Список литературы

Арутюнова 2005 — *Арутюнова, Н. Д.* Предложение и его смысл / Н. Д. Арутюнова. — М. : УРСС, 2005.

Моченов и др. 2005 — *Моченов, С. В.* Применение статистических методов для семантического анализа текста / С. В. Моченов, А. М. Бледнов, Ю. А. Луговских. — Ижевск : НИЦ «Регулярная и хаотическая динамика», 2005.

Караулов и др. 1982 — *Караулов, Ю. Н.* Русский семантический словарь. Опыт автоматического построения тезауруса: от понятия к слову / Ю. Н. Караулов, В. И. Молчанов, В. А. Афанасьев, Н. В. Михалев ; под ред. С. Г. Бархударова. — М. : Наука, 1982.

Рубашкин и др. 1998 — *Рубашкин, В. Ш.* Семантический (концептуальный) словарь для информационных технологий. Ч. 1 / В. Ш. Рубашкин, Д. Г. Лахути // НТИ. — Сер. 2. — 1998. — № 1. — С. 19–24.

Сокирко и др. 2005 — *Сокирко, А. Г.* Проект ДИАЛИНГ, COM-объект Goldrml / А. Г. Сокирко, Д. В. Панкратов. — М. : Диалог, 2005.

Финн 1999 — Финн В. К. О роли машинного обучения в интеллектуальных системах // НТИ. Сер. 2. 1999. № 12. — С. 1–3.

Salton 1989 — G. Salton. Automatic Text Processing. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

Salton et al. 1994 — G. Salton, J. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. Communications of the ACM, 37(2), February 1994.

Todd et al. — Todd A. Letsche and Michael W. Berry. Large-Scale Information Retrieval with Latent Semantic Indexing. URL: http://www.cs.utk.edu/ ~berry/sc95/sc95.html.

Проблемы описания и вопросы моделирования семантики слова в базах данных

И. М. Некипелова
Ижевский государственный технический университет, Россия

This article is devoted to the development in the field of modeling of the description of the word lexical value in the information retrieval system "Manuscript". This aspect is connected with the problem of use of the system for the implementation of linguistic research in the field of vocabulary and semantics and creation of the linguistic search system allowing the user making an exact idea about the word lexical value and its semantic relationships in the language and texts of the ancient manuscripts stored in a database.

В настоящее время актуальными в работе многофункциональных web-модулей транскрипций текстов являются разработки в области моделирования описания лексического значения слова и его семантических отношений. Творческая группа Удмуртского государственного университета и Ижевского государственного технического университета под руководством В. А. Баранова начала работы по созданию автоматизированного лексико-семантического модуля в информационно-поисковой системе «Манускрипт» (далее — ИПС «Манускрипт»). Это связано с необходимостью использования ИПС для проведения лингвистических исследований в области лексики и семантики и с необходимостью разработки лингвистической поисковой системы, позволяющей пользователю иметь точное представление о лексическом значении слова и его семантических связях в языке и текстах древних рукописей, хранящихся в базах данных ИПС «Манускрипт».

Особый интерес вызывают проблемы моделирования семантических и словообразовательных связей слов древних славянских и исходных древнегреческих текстов, поиск соответствий, хранение словообразовательных (морфемных и семантических) связей в базах данных и их использование. Практическое использование ана-

логичных морфологических свойств уже применяется в модуле выборок и запросов в системе «Манускрипт».

Нас интересуют типы лексического описания как основа моделирования лексических значений и семантики слова в базах данных. Для всех слов всех текстов, содержащихся в базах данных ИПС «Манускрипт», разработчиками программы создан модуль лингвистических словарей, которые все время пополняются новыми единицами — элементами словаря с уникальными значениями. «Словари — компоненты базы данных, содержащие информацию о форме, свойствах и значениях словарных единиц» [Баранов 2006: 5]. Модуль лингвистических словарей ИПС «Манускрипт» содержит словари с разными лингвистическими характеристиками связей и признаков слова. Этот модуль взаимодействует с единой базой данных, в которой хранятся иерархически структурированные данные.

