4
М-38

АКАДЕМИЯ НАУК СССР
Институт русского языка

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ПРЕДПРОЕКТНЫЕ ИССЛЕДОВАНИЯ

МОСКВА — 1988

ИРЯ
институт русского языка
им. в. в. виноградова



www.ruslang.ru

1 [illegible] экз.

АКАДЕМИЯ НАУК СССР
Институт русского языка

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА: ПРЕДПРОЕКТНЫЕ ИССЛЕДОВАНИЯ

4
М-38

1-й экз.

Москва - 1988

Машинный фонд русского языка:
предпроектные исследования.
Объем 10 а.л.
Отв. ред. член-корреспондент АН СССР Ю.Н.Караулов

АННОТАЦИЯ

Авторы статей, объединяемые сборником, поставили перед собой задачу охарактеризовать структуру и пути создания, ведения и использования фонда текстов русской письменной и устной речи на машиночитаемых носителях. Читатель вне сомнения обратит внимание на работы, посвященные проблеме анализа через синтез, вопросам автоматизации исследований поэтических текстов, системам ввода речевого сигнала в ЭВМ СМ-4. Для эффективного использования текстового фонда представляются существенными результаты исследований, посвященных проблемам теоретического программирования.

Сборник предназначен для специалистов, занимающихся разработкой и проектированием структурных компонент Машинного фонда русского языка.

ПРЕДИСЛОВИЕ

Предлагаемый вниманию читателей сборник является вторым в серии публикаций, отражающих состояние работ, связанных с проектированием Машинного фонда русского языка (первый сборник: "Машинный фонд русского языка: идеи и суждения", М.: Наука, 1986). Его материалы соответствуют стадии предпроектных исследований (1986 г.), следовательно, они в какой-то степени продолжают линии обсуждений этой задачи, намеченные в 1983 г. на Первой всесоюзной конференции, посвященной созданию Машинного фонда русского языка как системы комплексной автоматизации и информационного обеспечения лингвистических исследований и прикладных разработок.

В 1987 г. состоялась Вторая всесоюзная конференция по созданию Машинного фонда русского языка, на которой были подведены итоги предпроектных исследований и обсуждены задачи подготовки технических проектов его компонентов - лингвистических программно-источниковых пакетов.

В настоящий сборник были отобраны статьи авторов некоторых, наиболее продвинутых проектов. Его основная задача - сконцентрировать внимание разработчиков на основных конструктивных идеях, связанных с понятиями модульности программ и данных, "дружественности" программного обеспечения, тесной связи его проектирования с особенностями лингвистического материала и принципами работы филологов над источниками.

Редакционная коллегия

В.М.Андрющенко,
В.В.Морковкин

БАЗОВОЕ ЛЕКСИКОГРАФИЧЕСКОЕ ЗНАНИЕ И КОМПЬЮТЕР
(к постановке задачи создания лексикографических процессоров)

> "Наука - это часть наших знаний, которую мы сумели понять настолько хорошо, что можем обучить этому ЭВМ. Там, где мы не достигли еще такого уровня понимания, речь пока идет лишь о профессиональном искусстве. Формальная запись алгоритма или программы ЭВМ, по существу, позволяет нам выполнить весьма полезный тест глубины наших знаний, так как переход от искусства к науке просто означает, что мы поняли, наконец, как автоматизировать данную предметную область".
>
> Д.Кнут [1]

Можно соглашаться или не соглашаться с этим тезисом (все дело в том, как понимать слово "наука" - но благодаря ему в споре, является ли лексикография наукой, искусством или ремеслом, мы можем увидеть и рациональное зерно: ту часть "базового лексикографического знания" [2], которую следует передать ЭВМ.

То, что лексикография нуждается в компьютеризации, ясно давно: ведь лексикографический труд, как никакая другая лингвистическая работа, долог и тяжек, требует соблюдения большого числа разнообразных правил и приемов,

---

[1] Цит. по Громов 1986, 90.

[2] Это слово по отношению к лексикографии, может быть, более уместно, чем "автоматизация" (см. Андрющенко 1986), так как в последнем нейтрализуются смыслы "автоматическое" (целиком компьютером, без участия человека) и "автоматизированное" (человеком с помощью ЭВМ) составление словарей.

знание которых отличает профессионального лексикографа от любителя и от "лексикографа поневоле". В то же время все хотят, чтобы словарей издавалось больше, "хороших и разных". Теперь словари уже не только памятники культуры, интересные книги "для широкого круга читателей", они стали инструментами разнообразной работы с языком и над языком. Исследование и редактирование, обучение языку и перевод, поиск информации и автоматическая обработка данных, общение с ЭВМ требуют словарей. Лексикографическое знание не должно больше оставаться корпоративным, но и передача его может быть лишь "цеховой": лексикографию не изучают по книгам и статьям, ею овладевают на практике, в процессе составления словарей. Поэтому и компьютер в лексикографии целесообразно рассматривать как гибкий интеллектуальный обрабатывающий центр, помогающий начинающим лексикографам овладеть профессией и освобождающий искушенных от рутинных операций, но оставляющий достаточно простора для проявления индивидуального творчества.

Постараемся прежде всего сформулировать тот минимум лексикографических знаний, который нужно суметь передать ЭВМ, чтобы обеспечить широкомасштабную компьютеризацию словарного дела. В дальнейшем мы будем иметь в виду именно такую компьютеризацию, так как опыт частичных применений ЭВМ для решения отдельных лексикографических задач пока не привел к решительным сдвигам в области автоматизации этого вида лингвистической деятельности, самим своим характером, казалось бы, вполне подготовленной к алгоритмизации многих процессов.

В самом деле, создание словарей осуществляется, как правило, на основе картотек. Но ведь и ЭВМ может быть запрограммирована на выполнение любых картотечных операций. При подготовке словарей требуется получение, обработка и учет разнообразной информации. Но ведь и ЭВМ - это не что иное, как программируемая информационная машина. Словарь обладает сложной структурой, обычно его созданию предшествует написание проекта

или инструкции для составителей, что на вычислительном языке мы назвали бы моделированием. При применении компьютера написание и редактирование можно рассматривать как единый по своей природе процесс: в переводе на вычислительный язык написание - это редактирование и одновременно реализация модели. Представляя последовательность действий лексикографа как реализацию (под контролем и управлением компьютера) все более приближающихся к целевому словарю его моделей, мы можем в принципе алгоритмизовать весь процесс создания словаря - от замысла до рукописи. Последний участок - от рукописи до книги, - если рукопись электронная, вообще может быть пройден целиком автоматически.

Рассмотрим предварительно "начальные" составляющие лексикографической деятельности:

А. "Пред"-лексикографическая работа - обработка источников для составления словаря, подготовка исходной картотеки;

Собственно лексикографическая работа:

Б. Проектирование словаря [3];

В. Написание и редактирование словаря;

Г. "Пост"-лексикографическая работа - набор и верстка, издание книги.

Для каждой из этих составляющих информатика и вычислительная лингвистика имеют соответствующие им алгоритмические и программные средства:

---

[3] Мы вовсе не утверждаем, что подготовка источников непременно предшествует разработке проекта словаря - просто нам удобно рассматривать эти вопросы в такой последовательности. Однако подготовку источников и проектирование в общем случае допустимо рассматривать и как относительно независимые этапы. Проектировать словарь можно и не имея источников (их сбор и изучение осуществляются в процессе реализации проекта словаря), а можно и опираясь на уже собранные и предварительно обработанные источники, в том числе и такие, которые были использованы ранее для составления других словарей. Однако в "непрерывном" лексикографическом процессе наиболее типична ситуация, когда при подготовке нового словаря имеет место дополнительный сбор материала, который не прекращается и после выхода словаря в свет.

А′. Информационные системы как аналог картотек; автоматические конкордансы как источники текстовых данных;

Б′. Формально-грамматические анализаторы и синтезаторы информационных структур; автоматические словари как вычислительные модели словарей в полиграфическом исполнении;

В′. Диалоговые текстовые редакторы и системы управления словарными базами данных;

Г′. Автоматические спеллеры (программы, корректирующие орфографию текста, обеспечивающие правильность переносов), программы автоматического набора и верстки.

Задача состоит в том, чтобы из этих компонентов "сшить" систему автоматизированной подготовки словарей. Однако сделать это путем прямого использования подходящих типовых программных средств пока никому не удавалось; все известные нам попытки приводили лишь к частичному применению ЭВМ, а не к органичной и адекватной компьютеризации всего словарного дела. Ни уже давно отработанная техника составления частотных словарей, словоуказателей и конкордансов, включая генерацию карточек для словарных картотек, ни технология использования информационных систем для создания автоматических словарей, ни оригинальные методики создания и эксплуатации собственно автоматических словарей, ни средства диалогового ввода и редактирования текстов, ни технология автоматизации полиграфических процессов, ни какая-либо комбинация перечисленных процедур и сущностей пока не привели к созданию полноценных лексикографических систем (**Knowles** I983, I8I-I95, **Landau** I984, 272-293). Дело, видимо, заключается в том, что в таких программно-технологических комплексах пока не удается воссоздать носитель и "задатчик" собственно лексикографических знаний, который мы назовем лексикографическим процессором и попытаемся, насколько это возможно, описать его

необходимые свойства и функции [4].

Лексикографический процессор - это диалоговая программа, позволяющая:

а) готовить формализованный проект словаря, т.е. формулировать условия отбора слов в словник, определять схемы словарных статей, осуществлять комплектацию разделов будущего словаря;

б) задавать в соответствии с формализованным проектом для каждого очередного формируемого блока словаря его заголовок и "фрейм" (путем подсказки лексикографу его очередного шага и предоставления ему необходимой информации);

в) автоматически формировать стандартную информацию, например, морфологические и иные пометы, транскрипцию, лексикографические знаки и шрифтовые выделения, межстатейные ссылки, сокращения и т.д.;

г) производить автоматические проверки схем словарных статей, сигнализировать о расхождении в схемах однородных статей; производить групповые сравнения словарей, получать их объединения, пересечения и симметрические разности;

д) строить индексы к словарю, автоматически осуществлять "перевертывание" двуязычных словарей;

е) строить "проекции" словаря по заданному типовому фрагменту словарной статьи (например, выводить все заголовки, все цитаты, речения, определения с указанной пометой и т.п.);

ж) выполнять любую другую, формализуемую, но не предусмотренную пунктами а(-е) лексикографическую работу.

Лексикографический процессор, работающий совместно с пакетами А′ - Г′, должен быть той цементирующей основой, тем центральным и специфическим зерном, которое способно механическое объединение независимых и ограниченных в

---

[4] Подчеркнем, что нам не известен ни один реально сделанный лексикографический процессор; все дальнейшее изложение - это лишь наше предположение, каким он должен быть и как должна быть поставлена компьютеризация словарного дела.

своих функциях программных средств сделать адекватным инструментом компьютерного конструирования и изготовления словарей.

Рассмотрим теперь логику словарного дела с этих позиций.

Основные понятия и определения

Лексикография как отдельная область филологической деятельности разделяется на теоретическую и практическую. Практическая лексикография включает в себя, во-первых, создание словарей, а во-вторых, накопление, хранение и обработку первичных словарных материалов. Теоретическая лексикография характеризуется более сложной структурой. Она состоит из теории и истории лексикографии. Теория лексикографии, в свою очередь, объединяет по крайней мере семь компонентов, каковы: а) определение объема, содержания и структуры понятия "лексикография", б) словарная лексикология, в) учение о жанрово-типологическом разнообразии словарных произведений, г) учение об элементах и параметрах словарей, д) лексикографическое конструирование, е) учение о первичных словарных материалах, ж) учение о планировании и организации словарной работы [5]. Из названных семи компонентов наибольшее внимание традиционно привлекали и продолжают привлекать учение о жанрово-типологическом разнообразии лексикографических произведений и учение об элементах и параметрах словарей.

В контексте обсуждаемой проблемы особую важность приобретает именно жанрово-типологический аспект, поскольку в принципе каждый определенный в жанрово-типологическом отношении вид лексикографирования может претендовать на свою разновидность лексикографического процессора.

Жанрово-типологическое разнообразие лексикографических произведений создается в результате одновременного учета трех характеристик: объекта описания (какие языковые единицы описываются в словаре), характера описания

---

[5] Подробнее см.: Морковкин В.В. (в печати).

(как описываются включенные в словарь языковые единицы) и его цели (для кого и/или для чего предназначено соответствующее словарное описание). Жанровая определенность сообщается словарю единицами описания, в качестве которых могут выступать а) слова, б) единицы меньшие, чем слово (морфемы вообще и отдельные виды морфем), в) единицы большие, чем слово (свободные словосочетания, фразеологизмы, афоризмы и т.п.), г) комплексные единицы лексической системы (словообразовательные гнезда, лексико-семантические и лексико-фонетические объединения и т.п.). Типологическая конкретность словаря обусловлена тем, как именно разрешаются в нем такие дихотомии, как лингвоцентрический **vs**, антропоцентрический характер описания, объяснительный **vs**, демонстрирующий способ обнаружения значения заголовочных единиц, полиаспектная **vs**, аспектная направленность словаря, прямое **vs**, обратное (инверсионное) расположение заголовочных единиц и некоторые другие. Существенно важную роль в формировании типа словаря играет ориентация на определенную цель (для фиксации всего накопленного знания о каждой заголовочной единице; для осуществления продуктивных или рецептивных видов речевой деятельности; для наведения эпизодических справок и т.п.) и определенного адресата (для тех или иных категорий учащихся, для переводчиков и т.п.).

Центральными понятиями учения об элементах и параметрах словарей являются понятия макро- и микроструктуры словаря (Hausmann 1977, 3-4) и само понятие лексикографического параметра (Караулов 1981, 74-136).

Под макроструктурой понимают установленное для каждого данного словаря последовательность "черных" (описываемых словарем) слов и выражений, систему отношений между ними и комплектацию словаря, т.е. содержание и связи его частей, а под микроструктурой - множество типовых аспектов описания слова в словарной статье и систему отношений (например, взаимный структурный порядок подачи) фрагментов, соотносимых с установленными парамет-

рами, при условии соблюдения правил оформления фрагментов каждого типа.

Каждый словарь обладает своей индивидуальной макро- и микроструктурой, своим набором параметров и формой их подачи, но логика здесь одна: имеется перечень лексикографируемых единиц (словарных форм слов и/или словосочетаний) - словник; каждая такая единица является либо заголовком словарной статьи, либо заголовком подстатьи; каждая словарная статья и подстатья описывает свою заголовочную единицу выбранным для данного словаря набором атрибутов (аспектов, параметров) - орфографическим (всегда), акцентуационным (почти всегда), орфоэпическим (эвентуально), морфологическим (в той или иной мере всегда), семантическим, с точки зрения употребления (чаще всего комментированными цитатами), с точки зрения ее словообразовательных свойств (к сожалению, полно и всесторонне - весьма редко), приведением различного рода справок (этимологических - в неэтимологических словарях, источниковедческих, исторических - в неисторических словарях).

Создание любого словаря включает в себя две основные процедуры: формирование словника и интерпретацию отобранных языковых единиц. Содержанием словарной интерпретации является устранение неопределенности относительно заголовочных единиц. Дело в том, что каждая включенная в словник единица является единицей обобщенной, характеризующейся неопределенностью в области фонетических, акцентологических, морфологических, семантических и синтаксических свойств. Эта неопределенность обусловлена различием свойств лексических единиц с точки зрения языка и с точки зрения речи (дихотомия "язык - речь"). Действительно, на уровне языка любая лексическая единица представляет собой совокупность, с одной стороны, семантически незначимых вариантов, в том числе произносительных (ср. було[чн]ая и було[шн]ая), акцентологических (ср. ме́льком и мелько́м), морфологических (ср. горячий кофе и горячее кофе), синтаксических (ср. тосковать по нем и тосковать по нему), а с другой - семантически зна-

чимых вариантов, в том числе лексико-грамматических (ср. дом и дома и дому и домом и доме) и лексико-семантических (ср. войти в аудиторию и Аудитория встала). Включение в речь (т.е. в сферу функционирования) устраняет как семантически незначимую, так и семантически значимую неопределенность плана выражения и плана содержания лексической единицы. В результате "речевое слово" - это всегда вполне определенный лексико-фонетический (ср. було[чн]ая или було[шн]ая), лексико-акцентологический (ср. ме́льком или мелько́м), лексико-морфологический (ср. горячий кофе или горячее кофе), лексико-синтаксический (ср. тосковать по нем или тосковать по нему), лексико-грамматический (ср., дом или дома или до́му или домом или доме), лексико-семантический (ср., войти в аудиторию или Аудитория встала). Таким образом, устранение упомянутых видов неопределенности в рамках словарной интерпретации достигается прежде всего установлением корреляции между языковыми и речевыми свойствами заголовочных единиц.

Слово как основная языковая и предельная речевая единица характеризуется следующим набором требующих приведения в известность свойств: а) характер произношения, б) характер словоизменения, в) характер семантической структуры (неоднозначности), г) характер значения каждого лексико-семантического варианта, в том числе характер абсолютной ценности (денотативное и сигнификативное значение, включая страноведческий фон), характер относительной ценности (способность слова вступать в различные парадигматические отношения с другими словами, о его стилистической и статистической ценности, внутренней форме и др.), характер сочетательной ценности (способность слова сочетаться определенным образом с определенными словами), д) характер словообразовательной ценности (морфемный состав и словообразовательная структура).

Характер произношения заголовочных единиц отражается с помощью их сплошного или выборочного транскрибирования.

Сведения о словоизменении оформляются в виде грамматической характеристики. По месту расположения грамматическая характеристика может быть присловной и присловарной. Присловная грамматическая характеристика (она приводится непосредственно в словарной статье) бывает минимальной, восполняющей и полной. Минимальная грамматическая характеристика состоит в приведении только исходной и показательной (показательных) словоформ заголовочной единицы (показательная словоформа, которая в сочетании с исходной позволяет с наибольшей легкостью восстановить соответствующую словоизменительную парадигму). Восполняющая грамматическая характеристика кроме исходной и показательной словоформ включает в себя еще и восполняющую словоформу или восполняющие словоформы (восполняющая словоформа - та, которая обнаруживает какую-либо особенность словоизменительной парадигмы данного слова, отличающую последнюю от соответствующей образцовой парадигмы). Полная грамматическая характеристика состоит в приведении всех без исключения словоформ заголовочной единицы. Присловарная грамматическая характеристика - это оформленный в виде отдельной композиционной части словаря упорядоченный перечень образцовых парадигм, отражающих словоизменительные свойства заголовочных единиц. Каждая образцовая парадигма получает определенный индекс, с помощью которого обеспечивается взаимодействие между присловной и присловарной грамматическими характеристиками (или между заголовочной единицей и присловарной грамматической характеристикой). По ориентации на виды речевой деятельности грамматическая характеристика бывает активной (порождающей, кодирующей) и пассивной (аналитической, декодирующей).

Основными способами упорядочения значения многозначного слова являются исторический, логический и эмпирический. В согласии с историческим способом значения (как живые, так и уже вышедшие из употребления) располагаются и рассматриваются в последовательности

их появления в языке. Задача исторического упорядочения значений - воссоздание реальной семантической истории слова на основе всей суммы имеющихся письменных источников. В отличие от исторического логическое упорядочение не требует непременного отражения всех значений многозначного слова. Его задача, во-первых, отразить имеющиеся между значениями зависимости, а во-вторых, обеспечить одинаковую интерпретацию семантически родственных слов в отвлечении от особенностей их смыслового развития. В соответствии с эмпирическим способом разные значения упорядочиваются обычно на основе различных практически существенных, но с лингвистической точки зрения случайных факторов, таких, как частотность лексико-семантического варианта, его педагогическая ценность и т.п.

Ведущую роль в семантизации лексико-семантического варианта играет обнаружение главного компонента его значения - абсолютной ценности (соотносимого со словом смысла). Для называния этой операции используются три термина: определение, толкование и дефиниция. Определение - объяснение смысла с помощью словосочетания (простого или сложного) либо посредством отсылки к другому слову, а также текстовой отрезок, содержащий это словосочетание, эту отсылку. Толкование - объяснение смысла заголовочной единицы в филологическом словаре с помощью других слов (синонимов, антонимов) или словосочетаний (как простых, так и сложных), а также текстовый отрезок, состоящий из этих слов и словосочетаний. Дефиниция - объяснение смысла заголовочной единицы с помощью словосочетания (обычно сложного), согласованное с требованиями, предъявляемыми к логическим определениям, а также текстовый отрезок, состоящий из этого словосочетания. Толкование закреплено за филологическими словарями, дефиниция - за энциклопедическими (в том числе терминологическими) словарями, а определение может применяться по отношению как к тому, так и к другому. Термин "определение", как правило, не употребляется

вместо термина "толкование" в том случае, если соответствующий семантизирующий отрезок содержит одно слово. В нестрогом словоупотреблении все три термина часто используются недифференцированно без заметного ущерба для общего смысла высказывания.

Попытки создать строгую теорию описания значений лексикографируемых единиц приводят к понятию метаязыка толкований, состоящего из слов, используемых в объяснениях, синтаксических конструкций, по которым строятся объяснения, и правил выбора конструкций и заполняющих их слов в зависимости от свойств лексико-семантических вариантов. Если бы удалось найти формальные способы задания метаязыка толкований, можно было бы разработать автоматические методы порождения определений, которые давали бы однородной в семантическом отношении лексике однородные описания [6].

Система понятий (и называющих их терминов), различных условных знаков, шрифтовых выделений и т.п., с помощью которых описывается словарь, его структура и содержание составляют метаязык словаря: к метаязыку словаря относятся такие понятия, как "черное слово", "дефиниция" ("определение", "толкование"), "помета", "словник", "левая (правая) часть словаря", "словарный вход" и т.д.

"Среди используемых в лексикографии помет выделяются грамматические (напр., несов., прош., перех., кратк. ф., и т.п.), лексические (напр., о человеке, о транспорте и т.п.), семантические (напр., перен. и т.п.), функционально-стилевые (в том числе и терминологические, напр., газетн., офиц., спец., хим. и т.п.), стилистические (напр., разг., прост., высок. и т.п.), эмоционально-экспрессивные (напр., ирон., шутл., неодобр., презр., почтит. и т.п.), хронологические (напр., устар., истор., нов. и т.п.), статистические (напр., редко, реже, обычно, малоупотр. и т.п.), запретительные (напр.,

---

[6] Это один из главных признаков хорошего стиля лексикографирования.

не употр. и т.п.), ограничительные (напр., только [несов.] и т.п.) и некоторые другие" (Морковкин 1986, 102-117). Не только между различными словарями, но и внутри одного и того же словаря имеет место заметная непоследовательность использования одних и тех же помет, в особенности их сочетаний.

Значения нумеруются (если значение одно, то оно остается ненумерованным). В составе значения могут выделяться оттенки, которые в русских словарях, как правило, не нумеруются, а отделяются друг от друга каким-нибудь лексикографическим знаком (чаще всего - "параллельками").

Каждый оттенок значения обычно иллюстрируется примерами употребления. Примеры употребления могут быть цитатами или речениями (показательными словосочетаниями или предложениями, специально придуманными для иллюстрации употребления или достаточно распространенными, иногда устойчивыми выражениями, штампами и т.п.). Примеры выполняют, как правило, несколько функций: кроме того, что они иллюстрируют возможности употребления (и доказывают правильность описания), ими демонстрируются типовые лексические, семантические и синтаксические сочетаемости слова. Среди примеров могут встречаться и устойчивые сочетания, хотя для описания таких сочетаний обычно отводится специальная "зона" словарной статьи - так называемая "заромбовая часть" (часть словарной статьи, обычно последняя, если нет справочного отдела, размещаемая после лексикографического знака "ромб").

Если в словарную статью встраиваются и описания производных, то такие описания могут быть размещены после описания всех лексико-семантических вариантов и упорядочены по частям речи и по "линиям производности". Здесь следует заметить, что каждый лексико-семантический вариант в принципе обладает своей системой производности, поэтому, строго говоря, производные, описываемые после всех лексико-семантических вариантов, следует считать вынесенными "за скобки".

Теоретически описание каждого производного слова должно повторять схему описания отдельного слова, хотя фактически так не делается из-за риска превысить допустимую для восприятия степень сложности словарной статьи. Словари, содержащие описание производных внутри статей производящих, называются гнездовыми. Описание производных может быть заменено отсылками на статьи, заголовками которых они являются.

Корпус словаря может содержать отсылочные словарные статьи, в которых вместо описания лексико-семантического варианта находится отсылка на словарную статью, где этот или аналогичный ему (находящийся с ним в некотором стандартном отношении) вариант описан. Ссылки могут быть весьма разнообразны ("см.", "см. также", "то же, что и..." и т.д.), каждая их разновидность может иметь тонкие правила употребления, впрочем, часто нарушаемые.

Фразеологические единицы, в состав которых входит описываемое слово, как уже говорилось, не в фразеологическом словаре описываются в конце статьи. Схема их описания может быть аналогичной схеме целой статьи. Но во фразеологическую единицу входят несколько слов; обычно словари не помещают описания фразеологических единиц столько раз, сколько слов они содержат (служебные слова в расчет совсем не принимаются), а приводят только одно описание, давая на него из других статей отсылки. Выбор слова, в статье которого описывается фразеологическая единица, часто бывает весьма произволен и непоследователен.

Статья может заканчиваться справочным отделом, в котором помещаются различные "внесистемные" сведения. К ним относится этимология (в энциклопедических словарях, в словарях иностранных слов сведения о ней помещаются в начале статьи перед дефиницией); дата фиксации слова (в некоторых зарубежных словарях приводится дата фиксации каждого значения); словари, в которых слово было засвидетельствовано ранее; другие уместные сведения.

2-7018

1-й ту

Таким образом, описание структуры словарной статьи может быть дано в терминах непосредственных составляющих, и в общем случае оно может носить рекурсивный характер: некоторые составляющие словарной статьи (например, описания производных и фразеологических единиц) могут повторять схему целой словарной статьи. Поэтому в автоматических словарях такие описания обычно выносятся в отдельные статьи, а в статьи, в которых они могли бы быть приведены, вносятся соответствующие отсылки. Кроме того, статьи могут содержать отсылки к синонимам, антонимам, паронимам, конверсивам, именам родовых и видовых понятий (гиперонимам и гипонимам), другим типам лексических отношений. Эта система отношений образует часть макроструктуры словаря.

Другую ее часть образуют разделы словаря: руководство по его использованию, система индексов, называемых дополнительными входами в словарь, грамматический справочник и другие справочники, например, справочники имен собственных, сокращений и т.п.

В автоматических словарях эти части также имеют свои аналоги: руководству по использованию словаря соответствует его формальное описание, грамматическому справочнику - программы морфологического анализа и синтеза, а индексы - просто конструктивно необходимая часть автоматического словаря.

Наряду со словарями можно рассматривать и более крупные объекты - словарные системы, серии, ряды и семейства.

Основное свойство словарной системы - взаимодополнительность лексикографической информации словарей, входящих в нее. "Для русской лексикографии... систему словарей должны составить словарь современного русского литературного языка, исторический словарь русского языка XIX в., исторический словарь русского языка XУШ в., словарь русских народных говоров, словарь древнерусского языка XI-XУП вв. Эту систему должны дополнить словарь синонимов, фразеологический

словарь, словарь антонимов, словообразовательный словарь и другие виды специальных лексикографических изданий" (Сороколетов 1978, 40-41).

Словарная серия образуется по принципу дополнительности словника: в одну серию входят, например, а) тезаурусы информационно-поисковых "подъязыков", б) лингвострановедческие словари, в) двуязычные и многоязычные терминологические словари разных предметных областей и т.д.

Основное свойство словарного ряда - наследственность, поддержание и развитие лексикографической традиции. Почти все русские толковые словари представляют один ряд: более поздние словари, как правило, наследуют многие черты (и фактический материал) более ранних, хотя каждый, конечно, вносит и свой заметный вклад в описание лексики.

Наконец, отличительный признак словарного семейства - разнообразие способов подачи одних и тех же характеристик слов. Все ортологические словари, все морфемно-словообразовательные словари образуют свои словарные семейства.

Такова система основных лексикографических понятий, отражающих содержание традиционной лексикографической деятельности. Какую же долю в этой деятельности занимает точное знание?

Ответить на этот вопрос, хотя бы частично, мы сможем, создав полноценные, отвечающие реальным потребностям практики системы компьютеризации лексикографических работ. В качестве первого шага в этом направлении рассмотрим коррелирующие с основными понятиями лексикографии понятия информатики и вычислительной лингвистики.

Будем называть базой (лексикографических) данных совокупность взаимосвязанных наборов данных (файлов), содержащих относительно самостоятельные структурированные единицы данных. Такие единицы мы будем называть модулями (данных). Для лексикографии существенны

два основных типа модулей - словарные и текстовые модули. Словарные модули мы отождествим со словарными статьями (или группами словарных статей), текстовые - с отрывками текстов (выборками, эксцерпциями).

Системой управления (лексикографическими) базами данных мы будем называть совокупность программных средств, обеспечивающих создание баз из внешних источников (словарей, текстов или из данных, вводимых лексикографом), поддержание баз в рабочем состоянии, регулярные преобразования баз и операции над ними, их изменение и дополнение, удовлетворение информационных потребностей как лексикографов, так и потенциальных пользователей ("читателей словарей").

Системы управления базами данных функционируют на основе "знаний" об устройстве своих баз. Эти знания хранятся вместе с данными (являются частью базы данных) и называются схемами баз данных. В применении к лексикографии мы их будем называть схемами лексикографических источников. Чтобы стать доступными программам системы управления лексикографическими базами данных, схемы источников должны быть записаны на некотором формальном языке, пригодном для задания метаязыка произвольного словаря. В схемах должны быть описаны типовые структуры модулей, формы подачи лексикографической информации, ее взаимный порядок, взаимосвязи модулей и т.д. Будем считать, что модули членятся на фрагменты; фрагменты, структурно аналогичные модулям, мы будем называть субмодулями. Совокупности модулей образуют корпусы (текстовые корпусы - конструкции, достаточно распространенные в лексикографии; отдельные словари мы также будем считать корпусами); совокупности корпусов мы будем называть пакетами. Таким образом, пакеты могут представлять системы, серии, ряды и семейства словарей.

Лексикографические базы данных могут быть информационными и оперативными. Информационная база является источником лексикографических данных, оперативная

представляет модель подготавливаемого словаря или его части. Оперативная база данных может быть персональной базой отдельного лексикографа.

Автоматизированной лексикографической системой мы будем называть совокупность взаимодействующих информационных и оперативных баз лексикографических данных. Автоматизированная лексикографическая система управляется лексикографическим процессором.

Ниже мы рассмотрим логику словарной работы с применением автоматизированной лексикографической системы.

Замысел словаря

Работа над словарем начинается с ответа на вопросы "для чего?" и "для кого?". Обычно эти вопросы не различают, сводя их к одному вопросу "для кого?". Однако ответ на первый вопрос предопределяет преимущественно состав информации о слове, а на второй - ее форму. Ответы на оба вопроса вместе предопределяют как состав словника, так и характер описания его единиц. Для лексикографического процессора ответы на такие вопросы, конечно, являлись бы слишком общими (трудно было бы разработать формальную процедуру, проектирующую лексикографическую базу данных по таким аргументам, как "нужен словарь для перевода с русского на английский школьниками, родным языком которых является английский"), вместо них лексикографический процессор должен был бы задать ряд уточняющих вопросов и собрать приблизительно такую информацию:

А. Какими свойствами должны обладать (скорее, не должны обладать) слова, включаемые в словник;

Б) Какова типовая схема словарной статьи проектируемого словаря: какими могут быть заголовки; нужно ли давать транскрипцию, если "да", то полную или частичную; каков состав и форма грамматической информации; как отражать особенности изменения слов по отношению к типовым парадигмам; в какой форме представлять синтаксическую и лексическую сочетаемость; каковы допус-

тимые модели и формы словарных определений и подачи оттенков значений; как отражать отношения производности слов и вхождения их во фразеологизмы и т.п.

В. Какие индексы должны быть построены к словарю;

Г. Какие источники (тексты, словари, справочники) для проверки и комплектации какой информации будут использованы.

Ответ на последний вопрос определяет подготовку исходных баз лексикографических данных.

Базы лексикографических данных

Базы лексикографических данных могут быть текстовыми, словарными, документально-фактографическими и объектно-характеристическими.

Текстовая база данных естественно представляется конструкцией, называемой автоматическим конкордансом; словарная - автоматическим словарем; документально-фактографическая - информационной системой общего назначения, объектно-характеристическая - реляционными системами управления базами данных (СУБД). Общим свойством баз данных всех типов должна быть конструктивно предусмотренная возможность автоматического обращения к ним из независимого лексикографического процессора. Это естественно обеспечивается в автоматических конкордансах и словарях, но с трудом, или скорее вообще не обеспечивается, в типовых информационных системах общего назначения и в СУБД. Вот почему в вычислительной лексикографии приходится строить свои собственные специфические лексикографические системы управления базами данных.

Автоматический конкорданс - это один или несколько текстовых файлов, один или несколько словоуказателей к ним, содержащих также и статистику словоупотреблений, и система программ, обеспечивающих построение этой конструкции, внесение в нее изменений и дополнений, ее поддержание в рабочем состоянии и обращение к ней за справками, как правило, в диалоговом режиме или из другой

программы, в частности, из лексикографического процессора. При обращении к автоматическому конкордансу формируются условия поиска контекстов, их границы и форма выдачи. В ответ получаются цепочки литер, удовлетворяющие условиям поиска. Запрошены могут быть, естественно, словоформы слова, словарные формы слова (если в состав программ входит лемматизатор, т.е. программа, перерабатывающая текстовые формы слов в словарные), образы слов (например, слова, начинающиеся, или содержащие внутри себя, или кончающиеся определенными буквосочетаниями или сочетаниями букв определенных классов), словосочетания или их образы, слова, обладающие определенными статистическими свойствами и т.д. Длина контекста (цепочки литер) может быть указана числом знаков слева и справа от искомого слова, числом слов или положением между разделителями (маркерами) определенного вида, например, между маркерами конца предложения, или указанием вида и числа пропускаемых маркеров и т.п. образом. Контексты могут быть отпечатаны в различной форме: обычной пополосной печатью в две или три колонки, в виде стандартных словарных карточек, в виде заготовок словарных статей, выведены на терминал или на распечатку, в рассортированном по аргументу поиска порядке.

Автоматический словарь - это один или несколько словарных файлов (наборов данных, содержащих тексты или фрагменты словарных статей какого-либо введенного в ЭВМ или созданного в ней словаря), индексы к ним, описание макро- и микроструктуры этого словаря и система программ, обеспечивающая создание этой конструкции, ее поддержание в рабочем состоянии, ее преобразование и обращение к ней за справками, как правило, в диалоговом режиме или из другой программы, например, из лексикографического процессора. Обращение к автоматическому словарю аналогично обращению к автоматическому конкордансу с тем отличием, что аргументами могут служить не только слова, их образы и признаки, но и типы фрагментов словарных статей, а условия, накладываемые на контекст формулируются в терминах структурных составляющих словарных

статей. Выдаваемые контексты могут сортироваться по признакам аргументов поиска. Программы автоматического словаря должны обеспечивать также операции над словарями, например, слияние двух и более словарей, сравнение двух словарей (нахождение общей части и различий), изменение структур словарных статей (например, превращение гнездового словаря в негнездовой); индивидуальное и групповое редактирование словарных статей и т.п. Необходимо отметить, что в хорошем автоматическом словаре при вводе словарных статей автоматически восстанавливаются недостающие ссылки, а при создании словарных статей такие ссылки автоматически формируются; индексы к словарным статьям также строятся автоматически. Поскольку такие фрагменты словарных статей, как определения, примеры, фразеологические единицы могут быть рассмотрены как тексты (и вообще весь словарь может быть рассмотрен как текст), к ним могут быть построены такие же, как в автоматическом конкордансе, словоуказатели. Поэтому, например, такая лексикографическая задача, как соблюдение требования, чтобы в качестве "черных слов" встречались все слова, использованные в определениях, примерах и фразеологических или иных словосочетаниях, может быть выполнена автоматически.

Более того, идя по пути дальнейших обобщений, можно считать, что автоматический конкорданс является частным случаем автоматического словаря: словарная статья здесь состоит просто из набора контекстов для каждого из индексных слов. Но ведь это и есть исходная словарная картотека!

Возможна и такая конструкция: автоматический словарь является как бы индексом к текстовому корпусу. Словарь - корпус, относительно закрытый, текстовой же корпус открыт принципиально. Поэтому вполне возможно автоматический конкорданс строить так: создать его индекс на базе какого-либо словаря, а затем вводить тексты, присоединяя к индексирующим словам адреса их встречаемости в модулях. В такой конструкции создается возможность

постоянной проверки существующего словаря на текстах; кроме того, если ряд слов индекса замаскировать, то предотвращается создание ссылок на ненужные контексты.

Есть два обычных возражения лексикографов против использования конкордансов: их засоренность неинформативными примерами и фиксированная длина контекстов. Второе, как мы видели, снимается в автоматических конкордансах, первое же требует конструирования фильтра, подавляющего выдачу ненужных контекстов. Полноценным фильтром может быть, естественно, лишь достаточно развитый лингвистический процессор, т.е. программа, способная осуществлять автоматический лингвистический анализ и "понимание" текста на данном языке (уже упомянутый лемматизатор, разновидность морфологического анализатора - составная часть лингвистического процессора).

Но в автоматических конкордансах возможны и обычные эксцерпционные методы сбора материала. Эксцерпции могут рассматриваться как элементарные выборки (разумеется, без доверия к статистике, собранной на ансамбле из таких "выборок"), вводиться вручную в базу данных по мере их поступления и образовывать текстовой корпус, аналогичный по своей структуре и функциям обычным словарным картотекам. (При этом отпадает необходимость дублирования "карточек" и их раскладки!) Возможен также и автоматизированный (а с применением лингвистических процессоров и автоматический) отбор эксцерпций из текстовых корпусов, сформированных либо как ансамбли случайных выборок, либо как наборы полных образцовых текстов. При достаточно формальных требованиях к формированию таких корпусов и к спецификации релевантных фрагментов, отбор может быть запрограммирован и осуществлен непосредственно с полиграфических носителей.

Третьей исходной конструкцией, как мы уже говорили, является документально-фактографическая система, хранящая тексты документов (например, статей по лексикологии, или их рефератов, параграфов грамматики и т.п.) и обеспечивающая к ним доступ и получение либо текстов, содержащих нужные сведения (документальная компонента), либо

самих сведений (фактографическая компонента). Лексикографический подход к построению такой системы может и не отличаться от общепринятого, за исключением того, что нужно обеспечить ее взаимодействие с лингвистическим процессором. (Впрочем, это требование может иметь далеко идущие последствия в области конструирования информационных лексикографических систем, но мы не будем сейчас касаться этой проблемы, которая выходит далеко за рамки лексикографической тематики.)

Наконец, объектно-характеристической базой данных называется база, в которой представлены имена объектов описания и каждому имени сопоставлены значения описывающих его характеристик (атрибутов). В принципе и текстовые базы данных (автоматические конкордансы), и словарные (автоматические словари), и документально-фактографические информационные системы можно считать частными случаями объектно-характеристических баз данных и соответствующих им систем управления. Однако пока в вычислительной лексикографии понятие объектно-характеристической базы данных удобно рассматривать не в расширенном, а в суженном смысле, считая, что в таких базах не отражаются структурные отношения между характеристиками объектов (что неверно для текстов, словарей и документов). Примерами объектно-характеристических баз могут служить простые перечни слов (имен), снабжаемые какими-либо характеристиками - числовыми, классификационными и другими. Объектно-характеристические базы данных в лексикографии удобны как вспомогательные справочные конструкции, для построения которых излишне применять "тяжелую артиллерию" автоматических словарей и информационных систем. Здесь мы упоминаем о них лишь для полноты изложения.

Словник

Естественными источниками для формирования словника вновь проектируемого словаря являются словоуказатели автоматических конкордансов и индексы автоматических

словарей, документально-фактографических и объектно-характеристических систем. Но может быть использован и еще один, принципиально новый источник - программный генератор слов с заданными свойствами (синтезатор слов, разновидность лингвистического процессора).

С точки зрения длины словника словари варьируют в чрезвычайно широком диапазоне от 3 тысяч лексических единиц в малых учебных словарях до 200 тысяч в академических нормативных словарях и "сколько получится" в полных лексиконах языка (не один из которых, впрочем, не достиг еще миллиона словарных статей и не зарегистрировал действительно все слова, употребительные в обозримом языковом пространстве). Правила отбора словника в решающей мере зависят от жанра и типа словаря, а также от назначенного объема его словника. При этом в словарях фиксированной адресации с четко сформулированными специальными целями они обычно носят позитивный характер, а в словарях "для всех желающих узнать..." они чаще всего являются запретительными: не включаются слова, обладающие определенными свойствами, например, малоупотребительные, слишком специальные (научно-технические), диалектные и местные, грубопросторечные, имеющие определенные префиксы и/или суффиксы, интернационализмы (в малых иностранно-русских словарях), сложные слова с легко раскрываемым значением, имена собственные (кроме употребительных в переносном значении) и т.д. Полный перечень всех мыслимых ограничений на отбор словника, видимо, нельзя сформулировать в виде системы признаков, приписываемых словам в индексах лексикографических баз данных, однако разумное приближение к такой системе может быть достигнуто и на этом пути. Скажем, используя Макротезаурус информационных языков (Автоматизированная система 1982), можно формально задать степень "специализированности" термина его употребительностью в одном или немногих "подъязыках" и его положением в иерархической понятийной системе;

используя Словарь русских говоров, можно таким же образом исключать слова, обладающие ареальной ограниченностью; имея специальные справочники ограниченных лексических слоев (грубопросторечных и бранных слов и выражений, интернационализмов, имен собственных и т.д.), можно проверять по ним автоматически, не входит ли описываемая лексическая единица в какой-либо из таких списков, и, если необходимо, сигнализировать об этом лексикографу.

Наоборот, при формировании полных лексиконов такие справочники могут быть средством контроля полноты словника. Для общей лексикографии, видимо, источниками особой важности сейчас являются терминологические стандарты и номенклатуры промышленных изделий и сельскохозяйственных продуктов, данные словарных служб крупных информационных систем.

В специальном замечании нуждается понятие употребительности. Употребительным следует считать слово или выражение, которое либо встречается в текстах часто и регулярно, либо редко, но постоянно присутствует в сознании говорящих как неотъемлемая составляющая его "модели мира". Первым свойством обладают слова, имеющие достаточно высокий (больший 0,5) коэффициент употребительности по А.Жуйяну (Андрющенко 1978), вычисленный как нормированное отношение коэффициента вариации частоты слова к максимальному коэффициенту вариации для данного корпуса при данном числе и объеме частичных выборок. Такие коэффициенты легко получить в автоматических конкордансах, поэтому их индексы являются хорошим справочником для отсеивания малоупотребительных слов. Как правило, слова с низким коэффициентом употребительности имеют стилистические и/или тематические ограничения: в каких-то текстах они могут встречаться даже достаточно часто и регулярно, но в целом по совокупности их нельзя признать употребительными, так как имеется гораздо больше текстов, в которых они просто не появляются. Но среди этих слов могут встречаться (или вообще не встречаться в текстах) слова,

которые нужно признать употребительными по второму критерию (такие слова называются необходимыми, "постоянно находящимися в распоряжении" - Michéa 1953). К ним относятся названия частей тела, широкоупотребительных предметов обихода, названия основных цветов, временных интервалов, родственных отношений, явлений общественной жизни и т.п. Здесь употребительными в собственном смысле являются не слова, а предметы, понятия, представления, далеко не всегда называемые прямо и непосредственно в процессе общения. Для систематического выявления и учета соответствующих слов необходимы специальные методы, опирающиеся на анкетирование и связанные с построением лексико-понятийных полей. Такие методы могут быть реализованы в объектно-характеристических базах данных как стандартные процедуры, приводящие к построению полезных лексикографических справочников, например, ассоциативных словарей, используемых затем в процедуре отбора лексики в словники. Особое значение такие конструкции имеют для учебной лексикографии.

Еще раз подчеркнем, что непременным условием широкомасштабного компьютерного лексикографирования является требование, чтобы большая часть проверок осуществлялась автоматически и, так сказать, принудительно, помимо желания лексикографа, но последний не должен быть лишен возможности контролировать результаты таких проверок и вносить необходимые коррективы.

В целом "компьютерная" процедура отбора словника для проектируемого словаря могла бы выглядеть так:

1. Нужно сформулировать и на достаточно формальном языке задать общие требования к отбору слов: количество слов, которые следует отобрать, и свойства тех слов, которые не нужно включать в результирующий список;
2. Затем нужно сформировать запрос к словарным базам данных на выдачу заголовков словарных статей, удовлетворяющих указанным условиям, который должен

исполняться с параллельной проверкой по индексам текстовых и других баз данных системы;

3. Инициировать запрос и в качестве результата получить индекс к новой (пока еще пустой) словарной базе проектируемого словаря; в этот индекс должны войти все слова, удовлетворяющие условиям поиска, из всех источников, как словарных, так и текстовых, со ссылками на ту информацию, которая при них имеется.

Позднее, при формировании новой словарной базы индекс может расширяться за счет слов, "отлавливаемых" в текстовых частях формируемых словарных статей, и сужаться путем прямого редактирования.

Возможен и иной путь формирования словника: через морфемный или основный синтезатор слов (Мальковский 1985). Для этого нужно иметь достаточно полный морфемный словарь или словарь основ и соответствующий морфологический синтезирующий процессор. Синтезатор запускается с начала (по алфавиту) и образует слова - потенциальные заголовки словарных статей, которые проверяются по словарным и текстовым базам данных на отсутствие бракующих их свойств и снабжаются ссылками на имеющуюся информацию.

Схемы словарных статей

В общем случае типовая схема словарной статьи может быть представлена деревом составляющих и описана контекстно-свободной грамматикой или набором таких грамматик (Kolodyazhnaya I985). Она состоит из заголовка (левой части) и содержания (правой части), разделенных каким-либо знаком препинания или просто пробелом. Обычно переход от заголовка к содержанию сопровождается изменением шрифта: "черное слово" набирается полужирным шрифтом и, если имеются пометы, заголовок заканчивается светлым курсивом или шрифтом прямого начертания, но меньшего кегля, чем "фоновый" шрифт содержательной части словарной статьи. Поэтому естест-

венным разделителем между заголовком и содержанием во внутреннем представлении словарной статьи может быть код изменения шрифта; во внешнем представлении ему может соответствовать какой-либо специальный знак. Дальнейшие членения также, как правило, связаны с переходами к иному шрифту, либо с применением специальных лексикографических знаков ("параллелек", ромбов и т.п.). Сочетание характеристик шрифта (начертание, насыщенность, размер кегля) со знаками препинания и лексикографическими знаками является надежным "маркером" фрагмента словарной статьи, позволяющим опознать его содержание [8]: дефиниция, оттенок значения, иллюстративный пример, производное слово, грамматическая или стилистическая помета и т.д. При известной тренировке написание схем словарных статей проектируемого словаря не представляет трудностей.

Возможен случай, когда в одном и том же словаре используются несколько различных схем словарных статей, в зависимости от части речи или иных свойств "черного слова"; обычно отклоняющуюся от стандартной имеют отсылочные или частично-отсылочные словарные статьи (впрочем, формирование отсылочной словарной статьи может быть запрограммировано как редукция полной и выполнено автоматически по внутренней явной или неявной ссылке из ссылающейся статьи).

Схемы словарных статей могут быть поданы на вход генератора структуры ("фрейма") словарных статей, формируемых во внутреннем представлении в машине. Обычно это граф, вершинами (узлами) которого являются фрагменты словарных статей, а ребрами - наименования их типов. Если речь идет о вводе уже готового словаря в словарную базу данных, то входные словарные статьи могут быть проанализированы подпрограммой генератора, называемой сканнером, для них могут быть построены внутренние структуры, в узлах которых размещаются их

---

[8] При условии такого выбора сочетаний, при котором исключается их омонимия.

фрагменты. Структуры последовательно вводимых словарных статей связываются в цепочку, образующую подструктуру макроструктуры формируемого словаря, которая также представляется графом. Поэтому операции над словарями могут быть запрограммированы как операции над графами. Если речь идет о создании нового словаря, то генерируется пустая структура для каждой вновь формируемой словарной статьи, и генератор делает попытку заполнить ее вершины наличной информацией, выбираемой из словника, словарных, текстовых и других баз лексикографических данных. Если такая информация отсутствует, генератор задает вопрос лексикографу и получает информацию от него или оставляет вершину временно незаполненной. Так образуются "заготовки" словарных статей. На стадии редактирования лексикограф может заполнить оставшиеся пустыми вершины, или исключить их, или вообще переформировать всю словарную статью; при этом генератор словарных статей не должен допускать структур, не соответствующих запланированным схемам.

Написание словарных статей под управлением лексикографического процессора

Лексикограф может запросить лексикографический процессор приготовить в его индивидуальной модельной базе заготовки словарных статей к словам, свойства которых он может как-то задать: например, для всех союзов, междометий, предлогов, наречий времени, одушевленных существительных, многократных глаголов и т.п. Просмотрев на терминале или на распечатке заготовки статей, лексикограф может решить, оставить ли для них предварительно определенную схему или определить новую. Этот прием уместен в случае, когда целевой словарь - двуязычный, а исходный материал отобран по одноязычному (толковому) словарю. Кроме того, лексикограф может исключить часть исходного материала, если он заведомо знает, что он ему не потребуется, например,

лишние контексты, полученные из текстовой базы, лишние фрагменты словарных статей (при подготовке сокращенного варианта словаря) и т.п. Очищенный и преобразованный к следующему приближению к целевому, подготавливаемый микрословарь можно переместить в новую модельную персональную базу, на которой лексикограф может вести собственно редактирование/написание словарных статей. Последовательно, идя по структурной схеме словарной статьи сверху вниз, лексикографический процессор запросит от лексикографа ввод, коррекцию или согласие на представленный вариант очередного фрагмента, например, так:

```
ИМЯ                                    ДОРОГ/ОЙ,
   КЛАСС                               прил. IUб
   КР. ФОРМА
      МУЖ.                             м. дорог,
      ЖЕН.                             ж. дорога,
      СР.                              ср. дорого,
      СРАВН.                           I дороже,
      ПРЕВОСХ.                         I нет,
ОПИСАНИЕ
     #I.0 ЗНАЧЕНИЕ
          ТОЛК. Такой, который стоит больших денег,
имеет высокую цену.
          СИН. дорогостоящий , ценный, драгоценный,
 бесценный %книжн.% .
          АНТ. дешевый.
          СОЧЕТ. I. Д. вещь (товар, мебель, подарок,
машинаБ удовольствие, жизнь...)
          СОЧЕТ. 2. %%Дорог%% для отца.
          СОЧЕТ. 3. Баснословно... %%дорогой%%.
     #I.I ЗНАЧЕНИЕ.....
РАЗЕОЛ.
          %%Дорогой ценой%% - см. %%цена:%%
          %%Мал золотник да дорог%% %погов.%
-см. %малый%
.....
```

(Отступы соответствуют структурному членению словарной статьи, которое здесь в целях экономии места проведено не полностью.)

Слова, напечатанные прописными буквами, предъявляет лексикографический процессор. Это - наименования фрагментов словарной статьи, взятые из ее структурной схемы. Слова, напечатанные строчными буквами, выводятся лексикографическим процессором из словарной базы или, если соответствующая информация в ней отсутствует, вносятся лексикографом непосредственно с терминала. Знаки препинания и коды шрифтовых выделений генерируются процессором по схеме словарной статьи. В момент предъявления очередного фрагмента лексикограф может, запросив в словарных базах другие описания этого же слова, а в текстовой - примеры, использовать полученные данные для формирования статьи (конечно, не переписывая их, а только указывая, откуда их взять и куда вставить).

Возможен также вариант работы лексикографа под контролем процессора. В этом случае лексикограф просто пишет очередную статью, может быть, запрашивая комплектацию каких-либо фрагментов из имеющихся источников или программными средствами (например, автоматическое транскрибирование, генерацию морфологических помет), а лексикографический процессор следит за тем, чтобы не возникало нарушений схем словарных статей, и сигнализирует, если нарушения имели место, а также выполняет свою фоновую работу.

Такой метод работы позволяет разбить словарь на тематически и/или семантически однородные группы словарных статей, предоставить каждому члену авторского коллектива отдельную модельную базу и проводить параллельную работу по всему массиву; лексикографический процессор должен быть способен собрать все частичные результаты в новую модельную базу, например, главного редактора, который может организовать на ней окончательное редактирование словаря. Работа

в вычислительной сети позволит держать авторский коллектив не обязательно в одном месте, увеличивать по мере необходимости число авторов, организовывать конкурсное написание статей и т.д.

Внесение в тексты словарных статей транскрипции, грамматических помет, соблюдение правил расстановки и формы других помет, лексикографических знаков и управляющих кодов шрифтовыделений, расстановка и проверка ссылок, формирование ссылочных статей, проверка наличия в словнике слов, встречающихся в тексте словарной статьи, внесение таких слов в словник и генерация для них новых словарных статей, проверка цитат и указаний источников или их стандартизованное формирование, проверка идентичности описания одной и той же фразеологической единицы или многокомпонентных терминов в нескольких словарных статьях (или соблюдение принципа однократности описания таких единиц и замещение других описаний отсылками), выверка различных данных (реалий, терминов по стандартам, имен собственных, наименований жителей по местности и лиц по профессиям), разбивка "черных слов" на слоги, внесение в словник морфем и наименований лингвистических понятий, использованных в аппарате словаря и т.п. - вся такая работа может быть запрограммирована и, следовательно, может выполняться лексикографическим процессором. (Создание лингвистических процессоров позволит пойти еще дальше - по формальному семантическому описанию слова генерировать дефиниции, толкования, определения и речения, используя только словник проектируемого словаря.)

Использование баз данных, сформированных на основе имеющихся словарей синонимов, антонимов, омонимов, паронимов, сокращений, имен собственных и т.д. дает возможность при подготовке новых словарей снабжать их соответствующей информацией наиболее последовательным и полным образом, но позволяет решать также

и обратную задачу - вносить своевременно в такие справочники информацию из вновь создаваемых словарей. Собственно нет особой нужды в ведении отдельных баз данных для каждого из таких частичных лексикографических источников: вся их информация может быть сосредоточена в одном источнике - в Генеральном словнике. Генеральный словник может быть сам конструктивно оформлен как словарная база данных; ее модулем должен являться лексико-семантический вариант слова, информация к которому должна содержать параметры для порождения словарных статей как можно большего класса словарей.

Издательское редактирование и полиграфическое исполнение словаря

Собственно издательское редактирование по своей логике не должно отличаться от авторской работы по написанию/редактированию словаря. Здесь только смещен акцент в сторону контрольных функций, но если последние по большинству лексикографических операций выполняет компьютер, то что остается на долю издательского редактора? Он или становится не нужен, или, наоборот, может стать главным действующим лицом - заказчиком словаря. Но если так, то работа над "рутинными", "коммерческими" (в противоположность новаторским и академическим) словарями может быть организована принципиально иным, чем это сейчас принято, образом: вся подготовительная работа - разработка проекта словаря, подготовка стартового корпуса и источниковых баз данных - может быть проведена в издательстве, исходный материал выдан авторам в машинно-читаемой форме, в этой же форме результат получен в издательстве, здесь проверен и доработан. Это возможно, если издательство и авторы работают в одной и той же вычислительной среде - на одной и той же вычислительной технике, оснащенной одной и той же автоматизированной лексикографической системой.

После завершения процесса написания/редактирования словаря вся оставшаяся работа может быть осуществлена полностью автоматически: программы коррекции орфографии совместно с программами автоматического набора и верстки перерабатывают авторский корпус словаря в полиграфический корпус, воспринимаемый фото- и электроннонаборным оборудованием, который и выдается издательством в типографию для тиражирования. При надлежащем качестве работы все процессы, связанные с перепечаткой, считкой, держанием корректур, могут и должны быть исключены.

Подготовка новых (исправленных и дополненных) изданий словарей - это упрощенный вариант подготовки нового словаря, когда лексикографический процессор воспроизводит последовательно словарные статьи, а от автора/редактора требуется, если необходимо, внести в них локальные изменения. Источником данных для изменения могут быть авторские (редакторские) наблюдения, собранные в их персональных базах, или внутрисистемные данные: в промежутке между двумя изданиями могли накопиться новые тексты в текстовых базах, новые словари или изменения в старых словарных базах (выявлены новые слова, новые значения и особенности употребления и т.п.). Лексикографический процессор либо сразу обнаруживает изменения в корпусе переиздаваемого словаря, либо получает соответствующие данные из тех же внутрисистемных, но уже обновленных источников, что и при его создании.

При описанной постановке дела технический редактор должен принимать участие в создании словаря на правах его соавтора уже на самой ранней стадии его проектирования - на уровне разработки формализованного проекта, ибо в нем должны закладываться все технические параметры будущей книги: формат полосы, выбор шрифтов и прочие детали оформления.

