивным.

В. М. ЛЕЙЧИК

Следует обсудить некоторые сугубо практические вопросы создания Машинного фонда русского языка. Поскольку Машинный фонд русского языка будет полифункциональным, одна из важнейших задач Фонда состоит в том, чтобы удовлетворить определенные потребности народного хозяйства. Очевидно, в этом смысле было бы целесообразно начать организационную работу именно с того, чтобы подготовить разработку целевой комплексной программы по созданию Машинного фонда наподобие тех ста с лишним целевых комплексных программ по развитию научно-технического прогресса, которые сейчас разработаны и выполняются в нашей стране. В частности, уже сейчас можно было бы начать подготавливать терминологическое обеспечение решения народнохозяйственных задач, которые записаны в этих целевых комплексных программах. Представляется, что сейчас мы могли бы заинтересовать промышленность, народное хозяйство тем, что материалы Машинного фонда могут войти в терминологическое обеспечение целевых комплексных программ. Правильно говорят многие товарищи, в том числе В. И. Федосимов и другие, что, по всей вероятности, нужно сразу решать и организационную сторону вопроса. Создать Машинный фонд на основе энтузиазма и силами энтузиастов — это, конечно, нереально. А если будет заинтересованность со стороны народного хозяйства, если целевая комплексная программа будет зарегистрирована в ГКНТ и АН СССР, очевидно, мы сумеем сразу подвести материальную и организационную базу под решение этих вопросов.

В этом смысле хотелось бы поддержать предложения Ю. В. Рождественского и других товарищей, которые отмечали, что начать создание Машинного фонда можно было бы именно с ввода в Фонд тех терминологических материалов, которые у нас накоплены. Это прежде всего, конечно, информационно-поисковые тезаурусы, а их сейчас очень много, в частности, отраслевые информационно-поисковые тезаурусы, которые имеются практически во всех промышленных министерствах. К этому следует добавить, что в состав терминологического блока, который будет в Машинном фонде, нужно ввести также и материалы государственных и отраслевых стандартов на термины. В. М. Андрющенко упоминал государственные стандарты. Однако, кроме 600 государственных стандартов на термины, которые есть у нас в стране, имеется также достаточно большое количество отраслевых стандартов на термины. Сейчас число их приближается к двумстам, и постепенно это количество будет расти. Они содержат очень точную, очень конкретную информацию о научно-технических терминах, которые используются у нас в

народном хозяйстве. Хотелось бы к этому добавить, что, по всей вероятности, придется для решения этого частного вопроса принять некоторые общие решения при вводе информации в Машинный фонд. Представляется, что единицами Фонда, очевидно, должны быть разные вещи, в зависимости от того блока, который мы имеем в виду. Если это, предположим, художественная литература, то единицей Фонда должен быть текст, может быть — слово. Если говорить о терминологическом блоке, то здесь единицей должен быть научно-технический термин. При этом, конечно, нужно решить целый ряд сложных теоретических и прикладных вопросов, потому что мы не всегда знаем, где проходит граница термина. Но во всяком случае придется привлечь данные теории номинации. В терминологическом блоке единицей Фонда должна быть, точнее говоря, единица номинации, которая выражена в научно-техническом термине, а он может выступать либо в виде слова, либо в виде словосочетания, причем не всегда двухсловного, а иногда и многословного.

Далее хотелось бы сказать о том конкретном вкладе, который могут внести промышленные министерства, многие из которых уже сейчас работают очень активно над своей терминологией и разрабатывают государственные и отраслевые стандарты на термины, причем разрабатывают их в рамках своей отраслевой информационно-поисковой системы научно-технической информации наряду с тезаурусами. В частности, Министерство цветной металлургии СССР разработало информационно-поисковый тезаурус, который содержит 10 500 ключевых слов, в том числе 8690 дескрипторов. Эта работа уже завершена, тезаурус зарегистрирован во ВНИИКИ и ВИНИТИ. Одновременно ведется работа над отраслевыми стандартами на термины, относящимися к добыче руд, обогащению, металлургии и т. д. И здесь есть ряд нерешенных вопросов, которые также, очевидно, придется решать в процессе включения нашей информации в Машинный фонд. В частности, не увязаны по уровню разработки информационно-поисковые тезаурусы и стандарты, неясно, каковы будут единицы Фонда в блоках информации, взятой из тезаурусов и стандартов. Мы следуем правилу, что если имеется стандартизованный термин, то он должен являться дескриптором в информационно-поисковом тезаурусе. Значит, при передаче информации в Машинный фонд здесь должно быть достигнуто единство подхода. Однако все эти вопросы мы, очевидно, сможем решить только совместными усилиями. Нашим конкретным вкладом в Машинный фонд могут явиться отраслевой информационно-поисковый тезаурус и материалы отраслевых и государственных стандартов на термины. Но для такого участия промышленных министерств в создании Машинного фонда необходимо, чтобы была утверждена целевая комплексная программа, о которой было сказано в начале выступления.

Е. Е. ЛОВЦКИЙ

Я с большим интересом и удовлетворением выслушал выступления, прозвучавшие на конференции. Интерес мой объясняется тем, что я сам уже много лет занимаюсь машинным переводом, а удовлетворение у меня вызвал тот факт, что в выступлениях был четко сформулирован тезис о том, что если в основу Машинного фонда русского языка будет положена некая общелингвистическая модель — мысль сама по себе единственно, по-моему, правиль-

ная,— то в этой лингвистической модели грамматика и словарь должны быть согласованы друг с другом. При том, что эта мысль на первый взгляд кажется простой и очевидной, на самом деле она очень глубока и важна; настолько глубока и важна, что понять ее с первого раза бывает трудно. Впервые я столкнулся с ней в препринте З. М. Шаляпиной «О семантико-синтаксическом анализе в системе англо-русского машинного перевода». Правда, там была более глубокая, вернее, более категорическая мысль о том, что грамматика и словарь, вообще говоря, принципиально не отличаются друг от друга, и различие лежит только в степени общности той информации, которую мы помещаем в грамматику или в словарь, и что для записи информации и в грамматике, и в словаре следует использовать один и тот же формальный аппарат, и по форме своей грамматика и словарь не должны отличаться друг от друга. Только в словаре мы имеем лексические словарные статьи, в которых записываются индивидуальные свойства слов, а в грамматике мы имеем описание грамматических признаков, будь то морфологические, синтаксические или семантические признаки. Так вот, о важности этой идеи в лингвистическом плане говорилось достаточно подробно, я не буду это повторять. Я хочу только подчеркнуть, что мысль эта важна и в смысле программной реализации, потому что ставить на машину практически не отличающиеся друг от друга в формальном отношении массивы гораздо проще, чем реализовать словарь и грамматику, которые отличаются по форме друг от друга.

В заключение я хочу еще раз подчеркнуть, что единственно разумная политика в отношении Машинного фонда русского языка заключается в том, чтобы строить его на базе некой общелингвистической модели, и в основе такой модели должна лежать концепция согласованности грамматики и словаря.

В. В. МАРТЫНОВ

Не вызывает сомнения, что новое начинание чрезвычайно важно и полезно, что в зависимости от того, как пойдет дальнейшее развитие дел, многое и многое может измениться в нашей науке, по крайней мере в значительной ее части. Я хотел бы представить коротко то понимание Машинного фонда русского языка, или, как уже здесь говорилось, любого другого из национальных языков нашей страны, которое сложилось у нас в Минске в Институте языкознания, в секторе общего языкознания, в группе «Автоматизация научных исследований».

Мы занимаемся пока еще очень недолго несколькими аспектами этой проблематики, в частности проблематикой, связанной с языками представления знаний, конкордансами и фонетическими исследованиями. Но я хочу остановиться на первых двух, поскольку именно они представляют интерес для участников этой конференции.

Думаю, что такое грандиозное предприятие, как Машинный фонд русского языка, может быть представлено в виде некоторого генерализованного конкорданса русского языка. Что я имею в виду? Прежде всего, я думаю, основой для исследования языка во всех его проявлениях должен быть некоторый представительный корпус текстов. По-видимому, необходимо квалифицировать это понятие. Не только в отношении объема корпуса текстов,

и даже не столько в отношении объема, сколько в отношении представления разных сфер, в которых используется русский язык.

Мы делаем конкорданс белорусского языка XIX в. В нашем распоряжении мало текстов. В этих условиях мы, конечно, стремимся вводить все тексты, которыми в настоящее время располагаем, включая рукописные. Что касается русского языка, это, видимо, не представляется возможным осуществить где-то в обозримом будущем. Разработаны программы на основе двух экспериментальных текстов. Эти программы предусматривают следующее: прямой список словоформ, обратный список словоформ, два частотных списка, т. е. частотный и частотно-алфавитный и, что особенно важно и что не практиковалось в существующих формах конкордансов, комбинаторный список с меняющейся глубиной, т. е. на разную глубину справа и слева от данной словоформы. Методика выделения иллюстративного материала у нас еще недостаточно разработана. Она представлена в виде экспериментальных алгоритмов членения текста на некоторые синтаксические составляющие, на некоторые законченные синтаксические единицы. Предполагается, что в будущем могут сочетаться оба вида этой работы так, чтобы можно было уже расчлененный текст использовать как иллюстративный по отношению к словоформе. Думается, что фактически любые лингвистические задачи могут быть решены путем получения той информации, о которой говорилось выше и которая, как видно, сводится к статистико-комбинаторному описанию словоформы данного языка с достаточно полным представлением его в виде текстов.

Второй аспект, который мне кажется очень важным и продуктивным, сводится к следующему. Считается, что наука тогда достигает своей зрелости, когда она переходит от наблюдения к воспроизведению. Фактически зрелость, например, экспериментальной фонетики связывается с синтезом речи. То же самое можно сказать и о языке в целом, поскольку мы в настоящее время не имеем возможности синтезировать тексты на русском языке. Эта проблема оказывается весьма важной, и вот здесь вступает в игру вторая часть тех работ, о которых уже говорилось. На основе некоторого языка представления знаний создается модель, дающая возможность синтезировать тексты на данном языке в его канонизированной форме. Мы никогда не пользуемся неограниченным, близким к естественному варианту языком в целях решения каких-то прикладных задач. Мы всегда канонизируем, т. е. ограничиваем язык, но весь воспрос в том, что ограничиваем мы язык, как правило, по отношению к предметной области, а не к каким-то внутриязыковым характеристикам. Между тем язык представления знаний, определенным образом сконструированный, дает возможность создания такой модели, когда язык канонизируется независимо от предметной области. Я сейчас не могу вдаваться в подробности, могу сказать только то, что в конечном счете мы получаем образцы текста на литературном языке, которые вполне воспринимаются как репрезентирующие данный литературный язык, хотя они имеют очень существенные ограничения. Эти ограничения прежде всего сводятся к позиционной организации грамматики, т. е. для русского языка к строгой соотнесенности падежей и позиций. Порождение текстов — это синтез, который неотделим от анализа, и думается, что на основе такого рода методик можно было бы создать канонизированную форму русского языка, которая нам необходима для прикладных целей. Я лично не считал бы целе-

сообразным изучение опыта хаотического построения каких-то фрагментов русского языка в прикладных целях. Более целесообразно синтезирование русского текста на основе строго заданных правил, такое, чтобы при этом текст воспринимался как нормальный (отмеченный).

Подводя итог, хочу заметить, что представляется необходимым, во-первых, выяснить, что такое представительный корпус текстов для русского языка. Думается, что здесь должны быть представлены все жанры, включая грамматики и словари. Деловая и научная проза должна входить сюда полностью. Такого рода фонд текстов дает возможность решать не только лингвистические задачи, но и задачи информационно-поисковые.

И, второе, это — выбор модели, которая давала бы возможность синтезировать для наших прикладных целей тексты на русском языке, такие, чтобы они отвечали, с одной стороны, представлениям носителей русского языка об его узусе, а с другой — были удобны для обработки.

В. И. НАЗАРОВ

Насколько я понимаю, создание Машинного фонда должно решить две задачи: первое — проведение чисто лингвистических исследований, второе — создание словарей. Я остановлюсь на второй, близкой мне, проблеме. За последние десять лет в значительной степени увеличился выпуск словарей русского языка. Достаточно сказать, что если в 1974 г. мы давали 63 млн. п. л. выпусков и приблизительно 1300 т бумаги использовали на все словари, то в прошлом году мы использовали 1400 т бумаги только на словари русского языка. И тем не менее хотелось бы отметить, что потребности советских граждан в словарях русского языка удовлетворяются только на 10—15%. Это, пожалуй, самый дефицитный вид литературы, которая имеется в Советском Союзе. Поэтому мы делам все для того, чтобы из года в год увеличивать выпуск словарей русского языка. В прошлом году 1/3 всей бумаги потратили на словари русского языка, хотя сейчас готовим словари на 58 языках. Пожалуй, такого уникального издательства не найдешь ни в одной стране мира по количеству языков, которые находятся в работе, но тем не менее мы в большом долгу по словарям русского языка по количеству, и по типологии, и по ассортименту.

Я думаю, что проблема, которая здесь рассматривается, помогла бы решить ряд вопросов, которые мы перед собой ставим.

Какие это вопросы? Использование Машинного фонда дало бы возможность доводить действительно до самого современного уровня науки или развития языка выпуск того или иного издания, потому что машина уже сейчас позволяет почти полностью регламентировать подготовку словаря. А что такое словарь? Это самый сложный вид набора, который можно себе представить. Надо использовать до 5-ти выделений, а существующие литературные гарнитуры, которыми мы чаще всего набираем, сильно отстали от тех потребностей, тех задач, которые стоят перед словарным делом. В настоящее время разрабатывается новая словарная гарнитура, которая включает в себя 5 тыс. знаков. Это дело довольно-таки сложное, но когда мы ее подготовим, она будет использоваться как для высокой печати, так и для фотопечати.

В области русских словарей стоят огромные задачи, которые надо решать: с одной стороны, увеличить выпуск русских словарей для советских граждан, с другой — для иностранных.

Если говорить о будущем, то следует отработать общеметодологические проблемы нашего словарного дела. Это касается как общефилологических, так и научно-технических словарей. В этом отношении мы не найдем таких научных учреждений, которые занимались бы этими проблемами. Это в определенной мере сдерживает развитие типологии и расширение ассортимента выпускаемой продукции.

Сейчас мы готовимся выпускать семнадцатитомный словарь. Очень много проблем возникает с наборами. Высокий набор себя изживает, потому что это — горячий набор, он плохо действует на здоровье человека. Сейчас переходят к фотонабору, а фотонабором мы не можем выпускать многие наши словари, потому что фотонабор ограничен по части тех выделений, которые требуются для набора словарей. И это также проблема, которая сдерживает развитие словарного дела в стране.

Постановка рассматриваемой проблемы, если мы ее в ближайшие годы решим, конечно, изменит в целом положение словарного дела в нашей стране. Помимо того, что с помощью машины мы можем сделать наши словари более современными, мы можем также в значительной степени ускорить их выпуск. Вторая задача, которую можно было бы решить,— расширить ассортимент выпускаемых словарей, очень ограничен, например, выпуск толковых словарей.

Наша лексикография пошла по линии решения интернациональных проблем — у нас выпускаются двуязычные и многоязычные словари. Толковых словарей в том объеме, как, скажем, в Англии (я могу назвать до 30 толковых словарей, которые там выпускаются), нет. Но, естественно, нельзя, чтобы словарь Ожегова, скажем, находился на полке учащегося 4—5 класса, в то время как мы должны ему дать словарь на 10—20 тыс. слов, не более. Вот такого рода словари мы и должны в настоящее время готовить. Я думаю, что машина поможет нам решить вопрос расширения ассортимента тех же толковых словарей, не говоря уже о том, что можно было бы по заданной программе делать и типологически новые словари.

Хотелось бы подчеркнуть, что необходимо облегчить труд редактора, потому что редактор работает сейчас на уровне XVIII, в лучшем случае XIX в. и все это делается крайне ремесленно: картотека, затем машинописный оригинал, причем печатать на двух-трех языках довольно-таки сложно. Например, историко-этимологический словарь: никто его набрать не может, и никто не может подготовить машинописный оригинал. Поэтому машинная техника в какой-то степени, если ею правильно пользоваться, помогла бы решить и эти проблемы. Бывая за рубежом и встречаясь со своими коллегами, в том числе из Oxford University Press, я видел, что они очень неохотно делятся своими достижениями, и более того, просто не допускают к своим банкам данных, поэтому нам сложно судить о том, что у них имеется, но одно известное издательство предлагало нам выпускать русско-национальные словари буквально на всех наших национальных языках. У них в этом отношении техника значительно более развита. И в этом плане, конечно, данная конференция будет иметь историческое значение, потому что мы вступим в новый этап развития словарного дела (я говорю только о словарном деле).