Проблема хранения данных решается следующим способом: сведения о слове и его значении будут привязаны к элементам словаря, а не текста, при этом одному элементу может быть приписано несколько различных значений, любую единицу/словоформу/словосочетание можно также привязать к нескольким единицам словаря. Значение слова, его связи и признаки закреплены за основами. В электронной полнотекстовой библиотеке будут использованы данные различных исторических словарей для описания и лексико-семантического анализа языковых данных. Семантические отношения описываются связями, образующими сеть. «Связи — информация об отношениях между объектами текстов/рукописей и их единицами, а также о свойствах и значениях отношений» [Баранов 2006: 5]. Здесь важно подчеркнуть, что лексико-семантические отношения являются одной из самых сложных сетевых структур. Модель базы данных «Манускрипт» позволяет хранить сложные сетевые структуры, с помощью которых можно описать любые имеющиеся в тексте/рукописи объекты, их связи и оперировать ими. Устанавливаются связи единиц рукописей с единицами словаря, а единицам словаря присваиваются свойства и значения, которыми они обладают, при этом объекту лексико-семантических исследований, существующему во многих рукописях или текстах, специальный модуль (например, редактор OldEd) обеспечивает присвоение значения не

как объекту текста, а как объекту словаря. В конечном итоге можно связать любую словоформу и ее инвариант, хранящийся в соответствующем словаре.

Лингвистическим значением обладают различные по объему языковые единицы — от знака до словосочетания (знак, словоформа, фразеологизм, словосочетание).

Лексическим значением также обладают различные по объему языковые единицы — от слова до фразеологизма (слово, речевая формула, фразеологизм). «Все типы значений понимаются как дополнительные друг к другу, то есть как части (стороны, аспекты) целого» [Никитин 1997: 51]. Это связано с трудностью описания семантики слова в истории русского языка. Во-первых, слово и его связи дошли до современности лишь в определенном контексте, чаще в различных контекстах, однако и при наличии контекстов с разнообразными связями слова, мы не владеем информацией обо всех существовавших на то время семантических отношениях слова, то есть его языковых связях и характеристиках: у нас нет полного списка значений слов, формул, устойчивых выражений и т.п. Восстановлением этих связей занимаются многие ученые, но не всегда их мнения совпадают при интерпретации древнего текста. Во-вторых, при интерпретации связей слова, употребленных в древних текстах, мы не всегда можем говорить об адекватности подобной интерпретации, так как мы при описании неизбежно обращаемся к связям слова в современном языке, которых могло и не быть у слова на более ранних этапах развития языка. Иначе говоря, мы не в состоянии мыслить как древние русичи, а исследуем языковые единицы, опираясь лишь на современное мышление. Основную сложность изучения древнерусских текстов сформулировал В.А. Баранов: «К сожалению, до сих пор мы не всегда можем быть уверены в том, что наше понимание и интерпретация древнерусских текстов с точки зрения синтагматических связей, грамматического состава и семантики адекватны пониманию текста древними книжниками» [Баранов 2003: 16]. Особую сложность представляет исследование семантики текстов, поскольку «грамматическая система русского языка еще только формируется и как в частностях, так и в целом представляет иную, во многом отличающуюся от современной, систему» [Там же].

В настоящее время лексико-семантическая модель строится в виде сети. Редактор OldEd дает возможность установления и просмотра связей между единицами. При описании значения слова нас интересуют такие его характеристики, как содержание, структура, системные связи и т. п. Поскольку значения — это понятия, связанные со знаком, то понятия становятся семантическими единицами — значениями или частями значений (семами). Совокупности значений образуют семантическую систему языка (систему значений).

Первым шагом при моделировании связей и признаков слова является описание лингвистической типологии значений слова. Лингвистическая типология значений непосредственно связывает их со способом языкового выражения слов. «По сути дела, лингвистическая типология значения не имеет прямого отношения к содержанию и характеру выражаемого значения, а характеризует его по уровневой природе, выражающей его языковой единицы» [Никитин 1997: 67]. Лингвистическая типология значений прямо связывает значение со способом, характером его языкового выражения. Основными категориями лингвистической типологии значений являются значения грамматические и номинативные, а также морфологические, синтаксические и словообразовательные (как разновидности грамматических значений), лексические, фразеологические, словосочетательные (как разновидности номинативных значений). В основе различий лингвистических типов значения лежат различия в уровневой, или стратификационной, природе языковых единиц.

