УДК 82.085

С.В. ЛЕСНИКОВ
Сыктывкарский государственный университет

ГИПЕРТЕКСТОВЫЙ ЛИНГВИСТИЧЕСКИЙ УНИВЕРСУМ РУССКОГО ЯЗЫКА

Данная статья посвящена вопросу рассмотрения современного русского языка с точки зрения гипертекстового лингвистического универсума. Дается история вопроса, указывающая первые попытки машинной обработки лексических единиц литературных памятников. Говорится о необходимости конструирования лингвистического универсума русского языка в нелинейной форме.

The article involved is devoted to the problem of depicting modern Russian language from hyper textual linguistic universum point of view. The history of the problem is given; first attempts to make machine-made description of lexical units of literary monuments are described. The necessity to construct linguistic universum of Russian language in non-line form is being estimated.

По определению, «УНИВЕРСУМ (лат. universum – мировое целое, мир) – всеобщее <...> – множество, содержащее все элементы (объекты) какой-либо исследуемой области материального или духовного мира»*. При этом неполнота знаний (в нашем случае – языковая картина мира) предопределена на всех мыслимых уровнях – (для человека – конечность, компактность категорий пространство-время) от объема лингвистической информации до скорости и качества ее переработки. Кроме этого, специализация, углубленное познание в отдельной научной дисциплине, детализация естественным образом размывают общность даже этой одной конкретной дисциплины. Относительным выходом из парадокса неполноты знаний о мире как универсуме (согласуясь с неполнотой лингвистического универсума как продукта, отражающего в синтагматике и парадигматике языковые факты и явления) может служить некоторая предельно общая теория, позволяющая сконструировать, интегрировать в целостный объект ранее не связанные отдельные объекты (элементы)**.

Универсум – целостный объект конечно допускает принципиально различные членения, и в зависимости от основания членения можно получить качественно и количественно разнообразные множества «первичных» элементов, которые и представляют ступени иерархии конструируемого гипертекстового лингвистического универсума русского языка. На данном этапе исследования выделим следующие подуровни лингвистического универсума (анализ и синтез):

а) графокод (буквоэлемент) и графика (шрифт);

б) буквокод (буква, символ, знак, цифра, логограмма) и графемика (алфавит, азбука, система знаков, код);

в) морфокоды (морфема) и морфология + грамматика (морфемика);

г) орфограмматика (словоформа, лемма, слово, число, логема) и лексикология + семасиология (словарь, тезаурус);

д) орфосинтактика (композиция – словосочетание, синтагма, предложение, фраза, высказывание) и структурология (модель языка);

е) орфотектоника (текст, гипертекст; абзац, пункт, параграф, глава, раздел, часть, том, книга, библиотека; страница, строка, столбец, тетрадь, позиция) и феноменология (лингвистика, в нашем случае – русистика);

Кондаков Н.И. Логический словарь-справочник. М.: Наука, 1975. С.627.

**Карпов В.А.* Язык как система. Минск, 1992.

ж) предметно-семантические системы (литература; избранное, сочинения, собрание, свод; меморема) и системология (теория систем) (смотри, например, В.А.Карпов, Ю.А.Урманцев).

В процессе конструирования гипертекстового лингвистического универсума русского языка, в частности, используются:

а) «Словарь русских словарей /СРС/» (1100 словарей и энциклопедий в компьютерной форме, десять компакт-дисков, 7Гб). В личной (домашней) библиотеке предполагаемого научного руководителя проекта из имеющихся 1500 словарей русского языка – 800 пока не в компьютерной форме (в СРС из 1100 электронных словарей в личной библиотеке лишь 700 в книжной форме, а часть словарей актуализирована только в компьютерной, гипертекстовой форме).

Презентация «Словаря русских словарей» осуществлена 20 марта 2004 г. на международном конгрессе «Русский язык: исторические судьбы и современность» в МГУ им. М.В.Ломоносова.

«Словарь русских словарей» фактически состоит из: А) Книга: Лесников С.В. Словарь русских словарей: более 3500 источников / Предисловие проф. В.В. Дубичинского. Рецензенты: В.М.Андрющенко, Р.П.Рогожникова, Г.И.Тираспольский. М.: Азбуковник, 2002. 334 с. (500 экз. Опубликовано фактически в 2004 году). Б) 10 компакт-дисков (CD№ 01 Гизаурус ЛСВ, CD№ 02 МАС БАС2, CD№ 03 ССРЛЯ=БАС, CD№ 04 СРНГ Даль, CD СРС Филология, CD СРС РЯ Толковые, CD СРС МИР (символ история культура), CD СРС Термины, CD СРС Универсал, CD СРС ЭВМ Экономика). На каждом диске имеются поисковые программы. В) Бесплатная рассылка «СЛОВАРЬ РУССКИХ СЛОВАРЕЙ» http: // subscribe.ru/catalog/science.humanity.hypervault. Г) Каталог 10 компакт-дисков «Словарь русских словарей» (4000 источников, 1000 словарей, справочников, энциклопедий). В архивном файле.zip-100Kb.

б) «Всемирная литература от А до Я» (2672 автора, 200 тыс. файлов, содержащих полные тексты художественных произведений, десять компакт-дисков, 6Гб).

в) «Пресса 1995-2002. База данных материалов периодических изданий» (в основном полные подборки региональных газет, восемь компакт-дисков, 5,5Гб).