Я думаю, машинная техника в значительной степени помогла бы нам удовлетворить запросы советских и зарубежных читателей в первую очередь за счет расширения ассортимента и ускорения подготовки словарей.

А. С. НАРИНЬЯНИ

Относя себя к многочисленной категории заинтересованных в успехе этого большого дела, хочу выразить глубокое удовлетворение как самим фактом постановки проблемы, так и организацией представительной конференции. Работая последние десять лет в прямом контакте с лингвистами, мы много раз имели возможность убедиться, насколько все направление в целом, сама наука, страдает от междоусобицы и своего рода сектантства. Та консолидация, которая господствует среди ученых, отражает их научную зрелость, общее понимание того, что формулируемая программа касается всех, от ее реализации все выигрывают, она всем нужна.

Прежде всего необходимо подчеркнуть, что это — важнейшая государственная задача. Возможно, что справедливость такой постановки не всем присутствующим очевидна: здесь в основном говорилось о том, какую пользу может получить лингвист или лексикограф от применения ЭВМ. Однако основное значение этой программы не во внедрении вычислительной техники в лингвистику, а скорее наоборот. Идет научно-техническая революция, связанная с появлением ЭВМ. Мы вступаем сейчас в ее новый этап — начало общей информатизации общества. Вычислительная техника активно внедряется во все сферы человеческой деятельности, ее основная задача — автоматизация обработки информации, а подавляющая часть этой информации имеет форму текста на естественном языке. Естественный язык является основным носителем информации в человеческом обществе, таким образом, технический прогресс и естественный язык оказываются неразрывно связанными, составляя, образно говоря, амальгаму будущей информационной технологии.

Однако пока рецепта этой амальгамы у нас в руках нет. Нам известны лишь фрагменты общей картины, необходимо узнать во много раз больше. И без технической поддержки лингвистических и филологических работ, без их координации преодоление этого порога качества — невозможно. Таким образом, при выделении значительных ресурсов на создание Машинного фонда русского языка речь идет не столько о развитии фундаментальной науки, сколько о вкладе в важное прикладное направление. Исследования, связанные с естественным языком, уже сейчас необходимы не меньше, чем физика или техника. Если это пока не всем достаточно ясно, то через 5 или 10 лет это будет аксиомой, понятной каждому школьнику. Дело, которое сейчас начинается, имеет первостепенное государственное значение.

Не вызывает сомнения, что Фонд — это большая междисциплинарная программа. Отсюда следует, что основные трудности, кроме различных организационно-финансовых проблем, и даже, по-видимому, прежде них, состоят в том, что эта междисциплинарность порождает разноязычие: разницу формулировок целей, методик, терминологии, точек зрения и т. д. Чтобы работать вместе, надо научиться преодолевать этот междисциплинарный барьер. Он все время дает о себе знать: для одних Машинный фонд — это прежде

всего терминологические словари, для других, несомненно, важнее текстовые данные и т. д.

В этом плане на меня очень большое впечатление произвело выступление Ю. Н. Караулова, продемонстрировавшее глубокий личный интерес и настоящую ответственность за дело в целом. Именно такой подход дает надежду на успех в работе над поставленной сверхпроблемой, о которой можно сказать, что она по всем основным аспектам неразрешима, но разрешима в каждой точке и потому должна и может быть решена в целом.

Здесь хотелось бы сослаться на имеющийся опыт, хотя и меньший по масштабу. Уже более 8 лет активно работает межинститутский проект ДИАЛОГ, который объединяет два десятка групп, работающих по проблеме организации взаимодействия с ЭВМ на естественном языке. Группы — весьма разные по своим подходам, области интересов и специализации (лингвистика, информатика, искусственный интеллект). Если при организации проекта была иллюзия, что удастся достаточно быстро наладить координацию общей деятельности, подчинить всех одному плану и выйти на совместные работы, то восьмилетний опыт — я считаю, очень положительный — дал совершенно другие результаты. Люди, которые 8 лет назад говорили на разных языках, позиции которых были резко противопоставлены, научились понимать и уважать друг друга, постепенно двигаясь к осознанию права на существование различных подходов и их взаимодополняемости. В результате сформировалась если не интегральная координация, то по крайней мере весьма полезная структура двусторонних и многосторонних рабочих связей.

Мне кажется, что сходные принципы могли быть положены и в основу организации работ по Фонду русского языка. Думается, что не надо начинать с абстрактных споров о том, что такое Фонд и что к нему относится. Надо прежде всего определить, для чего и для кого он создается. Для меня ответ на этот вопрос достаточно очевиден: Фонд создается в первую очередь для тех, кто будет с ним работать прямо, и уже опосредованно, через них, для тех, кто будет использовать их результаты. Таким образом, затрагивается очень широкий спектр задач, и сразу же пытаться решить, что — Фонд, а что — не Фонд, пока явно преждевременно. Лучше начинать с принципа полной открытости: Фонд — это все, что сейчас делается в данной области, идет от реальных задач и, по-видимому, так или иначе может быть полезно.

Естественно, возникает вопрос, как организовать работу по Фонду, чтобы, несмотря на различия по задачам и методам, все результаты в сумме сводились к некоторому общему полезному знаменателю. Думаю, что прежде чем вводить стандарты, следует наладить учет — учет всего, что уже существует в машинной форме: текстовые массивы, словари, программное обеспечение и т. д. Эта информация о том, где и что уже есть, будет большим шагом вперед, в том числе и с чисто психологической точки зрения. Учет приведет к экономии усилий, к установлению полезных рабочих связей и к осознанию многомерности и сложности задач, связанных с Машинным фондом. По мере того как будет налаживаться учет, может начинаться стандартизация. Причем постепенная, сохраняющая принцип открытости Фонда. Например, создание единого словаря не должно связываться с введением жестких форматов, даже если эти единые форматы как угодно хороши. Надо представлять в этом словаре всю информацию о слове: в таком-то словаре это слово описывается так-то, такой-то группой — так-то, и т. п. Фонд дол-

жен вмещать все, что полезно: лексические и терминологические словари, конкордансы, архивы текстов и т. д.

Последний вопрос, это — техническое обеспечение работ по Фонду. Не хочется гасить общий оптимизм насчет возможностей вычислительной техники, но, думаю, все здесь совсем не так просто, как, возможно, представляется многим присутствующим. Когда речь идет об информационных массивах подобного объема, никакие магнитные носители и ЭВМ, даже самые лучшие, а тем более рядовые не могут рассматриваться как адекватное средство. По-видимому, надо искать какие-то комплексные средства, в частности те, о которых пока здесь совсем не упоминалось. Например, фотоносители: у них достаточно высокая плотность, надежность и уровень автоматизации; они гораздо доступнее и по стоимости, и по техническому сопровождению, чем большие ЭВМ. Кроме того, очень важно понимать, что разворачивание работ по Фонду должно вестись в расчете не на имеющуюся технику. Конечно, первое знакомство лингвиста с вычислительными машинами вызывает взрыв энтузиазма и надежд. Однако это — техника уже не сегодняшнего дня, а вчерашнего. Если перспективный план по Фонду нацелен на третье тысячелетие, то уже на ближайшие 10 лет надо планировать эти работы с расчетом на перспективу, т. е. на завтрашнюю технику и технологию. Только что говорилось о перфолентах на наборные машины; но, видимо, если спросить в соответствующих ведомствах, то окажется, что через 5 лет перфолент уже не будет. Следовательно, на ту технику, которая будет через 5 лет, и надо рассчитывать в координационных планах.

С. Е. НИКИТИНА

Хочу присоединиться к мнению тех, кто считает, что Машинный фонд русского языка должен быть ориентирован в очень большой своей части на лингвистов, т. е. удовлетворять неотложные нужды лингвистов, русистов. Может быть, не только русистов, но и филологов вообще. Мы разделяем мнение Н. Н. Волкова, который говорил об изящной экспериментальной системе поиска пословиц и о самом пословичном фонде, им сделанном. Это тем более отрадно, что были известны опыты создания машинных пословичных фондов, например, в Таллинском институте языка и литературы и в Армении, где есть две информационно-поисковые системы по песенному армянскому фольклору. И вот первая такая небольшая информационно-поисковая система по фольклорному тексту — русская. Как представляется мне, т. е. лингвисту, занимающемуся исследованием фольклорных текстов и языком фольклора, устройство такого машинного фольклорного фонда? Ю. Н. Караулов, касаясь вопроса системы словарей, отмечал, что в эту систему должен входить и словарь языка фольклора. Безусловно, этот словарь должен быть семантическим словарем с большим количеством парадигматических и синтагматических связей. И в самом устройстве этого словаря должна каким-то образом разрешаться та дилемма, которая опять же была поставлена Ю. Н. Карауловым,— дилемма между культурой и структурой. Дело в том, что язык фольклора — это очень своеобразный язык и, так сказать, трудно его с чем-нибудь сравнивать. С одной стороны, это как бы язык одного автора. Но почему-то язык этого одного автора известен во множестве диалектных разновидностей, количество текстов его не конечно. С другой стороны, если считать, что это каждая, так

сказать, разновидность диалектного языка, то оказывается, что исследовать его как диалектный язык невозможно, невозможно обращение к информантам. И в этом смысле оказывается, что к нему надо подходить как к мертвому языку, как это ни парадоксально. В этой ситуации приобретает громадное значение при исследовании языка наличие конкордансов, без которых в настоящее время просто невозможно, на мой взгляд, действительно, описывать язык фольклора и тем более делать словарь языка фольклора. Я приведу один пример, чтобы было понятно, насколько сложным должно быть семантическое устройство словарной статьи языка фольклора. Есть, предположим, слово *воля*, в фольклорных песенных текстах часто встречающееся. По крайней мере, у *воли* есть три значения: *воля* как атрибут человеческой личности, *воля* как свободная жизнь и *воля* как символ свободной девичьей жизни. Причем, кроме того, что само фольклорное слово является символом некоторого метаязыкового выражения 'свободная девическая жизнь', кроме того, оно имеет некоторую предметную, затекстовую соотнесенность, референтную, а именно— *волей* называется девичий головной платок, в некоторых местах — лента, а в некоторых, где *воля* имеет еще синоним, скажем, *зеленое деревце*. Но, кроме того, воля как внутритекстовый референт этой *воли* — это некоторое живое существо, полу-рыба, полу-птица, которое может кататься, сливаться, махать крыльями, наконец, улетать от своей хозяйки, не даваясь ей в руки. И этот смысл должен быть отражен в самой словарной статье слова *воля*. Возникает вопрос, каким образом можно разрешить этот конфликт между культурой и структурой. И здесь оказывается, что на помощь описанию фольклорного слова приходит аппарат, разработанный, можно сказать, в теоретической и прикладной семантике, в прикладной нисколько не меньше, чем в теоретической; вообще в современной, если угодно, лингвистической семантике, где разработан (при создании словарей-тезаурусов различного типа, комбинаторных словарей и словарей сочетаний) такой достаточно хороший, мощный аппарат уложения регулярных смыслов, как парадигматических, так и синтагматических. Прежде всего я хочу сказать об уже известных, по-видимому, в этой аудитории лексических функциях, которые фигурируют в толково-комбинаторном словаре. Их некоторая трансформация, некоторое преобразование — функции поэтические, связанные с определенным типом текстов, текстов фольклорных, в которых регулярно одни и те же смыслы выражены идиоматическим образом при разных словах, например, «что делать с воротами?» или «что делают ворота?» — типичное действие — они всегда *растворяются*, «что делать со слезами?» — они всегда *уливаются* или *проливаются*. Скажем, вот такие типические функции могут быть представлены в структуре такого словаря. И в этом смысле, делая язык фольклора, ориентируясь на такой способ описания и представления смысла, мы тем самым сокращаем то расстояние, которое могло бы существовать между словарем фольклора в человеческом варианте и словарем фольклора уже в варианте машинном.

Это дело будущего, а для того чтобы приступить к такого рода словарю, совершенно необходимо другое, что А. С. Герд назвал «регистрирующим фондом», а именно — необходимы частотные словари по разным жанрам фольклора и очень большое количество конкордансов. Причем надо сказать, что благодаря тому, что сочетаемость фольклорного слова устойчиво идиоматическая, она как бы «улавливается» на достаточно небольшом контексте, именно поэтому конкордансы здесь больше дадут, чем конкордансы художественных текстов.

Таким образом, каким представляется фольклорный фонд? Это, безусловно, некоторая представительная выборка по разным фольклорным жанрам, это, безусловно, система частотных словарей по разным жанрам, это — конкордансы. И на их основе новый словарь — уже как бы словарь интерпретирующий, как бы словарь, который может использоваться в последующих работах по моделированию текстов.

Вторая вещь, которая чрезвычайно нужна фольклористам,— это то, что связано с поиском бесконечного количества фольклорных текстов. Все их в машине иметь невозможно, но иметь некоторый справочный фонд, где такой текст можно найти, абсолютно необходимо. Для того чтобы его иметь, нужно разработать самый первоначальный язык описаний фольклорных текстов. Если бы можно было каким-то образом зафиксировать это в решении конференции с тем, чтобы дошло до фольклористов, тогда можно было бы устроить какое-то совещание по выработке такого языка.

И последнее. Уже отмечалось, что в Машинном фонде должна была быть информационно-поисковая система по лингвистике. Мне кажется, что при ориентации Машинного фонда несколько шире — на нужды филологии — соответствующая информационная система должна была бы быть по филологии, включая, кроме большого лингвистического ядра, безусловно, еще и текстологию, фольклористику и поэтику.

Ю. К. ОРЛОВ

Самое ценное в обсуждаемом Машинном фонде русского языка — это именно его многофункциональность, многоцелевая ориентированность. Я бы хотел заострить внимание на одной специальной возможности, которую может предоставить этот Фонд. Дело в том, что до сих пор лингвистика традиционно имела дело с тем, что не превышает размера предложения, и центральной единицей было слово, его связи с другими словами, его функции в предложении и т. д.

Сейчас, в сущности, только ЭВМ позволяют изучать более объемные языковые объекты, чем предложение. Тут у нас впервые появляется возможность построения и анализа некоторых количественных мер, когда мы можем эти объекты характеризовать теми или иными наборами чисел — какими-то многомерными векторами. Я занимаюсь тем, что можно было бы назвать количественной лингвистикой, которая рассматривает определенный текст или речевой поток как объект количественной природы, в котором можно просто считать те элементы, из которых он состоит — полное число его элементов, число разных элементов и т. д. Это можно считать на разных уровнях, чаще всего счет идет на словоупотребления и словоформы, или лексемы. И вот оказывается, что с точки зрения такого количественного подхода сильно отличаются друг от друга отдельно взятый текст и какие-то отрывки или наборы многих текстов в одну выборку. Я не буду говорить об этом подробно, приведу только один характерный пример. Есть частотный словарь чешского языка. В отличие от частотного словаря русского языка там приведены не только сводные данные по всему корпусу текстов, но и основные характеристики составляющих словарь отдельных текстов. Там оказалась такая странная картина — характеристики всего словаря совершенно не похожи на характеристики составляющих текстов. При этом нельзя сказать, что характеристики всего

словаря похожи на нечто среднее по всем текстам. Ничего подобного — они сильно смещены относительно характеристик отдельных текстов. Одна из самых общих характеристик — характеристика относительного богатства лексики, например, по всему словарю значительно выше, чем по составляющим этот словарь отдельным текстам.

