Wiener Slawistischer Almanach 52 (2003) 55-63

Владимир Дубичинский

СИСТЕМЫ МАШИННОГО ПЕРЕВОДА С ПОЗИЦИЙ КОМПЬЮТЕРИЗАЦИИ ЛЕКСИКОГРАФИИ

Современная лексикография сегодня уже невозможна без широкой компьютеризации. Постепенно традиционные методы заменяются компьютерной обработкой лексикографических данных. Кроме очевидной экономии времени, интеллектуальных ресурсов и др. компьютеризация даёт возможность поддерживать точность и тщательность контроля лексикографического процесса, оперативно исправлять допущенные, ранее не замеченные, ошибки, создавать новые перспективные в научном отношении лексикографические комплексы.

Важно заметить, что компьютер "живёт" умом и чувствами лексикографа, который использует его для совершенствования и ускорения обработки данных. Лишь в общении, когда лексикограф и компьютер ведут диалог в режиме полного взаимопонимания, машина преображается, она становится незаменимой в процессе создания и использования словаря.

Компьютеризация лексикографической деятельности заключается, прежде всего, в создании специализированных *машинных банков данных* и в разработке методов формирования этих банков, представления информации в банках и её использовании. Современная лексикография всё шире пользуется машинными банками данных, в частности, большими корпусами текстов на магнитных носителях, в которых компьютер по запросу осуществляет поиск нужных слов. На этой основе формируется целое новое направление лингвистики и лексикографии - *корпусная лингвистика и лексикография.*

Началом развития корпусной лексикографии русского языка можно считать осуществление с 1983 года в Институте русского языка РАН программы формирования машинных фондов.

Машинным фондом русского языка называется программа комплексной информатизации исследований в русистике, разработанная А.П.Ершовым и Ю.Н.Карауловым.

Под комплексной информатизацией научных исследований и прикладных разработок понимается: 1. последовательное оснащение современными вычислительными машинами с перспективой их объединения в

единую вычислительную сеть; 2. последовательное накопление на машинных носителях и в базах данных главнейших источников, необходимых для научного изучения языка и осуществления прикладных разработок; 3. создание программных средств, необходимых для подготовки научных трудов по филологии и проведения прикладных разработок; 4. развитие прикладных направлений (лексикографии, терминоведения, машинного перевода, автоматической обработки данных на естественном языке) как составной части академической и вузовской науки, являющихся, с одной стороны, проводником результатов фундаментальных исследований в практику, а с другой - источником новых идей и данных для фундаментальной науки.

В рамках проекта машинного фонда русского языка разрабатываются 9 фондов-составляющих (генеральный словник, словарный, текстовой, грамматический, терминологический, лингвогеографический, исторический, фонетический и лингвистический программно-источниковый фонды русского языка) и одна программная система - типовой, лингвистический программно-источниковый пакет UNILEX.

Генеральный словник и словарный фонд сосредоточивают в своих базах данных все словари современного русского языка. Словарный фонд создаётся одновременно и как информационная система по лексике, и как система автоматизации лексикографических работ.

Текстовой фонд - совокупность автоматических конкордансов (словарей особого типа, в которых для каждого словарного слова приведены все контексты его употреблений в определённом корпусе текстов) произвольных текстов. Такие конкордансы изготавливаются по мере надобности и хранятся в фонде.

Грамматический фонд — это информационная система по академическим грамматикам русского языка. Кроме грамматик, она включает в себя также другие справочные материалы по морфологии, синтаксису и словообразованию.

Терминологический фонд - информационная система по современной научно-технической терминологии.

Лингвогеографический фонд включает в себя автоматический вариант „Диалектологического атласа русского языка", а также словарные и текстовые справочные системы по диалектной лексике.

Исторический фонд - система автоматизации подготовки к изданию памятников русской письменности.

Ядром фонетического фонда является информационная система по фонотеке Института русского языка. С ней связана система автоматизации фонетических исследований.

Лингвистический программно-источниковый фонд включает в себя источники для изучения русского языка (на машинных носителях) и программные средства для их обработки.

Средствами комплектации фондов-составляющих являются так называемые лингвистические программно-источниковые пакеты, т. е. программные комплексы, управляющие крупными лингвистическими источниками. К ним относятся, например, автоматические конкордансы, автоматические словари, автоматический вариант „Диалектологического атласа русского языка", информационная система по „Краткой русской грамматике", процессоры русского языка и другие средства автоматизации и информационного обеспечения лингвистических исследований и разработок.

