Информационные технологии и письменное наследие

El'Manuscript-10

Международная научная конференция

Уфа, 28-31 октября 2010 г

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ
ИМ. М. АКМУЛЛЫ»

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Информационные технологии и письменное наследие

El'Manuscript–10

Материалы международной научной конференции
Уфа, 28–31 октября 2010 г.

Уфа, Ижевск
2010

Использование электронных баз данных в исторической лексикографии

О. А. Горбань, Е. М. Шептухина
Волгоградский государственный университет

The article deals with an experience of creation of electronic databases in historical lexicography. The authors consider the direct and reverse indexes of the Old Church Slavic and Old Russian languages.

Стремительное развитие в последние десятилетия компьютерной лингвистики и такой ее области, как компьютерная лексикография, представляет новые возможности сбора, хранения, обработки языкового материала, создания на основе этого традиционных «бумажных» и электронных словарей. Теоретические и практические вопросы компьютерной лингвистики, в том числе лексикографии, разрабатываются в научных трудах отечественных и зарубежных ученых (Ю. В. Рождественского, Р. Г. Пиотровского, А. С. Герда, А. Н. Баранова, Ю. Н. Караулова, В. В. Морковкина и мн. др.). Имеются определенные достижения в области создания электронных словарей и глоссариев, в основном современных языков. Что касается использования компьютерных технологий в исторических лингвистических исследованиях, в частности при изучении русского языка, то в этой области можно назвать работы по созданию корпусов древних текстов (напр., [Древнерусские]), электронному изданию памятников письменности (напр., [Путятина]), статистическому моделированию текстов [Герд, Федер, 2003], созданию электронных словарей древних языков и др. Из продуктов исторической компьютерной лексикографии русского языка в основном мы имеем электронные версии уже изданных типографским способом словарей, например, словаря И. И. Срезневского (на сайте http://imwerden.de), Словаря русского языка XVIII в. (на сайте http://feb-web.ru) и др.; о создании цифровой версии Словаря русского языка XI–XVII вв. см. в

[Филиппович, Чернышева, 1999]. Разработка этого направления стоит в ряду актуальных задач теоретической (исторической) и прикладной лингвистики и имеет широкие перспективы.

В докладе представлен опыт создания электронных баз данных как результата многолетних исследований истории русского и славянского глагола, проводимых в Волгоградском государственном университете (в лаборатории «Глагол», в НИИ истории русского языка).

Изучение глагольной лексемы основывается на комплексном подходе, который позволяет рассматривать ее семантическую структуру как единство взаимодействующих разноуровневых значений, организованных в пределах отдельного слова определенным способом в соответствии с системой данного языка и с закономерностями функционирования этой системы в речи (тексте) [Лопушанская, 1988, с. 5]. Учитывается лексическая, словообразовательная, формообразовательная, словоизменительная, валентностно-сочетаемостная семантика слова и его функционирующих словоформ. Анализ лексических единиц проводится при разграничении семантической деривации и семантической модуляции как переноса значения, результаты которого обнаруживаются при сопоставлении компонентов семантической структуры слова, сложившейся в системе языка, со смысловой структурой словоформы, функционирующей в тексте.

Комплексное исследование предполагает использование как традиционных научных методов, предусматривающих разноуровневый синхронно-диахронический анализ слов (компонентный, признаковый, контекстуальный, сопоставительный, количественный), так и инновационных программ (фиксация реальных и мнимых величин, моделирование их соотношения в рамках контекста, применение новых компьютерных технологий).

В 1998 – 2009 гг. НИИ ИРЯ ВолГУ проводил исследования в рамках научных проектов, поддержанных общероссийскими, региональными и внутривузовскими грантами: «История русского глагола» (грант РГНФ), «Матричная реконструкция семантической структуры русских глаголов XVIII–XX вв., отражающая взаимодействие литературного языка и нижневолжских диалектов» (грант РФФИ совместно с Администрацией

Волгоградской области), «Изменение смысловой доминанты языкового сознания древних русичей (по материалам летописных сводов XI–XVII вв.)» (грант РГНФ), грант ВолГУ; проектов, поддержанных ФЦП «Русский язык» Минобразования РФ «Подготовка нового поколения учебных пособий по русскому языку» (2000 г.), «Лексический состав русского языка XI–XIV вв. в словаре и тексте» (2001 г.), «Развитие семантико-грамматических классов русского глагола XI–XVII вв. как опосредованное отражение структуры сознания носителей языка» (2003 г.); фундаментального исследования по единому заказу-наряду «Теоретические основы изучения русского глагола: история и современное состояние» (ГБ-1.5.01).