Если бы берем отдельные тексты, то оказывается, что характеристики их эволюционируют со временем. Например, уже упомянутое относительное богатство лексики довольно монотонно и устойчиво растет со временем. Наиболее архаичные тексты обнаруживают и наиболее бедный словарь. Чем ближе к нашему времени, тем тексты лексически богаче. Можно было бы много говорить о причинах этого явления, но, к сожалению, здесь нет ни времени, ни места. Скажу только, что когда мы анализировали библейские тексты Нового завета, упорядочили их по возрастанию относительного богатства лексики, то оказалось, что это упорядочение очень хорошо совпадает с принятой хронологией этих текстов и даже обнаружилось одно интересное обстоятельство: два послания к Тимофею и одно к Титу, которые церковная традиция считает посланиями апостола Павла, по относительному богатству лексики резко отклонились от основной массы посланий Павла, которые образовали весьма компактную группу. Но в литературе хорошо известен факт, что эти три так называемые пастырские послания написаны лет на сто позже предполагаемой жизни Павла. Метод, видимо, пригоден для установления по крайней мере относительной хронологии текстов. Если бы у нас было множество данных этого рода по самым различным текстам, наверное, можно было бы построить и абсолютную хронологическую шкалу этого показателя с учетом языка, жанра, стиля и т. п. Но это только частный пример.

В связи со всем сказанным, я думаю, что основной единицей хранения в Фонде должна быть не лексическая единица, а текст. Если Фонд будет представлять собой набор текстов, каждый из которых будет иметь четко определенные границы, то в этом случае создание самых разнообразных словарных списков будет делом чистой техники. Мне бы хотелось заострить внимание именно на этом: обязательно основой Фонда должен быть именно набор дискретных, четко отделенных друг от друга текстов, а все остальное должно извлекаться оттуда, из этой первичной информации. Ни в коем случае нельзя заранее намешивать в единый корпус тексты в какой бы то ни было априорно заданной пропорции, а потом считать, что характеристики этого корпуса будут представлять собой некое «языковое среднее». Наоборот, там возникает то, что математики называют явлением смещения, и получаемые числа сами по себе будут совершенно не интерпретируемы. Учет этого смещения и интерпретация данных по наборам текстов возможны лишь в том случае, если известны характеристики отдельных составляющих набора текстов.

Мне кажется, что здесь возникает одно чисто практическое преимущество, которое есть смысл учесть. Если начать создание Фонда именно с ввода и обработки отдельных текстов, то полезная отдача Фонда может быть получена гораздо раньше, чем этот Фонд начнет функционировать как единое целое, так как можно будет уже на ранних этапах работы выпускать результаты «локальных обработок»: словари, конкордансы, словоуказатели по отдельным текстам. При этом, разумеется, желательно, чтобы это были тексты, имеющие либо филологическую, либо историческую, либо художественную или общекультурную ценность и т. д. Не только для языковых, но и для всех гумани-

тарных и общественных наук было бы просто замечательно, если бы мы располагали хотя бы такой серией конкордансов, которая была издана в Италии под руководством д-ра Марио Алиней. Они обсчитали и издали всю итальянскую классическую литературу начиная с раннего Возрождения. Мне выдали в библиотеке примерно полуметровый столб из этих конкордансов. Это просто машинные распечатки, воспроизведенные множительным способом. Мне кажется, что если бы поставить это дело на такие рельсы, то и сам процесс создания Фонда оживился бы, он бы не превратился в такое мероприятие, которое можно бесконечно долго обсуждать, но никак не дойти до его практической работы в задуманном варианте, а можно было бы начать с простейшей обработки отдельных текстов, а затем, постепенно накапливая эту машинную библиотеку и расширяя возможности обрабатывающих программ, постепенно расширять круг функциональных возможностей Фонда.

Разумеется, нужно подумать и об издательской базе, так как только благодаря публикации основных обработок можно будет сделать результаты работы Фонда действительно доступными для больших групп исследователей; индивидуальную же работу с Фондом вести только по узкоспециализированным задачам.

В. И. ПЕРЕБЕЙНОС

Создание Машинного фонда русского языка, так же как и машинных фондов языков национальных республик,— это вопрос, решение которого откладывать дальше нельзя. Без создания Машинного фонда нельзя по-настоящему подойти к решению проблемы лингвистического обеспечения ЭВМ, необходим он также для решения многих прикладных задач, связанных с автоматической переработкой текстов. Но дело не только в прикладных задачах. Машинный фонд необходим для дальнейшего глубокого изучения русского языка и языков национальных республик. Без применения ЭВМ невозможно ни обследование больших массивов текстов, ни составление статистических справочников, ни создание многих фундаментальных работ, для которых необходимо исследование большого материала. Чрезвычайно важно использование ЭВМ при решении проблем лингвистики текста. Поэтому инициатива Института русского языка в создании Машинного фонда русского языка очень важна, ее трудно переоценить.

Материалы, представленные на конференции, показывают, насколько глубого и широко задумано создание Машинного фонда. Возникает вопрос: кто будет руководить этим Фондом, где будет научный, вычислительный и организационный центр? Естественно, это должен быть не отдел или сектор в Институте русского языка и даже не вычислительный центр. Это должно быть учреждение типа института, связанное многими нитями с целым рядом заинтересованных организаций и учреждений — и научных, и технических, и особенно издательств, без привлечения которых я не мыслю себе создание Машинного фонда русского языка. Сейчас почти в каждом крупном издательстве есть специальная редакция, в задачи которой входит автоматизация редакционно-издательских работ. Есть такая редакция в издательстве «Советская энциклопедия», есть она и в издательстве «Наукова думка», у нас в Киеве, есть она и во многих других издательствах. Для автоматизации редакционно-издательских работ необходим морфологический, синтаксический, лексический

и стилистический анализ текстов, необходимо изучение закономерностей структурной организации текстов различных жанров, построение типологии текстов, получение их статистических характеристик с целью установления тех оптимальных образцов текста, на которые должен ориентироваться и к которым должен стремиться автоматический редактор. Так что издательства — это заказчики, формулирующие задачи и оплачивающие их реализацию, а наличие таких заказчиков — немаловажная сторона функционирования Машинного фонда. Помимо этого, издательские работники могут передавать в Машинный фонд перфоленты, получаемые при линотипном наборе и пригодные для ввода в ЭВМ, т. е. могут облегчить накопление массива текстов в Машинном фонде.

Вопрос о том, каким должен быть объем Машинного фонда, дискуссионный. Мне кажется, что сейчас трудно на него ответить, пока не решен вопрос о целях создания Машинного фонда. Все зависит от задач, которые ставят перед собой создатели Машинного фонда. Если преследовать цели чисто лексикографические — это одно, если ставить более широкие задачи исследования структуры языка и закономерностей строения речи — другое, если включать в круг задач исследования по диалектологии и палеографии — третье, если же рассматривать Машинный фонд еще и как сокровищницу культуры и хранилище библиографических раритетов—четвертое. Но во всех случаях Машинный фонд, помимо словарей, должен включать и достаточный массив текстов, который нужен и составителям словарей, особенно частотных, и всем исследователям языка и речи.

Что считать достаточным объемом массива текстов в Машинном фонде? Приведу один пример: коллектив, включающий более 20 кафедр английского языка вузов Советского Союза, с 1967 г. работает над составлением частотного словаря сочетаемости современного английского языка. Так как для учебных целей отбирается лексический минимум, то было решено изучать и включить в словарь сочетаемость первой тысячи самых частых полнозначных слов. Для этого создали картотеку на общем объеме текстов 10 млн. словоупотреблений. Оказалось, что такая выборка обеспечивает минимально достаточное количество примеров (900 и более) лишь для 400 слов, остальные 600 характеризуются такой частотой, которая не обеспечивает достаточной статистической достоверности данных о их сочетаемости. Некоторые же слова в такой большой выборке встретились всего 1—2 раза или вообще не встретились. А это ведь самые частые слова (по словарю Торндайка). Не думаю, что создатели Машинного фонда русского (и любого национального) языка решат ограничить исследование сочетаемости только первой тысячей самых частых слов. Значит, массив текстов в Машинном фонде должен составлять десятки, если не сотни, миллионов словоупотреблений.

Естественно, в таких выборках можно утонуть, если не организовать этот Фонд как фонд динамический, в котором часть текстов, статистически обработанных и исследованных, в каком-то компрессированном виде уходит в архив или вообще исключается из Фонда (а результаты их исследования хранятся в банке данных), и их место занимают новые тексты, т. е. в Машинном фонде все время должно осуществляться движение материалов — обработка, переработка, компоновка, перекомпоновка; иначе этот Фонд будет мертвым грузом, скажем, таким же, каким являются хранилища больших библиотек, из которых примерно 2/3 книг никто не затребует в течение если не нескольких лет, то во всяком случае года.

Значит, организация Машинного фонда и работа в нем должны быть с самого начала очень тщательно продуманы. Из доклада В. М. Андрющенко было видно, что работа по обдумыванию структуры Машинного фонда русского языка ведется очень интенсивно. Как представляется, одной из очень важных сторон такой работы должно быть обдумывание согласованности различных участков и подфондов, которые могут быть в разных учреждениях. Иными словами, первостепенным является решение проблемы координации работы коллективов — создателей Машинного фонда, чтобы не получилось так, что есть и один фонд, и другой, и третий, а использовать их очень трудно, так как они основаны на разных принципах. И поэтому, как мне представляется, началом организации Машинного фонда должно быть тщательное планирование координации, создание единой программы и вместе с тем обеспечение специализации различных аспектов Машинного фонда и соотношения отраслевых подфондов, этапности работы.

Машинный фонд русского языка может и должен служить образцом для создания машинных фондов национальных языков, так же как и машинные фонды национальных языков могут что-то внести в общую сокровищницу машинных фондов, делясь опытом и организационной работы, и работы непосредственно по созданию Машинного фонда. Следовало бы подумать и о совместимости машинных фондов национальных языков с Машинным фондом русского языка.

У нас тоже начаты работы по созданию Машинного фонда украинского языка. В Институте языковедения им. А. А. Потебни АН УССР над этим работает пока что небольшая группа под руководством доктора филологических наук М. М. Пещак, сотрудничающая с Кировоградским государственным педагогическим институтом. Эти коллективы начали работу с ввода в память ЕС ЭВМ одиннадцатитомного толкового словаря украинского языка. Словарь вводится полностью — с толкованиями, пометами и примерами-иллюстрациями. Это первый этап нашей работы, а над дальнейшими ее этапами мы сейчас думаем. Надеемся, что в создании нашего машинного фонда нам помогут и наши товарищи, работающие над созданием Машинного фонда русского языка, и товарищи, создающие фонды эстонского, казахского, грузинского и других языков. В свою очередь мы готовы и поделиться опытом нашей работы, и предоставить нашим коллегам интересующие их материалы.

М. М. ПЕЩАК

Если Машинный фонд русского языка рассматривать как научно-исследовательское учреждение, которое должно стать ведущим в области автоматизации лингвистических исследований (а именно так оно и представлено Ю. Н. Карауловым), то в обязанности Фонда должна была бы входить наряду с выполнением конкретных плановых тем большая координационная работа. Поэтому при Фонде русского языка полезно было бы создать Всесоюзный координационный совет, в состав которого входили бы ученые всех отраслей современного отечественного языкознания и ряда смежных наук. С самого начала существования данного совета целесообразно было бы в плановом порядке глубоко и всесторонне изучить работу всех коллективов страны, так или иначе занимающихся вопросами автоматизации лингвистических исследований, автоматического использования данных естественного языка. Осуществить эту

научно-организационную работу можно было бы путем заслушивания и обсуждения отчетов о содержании, результатах и перспективах исследований отдельных коллективов страны. Обсуждение отчетов помогло бы выявить в работе отчитывающихся коллективов то, что представляет научную ценность для Фонда русского языка как руководящего организационного центра. Координационный совет, таким образом, мог бы взять на учет и определить формы использования через Фонд русского языка всего, что имеется сегодня на местах по накоплению фактического материала, алгоритмов и программ. Понятно, что на начальном этапе координационному совету пришлось бы больше заниматься вопросами обобщения и распространения опыта лингвистов по использованию ЭВМ, чем обсуждением и утверждением теоретических основ, научных концепций выполняемых исследований. Однако этот вид научно-организационной работы Фонда русского языка сыграл бы неоценимую роль в объединении усилий большой армии исследователей, теперь настолько разрозненных, что актуальность проблем, которыми они занимаются, не всегда осознается во всем многообразии.

Всестороннее обсуждение отчетов в координационном совете Фонда русского языка способствовало бы также корректировке задач и аспектов исследований отчитывающихся коллективов, созданию в стране системы строго специализированных и взаимно дополняющих друг друга научных направлений. В сеть координированных организаций с целью обеспечения максимально объективного решения вопросов о специализации научных коллективов и налаживания их сотрудничества, очевидно, в обязательном порядке стоило бы включить все академические институты, в том числе и союзных республик, наиболее сильные языковедческие кафедры вузов страны, а также неакадемические НИИ.

Координационный совет при фонде русского языка мог бы сыграть необычайно важную роль в повышении уровня лингвистических исследований, даже таких, при выполнении которых не используют ЭВМ. В настоящее время над возможностями автоматизации лингвистических исследований не только задумываются, но и делают попытки применять ЭВМ в традиционной лексикографии, диалектологии, издании исторических письменных памятников и их обработке, при решении ряда этимологических задач, в стилистике, экспериментальной фонетике и т. п. Стимулирование процессов автоматизации во всех областях языкознания способствовало бы вовлечению в научный обиход все большего количества фактического материала и повышению уровня развития науки в целом.

Для украинских языковедов в связи с созданием Фонда русского языка особенно важным является то, что в нем планируется создавать автоматизированную систему — универсальную сокровищницу лексического состава русского языка как национального и как языка межнационального общения. В настоящее время Институт языковедения им. А. А. Потебни АН УССР работает над автоматическим созданием семантического словаря украинского языка. Мы исходим из того, что современные словари, предназначенные для людей, содержат достаточно совершенный аппарат описания лексической семантики и что машинная лексикография сможет успешно развиваться только при условии творческого использования этого формального аппарата. Поэтому в качестве исходной материальной базы для составления семантического словаря используется одиннадцатитомный толковый Словарь украинского языка, пол-

ный текст которого планируется записать на машинные носители на протяжении 1983—1990 гг.

Структура семантического словаря предусматривает обработку 135 тыс. слов и упорядочение их по матричному принципу на основании двухмерного распределения совокупностей формальных признаков, которыми сопровождаются толкования лексических значений этих слов как единиц словника в Словаре украинского языка (о принципах автоматической обработки лексикографических данных более подробно см. в кн.: *Клименко Н. Ф., Пещак М. М., Савченко И. Ф.* Формализованные основы семантической классификации лексики. Киев: Наукова думка, 1982. 250 с. На укр. яз.). Кроме традиционного назначения как очередного и нового типа лексикографического труда, семантический словарь включается в более сложную систему, которая будет работать на базе использования естественного языка в виде автоматического помощника издательского редактора.

В процессе подготовки материальной базы составления семантического словаря будут разрабатываться алгоритмы и программы: а) автоматизированной записи на машинные носители тех фрагментов словарной статьи, которые наиболее полно унифицированы; б) полного автоматического анализа формального аппарата лексикографической интерпретации значений в Словаре украинского языка; в) автоматического установления списков типовых формул толкования, их соотношения между собой и с единицами словника; г) технических приемов построения машинной матрицы, автоматического упорядочения в ней единиц словника и др.

В ходе подготовки семантического словаря будут решаться вопросы о возможностях его автоматической эксплуатации: а) использования в качестве первоосновы для изучения теоретических проблем современного украинского языкознания, в частности лексикологии и семасиологии; б) включения его в систему автоматизации издательского редактирования текстов; в) расширения объема словаря и совершенствования его структуры с пополнением материальной базы данных.

Составление и подготовка к печати семантического словаря планируется на 1990—1995 гг. В составлении словаря будут принимать участие Институт языковедения АН УССР (ведущая организация), Кировоградский пединститут, Ужгородский университет и Киевский филиал Украинского полиграфического института. Однако это не единственная лексикографическая работа, которая будет автоматически создаваться на базе одиннадцатитомного Словаря украинского языка. Путем несложного автоматического перебора единиц словника планируется подготовить несколько типов грамматических, словообразовательных, акцентуационных словарей, а также словарей фонетических и морфонологических чередований, паронимический словарь и др. Стилистические пометы в одиннадцатитомном Словаре послужат для автоматического составления нескольких стилистических словарей. Некоторые признаки формул толкований дают возможность создать, например, школьные тематические словари с заданным уровнем грамматичности и др.