В 1985-1996 в машинном фонде русского языка на машинных носителях и частично в базах данных накоплены текстовые источники русской литературы XIX-XX вв., главнейшие словари русского языка, „Краткая русская грамматика" (1989), некоторые другие материалы справочного характера, созданы текстовые корпусы поэзии, художественной прозы, общественно-политических и технических текстов; разработан программно-источниковый пакет UNILEX для персональных компьютеров, состоящий из 5 подсистем: подсистемы обработки лингвистических данных общего назначения, текстоориентированной подсистемы, словарной подсистемы, телекоммуникативной подсистемы и редакционно-издательской подсистемы. Каждая из этих подсистем может использоваться независимо от других.

Автоматизированные лексикографические системы, т.е. системы автоматизации подготовки и использования словарей, включают в себя программы и справочные данные, необходимые для лексикографической обработки текстов. В них используются текстовые редакторы для ввода и коррекции данных, программы контроля данных и запросов к системе, программы контроля орфографии и разметки входного текста, программы сегментации текста на слова, словосочетания, предложения и фрагменты словарных статей, программы лемматизации и подсчёта статистики словоупотреблений, программы загрузки, поиска и коррекции данных и др.

Введённые в систему тексты и/или словари размещаются в базах данных и снабжаются словоуказателями и другими индексами, позволяющими по слову или его характеристикам находить его контексты или словарные статьи, в которых оно описано. Результатом автоматической обработки текста в автоматизированных лексикографических системах являются частотные словари, конкордансы (словоуказатели с контекстами), чаще всего принимающие форму автоматических конкордансов, автоматические

моно- и многоязычные словари, размещаемые в базах данных и используемые программами лексикографических систем в качестве справочного материала при обработке новых данных. Поэтому такие системы являются развивающимися системами. Автоматические словари используются в системах автоматического перевода, а также в информационных системах и системах общения с компьютером на естественном языке в качестве справочников при подготовке и расширении словарей и уточнении грамматик этих систем.

В составе лингвистического обеспечения автоматизированных систем различают три группы функций автоматической обработки текста: автоматическое индексирование входных документов, составление поисковых предписаний по тексту запросов и автоматизированное ведение словарей системы. Ядром лингвистического обеспечения автоматизированных информационных систем являются *информационно-поисковые тезаурусы*, в терминах которых производится индексирование вводимых в систему текстов и запросов на их поиск. Индексирование текста заключается в составлении к нему поискового «образа», в котором указываются понятия, описываемые в тексте, и отношения между ними. Аналогично обрабатываются и запросы к системе. Сравнением поисковых предписаний с поисковыми образами документов осуществляется выбор текстов запрашиваемой тематики. Существуют и бестезаурусные системы, способные осуществлять поиск текстов по любым сочетаниям слов, встречающихся в них. В таких системах автоматически строятся словоуказатели к вводимым текстам.

Обобщив и проанализировав современный опыт российской лексикографии Харьковское лексикографическое общество пришло к выводу о создании так называемого *лексикографического автоматизированного рабочего места*. Лексикографическое автоматизированное рабочее место (ЛАРМ) представляется мне универсальной базой данных или совокупностью баз данных, которыми лексикограф может оперировать при создании электронных словарей.

На мой взгляд, разработка ЛАРМ – новый вид лексикографической деятельности, предполагающий творческое объединение усилий лингвистов-теоретиков, лексикографов-практиков и программистов. В настоящее время выдвинутый группой лингвистов Харьковского лексикографического общества термин «ЛАРМ» играет роль принципиального понятия компьютерной лексикографии, так как на основе определенной универсальной совокупности баз данных предлагается стандартизировать и унифицировать весь лексикографический процесс, направить лексикографическую деятельность в единое русло интернационализации электронно-словарных комплексов.

По идее харьковских лексикографов ЛАРМ должно включать:

- инвариантную электронную структуру словарной статьи, разработанную на основе баз данных толковых, идеографических, переводных и др. словарей, которые также являются составными частями соответствующих баз данных;

- в ЛАРМ необходимо наличие четко разработанной системы условных знаков и лексикографических помет, которые должны быть снабжены соответствующим программным обеспечением;

- компьютерную программу автоматического редактирования сканированных текстов и правки орфографических, грамматических, синтаксических и т.п. ошибок; и т.д.

Неотъемлемой частью ЛАРМ может стать АРМП (автоматизированное рабочее место переводчика), которое удовлетворяет всем лексикографическим и переводоведческим запросам лингвистов и помимо электронных словарных комплексов оснащается также современными системами машинного перевода.

В последнее время особую актуальность для переводной лексикографии приобретает *машинный (автоматический) перевод* – перевод текстов с одних естественных языков на другие с помощью компьютера.