Вторым шагом при лексическом описании слова является определение лексического значения слова, при котором необходимо давать прежде всего два основных типа описания значения слова: энциклопедическое и толковое, которые берутся из различных исторических словарей. Там, где возможно, будет дана этимология слова по результатам работы с этимологическими словарями.

Однако лексическое значение может быть отражением простого признака и этим исчерпываться. В таком случае оно имеет простую структуру слов, не разложимую на семантические признаки. Такие слова не имеют дефиниций в толковых словарях и могут быть истолкованы только косвенно — через синонимы или через употребление. Полный перечень этих слов до сих пор в научных

исследованиях не приведен. В этом случае следует давать толкование слова через синонимы, антонимы, отсылочное толкование и толкование через употребление слова. Во многих случаях это будет осуществлено с помощью комментирующей и справочной информации.

Не для каждого слова будут работать все типы значений. Следует разработать и описать правила ввода типов значения каждого слова, хранящегося в словаре. При этом надо дать однозначные значения, а значения должны быть сгруппированы в свойства.

При синонимическом и антонимическом описании слова предполагается связать соответствующие синонимы и антонимы, при необходимости — дать комментарий. Таким образом, должен быть сформированы полные синонимический и антонимический словарь и даны комментарии о контекстных синонимах и антонимах.

Отсылочное толкование должно быть построено по принципу связей слова с другими: отсылка к гипериониму, отсылка к производящему слову (с учетом деэтимологизированной связи) и пр. Именно здесь будут описаны все деривационные отношения: связь между омонимами, где значение производного слова должно быть описано через значение производящего слова. Семантические производные также должны быть разграничены по типам (метафора, метонимия, синекдоха и пр.). Таким образом, следует указывать тип отношений: безморфемный или морфемный.

При моделировании семантики слова в базах данных описание семантических изменений отражает деривационные связи слов, явления омонимии и другие словообразовательные связи семантических и морфемных производных дериватов с производящими словами.

При толковании слова важным является определение контекстных значений слова. Следует разработать точные критерии разграничения языкового и контекстного значений. Значение слова должно быть дано в абсолютном употреблении и в системных отношениях с другими словами. Словарная единица — одна, именно она обладает значением и семантическими отношениями, которые распространяются на все словоформы. При описании фразеологизмов и речевых формул должен учитываться прямой, обратный и инверсированный порядок слов, полная, усеченная и дополненная структура.

Третьим шагом поиска является описание различных дополнительных стилистических и коннотативных значений. Если несколько текстовых единиц обладают единой коннотацией, то они должны относиться к одной группе. Кроме того, у слова может быть несколько коннотаций.

Таким образом, если пользователя интересует лексическое значение слова, а далее и его лексико-семантические связи, то на первом шаге (шаг 1) следует выбрать «лексическое».

С одной стороны, в этом модуле возможности запроса достаточно точно определены и пользователь не может их изменить или расширить, с другой стороны, здесь есть свободные области запроса. Границы запроса можно оставлять достаточно широкими, а можно конкретизировать в предложенных границах запросов модуля.

Разрабатываемая лексико-семантическая система поиска позволяет самостоятельно выбирать текст, материал и область/тип лингвистического анализа с целью получения полного лексико-семантического описания слова (значение, структуру, связи, признаки, лексико-семантические характеристики слова).

Анализ материала может быть осуществлен в несколько шагов, при этом пользователь может сам выбирать, на каком шаге он хочет остановиться и отсечь отбор данных, подробное описание которых не представляет для него интереса. Так, если пользователь хочет установить словообразовательные связи между омонимами (производящее → производное), то он последовательно должен указать/выбрать следующие шаги (см. приложение): Шаг 1. Лингвистическая типология значений: «номинативное» → «лексическое»; Шаг 2. Типы описания значений: «отсылочное» → «к производящему слову» → «безморфемная связь». Далее, если его интересует только факт семантической деривации, он может остановиться и не продолжать уточнение типов семантической деривации, если же полученные результаты необходимо уточнить, то он может выбрать один из типов семантической деривации: метафору, метонимию или/и синекдоху. В первом случае в качестве результата запроса пользователю будут показаны все деривационные безморфемные связи между омонимами, во втором случае — только те связи, которые он запрашивал (метафорические, или метонимические, или синекдохичные).