На основании введенного в память ЭВМ одиннадцатитомного Словаря украинского языка в 1990—2000 гг. предусматривается создание автоматизированной системы, состоящей из накопителя лексикографического материала и подсистемы эксплуатации. Лексикографический материал будет пополняться за счет существующих и создающихся в украинском языкознании словарей:

одноязычных, двуязычных, терминологических, исторических, этимологического, фразеологического и др. При пополнении лексикографического материала разноструктурными словарями возникает проблема идентификации вводимых в ЭВМ данных. В связи с этим предусматривается создание библиотеки программ по упорядочиванию и унификации разноструктурных материалов в накопителе.

Подсистема эксплуатации будет заниматься: 1) созданием машинных словарей для обслуживания автоматизированных систем конкретного назначения; 2) составлением двуязычных, в первую очередь украинско-русского и русско-украинского, сопоставительных словарей идеографического типа и др.; 3) обеспечением экспериментальными материалами теоретических исследований по изучению истории и современного состояния украинского языка, его связей с близко- и отдаленнородственными языками. Для этого планируется разработка алгоритмов и программ автоматизации трудоемких вспомогательных этапов лингвистических исследований (в том числе составления диалектологических карт и диалектологических словарей).

Подсистема будет также состоять из службы справок по вопросам культуры речи и правописания, рекомендаций по совершенствованию делопроизводства, оформления и составления текстов средств массовой информации, процессов нормирования современного украинского литературного языка и регулирования языковой политики в республике.

Запланированная Институтом языковедения им. А. А. Потебни АН УССР работа в области машинной лексикографии, как нам кажется, будет во многом совпадать с работой Фонда русского языка. Поэтому, несмотря на то что украинские языковеды будут обрабатывать в основном украинский лексикографический материал, их опыт и результаты могут быть полезны и для Фонда русского языка. В свою очередь мы питаем большие надежды на то, что в Фонде русского языка многому сможем поучиться.

А. А. ПОЛИКАРПОВ

Я должен с большим удовлетворением для себя отметить, что то убеждение в возможных функциях замышляемого Машинного фонда, которое у меня сложилось, когда о его проекте было объявлено, было подтверждено во многих выступлениях и, по-моему, оно выражается в том, что основной функцией такого Фонда должна быть, несомненно, лексикографическая функция. И это не узкая задача, это — задача, которая является настолько широкой, насколько это можно себе представить. Причем такая задача, за которую можно взяться уже сейчас, поскольку и единица анализа при составлении этих массивов и их анализа известна, и в общем-то имеется некоторый опыт работы с такими массивами, и в этом смысле это будет как раз проявлением возможности поставить непосредственно с самого начала практические лексикографические и лексикологические задачи. Здесь и проблема метаязыка в определенном смысле решена тем, что имеется практический естественный язык толкований, язык дефиниций, которые удовлетворяют нас уже длительное время и могут удовлетворять нас поначалу и при образовании Фонда. Какие более конкретные задачи можно видеть при выполнении Машинным фондом этой функции? Та деятельность Машинного фонда, которая осуществлялась бы при выполнении этой функции, мне кажется, могла бы быть определена (об этом

уже говорилось на нашей конференции) как осуществление некой национальной лексикографической службы. В этом смысле, как мне видится, она должна решать две основных задачи. Это, во-первых, отслеживать динамику появления и отмирания слов, появления и развития новых значений у слов литературного и обиходного языка. Во-вторых, интегрировать деятельность отраслевых терминологических служб, для того чтобы непрерывно растущее лексическое богатство русского языка не осталось бы за пределами наших академических словарей. Причем мне представляется наиболее важной, наиболее принципиальной функцией подобного рода лексикографической службы функция отслеживаний изменений, динамики в литературном общеупотребительном языке. И это связано с тем, что данная сфера в большей своей части связана с обиходным, устным общением, с изменяющимся объектом, с быстротекущим объектом, и изменения эти необходимо фиксировать, чтобы они не были вовсе утеряны. В каких формах может осуществляться, на мой взгляд, именно эта следящая функция? Эта функция имеет, несомненно, культурный характер. Это фиксация того культурного нашего наследия, которое выражается в нашем обиходно-бытовом общении или в текстах классической литературы. Прежде всего, мне кажется, что необходим очень мощный текстовой банк, который нужно систематически исследовать с тем, чтобы с помощью дальнейшего ручного семантического анализа выяснить то состояние, которое мы имеем на данный момент в нашем литературном языке в употребительности тех или иных значений. Этот первоначальный массив текстов может быть обработан с целью получения данных о реализованных значениях, о частотах этих значений, может позволить сразу же внести некоторые существенные коррективы в данные толковых словарей, в готовящиеся проспекты толковых словарей о некоторых значениях, которые вообще выпали из употребления, в ранжирование значений, а во многих случаях подобного рода анализ может позволить нам вообще более точно судить даже о правильности разграничения значений в существующих в настоящее время толковых словарях. Дальнейшее заведение данных о значениях слов, которые были получены ручным образом, в машинную память позволяет нам получать частотно-семантические словари, алфавитные и собственно частотные, позволяет нам получать частотно-контекстологические словари, чтобы можно было проанализировать, какие контекстуальные условия сочетаются с реализацией тех или иных значений. Опять-таки эта контекстологическая информация может быть использована для увеличения полноты описания значений слов в наших толковых словарях, а также эта информация, несомненно, может быть использована в разработке автоматических контекстологических словарей. Дело в том, что разрешение многозначности — одна из ключевых проблем в системах автоматической обработки текстов, обработки лексической информации. Необходима система обнаружения несоответствий реальных текстов с теми значениями, с той семантизацией, которая имеется в толковых словарях. Это — тот аспект взаимодействия между текстовой реальностью и словами, который отмечал А. Я. Шайкевич и который мне представляется принципиально важным. В результате подобного функционирования лексикографической службы степень соответствия между текстами и словарями будет повышаться. В дальнейшем в результате первоначального накопления данных о реализации тех или иных значений и концептуальных признаков, которые существуют в реализации этих значений, мы можем автоматизировать процесс отслеживания реализа-

ции тех или иных значений, появления новых значений, которые раньше не существовали. Диагностика может осуществляться по появлению тех контекстуальных условий у знакомых нам слов, которых раньше у них не было. Тем самым проблемы отслеживания семантической динамики в какой-то степени могут быть облегчены, и на печать для последующего анализа семантизаторами, профессиональными словарниками, будут выводиться сомнительные случаи. Вот, как мне представляется, может работать национальная лексикографическая служба, и подобного рода попытки работы с массивами и выявление степени соответствия той информации, которая есть в толковых словарях, и того, что нам представляет живая текстовая реальность, — это те попытки, которые мы осуществляли на филологическом факультете. Мы имеем программу по распечатке текстового материала в виде конкордансов, около трех лет ведем работы по семантизации текстовых массивов. Семантизировано около 200 тыс. словоупотреблений, имеется масса случаев, когда нас большой семнадцатитомный словарь не удовлетворял, и подобного рода случаи мы рады, конечно же, показать профессионалам, которые работают над академическим словарем.

Н. Н. ПШЕНИЧНОВА

Организация Машинного фонда русского языка — дело очень большое, исключительно сложное и своевременное. Мы знаем, что на ЭВМ уже давно решаются важные, во многих случаях сложные, но отдельные, не связанные друг с другом лингвистические задачи. Создание Машинного фонда, как мне кажется, означает то, что лингвисты переходят к системному применению ЭВМ, ставится цель создать автоматизированную систему обработки данных русского языка. Эффективность этой системы определяется, в частности, тем, насколько основательно она будет опираться на информационные базы, или базы данных. Как и многие товарищи, считаю, что в базу данных должны войти разного рода тексты: литературно-художественные, научные, представляющие разговорную речь и т. д., разного рода словари, в том числе диалектные, исторические и т. д. Хочу обратить внимание на один вид текстов, которые тоже должны войти в базу данных русского языка. Так, А. С. Герд, в частности, заметил, что в Машинный фонд должны войти и диалектные тексты. Я считаю это правильным и хочу остановить внимание на таких диалектных текстах, которые являются текстами особого рода. Это материалы Диалектологического атласа русского языка. Материалы атласа — это огромное собрание записей диалектной речи, отражающей частные диалектные системы (или говоры) русского языка, территории наиболее древнего поселения русских, т. е. до XVI в. Это территория, где издавна шел процесс формирования основных диалектов русского языка. Материалы отражают говоры примерно 5 тыс. населенных пунктов. Для сравнения можно отметить, что это в 6,5 раз больше, чем количество населенных пунктов, вошедших во французский лингвистический атлас (в нем 639 населенных пунктов). Материал по каждой частной диалектной системе, по каждому населенному пункту хранится в Институте русского языка в виде записей в транскрипции, сделанных от руки в тетради. Каждая такая тетрадь — это ответы на вопросы «Программы собирания сведений для составления диалектологического атласа русского языка» (М.; Л., 1947). Каждый ответ имеет номер, соответствующий номеру Про-

граммы, и представляет собой перечень примеров, которые репрезентируют то или иное звено языковой системы, т. е. отвечают на вопрос Программы. Таким образом, во всех материалах Диалектологического атласа русского языка живая диалектная речь дана как бы в препарированном виде. Каждое фонетическое слово или сегмент фразы или фраза записаны в тех ответах на вопросы Программы, которые могут быть проиллюстрированы данным примером. Кроме такого материала, для многих частных диалектных систем приведены также записи в транскрипции спонтанной речи. Материалы Диалектологического атласа представляют собой огромную ценность для лингвистических исследований в самых разных аспектах. На основе этих материалов уже создан Диалектологический атлас русского языка. Однако созданный атлас отражает основные диалектные различия русского языка, материалы же атласа содержат большое количество и других сведений о русских говорах, относящихся ко всем языковым уровням. Кроме того, те диалектные различия, которые даны в атласе, могут изучаться и в другом плане, и другими методами. Материалы Диалектологического атласа позволяют разрабатывать самую широкую проблематику, относящуюся к исследованию частных диалектных систем и системы русского диалектного языка в целом как в плане синхронии, так и в плане диахронии, позволяют разрабатывать вопросы общего языкознания, т. е. материалы Атласа могут использоваться в самом широком плане, и они действительно постоянно находятся в работе: ими пользуются не только сотрудники нашего Института, но и работники многих других научных и учебных учреждений. В связи с этим я хочу отметить следующее. Научная ценность материалов Атласа ни у кого из специалистов не вызывает сомнения, и диалектологи, историки языка постоянно с ними работают. Но в результате материалы ветшают и могут прийти в негодность, а этого мы допустить не можем. Материалы собирались в течение примерно двух десятилетий, с 1946 г., на их собирание были потрачены огромные научно-организационные силы и материальные средства. Достаточно сказать, что к участию в собирании этих материалов были привлечены практически все вузы на территории европейской части СССР, имеющие кафедры русского языка. Научная ценность материалов повышается еще и в связи с тем, что этот материал никогда уже не может быть собран еще раз, так как то состояние русского языка, которое они отражают, никогда не может быть восстановлено. Наши материалы отражают состояние русского языка в форме его диалектов, в их соотношении между собой и с литературным языком, какое сложилось примерно к середине XX в. Сейчас это состояние русского языка существенно изменилось, что естественно для говоров как для любой живой, развивающейся языковой системы, особенно в современных условиях постоянного и активного воздействия на говоры литературного языка, с одной стороны, и междиалектного контактирования — с другой. Следовательно, наши материалы в случае их утраты не могут быть восстановлены. Материалы эти уникальны, следовательно, они должны быть сохранены. Если бы речь шла только о хранении материалов, то можно было бы использовать способ хранения в виде сильно уменьшенных фотокопий, где носителем информации является специальная фотопленка. Но, говоря о Машинном фонде русского языка, мы имеем в виду не только хранение, но и доступность материалов для их обработки на ЭВМ.

Все сведения о русских говорах, которые содержатся в материалах Диа-

лектологического атласа, должны войти в базу данных русского языка. Информация о говорах должна быть помещена в памяти ЭВМ. В ряде вузов и академических институтов хранятся материалы, отражающие русские говоры территории позднего заселения (например, в Саратове, Ростове-на-Дону, Казани, Уфе и др.).

Организация Машинного фонда русского языка должна, очевидно, привести к тому, что в целом в результате усилий многих коллективов будет создана единая система данных русского языка, т. е. будет создана информационная модель русского языка. В эту систему данных непременно должны войти сведения о частных диалектных системах, на основе которых может моделироваться система русского диалектного языка в целом или, иначе, его диасистема и могут быть получены сведения диахронического характера. Без этих данных, без данных частных диалектных систем информационная модель русского языка неизбежно была бы ущербной.

М. Н. РЕММЕЛЬ

Мне бы хотелось поделиться некоторым опытом, так как в нашем секторе вычислительной лингвистики Института языка и литературы Академии наук ЭССР фонд эстонского языка в существенной степени реализован и мы им пользуемся как в теоретических, так и в прикладных задачах. Наш фонд, разумеется, не такой грандиозный, как планируется здесь, но опыт его построения может быть полезен. Сначала остановлюсь на том, что мы имеем, затем перейду к замечаниям, к практическим советам.

Сектор имеет ЕС-1010 и СМ-4, связывает их СМ-52, которая имеет сразу две системы команд — и той, и другой машины. 10 дисплеев, которые могут работать одновременно, на прямых линиях от машин стоят в типографии фотонаборные автоматы. Есть банк словарей, от одного- до семиязычных, один из языков всегда эстонский. В общей сложности в банк входят 14 словарей, часть которых введена еще не полностью. Среди них, например, Грамматический словарь эстонского языка, Словарь диалектов эстонского языка, Большой русско-эстонский словарь, Большой англо-эстонский словарь, четырехтомная Энциклопедия для молодежи, Словарь сопоставлений прибалтийско-финских языков и др. Из грамматических алгоритмов реализованы фонология, морфология, морфологический анализ и синтез, а также историческая фонетика. Есть массив различных текстов.

Перейду теперь к некоторым выводам из проделанной нами работы и к некоторым советам. Мы считаем, что, во-первых, удачность или неудачность фонда зависит от числа и характера запросов в фонде, меньше это зависит от близости фонда к структуре языка или структуре лингвистики. Иными словами, мы считаем, что в первую очередь фонд должен, так сказать, работать. Во-вторых, мы считаем, что фонд сам должен себе зарабатывать деньги на существование — все услуги должны иметь цену, в том числе и в валюте. Такой опыт у нас имеется, даже фонд эстонского языка может себе зарабатывать валюту. Ю. Н. Караулов в своем выступлении задал вопрос: производственный или научный фонд? И ответил — научный. Я ответил бы так: по содержанию научный, по форме производственный. В-третьих, если у вас имеется научный по содержанию и производственный по форме фонд, то в первую очередь вы должны заботиться об интерфейсах, т. е. о сопряжении с дру-

ными фондами в мире, в Союзе. Должны заботиться и о физических интерфейсах, в первую очередь о фотонаборных автоматах, посредством которых продукты вашего фонда печатаются. Это тоже производство, причем выгодное. В Эстонии в настоящий момент ситуация такая, что все существенные с точки зрения фонда печатаемые словари и все энциклопедические словари идут через машины нашего сектора. Взамен мы получаем, кроме самих словарей, еще набранные вручную, но через фотонабор, перфоленты журналов, газет, художественной литературы. Это составляет наш текстовый массив. В-четвертых, необходимо ориентироваться на мини- или микромашины — центральный процессор большой машины всегда может быть составлен из нескольких микропроцессоров. Ваша польза будет в том, что весь комплекс обслуживается парой инженеров вместо двадцати при больших машинах при той же производительности. И стоит дешево.