Поскольку подготовка полиграфического корпуса - процесс, хотя и вполне формализуемый, но все-таки отно-

сительно дорогой по времени исполнения, целесообразно иметь специальные полиграфические базы данных, в которых организуется неполосное хранение будущей книги. В этом случае возможны локальные исправления непосредственно в полиграфическом корпусе без повторной его генерации. Так при стереотипных изданиях могут исправляться опечатки и технические погрешности.

В недалеком будущем, видимо, найдут применение и некнижные виды изданий словарно-энциклопедической и другой справочной литературы. Словари на оптических и гибких магнитных дисках, подключаемых к персональным компьютерам, будут тиражироваться в виде словарных баз данных или аналогов книг, моделируемых программными средствами. Компьютеры будут оснащаться синтезаторами и анализаторами речи, что открывает совершенно новые возможности для подготовки и издания учебной литературы. Такие издания со временем станут дешевле книжных и в некоторых случаях (специальные словари для служебного пользования, многотомные словарные издания) вытеснят последние с рынка. Словари общего назначения в электронном исполнении будут использоваться в информационных и экспертных системах, везде, где имеет место автоматическая обработка данных на естественном языке, и прежде всего в редакционно-издательской деятельности.

Поэтому широкомасштабная компьютеризация словарных работ, поддержанная ускоренным развитием вычислительной лингвистики, созданием разнообразных лингвистических процессоров, является одним из заметных и актуальных направлений ускорения научно-технического прогресса.

ЛИТЕРАТУРА

Автоматизированная система ведения информационных языков. Информационные материалы. ВНИИКИ. - М., 1982.- Вып. 4.

Алексеев П.М. Статистическая лексикография. - Л., 1975.

Андрющенко В.М. К вопросу об использовании коэффициента стабильности в качестве меры употребительности // Исследования в области вычислительной лингвистики и лингвостатистики. - М.: изд-во МГУ, 1978.

Андрющенко В.М. Вычислительная лексикография, ее возможности и перспективы // ВЯ, 1986. - № 3.

Громов Г.Р. Автоформализация профессиональных знаний // Микропроцессорные средства и системы. - 1986. - № 3.

Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. - М.: Наука, 1981. - С. 74-136.

Мальковский М.Г. Диалог с системой искусственного интеллекта. - М.: изд-во МГУ, 1985.

Морковкин В.В. О базовом лексикографическом знании // Учебники и словари в системе средств обучения русскому языку как иностранному. - М.: Русский язык, 1986.

Морковкин В.В. Об объеме и содержании понятия "теоретическая лексикография" (в печати).

Сороколетов Ф.П. Традиции русской советской лексикографии // ВЯ, 1978. - № 3.

Hausmann F.J. Einführung in die Benutzung der neufranzösischen Wörterbücher. - Tübingen, 1977.

Knowles F. Towards the machine dictionary // Lexocography: Principles and Practice / Ed. by R.R.K.Hartmann. - London: Academic Press, 1983.

Kolodyazhnaya L.I. Linguistis software for an automatic dictionary of common language // Symposium on automatic compilation of dictionaries (Tallinn, November 25-27, 1985). Summaries. - Tallinn, 1985.

Landau S.I. Dictionaries: The Art and Craft of Lexicography. - New York: The Scribner Press, 1984.

Micheá R. Mots fréquents et mots disponibles // Les Langues Modernes. - Vol. XLVII. - No. 4, 1953.

Л.И.Колодяжная

АВТОМАТИЧЕСКИЙ ГЕНЕРАЛЬНЫЙ СЛОВНИК РУССКИХ СЛОВАРЕЙ (методика создания, предварительные исследования, первые результаты)

I. Автоматический Генеральный словник (АГС) как подфонд Машинного фонда русского языка

Начало работ по созданию Машинного фонда русского языка можно отнести ко времени первой конференции, состоявшейся в феврале 1983 года. За это время определились основные концепции проекта, состав подфондов, организации-исполнители, этапы работы и сроки исполнения.

Автоматический Генеральный словник русских словарей относится к одному из более чем 20 подфондов Машинного фонда. Идея создания такого словника высказана в статье В.М.Андрющенко "Концепция и архитектура Машинного фонда русского языка": "Генеральный словник русских словарей... может быть создан как семейство однородных баз данных на основе Сводного словника, созданного в Словарном секторе Института русского языка ЛО АН СССР, и других "словнико-подобных" словарей, таких, как Грамматический словарь русского языка А.А.Зализняка, Орфографический словарь и др." (Андрющенко 1986, 26-44).

Функции АГС, с одной стороны, должны быть аналогичны функциям автоматических филологических словарей (см., например, Андрющенко 1981, 71-89), таким, как возможность работы со словарем в диалоговом режиме; проекции словаря по заданному списку элементов словарной статьи; выборка слов из словаря по заданным признакам; вычисление количественных параметров словаря; редактирование статей и пополнение словаря в диалоговом режиме. Очевидна аналогия подобных задач вообще задачам над какой-либо однородной совокупностью данных, представленным в виде базы данных на ЭВМ. Но такой агрегат данных, как словарная статья, имеет свои структурные особенности, поэтому разработка специализированных словарных СУБД имеет самостоятельное значение.

С другой стороны, поскольку АГС проектируется как система автоматических словарей "словникового" типа, наряду с перечисленными выше задачами могут возникать и самостоятельные задачи для каждого из составляющих словарей.

Для создания АГС как системы баз данных лексикографического типа (основная единица данных - словарная статья), необходимо иметь соответствующее программное обеспечение (СУБД) и сами данные: словари-источники на машинных носителях. АГС предполагается создавать в несколько этапов. На первом этапе (до 1990) предполагается создать АГС из словарей "словникового" типа, уже имеющихся в настоящее время на машинных носителях. В качестве программного обеспечения для создания и исследования АГС на первом этапе будет использован пакет программ словарно-ориентированной подсистемы системы УНИЛЕКС-2 (Колодяжная 1987а), выполняющей многие из перечисленных выше функций Автоматического словаря.

2. Принципы создания и исследования Автоматического филологического словаря в системе УНИЛЕКС-2

Словосочетание "Автоматический филологический словарь" будем считать синонимом таким словосочетаниям, как "словарная база данных" или "машинная версия филологического словаря". Под базой словарных данных мы традиционно понимаем, используя терминологию, принятую в теории баз данных, поименованную совокупность словарных статей, хранящихся во внешней памяти ЭВМ в таком виде, чтобы были возможны различные операции преобразования информации. Для того, чтобы осуществить такую возможность, необходимо представить каждую словарную статью в машинной базе в виде "агрегата данных", т.е. поименованной совокупности элементов и групп элементов, между которыми установлены определенные отношения.

В автоматизированной лексикографической системе УНИЛЕКС-2 разработана следующая модель структуры словарной статьи, учитывающая такие общие закономерности словарных статей различных словарей, как факультативность

элементов и групп элементов. В отличие от коммерческих данных, словарные статьи филологических словарей характеризуются существенной неоднородностью информационного состава. Это приводит к определенным трудностям при описании на формальном языке структурной схемы словарной статьи..

В системе УНИЛЕКС-2 схема словарной статьи (модель данных) имеет общие черты для различных филологических словарей и описывается в правилах подстановочной КС-грамматики. В словарной статье выделяются элементарные компоненты и составные. Из элементарных компонент составные строятся так, что каждая компонента данного уровня состоит из двух соседних компонент более низкого уровня. В исходном (например, книжном) тексте статьи либо выделяются, либо вносятся (при переносе текста на машинный носитель или программным путем) специальные символы-разделители, соответствующие иерархии компонент словарной статьи. В правилах формальной грамматики, описывающей схему статьи, нетерминальными символами считаются имена компонент (условные), терминальными - символы-разделители и значения элементарных компонент; начальным символом - условное имя всей статьи. При таком подходе, независимо от словаря, в формальной грамматике выделяются правила трех типов: обязательного вхождения одной компоненты в другую, факультативного и рекурсивного (примеры формального описания словарных статей конкретных словарей см. Колодяжная 1987б, 1987в, 1986).

Такой способ описания словарных данных позволяет, используя пакет программ системы УНИЛЕКС-2, создавать машинную базу филологического словаря и проводить его исследование. Операции преобразования информации, хранящейся в словарной базе, разделяются на основные (базовые) и сервисные. Основные операции позволяют работать с автоматическим словарем в диалоговом режиме, просматривая информацию каждой статьи (запрос - заглавное слово статьи), или конструируя новую статью, задавая список имен компонент базовой статьи в нужном порядке. К основным

операциям относятся также создание произвольной проекции словаря (по заданному списку имен компонент словарной статьи); выборка и печать подсловаря по заданным признакам, присутствующим в словарной статье; вычисление различных количественных параметров словаря (примеры реализации таких задач приведены в Колодяжная 1987а).

3. Словари-источники АГС

Согласно техническому заданию на создание Машинного фонда русского языка, в качестве поисковых признаков, хранящихся при каждой вокабуле в АГС, должны присутствовать следующие: признаки ударения и словоизменения, данные о произношении, признаки лексико-грамматической вариативности, статистические данные об употребительности, сведения об ограничении употребления, данные о морфемном составе и месте в словообразовательном гнезде, данные о лексико-семантической вариативности и т.п.

В связи с этим, система однородных словарных баз АГС должна создаваться из следующих словарей "словникового" типа: Сводный словник русских словарей под ред. Р.П.Рогожниковой, Грамматический словарь русского языка А.А.Зализняка, Словообразовательный словарь русского языка А.Н.Тихонова, Орфографический словарь русского языка, Частотный словарь русской разговорной речи, словарь "Слитно или раздельно?" Б.З.Букчиной, Л.П.Калакуцкой, словарь-справочник "Управление в русском языке" Д.Э.Розенталя, Словарь сокращений русского языка Д.И.Алексеева и др., словарь-справочник "Прописная или строчная?" Д.Э.Розенталя, словарь-справочник "Трудные случаи употребления однокоренных слов русского языка" Д.А.Бельчикова, М.С.Панюшевой, Словарь сочетаний, эквивалентных слову, Р.П.Рогожниковой.

Естественно, что при таком большом количестве словарей-источников АГС должен создаваться в несколько этапов. На первом этапе АГС будет состоять из тех словарей, которые уже заведены на машинные носители. К этим словарям

относятся: Грамматический словарь русского языка А.А.Зализняка (Зализняк 1977), Словообразовательный словарь русского языка А.Н.Тихонова (Тихонов 1985), словарь "Слитно или раздельно?" (Букчина, Калакуцкая 1985). Частотный словарь русской разговорной речи, созданный в МГУ (Овсиенко 1986). Кроме того, в порядке эксперимента, в АГС первой очереди будет включен начальный фрагмент (статьи на букву "А") Сводного словника русских словарей.

Для возможности создания словарной базы, для каждого из словарей системы АГС необходимо составить формальное описание структуры его словарной статьи. Эта работа сопряжена со значительными трудностями. Как правило, чтобы определить состав и порядок элементарных и составных компонент словарной статьи, недостаточно бывает информации из раздела "Описание словаря", включенного в любой книжный словарь. В таких описаниях не всегда четко определено, какие элементы словарной статьи обязательны, а какие факультативны, поэтому для составления адекватного формального описания практически требуется визуально просмотреть весь словарь и выбрать статьи с различным составом (различной информационной наполненностью правой части).

Несмотря на то, что АГС должен строиться из словарей "словникового" типа, т.е. таких, у которых правая часть менее развита, чем, например, у толкового словаря, для адекватного фиксирования различных признаков слова и для таких словарей должна быть проделана соответствующая работа, связанная с составлением формального описания словарной статьи. В настоящий момент такая работа проделана для Грамматического словаря и Сводного словника русских словарей. Подробнее об этом будет сказано ниже.

Необходимо сделать еще следующее замечание, касающееся первичной обработки словарей-источников. В настоящее время реально существуют два способа переноса текста словарей-источников (книга или рукопись) на машинный носитель (магнитная лента, диск): ручной ввод (флоппи-

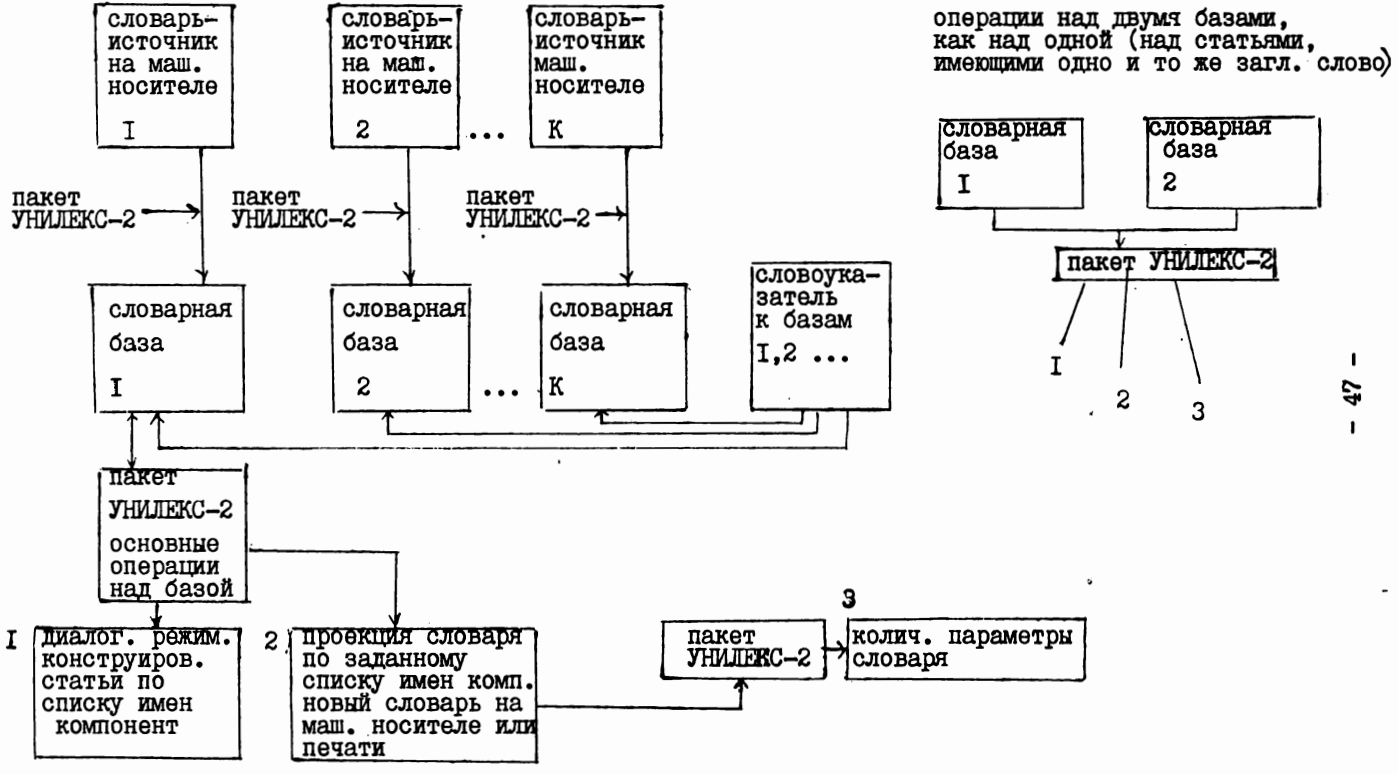
диски, дисплей) или перфолента (магнитная лента) при издании (переиздании) словаря. При ручном вводе возникает проблема замены некоторых книжных символов на те, что имеются на клавиатуре устройства. В тексте словаря на магнитной ленте (перфоленте), полученной из издательства, имеются дополнительные символы (перехода к другому шрифту). Поскольку формальное описание статьи, принятое в системе УНИЛЕКС-2, включает иерархические разделители, то перед созданием базы данного словаря может возникнуть необходимость написания интерфейсной программы, приводящей исходную разметку текста к разметке системы УНИЛЕКС-2.

4. Концептуальная схема АГС первой очереди

Создание и функционирование АГС первой очереди является в какой-то степени экспериментом (хотя и ожидается уже на этом этапе реальное исследование по какому-либо словарю). Это связано и с тем обстоятельством, что организации-исполнители (ИРЯ АН СССР и НИВЦ МГУ) не обладают пока достаточно мощными ЭВМ для возможности работы сразу со многими словарями системы АГС. Программы создания и исследования автоматического филологического словаря системы УНИЛЕКС-2 позволяют создавать словарные базы на ЭВМ типа ЕС-1022 и выше. Максимальная память, требующаяся для работы программ - 200К; объем словарной базы на диске зависит как от объема статей данного словаря, так и от сложности их структуры. Из опыта нашей работы по созданию базы Грамматического словаря А.А.Зализняка по предварительным оценкам, для базы из 100000 статей этого словаря потребуется 1,5 29-ти мегабайтных диска.

Архитектура АГС первой очереди и хранение его на внешних устройствах планируется так, чтобы была обеспечена возможность проведения операций как над каждым словарем, так и над каждыми двумя. Для удобства пакетной обработки базы больших словарей (более 10000 статей) предполагается делать составными, для того, чтобы ка-

СХЕМА АГС ПЕРВОЙ ОЧЕРЕДИ



кую-либо операцию над частью словаря можно было производить за один выход на ЭВМ (3-6 часов непрерывной работы). На рис. I представлена схема АГС первой очереди. Обозначены словари-источники, их базы, указаны основные операции над ними и взаимодействие данных с программами УНИЛЕКС-2.

5. Формальное описание словарной статьи Сводного словника русских словарей

Уже до начала проведения работ по созданию Машинного фонда русского языка, а следовательно, и идеи создания АГС, в словарном секторе ЛО ИРЯ АН СССР были начаты работы по созданию Сводного словника русских словарей под руководством д.филол. н. Р.П.Рогожниковой. В настоящее время Сводный словник, состоящий из более чем 163000 словарных статей, сдан в печать. Наиболее подробная информация о структуре статей Сводного словника и целях его создания изложена в работе Р.П.Рогожниковой "Сводный словник словарей и словоуказатели в словарной работе" (Рогожникова 1984).

Основная информация при слове в статье Сводного словника - это указание словарей, в которых содержится данное слово (всего выбрано 12 наиболее представительных русских словарей). Кроме того, в Сводном словнике фиксируются варианты написания слова, варианты ударения, омонимия; могут встречаться пометы при слове, указывающие на часть речи или значение. Одной из основных целей составления такого справочника является фиксация слов, не вошедших в словари русского языка (путем сопоставления различных словников со Сводным словником), а также изучение различных вариантов слова и выработка нормы (орфографии, произношения и т.п.). Очевидно, что создание машинной версии такого справочника позволит решать различные исследовательские задачи на ЭВМ. В частности, в докладе Р.П.Рогожниковой на 2-ой Всесоюзной конференции по Машинному фонду было отмечено: "Генеральный словник... позволит производить различные сопоставительные работы

со словниками словарей, словоуказателями текстов и т.п. И сам он, поскольку является сводом лексических единиц из различных словарей, имеет самостоятельное значение и может служить основой для отбора лексики как в словари общего, так и учебного типа" (Рогожникова 1987, 146).

Очевидно, что по своей актуальности именно Сводный словник должен быть центральным компонентом системы АГС. Пока мы не располагаем машинной копией этого справочника. Но с целью отработки методики автоматического решения некоторых исследовательских задач, решено создать машинную версию начального фрагмента словаря (статьи на букву "А"). Работа ведется под руководством Р.П.Рогожниковой в НИЦ МГУ.

В статье Р.П.Рогожниковой (1984) приводятся различные структурные типы статей Сводного словника, ориентируясь на которые можно составить формальную грамматику структуры статьи на языке системы УНИЛЕКС-2. В статье данного словаря были выделены следующие элементарные компоненты: заглавное слово, вариант заглавного слова, помета при заглавном слове, названия словаря, отсылка к другой статье. Ниже приведены примеры некоторых статей.

Изото́п Б ОЖ ОРФ
(Заглавное слово и названия словарей).

Бакенба́рд, бакенба́рда У
(Заглавное слово, вариант заглавного слова, название словаря).

Е́врика см. Э́врика
(Заглавное слово, отсылка к другой статье).

I. Кабачо́к (ум.-ласк. к I. Каба́к) у Б М ОЖ М-2
(Заглавное слово, помета при заглавном слове, названия словарей).

Епитрахи́ль (м.), эпитрахи́ль Б
(Заглавное слово, помета при заглавном слове, вариант заглавного слова, название словаря).

Кроме элементарных компонент, для написания формальной грамматики в статье были выделены составные компоненты. Ниже приведены условные имена компонент статьи, формальная грамматика со списком иерархических разделителей и машинные распечатки, являющиеся результатом работы программы автоматического анализа словарной статьи (построение дерева вывода, хранящегося в машинной базе).

Элементарные компоненты

ЗАГЛ - заглавное слово
ВАР - вариант заглавного слова
ПОМЗ - помета к заглавному слову
НАЗВ - название словаря
ОТС - отсылка к другой статье

Составные компоненты

СТАТС - статья сводного словника
ЗОНЗО - зона заглавного слова с отсылкой
ЗОНАЗ - зона заглавного слова
ЗОНАН - зона названий словарей
ЗОНАВ - зона вариантов заглавного слова
ЗАГП - заглавное слово с пометой

Грамматика статьи сводного словника

#∂ * ¤ ℓ

СТАТС=ЗОНЗО∂ЗОНАН
СТАТС=ЗОНЗО
ЗОНАН=НАЗВ*ЗОНАН
ЗОНАН=НАЗВ
ЗОНЗО=ЗОНАЗ*ОТС
ЗОНЗО=ЗОНАЗ
ЗОНАЗ=ЗАГП#ЗОНАВ
ЗОНАЗ=ЗАГП
ЗОНАВ=ВАРℓЗОНАВ
ЗОНАВ=ВАР
ЗАГП=ЗАГЛℓПОМЗ
ЗАГП=ЗАГЛ

БАКЕНБА´РД ☒ БАКЕНБА´РДА ∂ у #

Список значений элементарных лексических единиц

ЗАГЛ БАКЕНБА´РД
ВАР БАКЕНБА´РДА
НАЗВ у

КАБАЧО´К. I ♪ (ум.-ласк. к I. КАБАК) ∂ у * Б * М * ОЖ * М-2 #

Список значений элементарных лексических единиц

ЗАГЛ КАБАЧО´К. I
ПОМЗ (ум.-ласк. к I. КАБАК)
НАЗВ у
НАЗВ Б
НАЗВ М
НАЗВ ОЖ
НАЗВ М-2

6. Создание словарной базы Грамматического словаря русского языка А.А.Зализняка как компонента АГС

Машинная копия Грамматического словаря в последнее десятилетие активно используется, например, в работах по машинному переводу (Кулагина 1986) и искусственному интеллекту (Мальковский 1985). В этих работах информация, данная в словаре, преобразуется к той или иной модели морфологического анализа.

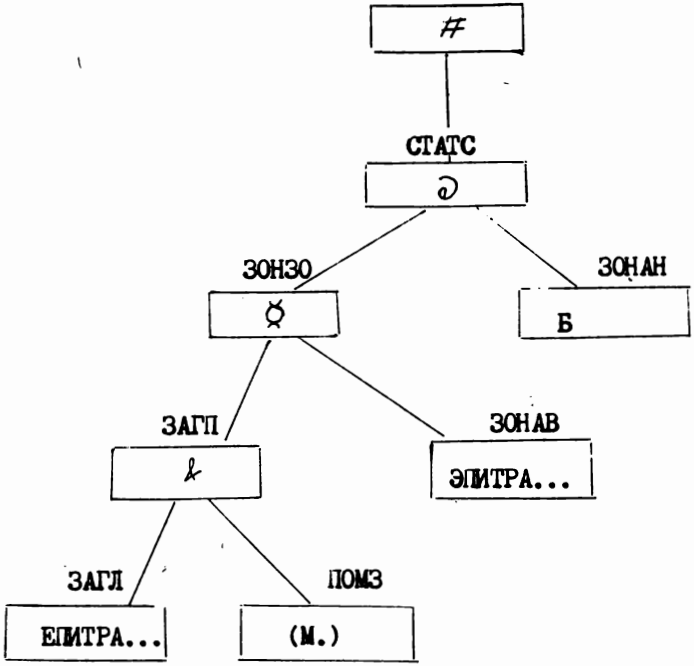
Функционирование Грамматического словаря в системе АГС не предполагает модификации информации, содержащейся в его статьях, а лишь использование этой информации как признаковой для того или иного слова. Машинная копия словаря А.А.Зализняка, созданная в Лаборатории по применению вычислительной техники в гуманитарных науках НИЦ МГУ (работа проводилась в начале 80-х годов под руководством Б.М.Андрющенко), отличается от книжного варианта, во-первых, заменой некоторых типографских знаков на символы, имеющееся на алфавитно-цифровом

Текст размеченной словарной статьи

ЕПИТРАХИ′ЛЬ ℓ(М.) ♂ ЭПИТРАХИ′ЛЬ ∂ Б #

Дерево синтаксического вывода:

Ширина дерева - 42



устройстве подготовки данных (таблица замен дана в приложении); во-вторых, каждая статья дополнена заглавным словом, расписанным по морфемному составу (эта работа описана В.П.Зайончковской - Зайончковская 1987).

Формальное описание структуры статьи Грамматического словаря на языке системы УНИЛЕКС-2

Структура словарной статьи этого словаря достаточно подробно дана в описании, предшествующем словарным статьям. Выбирая элементарные компоненты для формального описания, мы, естественно, включили все составные элементы, указанные в описании. Кроме элементарных компонент, указанных в книге, мы ввели зону, в которой хранится морфологическое разбиение заглавного слова, и пустую зону. Введение пустой зоны было необходимо по следующим причинам. Статьи Грамматического словаря по структуре можно разделить на статьи "регулярного" (простого) типа, состоящие только из заглавного слова, основного буквенного символа, индекса, зоны с морфологическим разбиением, и статьи "нерегулярного" типа, в которых, кроме перечисленных элементарных компонент, могут присутствовать, например, такие, как: вариант заглавного слова, вариант основного буквенного символа ограничения на словосочетания и т.п. Для написания грамматики статьи на языке УНИЛЕКС-2, необходимо знать последовательность расположения элементарных компонент. Но в Грамматическом словаре все компоненты, кроме заглавного слова и основного буквенного символа, факультативны. Кроме того, нет такой статьи нерегулярного типа, в которой присутствовали бы все возможные элементарные компоненты. "Пустая" зона введена для того, чтобы можно было в рамках одного формального описания учесть все возможные типы нерегулярных статей. "Пустая" зона присутствует только в нерегулярных статьях и располагается между регулярной частью статьи (заглавное слово, основной буквенный символ, индекс) и дополнительной информацией.

В статье машинной копии Грамматического словаря присутствуют все возможные символы, имеющиеся на алфавитно-цифровом дисплее (кроме некоторых букв латинского алфавита). Для описания на языке УНИЛЕКС-2 необходимо в исходный текст внести иерархические разделители, не присутствующие в тексте. Такими разделителями для словаря А.А.Зализняка были выбраны буквы латинского алфавита, не совпадающие по начертанию с русскими.

Ниже приведены условные обозначения элементарных и составных компонент для статьи Грамматического словаря, формальная грамматика и примеры работы программы автоматического анализа (разбор по элементарным компонентам и дерево вывода, хранящееся в базе) для некоторых статей.

Элементарные компоненты

ЗАГЛ - заглавное слово
ОБС - основной буквенный символ
ИНД - индекс
УКЗН - указание значения
ГЛПВ - глагол противоположного вида
НЕРЕГ - морфологически нерегулярная форма
ОТКЛ - отклонение (в словосоч.) загл. слова от норм. вида
ПОБС - помета перед основным буквенным символом
ОСОЧ - ограничение на словосочетание
ВГРАМ - вариант грамматической формулы
ВОБС - вариант основного буквенного символа
ВИНД - вариант индекса
ВЗАГЛ - вариант заглавного слова
ФЭГР - факультативный элемент грамматич. части
НП2 - наличие 2-гс предл. падежа
НР2 - наличие 2-го родит. падежа
МОРФ - разбиение слова на морфемы
ПУСТ - пустая зона

Последовательность расположения элементарных компонент

ЗАГЛ ПОБС ОСБ ВОБС ИНД БИНД ФЭГР НП2 НР2 ПУСТ ВГРАМ НЕРЕГ ВЗАГЛ ГЛПВ УКЗН ОСОЧ ОТКЛ МОРФ

Составные компоненты

СТАТЗ - вся статья
ЗЗНАЧ - зона значений
ЗІЗН - зона одного значения
ДОПГР - дополнительная грамматическая зона
ЗБСИН - зона букв. символа и индекса
ЗЗОБС - зона заглавного сл. и осн. букв. символа
ЗИНД - зона индекса
БР2 - зона индекса без НР2
БП2 - зона БП2 без НП2
ИНДВ - индекс с вариантом
ЗОБС - зона осн. букв. символа
ПОМББ - помета и букв. символ
БОТКЛ - зона "ДОПГР" без ОТКЛ
БОСОЧ - зона "БОТКЛ" без ОСОЧ
БУКЗН - зона "БОСОЧ" без УКЗН
БГЛПВ - зона "БУКЗН" без ГЛПВ
БВЗАГ - зона "БГЛПВ" без ВЗАГЛ

Грамматика статьи словаря А.Зализняка

Z W Y U R Q N L G

СТАТЗ=ЗЗНАЧ Z МОРФ
ЗЗНАЧ=ЗІЗН W ЗЗНАЧ
ЗЗНАЧ=ЗІЗН
ЗІЗН=ГРАМЗ Y ДОПГР
ЗІЗН=ГРАМЗ
ГРАМЗ=ЗБСИН U ВГРАМ
ГРАМЗ=ЗБСИН
ЗБСИН=ЗЗОБСRЗИНД
ЗБСИН=ЗЗОБС
ЗИНД=БР2 Q НР2
ЗИНД=БР2
БР2=БП2 N НП2
БР2=БП2
БП2=ИНДВ L ФЭГР
БП2=ИНДВ
ИНДВ=ИНД G ВИНД

ИНДВ=ИНД
ЗЗОБС=ЗАГЛ Q ЗОБС
ЗЗОБС=ЗОБС
ЗОБС=ПОМББ N ВОБС
ЗОБС=ПОМББ
ПОМББ=ПОБС L ОБС
ПОМББ=ОБС
ДОПГР=БОТКЛ U ОТКЛ
ДОПГР=БОТКЛ
БОТКЛ=БОСОЧ R ОСОЧ
БОТКЛ=БОСОЧ
БОСОЧ=БУКЗН Q УКЗН
БОСОЧ=БУКЗН
БУКЗН=БГЛПВ N ГЛПВ
БУКЗН=БГЛПВ
БГЛПВ=ББЗАГ L ВЗАГЛ
БГЛПВ=ББЗАГ
ББЗАГ=ПУСТ G НЕРЕГ
ББЗАГ=ПУСТ

Текст размеченной словарной статьи

АНТИП<ОД Q МО R IA Y Q (о человеке) W M R IA Y Q (О ПРЕДМЕТЕ) Z АНТИ✱ПОД✱ #

Список значений элементарных лексических единиц

ЗАГЛ		АНТИП<ОД
ОБС	МО	
ИНД	IA	
ПУСТ		
УКЗН		(о человеке)
ОБС	M	
ИНД	IA	
ПУСТ		
УКЗН		(о предмете)
МОРФ		АНТИ✱ПОД✱

Текст размеченной словарной статьи

АНТРАШ<А Q C R O Z АНТРЯАШ=А #

Список значений элементарных лексических единиц

ЗАГЛ АНТРАШ<А
ОБС С
ИНД О
МОРФ АНТРЯАШ=А

Текст размеченной словарной статьи

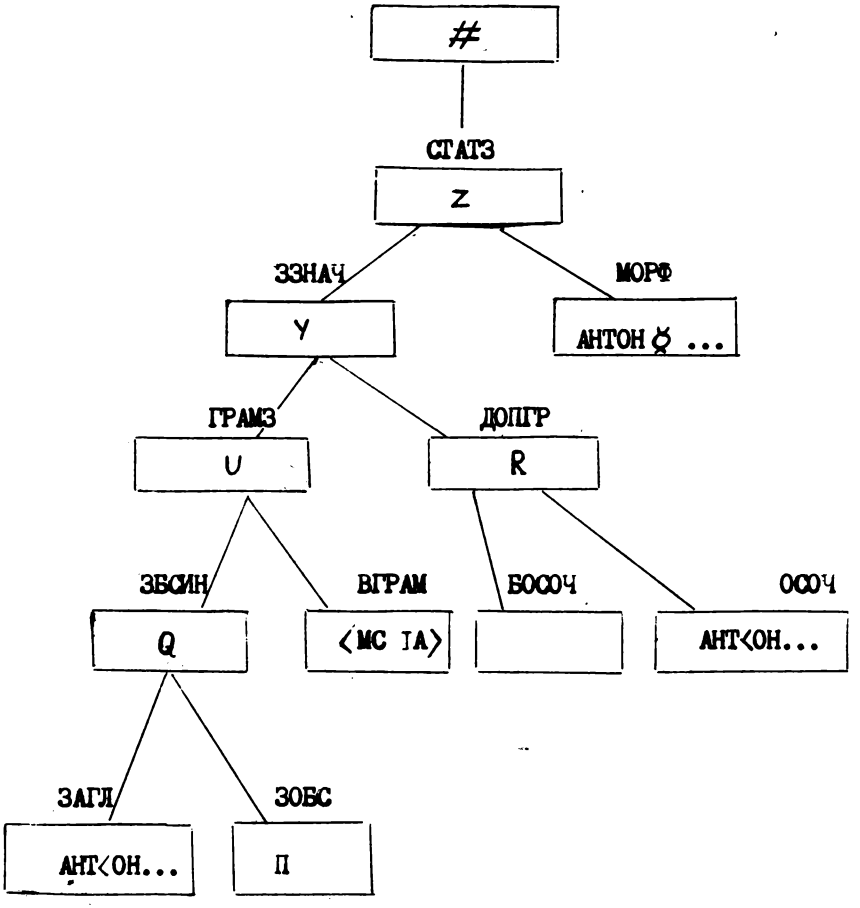
АНТ<ОНОВ Q П U <МС IA> Y R АНТ<ОНОВ ОГ<ОНЬ Z
АНТОНЯОВ=

Примеры вычисления некоторых количественных параметров фрагмента Грамматического словаря

Для создания машинной базы Грамматического словаря как компонента АГС (по методике УНИЛЕКС-2), исходная машинная копия словаря была переписана (программно) в прямом порядке по алфавиту заглавных слов. Массив исходных статей для удобства обработки в пакетном режиме был разбит на группы, включающие не более 10000 статей. Программным путем все статьи каждой группы были разделены на статьи регулярного и нерегулярного типа, при этом статьи регулярного типа автоматически размечались, согласно формальному описанию (программа разметки нерегулярных статей разрабатывается в настоящее время).

К ноябрю 87 года созданы по методике УНИЛЕКС-2 словарные базы регулярных статей для букв А, Б, В (около 9000 статей). Для словарной базы регулярных статей на букву "А" были вычислены следующие количественные параметры: распределение по частям речи; распределение по типам индекса словоизменения; распределение по типам

Дерево синтаксического вывода:



индекса внутри каждой части речи. Вычисление этих распределений проводилось при помощи программ системы УНИЛЕКС-2 (подробнее о методике см. Колодяжная 1987а). После создания полной базы предполагается вычисление этих, а также и других параметров для всего словаря. Ниже приведены соответствующие распределения.

Распределение регулярных статей на "А" по частям речи заглавных слов:

680	П
634	М
531	Ж
294	МО
96	ЖО
87	С
80	СВ-НСВ
24	НСВ НП
12	НСВ
11	МЕЖД.
5	Н
3	СВ-НСВ НП
2	СВ

Распределение регулярных статей на "А" по индексу словоизменения заглавных слов:

948	1А	7	6А	2	1А"2"	1	1А'
314	1*А	6	1*2А	2	1В-	1	2*А
307	7А	5	3*В	2	1В?	1	2С"1"
272	3А!ℓ	5	6*А	2	1	1	3*Аℓ
137	3А	4	1*А["1"]	2	2В	1	3С
121	2А	3	1А["2"]	2	3*А!ℓ	1	4А!
103	3*А	3	1В	2	3*2А		
68	8А	3	1С	2	3В		
37	5*А	3	1С"1"	2	4В-		
30	4А			1	1*А-		
27	0			1	1*А"1"		
10	5А			1	1А-		

Распределение заглавных слов регулярных статей на "А" по индексу внутри данной части речи:

254	Ж	I	М	7 А	5А
II7	Ж	I	М	I А	3С
68	Ж	I	М	8 А	3*А
37	Ж	I	М	3*А	2С"I"
36	Ж	I	М	3А	2*А
5	Ж	II	МЕЖД.	6А	
3	Ж	I70	МО	8Е	IА
2	Ж	60	МО	4В–	3А
2	Ж	35	МО	4А	5*А
2	Ж	7	МО	IВ–	2А
2	Ж	6	МО	0	I*2А
I	Ж	4	МО	I D	0
I	Ж	3	МО	5А	7А
I	Ж.	2	МО	I D	IА["2"]
64	ЖО	2	МО	3*А	3*2А
II	ЖО	I	МО	IА	3*В
8	ЖО	I	МО	5А	3*А
8	ЖО	I	МО	4А	IА"2"
5	ЖО	I	МО	7А	6А
536	М	I	МО.	IА	IА
39	М			3А	
I5	М			7А	
II	М			2А	
8	М			4А	
4	М			3*В	
3	М			IС"I"	
2	М			IС	
2	М			2В	
2	М			3В	
2	М			5*А	
2	М			IВ	
I	М			IА"2"	
I	М			IА["2"]	
I	М			6А	

5	H	
7	HCB	2A
3	HCB	4A
2	HCB	IA
I3	HCB HII	2A
9	HCB HII	IA
2	HCB HII	4A
3I4	II	I*A
272	II	3A!&
74	II	IA
6	II	4A
4	II	I*A["I"]
2	II	3*A!&
2	II	IB?
I	II	3*A&
I	II	IA′
I	II	IA–
I	II	I*A"I"
I	II	I*A–
I	II	4A!
30	C	7A
28	C	IA
2I	C	0
5	C	6*A
I	C	4A
I	C	IC
I	C	IB
2	CB	3A
80	CB-HCB	2A
3	CB-HCB HII	2A

ЛИТЕРАТУРА

В.М.Андрющенко. Автоматизированные лексикографические системы // Теоретические и прикладные аспекты вычислительной лингвистики. - М.: изд-во МГУ, 1981.

В.М.Андрющенко. Концепция и архитектура Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. - М.: Наука, 1986.

Букчина Б.З., Калакуцкая Л.П. Слитно или раздельно? - М.: Русский язык, 1985.

Зайончковская Б.П. Проблемы составления автоматического морфемного словаря на материале "Грамматического словаря русского языка" А.А.Зализняка // Вторая Всесоюзная конференция по созданию Машинного фонда русского языка. Тезисы докладов. - М., 1987.-С. 67-69.

Зализняк А.А. Грамматический словарь русского языка. - М.: Русский язык, 1977.

Колодяжная Л.И., Поликарпов А.А., Хотько Е.А. Принципы создания машинной версии словаря синонимов и исследование количественных характеристик его структуры // Техническое и алгоритмическое обеспечение информационно-преобразовательных систем. - Рязань, 1986.-С. 132-137.

Колодяжная Л.И. Автоматизированная лексикографическая система УНИЛЕКС. Словарно-ориентированная подсистема / Под ред. Н.В.Павлович. - М.: изд-во МГУ, 1987а.

Колодяжная Л.И. Возможности автоматического исследования филологического словаря на основе его машинной версии // Сочетание лингвистической и внелингвистической информации в автоматическом словаре. - Ереван: изд-во АН АССР, 1987б.-С. 156-159.

Колодяжная Л.И., Павлович Н.В. Проблемы формализации представления данных в автоматическом словаре поэтических образов. // Вторая Всесоюзная конференция по созданию Машинного фонда русского языка. Тезисы докладов. - М., 1987в.-С. 85-90.

Кулагина О.С. Морфологический анализ русских именных словоформ // Препринт № 10 Ин-та прикл. матем. им. М.В.Келдыша АН СССР. - М., 1986.-40 с.

Мальковский М.Г. Диалог с системой искусственного интеллекта. - М.: изд-во МГУ, 1985.-213 с.

Овсиенко Ю.Г. Современная разговорная речь как составная часть Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. - М.: Наука, 1986.

Рогожникова Р.П. Сводный словник словарей и слово-указатели в словарной работе // Теория и практика современной лексикографии. - Л.: Наука, 1984.

Рогожникова Р.П. Лингвистические проблемы создания Генерального словника русского языка в рамках Машинного фонда русского языка. Тезисы докладов. - М., 1987.- С. 145-148.

Тихонов А.Н. Словообразовательный словарь русского языка. - М.: Русский язык, 1985.

О.В.Загоровская
С.В.Лесников

ВИДЫ ЛЕКСИКОГРАФИЧЕСКОЙ ИНФОРМАЦИИ В АВТОМАТИЧЕСКОМ СЛОВАРЕ РУССКИХ ГОВОРОВ КОМИ АССР И СОПРЕДЕЛЬНЫХ ОБЛАСТЕЙ

Автоматический Словарь русских говоров Коми АССР и сопредельных областей (АСРГКА) представляет собой одну из составных частей Академического словарно-грамматического фонда Машинного фонда русского языка (МФ РЯ).

Методической основой создания АСРГКА являются известные положения Ю.Н.Караулова о том, что любые данные о языке могут быть представлены в лексикографической форме (а если лексикографическое воплощение некоторых подсистем языка и даже целых уровней еще слабо разработано, то это вовсе не означает, что оно принципиально недостижимо); любые лексикографированные данные о языке могут быть переведены в алгоритмизированную машинную форму (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 (Караулов 1986, 13); идеология МФ РЯ является, по существу, лексикографической, словарной; в основе способа хранения информации о слове должно лежать "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" (Караулов 1986, 14).

Формальная грамматика словарной статьи АСРГКА описана в нашей работе "Формальная грамматика словарной статьи автоматического Словаря русских говоров Коми АССР и сопредельных областей" (В печати. План 1988 г.).

В настоящей работе раскрывается конкретное наполнение основных элементарных и составных компонент словарной статьи названного автоматического регионального словаря и приводится ее иерархическая разметка.

Для наглядности описание элементарных и составных компонент словарной статьи АСРГКА предваряется схемами, содержащими графическую интерпретацию формальной грамматики словарной статьи АСРГКА с лексикографической разметкой.

Иерархическая разметка словарной статьи представляет собой предварительную специальную обработку имеющейся информации перед вводом ее в память компьютера на основе метода формализации (см., например, Гришин, Никольский 1982). При создании иерархической разметки словарной статьи АСРГКА использовались принципы, аналогичные изложенным в диссертационной работе Л.И.Колодяжной (Колодяжная 1986) и статье Ю.Д.Апресяна, И.М.Богуславского, Л.Л.Иомдина (Апресян и др. 1982). При этом учитывалась изначальная ориентация на автоматизированную лексикографическую систему UNILEX, разработанную в Лаборатории по применению вычислительной техники в гуманитарных науках НИВЦ МГУ (Андрющенко 1982) и являющуюся составной частью программного обеспечения АСРГКА.

В иерархической разметке словарной статьи АСРГКА используются обозначения # @ ℓ F J R ¬ ♂ ♀ I ж N > < +. Основой для выбора данных обозначения послужило предположение, что они не будут встречаться в самой записи элементарных компонент. Вместе с тем, следует отметить, что нарушение названного предположения не ведет к неисправимым ошибкам, т.к. в программное обеспечение АСРГКА включены средства проверки соответствия структуры размеченной словарной статьи заданной формальной грамматике.

Графическое изображение структуры словарной статьи АСРГКА дается в Приложении I.

Мнемонические обозначения компонент словарной статьи АСРГКА и сжатая форма ее иерархической разметки даны в Приложении II.

Раскрытие содержания компонент словарной статьи АСРГКА, предложенное в работе, не претендует на их строгое научное определение, т.к. с каждым научным понятием, как правило, связано целое поле исторически сложившихся значений и всякая попытка определения научного термина ведет к определенному ограничению данного поля (Налимов 1979, 139).

Описание основных компонент словарной статьи АСРГКА приводится в порядке, соответствующем значимости той или иной компоненты.

Значимость компонент словарной статьи АСРГКА тесно связана с явлением факультативности элементов составной компоненты, или зоны. В большинстве случаев факультативны правые подзоны, напр.:

СТАТ = ОСН @ ДОП

РНП = РН > НП

ОСЗНО = ОСЗГ № НОМ

Факультативность левых зон наблюдается только в составных компонентах ОТСЛ, ДОП, ФРН и ФОТС.

Отсутствие отдельных компонент или частей компонент в словарной статье определяется не только факультативностью. Оно возможно также (в силу их малой значимости) и в следующих случаях:

1. Информация отсутствует, но может быть восстановлена по предыдущей зоне или компоненте. (Так, например, может быть восстановлена информация о населенном пункте, где производился сбор материала по соответствующей зоне предыдущей иллюстрации.)

2. Информация отсутствует, но может быть восстановлена в результате последующих экспедиций или работы с экспедиционными источниками. (Так, например, может быть восстановлена информация о грамматических и стилистических особенностях диалектных словесных единиц.)

3. Информации в принципе быть не может в силу специфики определенного языкового явления (как, например, в случаях абсолютного отсутствия фразеологизмов у некоторых словесных единиц).

Ниже объясняются основные компоненты словарной статьи АСРГКА.

Составная компонента СТАТ (статья АСРГКА) представляет собой структуру, состоящую из заглавного слова и информации о его фонетических, грамматических, семантических, стилистических, сочетаемостных особенностях и фразеологических связях, вариантных отношениях и этимологии в соответствии с формальной грамматикой словарной статьи АСРГКА. См. схему I (Приложение I).

Составная компонента ОСН (основная часть словарной статьи АСРГКА) подразделяется на ЗЗГ (зона заглавного слова) и ЗДЕФ (зона дефиниций).

Составная компонента ЗГСЛ (заглавное слово) представляет собой структуру, в которую входит собственно заглавное слово словарной статьи (ОСЗГ) с номером омонима (НОМ), фонетическими (ФВАР) и словообразовательными (СВАР) вариантами, этимологическими сведениями (ЭТИМ). См. схему II. При раскрытии ОСЗГ, ОФВР (один фонетический вариант заглавного слова) и ОСВР (один словообразовательный вариант заглавного слова) используется современная русская орфография.

Составная компонента ЗДЕФ (зона дефиниций) представляет собой структуру, содержащую сведения обо всем объеме семантики слова, его основных и переносных значениях, оттенках значения. Названная составная компонента подразделяется на ДЕФ (дефиниции и ЗДЕФ. В данном случае используются рекурсивные описания, предполагающие эффективно распознаваемые структуры (о рекурсии см. Математическая энциклопедия 1986, 963).

По тому же принципу построены составные компоненты ЗОТТ (зона оттенков основного значения), ЗПРЗ (зона переносных значений), ЗФР (зона фразеологических сочетаний) и ЗИЛЛ (зона иллюстраций), которые могут состоять из соответствующей элементарной компоненты (содержащей сведения об одном оттенке основного значения, одном переносном значении, одном фразеологизме и одной иллюстрации) и составной компоненты (зоны от-

тенков основного значения, зоны переносных значений, зоны фразеологических сочетаний и зоны иллюстраций). (Структурно: 3 = 5 = 6).

Составная компонента ЗОЗН (зона одного основного значения) представляет собой структуру, содержащую информацию об основном значении слова (СЗН), его принадлежности к определенной части речи (ЧРОС), грамматических (ХАРОС) и стилистических (СТОС) характеристиках, способах определения значения, иллюстрациях (СИЛЛ), а также дополнительных сведениях этимологического, топонимического характера и разного рода комментариях. См. схему 3.

Составная компонента ПОМЗ (пометы к основному значению), содержащая информацию о грамматических и стилистических характеристиках заглавного слова, присутствует в словарной статье в том случае, когда названные характеристики являются общими для всех значений слова и их оттенков. В случае же различия грамматических и стилистических характеристик разных значений и оттенков значений заглавного слова соответствующие компоненты включаются непосредственно в зону определения каждого значения или оттенка значения.

Так, например, в словарной статье: БЛАДЕНЬЕ, -я, ср. I. Сила, здоровье. - На д'ев'ӑтом д'ес'ӑтк'е, да влад'ӗн'йа уш н'ет. 2. Способность владеть, действовать каким-либо членом тела. - Л'ежӳ, в ногах влад'ен'йа н'ет. - при заглавном слове раскрывается зона ПОМЗ.

В словарной статье: АЛИ. I. Разделит. союз - Год ӑл'и два жыл тут. 2. Присоединит. союз. - В'ел'ичай старшых ӑл'и начӑл'ство. 3. Поясним. союз. - А вот то называйец'ц'а т'епн'а ӑл'и ч'ерпӑк. 4. Усилит. частица в вопросит. предложениях. - Ӑл'и мы н'е сказывал'и? - при каждом значении дается компонента ПМОС, содержащая указания на соответствующие характеристики каждого значения заглавного слова.

Составная компонента ОИЛЛ включает в себя собственно иллюстрацию (СИЛЛ) и данные о ее источнике (ИСТ).

При этом информация, содержащаяся в СИЛЛ, приводится в упрощенной фонетической транскрипции исключительно на основе фактических полевых записей.

Составная компонента ЗОФ (зона одного фразеологизма) представляет собой структуру, в целом совпадающую с компонентой ОСН (за исключением информации об оттенках и переносных значениях фразеологизма).

ЛИТЕРАТУРА

Андрющенко Б.М. Автоматизированная лексикографическая система (Основные проектные решения) // Вычислительная лингвистика: Теоретические аспекты. Вопросы автоматизации лексикографических работ / Под ред. П.В.Демьянкова. - М., 1982.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Типы лексикографической информации в толково-комбинаторном словаре // Актуальные вопросы практической реализации систем автоматического перевода. Ч. П. - М.: изд-во МГУ, 1982.

Гришин А.В., Никольский Н.М. Системный анализ и диалог с ЭВМ в исследовании международных отношений (Глава 4, § 1. Формализация как предпосылка применения автоматизации в научных исследованиях). - М.: Международные отношения, 1982.

Караулов Ю.Н. Методология лингвистического исследования и Машинный фонд русского языка // Машинный фонд русского языка: идеи и суждения. - М.: Наука, 1986.

Колодяжная Л.И. Структура словарного текста в аспекте машинной лексикографии. Дис. ... канд. филол. наук. - М., 1986.

Математическая энциклопедия. Т. 4. - М.: Советская энциклопедия, 1986.

Налимов В.В. Вероятностная модель языка. - М.: Наука, 1979.

О.А.Казакевич

ТЕКСТОВЫЕ БАЗЫ ДАННЫХ УСТНОЙ РЕЧИ: СОЗДАНИЕ И ИСПОЛЬЗОВАНИЕ

Текстовые базы данных устной речи, или иначе, машиночитаемые корпусы текстов устной речи - явление достаточно новое, но уже прочно вошедшее в арсенал средств лингвистических исследований. Их возникновение связано с двумя направлениями современного языкознания: эмпиризмом, предполагающим ориентацию на конкретный языковой материал, изучение языковой реализации, ее описание и лишь на базе такого описания - индуктивные заключения о языковой компетенции, и установкой на преимущественное изучение устной речи как первичной формы языковой реализации, по отношению к которой письменная форма вторична.

Эмпирическое языкознание в своих исследованиях обычно опирается на корпус текстов. "Коль скоро лингвистика вообще претендует на то, чтобы считаться эмпирической наукой, она должна основываться на анализе объективно составленного и достаточно полного корпуса примеров речевых произведений носителей конкретного языка" (Henne, Rehbock 1982, 45). Корпус текстов - это некоторое множество текстов естественного языка, составленное таким образом, чтобы служить материалом для изучения тех или иных языковых аспектов.

При составлении корпуса, во-первых, возникает вопрос о том, какие типы языковой реальности должны быть в нем представлены (общелитературный язык, диалект, профессиональный язык, произведения определенного автора, устная речь и т.д.), во-вторых, вопрос о представительности корпуса, то есть о том, насколько выводы, сделанные по корпусу, могут быть распространены на представленную в нем языковую реальность в целом (Hess et al. 1983) [I].

[I] Исключение составляет исторический корпус, включающий все известные памятники некоторого мертвого языка, а также авторский корпус, включающий все произведения конкретного автора. Для таких корпусов вопроса о представительности, естественно, не возникает.

Очевидно, что объективный, представительный (и при этом обозримого объема, такой, что обработка его не вызовет особых проблем) корпус текстов может быть создан только на базе предварительных знаний объекта и области исследований (Henne, Rehbock I982). Неоднократно отмечалось, что не существует универсальных критериев оценки представительности корпуса (Bergenholtz, Schaeder I977; Bergenholtz, Schaeder I978; Schaeder I98I; Lenders, Willée I986). Всякий раз этот вопрос решается исходя из конкретных задач исследования.

Корпусоориентированные исследования не являются чем-то новым. Однако интерес к ним за последние два десятилетия заметно возрос. Не в последнюю очередь это связано с появлением машино-читаемых корпусов текстов: машино-читаемый корпус делает возможным проведение достаточно эффективного анализа большого количества лингвистических особенностей на большом объеме текстов (Biber I986 a).

Существующие в настоящее время машино-читаемые корпусы текстов можно достаточно условно подразделить на специализированные и многофункциональные. Среди специализированных основную массу составляют текстовые базы данных, предназначенные для дальнейшей лексикографической обработки: получения словоуказателей, конкордансов, различных словарей. Текстовые базы данных, входящие в лексикографические системы, должны быть достаточно большими. Крупнейшей текстовой базой данных является сейчас корпус текстов французского языка XIX-XX веков, составленный и введенный в ЭВМ в середине 60-х годов в Нанси с целью создания машинной картотеки для словаря французского языка XIX-XX веков (Gorcy I98I). Его объем - более 90 млн. словоупотреблений. Начало многофункциональным текстовым базам данных было положено знаменитым Брауновским корпусом (Francis I980). Многофункциональный корпус создается как материал для любых мыслимых лингвистических исследований. Средний объем многофункционального корпуса - I млн. словоупотреблений.

Первыми машино-читаемыми корпусами были корпусы письменных текстов, но вскоре постоянно возрастающий интерес к устной речи, с одной стороны, и успехи тексто-ориентированных исследований, опирающихся на машино-читаемые корпусы, с другой, привели к возникновению текстовых баз данных устной речи (Howes 1966; Berens 1979). В соответствии с установкой на учет в новых словарях и грамматических описаниях современных языков как устной, так и письменной формы существования речевых произведений, появляются и смешанные текстовые базы, включающие как устные, так и письменные тексты (Mentrup 1979; Quirk, Svartvik 1979). Однако до сих пор основная масса машино-читаемых корпусов представляют современные общелитературные языки в их письменной форме. Создание машино-читаемого корпуса устной речи - процесс гораздо более трудоемкий, чем аналогичная деятельность, связанная с письменными текстами: сбор материала и последующее транскрибирование магнитофонных записей - работа длительная и кропотливая. По подсчетам "отца" Брауновского корпуса Нельсона Френсиса, стоимость одного слова машино-читаемого корпуса устной речи порой может превосходить стоимость слова корпуса письменных текстов в 100 раз (Francis 1979) [2]. Не удивительно поэтому, что доля корпусов устной речи в общей массе машино-читаемых корпусов не слишком велика.

Насколько можно судить по публикациям, наиболее интенсивно корпусо-ориентированные исследования, в том числе и на машино-читаемых корпусах устной речи, ведутся сейчас в США, Великобритании, ФРГ и Швеции. Рассмотрим машино-

---

[2] Иногда, хотя и редко, корпус устной речи может попасть в руки лингвистов в готовом затранскрибированном виде. Например, для лингвистических исследований использовалась транскрипция магнитофонных записей Белого дома, опубликованная в качестве результата расследования Уотергейтского дела. К сожалению, низкое качество записи и транскрипции сильно снизили научную ценность данного издания (Francis 1979).

читаемые корпуса устной речи, созданные в этих странах, и ряд наиболее интересных, на наш взгляд, исследований, проводимых на материале этих корпусов.

Текстовые базы данных устной речи в США

Американские машино-читаемые корпусы текстов устной речи вряд ли можно назвать представительными. Однако полученные на них статистические данные безусловно интересны, прежде всего, как исходная точка для объективного сравнения особенностей устной и письменной речи.

Первый машино-читаемый корпус текстов устной речи американского варианта современного английского общелитературного языка был создан в начале 60-х годов Дэвисом Хаузом с целью получения частотного словаря разговорной речи (Howes 1966). Корпус объемом 250 тыс. словоупотреблений составлен из 50 интервью по 5.000 словоупотреблений каждое, записанных в период с 1960 по 1965 гг. в Бостоне. Среди информантов, давших по одному интервью, было 20 студентов-второкурсников и 20 пациентов клиники (с заболеваниями, не влияющими на умственную и речевую деятельность). Кроме того, у одного пациента клиники было взято 10 интервью для изучения стабильности частотных характеристик слов у индивидуума.

