

ПРОЕКТ ИНТЕРНЕТ-ВЕРСИИ СЛОВАРЯ
ЛЕКСИКО-СЕМАНТИЧЕСКИХ ВАРИАНТОВ РУССКОГО ЯЗЫКА
(МОРФОЛОГИЧЕСКАЯ ПАРАДИГМА, АКЦЕНТОЛОГИЯ, ДЕФИНИЦИЯ)
THE PROJECT OF THE INTERNET - VERSION OF THE DICTIONARY
OF LEXICO-SEMANTIC VARIANTS OF RUSSIAN
(A MORPHOLOGICAL PARADIGM, AN ACCENT, A DEFINITION)

Гипертекстовый тезаурус лексико-семантических вариантов /ЛСВ/ русского языка – множество слов, осмысляемых в их конкретном употреблении (реализации лексем в конкретных контекстах). Конструирование гипертекстового тезауруса ЛСВ русского языка в нелинейной форме с учетом реляционных, иерархических и сетевых парадигматических связей посредством реализации синтагматических связей в интерактивном режиме на ЭВМ оцифрованных словарных и текстовых материалов. <http://vault.syktu.ru/www.lsw.ru>

Сыктывкарский государственный университет
Сергей Владимирович Лесников serg@lsw.ru

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 - машинные картотеки письменных и устных источников современного немецкого языка; в Швеции в Гетеборгском университете; Киевский национальный университет (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 др. странах.

В России как наиболее авторитетных в области компьютерной лингвистики следует указать следующие коллективы ученых-лексикографов: МГУ им. М.В. Ломоносова (10 млн. словоупотреблений, корпус текстов русских газет конца 20 века А.А. Поликарпов, О.В. Кукушкина, Б.В. Виноградова, С.О. Савчук, др. 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 госуниверситет (компьютерная антология русского рассказа XX

века, Г.Я. Мартыненко, А.О. Гребенников, Е.А. Козлова, Е.И. Лазаренко, Т.И. Шерстинова); Саратовский госуниверситет (корпус Диалектологических текстов, В.Е. Гольдин); Казанский госуниверситет (компьютерный лингвографический фонд русского языка, К.Р. Галиуллин; электронная коллекция книг 18 века, В.В. Соловьев, А.В. Скоробогатов); Казанский гос. педагог. университет (синтаксический анализатор русских технических текстов, О.А. Невзорова, Н.В. Пяткин); Московский гос. лингвистический университет (информационные технологии и медиа лингвистика, Р.К. Потапова, В.В. Потапов); Нижегородский гос. педагог. университет (корпус текстов литературной критики произведений постмодернизма, Д.В. Гугунава); Новосибирский гос. педагог. университет (В.В. Кромер); Новосибирск. Институт филологии СО РАН (электронный корпус средневековых текстов, А.М. Лаврентьев); Петрозаводский госуниверситет (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", 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).

Тезаурус (гр. *tesauros* - сокровище, запас) – наиболее полный систематизированный свод (набор) данных о какой-либо области знания, позволяющий человеку или компьютеру в ней ориентироваться.

Гипертекстовый тезаурус лексико-семантических вариантов /ЛСВ/ русского языка – множество слов, осмысляемых в их конкретном употреблении (реализации лексем в конкретных контекстах).

Конструирование гипертекстового тезауруса осуществляется по следующему алгоритму: 1) сконструирован систематизированный глоссарий лексико-семантических вариантов русского языка из первоначально неупорядоченного списка словарных статей «Словаря русских словарей /СРС/» на основе

Comment [СВЛ1]: Hypertext тезаурус lexico-semantic variants /LSV/ Russian - set of words, осмысляемых в их concrete use (realization of lexemes in concrete contexts). Designing of hypertext thesaurus LSV of Russian in the nonlinear form in view of relational, hierarchical and network paradigmatic communications (connections) by means of realization of syntagmatic communications (connections) in an interactive mode on the computer dictionary and text materials.

той информации, которая содержится в этом СРС и в предварительных дефинициях; 2) в интерактивном режиме на компьютере формируется система дефиниций ЛСВ на основе тех семантических отношений, которые выявлены в процессе создания систематизированного глоссария ЛСВ; 3) в дальнейшем создается система согласованных глоссариев ЛСВ, отражающих, в допустимых русским языком пределах, выявленные семантические отношения и признаки, зафиксированные в дефинициях.

"Словарь русских словарей" фактически состоит из а) книги Лесникова С.В. Словарь русских словарей: более 3500 источников / Предисловие проф. В.В. Дубичинского. Рецензенты: В.М. Андрущенко, Р.П. Рогожникова, Г.И. Тираспольский. М.: Азбуковник, 2002. 334 с. (создано и опубликовано при финансовой поддержке в виде Российского фонда фундаментальных исследований /РФФИ/ №00-06-80176 «Гипертекстовый генеральный свод лексики русского языка» и проект «Словарь русских словарей» (подготовка и издание монографии, веб-сайт в Internet, электронная версия на компакт-диске) по федеральной программе «Русский язык» (постановление Правительства РФ от 30.12.2000 №1034, протокол №1 от 15.02.2001 заседания конкурсной комиссии и приказ Министерства образования Российской Федерации «О финансировании мероприятий федеральной целевой программы «Русский язык»); б) 10 компакт-дисков (CD№01 Гизаурис 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 и др. программы; в) бесплатной рассылки "Словарь русских словарей" <http://subscribe.ru/catalog/science.humanity.hypervault>; г) каталога 10 компакт-дисков "Словарь русских словарей" (4000 источников, 1000 словарей, справочников, энциклопедий). www.lsw.ru

Можно определить гипертекстовый тезаурус лексико-семантических вариантов /ЛСВ/ русского языка как множество, содержащее все элементы (объекты) какой-либо исследуемой области материального или духовного мира, т.е. в нашем случае – полный систематизированный свод ЛСВ русского языка. 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 размывает общность, даже этой одной конкретной дисциплины. Относительным выходом из парадокса неполноты знаний о мире в форме гипертекстового тезауруса

(согласуясь с неполнотой тезауруса 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). Тем самым возможно рассматривать конструируемый тезаурус ЛСВ как целостный объект, при этом допуская принципиально различные членения, а в зависимости от основания для членения, можно получить качественно и количественно разнообразные множества "первичных" элементов, которые и будут представлять ступени иерархии конструируемого гипертекстового тезауруса ЛСВ русского языка. На данном этапе исследования выделим следующие подуровни тезауруса (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); ж) предметно-семантические системы (литература; избранное, сочинения, собрание, свод; меморема) и системология (теория систем).

Конструирование гипертекстового тезауруса ЛСВ русского языка в нелинейной форме с учетом реляционных, иерархических и сетевых парадигматических связей посредством реализации синтагматических связей в интерактивном режиме на ЭВМ оцифрованных словарных и текстовых материалов и 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.

Пробная Интернет-версия словаря лексико-семантических вариантов русского языка (морфологическая парадигма, акцентология, дефиниция) доступна по адресу <http://vault.syktu.ru/www.lsw.ru>