ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 4 1987

ПИОТРОВСКИЙ Р. Г.

ЛИНГВИСТИЧЕСКИЕ АВТОМАТЫ И МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА

Инженерная лингвистика (ИЛ), являющаяся теорией и практикой использования лингвистических автоматов (ЛА) [1], сформировалась в результате взаимодействия таких научных стимулов и социальных потребностей, как стремление применить к исследованию языка объективные методы и новые идеи, выработанные в естественных науках и в математике, а также все растущие запросы информационной индустрии.

Создание Машинного фонда русского языка (МФ) является по своей сути инженерно-лингвистической задачей. Поэтому, планируя архитектуру и функционирование фонда, следует помнить о гносеологических, онтологических и прагматических особенностях инженерно-лингвистического подхода к изучению и описанию языка.

ИЛ иногда рассматривают как сугубо прикладную, подчиненную нелингвистическим задачам отрасль знаний. Однако это неверная оценка. Наряду с актуальными прикладными задачами ИЛ имеет свою совершенно самостоятельную, надстраивающуюся над узкотехнологическими задачами фундаментальную проблематику. В чем же состоит существо этой проблематики?

Чтобы лучше понять отличие проблематики ИЛ от методологических задач общего языкознания, рассмотрим соотношение биохимии и биофизики с такими классическими разделами биологии, как ботаника или зоология. Последние изучают явления жизни на наблюдаемом уровне. Что же касается биохимии и биофизики, то они исследуют жизнь на ненаблюдаемых микроуровнях — молекулярном и атомном. Сходным образом классическое языкознание изучает функционирование языка на наблюдаемом человеком уровне порождения и восприятия устной и письменной речи. Задача же ИЛ состоит в том, чтобы исследовать структуру языка и текстообразование на уровне предельной детализации лингвистических объектов и связей, некоторые из которых ненаблюдаемы (ср. такие лингвистические объекты, как дифференциальный признак фонемы или сема). Единственно возможным приемом изучения ненаблюдаемых объектов является метод моделирования. Отсюда следует, что МФ должен строиться как совокупность моделей тех лингвистических объектов, которые будут включены в этот фонд.

Какие же гносеологические требования следует предъявлять к лингвистическому моделированию? Первое требование определяет отношение лингвистической модели к ее оригиналу. ИЛ не может довольствоваться построением умозрительных гипотез, пусть даже облеченных в высокую математическую форму. Каждая модель должна быть подвергнута проверке с точки зрения того, как она объясняет реальность языка. Реали-

[1] Лингвистический автомат есть сочетание лингвистического алгоритма, реализующей его программы и компьютера.

зацию таких моделей, в том числе тех лексико-грамматических моделей, которые будут заложены в МФ, следует рассматривать не столько в плане практического внедрения, сколько с точки зрения проверки строгости и объяснительной силы данной модели. Поэтому теоретическим фундаментом ИЛ должна стать теория моделей воспроизводящих инженерно-лингвистических моделей (ВИЛМ). Второе требование относится к выбору принципа построения ВИЛМ. Сотни языковедов в нашей стране и за рубежом, работавших в 60—70-х годах в русле идеи порождающей грамматики и семантики, категорически высказывались в пользу построения чисто дедуктивных умозрительных моделей. К сожалению, как только эти модели подверглись экспериментальной проверке, обнаружилось, что их объяснительная способность крайне мала. Создатель первого в мире робота, способного принимать речевые сигналы и отвечать на них, Т. Виноград, жаловался на то, что «трансформационные алгоритмы способны обрабатывать лишь небольшие подмножества английского языка, да и те неэффективным образом» [1]. Можно показать, что существуют внутренние антиномии языка, препятствующие построению достаточно общих и «сильных» дедуктивных моделей [2].

Опыт построения и эксплуатации отечественных промышленных систем МП и логико-семантической обработки больших потоков иноязычных и русскоязычных документов [3] показал, что при построении МФ целесообразно ориентироваться на индуктивные или индуктивно-дедуктивные схемы, вырастающие на основе тщательного и целенаправленного изучения информационно-смысловой и синтактико-статистической структуры реального текста [4, 5].