В. Ш. РУБАШКИН

Хотелось бы высказать несколько соображений с позиций человека, причастного к разработке и использованию информационно-поисковых систем. Мне представляется — и я думаю, что это не только мое мнение, — сейчас развитие информационных систем, и документальных, и фактографических, вступает в период широкого использования морфологической и грамматической словарной информации. Можно назвать ряд систем, в том числе находящихся в промышленной эксплуатации, в которых такого рода информация используется уже сейчас, например, в отраслевых институтах информации электротехнической и угольной промышленности. И сейчас, в ближайшие годы, по-видимому, будет решаться вопрос: как, каким способом будет создаваться словарное обеспечение систем, использующих элементы морфологического и синтаксического анализа для решения задач информационного анализа текста. Будет ли это происходить так, как велась на предыдущем этапе разработка информационно-поисковых тезаурусов — по принципу «каждый у себя в огороде», или эта работа будет разумно централизована. Мне кажется, что при планировании работ, связанных с созданием Машинного фонда русского языка, нужно принять во внимание то, что есть работы разной степени срочности. В частности, работы, о которых здесь идет речь, следовало бы всячески форсировать, потому что лет через пять выполнять эти работы централизованным порядком будет, наверное, поздно — так или иначе они будут делаться заинтересованными организациями.

Добавлю к этому, что средства автоматического морфологического и синтаксического анализа в виде машинно-читаемых словарей и сопряженных с ними программных продуктов должны стать в ближайшее время стандартным инструментарием любой исследовательной группы, ориентированной на создание действующих экспериментальных моделей анализа текста. Ясно, что впредь нельзя всерьез заниматься семантикой текста, проблемами, скажем, анализа связного текста, не имея такой инструментальной базы. Дело здесь становится аналогичным положению с научным приборостроением: если нет готовых приборов и качественных материалов, силы и время уходят на их создание, а не на решение собственных задач; общий уровень работ при этом неизбежно занижается. К тому же исторически сложилось так, что именно русскому языку в этом плане менее всего повезло. Системы синтаксиче-

ского анализа до сих пор разрабатывались преимущественно в рамках работ по машинному переводу, и русский язык здесь рассматривается обычно лишь в плане синтеза. Это делает задачу еще более злободневной.

Для задач информационного анализа текста по целому ряду причин сейчас нет нужды стремиться к исчерпывающе полному синтаксическому разбору. В первую очередь представляют интерес средства анализа типовых конфигураций: выделение и анализ именных групп и распознавание связей «по валентности». Можно считать, что собственно морфологическая часть и соответственно описание связей согласования достаточно разработаны как методически, так и в материале. Хуже обстоит дело с «моделями управления». Хотя теоретическая основа здесь, видимо, достаточна, требуются определенные усилия, чтобы согласовать существующие взгляды и выработать единый методический подход к построению этой части словарного описания и соответственно к приемам анализа текста. Подчеркну, что ближайшая реальная цель, которая может быть обеспечена массовым словарным описанием, на мой взгляд,— это частичный синтаксический анализ. Видимо, полезным было бы создание рабочей группы специально по этим вопросам.

И еще одно замечание, уже о другом. Много замечаний было о необходимости включения в Машинный фонд русского языка представительного корпуса текстов как базы для выяснения контекстов употребления лексических единиц, статистических исследований и проч. Полагаю, следовало бы подумать и о создании достаточного корпуса синтаксически и даже семантически размеченных текстов. К тому можно привести много аргументов, назову два, которые представляются мне наиболее важными. Такой материал при достаточном объеме мог бы служить источником для получения алгоритмическим путем весьма разнообразной словарной информации и проведения различных, в том числе и семантических, классификаций лексики. Кроме того, существует далеко не беспочвенная точка зрения, согласно которой размеченный текст сам по себе или вместе с другими средствами может служить при использовании метода аналогий эффективным инструментом анализа текста.

Л. З. СОВА

По определению Ю. Н. Караулова, Машинный фонд — это искусственно конструируемый объект, т. е. объект, создаваемый в соответствии с образом, существующим в нашем сознании. Каждый выступающий имеет свой образ Машинного фонда, и объединить наши усилия мы сможем, только найдя основание сравнения для сопоставления этих образцов и сконструировав новый объект — обобщенное представление о Машинном фонде.

Попробуем это сделать. Как показывают выступления, русский язык при обсуждении вопроса о статусе Машинного фонда (МФ) являет себя в двух ипостасях: 1) как своего рода «языковой мир», представленный в объективной действительности наряду с миром животных или миром растений, и 2) как средство общения, орудие и инструмент познания. Преломляя эту дихотомию на конструирование МФ, мы можем говорить об объединении в рамках будущего МФ двух учреждений ($МФ_1$ и $МФ_2$): цель $МФ_1$ можно определить как собирание воедино «экземпляров» русского языка, их охрану и в конечном счете — создание «Красной книги государственного языка», государственного заповедника, оснащенного ЭВМ, т. е. $МФ_1$ — это сокровищница всего

лучшего, что накоплено в русском языке и литературе, память о том, что создано на русском языке человеческим сознанием, собрание реликвий, которые мы передаем потомкам. В отличие от $МФ_1$, $МФ_2$ может иметь функции научно-производственного объединения по переработке (в различных целях, определяющих структуру НПО) языковой информации с помощью ЭВМ.

Естественно, что вопрос о структуре, организации, характере планируемых работ, направлениях исследований и финансировании МФ попадает в этом случае в зависимость от целей создания указанных выше подразделений МФ и тех функций, которые на них возлагаются. Так, при организации хранилища русского языка ($МФ_1$) первоочередной представляется дихотомия «синхрония — диахрония» и выделение на ее основе в $МФ_1$ двух отделов: 1) библиотеки древнерусских памятников и 2) записей текстов на современном русском языке. Структуру каждого из этих отделов можно определить, исходя из оппозиций «устная—письменная речь», «язык как процесс и как результат», «литературный язык — диалект» и т. д. Учет этих параметров позволит действовать не стихийно, а целенаправленно, отбирая в $МФ_1$ все, что является релевантным для характеристики русского языка как феномена, развивающегося во времени и пространстве объективной действительности: памятники древнерусского языка и образцы современного языка (например, в виде хранящейся в Ленинграде семимиллионной картотеки цитат), записи текстов в письменной форме и образцы звучания устной речи (в нормативном и в диалектном произношении, например трехмиллионную картотеку русских говоров), произведения, относящиеся к различным жанрам и стилям (в частности, примеры деловой, научной и художественной прозы, публицистики, поэзии и т. д.).

В отличие от $МФ_1$, научно-производственное объединение по переработке информации с помощью ЭВМ ($МФ_2$) может иметь структуру, определяемую на основе совсем иных принципов. Так, во главу угла здесь можно поставить противопоставление языка как инструмента познания (в лингвистике, для лингвистики) и языка как средства коммуникации, обслуживающего в основном слои общества, не имеющие отношения к языковедам. Первый аспект приводит нас к созданию Института русского языка на базе ЭВМ, целью которого является разработка вопросов общего и русского языкознания с помощью ЭВМ, например исследование так называемых вечных проблем лингвистики о соотношении языка и мышления, частей речи и членов предложения, грамматических категорий, классов слов, а также более частных проблем, связанных с преподаванием и изучением русского языка. Второй аспект выводит за рамки собственно лингвистики и требует создания отдела $МФ_2$, который будет работать «на потребителя», нуждающегося в результатах переработки языковой информации. Образно говоря, в первом отделе $МФ_2$ ЭВМ ставится на службу лингвистики, во втором лингвистика — на службу другим наукам. Для функционирования обоих отделов должен быть создан третий отдел $МФ_2$ — по разработке собственно проблем автоматизации лингвистических процессов, результаты которых будут использоваться первым и вторым отделами $МФ_2$.

Если принципы, определяющие структурную организацию $МФ_1$, можно охарактеризовать как функционально-дескриптивные, то для решения вопроса о структуре $МФ_2$ наиболее существенными оказываются характеристики, вскрывающие различие возможностей человека и ЭВМ. Те же факторы обус-

ловливают и отличие нового Института русского языка от прежних институтов. Быстродействие ЭВМ и практически не ограниченные объемы вводимой в нее информации — эти параметры могут быть основными при определении тематики нового института. Например, этому учреждению будет по силам решить задачу о создании матрицы, в которой бы отмечались все потенциальные возможности всех слов русского языка к связям друг с другом, матрицы, без которой невозможно построить аналог ассоциативного поля, существующего в языковой интуиции говорящих на русском языке (ср. персептрон, конструируемый в Японии). До тех пор пока не будет такой матрицы, все наши теории сильного управления, валентностей, синтаксических структур, классов слов и пр. и пр.— воображаемое представление языковой деятельности, т. е. то, что видит лингвист, а не то, что есть на самом деле: пренебрегая исключениями, не замечая редких, малочастотных явлений, «экстраполируя» известное на неизвестное, в общем работая при «ручной» обработке текста по методу неполной индукции, мы отстаем от времени и забываем, что появилась возможность формулировать интересующие лингвистов проблемы в виде задач, сбор данных для решения которых можно возложить на ЭВМ и пользоваться при этом методом полной (математической) индукции.

Привлечение ЭВМ к лингвистическим исследованиям — это вопрос не только количества получаемой информации о языке, а качественного скачка в наших знаниях о тех явлениях, судить о которых мы могли при ручной обработке текстов лишь приблизительно. И здесь возможны любые открытия. Они нужны представителям других наук не меньше, чем нам, поскольку их мировоззрение базируется на языковом сознании, а в силу этого — на том, чему учат их языковеды. Так, обследование даже незначительной части русских существительных с помощью ЭВМ на предмет их сочетаемости с предлогами показывает, что в языковом сознании говорящих существует два времени и два пространства: 1) объективной действительности вне человеческого мозга и 2) «субъективной действительности» внутри мозга. Все явления описываются языком по отношению к этим двум системам, и каждое русское существительное как бы помечено размерностью такого сдвоенного пространственно-временного континуума. Это свойство «языковых клеток» отвечает за существование всех важнейших языковых категорий. Оно лежит в основе потенциальных возможностей словосочетаний и предложений к трансформациям, формирования классов слов, эволюции грамматического строя и многих других явлений.

В свете этих проблем $МФ_2$ выступает как институт, где человек формулирует задачи перед ЭВМ, решение этих задач вновь поступает к человеку, на основе этой информации он формулирует новые задачи для ЭВМ и т. д. ЭВМ становится не только средством обработки информации, но и орудием эвристического поиска, инструментом для прогнозирования научных направлений. Возвращаясь к вопросу о том, что включать в МФ, можно сказать: в «Красную книгу» ($МФ_1$) — все, что есть на сегодняшний день лучшего из памятников прошлого и образцов настоящего. С $МФ_2$ дело обстоит сложнее. «Первый» отдел $МФ_2$ — мечта нашего завтра, поэтому главное — не засорять его. Центральным здесь является вопрос о том, какой семантической информацией снабжать слова в будущем словаре МФ — сокровищнице нашего знания о мире, описании языковой интуиции говорящих на русском языке. Гете писал, что в споре рождается не истина, а проблема. Проблема, которую

эксплицировала наша конференция, — это поиск методов экстрагирования из текстов семантической информации, единственно необходимой для описания каждого слова русского языка, т. е. такой, с помощью которой ЭВМ сможет имитировать человека во всех языковых ситуациях. Одним из таких методов является имплицитное определение значения слова путем фиксации его значимости, описание его как пучка отношений со всеми остальными словами, т. е. построение матрицы словосочетательных возможностей всех слов русского языка, о которой шла речь выше. Семантическая информация должна быть не частью исходных данных МФ, а целью его создания, исходными же данными могут являться тексты на русском языке (например, часть текстов из $МФ_1$) и программа их анализа на предмет построения семантического словаря. Чем быстрее мы освоим этап ввода в МФ исходных данных, тем реальнее будет создание не только $МФ_1$, но и $МФ_2$. Однако как бы мы ни стремились вперед, нужны годы. Параллельно можно создавать «третий» отдел $МФ_2$ — собирать библиотеку программ для автоматизации лингвистических работ, объединять эти программы в комплексы, совершенствовать и унифицировать их и т. д. На базе этого отдела будет развиваться «второй» отдел $МФ_2$ в виде центра консолидации разработок по прикладной лингвистике и диспетчерского пункта со сводным каталогом задач, подъязыков и способов их решения в «дочерних» предприятиях. Этот процесс будет происходить, пока «первый» отдел $МФ_2$ не составит упомянутый выше Семантический словарь. После этого диспетчерский пункт сможет прекратить свое существование — его заменит «фабрика» по решению различных языковых задач для нужд потребителя с помощью универсального семантического словаря. Сколь далеки эти перспективы, зависит от того, какими темпами будет осваиваться terra incognita сегодня.

В. И. ФЕДОСИМОВ

Все выступавшие поддерживают идею создания Машинного фонда русского языка, но видят эту проблему каждый по-своему и связывают с ней надежды решить какие-то свои задачи. Поэтому первостепенной необходимостью является сейчас кооперация, координация, консолидация всех научных исследований и практических действий в данной области. Естественно начинать эту работу с целеполагания, с установления той цели, для которой мы хотим объединить результаты совершенно различных терминологических, лексикографических, текстологических работ. Мне хочется предложить такую цель, которая, на мой взгляд, включала бы все упоминавшиеся здесь вопросы автоматизации традиционной лексикографии, получения конкордансов, обратных, частотных словарей и т. д. Все это уже не проблема, и эти задачи успешно решаются в автоматизированных лексикографических центрах. Истинная же цель создания Машинного фонда русского языка состоит в обеспечении современных средств для управления сферой речевой деятельности нашего общества.

В настоящее время, когда под влиянием научно-технического прогресса развитие языка, особенно в лексической сфере, происходит ускоренными темпами, мы уже не можем полагаться на естественный ход этого развития. Поэтому создание Машинного фонда как искусственной среды, в которой русский язык смог бы в возможно большей полноте реализовать свою норму, в

которой можно было бы содействовать ее становлению, распространению и применению как в сфере фиксации, так и в сфере употребления с целью обеспечить единство понимания информации, циркулирующей в обществе, — вот что является главной целью планируемых мероприятий.

Из этой цели должны следовать и вышеперечисленные задачи, и многие другие, и, в частности, задача общественного использования научно-технических достижений одних людей другими людьми в другой области. Информация об этих достижениях должна быть передана от одной сферы деятельности в другую, от одних коллективов другим в такой форме, в которой она была бы доступна для восприятия в каждом месте, где ее использование может оказаться полезным. Однако развитие профессиональных языков, специфических форм представления информации в настоящее время достигло такой степени дивергенции, что ни один человек не может охватить всего многообразия средств и не в состоянии понять, оценить и применить интересующую его информацию из всех смежных областей. Одной из причин такого положения является не только концептуальное различие информации разных источников, но и просто различие в языковых средствах, в частности — лексических. Для того же чтобы организовать такой обмен информацией между различными профессиональными коллективами, необходимо выйти на новый уровень упорядочения, описания и нормализации языковых средств. Здесь возникает потребность и возможность осуществления новых видов лексикографической работы и проведение новых лексикографических принципов.

Для решения задачи обмена научно-технической информацией в нашей стране уже давно начата организация сети автоматизированных информационных центров, в недрах которых уже сейчас скопились богатые машинные лексические массивы, нормализующие язык представления информации. При создании Машинного фонда русского языка эти массивы следует объединить. Вопрос в том, как это сделать. Мне представляется, что мы не можем ставить перед собой задачу единомоментно и в одном месте собрать все относящиеся к русскому языку массивы и данные. Это невозможно не только из-за технологических или организационных ограничений, но просто из-за природы самого объекта — языка, который постоянно развивается как качественно, так и количественно. Поэтому Машинный фонд русского языка должен мыслиться как аппроксимация, как развивающаяся модель нашего обладания нашим языком. Мы сейчас должны построить не результат, а организовать процесс приближения к желаемому результату. А этот процесс уже начался стихийно путем накопления языковых фондов в различных автоматизированных информационных системах. Нужно лишь согласовать эти работы, которые и так уже ведутся в соответствии с планом и производственными заданиями конкретных организаций и ведомств. Поэтому основная трудность, которая нам видится при организации Машинного фонда русского языка, — это обеспечение дисциплины соразработчиков. Теперь, после этой конференции, мы становимся соразработчиками Фонда и должны подчиняться известной технологической дисциплине. Иначе говоря, мы должны соблюдать некоторые общие принципы, правила, ограничения, разрабатывать средства стыковки различных массивов, интерфейсы, обменные форматы, которые бы давали возможность нашим партнерам использовать наши работы сейчас, а нашим наследникам использовать наши работы в будущем. Самая главная проблема в создании таких общих принципов — это создание языка для описания

языковых данных (метаязыка) в автоматизированных системах. В настоящее время отсутствует достаточно полная и непротиворечивая система описания языковых единиц в традиционной лингвистике, а возникновение новых видов языковой деятельности приводит к необходимости инвентаризации таких единиц, которые традиционно не являлись объектом внимания лингвистов. Поэтому создание метаязыка для Фонда также невыполнимо одномоментно. Необходимо организовать сбор, регистрацию и упорядочение всех применяемых способов описания языковых объектов в каком-то едином для Фонда органе.