В качестве "зачина" обычно использовались фразы: "Расскажите, в какой области вы специализируетесь" (для студентов) или "Расскажите, что привело вас в больницу" (для пациентов). Если информант начинал обстоятельно отвечать, спрашивающий старался минимально вмешиваться, подавая лишь реплики, необходимые для продолжения монолога. Если разговор не клеился, спрашивающий предлагал новую тему, например, просил рассказать о семье.

Магнитофонная запись интервью была затранскрибирована на базе стандартной английской орфографии. Индивидуальные особенности произношения не учитывались. В транскрипции передавались общепринятые разговорные стяжения типа I'm, don't и т.п. Кроме того, отмечалось нес-

колько характерных для разговорной речи квази-слов: uhhuh 'да', uhun 'нет', huh, eh 'вопрос', hmmm 'размышление, нерешительность'.

Ввод в ЭВМ и подсчет частот был проведен в Вычислительном центре Массачусетского технологического института. В корпусе объемом 250 тыс. словоупотреблений зафиксировано 9699 различных словоформ, из которых 4097 встретились по одному разу (Howes 1966).

Корпус устной речи американского варианта английского языка Хартвига Даля создавался в семидесятые годы. Объем его - примерно I млн. словоупотреблений - ориентирован на Брауновский корпус. С самого начала ставилась задача сравнения данных письменной и устной речи.

Корпус состоит из записей психоаналитических сеансов. Из I5 магнитофонных кассет, попавших в распоряжение создателей корпуса, было отобрано случайным образом 225 сеансов (по I5 с каждой кассеты) для последующего транскрибирования и ввода в ЭВМ. Среди информантов как пациенты (7 женщин и 8 мужчин, их высказывания составляют 86% корпуса), так и психоаналитики (I женщина и I3 мужчин) из 9 городов США.

Пытаясь обосновать представительность корпуса, автор ссылается на корпус Хауза, проводя малообоснованную генерализацию: "Сбор текстов устной речи обычно проводится в монологической ситуации, когда у информанта просят рассказать о чем-нибудь по его желанию... Психоаналитический сеанс подобен экспериментальному монологу..." (Dahl 1979,IV).

Транскрибирование текстов производилось на основе стандартной орфографии. Все диалектные варианты слов записывались унифицированно. Исключение делалось лишь для неологизмов и очевидных случаев неправильного произношения из-за незнания слова или в результате путаницы - в основном это касается околонаучной лексики. Определенные трудности вызвало транскрибирование сленга и жаргонизмов.

При выборе того или иного способа транскрибирования приоритет всегда отдавался вариантам, обеспечивающим легкость идентификации и нахождения в базе данных. Так отдельные слова, составляющие фразеологическое сочетание, перед вводом в ЭВМ тем или иным способом спаивались. Например, было принято слитное написание allright (вместо обычного all right).

Для передачи слова, которое при транскрибировании и повторном прослушивании не удавалось распознать, был введен специальный знак z . Введение такого знака создатели корпуса считают принципиальным моментом, необходимым для адекватного представления устной речи, поскольку в реальном разговоре попадаются слова, которые слушающий не может идентифицировать. Слово z имеет довольно большую частоту.

Ни ударение, ни какие-либо иные просодические признаки в транскрипции не фиксировались. Помимо собственно текста в качестве комментария давалась некоторая дополнительная информация: отмечались паузы, длительное молчание, смех, бормотание, вздохи, кашель и тому подобные паралингвистические явления, сопутствующие речи.

Ввод в ЭВМ и статистическая обработка были осуществлены в ВЦ Нью-Йоркского медицинского центра.

Даже беглое сравнение частотных характеристик Брауновского корпуса [3] и корпуса Даля дает интересные результаты. Объем корпусов примерно одинаков - чуть больше I млн. словоупотреблений, но если в миллионном корпусе письменных текстов зафиксировано 50.406 различных словоформ, то в корпусе устной речи различных словоформ почти в три раза меньше - I7.87I.

---

[3] Полное название корпуса - Стандартный корпус текстов американского варианта современного английского языка Брауновского университета. Корпус состоит из прозаических текстов, взятых из книг, газет и журналов, изданных в США в I96I г. и представляет общелитературный язык образованных американцев (Francis I980). С I980 г. существует версия корпуса с грамматической индексацией.

Еще показательнее сравнение количества словоформ, покрывающих определенный объем текста для письменной и устной речи - см. Табл. I (Dahl 1979, VIII).

Таблица I

Сравнение количества различных словоформ, покрывающих фиксированный процент общего объема текста в корпусе Даля и Брауновском корпусе.

Процент общего объема текста, покрываемого данным количеством словоформ	Количество различных словоформ	
	Корпус текстов устной речи Даля	Брауновский корпус письменных текстов
50	42	I33
75	I83	I.788
80	270	2.827
85	436	4.572
90	848	7.955
95	2.243	I6.2I7
98	5.669	30.I24
I00	I7.87I	50.406

Различие между письменной и устной речью обнаруживается и при сравнении первого десятка самых частых слов - см. Табл. 2 (Dahl 1979; Kučera, Francis 1967).

Таблица 2

Самые частые слова корпуса Даля и Брауновского корпуса.

Ранг	Корпус устной речи Даля		Брауновский корпус письменных текстов	
	Слово	Абсолютная частота	Слово	Абсолютная частота
I	I	65.2I3		69.97I
2	and	38.020		36.4II
3	the	29.753		28.852
4	to	29.653		26.I49
5	that	27.558		23.237
6	you	26.598		2I.34I
7	it	20.542		I0.595
8	of	20.290		I0.099
9	a	I9.385		9,8I6
I0	know	I5.285		9,543

Британские текстовые базы данных устной речи

Крупнейшая текстовая база данных устной речи британского варианта общелитературного английского языка создана в Бирмингеме (Renouf 1986). К настоящему времени объем ее составляет 1,3 млн. словоупотреблений. Она является частью крупнейшего в мире машино-читаемого корпуса текстов британского варианта английского языка объемом примерно в 20 млн. словоупотреблений (Huizhong 1985).

Однако сбор устного материала продолжается. Сейчас особое внимание уделяется записи неформального разговора, для чего разработана новая методика, сочетающая высокое качество записи с естественностью получаемых данных. В студию звукозаписи приглашаются небольшие группы желающих по 6-8 человек (студенты или лица с высшим образованием) и им предлагается обсудить некоторую проблему или выполнить задание, связанное с описанием определенных сторон жизни. Работа ведется с целью создания нового учебного курса разговорного английского языка, учитывающего реальную языковую ситуацию (Renouf 1986).

Подробное описание Бирмингемского корпуса содержится в (Renouf 1984); к сожалению, эта работа осталась для нас недоступной.

Вторая по величине текстовая база данных устной речи британского варианта английского языка, Лондонско-Лундский корпус, изначально создавалась как часть специализированного корпуса текстов современного английского языка в рамках проекта "Обзор употребления английского языка".

Проект "Обзор употребления английского языка" (the Survey of English Usage) был начат в 1959 г. в Дурхэмском университете, с 1960 года реализация его перенесена в Лондонский университетский колледж. Проект предусматривал составление полного описания грамматического арсенала, которым располагает взрослый образованный носитель британского варианта английского языка, для которого этот язык является родным. Руководитель проекта -

известный британский лингвист Рэндол Квирк. Для реализации проекта за 20 лет была создана уникальная текстовая база данных, не имеющая параллелей не только для английского, но и для какого-либо другого языка мира (Quirk, Svartvik 1979).

Было решено включить в корпус как письменные тексты, так и тексты устной речи. При этом для учета различных функциональных стилей предполагалось использовать не только изданные тексты, но и рукописи. Общий объем Лондонского корпуса - I млн. словоупотреблений. Он состоит из 100 письменных текстов и 100 текстов устной речи, относящихся к периоду после 1950 г. Объем каждого текста - 5.000 словоупотреблений. В таблице 3 показано распределение текстов по типам (Quirk, Svartvik 1979, 207).

Очевидно, что никакой корпус текстов, по природе своей ограниченный, не может дать исчерпывающий материал для полного описания грамматики языка. Поэтому с самого начала планировалось проведение проверочных экспериментальных опросов носителей языка как необходимого источника расширения материала, содержащегося в корпусе, а также для изучения грамматических аспектов, вовсе не представленных в корпусе (Quirk, Svartvik 1979; Greenbaum, Quirk 1970).

Сначала весь материал корпуса был занесен на карты, причем последовательные отрезки текста, записанные на разных картах, частично перекрываются, чтобы обеспечить наличие адекватного контекста для каждой грамматической структуры. Для рукописных текстов не только сохранялись написание и пунктуация оригинала, но и отмечались последующие вставки или зачеркивания.

В качестве источника текстов устной речи использовались материалы радиопередач (записи публичных дискуссий, формальных и неформальных выступлений), магнитофонные записи, сделанные во время заседания различных комитетов, а также записи неформальных разговоров.

Таблица 3

Состав Лондонского корпуса современного английского языка

I. Письменные тексты (100)
   A. Изданные (46)
      Литература по искусствоведению - 6
      Научная литература - 7
      Учебная литература - 6
      Пресса: общие новости - 4; специальные репортажи - 4
      Административные и специальные материалы - 4
      Юридические и законодательные материалы - 3
      Проповеди - 5
      Художественная проза - 7
   B. Неизданные (36)
      Рукописи: художественные - 5; информационные - 6
      Письма: бытовые: близкому человеку - 6; равному - 3; находящемуся на социальной дистанции - 4
      Письма: не-бытовые: равному - 4; находящемуся на социальной дистанции - 4
      Личные дневники - 4
   C. Квази-устные тексты
      Драматические произведения - 4
      Написанное заранее официальное выступление - 3
      Новости, передаваемые по радио - 3
      Лекции: информационные - 4; художественные - 2
      Рассказы - 2
II. Тексты устной речи (100)
   A. Монолог (24)
      Подготовленное (но не записанное заранее) выступление - 6

- Спонтанные
  - выступление - 10
  - комментарий
    - спортивный - 4
    - неспортивный - 4

В. Диалог (76)

- Разговор
  - Записанный скрытым микрофоном
    - между близкими людьми - 24
    - между людьми, находящимися на дистанции - 8
  - Записанный явно
    - между близкими людьми - 22
    - между людьми, находящимися на дистанции - 6
  - Телефонный
    - между близкими людьми - 10
    - между людьми, находящимися на дистанции - 6

При разработке системы транскрипции в основу было положено орфографическое (а не фонетическое) представление, дополненное множеством просодических и паралингвистических признаков, воспринимаемых при прослушивании магнитофонной записи.

После внимательной сверки отпечатанного на картах текста с оригиналом делалось 200 копий каждой карты, и они поступали на "первичный анализ", в результате которого каждое грамматическое явление, отмеченное в тексте карты (от неопределенного артикля до сложного предложения) получало свою собственную карту, которая помещалась в соответствующее место картотеки. В таком виде корпус хранится в Университетском колледже в картотечных шкафах, каждый ящик которых имеет этикетку

определенной грамматической категории (нулевой артикль, именная группа, последовательности прилагательных, сравнительные конструкции и т.д.) и содержат все встречаемости данной категории в материале корпуса, независимо от типа текста.

На материале корпуса (и дополнительных экспериментальных данных) в 1972 г. была издана Грамматика современного английского языка (Quirk-Greenbaum 1972) как предварительный вариант грамматического описания. В 1985 г. вышла в свет Полная грамматика английского языка (Quirk et al. 1985).

Единственная в своем роде текстовая база данных была создана вручную. Однако сразу стало ясно, что пользование всем богатством грамматического и стилистического материала корпуса сопряжено с рядом неудобств. Прежде всего, размещение материала в ящиках шкафов сильно ограничило количество специалистов, получавших к нему доступ. В связи с этим в 1975 г. в Лундском университете было положено начало дочернему проекту "Обзор устной формы английского языка" (the Survey of Spoken English), целью которого являлся перевод текстов устной речи Лондонского корпуса в машино-читаемую форму. Ввод текстов в ЭВМ был закончен в 1979 г. С этого времени тексты устной речи Лондонского корпуса получили независимый статус в качестве Лондонско-Лундского корпуса. В 1980 г. часть Лондонско-Лундского корпуса была издана (Svartvik, Quirk 1980). Теперь для лингвистов в разных странах появилась возможность получить на выбор магнитную ленту с записью корпуса для последующей обработки на ЭВМ или печатную версию части корпуса.

Перед вводом корпуса устной речи в ЭВМ было решено опустить обозначение некоторых просодических и, в особенности, паралингвистических признаков как менее важных с точки зрения просодического и грамматического исследования устной речи. Таким образом, в машинной версии были сохранены базисные просодические различия, в то время как такие признаки, как темп, громкость, голосовые модификации и т.д. были опущены.

Ввод и обработка текстов в Лунде осуществлялись с видеотерминала. Тексты хранятся на магнитном диске, принадлежащем Лундскому университету. С помощью специальных программ вылавливались логические ошибки (например, опущенные с одной стороны скобки). Выявление прочих ошибок производилось по распечаткам.

Сейчас завершаются работы по грамматическому индексированию Лондонско-Лундского корпуса: каждому слову приписывается индекс, соответствующий определенному грамматическому классу. Это расширяет возможность получения по корпусу грамматической информации и информации о связи грамматических признаков с просодическими.

Поскольку не существовало адекватной полностью автоматической системы грамматического индексирования, а ручное индексирование корпуса объемом 500 тыс. словоупотреблений заняло бы слишком много времени, была разработана полуавтоматическая интерактивная система индексирования. Лингвист, сидящий за терминалом, имеет прямой доступ к базе данных. Специальные программы выводят на экран всю необходимую ему информацию. Программы помогают приписывать соответствующие грамматические классы наиболее частым словам и синтаксические функции словам определенных грамматических классов.

На материале Лондонско-Лундского корпуса уже выполнен целый ряд интересных исследований. Наиболее фундаментальная работа (Stenström 1984) посвящена вопросам и ответам в английской разговорной речи. На большом материале рассматриваются различные типы вопросов с ориентацией на прагматический аспект (вопрос с целью получения информации, ожидание подтверждения, предложение, риторический вопрос, восклицание, ожидание одобрения) и различные типы ответов (соответствующие и не соответствующие ожиданиям спрашивающего). Большое внимание уделяется структурным и смысловым элементам, обеспечивающим связность диалога.

В ряде работ рассматривается специфическая черта устной речи - так называемые словесные паузы, выражения типа 'значит', 'вот', 'ну', 'видите ли' и т.п. (Svartvik 1980; Svartvik, Stenström 1985; Erman 1986). Получают освещение и другие грамматические (Crystal 1980) и лексические (Aijmer 1984) особенности устной речи.

"Тот, кто создает корпус не только для себя, но делает его доступным для других исследователей, производит поистине неоценимый вклад в развитие науки" (Francis 1979, 122) - эти слова Нельсона Френсиса могли бы служить эпиграфом к описанию деятельности Международного машино-читаемого архива современного английского языка (the International Computer Archive of Modern English - ICAME). Международный архив был создан с целью: а) сбора и распространения информации о машино-читаемых английских языковых материалах и лингвистических исследованиях, завершенных или проводящихся на основе этих материалов; б) формирования архива машино-читаемых корпусов текстов английского языка, локализуемого в Бергенском университете (Норвегия), откуда копии этих корпусов могут быть отправлены заинтересованным лицам и организациям за определенную плату. Международный архив издает журнал "ICAME News", где публикуются материалы корпусоориентированных исследований. В настоящее время в Архив входят три корпуса: Лондонско-Лундский корпус устной речи, Брауновский корпус и Ланкастерско-Осло-Бергенский корпус письменных текстов британского варианта современного английского языка [4].

Опыт работы с корпусами Архива показал, что использование достаточно больших стандартных корпусов дает

---

[4] Ланкастерско-Осло-Бергенский корпус (сокращенно корпус LOB) является британским двойником Брауновского корпуса: объем - I млн. словоупотреблений, состоит из прозаических текстов, изданных в Великобритании в 1961 г. В 1985 г. закончена грамматическая индексация корпуса (Johansson 1985).

возможность значительно повысить качество изучения фактического употребления языка. Корпус такого типа может быть источником: I) больших текстовых выборок и большого количества текстовых выборок, что одинаково необходимо для достоверности всякого рода статистических данных, 2) текстовых выборок, взятых из различных типов текстов и 3) стандартизованной базы данных, пользоваться которой могут одновременно разные ученые, так что индивидуальные исследования могут быть повторены, а результаты, полученные в ходе различных исследований, могут непосредственно сравниваться (Biber 1986a).

В 1981-84 гг. в Лундском и в Ушсальском университетах проводились комплексные исследования в рамках проекта "Английский язык в устной и письменной форме". Программа исследований предполагала сопоставительное изучение устной и письменной форм английского языка на материале Лондонско-Лундского и Ланкастерско-Осло-Бергенского корпусов. Для осуществления проекта было выбрано два подмножества этих корпусов, представляющих, по мнению ученых, архетипы устной речи и письменного текста соответственно: спонтанный разговор и описательная проза. Объем полученного таким образом корпуса архетипов составил 200 тыс. словоупотреблений: 100 тыс. - устная речь и 100 тыс. - письменные тексты.

В рамках проекта был выполнен целый ряд интересных работ по сравнительному изучению лексики, грамматики и прагматики устной и письменной форм английского языка, причем практически все исследователи подчеркивали прагматические аспекты выявленных расхождений. В проекте приняли участие зарубежные ученые. Кроме того, на материале Лондонско-Лундского и Ланкастерско-Осло-Бергенского корпусов были составлены мини-корпуса по 25 тыс. словоупотреблений, которые рекомендовались студентам для выполнения курсовых и дипломных работ, тематика которых укладывалась в рамки проекта.

Результаты исследований были опубликованы в сборнике "English in Speech and Writing", в журнале "Studia

Linguistica и других изданиях. Сравнивались как отдельные лексические, так и грамматические аспекты. При этом исходной точкой мог служить план выражения, а затем рассматривалась семантика и особенности употребления выбранной формы (Aijmer 1986; Stenström 1986; Viitanen 1986) или грамматической конструкции (Bäcklund 1984; Bäcklund 1986), либо семантическая категория, и тогда рассматривались способы ее выражения и их распространение и особенности употребления (Tottie 1983; Altenberg 1984; Altenberg 1986; Tottie 1986; Hermerén 1986). Иногда в качестве дополнительного материала привлекался Брауновский корпус (Bäcklund 1984; Bäcklund 1986; Aijmer 1984), а также письменные тексты Лондонского корпуса (Tottie 1983).

Появились и методологические работы, в которых обсуждается закономерность сравнения различных типов текстов устной и письменной речи и достоверность получаемых результатов, в частности, ставится вопрос, действительно ли расхождения, отмечаемые при сравнении двух форм языковой репрезентации, отражают свойства этих форм, или влияние оказывают факторы, не зависящие от формы репрезентации, например, языковая интенция, степень объективности изложения и т.п. (Biber 1986a; Biber 1986b).

Текстовые базы данных немецкой устной речи

Для немецкого языка существует несколько машино-читаемых корпусов устной речи.

Самым представительным можно считать Фрейбургский корпус, составленный во Фрейбургском отделении Института немецкого языка на базе архива магнитофонных записей немецкой устной речи, существенную часть которого составляли записи радиопередач, а также записи, сделанные во время различных заседаний, конференций, слушания дел в суде и прочих публичных мероприятий (Berens 1979).

Из 820 магнитофонных записей, сделанных в конце 60-х - начале 70-х годов на территории ФРГ, общей протяженностью звучания в 580 часов было отобрано 222 текста различного

объема (от 175 до 16.390 словоупотреблений); общий объем текстов составляет примерно 600 тыс. словоупотреблений. Тексты были затранскрибированы в специально разработанной системе транскрипции, ориентированной на стандартную орфографию, и введены в ЭВМ.

Тексты корпуса распадаются на три группы:

1) монолог - 68 текстов: репортаж, распоряжение, доклад, устный рассказ и т.п.;

2) асимметричный диалог (интервью) - 87 текстов: интервью, консультация, экзамен, допрос, наведение справок и т.п.;

3) симметричный диалог (беседа) - 6 текстов: дискуссия, обсуждение, беседа, светский разговор и т.п.

На материале корпуса исследовались грамматические и лексические особенности устной речи.

В настоящее время Фрейбургский корпус, вместе с Мангеймовским корпусом письменных текстов ФРГ, ГДР, Швейцарии и Австрии (Schaeder 1978) и Боннским корпусом газетных текстов ФРГ и ГДР (Schaeder 1981) составил основу текстового банка лексикографической базы данных LEDA [5], созданной в Маннгейме для работы над новым полным междисциплинарным словарем немецкого языка (Teubert 1984). Предполагаемый объем текстового банка системы LEDA не менее 50 млн. словоупотреблений. Исходные теоретически обоснования к текстовому банку (Mentrup 1979) позволяют предположить, что представленность в нем устной речи не ограничится Фрейбургским корпусом.

Корпус текстов устной речи "Grunddeutsch", созданный в Стенфордском университете, состоит из текстов устной речи, записанных в ФРГ, ГДР, Швейцарии и Австрии в 1960-1961 гг., то есть, представляет основные географические варианты устной речи современного общелитературного немецкого языка. Корпус включает только диалогическую речь (симметричный и асимметричный диалог). Объем корпуса - около 600 тыс. словоупотреблений. Корпус

---

[5] LEDA = LExicographical DAta Base for German.

довольно интенсивно используется в учебном процессе, в первую очередь, для составления различных учебных пособий: немецко-английских словарей многозначных слов, словосочетаний, идиом, для получения всякого рода статистических данных, а также для исследований грамматики и семантики (Pfeffer 1979).

Специализированный Ульмский психоаналитический корпус содержит как устные, так и письменные тексты: наряду с записями психоаналитических сеансов, в корпус входят письменные отчеты о сеансах, при этом устная речь преобладает, составляя около 80% всего корпуса (Mergenthaler 1979). Перевод архива магнитофонных записей и отчетов на машинные носители был начат в 1975 г. на базе Ульмского университета. Сейчас объем корпуса превышает 3 млн. словоупотреблений. При этом постоянно ведется его пополнение новыми текстами. Транскрибирование магнитофонных записей осуществляется на базе стандартной орфографии. К текстам дается минимальный паралингвистический комментарий.

Корпус был задуман как эмпирическая база для научных исследований, но сейчас дает также иллюстративный материал на занятиях со студентами-медиками. Целью научных исследований является описание психоаналитического процесса. При этом, с одной стороны, корпус используется как источник документации для изучения различных клинических случаев, для систематического описания хода лечения и для оценки той или иной медицинской концепции, с другой стороны, как языковой материал, сам по себе являющийся предметом изучения, при этом рассматривается речь пациента, речь психоаналитика и их взаимодействие в психоаналитической беседе.

Рассмотрение ряда зарубежных текстовых баз данных дает некоторое представление о направлениях и масштабах работ в области использования машино-читаемых корпусов для анализа особенностей устной речи. Весьма плодотворным представляется сопоставительное изучение устной и письменной речи при наличии сопоставимых стандартных машиночитаемых корпусов устных и письменных текстов.

Создание текстовой базы русской разговорной речи (описание состава корпуса текстов русской разговорной речи, а также методики сбора и транскрибирования материала дается в Овсиенко 1976) в Лаборатории по применению вычислительных средств в гуманитарных науках МГУ открывает большие возможности для всестороннего изучения диалогической разновидности русской устной речи. К настоящему времени подготовлен частотно-семантический словарь лемм русской разговорной речи. Планируя дальнейшую работу с текстовой базой, безусловно, следует учитывать опыт зарубежных исследователей. К сожалению, для русского языка пока еще нет стандартного машино-читаемого корпуса письменных текстов, который мог бы послужить базой для сравнения с данными устной речи. Остается надеяться, что создание подобного корпуса войдет составной частью в проект Машинного фонда русского языка.

Литература

Овсиенко Ю.Г. Задачи описания разговорной речи и вопросы построения лингво-статистического справочника // Вопросы лингво-статистического справочника русской разговорной речи. - М.: изд-во УДН, 1976. - С. 3-9.

Aijmer K. 'Sort of' and 'kind of' in English con versation // Studia Linguistica. - Lund, 1984. - Vol. 38. - N 2. - P. 118-128.

Aijmer K. Why is 'actually' so popular in spoken English? // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 119-129.

Altenberg B. Causal linking in spoken and written English // Studia Linguistica. - Lund, 1984. - Vol. 38.- N 1. - P. 20-69.

Altenberg B. Contrastive linking in spoken and written English // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 13-40.

Bäcklund I. Conjunction-headed abbreviated clauses in English // Studia Anglistica Upsaliensia. - Vol. 50.- Uppsala, 1984.

Bäcklund I. Beat until stiff. Conjunction-headed abbreviated clauses in spoken and written English // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 41-55.

Berens F.J. Das Freiburger Korpus gesprochener deutscher standartsprache: Möglichkeiten seiner Auswertung und Beispiel der Untersuchung der Lexeme 'Freude' und 'freuen' und ihrer Ableitungen // Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. - Königstein, 1979. - S. 268-280.

Bergenholtz H., Schaeder B. Deskriptive Lexikologie // Zeitschrift für Germanistische Linguistik. - Vol. 5. - Berlin / New York, 1977. - S. 2-33.

Bergenholtz H., Schaeder B. Ausblicke auf eine deskriptive Lexikographie // Interdisziplinäres Deutsches Wörterbuch in der Diskussion. - Düsseldorf. 1978. - S. 116-172.

Biber D. On the investigation of spoken/written differences // Studia Linguistica. - Lund, 1986. - Vol. 40. - N 1. - P. 1. - 21(a).

Biber D. Spoken and written textual dimensions in English: Resolving the contradictory findings // Language. - Vol. 62. - 1986. - P. 384-414(b).

Crystal D. Neglected grammatical factors in conversational English // Studies in English Linguistics for Randolph Quirk. - London / New York, 1980. - P. 153-166.

Dahl H. Word frequencies of spoken American English. - Detroit, 1979.

Erman B. Some pragmatic expressions in English conversation // English in Speech and Writting. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 131-147.

Francis W.N. Problems of assembling and computerizing large corpora // Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. - Königstein, 1979. - P. 110-123.

Francis W.N. A tagged corpus - Problems and prospects. // Studies in Inglish Linguistics for Randolph Quirk. - London / New York, 1980. - P. 192-209.

Gorcy G. L'informatique et la mise en oeuvre du Trésor de la langue française, dictionnaire de la langue du 19-e et du 20-e siècle (1780-1960). - Pise, 1981.

Greenbaum S., Quirk R. Elicitation experiments in English. Linguistic studies in use and attitude. - Coral Gables, 1970.

Henne H., Rehbock H. Einführung in die Gesprächsanalyse. - Berlin / New York, 1982.

Hermerén L. Modalities in spoken and written English. An inventory of forms // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 57-91.

Hess K., Brustkern J., Lenders W. Maschinenlesbare deutsche Wörterbücher. Dokumentation, Vergleich, Integration. - Tübingen, 1983.

Howes D. A word count of spoken English // Journal of Verbal Learning and Verbal Behavior. - Vol. 5. - 1966. - P. 572-606.

Huizhong Y. The use of computers in English teaching and research in China // English in the World. Teaching and Learning the language and literatures. Papers of an international conference entitled 'Progress in English Studies' held in London, 17-21 September 1984. - London, 1985. - P. 86-102.

Johansson S. Grammatical tagging and total accountability // Papers on language and literature presented to Alvar Ellegard and Erik Frykman. - Gotheburg Studies in English. - Vol. 60. - Göteborg, 1985. - P. 202-220.

Kučera H., Francis W.N. Computational analysis of present day American English. - Providence, 1967.

Lenders W., Willée G. Linguistische Datenverarbeitung. - Opladen: Westdeutscher Verlag, 1986.

Mentrup W. Überlegungen zur Zusammenstellung und Verwendung eines Korpus für ein grosses interdisziplinäres Wörterbuch der deutschen Sprache // Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. - Königstein, 1979. - S. 182-203.

Mergenthaler E. Das Textkorpus in der psychoanalytischen Forschung // Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. - Königstein, 1979. - S. 131-147.

Pfeffer J.A. Drei deutsche Korpora im Institut für Grunddeutsch. Ziel und Stand ihrer Auswertung // Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. - Königstein: Scriptor, 1979. - S. 307-323.

Quirk R., Greenbaum S. A grammar of contemporary English. - London: Longman, 1972.

Quirk R., Svartvik J. A corpus of morden English // Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. - Königstein: Scriptor, 1979. - P. 204-218.

Quirk R., Greenbaum S., Leech G., Svartvik J. A Comprehensive Grammar of the English Language. - London: Longman, 1985.

Renouf A. Corpus development at Birmingham University // Aarts J., Meijs W. (eds). Corpora linguistics. Recent Developments in the Use of Computer Corpora in English Language Research. - Amsterdam: Rodopi, 1984. - P. 3-39.

Renouf A. The elicitation of spoken English // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 177-197.

Schaeder B. Eine Dokumentation maschinenlesbarer Textkorpora der Deutschen Gegenwartssprache // Interdisziplinäres Deutsches Wörterbuch in der Diskussion. - Düsseldorf, 1978. - P. 233-254.

Schaeder B. Lexikographie als Theorie und Praxis. - Tübingen, 1981.

Stenström A.-B. Questions and Responses in English Conversation // Lund Studies in English. - Vol. 68. - Lund, 1984.

Stenström A.-B. What does 'really' do? Strategies in speech and writing // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 149-163.

Svartvik J. Well in conversation // Studies in English linguistics for Randolph Quirk. - London, 1980. - P. 167-177.

Svartvik J., Quirk R. (eds). A Corpus of English Conversation // Lund Studies in English. - Vol. 56. - Lund, 1980.

Svartvik J., Stenström A.-B. Wordsl words, words: the rest is silence? // Papers on language and literature presented to Alvar Ellegard and Erik Frykman. - Gotheburg Studies in English. - Vol. 60. - Göteborg, 1985. - P. 342-353.

Teubert W. Applications of a Lexicographical Data Base for German // 10th International Conference on Computational Linguistics, 22th Annual Meeting of the Association for Computational Linguistics. Proceedings of COLING 84. 2-6 July 1984. - Stanford University, California: Association for Computational Linguistics, 1984. - P. 34-37.

Tottie G. The missing link? or why is there twice as much negation in spoken English // Papers from the Second Scandinavian symposium on syntactic variation.- Stockholm Studies in English. - Vol. 57. - Stockholm, 1983. - P. 67-82.

Tottie G. The importance of being adverbial. Adverbials of focusing and contingency in spoken and written English // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 93-118.

Viitanen O. On the position of 'only' in English conversation // English in Speech and Writing. A symposium. - Studia Anglistica Upsaliensia. - Vol. 60. - Uppsala, 1986. - P. 165-175.

И.А.Большаков

ЭФФЕКТИВНОЕ ЦЕЛОБАЙТОВОЕ КОДИРОВАНИЕ ТЕКСТОВ

Состояние проблемы. Проблема предельно компактного представления информации продолжает оставаться актуальной. В последние годы сжатие становится особенно важным для текстов на естественном языке, которые все чаще обрабатываются на ЭВМ. В машинах хранятся и между машинами пересылаются совокупности кратких сообщений, крупных документов, а порой - полные тексты газетных полос, брошюр и книг, и никакое совершенствование машинных носителей информации не успевает за потребностями практики. Особенно необходимыми, на наш взгляд, окажутся методы сжатия при хранении громадного корпуса рефератов научно-технических документов (Михайлов 1986) и составляющих Машинного фонда русского языка (Андрющенко 1985).

Неизбыточное (эффективное) кодирование сообщений, порождаемых источниками с заранее известными статистическими свойствами, было сформулировано как теоретическая проблема в рамках теории информации Шеннона (Шеннон 1963). Эта теория указала потенциальный предел сжатия, оформленный понятием энтропии, а позже были предложены конкретные виды малоизбыточных кодов, из которых двоичные коды Хаффмена обеспечивают компактность, близкую к пределу (Фано 1965, Хаффмен 1961).

Коды, по компактности близкие к пределу, принципиально неравномерны: более вероятные сообщения отображаются короткими кодовыми комбинациями, менее вероятные - более длинными. При этом ни одна кодовая комбинация не является начальной частью другой. Это свойство префиксности гарантирует правильное расчленение кодов при их декодировании.

Довольно быстро были выявлены недостатки известных малоизбыточных кодов. Волна публикаций по малоизбыточному кодированию, возникнув в 50-х годах, не иссякала

все 60-е годы, а в 70-х годах и особенно к рубежу 80-х годов, снова испытала существенный подъем (см. Cooper, Lynch 1982; Weiss, Vernor 1978; Mayne, James 1975 и литературу к Cooper, Lynch 1982).

В поисках практичных методов малоизбыточного кодирования исследователям пришлось учесть существенные различия текстовых материалов разных видов, например, текстов на естественных и формальных языках; совокупностей числовых данных; сложно структурированных документов, части которых являются либо текстами, либо числами, либо разносистемными кодами; упорядоченных словарных материалов.

Если интересоваться только текстами на естественных языках, то предложены разные принципы классификации методов их сжатия:

- Исходные буквенные цепочки могут быть как одинаковой, так и разной длины. То же может иметь место для результирующих кодовых комбинаций.

- В качестве исходных цепочек можно брать отдельные буквы, двухбуквенные сочетания (биграммы), многобуквенные сочетания, отдельные слова.

- Кодирование можно базировать на статистических основаниях либо на иных соображениях.

Выявлено (Cooper, Lynch 1982), что методы, основанные на частотности букв и биграмм, дают лишь умеренные коэффициенты сжатия (менее 2), пословное же сжатие сулит существенно лучшие результаты, особенно - при учете статистики.

И все же требует дальнейших объяснений та ситуация, что при хорошей теоретической изученности вопросов сжатия и сохраняющемся практическом интересе к проблеме малоизбыточное неравномерное пословное кодирование пока не нашло широкого применения в практике автоматизированного хранения и пересылки текстов. Попробуем предложить здесь наши объяснения.

I. Чтобы получить надежную статистику употребления отдельных словоформ в естественном языке, требуются обширнейшие статистические эксперименты. В таком языке, как русский, практику интересуют сотни тысяч разных словоформ, а из-за крайне медленно спадающего рангового распределения словоформ это требует накопления данных по текстам в десятки и сотни миллионов словоупотреблений. Если же исходить из статистики только конкретно сжимаемого текста (Mayne, James 1975), то этап сжатия превращается в очень трудоемкую операцию.

2. Реально тексты не являются последовательностями независимо появляющихся слов. Слова в них связаны синтаксически и семантически, причем эти связи до сих пор не удалось описать детально даже детерминистскими законами. Тем большие размерностные сложности возникнут, если попробовать связать сложнейшие детерминистские закономерности со статистикой употребления в предложениях различных поддеревьев взаимной подчиненности слов. Но именно с учетом синтаксиса и семантики в экспериментах с участием человека определилась четырех-пятикратная избыточность слов естественного языка (Яглом А., Яглом И. 1973), позволяющая надеяться на столь же высокий коэффициент сжатия текстов при эффективном их кодировании.

3. Строить коды Хаффмена для обширного ансамбля сообщений долго и хлопотно, а будучи построенными, они оказываются оптимальными только для данного конечного ансамбля (в данном случае - словоформ текста) и фиксированного вероятностного распределения на нем. Реально же язык является открытой системой, где всегда можно ожидать слово, ранее в текстах не встречавшееся. С другой стороны, источники текстовых сообщений не обладают ни стационарностью, ни эргодичностью. Наблюдая сколь угодно протяженный текст, мы не гарантированы от того, что какая-либо, даже хорошо известная словоформа так и не появится в нем. Статистика существенно зависит от жанра, предметной

7-7018

области и узкой подобласти, а само деление универсума знаний на области и замешивание последних при наборе единой статистики никогда не удается провести на безупречных логических основаниях.

4. Алгоритмов, автоматически генерирующих код по самой буквенной цепочке слова и наоборот, не предложено, и поэтому как передатчик, так и приемник сжатого текстового сообщения вынуждены хранить "дешифровочную таблицу", т.е. словарь, ставящий в соответствие в передатчике каждой словоформе свою кодовую комбинацию, а в приемнике - наоборот. Заметим, что та же проблема возникает и при использовании равномерных кодов, имеющих одинаковую длину для всех кодируемых словоформ. Но в машине словари приходится хранить на тех же носителях, что и сами тексты, исходные либо сжатые. Суммарная расходуемая память может и не отличаться существенно от объема исходного текста, без словарных "нагрузок". Это порой ставит под сомнение преимущества любого сжатого кода.

5. Хранить в словарях все словоформы таких флективных языков, как русский, крайне невыгодно. Чаще хранят основы лексем и грамматические признаки, характеризующие их словоизменительный и(или) словообразовательный класс. Объем словаря при этом уменьшается, но одновременно возникают проблемы морфологического анализа/синтеза и хранения либо генерации разных кодов для разных форм одной лексемы. Проблема совместной морфологической обработки и эффективного кодирования, насколько нам известно, не изучалась.

6. Чем сложнее и компактнее построены вспомогательные для проблемы сжатия текстов структуры данных, в первую очередь, - словарь, тем труднее обеспечить с его помощью высокую скорость кодирования/декодирования. Но скорость декодирования часто является критичной. На наш взгляд, именно ее

невысокая величина подорвала интерес к эффективному кодированию на машинах 60-х годов. В современных связных приложениях задержка на кодирование/декодирование для успеха дела не должна превышать долей секунды. Что же касается нынешнего интерактивного пользователя ЭВМ, то он будет склонен работать со сжатыми текстами только тогда, когда их декодирование осуществимо в реальном времени и не слишком заметно на фоне подкачки закодированного текстового файла с внешнего накопителя (диска, ленты) в оперативную память ЭВМ и пересылки его декодированного варианта на экран дисплея или в другой файл. Однако работы, которые изучают алгоритмическую сторону этой проблемы для ЭВМ 80-х годов, малочисленны.

7. Закодированный текст вместе с исходной его избыточностью теряет и помехоустойчивость. Нарушение в нем одного лишь бита может повести не только к замене одной словоформы на другую, но и к потере синхронизации при декодировании. В таком случае весь следующий за ошибкой текст будет декодирован неверно (если это вообще удастся). Для связных приложений это представляет существенную опасность, хотя в вычислительных приложениях появление таких ошибок маловероятно. Внесение в закодированный текст небольшой и контролируемой избыточности для нужд автоматической коррекции переводит одну нелегкую проблему - эффективное кодирование - в другую, еще более сложную - кодирование с коррекцией ошибок. При таком, более сложном кодировании требуется к тому же знать статистические свойства канала передачи данных.

8. Современные ЭВМ имеют память и процессор, явным образом ориентированные на укрупненную единицу - байт, причем почти повсеместно царствует байт восьмибитовый. В этих условиях обработка неравномерных

кодов с границами, которые никак не связаны с границами байтов или машинных слов, вызывает существенные затруднения.

Постановка задачи эффективного целобайтового кодирования. В сформулированных условиях наша работа предлагает эффективные неравномерные коды, имеющие целобайтовую структуру (далее - ЭЦБ-коды). Нами рассматривается также, пока теоретически, система, использующая предлагаемые коды для кодирования/декодирования естественноязыковых текстов.

К ЭЦБ-кодам нами предъявляются следующие требования:

- Возможность работы со словарями существенно разных объемов, т.е. в нижнем пределе - со словниками кратких сообщений объемом $2^6 - 2^{10}$ единиц, а в верхнем пределе - с политематическими словарями флективных языков объемом $2^{18} - 2^{24}$ единиц. (Под единицами здесь понимаются словоформы, т.е. конкретно входящие в текст буквенные цепочки от разделителя до разделителя.) Словарь составляется заранее.

- Минимум средней длины целобайтовых кодовых комбинаций. Это и есть основное функциональное свойство искомых кодов - их эффективность.

- Целобайтовая префиксность, гарантирующая декодируемость.

- Наличие уникального байтового кода для перехода к буквальному представлению последующей информации (тоже с байтовой структурой) и, возможно, для обратного перехода. Этим свойством должна быть учтена открытость языка, выражающаяся в возможности появления в текстах цепочек, словарем не охваченных. Имеются в виду слова-окказионализмы, редкие имена и аббревиатуры, числовые и иноязычные вкрапления, а также знаки препинания и сложные внутритекстовые разделители.

- Пренебрежение взаимосвязи между словами в тексте.

К системе, использующей ЭЦБ-коды, предъявляются дополнительные требования:

- Система должна включать морфологический словарь, хранящий словоформы и(или) основы лексем выбранного естественного языка. Основная структура словаря должна быть одинаково применима при кодировании и декодировании. Для обеспечения скорости работы допускаются, однако, дополнительные массивы, зависящие от направления преобразования.

- Структура словаря должна быть одинакова как для высокофлективных языков типа русского, так и для низкофлективных языков типа английского. Различия в флективности могут выражаться в разном среднем по словарю количестве форм на одну основу и вызывать чисто количественные различия словарей. (Конкретное количество форм у лексемы внутри данного языка зависит от изменительного класса этой лексемы и может колебаться от единицы до́ нескольких сотен.)

- Морфологический анализ (при кодировании) и синтез (при декодировании) всегда должны обеспечивать различимость разных форм и кодов.

- Должны быть предприняты все меры по снижению реально занимаемого словарем объема памяти.

- Должна обеспечиваться предельно возможная скорость кодирования и, особенно, декодирования текстов.

В соответствии со всеми выдвинутыми требованиями в последующих разделах будут построены и изучены эффективные целобайтовые коды, рассмотрены способы конкретного кодирования с их помощью словоизменительных парадигм естественного языка (в первую очередь - русского) и предложены основные принципы построения системы, практически воплощающей ЭЦБ-коды.

Математическая формулировка ЭЦБ-кодирования.
Для построения конкретного семейства эффективных кодов необходимо задаться ранговым распределением словоформ кодируемых текстов. Многочисленные эксперименты показывают (Herdan 1966), что это распределение приближенно подчиняется закону Ципфа

$$p(r) = C/r^{I+\varepsilon}, \quad I \leq r \leq N,$$

где r - ранг словоформы, N - объем словаря, $|\varepsilon| < 1$. Положим для простоты $\varepsilon = 0$ и в предположении $N \gg 1$ предельным переходом от суммы к интегралу определим нормировочную константу: $C \approx 1 / \ln N$.

По Шенноновской формуле энтропии

$$H = - \sum_{r=1}^{N} p(r) \log p(r)$$

(логарифм здесь и далее двоичный) легко получить соответствующее распределение (I) длину эффективного двоичного кода (Хаффмена) в зависимости от ранга словоформы

$$l(r) = \log \log N + \log r$$

с пределами ([ ] означает целую часть выражения)

$$l_{min} = [\log \log N], \; l_{max} = [\log \log N + \log N + 1]$$

Предельный переход в (2) к интегралу дает среднюю длину двоичного эффективного кода:

$$l_{ЭКК} = \log \log N + 1/2 \log N$$

Равномерные целобайтовые коды имели бы в тех же условиях длину

$$l_{РЦБ} = 8 \left[[\log N + 1] / 8 + 1 \right]$$

В таблице I даны полученные по формулам (3) - (6) характеристики оптимальных двоичных и равномерных целобайтовых кодов. Проигрыш в длине равномерного кода относительно мал (менее 50%) лишь при N , близком снизу к $2^{8к}$, где к = I, 2, ... В прочих случаях, а также асимптотически при $N \to \infty$ проигрыш ближе к 2. Это и свидетельствует лишний раз о полезности неравномерных целобайтовых кодов, которые по свойствам находились бы между описанными типами.

ЭЦБ-коды - это, по существу, коды с мощностью алфавита в $2^8 = 256$ символов. Хорошо известна (Фано I965) теорема кодирования, согласно которой ансамбль сообщений с (двоичной) энтропией H может быть закодирован 2^8-ичным

кодом так, что среднее количество $L_{ЭЦБ}$ байт на сообщение удовлетворяет двустороннему неравенству

$$H \leqslant 8 L_{ЭЦБ} < H + 8$$

Упрощенно говоря, при переходе к целобайтовым кодам можно добиться, чтобы увеличение средней длины не превышало размеров, связанных с "округлением" битовых кодов до ближайшего байта в большую сторону. Однако неравенства (7) дают лишь границы, ничего не говоря о конкретных способах приближения к низшей из них.

При построении конкретных ЭЦБ-кодов сначала нужно определить при любом N оптимальное количество однобайтовых N_1, двухбайтовых N_2 и т.д. кодовых комбинаций. Согласно формуле (4) и таблице I минимальная длина кодов Хаффмена превышает 8 бит только при астрономически больших $N \geqslant 2^{256}$, что далеко выходит за пределы наших интересов. Поэтому будем полагать, что у ЭЦБ-кодов всегда есть кратчайшие, однобайтовые комбинации:

$$1 \leqslant N_1 \leqslant 2^8, \quad N_1 \leqslant N$$

Многобайтовые комбинации из-за требования префиксности осуществимы только тогда, когда более короткие комбинации оставляют незанятыми ряд комбинаций своей длины в качестве префиксов, так что должны выполняться неравенства

$$0 \leqslant N_k \leqslant 2^{8k} - \sum_{j=1}^{k-1} 2^{8(k-j)} N_j, \quad k = 2, 3, \ldots$$

Подсчет средней длины ЭЦБ-кода в битах дает при распределении (I)

$$l_{ЭЦБ} = 8L_{ЭЦБ} = \frac{8}{\log N} \log \left\{ N_1 \left(\frac{N_1 + N_2}{N_1}\right)^2 \left(\frac{N_1 + N_2 + N_3}{N_1 + N_2}\right)^3 \ldots \right\},$$

где число сомножителей должно быть равно длине L (в байтах) максимальных кодовых комбинаций. Как подсказывает формула (4) и таблица I, для $N \leqslant 2^{24} \simeq 1{,}6 \cdot 10^7$ едва ли оптимальны более чем 4-байтовые ЭЦБ-коды, но, вообще говоря, конкретное значение L зависит от N и само должно быть результатом решения задачи оптимизации кода.

Требование минимума (10) ведет, как нетрудно видеть, к экстремальной задаче

$$\prod_{i=1}^{L} \sum_{j=1}^{i} N_j \to \max_{L,\, N_1, \ldots, N_L},$$

которую необходимо решать при условиях (8), (9) и общем правиле нормировки

$$\sum_{j=1}^{L} N_j = N$$

Нами сформулирована выпуклая задача целочисленного программирования с линейными ограничениями. Можно отнести ее также к задачам распределения ограниченных ресурсов. В зависимости от N некоторые координаты оптимума N_2, N_3,..лежат на нулевой границе либо положительны, находясь внутри области допустимости (Кюнце, Крелле 1965).

Приближенное нахождение оптимума. Высказанные выше суждения о ЭЦБ-кодах на основе свойств кодов Хаффмена уже не удовлетворяли канону строгости. Нахождение алгоритма строгого решения задачи (11) отложим до последующих работ, а ниже построим приближенное решение, формулируя частные подзадачи для отдельных участков значений N.

I) при $1 \leqslant N \leqslant 2^8$ точное однобайтовое решение очевидно:

$$N_1 = N, \; N_2 = N_3 = \ldots = 0.$$

2) При $2^8 < N < 2^{16}$ в предположении $L = 2$ имеем задачу линейного целочисленного программирования

$$N_1 \to \max,\ N_2 \leqslant 2^{16} - 2^8 N_1,\ N_1 + N_2 = N,$$

для которой удается получить аналитическое решение:

$$\left.\begin{aligned} M_1 &= [2^{-8}N - 1],\ M_2 = [(N - 1 - (2^8 - 1)M_1)\,2^{-8}], \\ N_1 &= 2^8 - M_1 - M_2,\ N_2 = N - N_1 \end{aligned}\right\}$$

3) При $2^{15} < N < 2^{24}$ в предположении $L = 3$ имеем выпуклую целочисленную задачу

$$N_1(N_1 + N_2) \to \max$$

с несколько уточненными условиями (8), (9):

$$\left.\begin{aligned} &1 \leqslant N_1 < 2^8, \\ &1 \leqslant N_2 < 2^{16} - 2^8 N_1, \\ &1 \leqslant N_3 < 2^{24} - 2^{16}N_1 - 2^8 N_2 \end{aligned}\right\}$$

4) При $2^{22} < N$ в предположении $L = 4$ имеем задачу того же типа

$$N_1(N_1 + N_2)(N_1 + N_2 + N_3) \to \max$$

с условиями

$$\left.\begin{aligned} &1 \leqslant N_k < 2^{8k} - \sum_{j=1}^{k-1} 2^{8(k-j)} N_j,\ k = 1, 2, 3, \\ &1 \leqslant N_4 \leqslant 2^{32} - 2^{24}N_1 - 2^{16}N_2 - 2^8 N_3 \end{aligned}\right\}$$

Для обеих задач (16) и (18) сохраняется нормировка (12).

Интервалы $2^{15} - 2^{16}$ и $2^{22} - 2^{24}$ значений N при поиске парциальных решений взяты перекрывающимися намеренно. Из получаемых для разных N решений предполагается выбрать дающее наименьшее значение $l_{ЭЦБ}$ - средней длины кода. Однако и лучшее из этих решений может не быть строго оптимальным из-за приближенности метода.

Пренебрежем в (I6) - (I9) дискретностью переменных и будем искать условный экстремум возникших непрерывных функций от N_i строго внутри области допустимости с последующим округлением результатов до целых.

В каждом байте ЭЦБ-кода могут быть битовые сочетания, которые завершают или не завершают здесь искомые кодовые комбинации. Незавершенные варианты обычно используются в дальнейших байтах (если таковые есть) в качестве префиксов более длинных кодовых комбинаций. Предельно возможное использование незавершенной комбинации состоит в том, что она берется префиксом для всех возможных 2^8 более длинных комбинаций. В непоследних байтах ЭЦБ-кода, как можно показать, префиксы всегда используются предельно.

Сделаем более сильное предположение, считая, что и самые длинные кодовые комбинации ЭЦБ-кода предельно используют оставленные для них префиксы. К примеру, если $L = 3$ и двухбайтовые комбинации оставили (среди 2^{16} возможных) k незанятых префиксов, то трехбайтовых комбинаций будет ровно $2^8 k$ (хотя, строго говоря, их может быть и меньше). При таком упрощении из каждой совокупности неравенств (I7) и (I9) можно оставить лишь последнее, превратив его в равенство.

Переходя в сделанных предположениях к координатам $N_I' = N_I$, $N_2' = N_I + N_2$ и логарифмируя (I6), имеем вместо (I2), (I6) и (I7) "гладкую" задачу условного экстремума

$$\ln N_I' + \ln N_2' \to \max,$$

$$2^{24} - 2^8 (2^8 - 1)N_I' - (2^8 - 1)N_2' = M,$$

решение которой получается дифференцированием:

$$N_I' = (2^{24} - N)2^{-8}(2^8 - 1)^{-1}/2,\ N_2' = (2^{24}-N)(2^8-1)^{-1}/2$$

Возвращаясь к исходным координатам и округляя результаты, имеем

$$N_I = \left[(2^{24} - N)2^{-8}(2^8 - 1)^{-1}/2\right],\ N_2 = \left[(2^{24}-N)2^{-8}/2\right],$$
$$N_3 = N - N_I - N_2$$

Аналогичным расчетом получаем приближенное решение для (I2), (I8), (I9):

$$N_I = \left[(2^{32} - N)2^{-16}(2^8 - 1)^{-1}/3\right],\ N_2 = \left[(2^{32}-N)2^{-16}/3\right],$$
$$N_3 = \left[(2^{32} - N)2^{-8}/3\right],\ N_4 = N - N_I - N_2 - N_3$$

Приближенные оценки показывают, что парциальные решения (I5) и (20) "сшиваются" при $N \approx 2^{16}/2$, а решения (20) и (2I) - при $N \approx 2^{24}/3$. Как раз возле этих значений становится отличной от нуля очередная координата N_3 и N_4 соответственно. Это примечательно, поскольку двухбайтовых кодов внешне достаточно вплоть до $N = 2^{16}$, а трехбайтовых - до $N = 2^{24}$. Более длинные кодовые комбинации становятся нужными при N, заметно меньших $2^{8к}$, к = 2, 3, ..., потому что полное использование более длинных комбинаций возможно лишь при отказе от более коротких, а это увеличивает среднюю длину кода (при уменьшении максимальной).

Свойства ЭЦБ-кодов. Формулы (I3), (I5), (20) и (2I) после их "сшивания" позволяют изучить ряд свойств ЭЦБ-кодов.

На рис. I даны зависимости от N непосредственно величин N_i, i = I, ..., 4. Закономерность изменения всех N_i одинакова. Становясь положительной при

некотором критическом значении N, величина N_i временно берет на себя прирост N, затем достигает максимума и начинает медленно спадать. Падение имеет характер нескольких участков квазипостоянства, между которыми имеются более крутые перепады. Последние близки к точкам сшивания парциальных решений $N \approx 2^{8k}/k$, $k = 2, 3, \ldots$

На рис. 2 построены зависимости от N вероятностей p_k кодовых комбинаций разной байтовой длины. Они вычислялись по формуле, вытекающей из распределения (I):

$$p_k = (\log \sum_{j=1}^{k} N_j - \log \sum_{j=1}^{k-1} N_j) / \log N,$$

где N_j брались из формул (I3), (I5), (20), (2I). Величина p_1 равна I вплоть до 2^8, а затем монотонно падает, колеблясь возле кривой $7/\log N$, прочие же p_k имеют участки нулевого значения, роста и падения, тоже приблизительно обратно пропорционального $\log N$.

По тем же формулам строится зависимость от N средней длины ЭЦБ-кода:

$$l_{ЭЦБ} = \begin{cases} 8 \text{ при } 1 \leqslant N \leqslant 2^8, \\ 8\left(2 - \frac{1}{\log N} \log\left(2^8 - \left[2^{-8} N - 1\right] - \left[\left(N - 1 - (2^8 - 1)\left[2^{-8} N - 1\right]\right) 2^{-8}\right]\right)\right) \\ \quad \text{при } 2^8 < N \leqslant 2^{15}, \\ \frac{8}{\log N} \log\left(2^8 \left(\frac{2(2^8-1)}{2^{24}/N - 1}\right)^2 N\right) \text{ при } 2^{15} < N \leqslant 2^{22}, \\ \frac{8}{\log N} \log\left(2^{24} \left(\frac{3(2^8-1)}{2^{32}/N - 1}\right)^3 N\right) \text{ при } N > 2^{22} \end{cases}$$

Численные характеристики ЭЦБ-кодов сведены в таблицу 2. Кроме величин N_k, p_k и $l_{ЭЦБ}$ здесь приведены проигрыши $l_{ЭЦБ}/l_{ЭКХ}$ в средней длине по сравнению с кодами Хаффмена и выигрыши $l_{РЦБ}/l_{ЭЦБ}$ по сравнению с равномерными целобайтовыми кодами. Замечаем, что

проигрыш по сравнению с кодами Хаффмена с ростом N убывает и, начиная с $N \approx 2^{13}$, не превышает 10%. По-видимому, можно строго показать, что асимптотически два этих типа кода по качеству не различаются.

На таблице 2 даны также выигрыши ЭЦБ-кодов по сравнению с обычным буквенным представлением текста. Эти показатели вычислялись в предположении, что средняя по тексту длина словоформы составляет около 7 букв (т.е. 7 байт), а вместе с одиночным разделителем - 8 байт. Для наиболее интересных, крупных словарей с $N \approx 2^{20} \approx 10^6$ словоформ теоретический выигрыш ЭЦБ-кодов по отношению к несжатому представлению текста близок к 4, а для меньших словарей - еще значительнее. Однако нашим расчетом не учитывались "накладные расходы" на кодирование внесловарных вкраплений. Поэтому решающую роль в определении глобального, общетекстового коэффициента выигрыша должен сыграть эксперимент над реально действующей системой сжатия и множеством разнообразных текстов.

При сравнении ЭЦБ-кодирования текстов высокофлективных языков типа русского и низкофлективных языков типа английского из-за существенно большего числа словоформ в первом случае (русский с английским - едва ли ближе, чем 8 : 1) коэффициент сжатия в русском получается несколько ниже (по нашим формулам - примерно на 10%). Глобально по тексту разница будет меньше, поскольку английские слова в среднем короче. Решающее слово здесь опять за экспериментом.

Построение кодовых комбинаций. Перед пояснением конкретных приемов построения ЭЦБ-кодов стоит напомнить, что коды Хаффмена не определяются однозначно. Коды, оптимальные для конкретного распределения кодируемых сообщения, образуют обширный класс эквивалентности. Отдельные элементы класса различаются тем, что для любой внутрикодовой позиции бинарный выбор можно сделать по-разному, взяв здесь либо 0, либо 1.

В ЭЦБ-кодах число степеней свободы еще больше. В пределах группы к-байтовых комбинаций допустимы все взаимных замен, и это не влияет на оптимальность. Отсюда вытекает следующий простой и рациональный способ конкретного выбора кодов.

Наиболее вероятные N_I словоформ отбираются и упорядочиваются. Способ упорядочения не влияет на оптимальность и поэтому выберем его потом для облегчения операций поиска. Возникшие номера словоформ $0, \ldots, N_I - I$, согласно предыдущим разделам не превышающие $2^8 - I$, в своем 8-битовом представлении и явятся искомыми однобайтовыми кодовыми комбинациями.

