ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 4 1986

МАТЕРИАЛЫ И СООБЩЕНИЯ

АСИНОВСКИЙ А. С., КУЗНЕЦОВА Е. Ж., ЛЮБЛИНСКАЯ М. Д., ЧЕРНЫШЕВА Л. В., ШЕВЧЕНКО Т. В.

К ВОПРОСУ ОБ АВТОМАТИЗАЦИИ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Создание автоматизированных информационных систем, т. е. систем, предназначенных для накопления, хранения, обработки и выдачи разнообразной текстовой информации, связано с кругом проблем технического, программного, математического и лингвистического обеспечения [1, с. 5]. Этим обусловлена необходимость союза лингвистов с программистами и математиками, союза, в котором заинтересованы обе стороны. Принципиальная возможность дальнейшего развития вычислительной техники и, прежде всего, «общения» с нею предполагает создание диалоговых систем, построенных на основе лингвистической информации. Создание ЭВМ с речевым вводом, организация передачи и хранения информации в ЭВМ по принципам естественного языка — это задачи, которые не могут быть решены специалистами по прикладной математике и вычислительной технике без помощи лингвиста. С другой стороны, идея применения ЭВМ для анализа текста привела к разработке методов математической и прикладной лингвистики, основанных на формализации задач и материала. За 30 лет развития вычислительной техники конструкция и возможности машин сильно изменились. Но только когда возникла возможность обработки литерных строк, т. е. ввод и вывод цепочек литер, а не цифр, ЭВМ стала доступна сравнительно широкому кругу лингвистов.

Существует представление, что если лингвист не занимается проблемами диалога с ЭВМ и не пользуется методами математической лингвистики, то вряд ли целесообразно, если вообще возможно, обращение к ЭВМ в его работе. Конечно, проблемы автоматического распознавания слуховых образов, автоматического перевода и реферирования, поиска информации не могут решаться без машины. Но ЭВМ может оказать большую помощь и при «традиционном» исследовании языка. При этом не существует принципиальных ограничений, связанных с методами исследования, языковыми уровнями, типами языков и степенью их изученности.

Само по себе обращение к ЭВМ не может дать решения ни одного лингвистического вопроса. Наряду с магнитофоном, спектрографом и т. п. ЭВМ — инструмент, предоставляющий огромные возможности для работы с материалом. С ее помощью исследование может быть автоматизировано. Безусловно, «ручная работа» при лингвистическом исследовании необходима и может дать очень много. Но часто бывают нужны сведения, получение которых требует неоправданно больших затрат времени. Например, лингвисту необходим словник, представленный в обратном алфавитном порядке. Для русиста такой проблемы нет — создан «Грамматический словарь русского языка» [2]. Но исследователям других, менее изученных языков, например палеоазиатских, приходится самим вручную создавать обратные и другие словари. Вот тут на помощь могла бы прийти ЭВМ, обладающая большим объемом памяти и скоростью выполнения операций: для обработки массива в 100 тыс. слов ей требуются не часы, а несколько минут. ЭВМ целесообразно применять для легко формализуемой обработки больших массивов информации — для создания различных справочников и указателей по текстам, обратных, частотных, словообразовательных, морфемных и других словарей. В этих случаях память машины заменяет картотеку.

Для решения многих задач необходима сортировка различных единиц накопленного языкового материала по определенным параметрам. В этом случае элементы различных уровней языковой структуры рассматриваются лишь с точки зрения их сходства или различия. Так, проведение исследований на графемно-фонемном уровне дает ценный материал для типологических исследований [3]; на морфемном уровне проводится статистико-дистрибутивное описание морфем [4]. Большие возможности дает ЭВМ лексикографам. На уровне слов и словосочетаний ведутся работы по составлению частотных словарей [6], словоуказателей и конкордансов, наиболее фундаментальные из которых описаны в обзорах [7, 8], причем словоуказатели могут быть получены традиционным образом на карточках [9].

Для того чтобы лингвист использовал ЭВМ как средство автоматизации, необходимо создавать банки лингвистических данных. Банк данных (БД) должен содержать языковой материал — тексты, словари, грамматики и алгоритмы их обработки. Эта информация записывается на перфокартах, перфолентах, магнитных лентах и дисках, т. е. внешних носителях машинной памяти.

Решение вопроса, на основе какого материала должен строиться БД, иногда сводится к выбору между словарем и текстом. Альтернативный подход в этом случае малоплодотворен. Для многих задач хранение словаря в памяти машины является предпосылкой автоматической обработки текстов, хотя лексикографу, безусловно, прежде всего важна работа с текстами, а фонолог и морфолог заинтересованы в анализе словарей.

Вводимая информация должна иметь по возможности универсальный характер, т. е. обеспечивать исследование в области звукового строя, морфемики, словообразования, грамматической и лексической семантики. Обеспечить это не просто, поскольку подготовка информации и запись ее в ЭВМ — весьма трудоемкое дело. Но только легко дополняемый и корректируемый БД, доступный многим исследователям для решения самых разных типов задач, обеспечит осмысленное применение ЭВМ, сохранение сил и времени лингвиста.