г) «Библиотека в кармане» (20 выпусков, 25 компакт-дисков, 17Гб, но часть материалов на разных CD повторяется, то есть реально после сортировки остается 5Гб).

д) Тематические энциклопедии и терминологические словари русского языка, отдельные подборки художественной литературы (русской, зарубежной, для школьников, студентов) – 100 компакт-дисков (70Гб).

Необходимо отметить, что компьютеризация лексикографических исследований в нашей стране активно осуществлялась в рамках проекта «Машинный фонд русского языка /МФ РЯ/» (Андрющенко В.М. Концепция и архитектура МФ РЯ. М.: Наука, 1989. См. работы В.М.Андрющенко, Ю.Н.Караулов, А.Я.Шайкевич, Л.И.Колодяжная, Ж.Г.Аношкина). В разработке Диалектологического подфонда МФ РЯ в качестве отв. исполнителя хоздоговорной НИР «Разработка и создание Автоматизированного Словаря русских народных говоров» и «Духовная культура рус. Севера. Словарь рус. говоров Республики Коми», науч. рук. НИР «Разработка и создание автоматизированной лексикографической системы» / АЛС / «ГОВОР» принимал участие С.В.Лесников.

По всей видимости, одним из первых в мировой лексикографии применил счетно-перфорационные машины для обработки литературных памятников (сочинения Фомы Аквинского, словоуказатели и конкордансы) в начале 50-х годов XX века Р.Буза (Германия). Для примера можно указать несколько иностранных коллективов, где впервые была поставлена и частично решена задача компьютеризированных национальных сводов лексикографических материалов: в США (Брауновский корпус текстов); во Франции в Институте французского языка – автоматизированная словарная картотека французского языка XVI-XX вв.; в Германии в Институте немецкого языка в Маннгейме – машин-

ные картотеки письменных и устных источников современного немецкого языка; в Швеции в Гетеборгском университете; в Киевском национальном университете (3 млн. украинских словоформ, Н.П.Дарчук, Л.А.Алексеенко); в Финляндии в Отделении славянских и балтийских языков и литератур Хельсинского университета (100 тыс. сл. статей, аннотированный корпус русских текстов ХАНКО, А.Мустайоки, М.В.Копотев).

Известны также емкие автоматизированные словари в виде терминологических банков данных: EURODICAUTOM (Люксембург, 300 тыс. сл. статей), LEXIS (ФРГ, 1,5 млн. терминов), TEAM (ФРГ, 1,5 млн. терминов), TERMDOC (Швеция), TERMIUM (Канада, Квебек, 1 млн. сл. стат.), аналогичные банки терминов имеются и в США, Италии, Эстонии, Мексике и других странах.

В России как наиболее авторитетные в области компьютерной лингвистики следует указать следующие коллективы ученых-лексикографов: МГУ им. М.В.Ломоносова (10 млн. словоупотреблений, корпус текстов русских газет конца XX века А.А.Поликарпов, О.В.Кукушкина, Б.В.Виноградова, С.О.Савчук, другие направления: В.В.Богданов; Л.В.Златоустова; Г.Е.Кедрова (Интернет-учебники); Ю.Н.Марчук; П.В.Гращенков, И.М.Кобозева; Н.В.Лукашевич, Б.В.Добров); Санкт-Петербургский государственный университет (компьютерная антология русского рассказа XX века, Г.Я.Мартыненко, А.О.Гребенников, Е.А.Козлова, Е.И.Лазаренко, Т.И.Шерстинова); Саратовский гос. университет (корпус Диалектологических текстов, В.Е.Гольдин); Казанский гос. университет (компьютерный лингвографический фонд русского языка, К.Р.Галиуллин; электронная коллекция книг XVIII века, В.В.Соловьев, А.В.Скоробогатов); Казанский государственный педагогический университет (синтаксический анализатор русских технических текстов, О.А.Невзорова, Н.В.Пяткин); Московский государственный лингвистический университет (информационные технологии и медиалингвистика, Р.К.Потапова, В.В.Потапов); Нижегородский государственный педагогический университет (корпус текстов литературной критики произведений постмодернизма, Д.В.Гугунава); Новосибирский гос. пед. университет (В.В.Кромер); Новосибирский институт филологии СО РАН (электронный корпус средневековых текстов, А.М.Лаврентьев); Петрозаводский государственный университет (грант РГНФ № 02-04-12015в, автоматизированная информационная система «Статистические методы анализа литературных текстов», В.Н.Захаров, А.А.Рогов, Ю.В.Сидоров, А.В.Король); Удмуртский госуниверситет (Ижевск) (система «Манускрипт», гранты РФФИ №02-07-90424в, 02-07-90318в, В.А.Баранов, А.А.Вотинцев, А.Н.Миронов, С.В.Ощепков, В.А.Романенко); Национальный корпус русского языка (лингвисты университетов и НИИ Москвы и Санкт-Петербурга, В.А.Плунгян, Д.В.Сичинава).

Конструирование лингвистического универсума русского языка в нелинейной форме с учетом реляционных, иерархических и сетевых парадигматических связей посредством реализации синтагматических связей в интерактивном режиме на ЭВМ позволит на основе новых информационных технологий при соответствующей классификации и систематизации объединить лексикографические материалы, обеспечить их оперативный ввод в научный оборот с целью оптимизации научных исследований в современной лексикографии.