Если объем кодируемого словаря превышает N_I словоформ, то далее строятся двухбайтовые комбинации для следующих по частоте встречаемости N_2 словоформ. С этой целью первая незанятая однобайтовая комбинация с арифметическим значением N_I дополняется справа нулевым байтом (иначе говоря, берется беззнаковое число $N_I \cdot 2^8$) и возникшая комбинация рассматривается как первая из двухбайтовых. В целом двухбайтовые комбинации занимают весь интервал двухбайтовых чисел от $N_I \cdot 2^8$ до $N_I \cdot 2^8 + N_2 - I$. Они и ставятся в соответствие указанным N_2 словоформам.

Если должны существовать и трехбайтовые комбинации, то первый незанятый двухбайтовый префикс $N_I \cdot 2^8 + N_2$ после умножения на 2^8 берется в качестве первой из трехбайтовых комбинаций, занимающих в целом арифметический интервал беззнаковых чисел от $(N_I \cdot 2^8 + N_2)2^8$ до $(N_I \cdot 2^8 + N_2)2^8 + N_3 - I$ и т.д.

Осталось выделить согласно требованию раздела 2 одну однобайтовую комбинацию для символа перехода к буквальному кодированию входящих в тексты внесловарных символов и символьных цепочек. В качестве такового удобно взять восьмибитовый нуль (нуль-символ), поскольку в кодах КОИ (ASCII) и ДКОИ (EBCDIC) естественноязыковые тексты этого символа никогда не содержат.

При этом приходится переместить последнюю по частоте встречаемости словоформу из однобайтовой группы в двухбайтовую, последнюю "двухбайтовую" словоформу - в трехбайтовые и т.д. К существенному снижению эффективности кода это не приведет.

Морфологическая подсистема кодирования

Программные средства морфологической обработки естественных языков существуют уже около 30 лет и продолжают совершенствоваться. Практика показала, что абсолютно универсальную программу, равно пригодную для анализа и синтеза, для множества языков и любых приложений, построить довольно трудно, да и едва ли нужно.

В нашей работе (Большаков 1985) указывалось, что в зависимости от приложений нужны морфологические анализаторы разных типов. Так, при автоматическом анализе запросов к базам данных и автоматическом переводе нужно преобразовать словоформу из текста в совокупность из одного или нескольких морфологических разборов-омонимов. Каждый из них представляется парой <имя лексемы, набор характеристик данной формы>. Под характеристиками понимается, например, падеж, число, лицо, время и т.п. С другой стороны, в системах автоматического обнаружения ошибок (автокорректорах) достаточен морфологический акцептор, который лишь опознает допустимость данной словоформы в языке.

Два указанных варианта - это крайние случаи. Для кодирования нужен некий промежуточный вариант. Разборов здесь, как и в акцепторах, не нужно, но необходима точная идентификация формы, чтобы ей можно было поставить в соответствие уникальный код. Это значит, что нужно найти любое допустимое по законам языка разложение словоформы на основу и цепочку суффиксов (далее - псевдофлексию) и оба полученных компонента использовать для построения уникальной кодовой комбинации. Для сравнения, акцептору автокорректора достаточно убедиться,

что псевдофлексия входит в некоторый класс псевдофлексий, допустимых при данном классе основ, конкретный же вид псевдофлексии несущественен.

Словарные структуры в задаче кодирования предлагается строить раздельно для групп словоформ с разной длиной ЭЦБ-кода. Для наиболее употребительной однобайтовой группы, обращение к которой будет весьма частым (более 35% при $N \leqslant 2^{20}$), лучше обойтись вообще без морфологии, тем более, что среди этих слов даже в русском языке много неизменяемых (предлоги, союзы и т.п.). В количестве, не превышающем 256 единиц (при интересующих нас объемах словарей - несколько менее 128), наиболее употребительные словоформы лексикографически упорядочиваются, и их порядковые номера берутся искомыми кодовыми комбинациями байтовой длины.

Морфологические преобразования используются в группах с двух- и многобайтовыми кодовыми комбинациями. Для каждой из этих групп строится свой парциальный словарь основ. Внутри своей группы основы подвергаются лексикографическому упорядочению. Способ организации словарей во всех группах предлагается взять одинаковым, ибо различия в длине кодов для этой организации несущественны.

Парциальный словарь размещается в совокупности блоков фиксированной длины. Основы располагаются внутри очередного блока (в сжатом виде) подряд и снабжены номерами своих морфологических классов. С каждой основой алгоритмически связывается столько кодовых комбинаций, сколько есть словоформ у основ данного класса. В итоге ЭЦБ-коды тоже оказываются лексикографически упорядоченными по парциальному словарю, причем с охватом всех форм, представленных в словаре основами.

Предлагается ввести массив-индекс для ускорения поиска нужного блока. В индексе даны парами буквенные цепочки основ, начинающих каждый блок, и кодовые комбинации, соответствующие начальным по порядку словофор-

мам этих основ. Поскольку и основы, и коды упорядочены, нужный блок можно быстро найти дихотомией при предъявлении для поиска основы либо кодовой комбинации.

Внутри блока просмотр всегда совершается последовательно, но сопутствующие операции зависят от направления преобразования.

При кодировании словоформы совершается ее морфологический анализ, сводимый к нахождению всех вкладывающихся в нее сзади псевдофлексий и поиску первой попавшейся основы, которая совпадает с оставшейся частью, а по классу допускает какую-либо отсеченную флексию.

Все различные псевдофлексии, допустимые для данного класса основ, пронумерованы подряд, начиная с нуля, причем порядок нумерации не существен. Псевдофлексия, допустимая для разных классов, может получить в каждом из этих классов разный номер.

Поскольку при последовательном просмотре блока легко вычислить по номерам классов уже просмотренных основ код $К_0$ начальной словоформы очередной основы, для получения кода преобразуемой словоформы достаточно сложить $К_0$ найденной основы с номером $К_ф$ найденной при анализе псевдофлексии в данном классе основ.

В процессе декодирования блок просматривается до тех пор, пока предъявленный код не окажется в интервале, начинающемся кодом $К_0$ начальной словоформы текущей основы и имеющем длину, равную числу псевдофлексий у основ данного класса. Код псевдофлексий $К_ф$, равный разности преобразуемого кода и $К_0$, по отдельным таблицам переводится в соответствующую буквенную цепочку, а затем совершается морфологический синтез, сводящийся к конкатенации основы и псевдофлексии.

Использование словарей основ дает существенный выигрыш в объеме памяти ЭВМ по сравнению с прямым хранением словоформ. Действительно, если полагать, что русская основа покрывает в среднем 30-35 форм (а такой величины можно достигнуть, если наряду с словоизменением

частично учесть и регулярное словообразование русских слов (Большаков 1985), то в крупных словарях (2^{19} - 2^{20} словоформ) в двухбайтовой группе нужно хранить примерно 1 тысячу основ, а в трехбайтовой - не более 30 тыс. Более длинные коды не понадобятся.

Сделаем несколько важных замечаний.

1) При кодировании словаря основ можно применить дополнительные способы сжатия, никак не связанные с ЭЦБ-кодированием, причем ряд из способов допускает совместное использование. Наиболее результативен здесь способ Купера (Cooper 1958). В целом проблема сжатия упорядоченных словарей заслуживает отдельного рассмотрения. Укажем лишь, что при комбинировании разных способов удается упаковать тысячу русских основ вместе с их классовыми номерами примерно в 4 Кбайта. Это значит, что в оперативной памяти типовых микро-ЭВМ уже можно разместить довольно обширные кодирующие словари.

2) Отнесение всех словоформ одной лексемы к одной кодовой группе в общем случае является отступлением от оптимальности. Известно, что разные формы лексемы могут иметь существенно разную вероятность появления в тексте. Замечены даже некие общие закономерности. Так, творительный и дательный падежи русских существительных в среднем встречаются значительно реже, чем именительный, родительный или винительный. При синтезе ЭЦБ-кодов мы неявно исходим из предположения, что колебания частот редко принимают межгрупповой характер. Конечно, ЭЦБ-коды в этом отношении явно устойчивее кодов Хаффмена, но полезна дополнительная экспериментальная проверка.

3) Совокупности форм разных лексем нередко пересекаются. Так, словоформа стекла принадлежит как существительному СТЕКЛО, так и глаголу СТЕКАТЬ. В той мере, в какой это зависит от составителя словаря, желательно сократить суммарный объем таких пересечений.

Возможности для сокращения существуют. Например, совокупность форм прилагательного ВЫДАЮЩИЙСЯ целиком вкладывается в совокупность форм глагола ВЫДАВАТЬ, и поэтому прилагательное можно просто не включать в кодирующий словарь, соответственно увеличив вероятность появления глагола. При практической работе такой системы из-за упорядочения основ и их последовательного просмотра всегда будет выявляться только один из возможных вариантов "толкования" и кодирования омонимичной словоформы. Ошибок при двусторонних преобразованиях, опирающихся на один и тот же словарь, не будет, но возникнут никогда не используемые кодовые комбинации, уменьшающие число тех уникальных словоформ, которые можно разместить в данной группе.

Наметки алгоритмов кодирования и декодирования. Алгоритм кодирования строится в виде цикла, тело которого включает два шага - кодирование внесловарного отрезка символов и кодирование словарного отрезка.

К очередному внесловарному отрезку сразу могут быть отнесены всевозможные внутритекстовые разделители (включая переходы на новую строку и знаки табуляции), знаки препинания, цифровые и буквенные цепочки иностранного алфавита (например, вкрапления в русский текст слов из латинских букв). Сюда же после проверки попадает и цепочка русских букв, если ей не находится соответствие в кодирующем словаре. Пока предполагается, что в закодированном тексте все внесловарные отрезки будут представляться буквально, т.е. как в исходном тексте, с предварительным выводом знака перехода к буквальному отрезку. Исключение можно сделать для самого простого и наиболее частого разделителя - одиночного пробела. Его можно просто подразумевать перед словарно верифицированным словом по умолчанию, вообще никак не отражая в закодированном тексте.

Словарным отрезком является очередная словоформа, которую удалось проанализировать с помощью одного из

кодирующих словарей. Сначала словоформа ищется среди словоформ однобайтовой группы. В случае успеха поиска в закодированный текст помещается ее байтовый код, а иначе начинается ее поиск с морфологическим разбором в "двухбайтовом" словаре основ. Если это удалось, в закодированный текст помещается соответствующий двухбайтовый код, а иначе выдвигается трехбайтовая гипотеза и т.д. Полная неудача свидетельствует, что слово системе не знакомо, и тогда оно присоединяется к предшествующему внесловарному отрезку (предусматривается задержка выдачи буквального отрезка, пока не выявится его завершение).

Алгоритм декодирования включает побайтовое перемещение внутри кодируемого текста специального указателя. Очередной байт, у которого расположен указатель, сначала сравнивается с кодом перехода к буквальному отрезку кода. Если такой код встретился, то весь отрезок, окаймленный ближайшими кодами перехода, прямо переносится в декодированный текст. На то же расстояние перемещается без всякой обработки и внутритекстовой указатель.

Если очередной байт (вне буквального отрезка) не является кодом перехода, то проверяется его принадлежность к однобайтовым кодам, для чего его содержимое арифметически сравнивается с однобайтовым порогом N_I. В случае, когда содержимое меньше порога, это действительно однобайтовый код, и из словаря словоформ извлекается соответствующая форма, а иначе байт считается лишь префиксом. Внутритекстовой указатель перемещается на одну байтовую позицию вдоль текста, а выделенная тем самым двухбайтовая величина (без знака) сравнивается с порогом $N_I \cdot 2^8 + N_2$. Если проверка успешна, с помощью словаря основ двухбайтовой группы производится синтез соответствующей словоформы, как это описывалось в предыдущем разделе. В ином случае рассматривается гипотеза трехбайтового кода и т.д.

Во всех случаях, когда в сжатый текст не было внесено ошибок, очередной шаг декодирования завершается успехом. Декодированная словоформа, сама по себе или с предыдущим пробелом, присоединяется к уже декодированной части текста, после чего цикл декодирования повторяется.

Как видим, операции преобразований внешне не сложны и в случае расположения всех словарей в оперативной памяти ЭВМ не требуют никаких обращений к дисковому накопителю. Это позволяет надеяться, что может быть достигнута достаточно высокая скорость декодирования (не менее 10 слов в секунду).

Заключение. Предложенный вид малоизбыточных кодов может иметь практическое значение как в вычислительных, так и в связных приложениях. Его предполагаемые достоинства - высокий коэффициент сжатия, целобайтовая структура и уменьшенная чувствительность к статистическому распределению отдельных слов.

Затраты на построение программ кодирования и декодирования однократны и в перспективе допускают чисто аппаратное воплощение на современной микроэлектронной основе, в корне отличной от изучавшейся в 60-х годах (Новик 1965). Лингвистические сложности построения морфологической подсистемы хотя и зависят от жанра и тематической области обрабатываемых текстов, но вполне преодолимы. Особенно хорошо сочетается система кодирования с другими системами, которые тоже вынуждены опираться на морфологический словарь.

ЭЦБ-коды можно использовать также для внутримашинного кодирования иероглифического письма. Возможны и приложения к формальным языкам (исходные тексты программ, записи формальнологических правил в системах искусственного интеллекта, сокращенное кодирование имен химических соединений и т.п.). Морфологический аспект при этом, естественно, исчезает.

Но пока настоящая работа ставит больше проблем, чем решает. Чтобы выяснить картину до конца, необходимо:

- построить строго оптимальный алгоритм определения числа кодовых комбинаций разной длины во всем интересном для приложений диапазоне объемов словарей;

- выяснить степень чувствительности предложенных кодов к закладываемой в них статистике и к возможным отклонениям реальной статистики текста от заложенной при синтезе кода;

- построить работающий морфологический словарь со структурами, допускающими легкое кодирование и декодирование;

- создать совершенные алгоритмы преобразований, чем будет завершена полная кодировочная система; при этом предстоит выбрать способ рационального кодирования упоминавшихся внесловарных элементов текста;

- провести серию экспериментов по кодированию/декодированию обширного корпуса текстов и выяснить реально достигаемые коэффициент сжатия и скорость преобразований;

- предложить методы обеспечения помехоустойчивости ЭЦБ-кодов, опирающиеся на достижения современной теории кодирования с исправлением ошибок и конкретные предположения о свойствах каналов связи между ЭВМ.

Полезно также сравнить ЭЦБ-кодирование с факторизационным методом сжатия текстов, предусматривающим выделение повторяющихся подцепочек в процессе самого кодирования (Mayne, James 1975). Достоинства и недостатки двух этих методов существенно различны. Метод факторизации явно уступает ЭЦБ-кодированию в скорости кодирования и компактности "дешифровочной таблицы", но зато обеспечивает адаптацию под конкретный текст и дает надежду на больший коэффициент сжатия в случае текстов, построенных из стандартных повторяющихся фрагментов.

Хотя словарные и алгоритмические затраты при кодировании, учитывающем взаимосвязи слов, будут существенно большими, проблема кодирования связного текста нашими предложениями тоже не снимается. В более сложных условиях свойства пословного кодирования послужат неким эталоном для сопоставлений.

ЛИТЕРАТУРА

Андрющенко В.М. Машинный фонд русского языка: постановка задачи и практические шаги // ВЯ, 1985. - № 2.

Большаков И.А. Упрощенный морфологический анализ при автоматической проверке правильности текстов // НТИ, 1985. - Сер. 2. - № 6.

Кюнце Г.П., Крелле В. Нелинейное программирование. - М.: Советское радио, 1965.

Михайлов А.И. Основные пути развития государственной системы // НТИ, 1986. - Сер. №. - № 1.

Новик Д.А. Эффективное кодирование. - М.: Энергия, 1965.

Шеннон К. Математическая теория связи // Работы по теории информации и кибернетике. - М.: ИЛ, 1963.

Фано Р. Передача информации. Статистическая теория связи. - М.: Мир, 1965.

Хаффмен Д.А. Метод построения кодов с минимальной избыточностью // Кибернетический сборник, № 3. - М.: ИЛ, 1961.

Яглом А.М., Яглом И.М. Вероятность и информация. - М.: Наука, Физматиздат, 1973.

Cooper W.S. The storage problem // Mech. Translat., 1958. - N 2.

Cooper D., Lynch M.F. Text compression using variable to fixed-length encodings // J. Amer. Soc. Inf. Sci., 1982. - V. 33. - N 1.

Herdan G. The advanced theory of language as choice and chance / Berlin et al.: Springer, 1966.

Mayne A., James E.B. Information compession by factorising common strings // Comput. J., 1975. - V. 18. - N 2.

Weiss S.F., Vernon R.L. A word-based compession technique for text files // J. Libr. Automat., 1978. - V. 11. - N 2.

Г.С.Цейтин

НА ПУТИ К СБОРОЧНОМУ ПРОГРАММИРОВАНИЮ

Современное практическое программирование еще не преодолело влияние установок, сложившихся в тот период, когда программа рассматривалась как единичный законченный продукт, настроенный на одну конкретную задачу (может быть, неоднократно решаему). Сегодня уже ясно, что установка на создание единичной программы для каждой конкретной задачи не позволяет удовлетворить быстро растущую потребность в программах, что главная задача современного программирования - это создание многократно используемых независимых элементов, пригодных для сборки разнообразных конкретных программ. Сегодня уже широко признается значение модульного программирования, пакетов прикладных программ, баз данных, распространенные языки и системы программирования оснащаются средствами поддержки модульности. Однако необходимый уровень независимости и многократной используемости пока не достигнут, а средства и подходы, по замыслу направленные на модульность, практически препятствуют ее расширению. Не случайно до сих пор не работает в полную силу идея государственного фонда алгоритмов и программ - большая часть создаваемых программ непригодна для распространения из-за чрезмерной ориентации на конкретное применение.

Само понятие модуля имеет по крайней мере два различных источника. Один - это уже названное стремление к выделению независимой единицы программистского знания, другой - потребность организационного расчленения крупных разработок. Модуль, закрепляемый за одним из низовых исполнителей большой разработки, вовсе не рассчитан на использование в ином окружении; напротив, задание на его разработку жестко увязано с конкретным проектом разрабатываемой системы, и изменения в одной части такого проекта могут затронуть по цепочке многие модули.

Вообще, подход к программированию по принципу "сверху вниз", рекламируемый в руководствах по структурному и

систематическому программированию на базе небольших примеров, а также популярный среди руководителей промышленных разработок, имеет ряд недостатков, проявляющихся именно на больших разработках. Отдельные части системы при таком подходе пишутся в расчете на использование только в специфическом для этой системы окружении и даже не могут быть полноценно оттестированы, пока не готова система в целом. Связи между отдельными элементами большой системы настолько перепутываются, что руководитель не может уследить за ними. По существу, разработка, превысившая некоторый критический уровень сложности, обречена на неудачу (как это признают, например, некоторые американские разработчики программного обеспечения стратегической оборонной инициативы). Использование специальных технологических систем, хранящих информацию о межмодульных связях в особым образом организованной базе данных, может несколько поднять границу сложности, свойственную такому методу организации разработок, но принципиально не меняет ситуации.

Представляется, что "барьер сложности", возникающий в связи с созданием больших программных систем, можно преодолеть, лишь систематически повышая исходный уровень, над которым надстраивается новая разработка, т.е. создавая и наращивая базу из постоянно используемых и хорошо опробованных программных модулей и инструментальных средств программирования. По сути дела, именно такой процесс накопления базы и обеспечил то, что нынешние возможности программирования существенно превосходят наш уровень, скажем, тридцатилетней давности.

Отсюда следует, что внимание и материальные ресурсы следовало бы направлять в первую очередь на развитие общих и инструментальных средств программирования, а не на "масштабные" уникальные разработки. Однако утверждению такого подхода не в последнюю очередь препятствует престижность крупных разработок и прочие выгоды, которые получает руководитель такой разработки, становясь главой большого коллектива. (По существу эта тенденция представ-

ляет собой разновидность затратного подхода в экономике.) Критическое отношение к "масштабным" разработкам встречается и в зарубежной литературе.

Если считать, что в идеале один модуль должен отражать одну специальную идею, один метод, то в любой программе можно увидеть, что вместо выражения этой идеи в чистом виде, в форме самостоятельного отрезка программы приходится на каждом шагу перемешивать в программе кусочки текста, мотивированные совсем не связанными друг с другом идеями. Достаточно сказать, что языки обычно требуют, чтобы были точно заданы тип и структура объектов, над которыми производятся действия, хотя для самой идеи, лежащей в основе этих действий, этот тип не имеет значения (сортировка, действия со списками и т.п.). Частичный выход из этого затруднения предлагает АДА в виде родовых объектов. Языки с абстрактными типами данных позволяют не определять тип в этом месте, но требуют, чтобы он был определен заранее. В целом на уровне традиционных языков программирования о четком разделении программы на исходные "идеи" не приходится и думать.

Если, несмотря на все недостатки, в написании конкретных модулей и пакетов все же утверждается установка на независимость и многократную используемость, то в разработке базовых средств программирования установка на глобальность до сих пор не преодолена. Концепции баз данных, с самого начала направленные на обеспечение независимости данных от конкретных прикладных программ, требуют, как правило, от каждой конкретной базы данных единой заранее разработанной структуры. Но в наибольшей степени установки на глобальность проявляются в языках и системах программирования, которые традиционно мыслятся как нечто всеобъемлющее, способное самостоятельно удовлетворить весь спектр потребностей пользователя. Такой подход отразился и в большом количестве штампов, бытующих в оценках языков программирования.

История развития языков программирования представляется с таких позиций непрерывным восхождением к все более и более универсальным, мощным и сложным языкам: ФОРТРАН, КОБОЛ, АЛГОЛ 6Ø, ПЛ-I, АЛГОЛ-68, АДА... Однако в такую концепцию не укладываются почему-то существующие и широко распространяющиеся языки с совсем несложными описаниями: ФОРТ (FORTH), РЕФАЛ, ПРОЛОГ, СМОЛТОК-8Ø (SMALLTALK-80). Можно было бы связать это с тем, что все это языки "нового принципа", уходящие в той или иной форме от фон-неймановских традиций, и что, может быть, просто они, вступив на новый путь, еще не успели набрать необходимой сложности. Однако и такой вполне традиционный по средствам язык, как МОДУЛА-2, также демонстрирует тенденцию языков к упрощению. Невольно приходит на ум сравнение с появлением первых млекопитающих в мире динозавров, использованное А.П.Ершовым в применении к персональным ЭВМ.

Языки программирования традиционно оцениваются в терминах "уровня". Однако понятие языка высокого уровня не однозначно. Под этим можно понимать либо использование синтаксических конструкций, позволяющих коротко записывать то, что длинно выглядит в объектных кодах, либо возможность рассматривать в виде единого объекта сложные структуры данных, либо освобождение пользователя от принятия решений технического характера (выбор конкретных команд или регистров, распределение памяти, выбор конкретного способа представления сложных данных), либо возможность использовать в готовом виде программы, выполняющие сложные действия над данными и, далее, самостоятельно пополнять запас таких программ, либо возможность поручить системе поиск пути к цели, оставляя за пользователем лишь задание этой цели, и т.п. Эти черты присутствуют в разных языках "высокого уровня" в разной степени: язык может иметь высокий уровень в одном отношений и низкий - в других. Различны и потребности разных пользователей в разных аспектах "высокого уровня": для программиста, решающего конкретную прикладную задачу, важно использование готовых подпрограмм, а для разработчика системного программного обеспечения

существенно, чтобы при сокращении технической работы по кодированию он мог сохранять контроль над представлением и размещением информации. Так что совсем не очевидно, что все эти разнообразные возможности необходимо сводить воедино в одном "супер-языке".

Язык программирования оценивают по представленному им запасу встроенных возможностей: можно ли в нем описывать массивы заранее не известного размера, допускаются ли рекурсивные процедуры, вариантные структуры, анализ строк по образцу, параллельные процессы, возможен ли доступ к командам операционной системы и т.п. Стремление заранее определить весь комплекс встроенных возможностей языка и свести все это в единую систему оказывается нелегкой задачей, и именно оно приводит к разрастанию языков до размеров динозавра. Однако мало кто задавался вопросом, зачем вообще нужно сводить все необходимые приемы программирования в единый нерасторжимый комплекс и не лучше ли дать потребителю возможность самому выбрать, что из всего этого ему нужно. Иначе говоря, нельзя ли низвести то, что сейчас рассматривается как фундаментальная конструкция языка, до положения отдельного модуля. Систематизация и ортогонализация запаса накопленных программистских конструкций сохраняют свою важность, но они перестают быть обязательной частью разработки системы программирования. Пока что, кажется, разработчики систем программирования готовы предложить рецепты модульности всем, кроме самих себя.

Можно оценивать язык программирования по направленности на определенный класс задач. Такая направленность обычно выражается в наличии готовых средств для часто встречающихся в этой области ситуаций. Современные универсальные языки в значительной степени преодолели необходимость в ориентации на частный класс задач, поскольку позволяют доводить универсальный язык до уровня специализированного путем добавления проблемно-ориентированного набора типов данных и подпрограмм. Поэтому не случайно при подготовке к разработке языка АДА выяснилось, что нет необходимости создавать отдельный язык для каждого ведомства. Другое дело, существует ли фактически необходимый запас готовых

процедур для некоторой прикладной области или его еще предстоит разработать. Наличие запаса готовых средств и традиций пользования языком действительно дает основания для предпочтения определенного языка в некоторой области (как ФОРТРАНА у физиков).

С другой стороны, приступая к работе в новой области или к созданию программного обеспечения специализированной системы, мы должны сами выбрать систему основных понятий и действий, которая будет отражена впоследствии в наборе типов данных и процедур. Это наиболее сложная и трудоемкая часть работы по освоению новой области, и если она сделана удачно, то потом уже гораздо легче программировать конкретные задачи.

Поэтому важную роль в оценке системы программирования играет то, как она позволяет выполнить первый, наиболее сложный этап работы, т.е. какие она допускает способы членения задачи на модули и оформления основных концептуальных единиц проблемной области в виде единиц системы программирования. Во всяком случае, эта характеристика языка важнее, чем предлагаемый им конкретный набор встроенных средств. Точно так же конкретное синтаксическое оформление программ, хотя и важно с психологической точки зрения (обозримость, предупреждение случайных ошибок), играет все же второстепенную роль по сравнению с возможностями организации модульности. В связи с этим вызывает сожаление сравнительное однообразие концептуальных средств модульности в распространенных языках программирования при пестроте синтаксического оформления и частных встроенных средств. Редко доступны или доступны в затрудненном режиме такие концептуальные единицы, как окружение (комплекс заранее инициализированных данных и процедур, аналог пакета в языке АДА), потоки (последовательности однотипных элементов, динамически порождаемых и обрабатываемых по одному), сопрограммы, запросы к динамическому окружению (т.е. получение процедурой информации, доступной в месте ее вызова). Отсутствие сопрограмм особенно затрудняет объединение мо-

дулей, потому что часто оказывается, что каждый из объединяемых модулей проще пишется, если именно его рассматривать как главную программу.

При глобальном подходе система программирования предлагает пользователю некоторый собственный мир понятий, в который пользователь должен полностью вложить свою задачу. А что этим понятиям соответствует в памяти машины или на машинных носителях, не всегда даже разъясняется пользователю. Навязывая определенные соглашения о способе использования ресурсов машины (памяти, регистров, режимов управления), система крайне затрудняет любое использование машинных возможностей вне этих рамок, в том числе взаимодействие с другой системой программирования. В частности, жесткое включение в язык программирования системы операций ввода/вывода может сильно затруднить обмен данными с другими системами.

В связи с этим поучителен опыт разработки и использования АЛГОЛА 68. Оказавшись вполне удобным средством для программирования задач, решаемых математическими методами, он вызывает разнообразные трудности. Как только приходится устанавливать связи программы с внешним миром, жестко определенная, вплоть до каждой проверки, система ввода/вывода может не согласовываться с особенностями реальной операционной системы, делая обмен в нужном формате либо со специализированными периферийными устройствами неэффективным или вовсе невозможным. Требования к элементарным типам данных и операциям над ними могут не согласовываться с возможностями оборудования или препятствовать их эффективному использованию. Установка на жесткий видовой контроль затрудняет взаимодействие с СУБД: хотя каждая конкретная база данных может использовать жестко определенные типы и структуры данных, сама СУБД работает с произвольными типами. Равным образом жесткий видовой контроль затрудняет описание на языке и обработку сложных структур данных с объектами разных типов (например, структуры программы на самом АЛГОЛЕ 68). Не содержит АЛГОЛ 68 и возможности обращения ко многим функциям операционной системы, в частности,

не определены средства обеспечения комплексирования программ с использованием стандартных средств операционной системы.

На практике, чтобы использовать преимущества АЛГОЛА 68 в реальных ситуациях, приходится разрабатывать на машинном языке дополнительные интерфейсы со средой и оформлять их в виде дополнительных процедур АЛГОЛА 68, а в отдельных случаях вносить и более крупные изменения в реализацию.

Установка на завершенность АЛГОЛА 68 привела и к другим трудностям. Международная рабочая группа, разрабатывавшая АЛГОЛ 68, постановила "заморозить" некоторое определение языка, и в результате оказались в двусмысленном положении разработанные впоследствии той же группой дополнения к языку (кстати, вводимые в этих дополнениях правила комплексирования программ, подобно соглашениям о вводе/выводе, полностью автономны и могут не согласовываться со средствами операционных систем, направленными на те же цели).

Влияние "глобалистских" тенденций проявляется даже в тех системах, которые специально направлены на разработку новых принципов модульности. Система ПРИЗ предназначена для построения программ на базе описания задачи в терминах вычислительных моделей. Она включает в себя решатель, который выбирает путь вычислений и строит программу решения задачи, и сам аппарат вычислительных моделей, позволяющий формировать исходные данные для этого решателя. Вычислительные модели представляют собой описание конкретной проблемной области, они являются достаточно независимыми и сменяемыми модулями этой системы. Напротив, решатель был задуман как единое, универсальное, уникальное в системе средство. В результате в ходе развития этой системы именно решатель приходилось неоднократно менять, приспосабливая его к все новым типам задач (решение систем уравнений, условные вычисления, рекурсии и т.п.). Однако ясно, что универсального механизма поиска решений для всех задач все равно не будет. Почему же не сделать решатель тоже сменяемым модулем и, может быть, допустить много взаимодействующих решателей?

Разновидностью требования многократной используемости программ является требование переносимости программы в другую операционную среду, на другое оборудование. На первый взгляд, наиболее надежным средством решения этой задачи представляется задание жестко определенного интерфейса между программой и средой, своего рода виртуальной машины, которая бы выступала перед программой в роли представителя среды и оставалась бы неизменной при любой замене среды. Однако отгородиться от всех изменений внешней среды подобной непроницаемой стеной не удается. Если в новом окружении старая система не может продуктивно использовать новые возможности, то она становится не очень полезной. (Автор этих строк однажды был свидетелем того, как на большой машине ЕС пропускали старую программу начисления зарплаты, написанную еще в системе ДОС: приходилось заменять тэзы в стойке управления дисками, чтобы машина воспринимала большие диски как 7-мегабайтовые; естественно, вся остальная работа на машине на это время прекращалась.) Уже упомянутая практика использования АЛГОЛА 68 показывает, как язык, первоначально задуманный как раз в стиле непроницаемого интерфейса, стал использоваться скорее как система "без нижнего уровня", позволяющая надстраивать привычные языковые конструкции над проблемно-ориентированной базой, разрабатываемой каждый раз самостоятельно.

Разновидностью переносимости является приспосабливаемость системы программирования к разнообразию символьных наборов реальных устройств подготовки и отображения информации, а также к использованию ключевых слов и идентификаторов на разных языках. Даже для такой сравнительно простой проблемы попытка ее решения по принципу закрепленного "виртуального" алфавита или набора ключевых слов встречается с серьезными трудностями.

Конечно же, реальный путь к переносимости не в том, чтобы закрываться от всех изменений среды, а в том, чтобы, наоборот, разрешать учитывать их в своей программе.

Стремление к независимости и свободной сочетаемости элементов программы пробивало себе путь в различных реальных разработках, иногда в результате сознательной установки разработчиков, иногда в качестве случайных удачных находок.

Интересен опыт развития операционных систем типа UNIX и разработанных в их рамках разнообразных инструментальных средств. Эти системы характеризуются не только переносимостью, т.е. высокой степенью независимости от конкретной архитектуры используемой ЭВМ, но и большим практическим опытом создания модулей, легко сочетающихся друг с другом без предварительной подготовки и легко переносимых в другое окружение. Немалую роль в этой переносимости сыграл, конечно, язык "С", непосредственно ориентированный на машинные структуры данных, а не скрывающий их от программиста собственными идеализированными структурами. Но дело не только в языке "С". В UNIX'Е сложились определенные традиции организации входных и выходных данных программ, ориентированные на естественную стыковку программ через файлы или "трубы", и имеется набор операций ввода/вывода, согласованный с этими традициями. Большую роль в приспособляемости программ к различным окружениям играют средства макрогенерации, которые дают возможность окружению управлять исходным текстом для транслятора. В частности, они позволяют назвать в первоначальном тексте "идею" некоторого действия просто словом, а окружение обеспечит для этого слова нужное макрорасширение. Таким образом, требуемое для окончательной программы переплетение "идей", пришедших из разных источников, обеспечивается посредством действий над текстом программ. Вообще, опыт и традиции систем UNIX учат ориентации прежде всего на символьную запись программ и данных как на их наиболее жизнеспособную и переносимую форму. С двоичным форматизованным представлением программ и данных мы имеем дело лишь после работы транслятора, и эта форма представляет лишь локальный интерес.

Важно отметить, что переносимость средств, разрабатываемых в системе типа UNIX, основывается не на жестко аксиоматизированной модели "виртуальной машины", правильная работа которой гарантируется лишь в случае точной ее реализации на любом оборудовании, а совсем напротив, на свободном сочетании и взаимодействии машинно-зависимых и машинно-независимых средств, при котором заранее не известно, что при переходе в новое окружение сохранится, а что заменится. Переносимость обеспечивается не тем, что система отгораживается от изменений окружения, а тем, что она хорошо расчленена на элементы, и изменения в окружении затрагивают лишь немногие из них.

Конечно, структура программы в UNIX'E еще очень далека от идеала расчленения программы на независимые "идеи", поскольку язык "С" (без макросредств) достаточно традиционен по своей организации, надо к тому же отметить, что и существующая степень переносимости ведет к определенному снижению эффективности по машинным ресурсам.

В развитии языков программирования хорошо известно направление, прямо нацеленное на выработку понятия модуля. Предлагаемые понятия модуля различаются во многих деталях, но объединены тем, что в последнее время называется объектно-ориентированным подходом. В качестве модуля выделяется не столько действие (процедура), сколько объект, включающий в себя определенные данные и процедуры, работающие с этими данными. Разные языки предлагают разные варианты этого понятия, отличающиеся статическим или динамическим формированием объектов, явной или неявной привязкой объекта к классу подобных ему объектов, деталями взаимодействия между объектами, но сам факт, что языки, исходившие из разных первоначальных установок (СИМУЛА, УТОПИСТ, КЛУ, АДА), выработали близкие понятия, свидетельствует о том, что объектно-ориентированный подход отражает какие-то реальные особенности структуры человеческого знания. Работы в области искусственного интеллекта также привели к выделению аналогичной по структуре единицы знания, называемой фреймом.

Вместе с тем реальные языки, реализующие объектно-ориентированный подход, в остальных аспектах часто остаются достаточно традиционными и не обеспечивают желательного уровня свободной сочетаемости модулей.

Традиционно в языках программирования сильны возражения против использования связей между модулями посредством символьных имен во время выполнения программы; считается, что в интересах эффективности все такие символьные связи должны еще на фазе трансляции или сборки заменяться машинными адресами или чем-нибудь подобным. За этим взглядом стоит, видимо, не только стремление к эффективности, но и убежденность в том, что имя всегда заменяет некоторый конкретный объект и имеет смысл лишь в связи с этим объектом. Этому подходу удобно следовать, пока мы остаемся в пределах математических задач, на базе которых и выросли традиционные языки программирования. Но реальные ситуации, возникающие за пределами таких задач, демонстрируют необходимость более широкого использования имен, в частности, установления связи между независимо построенными программными объектами путем использования одинаковых имен. Такой способ широко используется в операционных системах (файловые системы и основанные на них процедуры, разрешение взаимных ссылок между программами), в макрогенераторах (глобальные переменные).

Общий принцип здесь таков, что имя не связывается с конкретным объектом, а служит наименованием роли или функции, используемым для того, чтобы в разных ситуациях находить нужный объект, каждый раз свой, но одинаковый по назначению. В традиционных языках программирования этот прием также постепенно пробивает себе дорогу; выделение поля в структуре по имени, реакция на исключительные ситуации (одно и то же имя ситуации в разных окружениях вызывает разные реакции), родовые объекты в языке АДА. В системах, предназначенных для задач искусственного интеллекта, такое использование имен имеет уже давнюю традицию (язык ЛИСП, семантические сети).

Такое использование имен не нужно отождествлять с манипулированием строками: в памяти машины имя может быть представлено произвольным внутренним значком или адресом, его символьное представление нужно только для отождествления вне машины или при передаче информации между машинами. Это согласуется и с уже упомянутым подходом системы

Влияние старых традиций проявляется и в сложившихся представлениях о желательном размере модулей. Попытка строго расчленить текст программы на исходные "идеи", хотя бы при помощи уже существующих форм модульности, покажет, что, во-первых, обычный модуль (или процедура), содержит не одну "идею" и, следовательно, оптимальный размер модуля должен быть существенно меньше и, во-вторых, что связи между такими оптимальными модулями выглядят в традиционной записи очень громоздко (снова представим себе сортировку массива, элементы которого определены посредством абстрактного типа данных и каждое обращение к ним представляет собой вызов процедуры; а ведь и сама сортировка - это тоже еще не самая элементарная идея). Такого рода мысленные эксперименты приводят к представлению, что идеальный модуль должен содержать не более, чем несколько строк программного текста. Кроме того, становится ясно, что сборка программы из таких модулей приведет к большим накладным расходам на межмодульных стыках. Впрочем, последнее можно оценить и так, что современная техника трансляции уделяет больше внимания и достигает лучших результатов при внутримодульной оптимизации, чем при межмодульной. Это обстоятельство является одним из наиболее серьезных препятствий для приближения к оптимальному размеру модуля.

Необходимо сразу оговориться, что речь должна идти не столько о малом размере модуля в целом, сколько о малом размере единичной процедуры в модуле. Модуль может содержать не одну, а много взаимосвязанных процедур. А кроме того, еще и большое количество информации, необходимой для процесса сборки.

Заслуживает особого внимания тот факт, что уже упоминавшиеся небольшие, но жизнеспособные языки, при всех различиях в исходных принципах, характеризуются все как раз малым размером практически используемых процедурных единиц. Одно из руководств по языку ФОРТ прямо рекомендовало ограничивать одно определение "слова" (т.е. процедуры) не более, чем 7 + - 2 словами, имея в виду "психологическую константу" 7 + - 2, характеризующую объем кратковременной памяти у человека.

Малый размер процедур в этих языках стал приемлемым из-за того, что основные действия внутри процедур реализуются на обычных машинах путем интерпретации и потому не превосходят по эффективности межмодульные сопряжения. Другой подход проявляется в языке COL-2, где элементарные действия достаточно просты и традиционны, но предусмотрены развитые средства оптимизации программы в процессе ее сборки из модулей. Это обстоятельство также позволило пользоваться в этом языке процедурами размером в несколько действий.

Так что не исключено, что причина жизнеспособности всех этих языков не в отходе от принципов Дж. фон Неймана, а в высокой степени модульности.

Малый типичный размер процедуры во всех этих языках привел и к другому результату: отпала необходимость в сложном синтаксисе. В таком языке, как ФОРТ, синтаксис практически сведен на нет. Сложный синтаксис в "больших" языках мотивирован частично тем, что считается необходимым давать самостоятельное синтаксическое оформление каждому встроенному средству (например, вводятся особые обозначения для элементов массивов и для полей структур), частично просто желанием разнообразить оформление длинных процедур для облегчения восприятия текста (в АЛГОЛЕ 68 - разный синтаксис для формул и вызовов, выделенный и краткий стили написания условных и других предложений). В новых языках с упрощением синтаксиса отпадает необходимость в разработке громоздких трансляторов. На транслятор теперь можно

смотреть не как на уникальный элемент системы программирования, а тоже как на один из модулей. Уже не обязательно, чтобы все возможности системы обеспечивались одним транслятором. Можно, например, иметь один транслятор с входного языка, обеспечивающего доступ ко всем возможностям оборудования или операционной системы, и другой транслятор с машинно-независимого языка. Можно иметь входные языки и трансляторы для пользователей с разной квалификацией и сферой интересов, лишь бы все трансляторы давали совместимые по интерфейсу модули.

Может возникнуть сомнение в том, что разнообразные средства современных языков "высокого уровня", вводимые в них уникальным образом, с собственными особо описанными синтаксисом и семантикой, вообще могут быть сведены на уровень рядовых модулей. Здесь необходимо различать две стороны: обеспечение правильного выполнения требуемых операций и экономное использование ресурсов машины. Если иметь в виду только правильное выполнение действий, то ответ безусловно положительный. Например, вычислительная модель К.Хьюитта ("плазма"), сводящая все интерфейсы к простому механизму передачи сообщений между исполнителями ("экторами"), причем сами сообщения тоже являются исполнителями, успешно демонстрирует такую возможность. Даже понятие памяти, считающееся центральным в некоторых подходах к описанию семантики языков программирования, оказывается в модели К.Хьюитта лишь свойством определенных исполнителей, а многие привычные средства, например, циклы, воспроизводятся без использования памяти. Система СМОЛТОК-80, также основанная на концепции посылки сообщений, является практической системой, уже накопившей большой запас модулей для различных приложений.

Другой вопрос - как быть со средствами, направленными на машинную эффективность, например, со статической обработкой описаний типов данных. Мы уже видели, что подчинять всю работу статическим ограничениям на типы данных неудобно; в то же время хранить всегда информацию о типе данных во время обработки, выполняя нужные про-

верки и преобразования динамически, крайне неэффективно. Аналогичные трудности связаны и с использованием массивов (заранее известного или заранее неизвестного размера) и операциями над ними, включая контроль значений индексов. Можно, наконец, включить в эту группу все вопросы о контроле корректности данных, получаемых тем или иным модулем. Например, программа извлечения квадратного корня должна производить проверку знака аргумента, хотя эта проверка становится избыточной, если квадратный корень употребляется для вычисления расстояния между двумя точками по теореме Пифагора.

Природа затруднений во всех этих ситуациях одинакова. Решать их всегда в пользу машинной эффективности, налагая на автора программы жесткие ограничения или отключая динамические проверки после того, как программа "отлажена", нецелесообразно: наиболее дефицитным ресурсом в настоящее время является не машина, а время разработчика. Поэтому, если мы все же заинтересованы в оптимизации, надо научиться перекладывать ее на систему программирования. Концептуально операции над описаниями типов данных, а также разнообразные проверки на корректность должны всегда присутствовать в модулях. Дело системы - выделять случаи, когда эти действия можно перенести на фазу трансляции или сборки, исключив из готовой программы даже следы этих операций. Возможность такого подхода, правда, в небольшом масштабе, демонстрируется системой COL-2. В развернутом виде этот подход, если его удастся разработать, потребует широкого использования техники смешанных вычислений и трансформации программ с учетом присутствующих в них логических условий. Исходные логические условия для таких преобразований и оптимизаций должны, очевидно, присутствовать в самих объединяемых модулях. Таким образом, модуль становится не просто процедурной единицей программистского знания, но и включает в себя обширный запас сведений об интерфейсах в декларативной, в том числе логической форме.

Еще один очевидный вопрос - это можно ли обеспечить использование разнообразных модулей, не сформировав

единой базы с едиными жесткими соглашениями об интерфейсах. Не получится ли при этом, что, отрицая необходимость всеобъемлющей системы программирования, мы однако же требуем всеобъемлющих соглашений для сборки программ из модулей? Это действительно так, но лишь частично. Во-первых, базовые соглашения об интерфейсах, обязательные для всех модулей, должны в каждой конкретной реализации быть сведены к минимуму. Во-вторых, существующий опыт переносимости программ показывает, что нижний, базовый уровень тоже может быть сменяемым при сохранении взаимоотношений между модулями на верхних уровнях. И наконец, система сборки, удовлетворяющая высказанным требованиям, не должна быть и, скорее всего, не будет единственной в мире, так же, как не выработался единственный в мире язык программирования. Будет много систем, лишь частично совместимых по интерфейсу, будет процесс установления связей между этими системами и постепенного их сближения. Подчинение единому центру здесь совсем не обязательно. Для сравнения можно сослаться на мировую железнодорожную или телефонную сети, которые также не имеют единого проекта.

Другая аналогия - это структура обычного человеческого языка, где сосуществует много частных систем и правил, а многие правила, носящие общеязыковой характер, тем не менее имеют исключения. Перспектива иметь дело со столь неопределенной системой может показаться неудовлетворительной. Но если мы замахиваемся на такую задачу, как объединение всех человеческих знаний, представимых в машинной форме, то трудно рассчитывать на более простое решение, чем то, которое человечество уже выработало в применении к своей обычной деятельности.

Таким образом, для перехода к сборочному программированию на базе многократно используемых модулей нужно, по крайней мере, освоить уже известные принципы модульности, а именно:

- объектно-ориентированный подход;
- межмодульные интерфейсы в символьной форме;
- стиль программирования в модулях с процедурами малого размера.

В различных существующих системах эти возможности встречаются чаще всего порознь, или же вместе, но с потерей машинной эффективности.

По мере освоения этих возможностей могут, естественно, обнаружиться и новые направления.

Вот некоторые задачи, решение которых потребуется для создания техники сборочного программирования и поддерживающей его базы программных модулей.

I. Выработка стиля программирования, соответствующего принятым принципам модульности. Это не простая задача, потому что очень трудно расчленять то, что по привычке мыслится как единая программа традиционного стиля, на ее исходные концептуальные элементы. Если попытаться изучить с этой целью единичную программу, то, скорее всего, мы получим лишь тривиальное разложение на структурные части. Лишь когда рассматривается целый класс задач и известно, какие элементы постановки задачи могут варьироваться, можно прийти к действительно содержательному анализу. Эта задача по трудности аналогична задаче систематизации знаний в какой-либо области и выделения удобной системы основных понятий. Эта задача выглядит несколько по-разному в применении к общим приемам программирования и к конкретным прикладным областям.

I.I. В применении к общим приемам программирования эта задача предполагает выделение и оформление в качестве самостоятельных модулей некоторого общепринятого инвентаря приемов программирования, структур данных и основных алгоритмов. Здесь могут выделиться и такие единицы, которые не представлены в распространенных системах программирования, но обычно присутствуют в интуиции программиста, например, уже упоминавшиеся потоки. В выделяемых элементах придется учитывать не только процедурную, но и логическую информацию, необходимую для оптимизации в процессе сборки.

I.2. В специализированных областях задача выделения концептуальных модулей для программирования несколько проще, потому что в конечном счете такие модули должны соответствовать понятиям, используемым специалистами в этой области. Поэтому в таких областях задача разработки модульной базы для программирования равносильна задаче разработки средств программирования, ориентированных на специалистов. В целом ряде областей такую задачу удается успешно решать. Так, в задачах имитационного моделирования мы можем создать базу модулей, каждый из которых соответствует некоторому типовому элементу моделируемых систем, так что составление задания на моделирование для машины аналогично сборке моделируемой системы из элементов. Аналогичным образом облегчается работа разработчика трансляторов при использовании "систем построения трансляторов". Существуют также системы автоматизации конторской работы, успешно предоставляющие пользователю возможность работать в собственных понятиях.

2. Повышение эффективности межмодульных интерфейсов. Здесь возможно несколько не исключающих друг друга направлений.

2.I. Аппаратурные решения, т.е. разработка процессоров, реализующих межмодульные сопряжения схемно или микропрограммно.

2.2. Глубокие оптимизирующие преобразования программы в процессе сборки, в том числе отбрасывание неиспользуемых ветвей и проверок с заранее известным результатом, закрепление статических структур данных, статическое разрешение взаимных ссылок, выполнение действия над константами. В ряде случаев для проверки допустимости преобразований потребуется учитывать и преобразовывать логические условия, взятые из модулей. Сами алгоритмы преобразований тоже могут содержаться в модулях.

2.3. Разработка средств оптимизации генерируемого объектного кода для машин традиционной архитектуры, не привязанных жестко к конкретному типу машин.

3. Ведение большой базы программных модулей, идентификация модулей. Автоматическая и автоматизированная проверка взаимной пригодности модулей по описаниям интерфейсов, поиск модулей с требуемыми функциями. Учет частичной взаимозаменяемости модулей, версий одного модуля. Модули с "запасом прочности", т.е. с расширенным против обычно требуемого набором функций и проверок. Распределенная база модулей, совместимые базы на машинах различной архитектуры.

Высказанные здесь мысли неоднократно излагались в различной форме и обсуждались автором в течение последних нескольких лет. Очень полезными для понимания затронутых здесь вопросов были обсуждения на семинарах и рабочих группах, руководимых С.С.Лавровым, Э.Х.Тыугу, Е.А.Жоголевым, И.В.Поттосиным, работа в рабочих группах по языку АЛГОЛ 68 и операционной системе UNIX личные беседы с Л.К.Горским, Л.Б.Гримзе, А.Я.Диковским, А.П.Ершовым, А.Н.Масловым, С.Г.Масловым, В.Л.Темовым и многими другими. Автор надеется на дальнейшее обсуждение поднятых здесь вопросов и будет благодарен за любые отклики.

М.М.Железняков, Т.Н.Невлева,
И.М.Новицкая, Л.Н.Смирнова,
Г.С.Цейтин

ОПЫТ ПОСТРОЕНИЯ МОДЕЛИ ТИПА "ТЕКСТ - ДЕЙСТВИТЕЛЬНОСТЬ" С ИСПОЛЬЗОВАНИЕМ АССОЦИАТИВНЫХ СЕТЕЙ

I. Постановка задачи

Эта работа направлена на построение модели, описывающей семантику языка на достаточно глубоком уровне, вплоть до взаимодействия с действительным миром, и пригодной для построения практических систем, управляемых на естественном языке.

Необходимость выхода на уровень действий системы, воспринимающей текст, принципиально отличает эту постановку задачи от традиционных подходов, при которых исследование семантики сводится к толкованию смысла некоторых слов или конструкций посредством других слов или конструкций. С другой стороны, исследование реакции системы на введенный текст выглядит с точки зрения исследования семантики более узкой задачей, чем установление связей текста с системой понятий, описывающей действительный мир. Однако любая система понятий сама является результатом некоторой абстракции, до известной степени произвольной и зависимой от языка. Эти понятия не могут быть выражены иначе, чем посредством языка, так что есть большой риск, сводя смысл к описанию в терминах этих понятий, ограничиться простым перифразированием. Наконец, подход к исследованию семантики, основанный на изучении преобразований, сохраняющих смысл, без попытки раскрытия этого смысла в какой-то иной форме, упирается в трудности, связанные с тем, что некоторые виды смысловой эквивалентности связаны со знанием реальной ситуации: искусственное отделение таких видов эквивалентности сужает возможности исследования смысла. Кроме того, для создания практических систем исследования эквивалентности только на лингвистическом уровне недостаточно.

Известно, что во многих случаях при анализе текста без знания внеязыковой ситуации невозможно даже получить однозначное представление текста на морфологическом и синтаксическом уровне, поэтому в данной работе не ставится вопрос о построении автономного языкового процессора, учитывающего только языковые законы, с вынесением всей неязыковой обработки на последующую стадию. Лингвистические и нелингвистические знания будут здесь использоваться и взаимодействовать на всех этапах обработки.

Изучение смысла через моделирование реакций на текст имеет и существенный недостаток. Сегодня мы можем рассматривать осмысленное реагирование на текст лишь в пределах весьма узкой области, в которой мог бы функционировать реальный или воображаемый автомат. Соответственно, естественный язык сведется к очень узкому фрагменту, включающему лишь те высказывания, которые уместны и могут быть адекватно восприняты при работе с рассматриваемой системой. Это не просто ограничение на тематику текста или его словарь, а очень жесткое требование, делающее для большинства мыслимых сегодня автоматических систем соответствующий фрагмент языка просто малоинтересным. Положение осложняется тем, что при преобладающем сейчас подходе к разработке автоматических систем построению практической автоматической системы предшествует составление жесткого перечня ее функций, что влечет за собой по существу (а иногда и по форме) построение жесткого формального языка для обращения к системе. В этой ситуации использование естественного языка, даже когда оно технически возможно, становится неудобным для пользования системой.

Есть, тем не менее, ряд оснований для того, чтобы и в этих условиях исследовать механизмы функционирования естественного языка или его подъязыков. Во-первых, опыт показал, что даже для очень простых областей (подобных рассматриваемым в настоящей работе) переход от текста к функционированию системы требует решения новых лингвистических задач, ускользающих от внимания

при традиционном подходе. Во-вторых, практика построения автоматических систем на основе жесткой предварительной постановки задачи постепенно изживает себя: возникает необходимость создания систем с расширяемым набором функций и объединения разнородных независимо разработанных систем, а в этой ситуации жестко определенные формальные языки уже не столь эффективны, особенно при большой сложности рассматриваемой системы. В-третьих, есть основания считать (Цейтин 1982, 20-34), что и вообще язык представляет собой объединение многих относительно независимых подсистем для частных областей и ситуаций, поэтому от изучения функционирования языка на частных областях мы сможем прийти к пониманию общих законов языка на семантическом уровне.

В целом опыт работы в этом направлении, которое по аналогии с известной моделью "смысл - текст" можно назвать задачей "действительность - текст", или, конкретнее, "текст - реакция", не убеждает в том, что получение глубинного смыслового представления является необходимой промежуточной ступенью при переходе между действительностью и текстом. Это связано с тем, что необходимые в нашей задаче преобразования зависят как от языковых, так и от неязыковых знаний, и нет четкой границы между фазой, управляемой только языковыми знаниями, и фазой, управляемой только знаниями из конкретной предметной области вместе с описанием текущей ситуации.

В выборе направления производимых преобразований существенную роль играет целевая установка, определяемая характером решаемой задачи. В случае узкой предметной области эта установка в основном определяется самим выбором области (реакцией должна быть выдача управляющего воздействия, или запуск программы, или поиск решения задачи, или поиск документа по признакам и т.п.). Конечно, в пределах одной области эта установка тоже может меняться. От целевой ус-

тановки зависит не только выбор направления преобразований, но и то, какие признаки обрабатываемых объектов следует учитывать как существенные, а также то, какие комплексы языковых единиц нужно, с точки зрения данной задачи, рассматривать как одно целое, а какие преобразовывать порознь.

Из-за этого не представляется возможным определить некоторый всеобщий, не зависящий от конкретной задачи, процесс нормализации смыслового представления после получения глубинной синтаксической структуры (а в предлагаемой модели и синтаксический анализ не является независимым этапом). Вместе с тем все преобразования, выполняемые на пути от текста к реакции, можно описывать на языке "смыслов" и толковать как преобразования на смысловом уровне, однако эти преобразования не будут эквивалентными преобразованиями с точки зрения языка, потому что в них постоянно используются неязыковые знания. Определить "главное" смысловое представление текста, эквивалентное ему по смыслу, вероятно, не удастся.

В данной работе рассматривался ряд конкретных предметных областей, хотя такое рассмотрение большей частью не доводилось до полной машинной реализации. Одна из таких областей, доведенная до действующей программы понимания текста, связана с жесткой задачей управления небольшим пакетом прикладных программ при дополнительном ограничении, что каждая языковая команда должна соответствовать запуску ровно одной программы из пакета. Тем не менее, в связи с неготовностью реального пакета на момент исследования этой предметной области, некоторые требования к задаче приходилось домысливать в процессе разработки.

Некоторые предметные области рассматривались без детальной проработки лингвистической части, анализировалось только соответствие между понятиями (терминами) предметной области и выработкой реакции машины. Одна из таких областей - построение движений идеализированного робота - рассмотрена в статье: Цейтин 1985, 16-48.

Довольно богатый лингвистический материал представляет такая область, как арифметические задачи "на движение". Она не может рассматриваться как жестко определенная область из-за большого разнообразия сюжетов задач и неявно заданной, но необходимой для решения задачи информации, основанной на житейских знаниях, связанных с сюжетом задачи. Исследование на основе такой области проводилось в двух направлениях. Во-первых, не строя законченной модели области, можно "домысливать" описание области для каждой конкретной рассматриваемой задачи и на этой основе делать лингвистические наблюдения и строить рекомендации по структуре модели. Во-вторых, можно пойти по пути ограничения словаря и запаса сюжетов, знание которых необходимо для понимания задачи.

В конечном счете в этой области система должна прийти от формулировки задачи к ее решению. Промежуточным этапом служит преобразование задачи к системе взаимосвязанных "вычислительных моделей", каждая из которых описывает типовую ситуацию и взаимосвязи между встречающимися в ней величинами. Использованы, в частности, следующие вычислительные модели:

- равномерное движение (связь между скоростью, продолжительностью и путем);
- интервал времени (связь между начальным моментом, конечным моментом и продолжительностью);
- отрезок пути (связь между концевыми пунктами и величиной пути);
- составное движение, т.е. два последовательно выполняемых движения (связь между характеристиками движения в целом и составляющих движения).