Форма ввода информации определяется целями лингвистического исследования. Данные могут быть представлены в виде списков элементов, упорядоченных по заданным параметрам, таблиц, графиков, деревьев и т. д. Результаты проведенной обработки материала могут в свою очередь добавляться в БД или служить основанием коррекции уже записанного материала. Полученные результаты могут быть не сразу выведены на печатающее устройство, а извлекаться из памяти машины порциями, по мере необходимости, что делает возможной их поэтапную обработку.

Языковой материал в БД организуется различными способами. При этом какая-либо одна из форм представления материала обычно считается основной, а все остальные получаются с помощью сортировок и других операций. Одной из таких форм организации материала может служить базовый морфемный словарь, т. е. словарь исследуемого языка, представляющий каждый элемент в виде последовательности морфем. Элементами в таком словаре являются слова и словосочетания. На этой основе исследователь может получать списки слов, разделенных на морфемы, гнездовой словарь, набор словообразовательных формантов, набор аффиксальных морфем, идиоматические выражения с заданными словами или структурой и т. д. — материал для дальнейшего ручного или автоматического анализа. Базовый морфемный словарь пригоден для работы с лингвистическим материалом любого уровня языка. При исследовании звукового строя можно получить данные фоностатистического характера, а также исследовать фономорфологию на материале словника. Организация словаря позволяет получать также морфотактическую информацию. Наконец, если исследователь работает со словами или словосочетаниями, а не с морфемами, то он получит сведения, касающиеся основной словарной единицы как целого.

Обратимся к вопросу о том, как должна быть организована лингвистическая информация в базовом морфемном словаре и какие задачи с его

помощью можно решать. Морфемный словарь представляет собой списки индексированных корневых и аффиксальных морфем. Индекс каждой морфемы указывает на возможные сочетания ее с другими элементами слова и позволяет получать правильные словоформы с данной морфемой. Для корня такие ссылки формируют словообразовательное гнездо. Для аффикса ссылочный аппарат содержит информацию о том, в каких позициях и в каком аффиксальном окружении данный аффикс встречается в анализируемом словнике. Помимо этих служебных данных для каждого элемента можно указать грамматическую информацию (лексико-грамматическое значение, транспонирующую функцию, значение по согласовательным категориям и т. д.). Назовем некоторые параметры, по которым могут упорядочиваться списки морфем:

1. Фонетический облик и модель алломорфного варьирования. В упорядоченном по этому параметру списке фонетически сходные элементы будут находиться рядом (ср., например, русские корневые морфемы *волок-* и *ворот-*). Этот параметр связан прежде всего с кругом фономорфологических задач.

2. Позиционная морфотактическая характеристика морфемы, которая служит обеспечению исследований морфемной структуры слова.

3. Семантическая характеристика морфемы — система индексации, показывающая соотнесенность элементов списка с тем или иным выделенным семантическим классом.

Приведем примеры форм получения выходной информации: 1) словарь, организованный по алфавиту,— канонический вид с традиционными лексикографическими пометами и толкованием; 2) обратный словарь словарных форм слова, словоформ, основ; 3) гнездовой словарь; 4) словарь служебных морфем, характеризуемых по частоте встречаемости, позициям внутри слова, фонетической структуре, содержательным свойствам и т. д.; 5) сопоставительный словарь, в котором приведены указанные формы выхода информации параллельно для нескольких родственных языков.

Как уже отмечалось, каждая форма выхода является результатом переорганизации базового морфемного словаря. Возникает вопрос, какая лингвистическая информация необходима для создания такого словаря. Важно сразу же отделить задачи лингвистической работы от проблем, решаемых программистами и математиками при введении такого словаря в память ЭВМ. Представить тот или иной словник в виде морфем, т. е. задать списки,— задача лингвистическая. Как показывает практика, автоматическое членение слов на морфемы требует дальнейшей коррекции и доработки. Сама задача разделения слов на морфемы сложна и для многих языков является моментом первичного описания. Однако предметом машинного исследования могут быть результаты даже предварительного морфемного анализа. Автоматическая обработка позволяет легко выявить как *основные закономерности, так и списки исключений*, помогая проверить правильность проведенного морфемного членения. Что может дать такой предварительный словарь? Он позволяет получить исчерпывающую характеристику каждого элемента списка по заданным параметрам, а это — основа для коррекции словаря, исследовательский материал, который предоставляет лингвисту ЭВМ. В качестве примера такого рода работы можно сослаться на создание на базе «Русского словообразовательного словаря» [10] морфемного словаря распознавания в Лаборатории экспериментальной фонетики им. Щербы ЛГУ совместно с Институтом математики СО АН СССР [11].