Обратимся теперь к онтологическому аспекту построения разного вида ВИЛМ, который состоит в необходимости 1) постоянно учитывать те свойства речевой деятельности человека, которые резко противопоставляют эту деятельность языковым и речевым возможностям ЛА, и 2) не копировать подряд все детали лингвистического объекта, но выборочно моделировать в ЛА (и соответственно помещать в фонд) лишь те элементы и свойства этого объекта, которые оказываются абсолютно необходимыми для решения конкретных задач (в умении выделить эти элементы и свойства и состоит профессиональное искусство инженерного лингвиста) [2].

Непонимание соотношения онтологии естественного языка с существом «языка» ЛА дорого обходится «вычислительной» лингвистике: из сотен проектов, объявленных в конце 50-х и начале 60-х годов, до эксперимента и промышленной эксплуатации удалось довести к середине 80-х годов не более двадцати систем автоматической переработки текста (отметим, что в этих проектах по мере возможности учитывались указанные выше онтологические особенности естественного языка). Остальные дедуктивные проекты прекратили существование уже на начальных этапах своего развития. Исходя из всего сказанного, становится ясным, что современ-

[2] К этим свойствам в первую очередь относятся: а) метафорическая открытость и адаптивность естественного языка, которая основывается на способности человеческого сознания к бесконечной селекции и ассоциированию принимаемой информации; б) способность превращать при порождении текста обобщенное и размытое значение языкового знака в конкретное, четко очерченное значение речевого знака, а затем распознавать это значение путем сопоставления конкретных ситуативных условий употребления данного знака с его языковым значением. Современный ЛА полностью лишен этих возможностей: он использует в качестве своего «языка» либо заранее заданные номенклатуры языковых единиц, либо полностью лишенные метафорических возможностей исчисления [6].

ный ЛА нельзя рассматривать как автономно функционирующее устройство, черпающее из МФ всю необходимую для обработки русских текстов информацию. Полную обработку текста можно осуществить только в коммуникативной системе «человек — автомат — человек», обладающей обратной связью с человеком — потребителем информации.

Определим теперь основные принципы построения лингвистических автоматов ближайшего будущего,— принципы, которые закладываются в проектирование ЭВМ пятого поколения [7], в том числе таких компьютеров, в которых будет моделироваться самоорганизация и ассоциативное построение памяти (ЭВМ так называемого шестого поколения) [8]. Эти принципы должны быть учтены при построении и развитии МФ, которому суждено будет взаимодействовать с ЭВМ новых поколений. Таких принципов по крайней мере четыре.

1. Принцип интерактивной работы лингвистических автоматов. Смысл его состоит, с одной стороны, в том, чтобы обеспечить контроль человека над функционированием ЛА, а с другой стороны, чтобы разумно распределять лингвистические функции между ЛА и человеком. Так, например, при машинном переводе автомат должен осуществлять массовые рутинные и полурутинные операции по отождествлению единиц текста с единицами автоматического словаря, осуществлять простейший грамматический и смысловой анализ входного текста и находить эквиваленты результатам этого анализа на выходном языке. В то же время тонкие стилистические и семантико-синтаксические операции средствами выходного языка остаются пока в компетенции человека.

2. Принцип модульной архитектуры, который предусматривает построение ЛА в виде ансамбля программных модулей, каждый из которых воспроизводит определенный уровень и/или определенный аспект речемыслительной деятельности человека.

3. Принцип развития, который состоит в построении ЛА в виде открытой системы, развивающейся путем включения новых или обновления старых модулей. При этом указанные изменения, учитывающие идеи приложения теории нечетких множеств к моделированию процесса нововведения [9] (в нашем случае вторичного семиозиса), не должны вызывать существенных изменений в структуре ЛА. Реализация этого принципа особенно важна для построения эффективного МФ.

4. Принцип самостоятельного функционирования модулей ЛА. Реализация этого принципа позволяет отдельным модулям ЛА и их наборам работать автономно, выполняя при этом разнообразные виды автоматической переработки текста от самых примитивных (например, составление алфавитных, обратных или частотных словарей) до достаточно сложных семантико-синтаксических операций (например, составление реферата текста или решение с помощью ЭВМ, подвключенной к МФ, некоторой историко-лингвистической задачи) [10].