Переход от системы вычислительных моделей к решению поставленной арифметической задачи общеизвестен (Тыугу 1984).

Рассмотренный пример предметной области показывает, что реакция автоматической системы на текст не обязательно должна выражаться в физических действиях во

внешнем мире, это могут быть преобразования данных в памяти ЭВМ (например, запись информации в базу данных традиционной структуры). Внешне это можно было бы счесть похожим на преобразование текстов, как при традиционном лингвистическом подходе; однако здесь в качестве конечного результата будут выбираться такие формы представления информации, для которых переход к практическим приложениям принципиально ясен.

2. Сетевое представление информации

Для обработки текста и перехода от него к реакции автоматической системы необходимо предусмотреть форму представления информации о тексте и о ситуации, используемую в процессе обработки. Эта форма должна быть приспособлена для одновременного отображения языковых единиц различных уровней и реальных объектов рассматриваемой ситуации, а также должна быть открыта для расширения. Этим требованиям удовлетворяет представление информации в виде ассоциативной сети данных, где узлы могут изображать объекты любой природы, а дуги - отношения между этими объектами. Запас имен, приписываемых дугам, т.е. отображаемых отношений между объектами, ничем не ограничен, так же, как и типы объектов, включаемых в сеть.

Конкретный аппарат для представления ассоциативных сетей в ЭВМ и действий над ними, разработанный для этой цели, описан в работе Цейтин 1985. Центральным понятием предлагаемой для этой цели системы программирования является понятие модуля, который хранится в библиотеке системы и может многократно воспроизводиться в текущей сети. С программистской точки зрения модуль аналогичен пакету из процедур и данных, что позволяет естественным образом включать в информацию о тексте также и процедуры для его обработки. С позиций "искусственного интеллекта" модуль аналогичен введенному М.Минским понятию фрейма, отражающему некоторый тип объектов или

ситуаций в виде набора связанных с таким объектом или ситуацией ролевых позиций ("слотов"), взаимоотношений между ними и необходимых процедур обработки.

В предлагаемой модели модуль может быть связан с любым словом из словаря предметной области, любым понятием или конкретным объектом предметной области или с лингвистическим понятием. Если ряд модулей (например, для слов) имеет много общих характеристик (например, в отношении синтаксических средств), то вместо повторения однотипных характеристик во многих модулях всюду указывается ссылка на вспомогательный модуль, отражающий эти общие черты (соответствующий определенной лингвистической категории).

3. Представление объектов различных уровней

Структура ассоциативной сети позволяет представлять в ней одновременно объекты, отражающие разные уровни обработки текста. Рассмотрим здесь способы представления объектов на трех уровнях, которые условно назовем синтаксическим, семантическим и прагматическим.

К синтаксическому уровню относится представление слов текста и синтаксических связей между ними. Каждое слово текста представляется некоторым узлом в сети. Имеется связь типа "хозяин", изображаемая дугой от подчиненного слова к его синтаксическому хозяину. Если синтаксическая модель управления слова содержит указание на подчинение ему другого слова по определенному типу, то можно вводить также и дугу, ведущую к этому подчиненному слову и именуемую этим типом подчинения. Характерные черты структуры текста на этом уровне - это иерархическая организация текста, проявляющаяся в древовидной структуре синтаксических связей, и отсутствие явно выраженных связей между различными предложениями. Между словами можно также устанавливать связи, отражающие отношение замещения. Эти связи можно представлять особым видом дуг, направленных от замес-

тителя к антецеденту. Эти связи не следуют упомянутому ограничению, связанному с иерархической структурой, но тоже характеризуют объекты синтаксического уровня.

Объектами семантического уровня мы будем считать "референты" - предметы, действия и т.п., присутствующие в реальной или мыслимой ситуации, связанной с рассматриваемым текстом. Характерный признак такого объекта - способность участвовать в замещении в качестве нового референта замещающего слова. Известно, что таким референтом может стать не только объект, прямо называемый некоторым конкретным замещаемым словом; иногда замещаемого слова в тексте вообще нет, а референтом является объект, подразумеваемый в ситуации, вводимой каким-то другим словом. Мы будем включать в семантический уровень как прямо названные, так и подразумеваемые объекты, но в целях установления связи с синтаксисом будем отличать "прямые референты" слов от подразумеваемых при них "косвенных референтов" (косвенный референт некоторого слова может совпадать с прямым референтом другого слова).

Выделение референтов не является объективным свойством предметной области. Оно носит преимущественно лингвистический характер, определяется свойствами языка и могло бы происходить по-разному в разных языках. Более того, носителями некоторых языковых признаков, например, грамматического рода и одушевленности, являются именно референты, способные передавать этот признак через замещение, а не исходные слова. С разными словами в тексте могут связываться одни и те же референты; в частности, при замещении (кореферентном) прямые референты антецедента и заместителя совпадают.

Одной из постоянных характеристик референта является его семантический класс, один из конечного списка классов, предусмотренных для данной предметной области. Ситуации, учитываемые описанием предметной области, предписывают определенные семантические классы участ-

вующим в них объектам. Такая группировка по семантическим классам может показаться слишком грубым средством. И действительно, в конкретных правилах может потребоваться более тонкая характеристика объектов, но во многих случаях для ориентировки и использования самых основных правил необходима именно такая примитивная классификация.

Между референтами могут существовать связи, отражающие, например, отношение между ситуацией и участвующим в ней объектом или между объектом и его атрибутом (характеристикой). Эти связи могут образовывать достаточно сложный граф, выходящий за рамки отдельного предложения.

К прагматическому уровню будет относить объекты, непосредственно присутствующие в той ситуации, к которой должна относиться реакция машины (или на том уровне представления информации в ЭВМ, для которого дальнейшая обработка выполняется традиционными средствами программирования). Это могут быть ссылки на конкретные данные в памяти машины или конкретные программы обработки и т.п. Выделение прагматических объектов зависит, таким образом, от выбранной предметной области и целевой установки. Для системы, которая должна выдавать реакцию в реальной ситуации, к прагматическому уровню будут относиться те объекты, которые присутствуют именно в этой, а не какой-либо предполагаемой ситуации. Напротив, при решении арифметических задач, где часто ситуация, обозначаемая как реальная, сопоставляется с другой ситуацией, обозначаемой как нереальная, обе эти ситуации будут участвовать в обработке на прагматическом уровне, потому что по "правилам игры" в таких задачах законы движения одинаково соблюдаются как в реальных, так и в нереальных ситуациях. Для области, где основную роль играет работа с базой данных, прагматическими объектами будут, вероятно, те, которым соответствуют определенные записи базы данных или поля этих записей.

Прагматические объекты не будут иметь отдельного от семантических объектов представления в ассоциативной сети, но будут обладать дополнительными атрибутами, указывающими на конкретные структуры в памяти ЭВМ, отражающие текущую ситуацию. Очевидно, не каждый объект, представленный как референт, будет участвовать в обработке на прагматическом уровне.

4. Связи между различными уровнями

Связь между синтаксическим, семантическим и прагматическим уровнями мы будем описывать в виде процессов семантической и прагматической интерпретации. Под семантической интерпретацией будем понимать построение объектов и связей семантического уровня по известной синтаксической структуре, а под прагматической интерпретацией - установление при известной семантической структуре соответствий в реально обрабатываемой ситуации для тех из семантических объектов, для которых это необходимо. Такая форма описания не означает, что в реальной обработке текста получение полной синтаксической структуры должно предшествовать построению семантической структуры, а это последнее - переходу к реальной ситуации. При обработке текста все эти процессы могут проходить параллельно.

Особенность процесса семантической интерпретации - это переход от иерархически организованной синтаксической структуры к более сложному графу связей между референтами, и этот процесс поддерживается разнообразными языковыми средствами.

Первоначально референты (прямые и косвенные) вводятся в рассмотрение вместе со словами при помощи словаря. Извлекаемая вместе со словом словарная статья уже имеет заготовки для этих референтов, а в процессе последующей работы, когда выясняется, что разные слова текста входят в описание одной и той же ситуации, происходит отождествление некоторых референтов, первоначально введенных для разных слов независимо. На-

пример, для слов "поезд идет" косвенный референт слова "идет", обозначающий движущийся объект, будет отождествлен с прямым референтом слова "поезд" ввиду того, что в словарной статье слова "идет" уже установлена семантическая связь между движением и движущимся объектом, теперь окажется установленной связь между референтами двух разных слов.

Есть несколько видов синтаксической обусловленности устанавливаемых семантических связей.

А) Непосредственное задание семантической связи в модели управления слова. При этом предусматривается связь для прямого референта синтаксически подчиняемого слова и вместе с установлением синтаксической связи устанавливается и семантическая (семантическая связь из предыдущего примера детерминируется синтаксической связью сказуемого с подлежащим). Такая интерпретация синтаксической связи может задаваться в словаре не только по отношению к подчиненным словам, но и к управляющему слову (т.е. "модель управления" дополнится "моделью подчинения"). Такой способ характерен для слов, которые обычно не входят в модель управления своего хозяина (например, прилагательных).

Б) Установление семантической связи путем синтаксической трансформации, обычно с использованием особых правил для служебных слов. Сюда можно отнести установление связи через сильноуправляемый предлог или через преобразование из пассива в актив. В этом случае не предполагается фактическая перестройка синтаксической структуры, но семантические связи устанавливаются, как если бы она была произведена. Вот более сложный пример такого типа. "Поезд идет со скоростью V" - "Поезд идет, имея скорость V" - "Поезд идет" и "Поезд имеет скорость V" - "V есть скорость поезда". Далее по модели управления слова "скорость" его прямой референт становится атрибутом "скорость" того движения, которое совершает поезд.

В) "Обращенные" синтаксические связи - помещение слова, главного для некоторого фрагмента семантической структуры, не в вершину соответствующей части синтаксической структуры; а в одну из внутренних позиций. Таковы кванторные местоимения, слово "который", сравнительные степени прилагательных и наречий. Для установления семантических связей единственная трудность состоит в нахождении той "верхней" позиции, где должно было бы стоять главное слово (эта позиция обычно хорошо характеризуется синтаксическими признаками).

Г) Простые анафорические связи. При таких связях производится отождествление прямых референтов антецедента и заместителя (реально установление анафорической связи невозможно без ее проверки посредством такой интерпретации).

Д) Обстоятельства и другие "слабые" связи. Такие связи не определяют жесткого места в семантической структуре для референтов слов, вводимых посредством этих связей. Однако включают эти референты в некоторый "контекст", который используется для предпочтительного восстановления недостающих семантических связей в той части предложения, к которой присоединена данная синтаксическая связь. Пример: "в треугольнике медианы пересекаются в одной точке". Здесь для слова "медиана" требуется указание, что это медиана чего-то, и отсутствующий референт извлекается из обстоятельства "в треугольнике". Так как семантические классы позволяют произвести такое отождествление, для связей рассматриваемого типа место вводимого слова в семантической структуре не определено жестко, но определено приблизительно указаниям на ту часть синтаксической структуры, в интерпретации которой это слово можно использовать.

Е) Семантические отождествления, никак не обусловленные синтаксисом. Такие отождествления можно вводить (если совпадают семантические классы слов и попытки дальнейшей интерпретации не приводят к противоречию) на том основании, что обычно раз упомянутый объект участвует и в последующем развитии текста. Адресат

текста обычно производит молчаливо такое отождествление, чтобы "не умножать сущности без необходимости". Иногда, впрочем, такое отождествление диктуется потребностью получения о семантическом объекте дополнительной информации, которую можно найти только путем отождествления. Встречаются слова, сама семантика которых предполагает особую активность их косвенных референтов в поиске таких отождествлений, связанную с тем, что без соответствующего отождествления отношение, передаваемое этим словом, "повисает в воздухе" и слово становится семантически избыточным. Например, если в тексте сказано, что некоторое событие произошло "через час", то обязательно необходимо связать начальную точку отсчета этого часа с каким-то предшествующим событием. Указанная особенность относится именно к косвенным референтам, потому что связи для прямого референта могут быть установлены на основании синтаксических связей.

Прагматическая интерпретация состоит в операциях вывода или преобразования, обусловленных конкретной предметной областью, вплоть до выполнения внешних действий. Там, где это необходимо по смыслу, семантический объект должен быть отождествлен с конкретным объектом реальной ситуации. Если же не предполагается, что семантическому объекту в данный момент соответствует конкретный объект, то прагматическая интерпретация остается незавершенной и полученный системой текст может быть сохранен в этом виде на случай, если эта ситуация реализуется впоследствии.

Одним из существенных механизмов перехода от референтов к единичным реальным объектам является своеобразная "свертка" многокомпонентных сочетаний, которые целиком именуют один объект, хотя на всех уровнях до прагматического все их составные части функционируют самостоятельно (например, "выходной трансформатор усилителя низкой частоты левого звукового канала", "кнопка выключения питания", "моя левая рука" и т.п.).

Аналогичная свертка встречается и на уровне абстрактных понятий, когда объектом служит само это понятие, а не подпадающие под него предметы (например, "правильный треугольник" - это конкретное математическое понятие, а отдельное слово "правильный" ничего не означает).

Отличная от традиционной постановка задачи при разработке данной модели привела к новой трактовке и некоторых других языковых явлений, связанной с учетом прагматической интерпретации.

5. Последовательность действий при анализе текста

Модель анализа не предписывает и не может предписывать раз навсегда заданной последовательности действий при анализе. Носителями этой последовательности являются сами слова, точнее, их словарные статьи, включающие в себя и процедуры. "Централизованно" выполняются только первоначальный ввод предложения и поиск слов в словаре, соединенный с морфологическим анализом. После морфологического анализа может остаться неоднозначность на уровне выбора словарной статьи или определения изменяемых морфологических признаков слова. В первом случае в дальнейшей обработке могут независимо участвовать несколько словарных статей (но предусмотрено блокирование использования всех остальных статей, если начато использование одной из них). Во втором случае неопределенные морфологические признаки вообще не приписываются слову, но впоследствии эти признаки может приписать слову какая-то синтаксическая конструкция, и тогда будет произведена проверка на совместимость с морфологической информацией.

Основная обработка текста управляется гипотезами (или предсказаниями), присутствующими в словарной информации выбранных слов. Гипотеза включает в себя следующие атрибуты:

- базу, т.е. слово, из словарной статьи которого она возникла;

- цель, т.е. предсказываемый синтаксический или семантический объект вместе с его предсказываемыми атрибутами. Этот объект надо попытаться отождествить с некоторым реальным синтаксическим объектом или с некоторыми семантическими объектами;

- процедуру обработки ("внимания"), которая предусматривает определенный порядок поиска объекта, с которым отождествляется цель гипотезы, а также необходимые при этом дополнительные проверки;

- ранг, т.е. число, определяющее первоначальную очередность обработки этой гипотезы.

Процедуры при своей работе могут использовать механизм неудач и восстановлений. Этот механизм позволяет перед запуском какой-либо процедуры предусмотреть возможность неудачи при ее выполнении (в частности, неудачей может стать попытка отождествления объектов с несовместимыми характеристиками, а также явное "объявление" неудачи при определенных условиях). В случае возникновения такой неудачи выполнение процедуры прекращается, а состояние всей сети данных возвращается к существовавшему на момент запуска процедуры. Процедура, делающая попытку установления некоторой связи (путем отождествления цели гипотезы с другим объектом), вначале устанавливает эту связь, затем выполняет ряд проверок, и если хотя бы одна из этих проверок привела к неудаче, установленная было связь отменяется.

Проверки, выполняемые в ходе выполнения процедуры установления связи, помимо проверки отождествляемых объектов на совместимость, могут иметь следующий характер. Это может быть попытка установления некоторых дополнительных связей, подтверждающих исходную связь. Затем, возможно, семантическая и прагматическая интерпретация, которые также могут привести к неудаче. Разумеется, при прагматической интерпретации следует избегать совершения необратимых внешних действий до окончательного подтверждения анализа введенного предложения.

Требование установления дополнительных связей до окончательного принятия исходной связи отражает существующую в языке взаимообусловленность связей. Предусматривается два основных типа обращения к дополнительным связям: поиск конкретной связи, необходимой для завершения некоторой синтаксической конструкции, в частности, поиск хозяина для некоторого слова (указание на подтверждающую связь должно присутствовать в словарной статье) и поиск всех связей, заполняющих в предложении промежуток между связываемыми словами (с целью обеспечения проективности формируемой синтаксической структуры). Установление дополнительных связей производится через обращение к общему списку гипотез, где должны быть и те гипотезы, обработка которых приведет к установлению искомых связей (после обработки всех гипотез нужно выяснить, действительно ли требуемые связи установлены).

Обращение к списку гипотез здесь носит выборочный характер: обрабатываются лишь те из гипотез, которые могут привести к получению "полезной" связи. Это приводит к тому, что первоначально намеченный порядок обработки гипотез в соответствии с их рангами будет изменен, если это необходимо по существу гипотез.

Предусмотрено несколько типовых процедур "внимание" для гипотез, различающихся между собой способом и порядком поиска кандидата на отождествление с целью гипотезы. Для гипотез с синтаксическим объектом в качестве цели это:

- поиск конкретной лексемы;
- поиск на заданный синтаксический класс;
- поиск по семантическому классу прямого референта;
- поиск по позиции (влево или вправо от базы гипотезы, либо проверка только слова, соседнего с базой).

Для гипотез на семантический объект должны использоваться процедуры, учитывающие конкретные особенности используемых семантических классов.

Использование типовых процедур не исключает построения индивидуальных процедур для слов с особыми синтаксическими или семантическими свойствами.

Будем для краткости называть гипотезу синтаксической (семантической), если ее целью является синтаксический (семантический) объект.

В качестве примера взаимодействия гипотез разного характера рассмотрим анализ конструкций с предложными группами, подчиненными глаголу. Если предложная группа сильноуправляемая, то глагол предсказывает предлог (т.е. имеет синтаксическую гипотезу на него), а также имеет семантическую гипотезу на референт припредложного субстантива (предсказывает его семантический класс, а также задает его место в семантической структуре). Для слабоуправляемых групп предсказание на глагол, а также интерпретация установленной связи определяются предлогом вместе с припредложным субстантивом.

В предложении "поезд прибыл в три часа в город" глагол "прибыл" управляет двумя предлогами "в", один из них является сильноуправляемым ("прибыл в город"), другой - слабоуправляемым ("прибыл в три часа"). Слово "прибыл" имеет синтаксическую гипотезу на сильноуправляемый предлог "в". А при проверке связи, устанавливаемой по этой гипотезе, будет запущена гипотеза на припредложный субстантив из словарной статьи предлога. Кроме того, глагол задает еще и семантический класс прямого референта этого субстантива (географический пункт) и его семантические связи (конечный пункт движения).

Если при обработке гипотезы от глагола на предлог в качестве кандидата ошибочно выберется первый предлог "в" ("прибыл в три часа"), то этот вариант будет отклонен из-за несоответствия семантических классов. Другой ошибочный вариант, когда первому предлогу "в", в действительности управляющему группой "три часа", в качестве припредложного субстантива будет предложено

слово "город", не пройдет из-за нарушения проективности: промежуточная группа слов "три часа в" не сможет ни с чем связаться. Таким образом, будет установлена правильная связь "прибыл в город" и город будет отмечен как конечный пункт движения.

При проверке этой связи на проективность будут запущены все синтаксические процедуры для слов из промежутка между глаголом и предлогом, однако процесс установления связей здесь будет другим, связь между предлогом "в" и предложной группой "три часа" будет по-прежнему заключена по гипотезе, вводимой предлогом. Для гипотезы на хозяина, имеющейся у предлога, единственным кандидатом в силу проективности будет слово "прибыл". Интерпретация такой связи потребует отождествления прямого референта предложной группы с одним из референтов в пределах группы глагола: какой именно из этих референтов будет выбран, определяется семантическим классом (момент времени). Таким образом, получится, что "три часа" - это и есть момент завершения движения.

Взаимодействие полученных из словаря процедур по описанной схеме должно в случае успешного анализа завершиться построением синтаксической и семантической структур для предложения с включением элементов прагматической интерпретации. После этого возможна отработка реакции системы на введенное предложение (выполнение команды, выдача запроса на дополнительные сведения и т.п.). Завершившийся анализ отдельного предложения может оказать влияние на анализ последующего текста через сформированные семантические объекты и связанные с ними гипотезы. Семантические объекты могут быть использованы при анализе следующего предложения для семантических отождествлений. Остающиеся от предыдущего текста гипотезы могут сыграть активную роль в анализе нового предложения, если они будут включены в общий список гипотез и при работе сами выберут в новом предложении референты, отождествляемые с ранее введенными (это будет моделировать "предвзятое" вос-

приятие нового предложения под влиянием ожиданий, сформированных предыдущим текстом).

Таким образом, описываемая модель может учесть внутритекстовые связи, выходящие за пределы одного предложения.

6. Специальное использование общеязыковых средств в узкой предметной области

В узкой области, где значения, передаваемые некоторыми общеязыковыми лексическими или грамматическими средствами, часто оказываются лишенными смысла (ненужными для прагматики), эти средства могут приобретать новое, специфическое использование и значение. Вот некоторые примеры.

С точки зрения вычислительных моделей, обеспечивающих решение задач на движение, несущественны такие характеристики движения, как движущийся объект, соотношение между временем, когда происходит движение, и настоящим моментом, а также модальность, в которой сообщается о движении. Тем не менее в задачах о движении присутствует и упоминание движущегося объекта, и выражение времени и модальности. Две последние категории выражены не только потому, что язык не позволяет обходиться без них. Иногда они специально вводятся в формулировку задачи, но их цель состоит в том, чтобы просто указать, что некоторые два движения различны (то, которое уже совершено; то, которое предстоит совершить; то, которое могло бы совершиться при других условиях и т.п.). Таким образом, реальная (и сознательно используемая) функция категорий времени и модальности состоит в предотвращении отождествления референтов при семантической интерпретации.

Ссылка на движущийся объект имеет аналогичную функцию, но не только эту. Движущийся объект, игнорируемый на прагматическом уровне этой предметной области, может активно использоваться на семантическом уровне для управления попытками отождествления движений, а

также для переноса характеристик одного движения на другие при отсутствии явного указания на недопустимость этого.

Остается заметить, что разнообразные названия движущихся объектов сохраняют из своей "семантики" лишь указание на то, какие объекты отличаются друг от друга.

Другой пример такого рода связан с использованием актуального членения в строгих математических формулировках. Математическая строгость основана на игнорировании всех субъективных аспектов, поэтому актуальное членение здесь не может быть использовано по назначению, оно используется для выражения кванторной информации: при отсутствии противоречащих указаний тема связывается с внешним квантором общности, а рема - с внутренним квантором существования.

7. Учет некоторых видов межфразовых связей

Один из видов межфразовых связей связан с обязательностью реализации предикатных актантов. Предметная область может требовать для какого-либо предиката, чтобы определенное подмножество его актантов было реализовано в тексте в той или иной форме. Так, в задачах на движение требуют обязательной реализации атрибуты движения - путь, скорость, и время. Семантические объекты, соответствующие этим атрибутам, вводятся в качестве косвенных референтов слов со значением движений и являются целями гипотез, связанных с этими словами. В данной модели обязательность реализации соответствующих актантов выражается в требовании обязательного отождествления этих косвенных референтов с семантическими объектами, введенными за счет других слов. Часть таких отождествлений произойдет в результате установления синтаксических связей внутри предложения. Гипотезы, оставшиеся после этого нереализованными, моделируют явление, известное в лингвистике как "текстовое ожидание". Эти гипотезы будут участво-

вать в анализе следующих предложений и могут реализоваться там еще до установления некоторых внутренних синтаксических и семантических связей.

Обязательной реализации косвенных референтов могут требовать не только предикаты, но и некоторые другие слова. Например, "оставшийся" ("осталось", "остальной"), "обратный" и другие. В этих случаях в отличие от предикатов семантический класс предсказываемого объекта определяется не самим словом, а словами из его ближайшего синтаксического окружения.

Существует группа слов, которые не только требуют обязательной реализации косвенных референтов, но и могут с некоторой степенью достоверности предсказать их синтаксическое, а иногда и морфологическое оформление. Так, рассмотрим следующие косвенные референты слова "вышла" (в примере "машина вышла из пункта А"): движение, окончание движения, начальный пункт пути, конечный пункт пути. Окончание движения с большой степенью вероятности реализуется глаголами с приставками "при" или "до" (пришла, прибыла, достигла). А семантико-морфологическая реализация конечного пункта пути зависит от такой реализации начального пункта ("из пункта А в пункт Б" или "из Москвы в Ленинград", но не "из пункта А в Ленинград"). Таким образом, на последующие предложения могут распространяться гипотезы, предсказывающие не только семантические, но и синтаксические объекты.

Еще одним видом межфразовых связей, при котором требуется обязательное отождествление, являются анафорические связи. В отличие от рассмотренных выше видов межфразовых связей, где обязательного отождествления требуют косвенные референты, при анафорических связях обязательного отождествления требует прямой референт заместителя. В этом случае увеличивается также роль синтаксических признаков при поиске антецедента. Рассмотрим вначале случай, когда антецедент присутствует на поверхностном уровне.

Если заместитель является личным местоимением 3-го лица, то отождествляются прямые референты антецедента и местоимения, а целью анафорической гипотезы при заместителе является синтаксический объект, для которого заданы род и число, а также семантический класс его прямого референта. Этот семантический класс местоимение может получить в силу синтаксической связи с подчиняющим его предикатным словом. Идея индуцированных семантических признаков местоимения не нова в лингвистике. В описываемой модели такое индуцирование происходит естественным образом благодаря механизму отождествления узлов ассоциативной сети.

Если заместитель представлен сочетанием адъективного анафорического элемента ("этот", "тот", "такой же") с замещающим существительным, то отождествляются прямые референты антецедента и замещающего существительного соответственно. Целью гипотезы, вводимой адъективным анафорическим элементом, является синтаксический объект, у которого лексема совпадает с лексемой замещающего слова либо находится с ней в родо=видовых отношениях, или же у которого прямой референт имеет тот же семантический класс, что и прямой референт замещающего слова.

Встречаются и более сложные случаи замещения, когда антецедент не выражен на поверхностном уровне. Для таких случаев у анафорического слова должна быть предусмотрена семантическая гипотеза, целью которой является прямой референт замещающего слова и которая реализуется при отождествлении этого прямого референта с косвенным референтом некоторого слова (в данной предметной области - слова с семантикой движения), как в следующем примере: "турист прошел 16 км за 4 часа. За какое время он пройдет 20 км, если будет двигаться с той же скоростью?".

Таким образом, в данной модели заключение всех видов межфразовых связей оформляется в виде отождествления семантических объектов, но с разной степенью участия синтаксических признаков в процессе поиска.

8. "Параллельные" конструкции и сокращение

Конструкции языка, обычно трактуемые как состоящие из однородных членов, а также сходные с ними в некоторых отношениях сравнительные конструкции требуют с точки зрения семантической интерпретации несколько иного представления. В семантическом представлении для конструкций такого типа необходимо присутствие некоторого объединяющего объекта (ситуации, действия, предиката), для которого предусмотрено наличие двух подчиненных объектов ("мест предиката"), рассматриваемых как однотипные. Объединяющий объект может возникнуть из слов, выражающих операцию сравнения или какое-либо соотношение между однотипными объектами (например, встречу), или, наконец, из сочинительных союзов. Задачей интерпретации является восстановление семантических структур, занимающих предусмотренные для объединяющего объекта однотипные места.

При использовании в языке таких конструкций обычно первая из двух структур представлена в тексте полностью, а вторая - сокращенно. Можно было бы рассматривать такие тексты как результат особой сокращающей трансформации, примененной к первоначально полному представлению второй части. Но ввиду распространенности таких сокращений и в целях обеспечения анализа текста предлагается рассматривать такие конструкции не как сокращенную форму чего-то иного, а как самостоятельный, регулярный способ выражения конструкции, заполняющей второй место объединяющего объекта.

На синтетическом уровне мы будем считать, что эта сокращенная конструкция занимает некоторый связный участок предложения и состоит из одной или нескольких связных синтаксических групп, для которых не устанавливаются синтаксические связи с остальной частью предложений. Для каждой из этих синтаксических групп строится и соответствующая семантическая структура. Далее, для заполнения второго места объединяющего объекта снимается копия с заполнения его первого места. Затем каждая из се-

мантических структур второй части "вытесняет" некоторую часть из сделанной копии. Вытеснение состоит в том, что на место вытесняемого объекта ставится вытесняющий вместе со всеми связанными с ним объектами; при этом дополнительные признаки и связи вытесняемого объекта сохраняются лишь в случае, если они не противоречат новым признакам и связям. Вытесняемый объект должен быть семантически подобен вытесняющему (в простых случаях - иметь одинаковый семантический класс). А в случае, если вытесняемый объект является прямым референтом некоторого слова, вытесняющая синтаксическая группа должна быть оформлена так, как если бы она в синтаксической структуре занимала место этого слова. Встречаются случаи вытеснения элементов семантической структуры, не выраженных явно в тексте. Например, признака утвердительной модальности.

Это можно проиллюстрировать на примере следующего предложения, в котором неоднозначность интерпретации связана именно с выбором вытесняемого элемента: "ракета летела быстрее, чем думал конструктор" (пример С.Я.Фитиалова). Один из вариантов (ракета достигла определенной точки раньше, чем конструктор завершил обдумывание чего-то) получается, если в копии первого предложения вытесняется действие "летела" вместе с агентом "ракета" (но сохраняется указание на сравниваемый момент времени). При втором варианте (ракета летела быстрее в действительности, чем она летела в предположениях конструктора) в копии первого предложения сохраняются и действие, а игент, а вытесняется утвердительная модальность.

Аналогичный механизм вытеснения, вероятно, должен использоваться при семантической интерпретации некореферентного замещения.

9. Примеры взаимодействия языковых и предметных знаний при интерпретации

Границу между использованием общеязыковых знаний и знания конкретной предметной области провести порой до-

вольно трудно. Рассмотрим текст задачи, где, например, сказано: "Две машины выехали из двух городов... Каково будет расстояние между ними через 2 часа?". Как можно понять, что слова "между ними" относятся к машинам, а не к городам? Возможно, тут играет роль указание времени при вопросе о расстоянии, что было бы избыточным, если бы вопрос относился к расстоянию между неподвижными объектами. Такое рассуждение (кстати, довольно трудно формализуемое в рамках предлагаемой модели) предполагает, что при установлении анафорической связи мы во всех случаях должны избегать вариантов, не поддающихся интерпретации в данной предметной области. К языку в широком понимании такое рассуждение неприменимо, иначе мы никогда не могли бы сказать никакой нелепости. Могло ли в сделанном выборе антецедента для местоимения сыграть роль не рассуждение о постоянстве расстояния между городами, а просто то, что слово "машины" в первом предложении относится к теме, а "городов" - ни к теме, ни к реме? Или сыграли роль оба обстоятельства вместе? Более того, само соображение относительно неизменности расстояния можно (с некоторой натяжкой) истолковать и как языковое: для этого достаточно приписать слову "расстояние" два альтернативных "смысла" - расстояние между неподвижными объектами, и расстояние между объектами, из которых хотя бы один подвижен, с тем, чтобы указание на время считать допустимым только для второго варианта.

Еще пример: "поезду осталось пройти 50 км" для решения задач, содержащих подобное условие, обычно требуется использовать модель составного движения. Для этого со словами "осталось пройти" нужно связать три движения: то, что надлежало пройти; то, что уже пройдено (косвенные референты) и то, что осталось пройти (прямой референт). Движения, упоминаемые в других предложениях, могут отождествиться с одним из этих трех движений. Например, в предшествующем предложении могло быть сказано "поезд прошел..." или "поезд должен был пройти...". Единственный признак, по которому мы

можем установить, с каким из трех движений можно отождествить названное движение, является модальность (или время). Таким образом, здесь отождествлением движения управляет языковой признак. А можно ли считать чисто языковым явлением выделение трех движений по слову "осталось"?

Вот более сложный пример на ту же тему: "поезд задержался на I5 минут, поэтому, чтобы прибыть на станцию по расписанию, он увеличил скорость на...". Для использования этого текста при решении задачи приходится рассматривать следующее толкование слова "задержаться": начать некоторое движение позже, чем намечено. Тогда по словам "поезд задержался на I5 минут" мы сможем построить сеть, где будут заданы начало некоторого действия реальное, начало того же действия намеченное и разделяющий их временной интервал в I5 минут, затем по последующему тексту нужно будет отождествить это неназванное действие с движением от какой-то точки до станции (тоже в двух модальностях), а также отождествить модальность "намеченного" с указанием "по расписанию". Можно ли считать подобные приемы законным толкованием использованных языковых средств или это лишь частные уловки?

Следует заметить, что изобилие подобных случаев, делающее эти тексты интересными в качестве лингвистического материала, на самом деле носит несколько искусственный характер и связано все с теми же специфическими "правилами игры" в этой предметной области, которые требуют, чтобы автор задачи уложил сложные численные соотношения в достаточно компактную формулировку.

Еще один пример из более простой области - запросов к базе данных, содержащей сведения о кадрах некоторой организации: "указать такой отдел, в котором, чем старше сотрудник, тем больше он зарабатывает". Отображение такого текста на структуру базы данных не представляет особой сложности. Для сравнительной конструкции

здесь надо использовать модель, в которой выделяется множество объектов, для которого производится сопоставление, и два сопоставляемых параметра (возраст и оклад). С этой моделью можно легко связать программу, проверяющую указанное соотношение между возрастом и окладом. А также, если это желательно, перевод запроса на язык исчисления предикатов ("для любых двух сотрудников этого отдела, если возраст первого больше возраста второго...") Этот пример интересен тем, что показывает, что язык исчисления предикатов нецелесообразно использовать в качестве промежуточного представления между естественным языком и базой данных, потому что непосредственно по исходной формулировке (точнее, по модели для сравнительной конструкции) можно построить более эффективную машинную программу, чем по формуле исчисления предикатов. Действительно, в машинной программе можно использовать сортировку множества по одному из признаков, после чего проверить монотонность значений второго признака в отсортированной последовательности. Извлечь эту возможность из представления на языке исчисления предикатов довольно сложно.

Литература

Тыугу Э.Х. Концептуальное программирование. - М.: Наука, 1984.

Цейтин Г.С. О соотношении естественного языка и формальной модели // Вопросы кибернетики. - М., 1982.

Цейтин Г.С. Программирование на ассоциативных сетях // ЭВМ в проектировании и производстве. - Вып. 2. - Л., 1985.

М.Г.Мальковский,
И.А.Волкова,
Т.Ю.Грацианова

ВЗАИМОДЕЙСТВИЕ С ЭВМ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ: ПРОБЛЕМА РЕЧЕВЫХ ОШИБОК

Речевая деятельность человека неизбежно сопровождается теми или иными нарушениями языковых норм. Такие нарушения - вне зависимости от того, обусловлены они неполнотой знаний человека о языке или же случайными сенсомоторными "сбоями" (описки, опечатки, оговорки) - мы будем называть речевыми ошибками.

В идеале обработка речевой ошибки предполагает соотнесение ошибочной речевой единицы с полным описанием языка и с контекстом рассматриваемого коммуникативного акта. Лингвист (или другой специалист), занимающийся исследованием каких-либо теоретических аспектов проблемы речевых ошибок, например, их классификацией, и располагающий источниками, в которых содержится исчерпывающее описание единиц и правил того или иного естественного языка (словари, своды правил), находится в ситуации, достаточно близкой к такому идеалу. К тому же в теоретических исследованиях исходными объектами обычно являются тексты, в которых ошибки уже зафиксированы - факт ошибки известен.

В случае же повседневной речевой практики - непосредственного (диалог) или опосредованного (чтение текста) речевого взаимодействия рядовых носителей языка - ситуация иная. Лингвистические знания рядового носителя языка неполны, воспользоваться справочной литературой он может далеко не всегда, а сам факт ошибки никаким явным образом в анализируемом тексте не указан. Обнаружить речевую ошибку в этой ситуации непросто. Действительно, для получателя сообщения (слушающего) внешним признаком речевой ошибки служит появление в тексте какой-либо незнакомой ему речевой единицы. Однако такая "подозреваемая" речевая единица может

оказаться не только следствием речевой ошибки автора сообщения, но и правильной конструкцией или формой (например, просторечным вариантом или термином), не знакомой слушающему. С другой стороны, абсолютно правильная на первый взгляд единица может быть результатом речевой ошибки, обнаружить которую удается лишь на "высших" этапах анализа, например, при соотнесении содержания очередного сообщения с ранее полученной информацией.

Если одним из участников речевого взаимодействия является реализуемая на ЭВМ искусственная система (нас будем интересовать ситуация, когда она выступает в роли получателя сообщений - слушающего), положение становится еще более сложным. И лингвистические знания, и интеллектуальные способности (в том числе - в плане оперирования языком) такого "собеседника" пока весьма скромны. Однако, насмотря на возникающие трудности, игнорировать проблему речевых ошибок при создании подобных систем, как нам кажется, нельзя (Мальковский 1983). Как бы ни разнились характер использования и назначение систем взаимодействия с ЭВМ на естественном языке (системы машинного перевода, автоматического индексирования или реферирования, работающие в пакетном режиме; системы обеспечения диалога с машиной), оснащение таких систем средствами обнаружения и исправления речевых ошибок повышает устойчивость и эффективность их функционирования, облегчает (в случае диалоговых систем) процесс общения человека с ЭВМ.

В статье подход к проблеме речевых ошибок, рассматриваемой в контексте задачи создания систем взаимодействия с ЭВМ на естественном языке, иллюстрируется на примере диалоговой системы искусственного интеллекта TULIPS-2 (Malkovsky 1982 ; Мальковский 1985), средством общения с которой служит русский язык.

Система TULIPS-2 строится как система широкой ориентации. Это означает, что пользователи могут работать с системой в различных проблемных областях, для каждой из которых характерен свой набор объектов, связывающих их отношений, заданий по обработке этих объектов [I]. Обработку заданий пользователя (поиск ответа на вопрос, решение предлагаемой задачи) реализуют проблемно-ориентированные решатели задач, а диалог с пользователем обеспечивает языковой процессор системы.

Основополагающим принципом при разработке языкового процессора системы TULIPS-2 был принцип организации естественного, необременительного для человека диалога с системой. В соответствии с этим принципом пользователю системы должна быть предоставлена возможность вести диалог с ней (в любой из проблемных областей) в максимально удобной и привычной форме. Его внимание должно быть сосредоточено на содержательной стороне предлагаемых системе заданий, а не на языковых способах выражения этого содержания. Система же должна правильно понимать пользователя и в тех случаях, когда он недостаточно точен, употребляет в речи незнакомые системе единицы и конструкции языка. Возникающие при этом конфликтные ситуации система должна разрешать по возможности самостоятельно. Если в наиболее сложных случаях (число их должно быть минимизировано) языковому процессору для завершения анализа очередного сообщения все же нужна помощь пользователя,

---

[I] Например, в проблемной области "Синтез программ на Лиспе" (Мальковский 1985; Большакова, Мальковский 1986) объектами являются атомы, списки языка программирования Лисп, связанные отношениями вхождения в список и др., а задание на синтез предполагает генерацию программы на Лиспе, выполняющей те или иные преобразования исходных лисп-выражений.

инициируемый системой уточняющий диалог должен быть кратким, а для ответа на вопросы системы от пользователя не должны требоваться ни специальные лингвистические знания, ни знания об устройстве системы и используемых способах представления данных.

Очевидно, что обеспечение необременительного диалога требует весьма серьезного отношения к вопросам формирования используемого языковым процессором описания языка общения, т.е. языковых знаний системы (ЯЗС) [2]. Одной из необходимых предпосылок необременительного диалога является предварительное формирование и ввод в систему (в специальном режиме администратора) возможно более полного описания подмножества языка (языковых единиц и правил), используемого для общения с нею в рассматриваемых проблемных областях. Полнота соответствующего компонента ЯЗС - базовых языковых знаний - гарантирует эффективную работу языкового процессора в тех ситуациях, когда речевое поведение пользователя не вступает в противоречие с ожиданиями разработчиков (администраторов) системы.

Однако как бы тщательно ни формировались базовые ЯЗС, в анализируемых сообщениях всегда могут встретиться и незнакомые системе языковые единицы или конструкции, не описываемые известными ей правилами языка. В подобных случаях и возникают упоминавшиеся конфликты между ЯЗС и реально анализируемым текстом, конфликты, которые система должна разрешать, не отвлекая пользователя по пустякам. Отметим, что и пользователь может захотеть расширить используемое подмножество языка общения с системой, например, ввести новые термины или

---

[2] ЯЗС включают словари (основ, неизменяемых слов, флексий), правила грамматики, шаблоны синтаксических конструкций, описания моделей управления слов-предикатов, описания семантических значений языковых единиц и правила семантической интерпретации.

условные обозначения, позволяющие сократить длину набираемых на клавиатуре терминала текстов. Для обеспечения возможности пополнения ЯЗС непосредственно в ходе диалога выделен еще один компонент языковых знаний – открытые ЯЗС. Кроме того, система оснащена специальными модулями языковой адаптации и обучения языку, реализующими пополнение (а в необходимых случаях и коррекцию) открытых ЯЗС.

Очевидно, что новые сведения о языке, сообщенные системе пользователем или выявленные ею самостоятельно при исследовании структуры незнакомой речевой единицы и устранении конфликта, возникшего на этапе анализа, должны быть соотнесены с наличной совокупностью ЯЗС. Так, следует, в частности, проверить, не совпадает ли по внешнему виду новый языковой объект с некоторым уже известным системе объектом. Если, например, введенное пользователем условное обозначение совпало с известной лексической или морфологической единицей (сокращение "НА" для терминологического словосочетания "нормальный алгоритм" – с предлогом "на"), приходится формировать некоторый объединенный вариант словарной статьи; в том случае, когда совпадения нет, достаточно сформировать словарную статью нового объекта. Существенно более сложные действия требуется выполнить для проверки правильности (грамматичности) незнакомой речевой единицы, обнаруженной системой при анализе сообщения.

Для устранения конфликта, т.е. "превращения" незнакомого объекта в знакомый, возможны два пути.

Первый из них – адаптация ЯЗС – предполагает формирование описания незнакомого объекта (например, словарной статьи, связываемой с основой незнакомого системе слова) и временное включение этого описания в ЯЗС [3]. После

---

[3] Языковая адаптация предполагает сиюминутное изменение ЯЗС. Запоминание новых элементов ЯЗС (в том числе описаний, сформированных модулем языковой адаптации) мы называем обучением языку.

такого расширения языковых знаний конфликта уже нет, анализ сообщения может быть продолжен. Разумеется, адаптация ЯЗС допустима только в том случае, если незнакомая речевая единица грамматична, - т.е. построена по известным системе правилам языка из допустимых (с точки зрения ЯЗС) языковых единиц - как известных системе, так, возможно, и неизвестных.

Если же незнакомый объект построен с нарушением каких-то правил языка, зафиксированных в ЯЗС, реализуется второй из возможных путей - изменение незнакомой речевой единицы, т.е. исправление речевой ошибки пользователя. Управление в этом случае передается не модулю языковой адаптации, а модулю исправления речевых ошибок, который ищет преобразование, обратное тому, в результате которого возникла ошибочная форма, и выполняет это обратное преобразование.

Тем самым проблема речевых ошибок решается в системе TULIPS-2 в контексте более общей проблемы разрешения конфликтов при анализе сообщений пользователя. Всякий раз, когда языковой процессор встречает в сообщении незнакомый объект - т.е, возникает конфликтная ситуация, разрешение ее предполагает ответ на вопрос, является ли незнакомый объект правильным новым (в этом случае необходима соответствующая адаптация ЯЗС) или же искаженной формой знакомого объекта - результатом речевой ошибки. В принципе возможно, что и после исправления явной грамматической ошибки объект остается незнакомым системе (ошибка типа "ы после шипящей" в незнакомом слове "мышы"). Очевидно, что в этом случае после исправления ошибки должен быть вызван и модуль языковой адаптации - т.е. изменяется как анализируемый текст, так и языковые знания системы.

Классификация речевых ошибок

Поскольку проблема речевых ошибок рассматривается при создании системы TULIPS-2 в конструктивном плане - в аспекте обеспечения необременительного для пользова-

теля диалога с системой, - предлагаемая классификация речевых ошибок строилась таким образом, чтобы она была полезной именно с точки зрения действий, выполняемых при обнаружении и исправлении ошибок. Эта классификация учитывает, что для обнаружения ошибок могут использоваться разные способы, а на их появление система может реагировать по-разному.

Первый критерий классификации речевых ошибок, в соответствии с которым ошибки подразделяются на мотивированные и случайные, связан с понятием индивидуальной языковой модели (Мальковский 1985, 34). Индивидуальная языковая модель (ИЯМ) - это то подмножество языковых единиц и правил, которое усвоил и использует в своей речевой практике конкретный носитель естественного языка (как знаковой системы социального уровня) в ходе его усвоения приводит к тому, что в речи различных носителей языка начинают появляться некоторые индивидуальные особенности, либо вступающие в противоречие с языковыми нормами, либо нет. В первом случае мы и имеем дело с мотивированными речевыми ошибками - точнее, с ошибками, мотивированными особенностями ИЯМ конкретного носителя языка (конкретного пользователя системы) [4]. К ошибкам такого рода относятся, например, ошибки в словоизменении ("контейнера"), орфографические ошибки в основах ("слогаемое"), некоторые пунктуационные ошибки, смешение слов-паронимов ("смертный" - "смертельный"), нарушение лексической сочетаемости ("делать горе").

---

[4] Информация об особенностях каждого пользователя в употреблении языка хранится в специальных зонах долговременной памяти системы TULIPS-2 (зонах пользователей) и считывается в начале очередного сеанса работы с системой в ее оперативную память. Благодаря этому языковой процессор имеет возможность настроиться на ИЯМ пользователя, в частности, "вспомнить", какие условные обозначения, термины он ввел во время предыдущих сеансов, какие речевые ошибки он допускал.

Ошибки, обусловленные внешними по отношению к ИЯМ факторами: несвоевременное переключение регистра клавиатуры терминала, нажатие соседней клавиши, сбой на линии связи с ЭВМ, - мы будем называть случайными. Как правило, мотивированные речевые ошибки регулярно повторяются в речи пользователя, а случайные ошибки могут повторяться или нет. Отметим, что иногда отличить случайную ошибку от мотивированной сложно, так, употребление слова "представлять" вместо "предоставлять" в контексте "представлено право" может быть или результатом случайной ошибки (пропуск буквы) или результатом мотивированной ошибки (смешения слов-паронимов).

Мотивированные речевые ошибки могут различаться степенью серьезности (грамматичности). Помимо серьезных, абсолютно недопустимых грамматических ошибок - типа орфографических ошибок в основах или смешения слов - рассматриваются и ошибки, в результате которых появляются "полуграмматичные" формы ("контейнера", "сидевши"), которые имеют в словарях пометы: просторечное, устарелое, разговорное и т.п.

Следующий критерий классификации ошибок (мотивированных и случайных) связан с языковыми уровнями, нормы (правила) которых оказываются нарушенными в результате речевых ошибок. В соответствии с этим критерием учитываемые речевые ошибки можно классифицировать следующим образом:

I) орфографические ошибки: пропуск одной буквы, замена одной буквы, перестановка двух рядом стоящих букв, одна лишняя буква (отдельно рассматривается случай удвоения буквы), замена буквы русского алфавита буквой латиницы;

2) морфологические (словоизменительный уровень) ошибки: ошибки в окончаниях (флексиях) при склонении и спряжении слов (рассматриваются различные подклассы таких ошибок), употребление отсутствующих форм слов, несоблюдение чередований в основе, употребление незнакомых системе вариантов слов, испытывающих колебания в роде, одушевленности;

3) синтаксические ошибки: ошибки в моделях управления слов-предикатов, пунктуационные ошибки, нарушение нормативного порядка слов (в том числе - нарушение проективности), вставка пробела внутрь слова, пропуск пробела (отдельно рассматриваются случаи слитного и раздельного написания частиц "не" и "ни");

4) лексико-семантические ошибки: употребление слов в ненормативном значении, нарушение лексической сочетаемости.

Обнаружение и исправление речевых ошибок

Как уже отмечалось, признаком речевой ошибки (или ее возможности) служит возникновение конфликта, заключающегося в том, что из-за появления незнакомой речевой единицы процесс анализа сообщения заходит в тупик. Иногда момент обнаружения конфликта совпадает по времени с моментом проявления истинной его причины. Так, появление формы незнакомого системе слова сразу же сигнализирует о возникновении конфликтной ситуации, которая устраняется либо модулем языковой адаптации (если в сообщении действительно встретилось новое слово), либо модулем исправления речевых ошибок. Сразу же, как правило, обнаруживаются и конфликты при анализе знакомых слов, соответствующие учитываемым ошибкам в словоизменении, например, ошибкам в окончаниях при склонении и спряжении ("расширя" вместо "расширив") или несоблюдение чередований ("стригет" вместо "стрижет"); иногда удается сразу же исправить и грамматическую ошибку в незнакомом системе слове (ошибки типа "хорошый", "думаишь"). Важную роль при распознавании конфликтов играют предсказания, позволяющие учесть контекстные условия [5].

[5] Предсказание - описание ожидаемого результата работы некоторой процедуры анализа, которое позволяет ей провести "целенаправленный" анализ входных данных с учетом информации об уже проанализированных (объемлющими процедурами) фрагментах сообщения. Широкое использование аппарата предсказаний на всех уровнях анализа сообщений является одной из наиболее характерных черт языкового процессора системы

Однако в ряде случаев конфликт удается обнаружить только после завершения анализа очередной порции текста: отдельной именно группы, предложения или сообщения в целом. Так, если в результате орфографической ошибки незнакомое системе слово случайно совпало с формой знакомого слова, конфликт не проявится при морфологическом анализе ошибочной словоформы (она будет воспринята как форма знакомого слова), но может возникнуть позже, например, из-за того, что семантический класс слова окажется недопустимым с точки зрения контекста. В подобных ситуациях необходим возврат на один из предшествующих этапов анализа и поиск "точки", соответствующей истинной причине конфликта. В нашем примере необходимо вернуться на этап морфологического анализа конфликтной словоформы и рассмотреть другие пути анализа. Отметим, что иногда распознать конфликт может лишь пользователь, не соглашающийся с той или иной интерпретацией сообщения, которую нашла система. Его указания могут оказаться полезными и при исправлении ошибок.

При разрешении конфликтов (в частности, при обнаружении и исправлении ошибок) системе может помочь информация о вероятности различных причин конфликтов, наиболее характерных типах речевых ошибок. Она управляет как выбором одного из альтернативных путей разрешения конфликтной ситуации (языковая адаптация или исправление речевой ошибки), так и определением наиболее вероятного типа ошибки. Отметим, что при смене пользователей и/или проблемных областей эта управляющая информация может меняться [6]. Так, в проблемной области, в которой появление незнакомых системе слов маловероятно, при обнаружении такого незнакомого слова следует прежде всего проверить, не является ли оно искаженным ошибкой знакомым словом.

---

[6] Ее изменение реализуется в процессе настройки системы на сеанс - при активации данных, представленных в зонах очередного пользователя и очередной проблемной области.

Как правило, принять окончательное решение о причинах конфликта можно лишь после нахождения способа его разрешения. Поскольку в общем случае может существовать несколько причин конфликта, исправление ошибки обычно проходит в два этапа. На первом этапе принимается некоторое предварительное решение, ограничивающее область поиска способов исправления ошибки; на втором - проводится более тщательный поиск в этой суженной области. Проиллюстрируем эту схему на примере алгоритма исправления орфографических ошибок в основах слов.

На первом этапе работы алгоритма в качестве фильтра, ограничивающего область поиска, используется специально разработанная мера, отражающая буквенный состав слова. В соответствии с этой мерой определено расстояние между основами слов. Для сужения области поиска способа исправления ошибки отбираются основы знакомых слов, близкие основе исследуемого слова в смысле рассматриваемой меры, т.е. формируется множество основ-кандидатов. Затем к исследуемой основе применяются преобразования, обратные учитываемым орфографическим ошибкам до тех пор, пока полученная в результате одного из этих преобразований основа не совпадет с некоторым элементом множества основ-кандидатов. Если совпадение обнаружено (и грамматические признаки обеих основ совпадают), исследуемая основа отождествляется с соответствующей основой знакомого слова.

Очевидно, что несмотря на использование обширных базовых ЯЗС, тщательный анализ конфликтных ситуаций, их причин и способов устранения этих причин (способов разрешения конфликтов), модуль исправления речевых ошибок может находить неверные решения. Например, слово "странница" может быть автоматически преобразовано в знакомое системе слово "страница" (близкое исходному в смысле введенной меры и получающееся из него в результате преобразования "ликвидация удвоения буквы"). Возможно, что на последующих этапах анализа сообщения

никаких новых конфликтов не возникнет, и ошибка модуля исправления речевых ошибок долго не будет проявляться. В каких-то случаях подобные недоразумения не страшны, в других - неправильные действия модуля исправления речевых ошибок могут оказаться абсолютно недопустимыми.

Для учета этого обстоятельства, других особенностей проблемных областей, а также особенностей в употреблении языка отдельными пользователями или группами пользователей, в системе TULIPS-2 предусмотрены различные режимы работы модуля исправления речевых ошибок. Режим устанавливается в начале очередного сеанса по указанному пользователем значению параметра "степень самостоятельности" (этот параметр пользователь может менять и в ходе диалога) и по значениям других параметров, которые устанавливаются администратором системы для очередной проблемной области и класса пользователей. В зависимости от режима может быть разрешено или заблокировано автоматическое исправление любых речевых ошибок или отдельных классов речевых ошибок; можно предусмотреть и обязательное согласование всех попыток исправления ошибок с пользователем или обращение к нему за помощью при обнаружении любой ошибки. Другие параметры (пользователь может менять значение параметра "степень подробности") определяют, будет ли система информировать пользователя о факте обнаружения автоматически исправляемой ошибки и способах ее исправления и - если будет - сколь подробны будут ее комментарии.

Обнаружение и исправление речевой ошибки не исчерпывает всех действий по ее обработке. Иногда может оказаться целесообразным запомнить сам факт появления той или иной конкретной речевой ошибки. Такая статистическая информация, будучи записана в зону очередного пользователя, окажется полезной при анализе конфликтных ситуаций в последующих сеансах. Так, если для одного из пользователей характерна ошибка типа "перестановка рядом стоящих букв", есть все основания считать, что

именно из-за этой ошибки в набираемых им на клавиатуре терминала сообщениях часто появляются незнакомые системе слова. Запоминает система и факты появления грамматических ошибок.

В случае же мотивированных ошибок, ведущих к появлению "полуграмматичных" форм [7] система может согласиться с особенностями ИЯМ очередного пользователя, адаптировав соответствующим образом свои языковые знания. При этом формируется новое описание языковой единицы или новое правило, которое по окончании сеанса записывается в зону очередного пользователя (ожидается, что соответствующая форма может появиться повторно прежде всего в его речи).

Метазнания

Основная часть информации, используемой системой TULIPS-2 при обработке конфликтных ситуаций (а эта информация, как мы видели, достаточно обширна и разнообразна), представлена в метазнаниях системы о языке, включаемых в базовые ЯЗС наряду с описаниями конкретных языковых единиц и правил. Метазнания (названы они так потому, что используются при соотнесении новых речевых единиц с наличными языковыми знаниями и пополнении открытых ЯЗС, т.е. являются знаниями о языковых знаниях) представлены в системе как в декларативной, так и в процедурной форме. Обращение к метазнаниям оформлено как вызов специальных процедур метауровня, осуществляемый в момент обнаружения конфликта.

---

[7] Основным источником при описании этого класса ошибок послужила работа Граудина, Йцкович, Катлинская 1976. Среди других источников, использованных при разработке модуля исправления речевых ошибок, следует прежде всего упомянуть работы Красиков 1980, Цейтлин 1982, Трудности русского языка 1981.

Все учитываемые конфликты объединены в семь групп: появление незнакомого слова; противоречия при анализе знакомого слова; противоречия при анализе незнакомого слова; знакомое слово в непредсказанной форме; незнакомое слово в непредсказанной форме; неизвестна роль фрагмента; нет обязательного фрагмента - в зависимости от того, какие языковые уровни они затгаривают и какие процедуры анализа обнаруживают эти конфликты. Отметим, что один и тот же конфликт может входить в несколько групп, хотя в разных ситуациях способы его обнаружения могут различаться.

Так, например, конфликт "несоблюдение чередования" может возникнуть при анализе как знакомого, так и незнакомого слова. В первом случае конфликт обнаруживается по информации, содержащейся в словарной статье знакомой системе основы (здесь указывается, какие именно чередования происходят в основе при различных флексиях) и по флексии. Во втором - только по структуре (графемному и морфемному составу) основы и по флексии. Разумеется, при этом предложение системы исправить ошибку менее обосновано и может оказаться неверным (попытка исправить ошибку "несоблюдение чередования" и воспринять незнакомое слово "скула" как форму глагола "сковать" с характерным чередованием "ова" -"у").