Создание многоцелевой автоматизированной словарной системы осуществляется в настоящее время в сотрудничестве ЛО Института языкознания и Ленинградского Научно-исследовательского вычислительного центра АН СССР. Эта система должна предоставить возможность для разнообразного анализа обрабатываемого материала — текстов и словарей — в интерактивном (диалоговом) режиме [12]. Для ввода текстов используются перфоленты типографского фотонаборного устройства. Существующие программы дают возможность получать алфавитно-частотные и об-

ратные словоуказатели с выводом их на АЦПУ и накоплением в архиве системы. Алфавитно-частотный словоуказатель, записанный в архив системы, служит исходным материалом для создания карточек, на которых собраны все формы данного слова. При этом учитывается общая частота словоформ данного слова. На основе анализа полученного на ЭВМ словоуказателя к «Лирике» Вяземского и Сводного словника словарей, составленного в Большой картотеке словарного сектора ЛО Института языкознания [13], выявлено около 50 слов, не зафикисированных в современных толковых словарях.

Наиболее характерной чертой современного этапа развития автоматизации лингвистических исследований является создание автоматизированных систем обработки естественного языка [14—16]. Это направление применения ЭВМ связано с использованием развитого логического аппарата, что позволяет обратиться к решению системных и прикладных проблем различных уровней языка. При этом в машинных программах используются результаты чисто лингвистических исследований.

Перспективы использования ЭВМ как средства автоматизации лингвистических исследований представляются тесно связанными с развитием таких направлений, как автоматическое распознавание речи и машинный перевод [17]. Однако лингвист может надеяться на помощь ЭВМ при решении гораздо более широкого круга вопросов. Но оправдание таких надежд скорее всего возможно при использовании уже имеющихся результатов для задач традиционного языкознания и не в единичных случаях. В большой мере это может относиться к исследованиям малоизученных языков народов СССР, например, палеоазиатских или самодийских.

Таким образом, в процессе использования ЭВМ как средства автоматизации работы лингвиста выделяются три стадии: 1) подготовительная — введение в ЭВМ имеющейся в распоряжении исследователя информации о том, как устроено слово того или иного языка, какими содержательными характеристиками оно обладает; 2) основная — процесс исследования каждой языковой единицы с помощью ЭВМ по заданным лингвистом параметрам (по отношению ко всему объему записанного в памяти материала); 3) конечная — удовлетворительное обеспечение процесса познания языкового материала.

ЛИТЕРАТУРА

1. *Белоногов Г. Г., Кузнецов Б. А.* Языковые средства автоматизированных информационных систем. М., 1983.
2. *Зализняк А. А.* Грамматический словарь русского языка. М., 1977.
3. *Бектаев К. Б., Ризаев С. У.* О частотных списках графемно-фонемных сочетаний.— В кн.: Статистика казахского текста. Алма-Ата, 1973.
4. *Якубайтис Т. А.* Использование ЭВМ в лингвистических исследованиях.— ВЯ, 1979, № 3.
5. *Оливериус З. Ф.* Морфемы русского языка. Частотный словарь. Praha, 1976.
6. *Алексеев П. М.* Семантические частотные словари.—В кн.: Статистика речи и автоматический анализ текста. Л., 1977.
7. *Караулов Ю. Н., Молчанов В. И., Афанасьев В. А., Михалев Н. В.* Анализ метаязыка словаря с помощью ЭВМ. М., 1982.
8. *Герд А. С., Богданов В. В., Азарова И. В., Аверина С. А., Зубова Л. В.* Автоматизация в лексикографии и словари-конкордансы.— ФН, 1980, № 1.
9. *Вертель В.А., Вертель Е. В., Рогожникова Р. П.* К вопросу об автоматизации лингвистических работ.— ВЯ, 1978, № 2.
10. *Worth D. S., Kozak A. S., Johnson D. B.* Russian derivational dictionary. New York, 1970.
11. *Бондарко Л. В., Величко В. М., Загоруйко Н. Г.* Словообразовательный словарь и его использование для автоматического распознавания речи.—В кн.: Тезисы докладов и сообщений 12-го Всесоюзного семинара АРСО. Киев, 1982.
12. *Рогожникова Р. П., Чернышева Л. В.* Возможности использования автоматизированной системы в лексикографической работе.— В кн.: Информационно-вычислительные проблемы автоматизации научных исследований. М., 1983.
13. *Рогожникова Р. П.* Редкие слова в произведениях авторов XIX в.— ВЯ, 1982, № 1.
14. Computers in the humanities. Ed. by Mitchell J. Z. Edinburg, 1974.
15. *Попов Э. В.* Общение с ЭВМ на естественном языке. М., 1982.
16. *Ершов А. П.* Методологические предпосылки продуктивного диалога с ЭВМ на естественном языке.— ВФ, 1981, № 8.
17. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Крысин Л. П., Лазурский А. В., Перцов Н. В., Санников В. З.* Лингвистическое обеспечение в системе автоматического перевода третьего поколения. М., 1978.