Моделирование семантики слова первоначально будет представлено на материале Триодей XII–XIII веков. Результатом работы должен стать полный Словарь лексико-семантических отношений слов по всем рукописям, хранящимся в базе данных ИПС «Манускрипт», поскольку «модуль должен обеспечить использование единых словарей при фрагментировании и разборе текстов и рукописей» [Баранов 2006: 9].

Использование системы должно позволить за короткий промежуток времени получать материал для лингвистических и лингвотекстологических исследований, проводить его полный и точный лексико-семантический анализ, что позволит исследователю выйти за рамки анализа одного текста и/или текстов одного времени и сделать значимые фундаментальные выводы в области исторической лексикологии и словообразования.

Приложение

Структура описания значения слова
Типы единиц: знак, словоформа, речевая формула, фразеологизм
Шаг 1. Лингвистическая типология значений

- Грамматическое
  - морфологическое
  - синтаксическое
  - словообразовательное
- Номинативное
  - **лексическое**
  - фразеологическое
    - порядок слов
      - прямой
      - обратный
      - инверсированный
    - структура
      - полная
      - усеченная
      - дополненная/расширенная
    - степень слитности
      - речевая формула
      - фразеологизм
  - словосочетательное
    - лексическая валентность
      - семантическая валентность

грамматическая валентность
Шаг 2. Типы описания значений:
Энциклопедическое
Толковое
языковое
контекстное
Этимологическое
Отсылочное
к гипероним
к производящему слову
морфемная связь
безморфемная связь
метафора
метонимия
синекдоха
к эквиваленту в языке-источнике
Синонимическое
полные синонимы
частичные синонимы
контекстуальные синонимы
стилистические синонимы
Антонимическое
языковые антонимы
контекстуальные антонимы
стилистические антонимы
Шаг. 3. Дополнительные значения
Стиль
Коннотации

Список литературы

Баранов 2003 — *Баранов, В. А.* Формирование определительных категорий в истории русского языка / В.А.Баранов. — Казань: Изд-во Казанского гос. ун-та, 2003. — 390 с.

Баранов 2006 — *Баранов, В. А.* От оцифрованных коллекций средневековых рукописей к электронным многофункциональным интернет-библиотекам / В. А. Баранов // [данный сборник] — С. 3–9.

Никитин 1996 — *Никитин, М. В.* Курс лингвистической семантики : учеб. пос. для студ., асп. и препод. лингв. дисциплин в школах, лицеях, колледжах и вузах / М. В. Никитин. — СПб. : Науч. центр проблем диалога, 1996. — 760 с.

Технологии и проблемы кодирования транскрипции в русской фонетике на основе различных алфавитных систем

О. В. Овчинников
Ижевский государственный технический университет, Россия

Drawing up transcribitional binary latin-cyrillic system on the basis of coding Unicode according to the requirements of IPA (International Phonetic Alphabet): 1) revealing of signed unit of the transcribitional systems designated above alphabets; 2) comparison of the revealed pairs marks to coding Unicode; 3) recognition and completion of missing transcribitional pairs token from system IPA with adequate (exact or close) value; 4) "binding" of the revealed pairs transcribitional symbols under coding Unicode to one code value.

В настоящее время для записи звучащей речи используется несколько систем знаков на основе разных алфавитов. Для русского языка — это кириллица и латиница. Однако до сих пор не решен вопрос адекватной кодировки/декодировки речи в компьютерных системах. Кроме того, желание исследователей применить наборную транскрипцию для широкого круга задач с использованием информационных систем приводит к постоянному изменению и/или дополнению существующих перечней знаков. Поэтому задача *адекватного* перевода знаков одной транскрипционной системы в другую полностью не решена.

Целью нашей работы является составление транскрипционной бинарной латинско-кириллической системы на основе кодировки Unicode в соответствии с требованиями IPA (International Phonetic Alphabet).

Задачи: 1) выявление единства транскрипционных систем на основе латинского и кирилловского алфавитов; 2) сопоставление выявленных пар знаков с кодировкой Unicode; 3) выявление недостающих транскрипционных знакопар в системе IPA с адекватным (точным либо близким) значением; 4) «привязка» выявлен-