В ходе выполнения научных проектов, а также в предшествовавших исследованиях на основе изложенных выше теоретических принципов была разработана шкала признаков, по которой производилась обработка языкового материала, и созданы электронные базы данных, отражающие лексический состав старославянского, церковнославянского, древнерусского языков. Отдельные лексико-семантические группы глаголов занесены в базы данных с учетом их функционирования в контексте. Запись и обработка языкового материала осуществлялась с использованием системы управления базами данных Paradox for Windows.

Данные организованы в виде таблиц матричного типа, строки которых занимаются словами или словоформами с их характеристиками, столбцы содержат такие элементы, как «заголовочное слово (словоформа)» и параметры в соответствии с разработанной шкалой; количество параметров в разных базах варьируется. Так, при описании смысловой структуры глаголов движения в контекстах определяются 4 лексико-семантических параметра — 'среда', 'средство', 'способ', 'интенсивность' перемещения, а также лексико-грамматические, морфологические и другие признаки. Параметр 'среда перемещения' (интегральная сема) принимает значения 'перемещение по твердой поверхности', 'по воде', 'по воздуху' (дифференциальные семы) и т.д. СУБД Paradox позволяет по-разному организовать массив фактов (например, в прямом и обратном алфавитном порядке), выполнять выборки единиц по тем или иным признакам, в частности по морфемам (глаголы с приставкой при-, существи-

тельные с суффиксом -ств и т.д.), устанавливать соотношения величин и т.д. Все это расширяет возможности исследовательской работы и может быть отражено в лексикографических изданиях разного типа.

На данный момент результаты разработок воплощены в изданных словниках к словарям старославянского, церковнославянского, древнерусского языков. Наиболее простую структуру имеют Прямой и обратный словники к Словарю старославянского языка [Лопушанская, Горбань, 1997], содержащие перечни заголовочных слов основных словарных статей названного словаря. Словники к Словарю-индексу русской редакции древнеболгарского языка [Лопушанская, Горбань, 2001] дополнены сведениями о наличии лексем в словарях И.И.Срезневского, Л.Садник и Р.Айцетмюллера, в словаре Чехословацкой академии наук; соответствующая информация дается после значка ⊙, например: мати ⊙ Ср., SA., Slov.; съпросити ⊙ Ср.; страстоносъ ⊙. Здесь использована индексация, проведенная составителями Словаря-индекса.

В рамках проекта «Лексический состав русского языка XI–XIV вв. в словаре и тексте», реализованного коллективом НИИ ИРЯ ВолГУ совместно с ИРЯЗ им. В.В.Виноградова РАН, были созданы на основе электронных баз данных Словник-индекс и Обратный словник к Словарю древнерусского языка XI–XIV вв. Словник-индекс представляет собой перечень заголовочных слов словарных статей названного Словаря и является результатом сопоставления его данных со словарем И.И.Срезневского, Словарем русского языка XI–XVII вв., Словарем-индексом русской редакции древнеболгарского языка; индексация произведена составителями Словника. Во 2 томе издания публикуется Обратный словник к Словарю (составители С.П.Лопушанская и Е.М.Шептухина). Систематизация языковых единиц, предлагаемая в Обратном словнике, позволяет установить словообразовательные модели, определить их лексическое наполнение, в частности при исследовании закономерностей суффиксального словообразования разных частей речи.

В настоящее время ведется работа по созданию комментированного Тезауруса древнерусских глагольно-именных словосочетаний, включающего языковые единицы с указанием логи-

ко-семантических отношений между ними. Готовится к изданию справочное издание тезаурусного типа, в котором глагольная лексика систематизирована в соответствии с принадлежностью глаголов к одному из формально-семантических классов с учетом лексико-грамматической семантики. Эти и другие издания позволят решать многие актуальные проблемы истории русского языка.