С каждым конфликтом связано несколько возможных причин его возникновения, среди которых есть и причины, соответствующие мотивированным и случайным ошибкам. Причины конфликтов упорядочены по частоте ожидаемой встречаемости в текстах сообщений, причем в начале каждого сеанса происходит переупорядочение причин, отображающее особенности установленного режима работы модуля исправления речевых ошибок. При таком упорядочении учитываются значения параметров, управляющих режимом, и статистическая информация о причинах конфликтов, собранная системой в предшествующих сеансах работы с данным пользователем.

Каждой причине конфликта сопоставлены способ исправления соответствующей ошибки - т.е. процедура, реализующая необходимые действия, и метасписок, содержащий информацию, которая управляет поведением системы при исправлении ошибки. Изменение этой управляющей информации (как и переупорядочение причин конфликта) позволяет учесть особенности установленного в данном сеансе режима.

При исправлении речевой ошибки процедуры метауровня рассматривают все возможные причины конфликта (начиная с более вероятных в данном сеансе) и пытаются применить соответствующий способ устранения причины конфликта, т.е. исправления ошибки. Если способ оказывается неприменим (попытка исправить в незнакомом слове, в котором двойных букв нет, ошибку типа "удвоение буквы") или же его применение не приводит к успеху (в незнакомом слове обнаружены две одинаковые буквы, стоящие подряд, но и после удаления одной из них слово остается незнакомым), рассматриваются другие причины и способы. Если ни один из возможных вариантов исправления ошибки не привел к успеху, система (в зависимости от установленного режима) или обращается за помощью к пользователю, или отказывается от обработки очередного сообщения. Отметим, что и после успешной обработки конфликта (установления его причины и применения соответствующего способа) информация о конфликтной ситуации не уничтожается. Если на более поздних этапах анализа выяснится, что причина конфликта была установлена неверно, произойдет возврат в эту точку и будут рассмотрены альтернативные варианты и другие способы разрешения конфликта.

Примеры использования метазнаний

Рассмотрим конкретные примеры использования метазнаний при исправлении грамматических ошибок и при обработке "полуграмматичных" вариантов.

Пример I. Пусть на вход языкового процессора системы поступает фраза "Петя стригет глокую куздру".

В процессе морфологического анализа [8] словоформы "стригет" в ней будут выделены знакомая системе основа "стриг-" и флексия "-ет". Морфо-синтаксическая зона словарной статьи основы "стриг-" (для глагола "стричь") имеет следующий вид:

((8 28 (Ь) 27 (ЕТ ЕТЕ ЕШЬ ЕМ)) 9I20000).

Трактуется это описание так: "стриг-" - основа невозвратного глагола несовершенного вида (компонент 9I20000, четыре "0" означают, что значения грамматических признаков: время-наклонение, лицо, род и число - будут зависеть от конкретной флексии); в основе происходит чередование № 28 - "г" - "ч" при флексии "ь" (компонент 28 (Ь)); в основе происходит чередование № 27 "г" - "ж" при флексиях "-ет", "-ете", "-ешь" и "-ем" (компонент 27 (ЕТ ЕТЕ ЕШЬ ЕМ)); для глагола "стричь" характерен 8-й тип спряжения (компонент 8). В описании флексии "-ет" указано, что эта флексия встречается у глаголов 8-го типа спряжения несовершенного вида и означивает грамматические признаки следующим образом: время-наклонение - настоящее, лицо - третье, число - единственное.

Процедура морфологического уровня STFL , проверяющая сочетаемость основы с флексией, обнаруживает конфликт "несоблюдение правил чередования" (принадлежащий группе "противоречия при анализе знакомого слова"). Управление будет передано процедуре метауровня, которая рассмотрит (в указанном порядке) возможные причины этого конфликта: мотивированная грамматическая ошибка; появление "полуграмматичной" вариантной формы (ср. "достичь" - "достигнуть"); совпадение словоформы с формой

---

[8] Морфологический компонент языкового процессора системы TULIPS-2 описан в работах Мальковский I985, Мальковский, Волкова, Грацианова I983.

незнакомого слова (незнакомое существительное "трепет", которое может быть воспринято и как ошибочная форма известного глагола "трепать" - "треп-ет" вместо "треп-л-ет"); случайная орфографическая ошибка в основе.

Для первой (наиболее вероятной) причины будет рассмотрен метасписок, характеризующий соответствующий способ разрешения конфликта - исправление грамматической ошибки:

(ж ж НЧ 3 (:ШП2 ПРИ ФЛЕКСИИ . FL В ОСНОВЕ .ST ПРОИСХОДИТ ЧЕРЕДОВАНИЕ .ЧЕР) 5 ИО ( ) ( ) ж)

Его элементы трактуются следующим образом: ж - автоматическое исправление ошибки разрешено, ж - печать информации о конфликте разрешена, НЧ - ошибка типа "несоблюдение правил чередования"; 3 - степень серьезности ошибки; в заготовке текста, распечатываемого на экране терминала, фигурируют: шаблон печати № I2 (ОШИБКА В СЛОВОИЗМЕНЕНИИ!), явно заданные фрагменты текста и переменные FL, ST и ЧЕР (со значениями: флексия, основа, пара чередующихся цепочек букв); число 5 означает, что в случае, когда на исправление ошибки требуется согласие пользователя, будет распечатан тот же текст (затем будет задан вопрос: "ВЫ СОГЛАСНЫ? - (ДА/НЕТ)); символ ИО показывает, что будет исправлена ошибка; последние три элемента означают, что никакая информация о факте ошибки запоминаться не будет, а анализ сообщения после исправления ошибки должен быть продолжен.

В соответствии с этим описанием система исправит ошибку, проинформировав пользователя о своих действиях, и продолжит анализ фразы. Отметим, что первый, второй, восьмой и девятый элементы метасписка отражают особенности выбранного режима, в других сеансах эти элементы могут быть иными. Если, например, автоматическое исправление любых ошибок запрещено, первым элементом метасписка станет ( ).

Пример 2. Процессору предлагается фраза: "Петя любовался рельсой.", а в словаре указано, что "рельс" является неодушевленным существительным мужского рода, при-

надлежащим I-му парадигматическому классу (в формах единственного и множественного числа имеет флексии: "-ø", "-а", "-у", "-ø", "-ом", "-е", "-ы"о "-ов", "-ам", "-ы", "-ами", "-ах").

При анализе словоформы "рельсой" будет обнаружен (опять же процедурой STFL конфликт "флексия чужого парадигматического класса", возможные причины которого упорядочены следующим образом: колебания в роде существительных; колебания в формах именительного падежа множественного числа ("профессоры" вместо нормативного "профессора", но "лекторы", а не "лектора"); появление незнакомого слова ("логикой" - как форма слова "логика" имеет ту же основу, что и знакомое слово "логик"); мотивированная грамматическая ошибка; случайная ошибка в основе.

Для устранения первой причины конфликта (выбранной, как наиболее вероятной) следует создать копию словарной статьи слова "рельс", отличающуюся от известного варианта значением грамматического признака "род". Словарная статья, включающая оба варианта значения слова "рельс", будет по окончании сеанса записана в зону очередного пользователя.

Одна из возможных версий соответствующего метасписка может быть такова:

(( ) ж 2 (:ШП9 ЗНАКОМОЕ СУЩЕСТВИТЕЛЬНОЕ .PI РОДА УПОТРЕБЛЕНО В ФОРМЕ .P2 РОДА) 5 МЕ П ( ) ж)

Здесь: ( ) - автоматическая адаптация при установленном режиме не разрешена (требуется согласие пользователя); ж - выдача сообщения о конфликте, разумеется, разрешена; КР - "колебание в роде существительного"; 2 - степень серьезности преобразования ЯЗС; МЕ - признак формирования нового варианта словарной статьи; П - признак того, что этот новый вариант будет запоминаться в зоне пользователя. Поскольку адаптация ЯЗС производится только с согласия пользователя, на экране терминала будет напечатана и информация о выбранной трактовке конфликта (в том числе текст шаблона № 9 ВАРИАНТНАЯ ФОРМА), и вопрос (ВЫ СОГЛАСНЫ? - (ДА/НЕТ)).

Протокол диалога с системой TULIPS-2

Приведем в заключение протокол диалога с системой TULIPS-2 (в специальной проблемной области "Обучение языку"), иллюстрирующий работу модулей исправления речевых ошибок и языковой адаптации. Система TULIPS-2 реализована на языке программирования Плэнер для ЭВМ БЭСМ-6 [10]; пользователь работает за терминалом типа VT-340; интерактивный режим обеспечивает диалоговая система общего назначения ПУЛЬТ. Вызов системы TULIPS-2 осуществляется директивой Т2.

П: Т2

СИСТЕМА "TULIPS-2" ГОТОВА К РАБОТЕ
15.Ø5.86.
КТО?

П: БОГАЧЕВ

ПРОБЛЕМНАЯ ОБЛАСТЬ?

П: 3

ВЫ РАБОТАЕТЕ В ПРОБЛЕМНОЙ ОБЛАСТИ № 3 "ОБУЧЕНИЕ ЯЗЫКУ".
ЛЮБОЕ ВХОДНОЕ СООБЩЕНИЕ РАССМАТРИВАЕТСЯ МНОЮ
ЛИБО КАК КОНТРОЛЬНАЯ ИЛИ ОБУЧАЮЩАЯ ФРАЗА
(НА ЭКРАНЕ РАСПЕЧАТЫВАЕТСЯ НАЙДЕННАЯ ИНТЕРПРЕТАЦИЯ),
ЛИБО КАК ОЦЕНКА ИЛИ УКАЗАНИЕ.
СТЕПЕНЬ САМОСТОЯТЕЛЬНОСТИ СТАНДАРТНАЯ?

П: ДА

СТЕПЕНЬ ПОДРОБНОСТИ СТАНДАРТНАЯ?

П: НЕТ

ЗАДАЙТЕ, ПОЖАЛУЙСТА, СТЕПЕНЬ ПОДРОБНОСТИ: ЧИСЛО
ОТ Ø - МИН., ДО 5 - МАКС.; СТАНДАРТНОЕ ЗНАЧЕНИЕ = 2.

П: I

МОГУ ЛИ Я РАССЧИТЫВАТЬ НА ВАШУ ПОМОЩЬ НА ЭТАПЕ АНАЛИЗА?

П: ДА

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

П: ГЛОКАЯ КУЗДРА НАХОДИТСЯ В ЛЕСУ.

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "ГЛОКАЯ",
ОНО ВОСПРИНЯТО КАК ПРИЛАГАТЕЛЬНОЕ В ФОРМЕ

```
      ИМЕНИТЕЛЬНОГО ПАДЕЖА ЕДИНСТВЕННОГО ЧИСЛА.
      ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "КУЗДРА",
      ОНО ВОСПРИНЯТО КАК СУЩЕСТВИТЕЛЬНОЕ В ФОРМЕ
      ИМЕНИТЕЛЬНОГО ПАДЕЖА ЕДИНСТВЕННОГО ЧИСЛА
      (РОД: ЖЕНСКИЙ)
      ПОНЯТНО  (ВРЕМЯ АНАЛИЗА = 3.50000 СЕК.)
      =НАХОДИТЬСЯ=  (НАХОДИТЬСЯ)
        =СУБЪЕКТ= :  КУЗДРА — АТР → ГЛОКАЯ
        =МЕСТО=   :  ЛЕС (В)

-------------------------------------------------------------

        =ВРЕМЯ=   :  =СЕЙЧАС=/=ВСЕГДА=/=ИНОГДА=
        =ВИД=     :  =ДЕЙСТВИЕ ВЫПОЛНЯЕТСЯ=
П:  ВЕРНО
      НОВАЯ ИНТЕРПРЕТАЦИЯ ЗАПОМИНАЕТСЯ В ДАННОМ СЕАНСЕ.
      ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ
П:  ВЧЕРА КУЗДРА БЫЛА ВЛЕСУ.
      ОШИБКА!  НЕТ ОБЯЗАТЕЛЬНОГО АКТАНТА.
      ПРОПУЩЕН ПРОБЕЛ МЕЖДУ "В" И "ЛЕСУ".
      ВЫ СОГЛАСНЫ? - (ДА/НЕТ).
П:  ДА
      ОШИБКА ИСПРАВЛЕНА
      ПОНЯТНО  (ВРЕМЯ АНАЛИЗА = 4.18750 СЕК.)
      =НАХОДИТЬСЯ= (БЫТЬ)
        =СУБЪЕКТ=  :  КУЗДРА
        =МЕСТО=    :  ЛЕС (В)

-------------------------------------------------------------

        =ВРЕМЯ=   :  МАЙ — ДЕНЬ → 14
        =ВИД=     :  =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=
П:  ПРАВИЛЬНО
      ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ
П:  ВЧЕРА КУЗДРА ИГРАЛА ВЛЕСУ.
      ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "ВЛЕСУ",
      ОНО ВОСПРИНЯТО КАК НАРЕЧИЕ.
      ПОНЯТНО  (ВРЕМЯ АНАЛИЗА = 2.25000 СЕК.)
      =ИГРАТЬ=  (ИГРАТЬ)
```

=СУБЪЕКТ= : КУЗДРА
=ПАРТНЕР= : ?
=ИГРА= : ?

=ВРЕМЯ= : МАЙ -- ДЕНЬ -+ 14
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=
=ОБСТ-?= : ВЛЕСУ
К КАКОМУ ТИПУ ОТНОСИТСЯ ОБСТОЯТЕЛЬСТВО "ВЛЕСУ" (ВРЕМЕНИ, МЕСТА ИЛИ ОБР.ДЕЙСТВИЯ - ОТВЕТЫ: 1, 2 ИЛИ 3)?

П: НЕВЕРНО. ВО ФРАЗЕ БЫЛА ОШИБКА. ИСПРАВЛЕННЫЙ ТЕКСТ:
ВЧЕРА КУЗДРА ИГРАЛА В ЛЕСУ.
ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 1.75000 СЕК.)
=ИГРАТЬ= (ИГРАТЬ)
=СУБЪЕКТ= : КУЗДРА
=ПАРТНЕР= : ?
=ИГРА= : ?

=ВРЕМЯ= : МАЙ -- ДЕНЬ -- 14
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=
=МЕСТО= : ЛЕС (В)

При анализе второй и третьей фраз возник один и тот же конфликт - "появление незнакомого слова". Однако в первом случае выбранный системой способ его разрешения (формирование словарной статьи нового слова) привел к новому конфликту - на этапе синтаксического анализа выяснилось, что у глагола "была" не заполнена обязательная валентность "где?". Системе пришлось вернуться к анализу словоформы "влесу", которую ей удалось расщепить на знакомый (и ожидаемый) предлог "в" и форму знакомого слова "лес". Во втором случае недоразумение смог обнаружить лишь пользователь.

П: ПРАВИЛЬНО.
ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

П: КУЗДРА СТРИГЕТ МАШУ НА ТЕРРАСЕ НОВЫМИ НОЖНИЦАМИ.
"СТРИГЕТ" - ОШИБКА В СЛОВОИЗМЕНЕНИИ! ПРИ ФЛЕКСИИ

"-ЕТ" В ОСНОВЕ ПРОИСХОДИТ ЧЕРЕДОВАНИЕ: Г - Ж.
ОШИБКА ИСПРАВЛЕНА.
ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "ТЕРРАСЕ",
ОНО ВОСПРИНЯТО КАК СУЩЕСТВИТЕЛЬНОЕ В ФОРМЕ ПРЕДЛОЖНОГО ПАДЕЖА ЕДИНСТВЕННОГО ЧИСЛА. УКАЖИТЕ, ПОЖАЛУЙСТА, РОД ЭТОГО СЛОВА (ВОЗМОЖНЫЕ ОТВЕТЫ: М / Ж / С).

П: Ж

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 2.12500 СЕК.)
=СТРИЧЬ= (СТРИЧЬ)
=СУБЪЕКТ= : КУЗДРА
=ОБЪЕКТ= : МАША

=ВРЕМЯ= : =СЕЙЧАС=/=ВСЕГДА=/=ИНОГДА=
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЕТСЯ=
=ИНОСТР.= : НОЖНИЦЫ -- АТР → НОЖЫЕ
=МЕСТО= : ТЕРРАСА (НА)

П: КУЗДРА И МАША УШЛИ С ТЕРАСЫ.
ПОСКОЛЬКУ ОЦЕНКИ НЕ БЫЛО, СЧИТАЕТСЯ, ЧТО ПРЕДЫДУЩАЯ ФРАЗА ПОНЯТА ПРАВИЛЬНО. АНАЛИЗ ОЧЕРЕДНОГО СООБЩЕНИЯ: КАКАЯ ИЗ ОСНОВ: I. ТЕРРАС, 2. ТЕРАС - ВВЕДЕНА ВЕРНО (I ИЛИ 2, Ø - ЕСЛИ ОБЕ ВВЕДЕНЫ ВЕРНО)?

П: I

Просматривая список новых слов (встретившихся в данном сеансе), система установила, что основы "террас-" и "терас-" близки, а, следовательно, одна из них может быть ошибочной формой другой основы. После ответа пользователя на соответствующий вопрос система исправляет ошибку.

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 3.81250 СЕК.)
=ПЕРЕМЕЩАТЬ= (ПРИ ПОМОЩИ НОГ) (УЙТИ)
=СУБЪЕКТ= : КУЗДРА ← И → МАША
=ОБЪЕКТ= : КУЗДРА ← И → МАША
=ИСТОЧН.= : ТЕРРАСА
=АДРЕСАТ= : ?

=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЕНО=

П: ВЕРНО. КОНЕЦ.
НОВАЯ ИНФОРМАЦИЯ ЗАПОМИНАЕТСЯ В ОЧЕРЕДНОМ СЕАНСЕ.
РАБОТАЕТ БЛОК ОБУЧЕНИЯ
КОНЕЦ СЕАНСА

По директиве "КОНЕЦ" завершается очередной сеанс. Система запоминает словарные статьи новых слов: "глокий", "куздра" и "терраса". Запоминается также и то, что очередной пользователь допустил ошибки "пропуск буквы" и (дважды) "пропуск пробела". Если в последующих сеансах он будет допускать те же ошибки и суммарное количество их (по каждому виду ошибок отдельно) превысит некоторое пороговое значение, произойдет автоматическое переупорядочение причин конфликта "появление незнакомого слова".

Литература

Большакова Е.И., Мальковский М.Г. ЛИСС - диалоговая система синтеза программ // Экспертные системы. Материалы семинара. - М.: изд. МДНТП им. Ф.Э.Дзержинского, 1986. - С. 37-44.

Граудина Л.К., Ицкович В.А., Катлинская Л.П. Грамматическая правильность русской речи. - М.: Наука, 1976.

Красиков Ю.В. Теория речевых ошибок. - М.: Наука, 1980.

Мальковский М.Г. Диалог с ЭВМ на естественном языке: исправление речевых ошибок пользователя // Международный семинар по машинному переводу. Тезисы докладов. - М.: изд. ВЦП, 1983. - С. 128-129.

Мальковский М.Г. Диалог с системой искусственного интеллекта. - М.: изд-во МГУ, 1985.

Мальковский М.Г., Волкова И.А., Грацианова Т.Ю. Языковой процессор системы TULIPS-2: морфологический компонент // Разработка и применение лингвистических процессоров. - Новосибирск: изд. ВЦ СО АН СССР, 1983. - С. 18-36.

Пильщиков В.Н. Язык плэнер. - М.: Наука, 1983.

Трудности русского языка. Справочник. - М.: изд-во МГУ, 1981.

Цейтлин С.Н. Речевые ошибки и их предупреждение. - М.: Просвещение, 1982.

Malkovsky M.G. TULIPS-2 - Natural Language Learning System // COLING-82. Proc. of the Int. Conf. on Comp. Linguistics. - North-Holland P.C., 1982. - P. 191-193.

Ю.С.Мартемьянов,
Е.Н.Морозова

СВОЙСТВА СИНТЕЗИРУЮЩЕГО ОПИСАНИЯ В ПРИМЕНЕНИИ К АНАЛИЗУ ТЕКСТА

Для установления цели, средств и методики анализа текста на естественном языке (ЕЯ) необходимо исходить из достаточно полного и четкого представления о составе и структуре лексико-синтаксической информации, которая должна содержаться в правильно построенном ЕЯ-тексте. Одно из возможных представлений такого рода было получено в ходе разработки системы синтактико-лексического (СЛ) синтеза русского языка группой ФЛОС ВЦП.

Эта система предполагает исходные данные двух уровней глубины:

- базисные синтактико-лексические структуры, которые строятся подстановочной НС грамматикой и получают дополнительные осложнения при своих высказывательных узлах (гиперкатегориальные, информирующие о частичном употреблении порядка высказываний в системе составляющих - ср. "пресуппозиция", "эмфаза"), при отношениях (метакатегориальные, сообщающие о содержательной "роли" членов отношения в составе высказывания - ср. "агент", "пациенс") и при концевых категориях этих отношений (гипокатегориальные, указывающие на формально грамматические признаки рода - числа - лица - падежа и одушевленности);

- "модулированные" синтактико-лексические структуры, получаемые путем преобразования синтаксиса базисных структур при сохранении их лексики.

Соответственно, система включает:

- синтез глубинный, переводящий базисные структуры в модулированные, и

- синтез поверхностный, доводящий модулированные структуры до нормальных ЕЯ-предложений.

Глубинный синтез осуществляется посредством элементарных модуляций (около двух десятков), определенным

образом сочетающихся между собой. При модулировании в базисной структуре:

- изменяется взаимное расположение и связи частных подструктур - в сторону приближения к поверхностной форме;

- в частности - вычленяются ("экспонируются") такие высказывательные подструктуры, которые должны получить форму отдельных предложений, как следствие, узлы их первоначального вхождения помечаются признаками "текстового" повтора.

В ходе модулирования базисных структур, т.е. в результате применения синтаксических операций:

- пополняется набор синтаксических отношений (от 2-х до десятка) и соответствующих парных концевых категорий (с двух пар до десяти);

- концевые категории разных синтаксических отношений, базисных и модуляционных, группируясь в узлах структуры, образуют различные комбинации, моделирующие синтаксические типы слов ("части речи"), которых оказывается свыше 2-х десятков;

- при наличии определенной базисной лексики у двух смежных узлов лексема подчиненного узла объединяется с лексемой подчиняющего узла в сложную лексическую единицу; ср.: "(прошедшее + (завершенность + (начинательность + работает)))".

Мотивом модулирования является:

- с одной стороны - принципиальная речевая "анормальность" базисных структур, предназначенных только для четкой фиксации смысла; например, анормальность заключается в описательном, внесинтаксичном уточнении как высказывательной упорядоченности СЛ-структуры - с помощью гиперкатегорий "пресуппозиция (= тема)", "эмфаза", "сирконстанция", - так и признаков денотативной тождественности и текстовой повторности узлов;

- с другой стороны, мотивом модулирования оказывается определенная синтаксическая "дефективность" ЕЯ лексических единиц, в частности - несамостоятельность, аффиксальность их выражения; именно поэтому для поверх-

ностной реализации базисной лексики и требуется предварительная перестройка базисной синтаксической структуры, приводящая либо к появлению необходимых категориальных комбинаций (= "части речи"), либо к объединению базисной лексемы подчиненного узла с базисной лексемой узла подчиняющего.

Каждый из мотивов синтаксического преобразования плюс группа преобразующих модуляций составляют, соответственно, вход и выход некоторой трансформации.

Комбинации синтаксических категорий и сложные лексические единицы, полученные в результате трансформации (= мотивированного модулирования) базисных структур, составляют в свою очередь мотив, или вход, лексических трансформаций, осуществляемых поверхностным синтезом.

Поверхностный синтез, как уже сказано, имеет дело с модулированной вычлененной подструктурой (будущим предложением) и заключается прежде всего в преобразовании ее лексики:

- базисная лексика узлов подструктуры заменяется на поверхностную - в строгом соответствии с возникшими при модулировании комбинациями синтаксических категорий этих узлов и с их лексическим составом; заменяющее поверхностное слово оказывается, соответственно, либо синтаксическим дериватом базисной лексемы (читает → чтение или читающий), либо также емким аффигированным или супплетивным композитом, эквивалентным сложной лексической единице одного узла ("прошедшее + завершенность + начинательность + работает" = "заработал") или связному набору базисных лексем некоторого фрагмента структуры ("имеющий средством голографию" = "голографический");

- при лексических заменах особым образом учитывается синтактико-лексический контекст х заменяемой базисной лексемы, позволяющий выбрать для нее оптимальный фразеологический вариант (создает: дом → строит, проект → составляет; дом → строительство, чертеж → построение).

Лексические трансформации сопровождаются:

- введением при поверхностных словах определенных словоизменительных признаков - гипокатегорий числа, рода, лица, падежа, одушевленности;

- последующим распределением "несобственных" (управление) и "собственных" (согласование) гипокатегорий по определенным узлам;

- словоизменительным синтезом "гипокатегоризованных" слов с помощью обобщенных морфологических таблиц (120 типов);

- линейными преобразованиями структуры полученных словоформ, мотивированными либо требованиями самих словоформ, либо недореализованными гиперкатегориями и признаками текстового повтора.

Выводы для анализа

Анализ текста должен иметь объектом как его синтаксическую структуру, так и лексику и осуществляться в двух версиях, различающихся объемом анализируемой единицы (предложение - абзац) и глубиной результирующего представления.

Поверхностный анализ должен состоять в замене отдельного предложения на модулированную СЛ-структуру, т.е. на такую, где в базисном виде представлена только лексика, а выявленные синтактико-модуляционные отношения, признаки денотативного тождества и текстового повтора далее не раскрываются.

Углубленный анализ должен приводить к замене абзаца, т.е. нескольких предложений, связанных текстовыми повторами, на базисную структуру, где базисные слова объединены исключительно базисными синтаксическими отношениями (два вместо десяти), а модуляционные перестройки и возникающие при этом новые отношения либо элиминированы - если они мотивированы условностями поверхностной лексики или "нормальной" ЕЯ структуры (см. выше), - либо заменены на гипокатегории ("тема", "эмфаза", "сирконстанция"), выражающие коммуникативные интенции автора текста.

Главным средством поверхностного анализа должно стать синтактико-лексемное описание поверхностных словоформ:

- вместо синкретичных и не-операционных "частей речи" традиционного синтаксиса должны использоваться комбинации синтаксических категорий, являющихся концами соответствующих базисных или модуляционных отношений;

- каждой категориальной и гипокатегориальной характеристике словоформы должны быть сопоставлены линейные поисковые признаки, определяющие очередность, направление и способ поиска слов с парными синтаксическими категориями и субкатегориями; вынужденная нетипичность направления или очередности поиска могут быть указанием на особый мотив - например, линейно-выраженную гиперкатегорию ("тематичность" или "эмфатированность" слова с его зависимыми) или его текстовую повторность;

- словоформе-омониму со специфической синтаксической категоризацией должна быть поставлена в соответствие ее базисная лексема или так или иначе организованная совокупность базисных лексем; другими словами, с поправкой на омонимичность используются в обратном направлении лексические трансформации синтезирующего описания.

Аналогичным образом должны использоваться в обратном направлении - от выявленной группы модуляций к мотиву - и синтаксические трансформации при переходе к глубинному анализу:

- мотивы, связанные со специфическим деривативным или композитным строением поверхностной лексики, элиминируются вместе с устранением всех соответствующих модуляций;

- напротив, мотивы, связанные с поверхностной лексико-синтаксической реализацией гиперкатегорий, после восстановления базисной структуры сохраняются в ее узлах в качестве существенной коммуникативной информации.

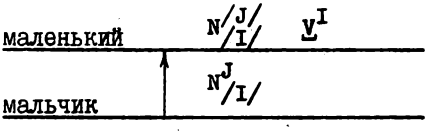
Все эти естественные следствия синтезирующего описания ФЛОС позволяют по-новому поставить и решить целый ряд традиционных проблем анализа - таких, как проблему "сильных" и "слабых" связей, трактовку "непроективных" структур, синтаксической омонимии и семантической интерпретации поверхностных отношений, а также обосновать адаптивность предлагаемого типа двухэтапного анализа для целей перевода между языками с различным синтактико-морфологическим строем.

Диалоговый синтаксический анализатор (ДСА) предназначается для автоматического синтаксического анализа текста на естественном языке.

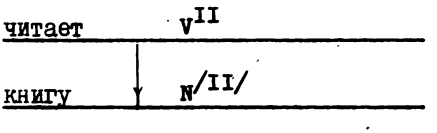
На вход ДСА поступает предложение естественного языка, каждое слово которого обработано программами морфологического анализа.

Результатом синтаксического анализа является структура (дерево подчинения), получаемая на основе синтаксической информации, содержащейся в словаре, и результатов морфологического анализа. При этом каждому слову из входного предложения в выходном дереве соответствует некоторое поддерево (в частном случае - отдельная вершина). Связи - стрелки между вершинами дерева направлены от корня поддерева (главного слова) к зависимым от него словам. Началу и концу каждой стрелки приписываются некоторые согласованные друг с другом (парные) синтаксические категории. Таким образом, каждая вершина дерева получает набор синтаксических категорий (приписанных концу входящей стрелки и началам исходящих). Связи упорядочены по тесноте. Эта упорядоченность выражается во взаимном расположении категорий, приписанных к началам стрелок, исходящих из вершины, чем определяется и упорядоченность самих этих стрелок. Таким образом, дерево содержит информацию не только о том, какое слово с чем связано, но и по какой категории и с какой теснотой.

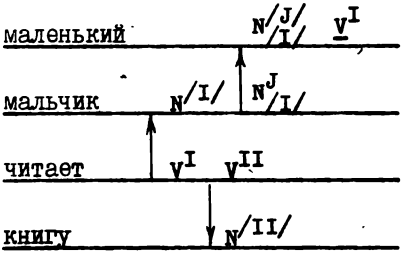
ДСА предусматривает два уровня анализа: уровень установленных связей и интерпретационный уровень. На первом уровне анализа устанавливаются предварительные связи между словами предложения с частичным упорядочением по тесноте связи. Последняя особенность означает, что связываются между собой не отдельные слова (это частный случай), а группы слов (уже связанные между собой с большей теснотой, чем теснота связи между группами). Так, например, в предложении Маленький мальчик читает книгу более тесной оказывается связь между словами маленький и мальчик, читает и книгу, а связь между последней парой и первой оказывается менее тесной. На этапе интерпретационного анализа происходит приписывание концам полученных связей - стрелок определенных синтаксических категорий; тем самым вершинам дерева оказывается приписан упорядоченный набор этих категорий. Вернемся к примеру маленький мальчик читает книгу. После установления связи между словами маленький и мальчик концам ее приписываются категории:



Концам связи между словами читает и книгу приписываются категории:



Наконец, после установления связи между группами слов маленький мальчик и читает книгу ее концам приписываются категории V^{I} и $N^{/I/}$:



Здесь слово читает - корень дерева.

В общем случае слову предложения может соответствовать в результирующем дереве целое поддерево. (Такие поддеревья содержатся в словаре.) Основная идея алгоритма синтаксического анализа: сборка результирующего дерева из словарных поддеревьев по определенным правилам на основе некоторой словарной информации.

Описание словарей

Словари синтаксического анализа строятся в соответствии с существованием двух уровней анализа.

Словарная статья каждого слова содержит:

- зону поверхностной лексики (само слово и возможная информация о том, что оно имеет омонимы);
- зону морфологической информации;
- ссылку на словарь таблиц поисковых признаков (см. ниже);
- ссылку на словарь внутренних поддеревьев (см. ниже);
- зону глубинной лексики (приписываемой вершинам поддерева), соответствующего данному слову поверхностной лексики).

Такая словарная статья поступает на вход морфологического анализатора. Непосредственным входом синтаксического анализатора (после морфологического анализа) является массив блоков информации, каждый из которых состоит из:

- словоформы, взятой из предложения (для омонимичных слов указывается также номер омонима в словаре);
- грамматических субкатегорий (род, число, лицо,

падеж, одушевленность), полученных в результате морфологического анализа;

- таблицы поисковых признаков (выбранной из соответствующего словаря по ссылке; при этом некоторые поля таблицы могут меняться с учетом результатов морфологического анализа);

- поддерева (выбранного по ссылке из словаря внутренних поддеревьев);

- глубинной лексики.

Существование двух разных словарей - словаря поддеревьев и словаря таблиц поисковых признаков - вызвано тем, что существуют целые классы слов поверхностной лексики, развертывающиеся при анализе в одинаковые поддеревья (возможно, с одинаковой таблицей поисковых признаков), и представления слов одного такого класса в результирующем дереве различаются между собой только приписыванием разной глубинной лексики к вершинам одинаковых поддеревьев.

Словарь таблиц поисковых признаков (ТПП) используется на уровне установления связей. Каждая имеющаяся в словаре ТПП соответствует некоторому поддереву из словаря поддеревьев (возможно, одному поддереву соответствует несколько разных ТПП). При этом в ТПП для некоторых категорий из данного поддерева занесена информация о требованиях к искомому слову, т.е. о том, с чем может быть связано слово по данной категории (так называемые поисковые признаки и функциональные параметры), а именно:

- является данный поисковый признак активным или пассивным (А или П) (т.е. требует ли появление данного слова в предложении наличия определенным образом связанных с ним слов и активизации поиска этих слов);

- каково направление поиска (Ал или Ап) (справа или слева от данного активного слова следует искать слово, связь с которым должна осуществляться по этому признаку);

- разрешается ли в процессе анализа изменять направление поиска на противоположное;

- какова теснота связи по данному поисковому признаку (и соответствующей категории);

- должно ли данное слово стать главным (АГ, ПГ) или зависимым (АЗ, ПЗ) по соответствующей связи (или, что то же самое, сопоставлено ли ему начало или конец соответствующей стрелки);

- здесь же помещается информация об управлении по данному поисковому признаку (т.е. данной категории); требование определенного падежа, наличие предлога и т.п.

Словарь поддеревьев используется только на уровне интерпретационного анализа. Он содержит поддеревья, соответствующие классам слов поверхностной лексики с приписанными к их вершинам категориями.

Кроме уже описанных словарей в ДСА используется словарь предложных конструкций, таблица морфологической омонимии, таблица парных категорий с информацией о том, какая обработка категорий может потребоваться при интерпретационном анализе.

Описание алгоритмов анализа

ДСА использует:

- общий алгоритм создания рабочего пространства и выбора пар слов, между которыми вероятна связь;

- частные алгоритмы установления связей по поисковым признакам и изменения информации в ТПП;

- алгоритмы обработки конфликтных ситуаций, возникающих при построении дерева;

- алгоритм интерпретации связей.

Частные алгоритмы и алгоритм обработки конфликтных ситуаций используются на уровне установления связей; алгоритм интерпретации - на интерпретационном уровне.

Общий алгоритм

Общий алгоритм создает рабочее пространство (РП), т.е. последовательно выбирает словарную информацию для каждого из слов предложения (начиная с первых двух слов и добавляя по одному слову на каждом очередном цикле анализа до тех пор, пока предложение не будет исчерпано).

У слов, находящихся в РП, в определенном порядке отыскиваются активные поисковые признаки, для каждого найденного такого признака в определенном порядке выбирается парное слово и вызывается соответствующий частный алгоритм установления связи между словами. В случае, если связь на данном РП установить не удается, соответствующий поисковый признак помечается как испробованный для данного РП. К обработке этого признака алгоритм вернется после расширения РП.

Если же частный алгоритм устанавливает связь по данному поисковому признаку, и эта связь подтверждается алгоритмами обработки конфликтных ситуаций и алгоритмом интерпретации, то соответствующий поисковый признак помечается как использованный. Связь по нему считается окончательной, если не будет отменена при обработке конфликтных ситуаций на одном из следующих циклов (при расширенном РП).

Расширение РП происходит после того, как на данном РП испробованы все использованные поисковые признаки.

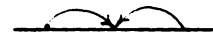
На разных этапах и разных циклах анализа предусматривается выдача промежуточного состояния РП и возможность вмешательства пользователя в режиме диалога (если пользователь считает, что анализ идет ошибочно).

Частные алгоритмы установления связи

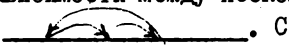
Частных алгоритмов установления связи столько, сколько активных поисковых признаков. Их назначение состоит в том, чтобы на данном РП найти слово с поисковым признаком, парным данному (по ТПП), при этом,

если необходимо, производится проверка на управление. Если такое слово найдено, то вновь установленная (потенциальная) связь с ним передается для проверки алгоритму обработки конфликтных ситуаций.

Алгоритмы обработки конфликтных ситуаций

Часто встречающейся при построении дерева конфликтной ситуацией является попытка подчинить слово, уже подчиненное по какой-то другой связи (ситуация "слуга двух хозяев"). Если конфликтующие связи при этом направлены в разные стороны (одна - от слова слева, другая - от слова справа) , то решением конфликтной ситуации будет отмена той из них, которая установлена по более слабому поисковому признаку (в соответствии с некоторой иерархией, упорядочивающей поисковые признаки по силе связи).

Если же эти связи направлены в одну сторону , то отменяется та, которая установлена между более удаленными друг от друга словами.

Довольно редко встречается конфликтная ситуация, при которой связи зависимости между несколькими словами образуют цикл: . Способ ее решения - отмена связи по самому слабому поисковому признаку (если есть несколько связей с одинаковыми поисковыми признаками, то требуются специальные проверки).

Возможная конфликтная ситуация "пересечение связей" (слово с номером **m**, например, зависит от слова с номером **s**, а слово с номером **q** - от слова с номером p, и **p < s < q < m**).



p s q m

Требуются специальные проверки синтаксических категорий, по которым установлены связи.

Еще одна конфликтная ситуация - "покрытие вершины":



- решается также после определенных проверок синтаксических категорий (в некоторых случаях последние две ситуации являются допустимыми и подлежат особой интерпретации).

Если возможность установления той или иной связи подтверждается алгоритмами обработки конфликтных ситуаций, то связь передается на интерпретацию.

Алгоритм интерпретации

На первом уровне анализа устанавливаются только предварительные связи на основании соответствия друг другу поисковых признаков двух слов. Из соответствия поисковых признаков, вообще говоря, не следует соответствие друг другу категорий, с которыми эти признаки связаны. Возможны и такие случаи, что поисковому признаку не соответствует никакая категория в словарном поддереве, но тем не менее связь между словами является правильной. Под интерпретацией как раз и понимается проверка соответствия друг другу категорий, приписываемых концам предварительной связи, а также определение категории, которая должна быть приписана концу связи в том случае, когда в словарном поддереве этой категории нет (например, в словарных поддеревьях существительных не содержится категорий, по которым с ними могут быть связаны определения - адъюнкты, и в результирующем дереве эти категории появляются именно в процессе интерпретации).

Соответствие друг другу категорий и возможность приписывания концам связей категорий, не содержащихся в словарных поддеревьях, определяется алгоритмом интерпретации с помощью таблицы парных категорий. В случае, если невозможно установить такое соответствие, связь считается неинтерпрети-

руемой и отменяется. При соответствии категорий связь признается окончательной для данного РП.

Алгоритмы изменения информации

Каждому частному алгоритму установления связи соответствует особый частный алгоритм изменения информации, обращение к которому происходит после того, как связь по данному поисковому признаку проинтерпретирована (т.е. подтверждена алгоритмом интерпретации). При определенных условиях меняются некоторые элементы скопированных во входные структуры синтаксического анализа ТПП, что во многих случаях позволяет ускорить установление связей между словами в предложении.

Особенности алгоритма синтаксического анализа

I. Главным достоинством предложенного алгоритма является его языковонезависимость. Вся информация о языке входного предложения содержится в словарях. Частные алгоритмы установления связей независимы от частей речи и имеют дело только с поисковыми признаками и функциональными приметами .

2. Предложенный алгоритм синтаксического анализа является обозримым в силу своей дифференцированности. Этапы анализа (построение и расширение рабочего пространства, установление связей, обработка конфликтных ситуаций, интерпретация, изменение информации) четко отделены друг от друга, что может позволить в случае ошибочного анализа быстро найти место ошибки.

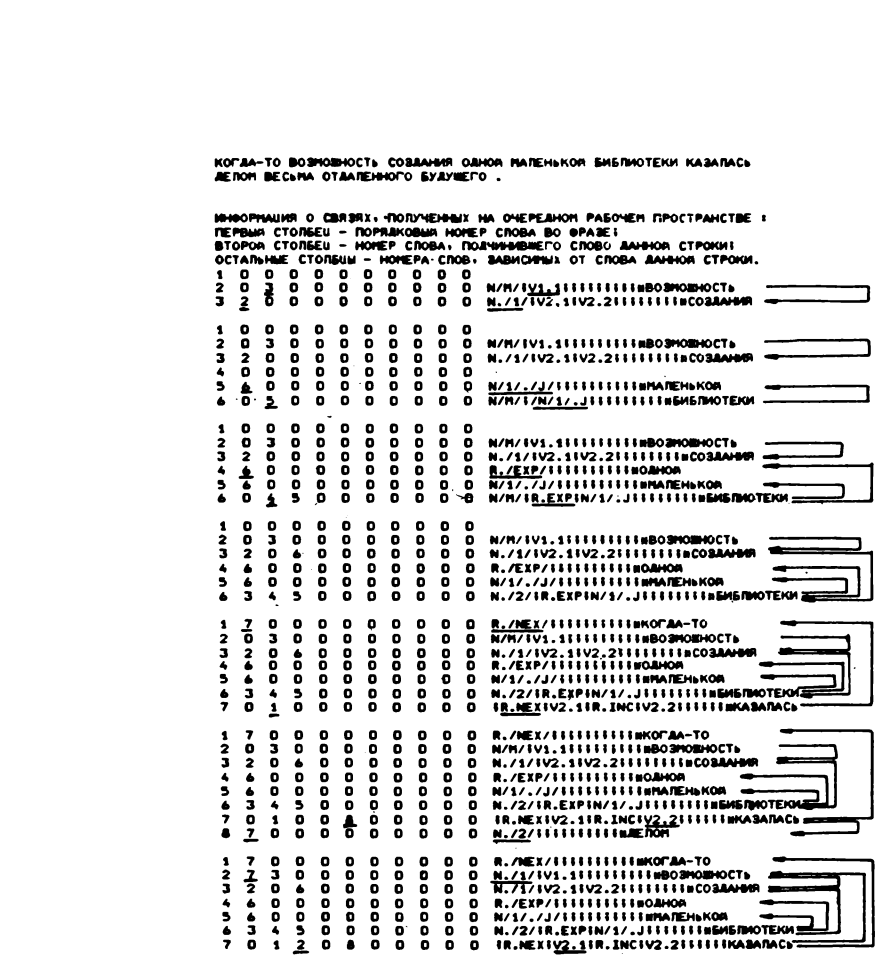
3. Такой алгоритм синтаксического анализа позволяет быстро отсеивать омонимы. Обработка омонимов происходит следующим образом: омонимы выбираются из словаря в определенном порядке, таком, что пригодность или непригодность омонимов, встретившихся первыми, проверить наиболее легко. Как

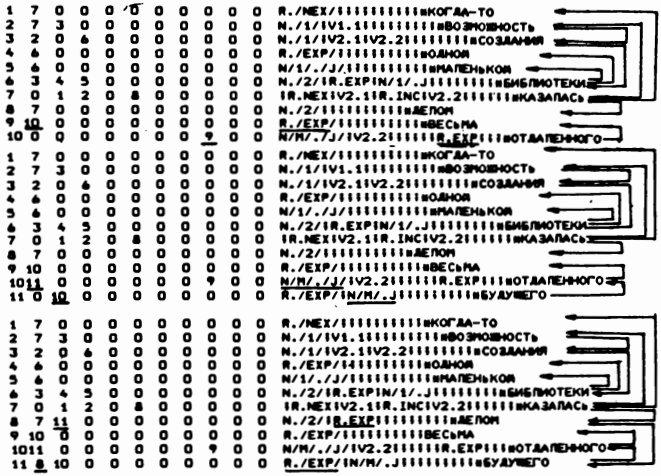
только с одним из омонимов установлена интерпретированная связь, остальные омонимы отбрасываются, т.е. отсев омонимов происходит на минимальном рабочем пространстве. В огромном большинстве случаев выбор омонима при этом оказывается правильным, хотя возможны и ошибки (при работе в режиме диалога в таких случаях можно предусмотреть вмешательство пользователя).

4. Возможна обратная связь с информацией (при установлении каждой окончательной для данного РП связи вызывается один из алгоритмов изменения информации).

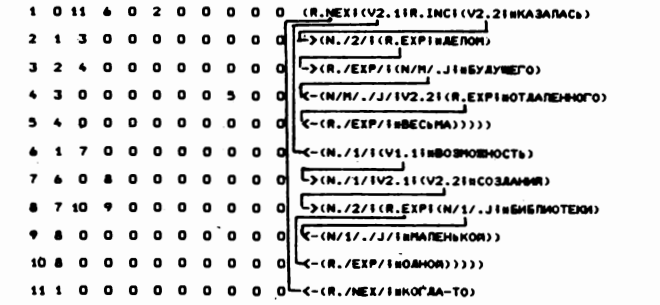
5. Результат анализа является единственным.

КОГДА-ТО ВОЗМОЖНОСТЬ СОЗДАНИЯ ОДНОЙ МАЛЕНЬКОЙ БИБЛИОТЕКИ КАЗАЛАСЬ
ДЕЛОМ ВЕСЬМА ОТДАЛЕННОГО БУДУЩЕГО .





ПОСТРОЕННОЕ ДЕРЕВО



О.Н.Гринбаум

АВТОМАТИЧЕСКАЯ СТРУКТУРИЗАЦИЯ ТЕКСТОВ (НА МАТЕРИАЛЕ РУССКОЙ КЛАССИЧЕСКОЙ ПРОЗЫ)

В статье описывается формально-пунктуационный метод структуризации художественной прозы, базирующийся на алгоритмическом подходе к основным правилам русской пунктуации. Метод разработан, реализован на ЭВМ и используется для автоматической разметки художественных текстов (выделение глав, страниц, абзацев, предложений, словоупотреблений) и маркировки абзацев по типам письменной речи. Проверка представленного здесь метода структуризации велась на материале полного текста романа М.Ю.Лермонтова "Герой нашего времени" и показала его высокую эффективность.

В первой части статьи обсуждается проблема структуризации художественной прозы применительно к формированию текстовой части Машинного фонда русского языка, описан метод структуризации и его программно-алгоритмическое ядро. Во второй части описаны алгоритм маркировки абзацев по типам письменной речи, алгоритм детализации видов прямой речи, а также система диалоговой корректировки структурированного текста. Представлены основные результаты структуризации полного текста романа "Герой нашего времени".

I. Формально-пунктуационный метод структуризации

Общие положения. Концепция Машинного фонда русского языка (МФ РЯ) предусматривает (Машинный фонд... 1986) разработку, создание и развитие двух основных частей фонда - словарно-грамматической и текстовой; их гармоничное сочетание или "сосуществование" (Шайкевич 1986) должно обеспечить функционирование фонда как многоцелевой научно-исследовательской и информационной системы. Создание текстовой части фонда, т.е. представительного корпуса русских текстов различных типов: литературно-художественного, научно-технического,

публицистического, фольклорного, организационно-распорядительного и др. (Караулов 1986; Герд 1986а) за последнее время из проблемы дискуссионного плана (Гиндин 1986) все более становится осознанной необходимостью и превращается в одну из аксиом компьютеризации филологических исследований. Об актуальности создания текстовой части МФ РЯ свидетельствуют и многочисленные зарубежные работы в области компьютерной лингвистики (подробнее см.: Тулдава 1986; Розенцвейг 1986).

Утвердившийся в компьютерной лингвистике постулат о примате текста при своей практической реализации встречает ряд сложных вопросов (см., например, Андрющенко 1986; Котов 1986; Герд 1986б), среди которых далеко не последним является вопрос о структуризации текстов. Потребность в проведении филологических, в особенности лингво-статистических исследований диктует необходимость включать в состав текстовой части фонда такие произведения художественной прозы, для которых выполнены - как минимум - следующие два условия:

1. Любой текст, хранимый в памяти ЭВМ, должен быть структурно-фрагментирован [1] с выделением томов, глав, страниц, абзацев, предложений и других структурных элементов.

2. Абзацы художественно-прозаических текстов должны быть разделены минимум на две категории: абзацы авторского повествования и абзацы с элементами прямой речи.

Первое условие существенно, например, при создании машинного фонда лексических единиц (Рогожникова 1985) или в разного рода филолого-статистических исследованиях (Тулдава 1977). Действительно, применяемое ныне правило определения контекста как определенного числа символов

---

[1] Мы исходили из того факта, что полные авторские тексты вводятся и хранятся в памяти ЭВМ в форме, идентичной печатному оригиналу. Это тем более справедливо для систем, сопряженных с фотонаборными автоматами (Андрющенко 1985).

вправо и влево от ключевого слова (Рогожникова 1985; Азарова и др. 1983) не может считаться безупречным и используется только потому, что текст, хранимый в памяти ЭВМ для получения карточек-цитат, не структурирован. При наличии же структурно-фрагментированного текста понятие "контекст" даже в формальных (компьютерных) системах имело бы неформальное (смысловое) значение, а сам контекст мог бы по желанию исследователя включать различное число предложений, абзацев или еще более крупных образований разной или одинаковой длины вправо и влево от ключевого слова.

Второе условие является следствием того, что авторское повествование в художественно-прозаическом тексте распадается на авторскую речь и речь персонажей. Без учета этого обстоятельства можно прийти к ошибочным выводам в любых филологических исследованиях, особенно в тех, которые основаны на использовании статистических методов (Лесскис 1962; Мартыненко 1983).

Если быть более точным, то следует отметить, что условие, указанное нами вторым (разграничение собственно авторских и не авторских фрагментов повествования) с необходимостью требуют предварительного решения первой задачи (структуризации текста). И это понятно, ибо создавая программу автоматической структуризации, можно было бы задать перечень формальных признаков, указывающих на начало прямой речи, но решить, где же она заканчивается - без указания, как минимум, границ абзацев - попросту невозможно. Также невозможно, в общем случае, отделить заключенную в кавычки прямую речь от предложений с цитатами, речениями, различными названиями и т.п., если отсутствуют сами границы предложений. Структурирование же вручную произведений художественной прозы - даже в режиме диалога с компьютерной системой - занятие архисложное, особенно если учесть масштабы этой работы. И, напротив, при наличии предварительно структурированного текста

работа исследователя по его корректировке, проводимая в режиме диалога с ЭВМ, будет иметь достаточно высокую практическую эффективность.

Таким образом, решение проблемы структуризации прозаических художественных текстов, ввиду своей сложности и чрезвычайной трудоемкости, не может рассматриваться иначе, чем в ракурсе автоматической обработки.

Формально-пунктуационный метод структуризации прозаических текстов. Содержательно задача структуризации прозаических текстов формулируется (Гринбаум, Мартыненко, Фитиалов 1985) следующим образом: текст, введенный в память ЭВМ в виде, идентичном печатному оригиналу, должен быть структурирован формальным способом с указанием:

а) границ страниц, абзацев, предложений, словоупотреблений;

б) принадлежности структурных элементов текста (абзацев и предложений) к одному из типов письменной речи (авторская или прямая речь).

Минимальное требование к введенному в ЭВМ тексту - выделенность прописных букв, поскольку большинство современных ЭВМ еще не позволяют одновременно пользоваться строчными и прописными буквами.

Задача структуризации включает в качестве самостоятельных следующие подзадачи:

а) разработка программно-алгоритмического комплекса для структурной фрагментации текста;

б) разработка алгоритмов и программы формального разграничения и классификации абзацев по типам письменной речи - авторской и прямой;

в) разработка алгоритма и программы детализации абзацев прямой речи с выделением абзацев собственно прямой речи, абзацев "смешанной" речи (содержащих и авторскую, и прямую речь) и абзацев "вложенной прямой" речи (прямая речь внутри абзаца собственно прямой речи);

г) построение эффективной диалоговой системы, позволяющей легко проверять правильность и - при необходимости - корректировать структурированный художественный текст.

Для решения первых трех подзадач был разработан и реализован формально-пунктуационный метод, основанный на:

а) детальном анализе правил русской пунктуации;

б) трансформации этих правил в формально-логическую схему распознающего автомата - алгоритмическое ядро всего метода структуризации [2].

Разработка формально-пунктуационного метода структуризации художественной прозы натолкнулась на ряд существенных противоречий между нечеткостью (размытостью) пунктуационной системы, с одной стороны, и строгостью формализованной записи алгоритма - с другой. Нечеткость использования знаков препинания, отсутствие обязательной регламентации и почти стихийный процесс исторического развития пунктуации (Шапиро 1974) - все это, видимо, тормозило развитие работ по формальной структуризации текстов. Тем более, что художественные тексты, увидевшие свет в разное время (например, даже с прошлого века и по сегодняшний день) несут в себе не только характерные для своей эпохи пунктуационные особенности, авторские предпочтения и вольность при употреблении тех или иных знаков, но и редакторскую правку последующих изданий.

Такое наслоение многих "мешающих" факторов резко усложняет процесс формализации пунктуационных правил. Но поскольку главной задачей при построении формально-логической схемы структуризации была задача вычленения предложений, из рассмотрения были исключены

---

[2] В нашем исследовании отсутствует сопоставительный анализ при выборе наилучшего метода решения, поскольку альтернативные методы (кроме ручных) нам неизвестны.

многие знаки, не относящиеся к разграничительным знакам препинания, в том числе запятые, двоеточия, точки с запятой. Однако некоторые из них были впоследствии использованы нами при разработке процедуры маркировки абзацев и детализации видов прямой речи; это - тире, двоеточие, кавычки, скобки.

Выделение предложений построено на анализе правого и левого окружения знаков препинания, потенциально выступающих в качестве разграничителей предложений: точки, вопросительного и восклицательного знаков, многоточия. Необходимость анализа окружения таких знаков диктуется тем обстоятельством, что в художественных текстах разграничительные знаки ".", "?" и "!" помимо своего традиционного места в конце предложения часто встречаются и в середине предложений, тогда как для других типов текстов это практически исключено.

Итак, построение формально-логической схемы структуризации есть - в самом общем виде - задание последовательного перечня ситуаций (состояний распознающего автомата), каждая из которых предопределяет один из возможных результатов анализа (элемент множества возможных решений). Общая схема принятия решения относительно выделения из текста очередного предложения выглядит следующим образом:

а) распознается символ, принадлежащий множеству потенциальных разграничителей предложений;

б) производится анализ правого или левого окружения выделенного символа (поиск на дереве альтернатив пунктуационных ситуаций);

в) решение о выделении очередного предложения принимается лишь в том случае, если поиск на дереве альтернатив завершился в терминальной вершине, соответствующей одной из решающих ситуаций.

Для примера рассмотрим две простейшие ситуации (значения двух терминальных вершин):

⟨знак + пробел + прописная буква⟩ (I)

и ⟨знак + пробел + буква (не прописная)⟩, (2)

где знак " есть либо ".", либо "?", либо "!", а "+" используется для обозначения операции сцепления (конкатенации). Ситуация (I) является решающей, а ситуация (2) таковой не является. Иллюстрацией к (I) может служить следующая строка текста:

ции, нам отвели ночлег в дымной сакле.
Я пригласил (3)

(Лермонтов М.Ю. Герой нашего времени)

и к ситуации (2) - другая строка из того же произведения:

чем у Бэлы; а какая сила! скачи хоть на
пятьдесят (4)

После распознавания в строке (3) разграничительного знака ".", к очередному i=ому выделяемому предложению будет отнесена вся левая часть строки (3) вплоть до пробела перед буквой "Я"; этой буквой начнется следующее (i + I)-ое предложение, в которое будут включены два слова строки (3), расположенные правее знака ".", и все последующие слова вплоть до выявленного центра (i + I)=ой решающей ситуации. В строке (4) решающей ситуации нет, следовательно, эта строка целиком войдет в очередное выделяемое предложение.

Маркировка абзацев и предложений по типам письменной речи, а также детализация видов прямой речи производятся аналогичным образом, хотя и требуют других решающих правил для других пунктуационных ситуаций.

Анализ и формализация правил русской пунктуации. Формально-пунктуационный метод структуризации художественной прозы базируется на алгоритмическом подходе к основным правилам русской пунктуации. К этим правилам мы, в первую очередь, относим такие правила, которые позволяют проводить разграничение (выделение) предложений в тексте, исходя только из их графического (печатного) изображения.

Известные правила пунктуационного оформления конца предложения были подвергнуты нами тщательному изучению на предмет составления перечня всех возможных и встречающихся в реальных художественных текстах ситуаций - таких, в которых символы окончания предложения (".", "?", "!") действительно означают конец предложения, а также ситуаций, в которых это места не имеет. В результате анализа выявлено три варианта ситуаций, безусловно фиксирующих конец предложения:

А. ZAG - ситуация: знак перед заголовком или подзаголовком.

Б. АО - ситуация: знак перед началом нового абзаца (перед появлением в тексте абзацного отступа).

В. FIN - ситуация: знак внутри абзаца в окружении других знаков или символов, образующих вместе с ним решающую ситуацию.

Вариант А очевиден: заголовок (подзаголовок) не может разрывать предложение на части, следовательно, умение формально выявлять в тексте заголовки (подзаголовки) позволит безошибочно фиксировать конец предшествующих им предложений.