Если согласиться с только что описанной стратегией построения ЛА, то основные технологические задачи построения МФ русского языка, который должен будет стать в первую очередь универсальной лингвистической информационной базой данных для всех автоматов, работающих с русским языком, можно свести к следующим пунктам:

1) должны быть сформулированы информационные, вероятностные и системные принципы отбора и включения в МФ лингвистических единиц (основ, морфем, словоформ, словосочетаний), а также их парадигматических связей и валентностей, особое внимание должно быть обращено

а)
**** ОПЕРАТОР
PERSONAL COMPUTER ACC 80000

ЛИЧНЫЙ ЭВМ ACC 8000

**** ОПЕРАТОР D
FOR BUSINESS APPLICATIONS PERSONAL COMPUTER ACC 8000 PROVIDES A WIDE RANCE OF AC COUNTING, FINANCIAL PLANNING AND MANAGEMENT CONTROL.

ДЛЯ ДЕЛОВЫХ ПРИМЕНЕНИЙ ЛИЧНЫЙ ЭВМ ACC 8000 ОБЕСПЕЧИВАЕТ ШИРОКИЙ ДИАПАЗОН ВЕДЕНИЕ ОТЧЕТНОСТИ (РАСЧЕТНЫЙ, РАССЧИТЫВАЯ), ФИНАНСОВОЕ ПЛАНИРОВАНИЕ И УПРАВЛЕНИЕ КОНТРОЛЬ (КОНТРОЛИРОВАТЬ).

б)
***** ОПЕРАТОР L
ИСПОЛЬЗОВАНИЕ ПРОСТЫХ МИКРОКАЛЬКУЛЯТОРОВ ПРИ ТАКТИЧЕСКИХ РАСЧЕТАХ

THE USAGE OF SIMPLE MICROCALCULATORS BY(IN) TACTICAL CALCULATIONS

***** ОПЕРАТОР
МИКРОКАЛЬКУЛЯТОРЫ — ЭТО ПЕРЕНОСНЫЕ ВЫЧИСЛИТЕЛЬНЫЕ УСТРОЙСТВА МАЛОГО РАЗМЕРА И НЕБОЛЬШОЙ МАССЫ (50—300 Г), ВЫПОЛНЕННЫЕ НА БАЗЕ ЭЛЕКТРОННЫХ ЭЛЕМЕНТОВ И ПРЕДНАЗНАЧЕННЫЕ ДЛЯ ИСПОЛЬЗОВАНИЯ В ЛЮБЫХ УСЛОВИЯХ. ОНИ ЗНАЧИТЕЛЬНО ПОВЫШАЮТ ПРОИЗВОДИТЕЛЬНОСТЬ ТРУДА И ОБЛЕГЧАЮТ ВЫПОЛНЕНИЕ ОПЕРАЦИЙ, СВЯЗАННЫХ С ОТНОСИТЕЛЬНО НЕСЛОЖНЫМИ ВЫЧИСЛЕНИЯМИ.

MICROCALCULATORS — IT IS PORTABLE CALCULATING DEVICE OF SMALL SIZE AND SMALL WEIGHT (50—300 G), BUILT ON THE BASIS OF ELECTRONIC ELEMENTS AND INTENDED FOR USAGE IN ANY CONDITIONS. THEY CONSIDERABLY INCREASE LABOUR PRODUCTIVITY AND FACILITATE PERFORMING OF OPERATIONS, CONNECTED WITH RELATIVELY SIMPLE CALCULATIONS.

Рис. 1. Фрагмент результатов работы системы СИЛОД по переводу английских (а) и русских (б) текстов. Возможности системы СИЛОД демонстрировались в вычислительных центрах Индии, а также показывались представителям информационных служб Финляндии, Венгрии и ФРГ.