Список литературы

Герд, Федер, 2003 — Герд А. С., Федер В. Церковнославянские тексты и церковнославянский язык. СПб.: Изд-во С.-Петерб. ун-та, 2003. 212 с.

Древнерусские — Древнерусские берестяные грамоты [Электронный ресурс] // Рукописные памятники древней Руси: [сайт]. URL: http://gramoty.ru (дата обращения: 19.09.2010).

Лопушанская, 1988 — Лопушанская С. П. Изменение семантической структуры русских бесприставочных глаголов движения в процессе модуляции // Русский глагол (в сопоставительном освещении) / Волгогр. гос. ун-т. Волгоград: Изд-во ВПИ, 1988. С. 5–19.

Лопушанская, Горбань, 1997 — Лопушанская С. П., Горбань О. А. Прямой и обратный словник к Старославянскому словарю (по рукописям X–XI веков) / под ред. Р. М. Цейтлин, Р. Вечерки и Э. Благовой. М.: Русский язык, 1994. 842 с. — Волгоград: Изд-во Волгоград. гос. ун-та, 1997. 178 с.

Лопушанская, Горбань, 2001 — Лопушанская С. П., Горбань О. А. Прямой и обратный словник к Словарю-индексу русской редакции древнеболгарского языка конца XI – начала XII в. / под ред. Имре Х. Тота: в 3 т. Сегед, 1989–1995. — Волгоград: Изд-во Волгоград. гос. ун-та, 2001. 92 с.

Путятина — Путятина Минея [Электронный ресурс] // Манускрипт: славянское письменное наследие: [сайт]. [2004]. URL: http://io.udsu.ru/ptm (дата обращения: 19.09.2010).

Словник-индекс и Обратный словник к Словарю древнерусского языка (XI–XIV вв.). [В 10 томах. М., 1988 – 2001 и след.] : в 2 т. Т. 1: Словник-индекс; Т. 2: Лопушанская С. П., Шептухина Е. М. Обратный словник. Москва; Волгоград: Издательство Волгоград. гос. ун-та, 2002. 450 с.; 292 с.

Филиппович, Чернышева, 1999 — Филиппович Ю., Чернышева М. Историческая компьютерная лексикография — terra incognita в компьютерном мире // Компьютерра. 1999. № 45. URL: http://offline.computerra.ru/1999/323/3379 (дата обращения: 19.09.2010).

Мультимедийный корпус русского языка: опыт создания и использования[1]

Е. А. Гришина, С. О. Савчук
Институт русского языка им. В. В. Виноградова РАН, Москва

The paper introduces the Multimodal Russian Corpus (MURCO) created in the framework of the Russian National Corpus (RNC). MURCO provides the users with a great amount of phonetic, orthoepic, and intonational information related to Russian. Moreover, the deeply annotated part of MURCO contains data on the Russian gesticulation, speech act system, the types of vocal gestures and interjections in Russian etc. The total structure of MURCO, the types of annotation and the possibilities of use of them in studying and teaching Russian are described.

Мультимедийный корпус русского языка — это электронный ресурс, предназначенный для изучения звучащей речи, «погруженной» в обстоятельства ее произнесения. Основу корпуса составляют видео- и аудиозаписи текстов, выровненные с их расшифровками, что позволяет исследовать не только языковые единицы, но и речевые действия говорящего в различных ситуациях общения, и его неречевое поведение (мимику, жесты, позы). Подобное представление звучащей речи в виде корпусов для русского языка только начинается (см. [Степанова и др., 2008; Гришина, Савчук, 2008]), а в виде большого общедоступного корпуса производится впервые [Grishina, 2010].

В настоящее время основу корпуса составляют видеоматериалы из отечественных фильмов и аудиозаписи публичной и непубличной устной речи. Технология подготовки материалов для корпуса предполагает расшифровку видео и аудиоматериалов, произведенную с высокой степенью подробности (т. е.

[1] Работа выполнена при поддержке Программы ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей» и РФФИ (грант 10-06-00151-а, 08-06-00371-а).