Вариант Б тоже достаточно прозрачен, но требует некоторой детализации. Дело в том, что предложения, состоящие из прямой речи и предшествующих ей вводящих слов автора, нередко размещаются на разных строчках печатного текста; в таких случаях вводящие слова заканчиваются двоеточием, а прямая речь оформляется абзацным отступом (АО), за которым следует тире. Например:

Она вдруг почувствовала себя оскорбленной и сказала холодно:

- Нам нужно расстаться на некоторое время, а то от скуки мы можем...

(Чехов А.П. Попрыгунья)

Наличие абзацного отступа, таким образом, сигнализирует о конце предложения лишь в том случае, если новая

строка (включающая этот АО) не начинается с тире. Если же после абзацного отступа следует тире, а предыдущая строка заканчивается двоеточием, то АО-ситуация

⟨Двоеточие + АО + тире + ПБ⟩ ,

где ПБ - прописная буква, решающей не является. Если двоеточия в конце строки, предшествующей АО, нет, то ситуация:

⟨АО + тире + ПБ⟩

является решающей.

Вариант В из всех трех является наиболее сложным, поскольку выявление FIN -ситуации производится внутри абзаца и, следовательно, может базироваться только лишь на анализе взаимного расположения знаков и символов вокруг потенциального центра решающей ситуации. Эти центры, как мы уже отмечали, образуются знаками ".", "?" и "!". Обозначим множество таких знаков препинания через Z , тогда $Z = (Z_1, Z_2, Z_3)$ и Z_1 = ".", Z_2 = "?", Z_3 = "!".

Любая FIN -ситуация рассматривается нами как последовательность трех более простых ситуаций F, I и N (FIN = F + I + N):

I. F -ситуация ("конец предложения") фиксируется при наличии одного из элементов множества F = Z + T, где T = (пусто, кавычки, закрывающая скобка). Иначе, F = (Z, Z", Z)) , что соответствует любому набору символов, существенному для формального выявления в тексте ситуации "конец предложения".

2. I-ситуация ("интервал") определяется теми символами, которые используются для заполнения интервала между двумя предложениями, расположенными внутри одного абзаца: I = (i_1, i_2) и i_1 ="пробел", i_2="тире".

3. N -ситуация ("начало предложения") фиксируется при наличии одного из элементов множества N = P + E, где P=(пусто, кавычки, открывающая скобка), а E - мно-

жество всех прописных букв. Иначе, N =(E, "E, (E), что соответствует набору символов, используемому для графического изображения начала любого предложения, находящегося в середине абзаца.

Приведем примеры, поясняющие введенные формализмы.

Пример I.

> Вдоль оврага - по одной стороне опрятные амбарчики, клетушки с плотно закрытыми дверьми; по.другой стороне пять-шесть сосновых изб с тесовыми крышами. - Над каждой крышей высокий шест скворечницы...
>
> (Тургенев И.С. Стихотворения в прозе. "Деревья")

В этом отрывке тире встречается три раза: в первых двух случаях - в середине предложения, а в последнем случае - между двумя предложениями. FIN -ситуация для последнего случая выглядит следующим образом:

f ="точка", i ="тире", n ="Н"; тогда в целом

fin =⟨точка + тире + "Н"⟩ .

Пример 2.

> Придет, бывало, к нам в деревню; я то ему племянником довожусь. - "Что, брат Вася, - скажет, - приди-ка, брат, переночуй-ка у меня!"
>
> (Тургенев И.С. Три встречи)

В этом случае FIN -ситуация формируется так: f ="точка", i ="тире", Р="кавычки", e ="Ч". Следовательно: fin =⟨точка + тире + кавычки + "Ч"⟩ .

Пример 3.

> "А своими наемными работниками ты доволен?" -
> "Да, - процедил сквозь зубы Николай Петрович, -..."
>
> (Тургенев И.С. Отцы и дети)

Здесь FIN -ситуация образована следующей последовательностью символов: Z ="?", t ="кавычки", i ="тире", р="кавычки", e ="Д". Следовательно: fin = <? + кавычки + тире + кавычки +"Д" > .

Пример 4.

> Да ко всему-то впридачу, кроме позора-то, ненавистную жену ввести в дом! (Потому что ведь ты

меня ненавидишь, я это знаю!)

(Достоевский Ф.М. Идиот)

Здесь: fin = ⟨! + пробел + "П"⟩.

Итак, любая FIN -ситуация представима последовательностью трех более простых ситуаций F ("конец предложения"), I ("интервал") и N ("начало предложения"). Этот факт отражает таблица формализованного представления правил выявления ситуации "конец предложения" (табл. I), где по вертикали расположены F, I и N - ситуации, а по горизонтали - их возможные значения.

Таблица I

Формализованное представление правил выявления ситуации "конец предложения" в середине абзаца

F -ситуация (конец предложения)		Значения F -ситуации								
		Z			Z"			Z)		
I- ситуация (интервал) и ее значения	пробел	+	+	+	+	+	+	+	+	-
	тире	+	+	-	+	+	-	+	-	-
N -ситуация (начало предложения)		E	"E	(E	E	"E	(E	E	"E	(E
		Значения N -ситуации								

В табл. I зафиксировано всего восемнадцать возможных значений FIN -ситуации. Знаком "+" отмечены такие последовательности символов, которые, реально встречаясь в художественных текстах, позволяют с гарантией фиксировать в середине абзаца ситуацию "конец предложения". Знаком "-" отмечены те ситуации, которые в реальных текстах практически не встречаются; они, тем не менее, тоже являются решающими FIN -ситуациями.

Потенциальные возможности графических средств оформления конца - начала предложения оказываются, как мы видим, избыточными: только 70% этих возможностей используется в графике русского языка. В более общем

плане следует сказать, что пунктуационная избыточность графических средств русского языка, подтвержденная в настоящем исследовании при рассмотрении формальных способов структуризации художественной прозы, представляет собой еще один пример избыточности выразительных средств языка в целом.

Подведем некоторые итоги. Рассматривая варианты фиксации в тексте ситуации "конец предложения" и имея в своем распоряжении только изобразительные средства письменной речи, т.е. буквы алфавита и знаки препинания, мы выделили три возможных случая: ситуация "конец предложения" фиксируется перед заголовком (ZAG -ситуация), перед началом абзаца (АО-ситуация) и внутри абзаца (FIN -ситуация). Каждый из этих случаев сравнительно легко формализуем: надо позаботиться лишь о том, чтобы иметь доступ одновременно к двум смежным строчкам анализируемого текста; для ZAG- и АО-ситуаций - просто потому, что заголовки (и начала абзацев) всегда расположены ниже текущей (анализируемой) строки текста; для FIN -ситуаций - потому, что внутри абзаца текущее и новое предложения могут располагаться на разных строчках, а знание пунктуационно-символьной обстановки на границе двух предложений обязательно.

Важной представляется не только сама возможность формализации правил выявления конца предложений, но и надежность метода, т.е. полнота охвата (фиксации) реальных текстовых ситуаций.

Проверка возможностей формально-пунктуационного метода структуризации прозаических текстов проводилась на материале романа М.Ю.Лермонтова "Герой нашего времени". Структуризация всего произведения (150 страниц текста), выполненная в автоматическом режиме на ЭВМ ЕС-1035, дала ничтожный процент ошибок: из 3090 выделенных предложений лишь в двух

случаях зафиксирована излишняя детализация текста. В обоих случаях причиной явилось наличие знака препинания "точка" перед именем собственным:

Его звали... Григорьем Александровичем Печориным. (4)

... на кресте написано крупными буквами, что он поставлен по приказанию г. Ермолова, а именно в I824 году. (5)

Оба случая соответствуют одной и той же FIN -ситуации: <знак + пробел + прописная буква> . Эта ситуация для большинства предложений является решающей, что и отмечено в табл. I; в предложениях же (4) и (5) после разграничительного знака "точка" следует пробел и имя собственное, что формально (но здесь - ошибочно!) воспринимается как ситуация "конец предложения". Тем не менее, второй случай в принципе исправим: для этого в алгоритм структуризации следует ввести блок анализа стандартных сокращений, что несколько увеличит время работы программы, которая реализует этот алгоритм, но позволит избежать излишней ручной корректировки структурированного текста. Предложение (4) для своего формального выделения из текста требует не только дополнительного использования программы синтаксического анализа, но и специального решателя, указывающего, какие выделенные из текста и следующие друг за другом предложения составляют единое целое, а какие - нет. Затраты на проведение такого анализа (а его надо будет проводить для каждых двух смежных предложений текста!), а также нерешенность проблемы автоматического синтаксического анализа предложений сложной структуры лишний раз убеждают в правильности выбранного нами подхода к этому вопросу: исправление ошибок, допущенных при автоматической структуризации художественных текстов, должно производиться в режиме диалога "человек - ЭВМ" с помощью диалоговой системы корректировки структурированного текста.

Алгоритмическое ядро метода структуризации. Формально-пунктуационный метод, основу которого составляет алгоритм вычленения предложений, базируется на автоматическом поиске одной из трех решающих ситуаций: ситуации "конец предложения" перед заголовком (. ZAG -ситуации), ситуации "конец предложения" перед новым абзацем (АО-ситуации) и ситуации "конец предложения" в середине абзаца (FIN -ситуации).

Попытка отнести текущую ситуацию в строке к одной из этих трех решающих ситуаций производится последовательно: сначала делается попытка отнести ее к ZAG -ситуации (т.е. определить, не является ли очередная строка заголовком), затем - к АО-ситуации; если же обе эти попытки окажутся неудачными, то уточняется принадлежность текущей ситуации к одной из FIN -ситуаций.

Поиск ZAG -ситуации осуществляется просмотром двух строчек текста - текущей, в которой зафиксирован знак "Z" (либо ".", либо "?", либо "!"), и последующей (иногда - двух последующих) строки, смежной с текущей строкой. Выявление заголовков опирается на ряд особенностей их графического начертания и расположения печатного материала на страницах текста. Сюда, в первую очередь, следует отнести:

а) сдвиг начала заголовка относительно левого поля текста на величину, большую, чем абзацный отступ;

б) оформление начала заголовка прописной буквой;

в) отсутствие (как правило) в заголовке разграничительных знаков препинания;

г) оформление следующей за заголовком строки в виде строки с абзацным отступом;

д) уменьшенное (как правило) значение длины текущей строки по сравнению со средней длиной строки печатного текста (для данного издания) в том случае, если эта строка предшествует заголовку.

Совокупность таких особых примет, сопровождающих появление в тексте заголовков, вполне достаточна для

их формального распознавания в художественных текстах. Это, конечно, не означает, что разработанный нами алгоритм безошибочен при идентификации всех вариантов заголовков и подзаголовков в любых прозаических текстах; существуют, однако, два способа преодоления возможных ошибок:

а) корректировка структурированного текста в режиме диалога "человек - ЭВМ";

б) пополнение списка условий для формального выявления заголовков, такими, например, как наличие в строке только цифровых символов или наличие в строке только слова "Глава" с последующим набором цифровых символов.

Введение новых признаков для идентификации заголовков может быть сделано достаточно просто ввиду используемого нами модульного принципа построения всего программно-алгоритмического комплекса.

Поиск АО-ситуации (абзац) не представляет особой трудности, поскольку наличие в строке текста абзацного отступа является и необходимым, и достаточным для этого условием.

Поиск FIN -ситуации ("конец предложения" в середине абзаца) осуществляется согласно решающим правилам, представленным в табл. I: считывается очередная строка текста и, если в ней обнаруживается одна из F -ситуаций ("конец предложения"), то уточняется значение I-ситуации ("интервал" между словами), а затем производится поиск одного из возможных значений N -ситуации ("начало предложения"). Если для найденной последовательности символов, образующих (f + i) - ситуацию, не подтвердится ни одно из возможных значений N -ситуации, то FIN -ситуация не фиксируется; в противном случае производится формирование очередного выделяемого предложения и затем - несколько упрощая реальную работу алгоритма - оно записывается во внешнюю память ЭВМ. Оставшаяся часть текущей строки - справа от i-символа в FIN -ситуации - пере-

писывается в поле памяти, отведенное для хранения очередного предложения, и вновь подвергается анализу на наличие FIN -ситуации.

Таким образом, алгоритм формально-пунктуационного метода включает две циклически выполняемые операции: поиск решающей ситуации для каждой строки текста (внешний цикл) и - при наличии в данной строке хотя бы одной FIN -ситуации - поиск других возможных в этой строке FIN -ситуаций (внутренний цикл).

Программа, реализующая этот алгоритм, составлена и отлажена на языке ПЛ/I в среде ОС ЕС; ее объем - около 500 операторов языка ПЛ/I; среднее время на вычленение одного предложения - 2,4 секунды; структурирование полного объема романа М.Ю.Лермонтова "Герой нашего времени" (I50 страниц текста) заняло - с учетом работы программ маркировки абзацев и детализации видов прямой речи - 2,I часа работы ЭВМ ЕС-I035.

2. Детализация структурированного текста

Маркировка абзацев и предложений по типам письменной речи. Филолого-статистическому исследованию художественной прозы должно, на наш взгляд, предшествовать разграничение речевого материала на две категории - авторскую речь и чужую речь. Чужая речь включает прямую речь, диалоги, прямую речь внутри прямой речи и другие виды неавторского повествования. Имея в своем распоряжении только лишь формальные средства - знаки пунктуации и символы графической системы языка, мы предприняли попытку найти такие изобразительные средства, которые позволили бы автоматически (и одновременно с процессом структуризации) вести разграничение предложений по типам письменной речи. Представленный ниже метод автоматической маркировки структурных элементов художественной прозы - абзацев и предложений - базируется на строгом описании средств оформления прямой речи и ее пространственном расположении на страницах печатных изданий.

В алгоритме маркировки основным структурным элементом текста считается абзац: все предложения абзаца, последовательно вычленяемые из текста программой структуризации, подвергаются анализу на предмет наличия в них прямой речи; решение отнести тот или иной абзац к одному из типов письменной речи принимается по завершении анализа всех предложений данного абзаца. Сделано это с одной целью - разработать программу, обеспечивающую однородность совокупности лингвистических данных, подвергаемых статистической обработке, поскольку любые виды прямой речи, включенные в состав авторских абзацев, должны изменять маркер этих абзацев и исключать тем самым их участие в дальнейших исследованиях авторской речи. Такие абзацы могут впоследствии составить исходный материал для углубленного анализа стилистических приемов и особенностей преломления авторской манеры письма в речи персонажей (Кожевникова 1977).

Для маркировки абзацев и предложений будем пользоваться следующей системой обозначений:

%А% - авторский абзац (точнее - его первое предложение);

%П% - абзац прямой речи (его первое предложение);

%а% - предложение авторского абзаца (не первое);

%п% - предложение абзаца прямой речи (не первое).

Тогда появление элементов прямой речи в любом предложении авторского абзаца должно изменять тип этого абзаца (маркер первого предложения) с %А% на %П%, а каждое из последующих предложений, входящих в состав данного абзаца, должно метиться как %п%.

Пример I. Рассмотрим последовательность маркировки следующего абзаца:

> В этот вечер Казбич был угрюмее, чем когда-нибудь, а я заметил, что у него под бешметом надета кольчуга. "Недаром на нем эта кольчуга, - подумал я, ух он, верно, что-нибудь замышляет".
>
> (Лермонтов М.Ю. Герой нашего времени)

В первом предложении этого абзаца нет элементов прямой речи, следовательно, оно после вычленения из текста будет помечено маркером %А%. Второе предложение - предложение с прямой речью, поэтому его маркер получит значение %п%, а маркер первого предложения будет изменен на %П%.

Перейдем теперь к описанию средств графического оформления и пространственного расположения элементов прямой речи.

Как известно (см., например: Шапиро 1974), существуют два варианта построения текстов с прямой речью:

а) прямая речь оформляется в виде отдельных реплик участников разговора; в тексте такие реплики начинаются с новой строки;

б) прямая речь входит в состав абзаца авторского повествования; в этих случаях прямая речь обычно заключается в кавычки.

Построение алгоритма маркировки абзацев для первого варианта не вызывает особых затруднений: наличие абзацного отступа, за которым всегда следует тире, однозначно определяет всю ситуацию в целом. Первое предложение такого абзаца, как мы уже отмечали, метится маркером %П%, а все последующие - маркером %п%.

Здесь необходимо сделать небольшое уточнение, и мы об этом уже говорили: ситуация ⟨АО + тире⟩, зафиксированная в начале строки, требует дополнительной проверки на отсутствие двоеточия в конце предыдущей строки; поэтому при маркировке абзацев сначала проверяется наличие этих двух условий, а лишь затем первое предложение метится как %П%.

Уточним: ситуация ⟨АО + тире⟩ является необходимым и достаточным условием фиксации в тексте прямой речи, но потребность в детализации прямой речи по ее типам (см. ниже) приводит к необходимости детализировать и сам алгоритм маркировки. Поэтому, если в конце строки стоит двоеточие, а следующая строка начинается с ⟨АО + тире⟩, то это является свидетельством появле-

ния в тексте "смешанной" речи, т.е. прямой речи внутри абзаца авторской речи и метится как %С%.

Второй вариант построения текстов с прямой речью использует кавычки - ключевой элемент для фиксации и выделения прямой речи. И если бы те же кавычки не использовались для обозначений в тексте заимствований (названий, цитат, пословиц, поговорок и др.), то алгоритм маркировки был бы для этого варианта просто тривиален.

Пример 2.

> Петушков встрепенулся, - а солдат вытянулся, пожелал ему "здравья" и вручил ему большой конверт, запечатанный казенной печатью.
>
> (Тургенев И.С. Петушков)

Это предложение, конечно же, не должно помечаться как %П%, ибо в нем нет прямой речи; алгоритм маркировки должен отличать такие случаи и относить их к авторской речи.

Для того, чтобы реализовать это требование, можно было с самого начала двигаться в двух направлениях: либо проанализировать и выявить характерные (формальные, отличительные) признаки для предложений с заимствованиями, либо то же самое уточнить и формализовать для предложений с прямой речью. Первый способ, кажущийся вначале более предпочтительным, на самом деле алгоритмизации не поддается. Причина здесь одна - отсутствие каких-либо ограничений на использование в тексте цитат и прочих заимствований, что в отношении поиска отличительных пунктуационных особенностей делает задачу бесперспективной.

Поэтому выбран был второй путь - проанализированы и формализованы правила графического оформления прямой речи с использованием кавычек. Для того, чтобы описать эти правила в сжатом виде, обратимся к уже использованным ранее формализмам: пусть N -ситуация описывает последовательность символов, расположенных в начале

предложения, F-ситуация описывает последовательность символов, находящихся в конце предложения, а S-ситуация (это новое обозначение) - наличие определенных символов в середине предложения. Знаком Z по-прежнему будем обозначать любой из разграничительных знаков препинания Z =(".", "?", "!"), символом E - любую прописную букву, знаком "V" - логическую операцию "ИЛИ", а аббревиатурой "nil" - отсутствие кавычек в N- или F -ситуациях. Тогда правила фиксации прямой речи, оформленной с помощью кавычек, могут быть сведены к четырем случаям, представленным в Табл. 2.

Таблица 2

Формализованное представление правил выявления прямой речи

F-ситуация (конец предлож.)	⟨Z"⟩ V ⟨"Z⟩		nil	
N-ситуация (начало предлож.)	⟨"E⟩	nil	⟨"E⟩	
S-ситуация (середина предлож.)	⟨-...-⟩	⟨: "E⟩	⟨" , -⟩	⟨Z" -⟩
ВАРИАНТ	I	2	3	4

Вариант I реализует тот случай, когда прямая речь разрывается словами автора на две части. Здесь S-ситуация должна содержать два тире, разделенных некоторым текстом.

Пример 3 (вариант I, F = ⟨Z"⟩).

"Может быть, - подумал я, - ты оттого именно меня и любила: радости забываются, а печали никогда..."

(Лермонтов М.Ю. Герой нашего времени)

Следующий, второй вариант реализует тот случай, когда прямая речь находится в конце предложения и следует за словами автора.

Пример 4 (вариант 2, F = ⟨"Z⟩).

Староста сперва проворно соскочил с лошади, поклонился барину в пояс, промолвил: "Здравствуйте, батюшка Аркадий Павлович".

(Тургенев И.С. Бурмистр)

Варианты 3 и 4 реализуют случаи, когда прямая речь находится в начале предложения, а слова автора следуют за ней.

Пример 5 (вариант 4).

"Скорей, скорей в город за лекарем!" - кричал Владимир.

(Пушкин А.С. Дубровский)

Те же правила выявления прямой речи применимы и в более сложных случаях, а именно тогда, когда прямая речь состоит не из одного, а из нескольких предложений. При определении типа абзаца те предложения, которые расположены внутри кавычек, рассматриваются как одно целое, а знаки препинания ".", "?" и "!" в расчет не принимаются.

Пример 6 (вариант I).

"Может быть, мы никогда больше не увидимся, - сказал он мне. - Перед разлукой я хотел бы с вами объясниться".

(Пушкин А.С. Выстрел)

Здесь прямой речью объединены два предложения. Анализ текста, расположенного между кавычками, приводит к следующему результату: f = ⟨кавычки + "."⟩, n = ⟨кавычки + "М"⟩, S = ⟨тире + "сказал он мне" + тире⟩; точка после слова "мне" игнорируется. Поэтому факт наличия прямой речи будет зафиксирован согласно правила I (см. табл. 2).

Пример 7 (вариант 3).

"Наплюй на него, бабонька. Была бы шея, а ярмо будет", - с нескрываемым сожалением посоветовала одна.

(Шолохов М.А. Тихий Дон)

В этом примере f = nil, n = ⟨кавычки + "Н"⟩, а S = ⟨кавычки + "," + тире⟩, что соответствует

варианту 3 таблицы 2. Размеченный текст для данного примера будет состоять из двух предложений:

1. %П% "Наплюй на него, бабонька.

2. %п% Была бы шея, а ярмо будет", - с нескрываемым сожалением посоветовала одна.

Разработанные и представленные выше формальные правила выявления прямой речи охватывают три наиболее употребительных варианта построения предложений с вводящими словами автора, а именно, когда слова автора а) предшествуют прямой речи; б) следуют за прямой речью и в) включаются в прямую речь, разделяя ее на части.

Существует, как известно, еще один вариант, когда вводящие слова автора включают в себя прямую речь. Такой способ построения предложений с прямой речью используется относительно редко; тем не менее он тоже был подвергнут формализации и оказалось, что некоторых изменений в уже имеющихся правилах достаточно для обработки всех вариантов построения предложений с вводящими словами автора.

Пример 8.

> На вопрос мой: "Жив ли старый смотритель?" никто не мог дать мне удовлетворительного ответа.
>
> (Пушкин А.С. Станционный смотритель)

В начале этого предложения кавычки отсутствуют, т.е. N -ситуация не фиксируется (n = nil); в середине предложения распознается ситуация S = ⟨: "Ж⟩ , а ситуация ⟨ z"⟩ , которую мы (см. вариант 2 табл. 2) должны были бы искать и обнаружить в конце предложения, находится здесь в его середине.

В следующих двух примерах ситуация аналогична, но N -ситуация находится не в начале, а в середине предложения.

Пример 9.

Я только тогда выпрямился и подумал: "Зачем это отец ходит ночью по саду?" - когда опять все утихло вокруг.

(Тургенев И.С. Первая любовь)

Пример 10.

И только когда он шептал: "Мама! Мама!" - ему становилось как будто легче...

(Чехов А.П. Степь)

Этих примеров вполне достаточно, чтобы убедиться в следующем: в рассматриваемой нами конструкции вводящие слова автора, находящиеся левее прямой речи, отделяются от нее двоеточием. Следовательно, вычленение предложений с вводящими словами автора, которые включают в себя прямую речь, должно производиться при наличии следующей ситуации (вариант 5):

$(S_I = \langle :"E\rangle) \& ((S_2 = \langle ",\rangle) \vee (S_3 = \langle Z"\rangle))$,

где знаком "&" обозначена логическая операция "И", а ситуация S_I должна предшествовать появлению в тексте ситуации S_2 или ситуации S_3.

Таким образом, мы описали еще один, пятый по счету, вариант фиксации прямой речи, если последняя заключается в кавычки. Соответствующее этому варианту правило формализованного выявления прямой речи выглядит следующим образом:

$(F = nil) \& (N = nil) \& S_I \& (S_2 \vee S_3)$.

Представим теперь результаты проверки программы маркировки абзацев и предложений на материале романа М.Ю.Лермонтова "Герой нашего времени".

В тексте романа выделено 447 авторских абзацев и 723 абзаца прямой речи (каждая реплика диалога - это новый абзац прямой речи). В абзацах авторской речи зафиксировано 1235 предложений, а в абзацах прямой речи - 1674 предложения. Вариант 5 (слова автора включают в себя прямую речь) встретился в тексте только один раз, ошибок при выделении прямой речи не

обнаружено, предложения с заимствованиями (например, "оказия", "черта", "последняя туча рассеянной бури" и другие) обработаны программой маркировки правильно.

Эти результаты показывают, что разработанный нами алгоритм маркировки абзацев и предложений обладает достаточно высокой "лингвистической" надежностью. Критически подходя к оценке качества полученных результатов, следует еще раз отметить те исторически сложившиеся факторы (нечеткость пунктуационной системы языка, авторская вольность при расстановке знаков препинания и др.), которые не позволяют гарантировать абсолютную надежность формальных методов структуризации художественной прозы. Поэтому мы вновь подтверждаем наш тезис о том, что:

1) ошибки при формально-пунктуационном членении сплошного текста на абзацы и предложения, так же, как и при их маркировке, не могут быть полностью исключены;

2) устранение этих ошибок должно вестись исследователем-филологом в режиме диалога "человек - ЭВМ";

3) формальная структуризация художественных текстов и маркировка абзацев призваны и в решающей степени облегчают этот ручной, поистине титанический труд;

4) облегчение труда исследователя, достигаемое за счет применения разработанного и описанного здесь комплекса программ таково, что позволяет перевести задачу использования ЭВМ в филолого-статистических исследованиях авторской и неавторской манеры письма из области желаемого в разряд реальных практических действий.

Детализация видов прямой речи. Внимательное рассмотрение вопросов, связанных с появлением прямой речи в художественных текстах, подводит нас к необходимости продолжить ее дальнейшую конкретизацию. В самом деле, первый вид прямой речи - прямая речь внутри авторского абзаца - встречается в текстах художественной прозы весьма и весьма часто. Такое сочетание авторской и прямой речи внутри одного абзаца позволяет

оживить повествование, приблизить его к устной речи, а при необходимости - легко изменить направленность изложения (Кожевникова 1977).

Другой вид прямой речи - диалог персонажей, используя который автор напрямую сталкивает своих героев, вкладывает в их слова различные оттенки и отношения к происходящим событиям, друг к другу, к историческим и культурным традициям и т.д. И если первый вид прямой речи может быть в большинстве случаев заменен косвенной речью, что, скорее всего, приведет к потере образности и эмоциональной окраски повествования, но сохранит смысловой потенциал высказываний, то в диалогической речи подобная замена практически исключена.

Итак, мы различаем и, соответственно, по-разному маркируем:

а) собственно прямую речь, т.е. прямую речь без авторских слов, т.е. речь диалогическую; для нее маркерами являются %П% и %п%;

б) "смешанную" речь, т.е. прямую речь внутри авторских абзацев; все предложения таких абзацев метятся как %с%, а первое - как %С%;

в) "вложенную" прямую речь, т.е. прямую речь внутри абзаца собственно прямой речи, которая маркируется как %В% и %в%.

Все три вида прямой речи характеризуют, на наш взгляд, особые стилистические приемы, которыми в той или иной степени владеет каждый автор и, более того, которые должны изучаться каждый в отдельности и лишь затем - все вместе, в общей своей совокупности как интегрированный набор методов, приемов, средств авторской манеры письма.

В связи с этим алгоритм маркировки абзацев и предложений включает в себя отдельный блок, накапливающий сведения о типе предложений одного абзаца и, при необходимости, изменяющий их маркеры по следующей схеме:

а) если в абзаце, первоначально имевшем маркер %А%, встретится прямая речь, то все маркеры предложений этого абзаца заменяются на %с%, а для первого предложения - на %С%.

б) если в абзаце, имевшем маркер %П%, встретится прямая речь, то все маркеры должны измениться на %в% и %В%.

Пример II.

> Я отдал бабе корзинку. Двери заперли, но все еще слышались рыдания и крик: "Нина!" Я сел в пролетку и приказал извозчику ехать не спеша к Невскому.
>
> (Чехов А.П. Рассказ неизвестного человека)

Маркер первого предложения, вначале имевший значение %А%, будет - после вычленения из текста второго предложения - заменен на %С%, а все последующие предложения абзаца будут помечены маркером %с%.

Скажем теперь о результатах проверки алгоритма детализации видов прямой речи: в тексте романа "Герой нашего времени" выделено 619 абзацев собственно прямой речи (1153 предложений), 85 абзацев "смешанной" речи (422 предложения) и 19 абзацев вложенной прямой речи (99 предложений). Ошибок при работе алгоритма детализации не обнаружено; что же касается дальнейших статистических исследований структурированного и размеченного таким образом текста, то подобные исследования могут и должны строиться с учетом возможности детализации видов прямой речи или же без нее исходя из конкретных филологических задач.

Это, в частности, означает, что принятая нами схема детализации видов прямой речи не может считаться окончательной: уточнение видов речи и соответствующая этим видам маркировка абзацев и предложений будут определяться конкретными потребностями в проведении филологических исследований. Однако и сейчас диалоговая система корректировки текстов "DIAKOR" (см. ниже)

позволяет, при необходимости, вводить и использовать новые маркеры для абзацев и предложений структурированных текстов, например, выделять косвенную, полупрямую и несобственно-прямую речь.

Диалоговая система корректировки структурированного текста. Система "DIAKOR" разработана как автоматизированное рабочее место филолога для выполнения работ по корректировке структурированных прозаических текстов. Необходимость в разработке такой системы диктовалась тем обстоятельством, что для выполнения подобных филологических работ нужна простая, доступная любому пользователю-непрофессионалу компьютерная система; что же касается известных универсальных диалоговых систем "JEC", "PRIMUS", "ОКО" и др. (Зверев... 1986), то они не только сложны для непрофессионала, но и не обеспечивают выполнения ряда специальных операций, например, перенумерации предложений структурированного текста. Система "DIAKOR" реализована в среде ОС ЕС и предоставляет пользователю возможность работать в режимах чтения, записи, удаления и корректировки (включая перенумерацию) предложений, а также в режиме помощи "HELP" для получения пользователем основных навыков диалога с ЭВМ.

Предложения структурированного текста выступают в этой системе как квант информационного потока при обмене сообщениями между пользователем и ЭВМ. Организация такого обмена производится следующими основными командами:

1. RR - читать предложение на экран дисплея;
2. WW - писать предложение в память ЭВМ;
3. DD - удалить предложение;
4. RN - перенумеровать текст;
5. HH - перейти в режим "HELP".

Выбор автора повествования и настройка системы на определенный текст этого автора производятся пользователем в режиме "меню". Вначале на экране дисплея появляется "СПРАВКА ОБ АВТОРАХ", например:

```
*        СПРАВКА ОБ АВТОРАХ                    *
*            = 0 - конец работы                *
*            = I - Лермонтов М.Ю.              *
*            = 2 - Чехов А.П.                  *
*            = 3 - Толстой Л.Н.                *
* Внимание! Для выбора автора замените         *
*      = на * и нажмите на <ВВОД>              *
```

После выбора одного из авторов на экране появляется "СПРАВКА О ТЕКСТАХ", в которой указаны хранящиеся в ЭВМ структурированные тексты этого автора, например:

```
*   ЛЕРМОНТОВ М.Ю.                             *
*          СПРАВКА О ТЕКСТАХ                   *
*   = 0 - конец работы с текстами данного автора
*   = I - Герой нашего времени                 *
*   = 2 - Княгиня Лиговская                    *
*   = 3 - Панорама Москвы                      *
*      Внимание! Для выбора текста замените    *
*          = на * и нажмите на <ВВОД>          *
```

Далее система переходит в основной режим - режим чтения и корректировки структурированного текста.

Отличительной особенностью системы "DIAKOR" является возможность работы с записями переменной длины индексно-последовательных файлов. При структуризации текста каждое вычленяемое предложение получает свой номер (ключ) записи, а длина самой записи варьируется в зависимости от длины предложения и может достигать 6000 символов. Для работы с такими записями потребовалось разработать специальные многоэкранные средства связи и поддержки диалога. Наиболее сложным в системе "DIAKOR" является режим перенумерации предложений. Действительно, каждое предложение структурированного текста записывается в память ЭВМ под определенным номером (ключом), который соот-

ветствует порядковому номеру предложения от начала текста. Поэтому при удалении или вставке нового предложения вся последовательность номеров предложений (начиная с данного номера) должна быть изменена; это требует автоматического слияния двух файлов (основного файла и файла корректуры) с одновременной перенумерацией записей (предложений). Для пользователя, однако, этот режим ничем от других режимов не отличается и по его завершении пользователь получает сообщение о возможности продолжать работу.

В Приложении приведен отрывок романа М.Ю.Лермонтова "Герой нашего времени" (т.е. отрывок текста, введенного в ЭВМ в виде, идентичном оригиналу) и соответствующий ему фрагмент структурированного текста.

Выводы. В работе представлен формально-пунктуационный метод структуризации художественной прозы, образующий вместе с системой диалоговой корректировки текстов программно-алгоритмический комплекс подготовки и формирования текстовой части МФ РЯ. Работоспособность этого комплекса и его высокая "лингвистическая" надежность подтверждены при структуризации полного текста романа М.Ю.Лермонтова "Герой нашего времени". Положено только самое начало большой филолого-компьютерной работе над художественными текстами: и потому, что методы структуризации художественной прозы, а также их компьютерная реализация могут и будут продолжать совершенствоваться, и потому, что те же проблемы существуют для других видов текстов, и по многим другим причинам (см., например, Машинный фонд... 1986). Следующий шаг, очевидно, должен и будет состоять в детальной инвентаризации филологических потребностей (Герд 1986) для углубления начатых исследований в рамках МФ РЯ.

Отрывок из романа М.Ю.Лермонтова
"Герой нашего времени"

А. Фрагмент неструктурированного текста

- ж А что ж такое она пропела, не помните ли?

- ж Да, кажется, вот так: "ж Стройны, дескать, наши молодые джигиты, и кафтаны на них серебром выложены, а молодой русский офицер стройнее их, и галуны на нем золотые. ж Он как тополь между ними; только не расти, не цвести ему в нашем саду". ж Печорин встал, поклонился ей, приложил руку ко лбу и сердцу и просил меня отвечать ей; я хорошо знаю по-ихнему и перевел его ответ.

ж Когда она от нас отошла, тогда я шепнул ж Григорью ж Александровичу: "ж Ну что, какова?" - "ж Прелесть! - отвечал он. - ж А как ее зовут?" - "ж Ее зовут ж Бэлою", - отвечал я.

ж И точно, она была хороша: высокая, тоненькая, глаза черные, как у горной серны, так и заглядывали к вам в душу.

Б. Фрагмент структурированного текста

=I96= %П% - ж А что ж такое она пропела, не помните ли?

=I97= %В% - ж Да, кажется, вот так: "ж Стройны, дескать, наши молодые джигиты, и кафтаны на них серебром выложены, а молодой русский офицер стройнее их, и галуны на нем золотые.

=I98= %в% ж Он как тополь между ними; только не расти, не цвести ему в нашем саду".

=I99= %в% ж Печорин встал, поклонился ей, приложил руку ко лбу и сердцу и просил меня отвечать ей; я хорошо знаю по-ихнему и перевел его ответ.

=200= %С% ж Когда она от нас отошла, тогда я шепнул ж Григорью ж Александровичу: "ж Ну что, какова?"

=201= %c% - "ж Прелесть! - отвечал он.
=202= %c% - "ж А как ее зовут?"
=203= %c% - "ж Ее зовут ж Бэлою", - отвечал я.
=204= %A% ж И точно, она была хороша: высокая, тоненькая, глаза черные, как у горной серны, так и заглядывали к вам в душу.

Литература

Азарова И.Б., Горохова С.И., Григорьев Г.Г., Кузнецова Е.А. Разработка автоматических словоуказателей и конкордансов для художественных текстов // Структурная и прикладная лингвистика. Вып. 2. - Л.: Изд-во ЛГУ, 1983. - С. 187-190.

Андрющенко В.М. Машинный фонд русского языка: постановка задачи и практические шаги // ВЯ, 1985. - № 2. - С. 54-64.

Андрющенко В.М. Концепция и архитектура Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. - М.: Наука, 1986. - С. 26-43.

Бусленко В.Н. Автоматизация имитационного моделирования сложных систем. - М., 1977. - С. 240.

Герд А.С. Типы русских текстов и организация Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. - С. 67-74. (а)

Герд А.С. Русская морфология и Машинный фонд русского языка // ВЯ, 1986. - № 6. - С. 90-96. (б)

Гиндин С.И. Выступление в дискуссии // Машинный фонд русского языка: идеи и суждения. - С. 175-179.

Гринбаум О.Н., Мартыненко Г.Я., Фитиалов С.Я. Об одном эвристическом алгоритме структуризации художественных текстов // Автоматизированные системы переработки текстовой информации (АСПТИ). - Тезисы докл. - Львов: УНИИПИ, 1985. - С. 52-54.

Ершов А.П. Машинный фонд русского языка (Внешняя постановка вопроса) // ВЯ, 1985. - № 2. - С. 51-53.

Зверев В.И. и др. Алфавитно-цифровые дисплеи ЕС-7920 в диалоговых системах. - М.: Наука, 1986. - 240 с.

Караулов Ю.Н. Методология лингвистического исследования и Машинный фонд русского языка // Машинный фонд русского языка: идеи и суждения. - С. 13-25.

Кожевникова Н.А. О соотношении речи автора и персонажа // Языковые процессы современной русской художественной литературы. Проза. - М.: Наука, 1977. - С. 7-98.

Котов Р.Г. Машинный фонд русского языка и автоматизация словарно-терминологического обеспечения информационной службы // Машинный фонд русского языка: идеи и суждения. - С. 84-89.

Лесскис Г.А. О размере предложений в русской научной и художественной прозе 60-х годов XIX века // ВЯ, 1962. - № 2. - С. 78-95.

Мартыненко Г.Я. Многомерный синтактико-статистический анализ художественной прозы // Структурная и прикладная лингвистика. Вып. 2. - Л.: Изд-во ЛГУ, 1983. - С. 58-72.

Машинный фонд русского языка: идеи и суждения. - М.: Наука, 1986. - 240 с.

Пиотровский Р.Г. Психологические, теоретические, организационные и технические аспекты создания Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. - С. 146-148.

Рогожникова Р.П. Машинный фонд русского языка и словарное дело // ВЯ, 1985. - № 4. - С. 54-60.

Розенцвейг В.Ю. Опыт создания национальных лексикографических служб за рубежом // Машинный фонд русского языка. - С. 75-83.

Тулдава Ю.А. О квантитативных характеристиках богатства лексического состава художественных текстов // Уч. зап. Тарт. ун-та, вып. 437. LINGUSTICA. - Тарту, 1977. - С. 159-175.

Тулдава Ю.А. О машинных фондах русского языка за рубежом // Машинный фонд русского языка: идеи и суждения. - С. 141-142.

Шайкевич А.Я. Выступление в дискуссии // Машинный фонд русского языка: идеи и суждения. - С. 226-230.

Шапиро А.Б. Современный русский язык. Пунктуация. - М., 1974. - 287 с.

16-7018

М.А.Краснопёрова

ОБ АВТОМАТИЧЕСКОМ ИЗУЧЕНИИ РИТМИКИ РУССКОГО СТИХА (ОБЗОР)

Введение.

Под ритмикой мы понимаем то, что имеет отношение к ритмической структуре (РС) текста [1]. Со времен А.Белого основной способ изучения ритмики - это счет. Центральное место занимает счет по тексту. Подсчитываются частоты различных компонентов, из которых складывается РС. При анализе ритмики стиха наиболее распространенными единицами счета являются так называемые формы, варианты и ударения на сильных местах [2]. Кроме этих компонент подсчитываются многие другие. Часто их можно представить как совокупности или последовательности из форм и вариантов, см. далее, например, 2.3. Результаты подсчетов могут быть дополнены обработкой по формулам, выражающим совокупные характеристики ритмики. Например, показатель изменчивости ритма, отчетливость метра и т.д., см. 2.2.

В качестве вспомогательного средства используются теоретические расчеты. Примером может служить широко известная в стиховедении языковая модель размера (ЯМ), окончательный вариант которой был предложен академиком А.Н.Колмогоровым [I9]. В популярной интерпретации эта модель показывает, какими должны быть частоты форм,

---

[1] РС целого текста или его части определяется как последовательность ударных и безударных слогов, разделенная словоразделами.

[2] Сильные и слабые места устанавливаются скандированием строк. Если при этом не происходит переакцентуации слов, то те позиции, на которые падает скандирующее ударение, называются сильными местами (-), остальные - слабыми (Земли полуднем раскаленной, Не освежила ночи мгла). Формами называются конфигурации, различающиеся положением реальных ударений на сильных местах, вариантами - формы, разделенные словоразделами. (Примеры приводятся в Приложении).

вариантов и других ритмообразующих элементов, если поэт пишет стихи как прозаик прозу, отличаясь от него лишь стремлением соблюсти размер.

И, наконец, сравнение. Сравниваться могут текстовые показатели, например, для произведений разных авторов. эпох, для стиха и прозы и т.д. Текстовые данные могут сопоставляться с теоретическими. В качестве инструмента сравнения могут использоваться статистические критерии: χ^2 критерий нормального распределения, Стьюдента и др.

Обобщенную схему подсчетов при анализе ритмики стиха можно представить на рис. I.

Для получения надежных результатов требуется выполнить, как правило, весьма громоздкую работу. Это подводит к мысли об использовании ЭВМ. Использование ЭВМ облегчает то, что единицы счета легко формализуемы. Основными определителями являются обычно место и число ударных и безударных слогов и словоразделов [3].

Насколько нам известно, первая программа по автоматическому анализу ритмики русского стиха была сделана в 1969 году [7]. В 1970-х - 1980-х годах появился ряд других работ по этой тематике. Все эти работы можно расклассифицировать по их целям, а именно: I) получение справочного материала; 2) внутристиховедческий анализ; 3) изучение свойств ритмообразующего материала; 4) исследование механизмов порождения и восприятия ритма. Предлагаемое деление условно, результаты, получаемые исследователями одного из этих направлений, могут использоваться и для решения

---

[3] Вопрос о распознании единиц метра существенно усложняется, если ставить задачу автоматического определения ударений. С позиций ритмики решение этой задачи важно потому, что должно привести к единообразному акцентуированию. Дело в том, что так называемые двойственные слова (местоимения, числительные и др.) в одних случаях атонируются, в других оказываются более выделенными (Он был простой и честный малый; Он по-французски совершенно мог изъясняться; Но он измучился...) Ср. разрабатываемые А.Н.Колмогоровым и Прохоровым три системы записи ритма [20].

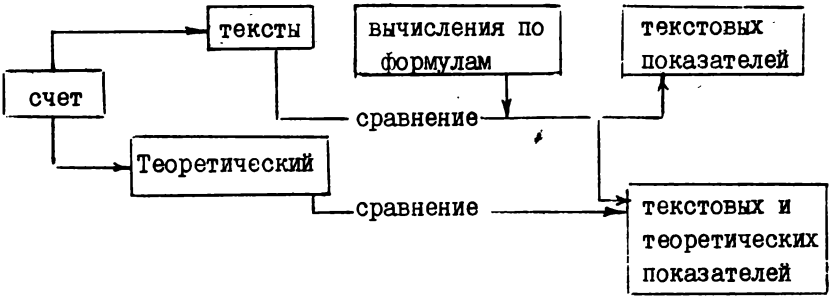

тексты
вычисления по формулам
текстовых показателей
счет
сравнение
Теоретический
сравнение
текстовых и теоретических показателей

Рис. I.

иных задач.

В машинном варианте на материале русского стиха задача уточнения атрибуции и датировки текста с помощью изучения ритмики не ставилась. Однако применявшиеся методы ее решения нетрудно адаптировать для использования на ЭВМ. Мы рассматриваем эту задачу как одну из реальных областей приложения действующих программ.

В настоящее время сделаны первые шаги по созданию языка, управляющего анализом ритмики, и его машинного интерпретатора. На этом языке должен указываться подлежащий учету ритмический объект. Эта работа связана с построением единой системы автоматического анализа ритмики. Хотя она и не доведена еще до машинной реализации, мы рассмотрим ее подробно, поскольку она является первым подступом к задачам такого рода. Перейдем теперь к более конкретному описанию сделанного.

I. Получение справочной информации. Программное обеспечение этих работ создано О.П.Исаковым. Задания на счет и тексты на машинных носителях обеспечиваются С.В.Калачевой. Конечной целью работы является создание ритмических справочников, где для каждого произведения будут указаны сведения о его размере и определенная информация, касающаяся ритмообразующих элементов и некоторых сопровождающих признаков.

На вход подается текст с расставленными ударениями и ограничителями конца строф. На выходе получают:

I) сведения о метрах и размерах произведения [4] в условной записи. Например, Я4 (четырехстопный ямб) или ПМК5 (полиметрическая композиция 5). В случае неоднородности размера приводится развернутая форма -

---

[4] Под метром здесь понимается родовое понятие: ямб, хорей и т.д., под размером - видовое: четырехстопный ямб (Я4), трехстопный ямб (Я3) и т.д.

названия метров в порядке их встречаемости и количества охватываемых строк (4ж8, Х*4, Хж4, Дж3 и т.д.). Дается информация о последовательности размеров, составляющих произведение, наполняющих один и тот же метр, о размере каждой отдельной строки, о соответствии между номерами строк и метрами. Приводятся числовые данные об общем количестве метров, размеров и внеметрических строк, о числе размеров каждого метра, о числе строк различной стопности и др.

2) сведения о строфической организации. Указывается число строф, характер клаузул [5] по строкам каждой строфы, количество метров в каждой строфе и др.

3) сведения о словоразделах и словах. Указываются положения словоразделов в каждой строке, число слогов и длина заударной части в каждом отделяемом ими слове, общее число словоразделов после каждого слога. Приводятся данные о количестве графических слов в каждой строке, их сумме и средней длине, о частотах фонетических слов [6] каждого типа и их средней длине. Кроме описанной информации даются построчные и общие сведения о количестве ударений, о числе спорных ударений, о количестве слогов, согласных и др. Для каждой строки указывается вид анакрузы [7], запись, соответствующая ритмической форме, часть предложения, к которой строка относится (начало, продолжение, конец).

---

[5] Клаузулой называются слабые слоги, находящиеся за последним сильным.

[6] Фонетическим словом считается акцентная группа, объединенная одним ударением. Оно может содержать более одного графического слова: он знал.

[7] Анакрузой называются слабые слоги, находящиеся перед первым сильным.

Выводимая информация не содержит традиционных сведений о частотах форм, вариантов, профилях ударности. Необходимые операции, однако, нетрудно подключить в имеющиеся программы.

Принципиальную трудность в этой работе представляет задача распознания метра. Имеющаяся публикация [5] отражает простейшую простую форму этой задачи - машинный анализ ямбических текстов. В ходе этого анализа ямб отличается от неямба и определяется конкретная разновидность ямбического метра: а) стопность, если произведение написано одним размером; б) повторяющийся период, если оно построено на регулярном чередовании стопностей; в) преобладающий размер, если имеется переходная метрическая форма (75% одного размера); г) состав и процентное соотношение размеров в случае вольного ямба.

Действующие сейчас программы анализируют и другие классические метры (хорей, дактиль, амфибрахий, анапест). Ведется работа по распознаванию неклассических метров.

2. Внутристиховедческий анализ. Исследования этого рода, выполняемые с помощью ЭВМ, можно условно разделить на три группы: I) выявление принципов ритмической организации текста, 2) типологический анализ, 3) эволюционный анализ.

2.I. Выявление принципов ритмической организации текста. Наряду с очевидными формами организации ритмики, такими, как подчиненность метру, разбиение на строфы, в произведении могут быть и неочевидные. Наиболее ранние работы по выявлению этих форм по весьма несложной методике были сделаны нами. Подсчитывались частоты акцентных групп, начинающих пары соседних строк. Составляющие члены этих групп по характеру РС могут быть а) одинаковыми

Не верь толпе пустой и лживой
Забудь сомнения свои

б) с одинаковой предударной и разной заударной частью

Цветущее | блаженство мая
Румяный | свет, златые дни

в) с разной предударной частью

И мне | наскучил их несвязный
И оглушающий | язык

Были выявлены закономерности распределения этих структур, см., например, [00].

Г.С.Басюточкин применил к стихотворному тексту спектральнй анализ стационарных временных рядов. Объектом исследования был верлибр "Александрийских песен" М.Кузьмина (стихотворение "Три раза я его видел лицом к лицу") и некоторые произведения других поэтов, не имеющие явной упорядоченности в слоговых длинах строк [4]. Стихотворный текст рассматривался как стационарный временной ряд, состояние которого в каждый момент характеризуется строчной длиной. Спрашивалось: можно ли разбить его на компоненты, имеющие сходную внутреннюю организацию по упорядоченности этого параметра?

Наиболее интересные результаты показало стихотворение с наименее выраженной ритмической организацией (Три раза...). Обнаружена тенденция к повторению сходных групп примерно по три, по пять и по девять строк. Выдвинута гипотеза о связи между внутренней упорядоченностью групп от семи до девяти строк и их содержательным построением.

Применение спектрального анализа требует большой вычислительной работы и может стать практическим орудием стиховеда только при использовании ЭВМ.

2.2. Типологический анализ. В.С.Баевский [I, 2] разработал методику изучения РС стихотворного текста на основании ее комплексных характеристик. В отличие от бинарного противопоставления слогов по ударности - безударности вводится многоступен-

чатая шкала выделенности. Выделенность слога определяется некоторым баллом в зависимости от положения в слове, в строке, в зависимости от типа самого слога и слова, которому он принадлежит. На основании этой градации вычисляются средние выделенности сильного и слабого слогов на определенном месте строки и безотносительно к нему. Далее рассматриваются следующие показатели: I) метрическая отчетливость ритма (m_t) - разность между средними выделенностями сильного и слабого слога; 2) ритмическая характеристика строки - среднее отклонение от средних выделенностей слогов (σ_j); 3) изменчивость ритма (σ) - среднее для всего произведения.

С этих позиций были проанализированы произведения советских поэтов 1956-65 гг., написанные четырех- и пятистопным ямбом и хореем. Для сравнения привлекались данные по русской поэзии XIX - начала XX-го века. Основная цель исследования состояла в классификации поэтов по указанным характеристикам их ритмики. Наблюдаются две ведущие тенденции: продолжение естественной эволюции и возвращение к ритмическим традициям прошлого. Классификация осложняется тем, что одни и те же поэты могут проявлять разные тенденции в отношении разных параметров и размеров.

Программа на ЭВМ строилась в форме диалога с исследователем. Исследователь набирал определенную последовательность программных модулей, обеспечивающих ввод, необходимую обработку текста и вывод результатов. При изучении отклонений от средних показателей (σ_j, σ) варьировался сопоставляемый фон: отдельное стихотворение, ансамбль стихотворений одного поэта, написанных данным размером в одно время, все стихотворения этого поэта, выдержанные в одном размере и т.д. Осуществить трудоемкие вычисления на большом текстовом материале (442 стихотво-

рения, II 282 строки) оказалось возможным только с помощью ЭВМ.

Естественно, что та же методика может быть использована при других целевых установках.

2.3. Эволюционный анализ на ЭВМ подобная работа была выполнена Б.Ф.Каюмовой [6, 7]. Изучалась эволюция ритмики Я4 в ХУШ - начале XIX вв. (с 1745 по 1820 гг.). До того ритмика Я4 уже была предметом рассмотрения исследователей стихотворной речи (А.Белый, К.Ф.Тарановский, М.Л.Гаспаров и др.). Основные результаты, касающиеся ритмической эволюции этого размера, получены К.Ф.Тарановским [21-23]. Необходимые подсчеты выполнялись всегда вручную.

На обширном материале (по ХУШ в. около 43, а по XIX в. - около 10,5 тыс. строк) В.Ф.Каюмова проанализировала соотношение реальных и теоретических частот различных ритмообразующих элементов на разных этапах изучаемого периода.

Наряду с традиционными параметрами, такими, как формы и профили ударности, рассматриваются параметры, не подвергавшиеся ранее последовательному эволюционному анализу: 1) варианты с предударным сильным слогом (Сапфиры, адаманты блещут ⏑ ⏓ ⏑ | – ⏑ ⏑ ⏑ | ⏑ ⏑) и варианты с заударным сильным слогом (Бесчисленность твоих красот ⏑ ⏑ ⏑ – | ⏑ ⏑ | ⏑ ⏑); 2) двустрочия определенного вида, составленные неполноударными формами (Колика ныне добродетель ⏑ ⏑ ⏑ ⏑ ⏑ – ⏑ ⏑ ⏑ Российский украшает трон ⏑ ⏑ ⏑ – ⏑ ⏑ ⏑ ⏑). Рассматривались также отношения между частотами разных вариантов, отношения между реальными и теоретическими частотами выделенных двустрочий. Все подсчеты выполнялись на материале, разбитом по пяти-, десяти- и пятнадцатилетиям. Для двустрочий различались пары, образованные нечетными-четными и четными-нечетными строками строфы, что вызвано представлением о

разнице в связанности их элементов. Для обеих совокупностей выполнялись раздельные и общие подсчеты. Все произведения условно разбивались на два жанра-оды и не-оды, по ним также рассматривались раздельные и общие показатели.

В работе исследовался вопрос о том, что определяло развитие ритмики Я4 в изучаемый период. Полученные результаты развивают представление об основополагающей роли сочетания языковых факторов и особенностей процесса восприятия и порождения РС, ср. [13]. Литературная специфика произведений выступает как фактор, модифицирующий глобальный процесс, но не меняющий его сути. Выясняется особая роль процессов становления и распада жанров: в период становления не-одических жанров (60-е - 70-е годы) ударность сильных слогов здесь повышена по сравнению со средними данными, в период распада од (90-е годы) некоторые ритмические показатели в них - вопреки характеру эволюции - обнаруживают тенденцию к возвращению своего прежнего уровня. Большой материал позволил подробно исследовать перебойный период в развитии ритмики (80-е и отчасти примыкающие годы). В этот период движение процесса регрессивной акцентной диссимиляции (процесс РАД) [8] было нарушено противоположными тенденциями. Выявилась своего рода спиралевидность развития: после перебоя эволюция снова идет в направлении процесса РАД, но исходными для нее являются показатели перебойного периода.

Частотная эволюция двустрочий поддерживает представление о наличии слабой зависимости между их членами (в соответствии с моделью восприятия и порождения РС, см. [4]). Это представление корректирует распространенный тезис об отсутствии статистической зависимости между ритмическими строками.

---

[8] Процесс РАД предполагает постепенное повышение частоты ударений на втором и понижение на первом и третьем сильном слоге [13].

Последовательная хронологическая детализация при рассмотрении всех параметров РС позволила уточнить характер их частотной эволюции, особенности взаимодействия и функциональной нагруженности. На основании полученных данных выдвигается гипотеза об имманентных свойствах некоторых ритмических элементов ("облегчающее" воздействие на ритм вариантов с предударным сильным слогом и отягчающее вариантов с заударным сильным слогом [9] и др.).

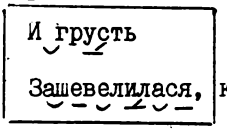
3. Изучение свойств ритмообразующего материала. Эти исследования выполнялись нами и под нашим руководством Н.А.Боголюбовой, Т.Н.Скурихиной и Т.Б.Шлюшенковой. Задача состояла в создании своего рода лингвистики ритма. С этой целью выделялись сравнительно простые элементы, могущие претендовать на роль определенных единиц ритменного текста. С помощью ЭВМ изучались некоторые устойчивые свойства этих элементов. Метод изучения - дистрибутивно-статистический. Фиксировались определенные признаки и подсчитываются частоты изучаемых единиц и этих признаков в одних и тех же условиях. Реальные частоты сравнивались с теми, которые имели бы место в случае независимости ритмообразующих единиц от выделенных признаков. На основании результатов сравнения и, возможно, определенной последующей обработки (например, сопоставления с показателями прозы) делаются выводы об изучаемых свойствах. Роль машины сводилась к нахождению отрезков текста, по которым должен вестись счет, распознанию типа ритмообразующих единиц и подсчету частот. Расчет теоретических показателей и сравнение с ними проводились вручную.

До сих пор единицами исследования были РС, начинающие пары соседних строк. Эти структуры двучленны: один член принадлежит первой строке пары, другой -

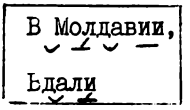
---

[9] Эти гипотезы расходятся с мнением М.Л.Гаспарова [16].

второй. В зависимости от отношения между членами выделяются два основных класса. Первый отличает менее обремененным вторым зачин. Во втором члене либо ударение находится дальше от начала, чем в первом, либо при одинаковом положении ударений заударная часть короче.

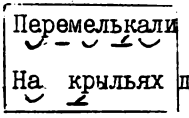


на дне старинной раны
как змей...

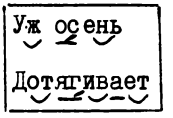


в глуши степей
Италии своей...