На основе единого языка описания языковых данных, системы интерфейсов, коммуникативных форматов и других средств связи отдельных лингвистических массивов должна возникнуть распределенная большая система представления языковых данных, характеризующаяся наличием центральных и периферийных органов, подсистемами управления, развития, информационного обслуживания и др.

Строить такую систему следует на организационной основе достаточно высокого уровня. Здесь прозвучало мнение, что Машинный фонд русского языка можно строить на общественных началах. Но я думаю, что только на общественных началах построить Фонд не удастся. Мы здесь собрались на достаточно авторитетной конференции и вполне можем обратиться в ГКНТ, в Президиум Академии наук, в Минвуз с предложением разработать совместную целевую комплексную научно-техническую программу по созданию и развитию Машиного фонда русского языка. В этой программе должна быть предусмотрена организация таких государственных служб, которые занимались бы вопросами регистрации, аттестации и сертификации лингвистических данных и лингвистических процессоров.

Мне еще хотелось бы рассказать о практической деятельности, которая сейчас ведется в области автоматизированных систем научно-технической информации и терминологии и которая представляет определенный интерес в связи с проблемой создания Машинного фонда русского языка. В нашей стране сейчас функционирует примерно 200 автоматизированных систем, в которых ведется поиск научно-технической информации, и для этого в большинстве систем разработано лингвистическое обеспечение, включающее автоматизированные массивы профессиональной лексики. Все эти отраслевые и региональные иформационные центры объединены в Государственную автоматизированную систему научно-технической информации (ГАСНТИ). В рамках ГАСНТИ создан орган по сбору, регистрации и упорядочению языковых средств отдельных информационных центров, который должен обеспечить их совместимость и разработку единых общегосударственных языковых средств представления, поиска и распространения научно-технической информации. Этот орган, называемый Автоматизированной системой ведения информационных языков (АСВИЯ), имеет в своем составе централизованный фонд данных по информационным языкам, функционирующий во ВНИИКИ. В настоящее время здесь собрано около сотни информационно-поисковых тезаурусов, которые являются нормативными словарями описания научно-технических документов, более сорока рубрикаторов научно-технической информации, которые задают лексику описания тематики информационной деятельности, и другие материалы. ВНИИКИ издает ежеквартальный бюллетень АСВИЯ, в котором помещается оперативная информация о деятельности и состоянии системы. На бюллетень могут подписаться все заинтересованные организации.

На машинных носителях информации сейчас в центральном фонде АСВИЯ собран массив более 400 тыс. словарных статей, главным образом информационно-поисковых тезаурусов. Кроме них, в массиве представлен Рубрикатор ГАСНТИ — главный инструмент описания тематики информационной деятельности в стране, Словарь стандартизованной терминологии, общесоюзные классификаторы технико-экономической информации, обеспечивающие общение АСУ. В этом массиве имеются средства автоматической и автоматизированной обработки лексикографических данных, включающие возможность составления различных тематических указателей, конкордансов, анализ пересечения лексических полей, статистический анализ, поиск термина по его лексической окрестности, выявление систем порождающих терминов. Эти возможности существенно расширяют сферу использования такого автоматизированного лексического фонда по сравнению с традиционным словарем.

Важным достижением с точки зрения объединения автоматизированных лексических служб является разработка в процессе создания АСВИЯ лексикографического коммуникативного формата ФОЛИЯ (Формат обмена лексикой информационных языков). Этот формат, основанный на принципах разработки международного лексического формата MATER, был утвержден в 1980 г. ГКНТ, а в настоящее время его более развитая версия проходит стадию утверждения в качестве государственного стандарта СССР и стандарта СЭВ. Этот формат является примером того, как мы практически подходим к построению языка для описания языковых данных — путем фиксации средств представления их при обмене. В то же время средства представления данных внутри каждого фонда, я считаю, не нужно регламентировать, так как выбор их определяется специфическими задачами каждой конкретной лексикографической службы. Аналогичным образом, я думаю, мы со временем создадим единый язык запросов на информацию о лексикографических данных. Тогда в условиях объединения лексикографических центров каналами оперативной связи поиск необходимых сведений можно будет осуществлять так же просто, как сейчас осуществляется поиск абонента телефонной сети. Здесь основная трудность состоит вовсе не в физическом объединении центров линиями связи, а в концептуальных трудностях разработки языка общения автоматизированных центров.

Помимо лексикографического фонда АСВИЯ, во ВНИИКИ функционирует Автоматизированная система информационно-терминологического обслуживания, которая содержит в себе нормативную лексику, установленную отечественными и международными терминологическими стандартами, а именно рекомендуемые и запрещенные термины с их определениями и иноязычными эквивалентами. Система позволяет проводить тематический, классификационный и лексический поиск терминов по большому перечню признаков.

Во ВНИИКИ совместно с рядом всесоюзных информационных центров вводится в эксплуатацию Макротезаурус ГАСНТИ. Это объединенный сводный лексический фонд взаимодействующих центров научно-технической информации, который будет лексической основой для общесистемного языка обмена научными данными, будет представлять собой язык-посредник, позволяющий сопрягать сведения, содержащиеся в различных фондах и записанные в различных формах.

Такая функция — быть языком-посредником при общении различных профессиональных коллективов, по нашему мнению, и является основной зада-

чей будущего Фонда русского языка. При этом посредничество не должно сводиться к пассивному сопоставлению знаков и выражений, а должно быть активным, направленным на состав языковых средств локальных профессиональных коллективов и стимулировать конвергентные процессы их развития, компенсируя нежелательные языковые явления, порождаемые растущей специализацией знания и профессиональной деятельности.

Ю. Н. ФИЛИППОВИЧ

Мы очень заинтересованы в успешном решении хотя бы части задач, поставленных в докладах и выступлениях. Заинтересованность объясняется трудностями, появившимися в процессе создания словарей диалоговой системы, разрабатываемой для решения задач управления в нашей организации. Создаваемый словарь представляет собой двухкомпонентную систему машинных словарей, реализованных в виде независимых локальных баз данных. Одна из баз — это словник, другая — тезаурус. Организация, которая передает нам программное обеспечение компонентов диалоговой системы, не может передать сам словарь по многим причинам. Вот некоторые из них: мы оказались в разных ведомствах, предметная область оказалась различной, кроме этого мы используем различные системы управления базами данных (СУБД). Следует назвать и более глубинные проблемы, которые возникают в связи с нашими трудностями и которые являются наиболее типичными для большинства таких же пользователей: проблема определения временных и ёмкостных параметров технических средств и программных систем, являющихся основой словарных баз данных; проблема автоматизации проектирования словарей в виде баз данных; проблемы первоначальной загрузки, корректировки, обучения, контроля содержания, надежности баз данных систем машинных словарей, разработки средств их восстановления и др.

С сожалением следует отметить, что направления дискуссии не связаны с рассмотрением данных вопросов, а мы предполагали поделиться своим опытом в решении именно этих проблем.

Для решения тех серьезных задач, которые выдвигает дискуссия, необходимо большее единство во взглядах лингвистов, филологов, литературоведов и других специалистов на цели и задачи создания Машинного фонда русского языка в сравнении с тем, которое обнаружилось в докладах и выступлениях. Кроме того, наряду с заинтересованностью в решении конкретных задач должен присутствовать и реализм. К сожалению, этих двух чувств некоторым выступающим не хватало. Идеи и предложения, высказанные в докладах и выступлениях, по нашему мнению, развиваются на трех уровнях.

Первый уровень — это использование ЭВМ в лингвистике, филологии и других смежных с ними направлениях. Я сомневаюсь, что цель столь представительного общества — обсудить возможность и необходимость использования ЭВМ в научных исследованиях. При использовании ЭВМ в различных задачах наиболее острыми являются проблемы подбора соответствующих технических средств и программного обеспечения. Для некоторых из отмеченных в выступлениях задач можно назвать очевидные трудности. Для тех, кто хочет сканировать кириллицу, необходимо отыскать соответствующее техническое средство. Для тех, кто хочет на машинные носители (магнитные ленты и диски) перенести литературные произведения, основной проблемой явля-

ется подбор средств, имеющих удовлетворительные характеристики долговечности, надежности и компактности хранения данных. Имеющиеся сведения о номенклатуре внешних устройств ЭВМ и их характеристиках не позволяют сделать положительного вывода о возможности эффективного решения этих задач. Еще один пример. Сделайте словарь, например, для лексикографических исследований и попытайтесь подарить его коллеге, занимающемуся аналогичной проблемой. Он спросит у вас: «Вы какую СУБД использовали?» К примеру вы ответите: «СУБД ИНЕС». А он ответит: «У нас используется БАНК или ОКА. К сожалению, Баш подарок мне не пригодится». Поэтому о взносах, о которых здесь говорили, нужно подумать, а нужны ли эти взносы, и если нужны, то в какой форме?

Второй уровень — это автоматизация научных исследований и управление процессами в данной отрасли знаний. Опыт разработки и внедрения АСУ различных уровней и типов показывает, что успех во многом связан с подготовленностью объекта к автоматизации. Применительно к сегодняшнему обсуждению предполагается некоторая образованность в этой области и в первую очередь знание типичных ошибок, которые допускаются при решении задач автоматизации: сведение автоматизации сложных процессов к установке ЭВМ или созданию вычислительного центра в этой организации; стремление получить из ЭВМ больше информации, чем туда поместили, и др. Это наиболее часто встречаемые ошибки, о которых многие не вспоминают или даже не знают. Оценивая возможности автоматизированных систем научных исследований, не следует ограничиваться только данными, хранимыми в ЭВМ. Необходимо иметь в виду создание баз знаний на основе систем с элементами искусственного интеллекта, но рассчитывать на новые знания, видимо, не следует.

Корректность в используемой терминологии — один из признаков серьезного и заинтересованного подхода к делу. К сожалению, в некоторых выступлениях мы слышали про «магнитофонные ленты», «машинников» и др.

Как положительное следует отметить некоторые разумные догадки, истинность которых уже доказана практически во многих других областях знаний. Было высказано предложение (оно прозвучало в выступлении киевлян) отдельным пользователям предоставлять не весь Фонд, а только его часть — активный словарь. Данное предложение может быть реализовано в рамках концепции локализации предметной области пользователей и иерархического принципа построения памяти вычислительной системы. В другом выступлении отмечалось, что при создании Фонда русского языка следует сформулировать цели и задачи, качественно отличные от тех, которые ставятся исследователями сегодня. Говорилось о задачах завтрашнего дня. Данную идею содержит принцип создания новых задач при проектировании автоматизированных систем управления и обработки научной информации.

В целом следует отметить, что ряд предложений уже реализован в реально действующих системах и с ними следует ознакомиться.

Третий уровень — это разработка Машинного фонда русского языка в тех аспектах его применения, которые были намечены в основных докладах. Если речь идет о фонде как некоторой базе данных, то необходимо назвать трудности, которые существуют на сегодня для такой реализации. Во-первых, отсутствует опыт разработки и многоаспектной эксплуатации лингвистических баз данных (ЛБД) объемом более 20—30 тыс. терминов. Исключение, по на-

шему мнению, составляет ЛБД систем «ПОЭТ» и «АИСТ» (руководитель работ Э. В. Попов). Во-вторых, лишь немногие СУБД из имеющихся в организациях, занятых реальной обработкой данных, способны эффективно поддерживать большие и сверхбольшие базы данных. В-третьих, разработка и эксплуатация базы данных Машинного фонда русского языка потребовала бы решения не только сложных технических проблем, но и создания соответствующей организации, ведущей эту разработку и эксплуатацию, а также подготовки соответствующих специалистов в вузах.

Если Фонд реализовать на основе файловых систем, то к перечисленным трудностям в соответствующей интерпретации следует добавить явное несоответствие целей использования Фонда многими различными пользователями в отличных режимах и эффективности реализации их трансакций. Однако наибольший опыт имеется именно здесь. Например, машинная реализация частотного словаря семантических множителей русского языка (руководитель работ Ю. Н. Караулов), машинный словарь системы «ПСИХЕЯ» (руководитель работ Ф. И. Рыбаков) и др.

Таким образом, необходимо более четко определить, кто за что здесь фактически выступает. Кто выступает за ЭВМ, кто — за автоматизацию, а кто — за Фонд. Если это не выяснить, то трудно не только обсуждать предложения, но и решать практические задачи. В рамках данного непротиворечивого, подчеркиваю, размежевания и следует формировать силы.

Так как проблем было поднято очень много и в ближайшее время их все не решить, то полезно создать ряд комиссий. Одну из них можно направить на изучение и координацию предложений по разработке и реализации технической концепции Машинного фонда русского языка. Данный вопрос является первостепенным и наиболее важным. Трехдневный запас энтузиазма, которым каждый из нас обогатился, быстро рассеется, когда системотехники, программисты и др. не смогут быстро и эффективно реализовать поставленные задачи. На первое время, пока нет комплексной программы, необходим заказчик со средствами. Таким заказчиком является промышленность. Необходимо из всего комплекса изменяющихся и по-разному актуализирующихся целей выделить несколько, реальных для сегодняшнего дня, но тех, которые могут заинтересовать промышленные организации. Мы уверены, идеи создания Машинного фонда русского языка будут воплощены на практике, если союзником лингвистов и филологов будут представители промышленности.

А. Я. ШАЙКЕВИЧ

Хотелось бы сказать несколько слов о некоторых общих чертах Машинного фонда. Ю. Н. Караулов построил свое выступление как некоторый набор альтернатив, которые диалектически увязывались одна с другой. Но первая возможность, первая альтернатива была сформулирована так: что это должно быть — научное целое со своим научным лицом или это только склад? Так, в заостренной форме, был поставлен вопрос. Склад, чисто информационная система, или это поле деятельности для лингвистов, для исследователей. Ю. Н. Караулов здесь, уже диалектически не объединяя две стороны дела, вполне высказался за научный вариант. Это должно быть нечто научное, что позволяет получать новые знания и содержит в себе, в частности, некоторые

алгоритмы научно-исследовательского характера. На этом мне хочется остановиться.

Мне кажется, что если Машинный фонд будет создан даже как инвентарь, как хранилище, то научно-исследовательские процедуры неизбежно за ним последуют. Их можно даже не планировать. Как только у вас есть возможность в машине оперировать с чем-то, то хотите вы этого или нет, вы как исследователь тоже начинаете эти процедуры разрабатывать и применять. Казалось бы, можно так рассуждать: тем лучше, давайте заранее и введем эти некоторые научные процедуры. Но я боюсь, что введение некоторых научных процедур как элемента, равноправного инвентарю, нас может заранее сориентировать не в том направлении, не даст нам возможности потом выбирать разные пути. Опасность в том, что затушуется различие между бесспорным материалом и результатами какого-то исследования, придуманного конкретным лингвистом, который и несет ответственность за то, что получается. Вводить такие результаты в систему наравне с текстами и со словарями, по-моему, неверно. Здесь мы имеем дело с каким-то косвенным знанием о речи и о человеке, и вводить эти вещи должны конкретные исследователи в ходе своего исследования на базе системы. Когда они будут работать с Машинным фондом, пусть они там что угодно и вводят. Можно было бы в машине делать учет заказов на подобного рода исследования, если окажется, что такой-то алгоритм популярен, один за другим исследователи его повторяют, ну, тогда, понятно, его хорошо бы ввести в машину и там его держать как постоянный рабочий материал.