на выявление и фиксацию имплицитных (не наблюдаемых непосредственно) семантических связей;

2) все нечеткие лингвистические объекты, подлежащие включению в фонд, должны преобразовываться по единой, заранее разработанной процедуре в дискретные «машинные» лингвистические единицы (это в первую очередь относится к значениям основ, словоформ и словосочетаний), особое внимание должно быть обращено на деятельное кодирование этих единиц;

3) МФ должен строиться как открытая база данных, которую можно легко, без существенной структурной перестройки пополнить и корректировать как за счет введения новых, так и за счет устранения ставших ненужными лингвистических единиц.

И в заключение несколько слов о прагматическом аспекте планирования МФ.

МФ русского языка весьма дорогостоящий проект. Поэтому при планировании его создания и развития нельзя не учитывать потребностей

отечественной и зарубежной информационной индустрии, которая всегда будет надежным источником финансирования МФ.

В настоящее время в промышленную и экспериментально-промышленную эксплуатацию вышло несколько систем автоматической переработки научно-технических и деловых текстов (ср. рис. 1). Назовем здесь экспериментально-промышленные системы англо-русского МП, построенные под руководством К. Б. Бектаева, Л. Н. Беляевой и С. В. Соколовой, В. В. Гончаренко, Ю. Н. Марчука и Б. Д. Тихомирова, а также промышленную систему переработки иероглифического текста, выполненную под руководством С. Г. Пучкова и С. В. Соколовой. Отсутствие единого МФ русского языка не дало возможности сделать эти системы совместимыми в их выходной (русской) части. А ведь такая совместимость позволила бы сэкономить много времени, средств и творческих усилий создателей этих систем.

На международном рынке особый интерес привлекают алгоритмы русско-иноязычного МП (например, русско-английский, русско-финский, русско-арабский и т. п.).

Учитывая прагматические перспективы развития фонда, было бы желательным учесть интересы реального промышленного МП в том смысле, чтобы заложить в МФ информацию, которую можно было бы легко переносить в двуязычные алгоритмы. При этом следует учесть, что основной инженерно-лингвистической особенностью наших конкурентоспособных систем является компактность их словарных статей, обеспечивающаяся строгим отбором в них такой информации, которая абсолютно необходима и релевантна для нужд анализа русского входного текста (в первую очередь) и для его синтеза (во вторую очередь). Речь здесь, разумеется, идет о деловой, публицистической и научно-технической прозе. При этом имеются в виду не раритеты, а нормативные формы слова (здесь важно также определить разумный баланс между машинными основами и флексиями). Крайне важны сведения о многозначности и омонимии словоформ в указанных стилях, об их типовых валентностях.

Одновременно можно было бы подумать и о такой лингвистической информации русских входных словоформ, которая необходима для синтеза выходного текста на корневом (например, английском или китайском), агглютинирующем (например, венгерском или японском) или фузионном с внутренней флексией (например, арабском) языках.

ЛИТЕРАТУРА

1. *Виноград Т.* Программа, понимающая естественный язык./Пер. с англ. М., 1976. С. 74.
2. *Piotrowski R.* Text processing in the Leningrad research group «Speech statistics». Theory, results, outlook // Literary and linguistic computing. 1986. V. 1. № 1.
3. *Чижаковский В. А., Бектаев К. Б.* Статистика речи. 1956—1986.: Библиографическое пособие. Кишинев, 1986.
4. *Fu K. S.* A step towards unification of syntactic and statistical pattern recognition // IEEE Transactions of pattern analysis and machine intelligence. 1986. V. 8. № 3.
5. *Piotrowski R.* Text—Computer — Mensch. Bochum, 1984.
6. *Мельников Г. П.* Системология и языковые аспекты кибернетики. М., 1978.
7. *Симонс Дж.* ЭВМ пятого поколения: компьютеры 90-х годов. М., 1985.
8. *Kohonen T.* Self-organisation and associative memory. Berlin, 1984.
9. *Nekola J., Novak V.* The application of fuzzy set theory in innovation process modelling // Fuzzy information knowledge representation and decision analysis proceedings: IFAC Symposion, Marseille, 19—21 July 1983. Oxford, 1984.
10. *Altmann G.* Das Piotrowski-Gesetz und seine Verallgemeinerungen // Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte/ Hrsg. von Best K.-H., Kohlhase J. Göttingen, 1984.