Как известно, первые работы по автоматизации перевода появились в США в начале 50-х гг. XX в. Первый публичный эксперимент по автоматическому переводу был проведён в Джорджтаунском университете в 1954 г. В середине 50-х гг. работы начались во многих странах, в том числе и в СССР: в 1954 г. в Москве начали работать группы И.К.Бельской и Д.Ю.Панова (англо-русский перевод) и А.А.Ляпунова и О.С.Кулагиной (французско-русский перевод).

В настоящее время в разных странах создано большое количество экспериментальных и практических систем автоматического перевода. Из коммерческих систем наиболее распространена SYSTRAN, для которой известно около 15 версий для разных пар языков. Все эти версии снабжены большими словарями (например, в словаре для пары «русский – английский» 200 тыс. слов общей лексики и 200 тыс. терминов). Перевод, полученный в результате использования этой системы, требует существенного редактирования. Работают также системы ATLAS, LOGOS, LITRAS и др.

Примерно к 1995 г. появились коммерческие системы автоматического перевода в России, например, система англо-русского (и обратно) перевода STYLUS, комплект переводческих программ ЛЕКСИКОН, серия систем машинного перевода SILOD-MULTIS, созданная в Российском государственном педагогическом университете им. А.И.Герцена (Санкт-Петербург) и др.

Основными центрами компьютерной лингвистики и лексикографии в Украине являются Киев, Львов и Харьков.

В Институте кибернетики им. В.Глушкова НАН Украины (Киев) разработана концепция интегрального словаря, на основе которой создается автоматизированная система ведения интегральных словарей АСВИС. На сегодняшний день из этой программы создана подсистема формирования словарей СИФОРС, которая ориентирована на ведение терминологических баз данных.

Близкая по духу концепция создания лексикографического процессора осуществляется сегодня в Украинском языково-информационном фонде (Киев).

В Институте прикладной информатики НАН Украины (Киев) создана адаптивная лингвистическая система АЛИСА, которая представляет собой естественноязыковой лингвистический процессор, ориентированный на выполнение целого ряда функций, в частности автоматизированного создания словарей, тезаурусов, фразеологических, терминологических баз данных и др. На основе этой же системы построены украинский автокорректор ТВИР и самый популярный украинский спелчекер РУТА.

Во Львовском политехническом университете созданы: а) система поддержки многоязычных терминологических словарей «СЛОВО», в которой отработаны системные вопросы технологии подготовки словарей к изданию; б) многоязычный банк стандартизированных терминосистем.

Львовский экономический институт разработал автоматизированную систему для создания и сопровождения многоязычных терминологических словарей.

Созданные в Харькове успешно работают по всей Украине системы ПАРС (переводная англо-русская система), ее модификации ПАРС/D (для русско-немецкого и немецко-русского автоматического перевода) и РУМП (русско-украинский и украинско-русский машинный перевод). Система ПАРС оснащена целым комплексом терминологических компьютерных словарей почти по всем областям человеческой деятельности (на много десятков тысяч терминов каждый).

Остановлюсь подробнее на кратком описании *системы* РУМП, разработанной харьковским ученым М.С.Блехманом. Она работает на IBM-совместимых персональных компьютерах и переводит тексты с русского языка на украинский и с украинского на русский.

РУМП обеспечивает связный перевод текстов не только обиходного (общеязыкового), но также и терминологического характера, покрываемых авиационным (10 тыс. терминов), экологическим (15 тыс. терминов), компьютерным (14 тыс. терминов) и финансовым (12 тыс. терминов) электронными словарями.

Система поддерживает двухсторонние украинско-русские и однозычные грамматические словари. Например, если ввести русское слово с его украинским эквивалентом, система установит украинско-русское соответствие. Иными словами, один и тот же словарь используется как для украинско-русского, так и для русско-украинского перевода.

Словарная статья системы РУМП представлена в традиционном виде «бумажных» словарей: сначала приводится заголовочное слово с одним или несколькими переводными эквивалентами, а затем – словосочетания.

Пользователь может просмотреть как украинскую, так и русскую часть словаря; его можно «перелистать» вверх и вниз; есть возможность вносить в словарь редакторские правки.

Система работает в многозадачном режиме, будучи совместима с широко используемыми текстовыми процессорами. Уникальным свойством РУМП является автоматическое грамматическое кодирование слов, вводимых в словарь: пользователь вводит слово в его канонической форме, а система определяет его грамматические характеристики, включая тип склонения/спряжения.

В процессе перевода РУМП отличается следующими особенностями:

- одновременное использование до четырех словарей с заданием их приоритетов;
- перевод как всего текста целиком, так и его фрагмента, заданного пользователем;
- выделение многозначных слов в тексте перевода звездочками, что дает возможность выбрать из предлагаемых системой наиболее адекватный переводной эквивалент;
- выделение ненайденных в словарях РУМП слов;
- ввод новых слов и словосочетаний прямо из текста в словарь.