Есть разная степень достоверности материалов, вводимых в Машинный фонд. Мне тоже, как и С. И. Гиндину, кажется, что тексты — то единственное первичное, достоверное, фактическое в языке, что вводится в машину. Словарная традиция надстроена над этими текстами, а кроме того, она есть продукт собственного развития лексикографической деятельности как особого вида деятельности вообще, обусловленный не только самим языком, которому посвящена данная лексикография, но и какими-то филологическими традициями, например, обычаем давать определения в словаре. Это есть некоторый исторически конвенциональный элемент. Мне кажется, что запас, созданный лексикографией прошлого, частично должен быть введен в Машинный фонд. Здесь, конечно, надо вводить не только то, что создается сейчас, но, безусловно, надо ввести, скажем, шахматовский словарь (хотя он не завершен) и бодуэновское издание словаря Даля. Я здесь никакого конфликта с прошлым не вижу и не заметил его в сегодняшних докладах. Но все-таки при этом нужно думать об одной проблеме, которая возникает внутри Машинного фонда и которая почти не затрагивалась сегодня. Вопрос о взаимодействии лексикографической и текстовой части. Прозвучала мысль, что текстовая часть служит иллюстрацией для лексикографии, а между тем небезынтересен вопрос для самой лексикографии: насколько гармоничны словари с текстами. Что я имею в виду? Представим себе, что я решил в качестве толкования животных в сказках (например, в русских сказках) давать их зоологические номенклатурные наименования. Это нелепо совершенно, и в рамках не зоологических текстов, а текстов фольклорных, совершенно искусственно и ненужно. Ясно, что в данном случае метаязык лексикографического описания явно не гармонировал бы с текстом. Возьмем теперь не такую искусственную ситуацию, а более обычную. Предположим, у нас есть несколько словарей.

Мы их завели в машину. И вот оказывается (очень обычная ситуация), что для одного и того же слова разные словари дают разный перечень значений, разное их число, разные формулировки. Можем ли мы сказать, что какой-то из этих словарей гармоничнее соотносится с текстами, чем другой? Думаю, что в машине, в частности, можно предусмотреть пути решения этого вопроса. Если я имею заданную систему словарных значений и имею полный на основе данной выборки перечень всех контекстов употребления слова, я начинаю последовательно все эти контексты проецировать на словарные толкования, и один словарь мне дает полную расписку — такую, что я в каждом отдельном случае могу сказать: «Ага, это значение № 1, а это значение № 2, а это — № 3». А другой словарь меня через случай употребления ставит в тупик. Я затрудняюсь, не знаю, какому значению приписать данный контекст. Вот тогда я скажу: первый словарь хорошо, гармонично с данными текстами распределил полисемию, а второй — неудачно. Я думаю, что в самой машине возникнут большие исследовательские возможности для выработки методов самой лексикографии. Если взять конкретный пример, по-моему, вывод о том, что «пенсне» и «пловец» находятся в одном поле, это плохой вывод, он не гармонирует с тем, что я как носитель языка сам произвожу и слышу от других. Я могу с уверенностью сказать, что в данном случае алгоритм обработки лексикографического материала меня привел к выводу, который не гармонирует с речевым опытом. Почему — это другой вопрос, это уже тема отдельного обсуждения. Я думаю, что в Машинном фонде, в машинной субстанции, что ли, эти две основные части — текст и словарь — найдут со временем гармоничное сосуществование.

Мне кажется, что всем нам, лингвистам, занимающимся машинами (если Машинный фонд осуществится), нужно психологически перестроиться в одном отношении. Нам всем нужно привыкнуть самим вводить наш материал. Перестать смотреть на свой труд как на труд мыслителя исключительно и постараться работать и руками тоже. Скажем, по моим расчетам, если вы два часа в день будете работать в этом качестве, то за два года вы введете все тексты Пушкина: во-первых, доказано, что специалист, работающий для себя, вводит этот текст несравненно лучше того лаборанта, который делает это за деньги, — раз, и, кроме того, лаборантов достать очень трудно, а исследователи — мы все с вами. Пушкин сто́ит того, чтобы потратить два года по два часа в день на ввод его в машину. Жертва очень маленькая. Не будут платить — неважно. Если нельзя в рабочее время — нужно вводить после работы бесплатно.

Теперь я хотел бы вернуться к некоторым идеям, которые высказывались ранее и получили развитие в выступлении А. С. Герда, но все-таки еще требуют некоторых уточнений. Я снова хотел бы вернуться к проблеме соотношения, с одной стороны, текстов и словарей, а с другой стороны— разного рода текстов, которые вводятся в машину. Причем соотношение это должно варьировать в зависимости от жанров и от характера видов речевой деятельности. Начну с последнего, с вопроса о том, что именно надо вводить в машину. В. Ю. Розенцвейг очень красноречиво рассказал о зарубежном опыте, в частности об опыте французского лексикографического центра в Нанси, который материализовался в виде «Trésor de la langue française». Действительно, для всех людей, любящих словари, независимо от их специальности, филолог или не филолог, этот словарь — просто интеллектуальное «объ-

единение». К сожалению, в XX в. редки словари (а они были довольно часты в XVIII в. и «Trésor de la langue française» к этому типу относится), которые можно читать для своего удовольствия, вечером сидеть и просто вместо того, чтобы смотреть телевизор, читать словарь подряд. Этот словарь относится к подобному типу. Но, перенимая лучшие черты этого французского опыта, мы должны избежать некоторых недостатков. При всем моем восхищении этим словарем, я вижу в замысле его некоторые неучтенные возможности. Начну с характера вводимых текстов. Я сейчас буду говорить не о деловых и прикладных текстах, которые там представлены в изобилии, а о текстах, традиционно относимых к художественной литературе. 70 млн. слов текста, которые были введены в машину и составили художественную часть выборки, представляют произведения тысячи авторов, причем в соответствии с французской традицией это были авторы первого ранга. По-видимому, там была какая-то внутренняя анкета, жюри одобрило таких-то авторов, сочло их достойными включения в словарь и игнорировало остальных. Эта традиция характерна как для французской, так и для русской лексикографии. Я думаю, что среди текстов художественной литературы, вводимых в машину, должно быть два рода (это отмечал С. И. Гиндин, и я это тоже повторяю). С одной стороны, бесспорные тексты такого типа, которые все мы считаем элементом истории культуры данного народа и которые должны быть в идеале, если не сразу, то через некоторое время, введены полностью. Ну, скажем, какой-то круг авторов, который, между прочим, обычно и фигурирует в цитатах-иллюстрациях словарей. Какой-то круг таких авторов есть и за пределами художественной литературы. Конечно, значение работ В. И. Ленина таково, что ленинские тексты относятся к этому разряду[6]. Так вот, с одной стороны, должны быть введены подобные тексты, а с другой стороны, необходимо иметь очень дифференцированную и очень широкую выборку «средних», которые составляли бы некий средний лингвистический фон и которые далеко не всегда попадали бы в примеры печатного словаря, но которые участвовали бы в создании Машинного фонда. Этот средний фон чрезвычайно существен. Только на его основе мы могли бы получить стандартные стратифицированные стилистические узкожанровые выборки, но на более широкой и более дифференцированной основе. Они и дали бы филологам возможность оценивать те вершины литературы, которые обычно приковывают к себе внимание. Без этого понять роль и семантику индивидуальных авторов часто очень трудно. Это второй пласт. Если перенестись в область деловой прозы, то доля этой второй, собственно выборочной, статистически оправданной части, будет неизмеримо больше, так как в деловой прозе у нас останется мало текстов культурно-эпохальных, как элементы культуры безусловно и навеки сохраненных, но зато у нас будет множество небольших выборок, даже не текстов, а небольших кусочков по 100—200 слов текста, которые в совокупности намного лучше дадут возможность представить соответствующий жанр.

---

[6] Замечу, кстати, одно удивительное обстоятельство: почему, при том, что в г. Горьком был на машине получен конкорданс по «Материализму и эмпириокритицизму», где для каждого словоупотребления давалось целое предложение, этот опыт был предан забвению, а в Институте русского языка, как я понял, пошли по пути ручного расписывания, явно устаревшего в наше время.

Итак, мы можем вводить полные тексты целиком, мы можем вводить выборочно громадное число текстов, представленных очень небольшими кусочками (это последнее статистически обосновано), но остается и третья возможность, которая просто совпадает с тем, что делали лексикографы и в прошлом. А именно, полагаясь на указанные выборки, представляя себе состав текстов, уже введенных в машину, лексикограф как бы по старинке отбирает интересные примеры из остальной литературы и вводит их в машину. Дело в том, что как бы мы ни увеличивали выборку (даже до многих сотен миллионов словоупотреблений), там будут отсутствовать очень многие слова, которые представлены в словарях. Мне кажется, что этот третий источник, чисто традиционный по технике подготовки, будет очень важным, без этого в наших словарях оставались бы колоссальные пробелы. Я не знаю, входит ли в состав ленинградской семимиллионной картотеки, и если входит, то в какой степени, например, анализ наших энциклопедий начиная от Брокгауза и Ефрона и кончая современными советскими. Но во всяком случае, когда я сейчас открываю энциклопедию Брокгауза, я нахожу там множество слов, которых в опубликованных словарях я не нахожу. Некоторые ушли из языка, некоторые сохраняются, но в опубликованных словарях их нет. Смысл того, о чем я говорю, заключается не в том, чтобы целиком ввести Брокгауза и Ефрона, а как раз такого рода литературу ввести выборочно, немного, а остальное доделать человеческими руками для того, чтобы с контекстами эти слова попадали в Машинный фонд.

В связи с последним я касаюсь области, уже связанной не с художественной литературой, а с терминологией. Терминология постоянно пополняется, но проходит довольно большой срок, прежде чем новые термины оседают в словарях. Необходима особая служба слежения за новыми терминами. Поскольку наука интернациональна, эта служба слежения не может ограничиваться одним языком. Фиксируя, например, появление нового английского термина, мы можем заранее думать о его русском эквиваленте, забегая вперед, как бы прогнозировать развитие русской терминологии. Здесь задачи Машинного фонда смыкаются с задачами центров научной информации.

С. Д. ШЕЛОВ

Задача создания Машинного фонда русского языка многоаспектна и многогранна. На конференции обсуждалась сама задача как таковая, ставился вопрос о том, какие рамки поставить, какую последовательность этапов решения задачи необходимо наметить (Р. Г. Котов, Ю. В. Рождественский, А. С. Герд). По-видимому, устанавливается и определенная стратификация задач и тех пользователей, которым Машинный фонд русского языка будет необходим. В частности, наметилось такое деление: во-первых, это — широкая аудитория лиц, для которых характерна профессиональная ориентация на слово, я бы так сказал. Это — работники издательств, редакторы, журналисты, работники информационных служб (от себя добавлю: и библиотечных, ибо разработка любого каталога — это в том числе и работа со словом, будь то предметный, систематический или какой-нибудь другой каталог). Это, конечно, — преподаватели и переводчики. Во-вторых, это — сами языковеды. По-видимому, первоначальными являются задачи, связанные с обслуживанием первой категории лиц, а задачи исследовательские (обеспечение специаль-

ной информацией лингвистов) отодвигаются как будто на второй план. Конечно, само это противопоставление очень относительно, потому что лингвистам нужны и «обычные» справочники о языке, как и рядовым носителям языка.

Очевидно, что столь обширный проект, как создание Машинного фонда русского языка, не следует мыслить как одноразовое объявление (в близкой или отдаленной перспективе) о том, что создан Машинный фонд и что теперь им можно пользоваться, как мы пользуемся предметом домашнего обихода. От Машинного фонда следует и должно ожидать практическую выгоду уже в ходе его создания, а его ведение должно быть постоянным и непрерывным процессом. Какой может быть практический выход от Машинного фонда уже на этапе его создания? Я думаю, первоначально это могут быть списки слов или словосочетаний, это могут быть простые словари и, наконец, это может быть использование Фонда как «адресной книги», когда Фонд отсылает к тому или иному источнику, где можно получить необходимую справку.

Особый интерес в этом смысле представляет связь сегодняшней тематики с задачами терминологии. Не случайно о терминах говорилось больше, чем о любых других пластах лексики. Это, на мой взгляд, объясняется двумя обстоятельствами.

Во-первых, привлечение терминологического пласта лексики к разработке Машинного фонда русского языка определяется самой природой термина. От термина требуется если не однозначность, то по крайней мере ясность. Здесь, думается, существует единство мнений. Я далек от мысли, что проблематика научно-технического документа исчерпывается терминологической проблематикой. Безусловно, нет. В частности, мне очень близка мысль о том, что в научно-технических документах существуют регулярные лексико-семантические сдвиги, которые нельзя отнести к терминологическим, но которые тем не менее очень интересны и важны. Простой пример. Если у вас в руках книга на археологическую тему, то скорее всего, что фраза «амфоры такой формы известны также и в Западной Грузии» должна пониматься не таким образом, что в Западной Грузии знают о таких амфорах, а таким образом, что эти амфоры были найдены в Западной Грузии. Слово «известен» в текстах археологической тематики используется именно так очень часто. Случаи такого использования лексики невозможно расценить как терминологические (при всей интуитивности понятия «термин»). Но если говорить о терминологии в том общем понимании, которое имеется у присутствующих, то требование ясности и определенности значения термина, по-видимому, справедливо.

Во-вторых, привлечение терминологического пласта лексики к разработке Машинного фонда объясняется известным «организационным» (Р. Г. Котов) аспектом. Действительно, терминология лучше других слоев лексики подготовлена к этой работе. Существует много общегосударственных задач (они перечислялись Р. Г. Котовым, и нет необходимости их повторять), при решении которых терминология, выражаясь канцеляризмами, так или иначе «задействована» и которые уже вызвали к жизни существование соответствующих машинных фондов.

Таковы причины важности терминологической проблематики в обсуждаемом здесь проекте. Какова, однако, традиция терминологических исследований и чем может быть Машинный фонд ей полезен?

Я бы выделил три группы специалистов по терминологии и их исследований. Первая — это языковеды, которые занимаются специальной лексикой, интересуясь главным образом ею как частью общелитературного языка. Сюда входит определение этимологического состава терминологии, использование терминологии в языке художественной литературы и т. п. Вторая группа специалистов традиционно называется терминологами. Различия между первой и второй группой очень зыбкие, но, наверное, условно эту границу следует провести так. Если специалист в своей работе использует достаточно глубокое знание предметной области, то он становится терминологом, а если он ограничивается семантикой общеязыковой, то это просто лингвист, изучающий терминологию. Терминологи в свою очередь выполняют два вида работ. Один из них я бы назвал описательным, традиционным. Второй можно было бы назвать прескриптивным. Это предписывающий вид работ, т. е. создание в известной мере нормативных документов; нормативность может быть разной — от сборников рекомендуемых терминов до ГОСТов, но работа должна обязательно заканчиваться материалами такого рода. Наконец, существует и третья группа исследований. Вопросы терминологии изучаются в философских работах, где они традиционно входят в раздел «Логика и методология науки». Казалось бы, что, исследуя один и тот же объект, эти три группы исследований должны тесно взаимодействовать. Каково же реальное положение? По нашим наблюдениям, взаимодействие здесь слабое, и это, безусловно, приводит к отрицательным результатам. Назову лишь некоторые нерешенные научные проблемы.

Так, отсутствует какая бы то ни было логико-семантическая классификация терминов, кроме той, о которой упомянул Ю. В. Рождественский, а именно классификации по тематическому признаку: терминология биологии, терминология электротехники, терминология лингвистики и т. п. Такое членение специальной лексики, конечно, существует, но какой-либо внутренней классификации, к сожалению, нет. Правда, существуют два-три термина типа «номенклатура» или «профессиональная лексика», которые намечают некоторое членение специальной лексики, но единства в понимании этих терминов (а следовательно, и самого членения терминологии) нет.

Другой вопрос. Терминологией иногда называют классы специальной лексики, связанные устойчивыми ассоциациями с техническими и теоретическими дисциплинами: терминология ядерной физики, терминология электроники и т. п. В то же время существует традиция связывать терминологию со специальными пластами лексики, которые как будто лежат в совершенно иной плоскости. Так, говорят: «терминология родства», «терминология свадебного обряда», «терминология пчеловодства» и т. п. Спрашивается, в какой мере в этих двух случаях речь идет об одном и том же объекте, называемом «терминология»?

Многие вопросы не решены и для нормативной, прескрептивной терминологической работы. Ряд недостатков был отмечен в недавно вышедшем сборнике Института русского языка АН СССР о языке терминологических стандартов. Имеются и организационные недочеты, которые были показаны Горьковской группой исследователей терминологии и которые состоят в том, что стандарты используются недостаточно.