Второй характеризуется более обремененным вторым зачином. Такие РС характеризуются противоположным расположением членов.



наши взлеты
дружбы и вражды...



семьдесят восьмую
старый век...

и т.д. Рассматриваются также некоторые другие типы РС. В дальнейшем совокупность изучаемых единиц предполагается расширить за счет привлечения структур, образуемых срединными и конечными частями двустрочий.

С помощью машины изучались следующие свойства выделенных единиц.

I. Отношения между ними по степеням контрастности (степени различия составляющих членов). Опорный признак, характеризующий степень контрастности, - сила межстрочной паузы, разделяющей члены структуры. Использовалась градация по знакам препинания, предложенная Б.В.Томашевским: отсутствие знака - 0; пауза, соответствующая запятой, - I; пауза, соот-

ветствующая точке с запятой, - 2, точке - 3. Предполагалось, что пауза может смягчить или усилить контрастность. На основании частотных распределений определялась ее функциональная нагрузка, что служило поводом для выводов о контрастности [19].

2. Отношение РС к семантике наполняющей ее речевой конструкции. Выявлялась сама возможность тяготения РС к определенному смыслу [12]. Опорными признаками служили смыслы "внутрь" и "внутри", выраженные непосредственно этими словами или семантикой предлога "в". Сравнивались показатели стиха и прозы. Получены аргументы в пользу того, что РС с менее обремененным вторым зачином имеют предрасположенности к выделенным смыслам.

3. Отношение РС к коммуникативным типам предложения (повествовательное, вопросительное, восклицательное). Опорными признаками служили знаки точки, вопроса и восклицания, ближайшие справа к наблюдаемым РС. Сопоставлялись стихи и проза. Получены аргументы в пользу того, что РС с более обремененным вторым зачином имеют предрасположенность к вопросу и восклицанию.

Упомянем еще об одной работе, пока выполнявшейся вручную, предполагается, что впоследствии она может быть выполнена на ЭВМ. Работа посвящена изучению семантических отношений между выделенными РС (тяготению к сходной или различной семантике либо отсутствию выраженных тенденций). В этом исследовании используется семантический словарь Ю.Н.Караулова [26] и синонимический словарь З.Е.Александровой [25]. Важно подчеркнуть, что речь идет не об отношениях между членами РС, а об отношениях между парами двучленных РС. Рассматриваются всевозможные пары этих структур. Подсчитываются попадания слов-носителей их вторых членов в одни и те же статьи словарей [10].

[10] Предполагается, что недостатки семантического словаря не влияют существенным образом на распределения РС слов по статьям дескрипторов.

Полученные данные сопоставляются с теоретическими. То же делается для прозы. После этого выполняется сравнительный анализ результатов, на основании которого выносятся суждения о семантических отношениях. В работе используются данные по произведениям разных авторов и выявляются отношения, имеющие устойчивый характер. Полученные данные позволяют предполагать наличие относительно устойчивой системы семантических предрасположенностей выделенных ритмообразующих единиц.

В дальнейшем предполагается а) уточнить полученные результаты в ходе анализа нового материала, б) расширить корпус изучаемых единиц неначальными частями строк, в) изучить не только семантические отношения РС, но и конкретный характер их семантических предрасположенностей.

4. Исследование механизмов восприятия и порождения ритма. Модель восприятия и порождения РС стихотворного текста, предложенная в [I4], сама едва ли может быть полностью реализована на ЭВМ. Некоторые ее свойства, однако, могут быть испытаны с ее помощью. Из модели выносятся для машинной проверки две идеи: идея преемственности и идея слабой зависимости.

Идея преемственности по сравнению с тем, как она представлена в модели, взята в существенно упрощенном виде. Основной акцент делается на механизме слабой зависимости и его функции в эволюционных изменениях РС. Слабая зависимость вытекает из фундаментального свойства модели: принципа, согласно которому РС каждой строки при восприятии и порождении текста попадает на один и тот же участок памяти, так называемое стандартное место. Каждая следующая структура вступает во взаимодействие с тем, что было оставлено предыдущим. Это взаимодействие и обуславливает зависимость между ритмическими строками. Согласно модели это положение влияет на их характер в процессе порождения. Вероятность

одних строк увеличивается, других - уменьшается. Например, с позиций модели после строки с пропуском метрического ударения типа:

Капель до середины дня

уменьшается вероятность полноударной строки типа:

Потом морозом землю скомкав

и увеличивается вероятность неполноударной типа

Но заморозком землю скомкав

Степень изменения зависит от характера взаимодействующих РС. Она задается определенными формулами. Приведенные соображения и были положены в основу программы на ЭВМ [ ] . Назначение программы состояло в том, чтобы предсказывать частоты ритмических форм в разные периоды XIX-го века на основе данных о предшествующих периодах. В машину закладывались данные по XVШ-му веку, она предсказывала показатели по двум ближайшим десятилетиям XIX-го века. Затем закладывались данные по произведениям старшего поколения поэтов XIX-го века, она предсказывала аналогичные показатели для младшего. Полученные результаты сравнивались с известными статистическими данными К.Ф.Тарановского по текстам того же времени [21]. Реальные и рассчитанные частоты оказались весьма близкими [14].

Проведенный эксперимент имел еще и другое назначение. Исследователи стиха обнаружили, что в русской поэзии на протяжении XVШ-го - XIX-го вв. ритмика четырехстопного ямба эволюционировала от так называемого рамочного профиля ударности к альтернирующему [15, 21, 23]. Рамочный профиль характеризуется тем, что крайние сильные слоги ударяются чаще, чем средние. Графическое его изображение таково:



Наиболее выраженный вид он принимает в седьмой форме: И подданных не презирайте ⏑ ⏑́ ⏑ – ⏑ – ⏑ ⏑́ ⏑ . Альтернирующий тип характеризуется тем, что четные сильные слоги ударяются чаще, чем нечетные (⏑́ –). Крайнее выражение он принимает в шестой форме: Адмиралтейская игла ⏑ – ⏑ ⏑́ ⏑ – ⏑ ⏑́ . Было установлено, что переход от рамочного к альтернирующему типу шел в соответствии с процессом РАД (см. сноску на с. 251). Согласно данным К.Ф.Тарановского, установившийся альтернирующий тип во второй половине XIX-го века приобретает еще более выраженную форму. Возникает вопрос о причинах описанных явлений. Была выдвинута гипотеза, что основными причинами являются как раз те самые факторы, которым в модели соответствует механизм слабой зависимости и преемственности. При этом указывались те конкретные языковые явления, через которые зависимость получает наиболее выраженный характер. Это ударения слов, которые содержат безударные сильные слоги [13]. Полученные результаты говорят в пользу выдвинутых предположений. Проведенный эксперимент послужил еще одной цели. Возник вопрос, к чему придет ритмика изучаемого размера, если те внутренние факторы, которые в этом эксперименте задействованы, не подвергнутся никакому модифицирующему влиянию. Оказалось, что все прочие ритмические формы будут постепенно поглощаться той, которая идеально воплощает альтернирующий тип, шестой формой четырехстопного ямба: ⏑ – ⏑ ⏑́ ⏑ – ⏑ ⏑́ (адмиралтейская игла).

Но развитие стиха до этого не дошло. Альтернирующее движение было направлено вспять. В начале XX-го века наблюдается заметно выраженное противоположное движение [22]. Тем не менее альтернирующий тип никогда не уступил окончательно своих позиций. Изменения шли то в сторону его усугубления, то в сторону сглаживания, то в сторону весьма выраженного разде-

ления стихов на два вида - рамочный и альтернирующий. Но и в этом последнем случае, судя по средним данным, альтернирующий тип преобладал [16, 17].

Что подсказывает на этот счет модель восприятия и порождения РС? В модели учтены явления, модифицирующие спонтанные действия порождающего механизма. К ним относятся фильтры и стимуляторы ритмических эффектов. Первые уменьшают вероятность определенных ритмических эффектов, возникающих при взаимодействии строк, вторые - увеличивают. Эволюционное движение ритмики четырехстопного ямба в XVIII-м и большей части XIX-го века было направлено, в основном, в сторону увеличения гладкости стиха, шло от стихов типа: Но вышний зрак свой отвращает От Готфских ослепленных стран... к стихам типа: Земли, полуднем раскаленной, Не освежила ночи мгла... С позиций модели на тот внутренний спонтанно действующий процесс, который сам по себе способствует гладкости, в определенный момент начинают накладываться фильтры ритмических эффектов, сильно ее нарушающих. Сочетания типа Кочующие караваны В пространстве брошенных светил... благодаря этим фильтрам существенно ограничиваются. Эти фильтры, правда, в упрощенном виде, учтены в машинном эксперименте. Там подобные сочетания считаются запретными.

Если обратиться к началу двадцатого века, то в это время набирают силу противоположные тенденции. С позиций модели на спонтанно действующий процесс, ведущий к увеличению гладкости ритма, начинают накладываться стимуляторы ритмических эффектов, ее нарушающих. Это и вызывает пульсирование четырехстопного ямба вокруг альтернирующего рисунка и даже разграничение с ним.

Действие этих факторов, накладывающихся на спонтанно функционирующий механизм, в сочетании с вероятностными свойствами языка и определяет с позиций модели характер эволюции ритмики одного размера.

Что касается выполненного эксперимента, то он моделирует изучаемый процесс лишь в его срединной части: начиная от действия фильтров, если говорить терминами модели, и, кончая раньше, чем альтернирующее движение пошло вспять. Реализованная программа может стать основой дальнейших исследований. Можно менять исходный материал, уточнять формулы, определяющие вероятности строк, варьировать влияния фильтров и стимуляторов, добиваясь адекватности реальным фактам. Это будет способствовать более глубокому проникновению в подлинный механизм. Подобные исследования целесообразно строить в виде диалога машины и человека, подсказывающего ей, как и какие факторы следует ввести в действие.

5. К вопросу об атрибуции и датировке текста. Ритмика русского стиха не были еще предметом автоматического обследования, специальной целью которого было бы уточнение авторства и даты создания текста. Однако уже те материалы, которые получаются при справочном и эволюционном анализе могут принести существенную пользу в этих вопросах. Тот факт, что распределение ритмообразующих элементов связано со временем написания произведения, а отклонения характеризуют индивидуальные особенности, сам по себе является уже довольно мощным средством для достижения этих целей. Приведем два "немашинных" примера, подтверждающие высказанную мысль.

I) Две знаменитые оды М.В.Ломоносова "Вечернее размышление о Божием Величестве при случае великого северного сияния" и "Утреннее размышление о Божием Величестве" датируются обе 1743-м годом. Дата первого произведения известна доподлинно. Второе датируется тем же годом, так как составляет с ним как бы одно целое. В.М.Жирмунский обратил внимание на разницу в их ритмике [I8]. Первая ода имеет лишь четыре безударных сильных слога на 48 строк, вторая - 32 на 42 строки. Это заметно при чтении:

Лице свое скрывает день
Поля покрыла мрачна ночь
Взошла на горы мрачна тень
Лучи от нас склонились прочь
("Вечернее...").

Уже прекрасное светило
Простерло блеск свой
до земли
И Божии дела открыла.
Мой дух с вниманием внемли...
("Утреннее...")

Известно, что для раннего Ломоносова было характерно стремление использовать полноударные строки. Их число до 1745 г. превышало 75%. В 1745 г. наблюдается резкий перелом в ритмике Я4. Число полноударных строк в произведениях М.В.Ломоносова в среднем упало ниже 35%. Отсюда вывод: либо "Утреннее размышление..." было написано не раньше 1745 г., либо оно переделано зрелым Ломоносовым и дошло до нас только в переработанном виде. Сам Жирмунский указал на вторую возможность и подчеркнул еще разницу в строфическом оформлении произведений.

2) Весьма выразительны выводы Б.В.Томашевского, касающиеся продолжения "Русалки" А.С.Пушкина, предложенного Д.П.Зуевым. Он сравнил распределение ударений и словоразделов в пушкинском бесцезурном пятистопном ямбе с аналогичными показателями предполагаемого продолжения. В частотах словоразделов, выражающих собой довольно тонкие, трудно имитируемые свойства текста, были обнаружены заметные различия. Это позволило Томашевскому сделать вывод, что данное произведение Пушкину не принадлежит [24].

Можно привести и другие примеры использования частотных данных по ритмике текста для уточнения авторства и даты создания произведения. Естественно, что числовые показатели типа тех, по которым проводился анализ в приведенных примерах, могут быть получены с помощью ЭВМ.

6. О языке для описания РС. Если иметь в виду некую единую систему, могущую работать с разными ритмообразующими элементами, надо научить ее распоз-

навать те из них, которые интересуют исследователя в данный момент. С этой целью сделана попытка создать язык для описания РС и алгоритм, интерпретирующий его на ЭВМ [4]. Алгоритм записан на Algol'e-60, но не доведен еще до машинной реализации. Сам язык требует усовершенствований. Мы считаем возможным дать здесь информацию об этой работе, так как имеющийся опыт может быть использован в дальнейшем.

Объектами описания являются ритмические характеристики стихотворной строки. Например, формы, варианты, число слогов до первого сильного слога, число слогов за последним сильным слогом, количество ударений и т.д. Эти характеристики не задаются заранее, а могут быть сконструированы в выражении, записанном на созданном языке. Назначение интерпретирующего алгоритма состоит в том, чтобы, имея подобное выражение, распознать нужную строку. Он может работать на тексте, закодированном так, чтобы указывались ударения, словоразделы, сильные и слабые места. Областью его действий может быть, разумеется, не только текст, но и любая другая совокупность ритмических строк. Если, например, требуется вычислить теоретические вероятности каких-то РС и работает программа, порождающая всевозможные варианты данного размера [IO], он может выделить из них необходимые.

Предлагаемый язык использует в качестве основы язык узкого исчисления предикатов математической логики. Он состоит из трех уровней:

(I). Предметные постоянные и базовые предикаты. К предметным постоянным относятся сильные и слабые слоги с ударениями и без них, атонированные слоги и словоразделы (I, H, D, W, A, S соответственно). Все они распознаются непосредственно по своим кодам.

К базовым предикатам относятся предикаты принадлежности ($\in (x, y)$), непосредственного примыкания

[IO] Для классических размеров такая программа была составлена Г.С.Цейтиным.

(Q (x, y)), тождества (=(x, y)), равенства (=(x, c)). Каждому базовому предикату сопоставляются процедуры, поверяющие его истинность на ритмических строках.

Элементы этого уровня первичны в том смысле, что им соответствуют непосредственные действия с текстами. Все элементы других уровней в конечном итоге выражаются через эти, а также языковые средства узкого исчисления предикатов: кванторы, логические связки и скобки.

(2). Позиционные и количественные предикаты. Позиционные предикаты в соответствии со своим названием указывают положение одних компонентов строки относительно других. Имеется шесть позиционных предикатов: непосредственного вхождения между (Q(x, y, z)), следования после (Aft(y, x)), нахождения между (Bet(x, y, z)), сочленения (ж(z, x, y)), начальный маркер (Б(x)), конечный маркер (Е(x)). Каждому позиционному предикату ставится в соответствие выражение, которое, кроме символов, указанных в п. I, может содержать символ другого позиционного предиката. Например: $Bet(x, y, z) \leftrightarrow \exists u(\in(y, u) \wedge \Theta(x, u, z))$.

Среди количественных предикатов различаются предикаты числа и номера. Числовые делятся, в свою очередь, на предикаты однократного и многократного вхождения (I(x, y), M + I (x, y)).

$I(x, y) \leftrightarrow = (x, c) \wedge \in(x, y) \wedge \forall z(=(z, c) \wedge \in(z, y) \rightarrow \Xi(z, x))$

(c - предметная постоянная).

При описании многократного вхождения используется символ однократного, выражение для предиката номера содержит символ многократного вхождения.

(3). Термины. Приводимые предикаты соответствуют некоторые часто употребляемым терминам, определяющим ритмические компоненты строки: слог, (Syl(x)), анакруза (А(x)), клаузула (К(x)), ритмическое слово (v(x)), комплекс (С(x)), группа слогов от ударного до ударного включительно, междударный сегмент (Sm(x)), стихотворная строка (Str(x)). В их описании

используются предметные постоянные, базовые и позиционные предикаты. Например:

$A(x) \leftrightarrow \exists u(B(u) \wedge \exists v(=(v, I) \; v = (v, H) \wedge \Theta(u, x, v))) \wedge \neg \exists w(=(w, H) \vee = (w, I) \wedge \in (w, x))$, -

(объект, находящийся непосредственно между началом строки и сильным слогом, не содержащий сильных слогов).

Любое выражение, накладывающее условие на искомую строку, может быть записано с помощью символов описанного языка. Например: Пусть требуется найти строку, соответствующую шестой форме Я4 без дополнительных ударений (очарование мгновений $\cup - \cup \angle \cup - \cup \angle \cup$). Необходимое условие запишется так:

$\exists x(Str(x) \wedge \exists y(=(y, I) \; v = (y, H) \wedge Ч(y, x) \wedge \exists z(=(z, I) \wedge 2(z, x) \wedge \exists w(=(w, S) \wedge I(w, x) \wedge \exists v(A(v) \wedge =(v, c))))))$

Интерпретирующий алгоритм рассчитан на вычисление истинностного значения подобной формулы на конкретных РС. Определяемые им действия сводятся к поиску очередного базового предиката, нахождению его значения, вычислению с его помощью значения той или иной части формулы. И так до тех пор, пока не будет определена истинность или ложность всего выражения.

Дальнейшая работа, связанная с управляющим языком и его интерпретацией должна быть направлена на то, чтобы пользователь мог формулировать свое требование в естественной речевой форме. Машина должна будет переводить эту формулировку в выражение типа описанных и вычислять их значения.

Мы описали известные нам исследования, связанные с автоматическим изучением ритмики русского стиха. Цели и методы этих исследований могут быть задействованы в специальных программах, которые войдут в специальный подфонд Машинного фонда русского языка.

Литература

Баевский Б.С., Осипова Л.Я. Алгоритм и некоторые результаты статистического исследования альтернирующего ритма на ЭВМ "Минск-32" // Машинный перевод и прикладная лингвистика, вып. 17. - М., 1974. - С. 174-195.

Баевский Б.С., Осипова Л.Я. Исследование стихотворного ритма с использованием ЭВМ // Структурная и математическая лингвистика, т. 2. - Киев, 1974. - С. 11-19.

Васюточкин Г.С. О некоторых аспектах применения математики и ЭВМ в стиховедении // НТР и развитие художественного творчества. - Л., 1980. - С. 202-217.

Васюточкин Г.С., Красноперова М.А. О языке для описания ритмических структур в одной специализированной ИПС // Семантика естественных и искусственных языков в специализированных системах // Тез. докл. - Л., 1979. - С. 39-40.

Калачева С.В., Исаков О.П., Бирюлин А.С. О возможностях машинного исследования стиха // Вестник МГУ. - № 1, 1986. - Сер. 9. - Филология.

Каюмова В.Ф. Эволюция ритмики русского стиха в XVIII-м - начале XIX в. (на материале четырехстопного ямба: Автореф. дис. канд. филол. наук. - Л., 1984.

Каюмова В.Ф. Профили ударности и жанры в русской поэзии XVIII - начала XIX в. (на материале четырехстопного ямба) // Вестник ЛГУ, 1984. - Сер. истории, языка и литературы. - С. 77-89.

Красноперова М.А. Об использовании ЭВМ для изучения метра и ритма стихотворных текстов // Информационные процессы, эвристическое программирование, проблемы нейрокибернетики, моделирование автоматами, распознавание образов, проблемы семиотики (материалы У Всесоюзного симпозиума по кибернетике). - Тбилиси, 1970. - С. 346-347.

Красноперова М.А. Замечания о ритме строфы "Евгения Онегина" // Материалы XXУ научн. студенч. конф. Литературоведение, лингвистика. - Тарту, 1970. - С. 59-63.

Красноперова М.А. О программе для статистического обследования стихотворных текстов // Методы вычислений, вып. 9. - Л., 1974. - С. 169-178.

Красноперова М.А. Об отношениях между степенями контрастности ритмических структур // Лингвистические проблемы функционального моделирования речевой деятельности, вып. 3. - Л., 1976. - С. 145-158.

Красноперова М.А., Боголюбова Н.А. Испытание на связь между ритмическим эффектом и смыслом // Лингвистические проблемы функционального моделирования речевой деятельности, вып. 4. - Л., 1979. - С. 176-193.

Красноперова М.А. К вопросу о законе регрессивной акцентной диссимиляции и его причинах // **Russian Literature,XII, 1982. - North-Holland P.C. - N 2. - P.217-226.**

Красноперова М.А. Модель восприятия и порождения ритмической структуры стихотворного текста // Вопросы кибернетики. Общение с ЭВМ на естественном языке, вып. 80. - М., 1982. - С. 124-140.

Другая использованная литература

Белый А. Символизм. - М., 1910.

Гаспаров М.Л. Современный русский стих. - М., 1974.

Гаспаров М.Л. Очерк истории русского стиха. - М., 1984.

Жирмунский В.М. О национальных формах ямбического стиха // Теория стиха. - Л., 1968. - С. 7-23.

Колмогоров А.Н. Пример изучения метра и его ритмических вариаций // Теория стиха. - Л., 1968. - С. 145-167.

Колмогоров А.Н., Прохоров А.В. К основам русской классической метрики // Содружество наук и тайны творчества. - М., 1968. - С. 397-432.

Тарановски К. Руски дводелни ритмови. - Београд, 1953.

Тарановски К. Руски четворостопни амб у првим децени ама XX века // Јужнословенски филолог. - Т. 21 (1955-1956). - С. 15-44.

Тарановский К.Ф. О ритмической структуре русских двухсложных размеров // Поэтика и стилистика русской литературы. - Л., 1971. - С. 420-428.

Томашевский Б.В. О стихе. - Л., 1929.

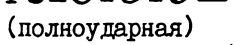
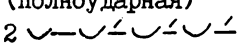
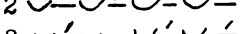
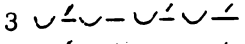
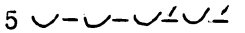
Словари

Александрова З.Е. Словарь синонимов русского языка. - М., 1968.

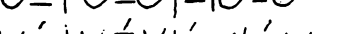
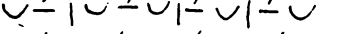
Караулов Ю.Н., Молчанов В.И., Афанасьев В.А., Михалев Н.В. Русский семантический словарь. - М., 1982.

Приложение

Список форм Я4:

№	схема	пример
I		Мороз и солнце: день чудесный
(полноударная)		
2		И понесется в мир свободы
3		Коснулся высоты заветной
4		И день грядущий настает
5		И не договорил опять
6		На православную дорогу
7		Нечаянностях впопыхах
8		И стереомагнитофон

Примеры вариантов

Мороз и солнце: день чудесный	
Открой сомкнуты негой взоры...	
На мутном небе мгла носилась	

О.П.Исаков

АВТОМАТИЗИРОВАННАЯ СИСТЕМА СТИХОВЕДЧЕСКОЙ ОБРАБОТКИ ПОЭТИЧЕСКИХ ТЕКСТОВ НА ЭВМ

I. Введение

Система автоматизированной обработки поэтических текстов "Ярославль" (СОПТ"ЯР"), краткое введение в которую дается в работе, должна определять размер произведения и слагающих его строк, выявлять метрические, ритмические и строфические характеристики стиха, вычислять самые различные статистические показатели по ряду произведений, строить метрические и ритмические справочники с помощью ЭВМ.

2. Структура системы обработки поэтических текстов.

Структура системы изображена на рис. I. Ее основные части: база поэтических текстов, система управления базами данных, комплекс программ обработки поэтических текстов "МЕТР 5", комплексы программ "СТРОФИКА", "РИФМА", фонематический транскриптор, комплексы программ "МЕТР. СПР.", "РИТМ.СПР", реализующие построение ритмических и метрических справочников, база данных результатов исследований.

Наиболее законченной частью системы является комплекс программ "МЕТР 5". Комплекс опробован на данных, подготовленных стиховедческой группой МГУ под руководством профессора Калачевой С.Б., обработано около 200 произведений А.Пушкина, В.Маяковского, А.Фета, Г.Державина, Ф.Тютчева, Н.Некрасова, что примерно 7000 силлабо-тонических строк. Алгоритмы разработаны Исаковым О.П. В программировании принимал участие Бирюлин А.С. Ниже будут изложены основные методы, лежащие в основе комплекса "МЕТР 5". Методики определения размера произведения и переходных метрических форм будут изложены в отдельной статье.

3. Модели стихотворных произведений.

Перечислим основные задачи, решение которых должен обеспечить комплекс программ "МЕТР 5": I) определение ритма и размера стихотворных строк; 2) определение размера произведения; 3) выявление метрических определителей (анакруза, клаузула, цезура); 4) выявление ритмических определителей (пропуск метрических ударений, отягчение и перестановка ударений, словоразделы). Классификация определителей заимствована у В.Жирмунского [5]. Кроме того, вычисляется ряд статистических показателей, характеризующих произведение в целом: средняя длина лексического и фонетического слова, мера предрасположенности произведения к идеальным метрам, степень ритмичности произведения и т.д. Некоторые из них будут описаны в данной статье.

В зависимости от решаемых задач строится 3 разных математических модели: модель-I, модель-2, модель-3. Каждая из моделей выполняет определенные функции. Например, модель-I используется для выявления метрических и ритмических определителей стиха, ритма и метра стихотворных строк. Эти показатели составляют основу для модели-2. В свою очередь, модель-2 используется для построения модели-3. При помощи последней, ЭВМ определяет размер произведения, выявляет переходные метрические формы. Будем различать математическую модель стихотворной строки и математическую модель произведения. Последняя помимо моделей стихотворных строк содержит характеристики всего произведения в целом, например, цезуру.

Все модели ЭВМ строит автоматически в своей памяти по определенному алгоритму. На рис. 2 приведена схема обработки стихотворных произведений.

4. Подготовка поэтических текстов на машинных носителях.

Подготовка данных на машинных носителях .
Она включает следующие основные операции.

I. Данные заносятся на перфокарты или магнитную ленту с фиксированным форматом записи.

2. Каждое произведение сопровождается "управляющей перфокартой", содержащей код языка, фамилию автора, название и дату написания произведения.

3. Конец произведения отмечается символом "777" на отдельной перфокарте.

4. Ударения в лексических словах отмечаются символами " ж "и" + ", которые ставят впереди ударяемой буквы. Символ "ж" означает сильное ударение, а символ "+" - спорное. Последний был предложен группой стиховедов МГУ.

5. Для правильной расстановки словоразделов, по усмотрению пользователя, группа энклитик выделяется с помощью знака "#", который ставится между словами, входящими в группу. Например, "зжа#день, пжо#полю" и т.д.

6. Конец строки отмечается пятью и более пробелами. Если на одной перфокарте набивается несколько стихотворных строк, конец строки отмечается символом

7. Если имеются строфы, их отмечают символом

Ниже дается пример стихотворения, подготовленного на машинном носителе:

I. ´А.С.Пушкин´,´Приют любви он... ´, ´I824´
Прижют любжви, он вжечно пжолн
Прохлжады сжумрачной и влжажной
Т+ам никогджа стеснженных вжолн
Не умолкжает гжул протжяжный
777

5. Построение модели I

Структура модели I стихотворной строки приведена на рис. 3. Объединение моделей I всех стихотворных строк даст модель I произведения. Рассмотрим составные части модели I и принципы их построения.

I. Число слогов соответствует числу гласных букв в стихотворной строке.

2. Число "ж" соответствует числу символов "ж" в строке.

3. Число "+" соответствует числу символов "+" в строке.

4. Массив М1 - массив переменной длины. Размер массива определяется числом слогов в стихотворной строке. Все лексические слова в стихотворной строке нумеруются по порядку слева-направо. Каждому слогу строки ЭВМ присваивает номер того слова, которому он принадлежит. Например, стихотворной строке "Журч*а еще беж*ит за м*ельницу руч*ей", содержащей пять слов, поставлен в соответствие следующий массив: (I,I,2,2,3,3, 4,5,5).

5. Массив М2. Массив переменной длины. Его размер определяется числом слов в строке. Элементами массива является число слогов в словах. Например, стихотворной строке "*Эти г*оры, *эти д*олы" будет поставлен в соответствие следующий массив: (2,2,2,2).

6. Массив связности слов строки. Массив переменной длины по числу слов в строке. Характеризует связь слов строки отношением энклитики. Словам, связанным отношением энклитики в этом массиве, соответствует символ "I", не связанным - "0". Например, стихотворной строке "И л*ёг и засн*ул он бл*из в*ерной осл*ицы" будет поставлен в соответствие следующий массив: (0,0,0,I,I,0,0,0).

7. Модель ударности слогов I. Строка, состоящая из символов "0" и "I", переменной длины по числу слогов в строке. Если слог сильный, т.е. помечен символом "ж", ему соответствует символ "I", если слабый или безударный - символ "0".

8. Модель ударности слогов 2. Используется для определения метра и размера стихотворной строки. Представляет собой последовательность символов "0", "I", "2", "9" переменной длины, по числу слогов в строке. Модель строится автоматически по специальной программе, реализующей алгоритм преобразования J текущего слога строки.

Преобразование J устанавливает различие между ударными слогами в зависимости от длины слов, которым принадлежат эти слоги. Преобразование J определено на множестве всех слогов и имеет множеством своих значений символы (0,I,2,9). Преобразование J описывается следующим образом:

I. Если w обозначает безударный слог, то J(w) = 0

2. Если w обозначает ударный слог, то:

2.I. J(w) = I, если слог принадлежит односложному слову.

2.2. J(w) =2, если слог принадлежит двусложному слову.

2.3. J(w) =9, если слог принадлежит более чем двусложному слову.

Например, стихотворной строке "Сег*одня пролетари*ат гр*ом голос*ов раск*уй" будет поставлена в соответствие следующая модель: "0 9 0 0 0 0 0 9 I 0 0 9 0 2".

6. Анализ модели I

Модель I произведения используется для определения метра и размера стихотворных строк, анакрузы, каталектики, словоразделов и других показателей.

6.I. Определение метра стихотворных строк.

Для определения метра стихотворных строк используется модель ударности слогов - 2 (см. п. 5). Если метр предполагается двусложным, модель условно разбивается на пары символов. Если в модели нечетное число символов, к последнему символу модели автоматически приписывается символ пробела "␣". Если метр предполагается трехсложным, модель ударности слогов - 2 разбивается на "тройки" символов. Если число символов не делится нацело на три, то для образования последней тройки символов к концу модели достраивается нужное число пробелов. Будем отличать пару (тройку) символов, в которой находится последний от начала строки символ "I", "2" или "9",

и называть ее последней значащей парой (тройкой). К ней предъявляются особые требования, т.к. последняя значащая пара (тройка) является константой стиха.

Определение метра и размера стихотворных строк с помощью ЭВМ заключается в последовательном просмотре и анализе всех пар (троек) модели ударности слогов - 2. Каждая текущая пара (тройка) символов проверяется на правила, ограничивающие употребление сверхсхемных и пропуск схемных ударений. При предлагаемом способе кодирования эти правила выглядят следующим образом.

6I.I Двусложные метры. Ямб.

(а). В текущей паре символов ху не допустимы комбинации 2у, 9у, у $\in$ (0,I,2,9).

(б). В последней значащей паре ху допустимы комбинации, у которых х $\in$ (0,I), у $\in$ (I,2,9). Все остальные комбинации недопустимы.

Хорей.

(а). В текущей паре символов ху не допустимы комбинации символов х2, х9, где х $\in$ (0,I,2,9).

(б). В последней значащей паре ху допустимы комбинации, у которых х $\in$ (I,2,9), у $\in$ (0, "␣"). Все остальные комбинации недопустимы.

Если просмотрены все пары из модели ударности слогов - 2 и нигде не нарушено правило (а), (б), строка считается ямбической (хореической) в зависимости от того, на какой метр она определялась. В случае хотя бы одного нарушения правила (а) или (б), строка считается не ямбической (не хореической). В ходе определения метра каждой двухсложной строки ЭВМ ведет учет ямбических (хореических) стоп, пиррихиев и спондеев, ямбов в хорее и хореев в ямбе.

Замечание I. Согласно правилу (а), стопа вида "I0" является допустимой в ямбе, а стопа вида "0I" - в хорее. Поэтому ЭВМ "не защищена" от ошибки, если встретит чисто хореическую строку, у которой модель ударности слогов 2 имеет вид "I 0 I 0 I 0 ...". Чтобы исключить подобные ситуации, вводится правило, общее для двусложных метров,

ограничивающее число употреблений хореев в ямбе и, наоборот, ямбов в хорее:

(в). Число стоп p противоположного метра ограничено длиной строки L , исчисляемой в стопах:

$p \leqslant 1$, если $L \leqslant 3$

$p \leqslant 2$, если $L \leqslant 5$

$p \leqslant 3$, если $L \geqslant 6$.

Если число стоп противоположного метра превысило допустимое число, согласно правилу (в), строка считается не ямбической (не хореической). Область действия правила (в) - строки, состоящие из односложных слов, которые практически не встречаются в 4-5 сложном ямбе (хорее). Поэтому, наиболее вероятно использование правила (в) в 2-3=стопном ямбе (хорее).

6.I.2. Трехсложные метры.

Поскольку у всех трех метров, процедуры распознавания аналогичны, опишем их в виде одной процедуры, оговаривая каждый метр особо.

(I). В текущей тройке символов **xyz** не допустимы следующие комбинации: (а) y = 9 или **z** = 9 - дактиль; (б) **x** = 9 или **z** = 9 - амфибрахий; (в) x = 9 или y = 9 - анапест.

(2). Если в текущей тройке символов **xyz** пропущено схемное ударение (x = 0, y = 0 или **z** = 0), то для последовательности символов **x'y'z'xyzx"y"z"** должны выполняться условия (закон переакцентуации): (а) если x = 0, то y' = 0, **z'** = 0, y = 0, **z** = 0 (дактиль); (б) если y = 0, то **z'** = 0, x = 0, **z** = 0, x" = 0 (амфибрахий); (в) если **z** = 0, то x = 0, y = 0, x" = 0, y" = 0 (анапест).

(3). В последней значащей тройке символов **xyz** должно выполняться условие: (а) $x \in \{1,2,9\}$ и **yz** $\in \{0,0,0\ _,\ _\ _\}$ - (дактиль); (б) $y \in \{1,2,9\}$ **x** = 0, **z** $\in \{0,\ _\}$ - (амфибрахий); (в) $z \in \{1,2,9\}$ x = 0, y = 0 - (анапест).

Если просмотрены все тройки символов и при этом ни в одной из них не нарушены правила (I) - (3), фиксируется,

что метр строки - дактиль, амфибрахий или анапест (в зависимости от того, какой метр предполагался в этой строке). При этом по каждой строке автоматически учитывается число трехсложных стоп, трибрахиев, амфимакров, антибакхиев, кретиков.

6.I.3 Описание ритма стихотворных строк.

После определения размера каждой стихотворной строки, строится таблица по типу таблицы I, где в первой колонке печатается номер строки, во второй - текст, а в третьей - ее размер. Символы "П", "Х", "С", "Я", заключенные в скобках, обозначают пиррихии, хорей, спондей, ямб, а цифра, стоящая рядом с ними, - номера стоп, в которых они встретились. Например, произведение А.С.Пушкина "Сатирик и поэт любовный" будет описано следующим образом.

С целью выявления двойственных строк (строк, принадлежащих сразу нескольким метрам) строится таблица по типу таблицы 2, где в I-й колонке печатается номер строки, во 2-й - соответствие метров и строк, а в третьей уточняется ритмический рисунок каждой строки. Символы "Я", "Х", "Д", "АМ", "АН", "НКЛ" обозначают ямб, хорей, дактиль, амфибрахий, анапест и неклассическую (не силлабо-тоническую) строку соответственно. Если строка принадлежит определенному метру, то на пересечении строки и столбца с данным метром ставится символ "I". Двойственные строки типа "В зав*етную *их цитад*ель" могут содержать несколько символов "I" в таблице соответствия метров и строк.

В результате обработки произведения А.С.Пушкина "Сатирик и поэт любовный" будет построена таблица 2 соответствия метров и строк.

Кроме того, по каждому произведению строится таблица распределения ямбов, хореев, пиррихиев и спондеев по номерам стоп. Например, по примеру, указанному выше, будет построена таблица 3.

Для трехсложных метров таблицы I,2,3 аналогичны таблицам двусложных метров с той разницей, что вместо спон-

дея дается бакхий, который обозначается символом "Б", а сами трехсложные метры обозначаются символами "Д" (дактиль), "М" (амфибрахий), "Н" (анапест).

6.2. Выявление метрических определителей.

К метрическим определителям относятся: анакруза, каталектика, словораздел. Для автоматического распознавания метрических определителей используется модель ударности слогов-I, массив номеров слов, которым принадлежат слоги, массив связности слов. После завершения обработки всего произведения, ЭВМ строит линейку анакруз, линейку каталектик, строит таблицу, содержащую подробную информацию о каждом словоразделе. Эти показатели являются составной частью модели-I всего произведения.

6.2.I. Анакрузы.

В модели ударности слогов-I слева направо подсчитывается число символов "О", расположенных левее ближайшего символа "I". Ставя в соответствие каждой строке числовой показатель ее анакрузы, получим линейку типов анакруз всего произведения. Например, обработав произведение Г.Державина "На смерть Катерины..." ЭВМ построит следующую линейку анакруз: 2 2 0 0 2 0 0 I 0 0 2 0 0 3 I 2 I I I I 0 0 0 0.

6.2.2. Каталектика.

В модели ударности слогов-I подсчитывается число символов "О", расположенных правее последнего символа "I" в строке. Если искомое число $N = 0$, то каталектика мужская и ее тип кодируют символом "М", при $N = I$ - женская "Ж", при $N = 2$ - дактилическая "Д", при $N = 3$ - гипердактилическая "Г", при $N = 4$ - пентоническая "П", если $N > 4$, тип каталектики кодируют символом "С". Ставя в соответствие каждой стихотворной строке произведения тип ее каталектики, получим линейку каталектик. После обработки произведения Г.Р.Державина "На смерть Катерины..." будет построена следующая линейка каталектик: Д Д Д Ж Д М Д М Ж М Ж М Д М Д М Д М Д М М М М.

6.2.3. Словораздел.

В звучащей речи лексические слова, не имеющие ударений, примыкают к словам, имеющим сильное ударение, образуя с ними фонетические слова [4]. Под словоразделом понимается граница между фонетическими словами в стихотворной строке [5]. Например, стихотворная строка "Когд*а / не в ш*утку / занем*ог" содержит 5 лексических и 3 фонетических слова. В большинстве случаев граница словораздела устанавливается последним слогом слова, имеющего сильное ударение. Например, "Кип*ит / вод*а, / рев*ёт / руч*ьём". Исключение составляют энклитики, например: п*од ноги, з*а нос, со дн*а н*а день, з*а день. Автоматическое распознавание энклитик с помощью ЭВМ - предмет будущих исследований. В настоящее время группу лексических слов, объединенных отношением энклитики, по своему усмотрению, выделяет сам исследователь. Для этой цели используется знак "#", который ставится между лексическими словами, связанными между собой отношением энклитики, например: з*а #ухо, со #дн*я #н*а #день. В случае отказа от употребления знака " # ", возможны ошибки при определении словоразделов. Покажем наиболее типичные из них. Положение "машинного" словораздела в данных примерах дается знаком " | ", правильное положение знаком " ¦ ".

"Вся ж*изнь | его ¦ как*ой-то | т*яжкий | с*он | "
А.С.Пушкин "Коварность", Я5 цезурированный.

"Вед*и | вед*и | мен+я ¦ под л*иповые | сен*и |
Да вн*овь | ув*ижу | я ¦ ковр*ы | густ*ых | луг*ов | "
А.С.Пушкин "Царское село", Я6 цезурированный.

"И л*ёг | и засн*ул | он ¦ близ в*ерной | осл*ицы | ".
А.С.Пушкин "Подражание корану" ("И путник усталый..."), АМ4 цезурированный.

Процесс автоматического распознавания словоразделов основан на следующих двух правилах: 1) если лексическое слово имеет сильное ударение и не связано с соседними словами отношением энклитики, граница словораздела определяется последним слогом этого слова; 2) если группа

лексических слов связана отношением энклитики, положение словораздела определяется последним слогом слова, замыкающего группу.

Тип словораздела определяется аналогично типу каталектики, с той разницей, что подсчет символов "О" в фонетическом слове ведется до ближайшей справа границы словораздела. Информация о словоразделах печатается в виде таблицы (см. таблицу 4), в которой по каждому словоразделу дается номер слога, которым заканчивается словораздел, его длина в слогах и тип словораздела. Например, обработав стихотворение А.С.Пушкина "Из письма Родзянке", ЭВМ построит таблицу 4.

Кроме того, ЭВМ строит таблицу распределения словораздельных форм по всему произведению. Например, обработав произведение А.С.Пушкина "К Сабурову", ЭВМ построит таблицу 5.

7. Анализ модели 2

Структура модели 2 стихотворной строки приведена на рис. 4. Модель 2 используется для определения различных показателей, характеризующих произведение в целом: средняя длина фонетического слова, распределение типов фонетических слов и силы словоразделов по частоте, цезура произведения и др.

7.I. Ритмическое богатство стиха.

Важной характеристикой, получаемой в ходе исследования модели 2, является распределение типов фонетических слов по частоте их встречаемости в произведении. Это распределение Б.Томашевский назвал "ритмическое богатство стиха" [3]. ЭВМ автоматически систематизирует все типы фонетических слов, по всему произведению. Результат представляется в виде таблицы, в I-ом столбце которой даются типы фонетических слов (О - означает безударный слог, I - ударный), во второй графе - частоты, а в третьей - процент от общего числа фонетических слов. Например, после обработки на ЭВМ стихотворения А.С.Пушкина "Второе послание Цензору", будет построена таблица 6.

7.2. Средняя длина лексического и фонетического слова.

Средняя длина лексического слова $L_{\text{ср. лек.}}$ определяется по формуле:

$$L_{\text{ср. лек.}} = \frac{\text{Число слогов}}{\text{Число лексических слов}}$$

Для этой цели используется модель I произведения, содержащая данные о числе слогов и лексических слов всех стихотворных строк.

Средняя длина фонетического слова $I_{\text{ср. фон.}}$ определяется по формуле:

$$L_{\text{ср. фон.}} = \frac{\text{Число слогов}}{\text{Число фонетических слов}}$$

Поскольку к лексических словам относятся предлоги, союзы и местоимения, как правило, выполняется неравенство: $L_{\text{ср. фон.}} > L_{\text{ср. лек.}}$ Например, после обработки стихотворения А.С.Пушкина "Андре Шенье", были получены следующие значения: $L_{\text{ср. фон.}} = 2.5$, $L_{\text{ср. лек.}} = 2.03$.

7.3. Определение цезуры.

Для визуального определения цезуры удобно использовать таблицу расположения словоразделов по позициям стихотворной строки, форма которой была предложена профессором МГУ Калачевой С.В. Например, в результате обработки стихотворения А.С.Пушкина "Из письма Плетневу" будет построена таблица 7.

Цезура понимается как словораздел, занимающий во всех стихах некоторого текста постоянную позицию [I]. Введем понятие силы словораздела P (j) на j-ом слоге стихотворного произведения. Сила словораздела определяется по формуле:

$$P(j) = \frac{N(j)}{L} \cdot 100, \quad j = 1, 2, 3 \ldots \quad (4)$$

где: N(j) - число словоразделов по всему произведению, заканчивающихся на j-ом слоге; L - число строк произведения.

Для автоматического распознавания цезуры с помощью ЭВМ определим как цезуру тот словораздел, чья сила подчинена неравенству (5):

$$80 \leqslant P(j) \leqslant 100 \qquad (5)$$

Левая граница в неравенстве (5) определялась эмпирически, и в дальнейшем будет определяться более точными методами. В таблице 8 собраны значения силы словоразделов на всех позициях стихотворных строк различных произведений. Из нее видно, что ряд произведений, представленных в таблице, имеет цезуру, кроме "Ласточки" Г.Державина. Обрабатываемые данные не содержали информацию об энклитиках, поэтому сила словораздела, задающего цезуру у некоторых произведений, меньше ста процентов. Ошибки при определении словоразделов возникали в тех случаях, когда за управляющим словом следовало подчиненное ему местоимение, союз или частица. Из таблицы 8 видно, что отклонение от реальной картины не велико и ЭВМ успешно справляется с определением цезуры.

Литература

Русское стихосложение XIX в. - М.: Наука, 1979.

Б.Томашевский. О стихе. - М.: Прибой, 1929.

Современный русский язык. - М.: Высшая школа, 1976.

Ь.Жирмунский. Теория стиха. - Л.: Советский писатель, 1975.

С.В.Калачева, О.П.Исаков, А.С.Бирюлин. О возможностях машинного исследования стиха. / Вестник Московского государственного университета. - Сер. 9. - Филология. - № 1, 1986.

19-7018

Таблица I

I Сат*ирик и по*эт лю*бовный,	Я4(П2)
2 Наш ариста*кип и асмод*ей,	Я4(П1,П3)
3 Т+ы не плем*янник *Анны Льв*овны,	Я4(П1)
4 Пок*ойной т*ётушки мо*ей.	Я4(П3)
5 Пис*атель н*ежный, т*онкий, *острый,	Я4
6 М+ой д*ядюшка - не д*ядя тв*ой,	Я4(П2)
7 Но, м*илый, - м*узы н*аши с*ёстры,	Я4
8 Ит*ак, ты вс*ё же бр*атец м*ой.	Я4

Линейка метров = Я Я Я Я Я Я Я Я
Линейка размеров = 4 4 4 4 4 4 4 4

Таблица 2

Таблица соответствия метров и строк

	Я	Х	Д	АМ	АН	НКЛ	
I	I	0	0	0	0	0	0
2	I	0	0	0	0	0	0
3	I	0	0	0	0	0	0
4	I	0	0	0	0	0	0
5	I	0	0	0	0	0	0
6	I	0	0	0	0	0	0
7	I	0	0	0	0	0	0
8	I	0	0	0	0	0	0

Ритмический рисунок строки

Я	П	Я	Я	+
П	Я	П	Я	
П	Я	Я	Я	+
Я	Я	П	Я	
Я	Я	Я	Я	+
Я	П	Я	Я	
Я	Я	Я	Я	+
Я	Я	Я	Я	

Таблица 3 .

Распределение видов стоп по их номерам

Стопа	Я	Х	П	С
I	6	0	2	0
2	6	0	2	0
3	6	0	2	0
4	8	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
IO	0	0	0	0
Всего	25	I	5	I

Таблица 4

Информация о словоразделах

Виды словоразделов		Информация о каждом словоразделе		
I	МДМ	2 2 М	6 4 Д	8 2 М
2	ЖЖЖ	3 3 Ж	5 2 Ж	9 4 Ж
3	МГЖ	2 2 М	7 5 Г	9 2 Ж
4	ЖММ	3 3 Ж	6 3 М	8 2 М
5	МЖЖ	2 2 М	5 3 Ж	9 4 Ж
6	МММ	4 4 М	6 2 М	8 2 М

Таблица 5

Ритмическая форма - I	Частота
I	2
ЖМДЖ	I
ДММЖЖ	I
МДМ	I
ЖМ	I
МЖМЖ	I
ЖММ	I

В графе I дается тип ритмической формы, а в графе 2 - ее частота.

Таблица 6

Таблица количества фонетических слов

Бид фонетического слова	Кол-во	%
I	9	3.27
IO	I0	3.64
OI	47	I7.09
IOO	I5	5.45
OIO	74	26.9I
OOI	30	I0.9I
OIOO	2I	7.64
OOIO	I9	6.9I
OOOI	I4	5.09
OIOOO	3	I.09
OOIOO	2	0.73
OOOIO	I2	4.36
OOOIOO	II	4.00
OOOOIO	2	0.73
OOOOOI	4	I.45
OOOOOIO	2	0.73
Итого	275	I00

Таблица 7

Строки	Позиции словоразделов по номерам слогов								
I			3		5		8		
2		2				6		9	
3			3		5		8		
4						6		9	
5						6	8		
6				4		6		9	
7		2		4		6	8		
8					5		8		
9		2				6		9	
Всего	0	3	2	2	3	6	0	5	4

Таблица 8

№	Автор и название произведения	Размер произведения	Р(1)	Р(2)	Р(3)	Р(4)	Р(5)	Р(6)	Р(7)	Р(8)	Р(9)	Р(10)	Р(11)	Р(12)	Р(13)	Р(14)	Р(15)
1	А.Пушкин "Второе послание к цензору"	Я6	1.38	16.6	45.8	18.05	12.5	83.3	∅	8.33	52.7	36.11	8.3	50	50		
2	А.Пушкин "Желание славы"	Я6	6.66	30	46.6	23.33	10	93.3	3.33	6.66	13	36.66	13.33	50	50		
3	А.Пушкин "Подражение корану" (и путник усталый...)	М4	∅	27.77	50	19.4	2.77	94.4	2.77	33.33	47.2	19.44	66.66	33.33			
4	Г.Державин "Ласточка"	ПМФ	4	16	52	24	30	54	16	50	50						
5	А.Пушкин "Издавна мудрые искали...	Я4	∅	12.5	81.25	12.5	37.5	43.75	12.5	50	50						
6	А.Пушкин "Желание"	Я5	∅	8.3	33.3	100	∅	8.3	50	50	16.6	50	50				

Таблица 8. Продолжение

7	Н.Некрасов "Я зато глубоко презираю себя..."	Н4	Ø	Ø	50	43.7	Ø	93.7	Ø	Ø	25	62.5	I2.5	I00			
8	А.Пушкин "Я был свидетелем..."	Я6	Ø	8.3	4I.6	25	I6.6	I00	Ø	33.3	50	33.3	Ø	50	50		
9	А.Пушкин "Царское село"	Я6	Ø	27.3	50	27.3	9	9I	4.5	I3.6	4I	4I	22.7	50	50		
I0	А.Пушкин "Коварность"	Я5	7.I	I4	3.5	96.4	Ø	2I.4	42.8	39.2	I7.8	50	50				

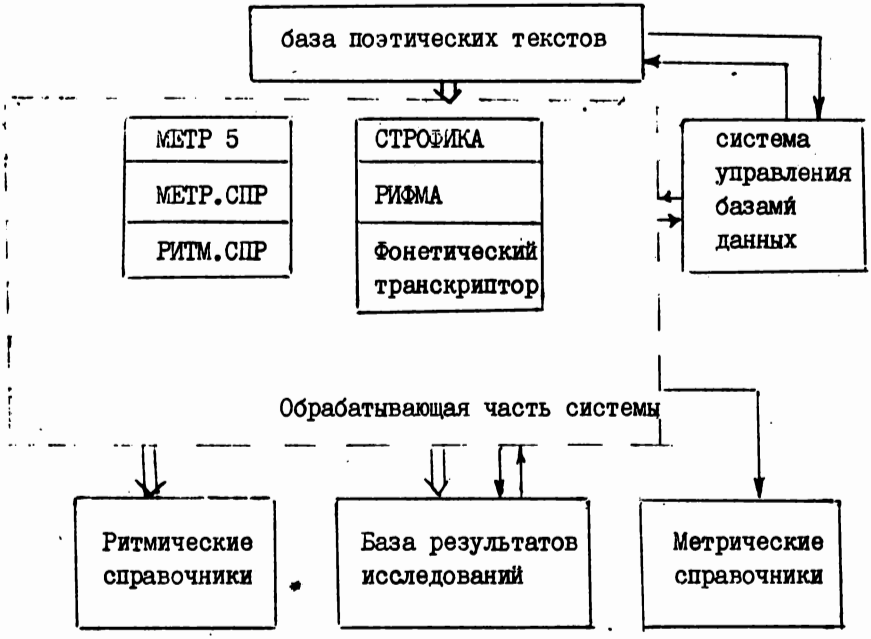


Рис. I. Структура системы обработки поэтических текстов.

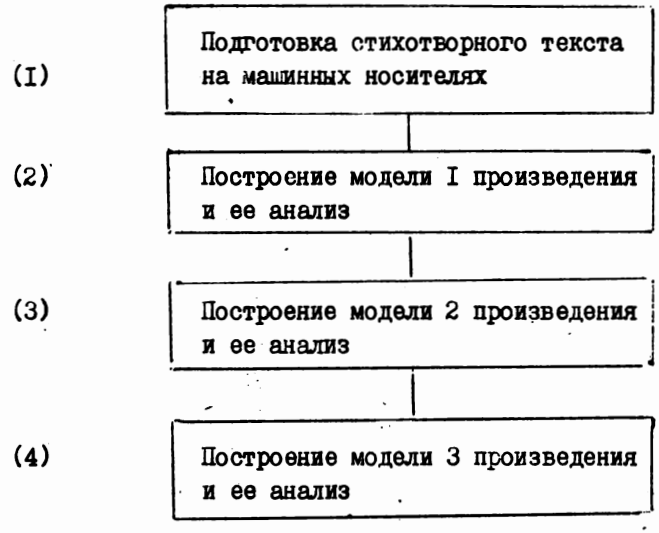


Рис. 2. Схема обработки стихотворных произведений на ЭВМ.

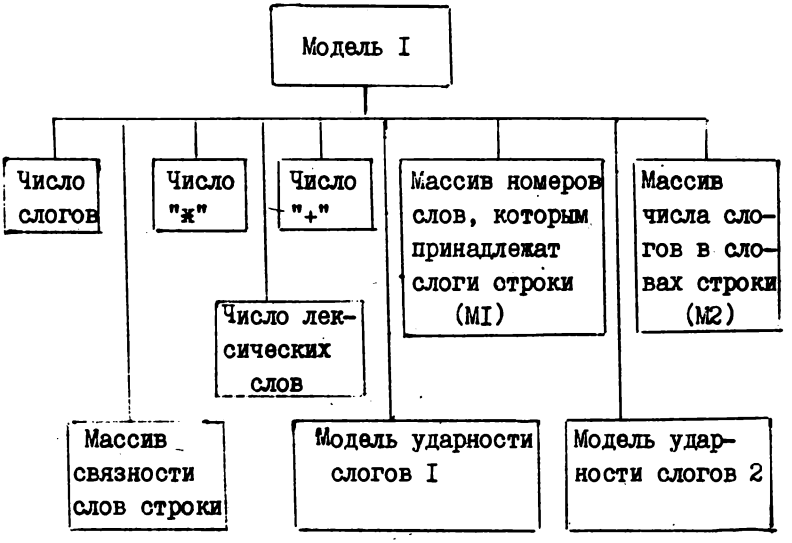


Рис. 3. Структура модели I стихотворной строки.

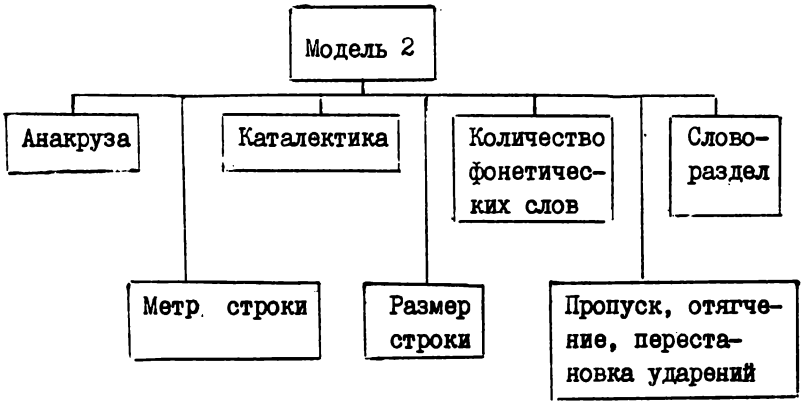


Рис. 4. Структура модели 2 стихотворной строки.

ИПРЯ
Институт русского языка

Оглавление

В.М.Андрющенко, В.В.Морковкин. Базовое лексикографическое знание и компьютер 4
Л.И.Колодяжная. Автоматический Генеральный словник русских словарей (методика создания, предварительные исследования, первые результаты). 41
О.В.Загоровская, С.В.Лесников. Виды лексикографической информации в автоматическом словаре русских говоров Коми АССР и сопредельных областей. 64
О.А.Казакевич. Текстовые базы данных устной речи: создание и использование. 71
И.А.Большаков. Эффективное целобайтовое кодирование текстов. 95
Г.С.Цейтин. На пути к сборочному программированию. . . . 120
М.М.Железняков, Т.Н.Невлева, И.М.Новицкая, Л.Н.Смирнова, Г.С.Цейтин. Опыт построения модели типа "текст - действительность" с использованием ассоциативных сетей. 140
М.Г.Мальковский, И.А.Волкова, Т.Ю.Грацианова. Взаимодействие с ЭВМ на естественном языке: проблема речевых ошибок. 168
Ю.С.Мартемьянов, А.Н.Морозова. Свойства синтезирующего описания в применении к анализу текста. . . . 192
О.Н.Гринбаум. Автоматическая структуризация текстов (на материале русской классической прозы). 209
М.А.Красноперова. Об автоматическом изучении ритмики русского стиха (обзор). 242
О.П.Исаков. Автоматизированная система стиховедческой обработки поэтических текстов на ЭВМ. 268



МПО "1-я Образцовая тип. им.А.А.Жданова". Зак. 7018. Тир. 300.