Приведу несколько примеров переводного теста, проведенного с системой русско-украинского машинного перевода РУМП:

1. Не стоит ли забыть всё, что было раньше? (Чи не варто забути все, що було раніше?) – РУМП: *Не коштує чи забути всіх, що було раньше?*

2. Эти карманные часы очень дорогие (Цей кишеньковий годинник дуже дорогий) – РУМП: *Ці карманные години дуже дорогі.*

3. Не стоит так сердиться на своих друзей, которые говорят… (Не варто так гніватися на своїх друзів, які кажуть...) – РУМП: *Не коштує так сердиться на своїх друзей, що говорять... .*

4. Они знают толк в этом деле (Вони розуміються на цій справі) – РУМП: *Вони знають толк в цій справі.*

5. Фирма находилась по адресу (Фірма знаходилася за адресою) – РУМП: *Фірма знаходилась за адресою.*

6. Рассматривающиеся в текущем году на общем собрании задания (Завдання, що розглядаються у цьому році на загальних зборах) – РУМП: *Що розглядаються в поточному році на загальному зібранні завдання.*

7. Мальчик подошёл к первому солдату, а мужчина – ко второму (Хлопчик підійшов до першого солдата, а чоловік – до другого) – РУМП: *Хлопчик підійшов до першого солдата, а чоловік – до другого.*

Как можно увидеть из приведенных примеров, система машинного перевода РУМП еще далека от совершенства. В данной статье я не буду подробно анализировать недостатки лексического и грамматического характера. Однако, учитывая мнение специалистов о принципиальной недостижимости 100% качества перевода (это сравнимо лишь с полным моделированием искусственного интеллекта), а также о возможности редактирования и подбора синонимичных вариантов, сырой переводческий материал, который предоставляет РУМП, на несколько порядков облегчает и ускоряет действия переводчиков.

С 70-х гг. наблюдается тенденция к интеграции всех подходов к конструированию систем автоматической обработки текста в рамках *конструирования искусственного интеллекта* – направления в информатике, связанного с созданием сложных человеко-машинных и робототехнических систем, моделирующих человеческую деятельность в различных сферах и предметных областях. В таких системах текст на естественном или искусственном языке является как источником накопления знаний системы, так и источником данных для выбора её поведения, а также средством взаимодействия системы с человеком. Здесь функции редактирования всё больше сливаются с функциями содержательной обработки, образуя единый аппарат понимания текстов.

Компьютеризация открывает возможности для автоматизации наиболее сложных областей человеческой деятельности, требующих затрат прежде всего интеллектуального труда, таких, как редакционно-издательские процессы, извлечение информации из текстов, медицинская и техническая диагностика, экспертная деятельность, проектирование машин и сооружений, изготовление проектной документации, управление социально-экономическими системами. Во всех этих случаях *автоматическая обработка текста* играет первостепенную роль.

Однако в таких массовых «промышленных» применениях автоматическая обработка текста должна опираться на мощную информационную поддержку в виде автоматизированных словарных картотек, автоматических словарей, грамматик и других форм представления лингвистических данных в компьютере. Разработка таких систем приобретает форму машинных фондов национальных языков (например, Машинный фонд

русского языка, Машинный фонд украинского языка), национальных автоматизированных лексикографических служб и т. п.

Как убедительно доказывает весь ход научно-технического прогресса, компьютеризация словарной деятельности существенно расширяет возможности лексикографов. Безостановочное развитие компьютерных технологий подсказывает необходимость полной компьютеризации словарных исследований. И это подтверждают сейчас новые направления лексикографии: создание словарных картотек на основе компьютерных баз данных, электронное построение словарных статей и автоматическая обработка лексического материала, составление печатных словарей на компьютерной основе и создание собственно электронных словарей (без их бумажных аналогов) и мн. др.

И мне кажется, именно развитие компьютерных словарных исследований, наряду с глубокими традиционными лингвистическими и переводоведческими открытиями, определяет широкое будущее лексикографии.

Литература

Дубичинский, В.В., Самойлов, А.Н. 2000. *Словари русского языка*, Харьков.

Дубичинский, В.В. 1998. *Теоретическая и практическая лексикография*, (=Wiener Slawistischer Almanach, Sonderband 45), Wien-Charkov.

Пещак, М.М. 1999. *Нариси з комп'ютерної лінгвістики*, Ужгород.

Широков, В.А. 1998. *Інформаційна теорія лексикографічних систем*, Київ.