И вот здесь, представляется, создание Машинного фонда было бы чрезвычайно важным и полезным как звено, которое организует все эти области

в смысле и чисто научном, и организационном. Представляется перспективной картина согласования той лексики, которая традиционно толкуется в словарях как специальная (ее процент не так уж мал), с теми нормативными документами, которые регламентируют ее употребление (сборники рекомендуемых терминов, ГОСТы и т. п.). Таким образом, для той части лексики, которая стандартизована или содержится в каких-то нормативных списках в Машинном фонде, будет ссылка на соответствующие документы. Следовательно, нет необходимости все терминологические стандарты и все нормативные тексты вводить в память машины. Гораздо проще иметь просто ссылку на соответствующий источник, что тем более целесообразно ввиду того, что машинные фонды нормативных терминологических документов уже создаются в некоторых организациях (например, во ВНИИКИ). Помимо известной популяризации нормативных документов в области специальной лексики, это позволит Машинному фонду русского языка служить важным связующим звеном между общеязыковой лексикой и специальной терминологией.

Именно так можно охарактеризовать ход обсуждения, конкретность предложений и выводов.

Идея создания Машинного фонда русского языка получила горячую поддержку. Если и были сомнения, споры, то они касались объемов и сроков работ, ранжирования задач и целей по степени важности и реализуемости, организационных вопросов и вопросов технического оснащения.

Если стремиться увязать в одном проекте все внесенные предложения, то мы действительно приходим к самому общему представлению о Машинном фонде как о системе комплексной автоматизации лингвистических исследований и прикладных разработок. В слове комплексный заключены и смысл, и цель, и характер автоматизации.

Автоматизация исследований и разработок в языкознании, как и повсюду, есть средство повышения производительности труда и качества продукции. Потребности сегодняшнего дня таковы, что русистика не может решать прикладные задачи в области проектирования лингвистического обеспечения информационных систем, если ее исследовательская база (лингвистические источники) и сам процесс проектирования не будут автоматизированы. Поэтому выдвинутое противопоставление «лингвистика помогает технике или техника работает на лингвистику» не кажется нам эвристически полезным. «Техника» не может работать на лингвистику сама по себе — только языковеды, вооруженные техникой и соответствующими методами работы, могут решать свои исследовательские задачи в соответствии с требованиями ускоренного научно-технического развития нашей страны. Мы вовсе не утверждаем, что все лингвистические задачи требуют для своего решения технического оснащения. Мы утверждаем лишь, что для любой лингвистической задачи автоматизированное информационное обеспечение полезно, а в прикладных разработках, в лексикографии оно теперь уже строго необходимо. При достигнутом языкознанием уровне детальности и систематичности в учете многообразных взаимодействующих лингвистических явлений и при современных требованиях к объемам привлекаемого материала автоматизация информационной базы и алгоритмизация методов — единственный путь построения конструируемых объектов. Среди таких задач на пер-

вое место, несомненно, выдвигаются задачи создания и оперативного пополнения словарей действующих информационных систем в различных отраслях науки и промышленности.

Поскольку автоматизация — средство, а не метод, никакой опасности «истолкования метода в духе операционализма» не существует, но действительно существует опасность отрыва русистики от задач современной лингвистической практики.

Материалы сборника подтверждают ту аксиому информатики, что автоматизация только тогда дает ожидаемый эффект, когда она комплексная. Поэтому автоматизация только введения текстовых источников (по существу — создание новых словарных картотек) не даст нужного результата, если не будут параллельно решены вопросы создания словарных баз, согласования словарной информации с грамматической, не будут решаться вопросы систематического сбора языковой статистики, не будут построены системы информационного обеспечения разработок. С другой стороны, призыв сузить задачу Машинного фонда русского языка до создания автоматического терминологического словаря русского языка без решения вопросов его текстового и информационного обеспечения так же неприемлем, как и призыв ограничиться исследовательскими задачами, не решая задач производства лингвистической продукции, или наоборот.

Автоматизация затрагивает прежде всего словарное дело. Это естественно: именно в словарном деле мы имеем сейчас и опыт, и образцы м а с с о в о г о применения вычислительной техники и комплексности ее использования — от сбора источников до получения печатного продукта.

Но и в автоматической лексикографии сейчас ощущаются трудности, связанные с необходимостью использовать языковые процессоры для автоматической обработки входных данных и анализа запросов к словарным базам. С другой стороны, использование автоматизированных систем редактирования и набора также требует применения языковых процессоров для полноценной работы с текстом.

Автоматизируя лишь одни аспекты лингвистической деятельности и оставляя в стороне другие, мы рисковали бы надолго закрепить вре́менные, промежуточные результаты и приемы исследовательской и конструкторской работы в качестве постоянных компонентов основной деятельности. Так автоматизация только введения текстовых данных грозит закрепить частотные и обратные словари, словоуказатели и конкордансы в качестве единственного и конечного результата вычислительной лексикографии. Создание автоматических словарей, не оснащаемых настоящими языковыми процессорами, ограничивает возможности их развития в полноценный эквивалент больших национальных словарей. Поэтому призыв ограничить Машинный фонд русского языка одним автоматическим терминологическим словарем эквивалентен призыву создать большой, но лингвистически недостаточно обоснованный и оснащенный словарь.

Одна из несомненных тенденций научно-технического развития состоит во все большем использовании средств вычислительной техники в качестве *изготовителя* и *хранителя текстов*. Следствием этого может явиться то, что через некоторое время бо́льшая часть источников для изучения современного русского языка будет существовать только в машинной форме, и лингвисты, не вооруженные вычислительной техникой, просто не смогут их изучать. Более того, в связи с созданием средств автоматизации редактирования и набора в издательской практике все большее применение будет находить прием в издательство рукописей, выполненных на машиночитаемых носителях. Это означает, что такие лингвистические труды, как крупные словари (с усложненным набором) или диалектологические атласы, созданные вручную, смогут быть изданы только ценой больших усилий и затрат. Следовательно, такие труды с самого начала должны делаться с помощью ЭВМ и представляться в издательства в компьютерной форме.

Справедливо обращалось внимание на то, что чем ближе к нам изучаемый период развития языка, тем более высокие требования он предъявляет к учету языкового материала и качеству его исследования. Приведенные П. Н. Денисовым простые расчеты убеждают нас в том, что адекватное представление о современном русском языке, языке конца XX в., может дать картотека, насчитывающая порядка 100 млн. цитат, и словник объемом 300—400 тыс. слов. Но это — такой объем, который обработать ручными методами уже невозможно за время, достаточно близкое к характеризуемому словарем периоду существования языка. Конечно, правы и те, кто считает, что для современной лексикографии метод частичной выборки, смещающий статистику употребления и дающий неполное представление об авторских и стилевых особенностях, уже недостаточен. Следовательно, должны быть заложены новые картотеки, соответствующие возросшим требованиям к используемым лингвистическим источникам как по качеству материала, так и по качеству аппарата, обслуживающего картотеку.

В то же время можно и нужно повысить эффективность использования и качество старых картотек. А. С. Нариньяни обратил внимание на целесообразность использования фотоносителей. Действительно, современные средства микрографии, соединенные с вычислительными средствами, способны уместить многомиллионную картотеку Института русского языка АН СССР на нескольких десятках тысяч микрофиш и сделать ее доступной с читально-копировальных устройств, работающих под управлением ЭВМ. Задача создания таких автоматизированных картотек на базе существующих реальна уже сейчас, и ее следует рассматривать в качестве одного из мероприятий комплексной автоматизации лингвистических исследований.

В понятие комплексности автоматизации входит также автоматизация исторических и диалектологических, социо- и психо-

лингвистических исследований, исследований разговорной речи. Диалектологический и устный материал уникален в каждый момент времени, и потеря данных сегодняшнего дня невосполнима для будущих поколений. Много говорилось о культурном значении Фонда. Мы не можем оставить нашим потомкам смещенное представление о нынешней лингвистической культуре и языковом состоянии: сочетание синхронии и диахронии, комплексное и системное изучение всех форм существования языка должно быть сохранено и в компьютизированном лингвистическом источниковедении, и в исследовательской практике.

Вот почему мы — за комплексную автоматизацию, за Машинный фонд русского языка, включающий в себя и «культурные», и «прикладные» фонды.

Как же в таком случае совместить эту тотальность подхода с реальностями сегодняшнего дня? Конечно, правы те, кто опасается недостатка в силах и средствах, в организационных и координационных усилиях, кто опасается нарушения целостности, совместимости и сохранности данных в компьютерной среде, кто видит и другие многочисленные трудности на пути к цели. Ответы на эти вопросы могут быть только самыми общими. Прежде всего, конечно, определенную гарантию дисциплины создания и использования Машинного фонда даст научно-техническая программа, в которой сформулированы проектные задания и сроки их исполнения по основным компонентам Фонда. Далее, определенную роль должны сыграть закрепленные в государственных стандартах и методиках содержательные требования к исполнению проектов, уточненные применительно к Машинному фонду русского языка головной организацией-исполнителем. Такой организацией, по общему мнению, должен стать Институт русского языка АН СССР.

Само собой разумеется, что Машинный фонд русского языка станет реальностью, если его создание будет соответствующим образом финансировано. Важным условием финансирования является более тесное сотрудничество лингвистических коллективов и организаций с промышленными ведомствами, прямое выполнение их заказов.

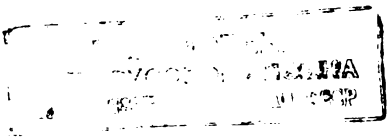
Исходя из того, что в идее Машинного фонда русского языка сконцентрировалось большое количество различных целей и задач, совместное и параллельное решение которых и создает комплексную автоматизацию лингвистических исследований, конструктивно такой Фонд не может быть иным, как совокупностью локальных баз данных, перерастающей со временем в распределенный лингвистический банк в сети ЭВМ. Возникает вопрос, необходимо ли стремиться к полной информационной совместимости этих баз? Наш ответ: практически это невозможно, да и теоретически не нужно. В самом деле: мы собираемся строить лингвистическую систему, а раз так, значит, со временем каждая из подсистем должна обладать своим языковым процессором. Следовательно, каждая подсистема (база данных)

может рассматривать другую базу данных как текст, поданный на вход ее языковому процессору, который перерабатывает ее содержимое в представление своей базы. Теперь вопрос состоит в том, как построить такой процессор, точнее, серию последовательно улучшающихся процессоров такого типа. Таким образом, мы сформулировали вопрос, который, возможно, послужит одной из тем следующих симпозиумов и конференций по проблемам создания Машинного фонда русского языка.

Решение этого вопроса снимет и другую часть опасений, связанных с изменчивостью техники и средств программирования. Говоря лингвистическим языком, в Машинном фонде русского языка должна быть обеспечена лингвистическая непрерывность диахронии и синхронии, их взаимодействие за счет естественной внутренней переводимости, а не за счет искусственной консервации или декретирования.

Наконец, необходимо остановиться на вопросе, волнующем большинство сочувствующих идеям Машинного фонда языковедов: как быть с вводом данных? Скажем прямо, надежды лингвистов на появление в скором будущем надежных и универсальных оптических читающих устройств не кажутся нам оправданными. Более того, исходя из тенденций научно-технического развития, можно ожидать, что, поскольку бóльшая часть информации, циркулирующей в обществе, будет иметь компьютерную форму уже в момент ее возникновения, необходимость в таких устройствах будет становиться все меньше. Конечно, какая-то часть информации, занесенной на бумажные носители стандартизованными шрифтами, может быть введена в память ЭВМ с помощью оптических читающих устройств, но реальная доля такой информации в актуальных исследованиях не слишком велика. Поэтому проблему ввода надо решать, одновременно используя все возможности, но основным вкладом должен стать только вклад самих лингвистов-исследователей и разработчиков, сделавших ЭВМ инструментом своей повседневной работы.

Редколлегия



ИРЯ
институт русского языка
им. в. в. виноградова

СОДЕРЖАНИЕ

Предисловие . 3

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ПОСТАНОВКА ЗАДАЧИ 7

А. П. Ершов. Машинный фонд русского языка: внешняя постановка . . 7

Ю. Н. Караулов. Методология лингвистического исследования и Машинный фонд русского языка 13

В. М. Андрющенко. Концепция и архитектура Машинного фонда русского языка 26

Б. Ю. Городецкий. К созданию Машинного фонда русского языка (определение, применения, актуальные проблемы) 44

Р. П. Рогожникова. Машинный фонд русского языка и словарное дело . 58

А. С. Герд. Типы русских текстов и организация Машинного фонда русского языка 67

В. Ю. Розенцвейг. Опыт создания национальных лексикографических служб за рубежом 75

Р. Г. Котов. Машинный фонд русского языка и автоматизация словарно-терминологического обеспечения информационной службы 84

Ю. Д. Апресян. Некоторые следствия из единой модели языка для машинных фондов 90

Н. Н. Леонтьева. Об информационной системе словарей Машинного фонда русского языка 109

И. А. Большаков. Составляющие и принципы формирования программного обеспечения для Машинного фонда русского языка 125

Л. П. Жуковская. Изучение древних рукописей и Машинный фонд русского языка 132

А. П. Журавлев. Ореольная семантика в Машинном фонде русского языка 136

Ю. Г. Овсиенко. Современная разговорная речь как составная часть Машинного фонда русского языка 139

Ю. А. Тулдава. О машинных фондах русского языка за рубежом . . 141

С. В. Базарнова, Р. Р. Мдивани. Терминологическая база данных Автоматизированной информационной системы по общественным наукам ИНИОН АН СССР 142

Р. Г. Пиотровский. Психологические, теоретические, организационные и технические аспекты создания Машинного фонда русского языка . . 146

И. П. Белецкая. О создании банка данных для синтаксиса русского языка 149

Ю. В. Ванников. О некоторых лингвистических аспектах проблемы создания Машинного фонда русского языка 151

www.ruslang.ru

ДИСКУССИЯ О МАШИННОМ ФОНДЕ РУССКОГО ЯЗЫКА 155

От редакторов 155

Выступления участников дискуссии: А. Б. Антопольский (155), А. С. Асиновский (157), К. Б. Бектаев (158), Г. Г. Белоногов (162), В. Н. Белоозеров (163), Л. Н. Беляева (165), Л. В. Бондарко (168), В. В. Бородин (169), Е. М. Верещагин (170), И. Н. Волкова (172), А. С. Герд (175), С. И. Гиндин (176), В. П. Григорьев (179), М. П. Данилов (181), П. Н. Денисов (182), А. В. Зубов (184), Т. Ф. Иванова (186), Н. П. Козленко (189), Н. З. Котелова (189), В. М. Лейчик (194), Е. Е. Ловцкий (195), В. В. Мартынов (196), В. И. Назаров (198), А. С. Нариньяни (200), С. Е. Никитина (202), Ю. К. Орлов (204), В. И. Перебейнос (206), М. М. Пещак (208), А. А. Поликарпов (211), Н. Н. Пшеничнова (213), М. Н. Реммель (215), В. И. Рубашкин (216), Л. З. Сова (217), В. И. Федосимов (220), Ю. Н. Филиппович (224), А. Я. Шайкевич (226), С. Д. Шелов (230)

Машинный фонд русского языка: от идеи к проекту . 234

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА:

идеи и суждения

Утверждено к печати Институтом русского языка АН СССР

Редактор издательства Л. М. Гаврилова.
Художник Ф. Н. Буданов. Художественный редактор Г. П. Валлас.
Технические редакторы И. Н. Жмуркина, Н. П. Кузнецова. Корректор М. Ф. Сафонова

ИБ № 32162

Сдано в набор 16.04.86. Подписано к печати 30.07.86
Формат 60×90¹/₁₆. Бумага книжно-журнальная.
Гарнитура литературная. Печать высокая. Усл. печ. л. 15. Усл. кр. отт. 15,25.
Уч.-изд. л. 17,8. Тираж 1900 экз. Тип. зак. 4836. Цена 2 р. 80 к.

Ордена Трудового Красного Знамени издательство «Наука»
117864, ГСП-7, Москва, В-485, Профсоюзная ул., 90
2-я типография издательства «Наука» 121099, Москва, Г-99, Шубинский пер., 6

155
. 155
Аси-
озе-
дин
75),
81),
Коз-
кий
яни
06),
13),
Фе-
26),
. 234
Ф. Сафонова

2 р. 80 к.