Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике[1]

Automatic analysis of terminology in the Russian text corpus on corpus linguistics

Митрофанова О. А. (alkonost-om@yandex.ru)
Санкт-Петербургский государственный университет (СПбГУ)

Захаров В. П. (vz1311@yandex.ru)
Санкт-Петербургский государственный университет (СПбГУ),
Институт лингвистических исследований РАН (ИЛИ РАН)

В докладе рассматриваются результаты анализа русскоязычной терминологии корпусной лингвистики, полученные при совмещении ручной и автоматической обработки специального корпуса текстов. Особое внимание уделяется выявлению однословных и неоднословных терминов, использованию лексико-грамматических шаблонов для описания внутренней структуры терминов, а также терминообразующих контекстов.

1. Постановка проблемы, цели и задачи исследования

В многообразии жанров корпусов текстов особое место занимают корпусы специальных, прежде всего, научных текстов, отражающие знания по отдельным предметным областям. Особенности данных корпусов — наличие жёстких ограничений по типу и тематике текстов, входящих в их состав; формализованность содержания текстов, опирающегося на логико-понятийную схему предметной области; высокая структурированность словаря текстов за счёт насыщенности терминами; очевидное влияние научного стиля на лексико-семантические, морфологические, синтаксические параметры текстов в корпусе [Герд 2005]. Сочетание указанных особенностей специальных текстов делает их хотя и сложным, но всё же весьма привлекательным материалом для исследования. Многие проблемы, возникающие при работе со специальными корпусами текстов, не имеют очевидных и однозначных решений. Таковы вопросы о том, что считать термином той или иной области знаний, как описать, представить значения и связи терминов в терминосистеме, как разработать специальный корпус текстов, как выделить термины из текстов в таком корпусе и др. Следует подчеркнуть, что полное решение данных вопросов выходит за рамки нашего исследования; в процессе работы с терминологией мы используем нестрогое понимание термина как лексической единицы, характерной для некоего текста или множества текстов.

Результаты анализа корпусов текстов, сформированных для отдельных предметных областей, имеют высокую прикладную ценность. Специальные корпусы текстов и извлечённые из них данные востребованы как в научно-технической лексикографии (при составление терминологических словарей, классификаторов, рубрикаторов), так и в сфере автоматической обработки текстов (при автоматическом индексировании и реферировании документов, автоматической классификации и кластеризации документов, в информационном поиске и машинном переводе). На основе специальных корпусов текстов создаются и пополняются терминологические базы и банки данных, терминологические тезаурусы, формальные онтологии для отдельных предметных областей, многоязычные терминологические ресурсы.

Особенно важны исследования специальных корпусов текстов для развивающихся областей знаний, и одной из таких областей является сама кор-

[1] Работа выполнена при частичной финансовой поддержке гранта РГНФ (проект номер 07-04-00161а).

пусная лингвистика. Существуют системные описания терминологии корпусной лингвистики для английского [Baker et al. 2006], а также для ряда других языков, в том числе и славянских: см., например, соответствующий раздел в терминологической базе данных для словацкого языка, разрабатываемой в Институте языкознания Л. Штура (Братислава, Словакия) (URL: https://data.juls.savba.sk/std/) [Levická 2007; Šimková 2006]. Однако в русскоязычных терминологических ресурсах данная предметная область до недавнего времени не была представлена.

С 2002 г. на кафедре математической лингвистики СПбГУ и в ИЛИ РАН осуществляется проект, **целью** которого является создание корпуса русскоязычных текстов по корпусной лингвистике и разработка лингвистических ресурсов на основе данного корпуса. В рамках проекта проводится многоаспектное исследование содержания и структуры текстов в корпусе, что предполагает решение ряда **задач**, среди которых

- извлечение, анализ и систематизация терминологии корпусной лингвистики,
- классификация терминов в корпусе,
- разработка формальной онтологии по корпусной лингвистике,
- тематическая рубрикация текстов в корпусе,
- подготовка данных для компьютерного тезауруса по корпусной лингвистике.

Отдельные результаты работы, полученные к настоящему времени, освещены в ряде публикаций: см., в частности, [Виноградова, Митрофанова, Паничева 2007; Виноградова, Митрофанова 2008; Mitrofanova et al. 2007]. В данной статье обсуждается один из аспектов данного проекта, а именно, проблема автоматизации извлечения терминов, анализа и систематизации терминологии корпусной лингвистики.

2. Исходные лингвистические данные

В состав русскоязычного корпуса текстов по корпусной лингвистике входят тексты различной тематики, отражающие широкий спектр проблем корпусной лингвистики: определение корпусной лингвистики как особой области научной деятельности, противопоставление её другим направлениям лингвистики и языковой инженерии; определение корпуса в соотнесённости с другими типами лингвистических данных; различные аспекты создания и использования корпусов; процедуры, выполняемые при работе с корпусом (разметка, типы разметки, поиск в корпусе); типология корпусов; корпусы текстов с позиций разработчиков и пользователей; взаимодействие корпусов и корпусориентированных лингвистических ресурсов и пр. Ядро корпуса составляют материалы научных конференций по корпусной лингвистике [КЛ и ЛБД 2002, КЛ 2004, КЛ 2006, КЛ 2008], отдельные статьи, учебные пособия, монографии и другие научные материалы. Корпус периодически пополняется новыми документами. Материалы корпуса хранятся в текстовом формате, наряду с этим у разработчиков корпуса существует доступ к файлам с оригинал-макетами. В ходе подготовки текстов статей к размещению в корпусе производится 1) графематический анализ, направленный на выделение и удаление нетекстовых элементов (таблиц, рисунков, формул, гиперссылок, числовых данных и пр.) и иноязычных вкраплений, 2) морфологический анализ (лемматизация, полная морфологическая разметка), 3) метаразметка, которая предполагает фиксацию основных параметров каждой статьи в её паспорте. Наряду с библиографическим описанием эксперты включают в число параметров статьи и наборы из 10 выделяемых вручную терминов-дескрипторов, позволяющих диагностировать тематическую принадлежность текста и проверить данные автоматического анализа. Например:

Текст:

И. С. Николаев, А. С. Герд, И. В. Азарова. «Корпус данных в проекте “Комплексная модель формирования культурного ландшафта и историко-культурной зоны Ингерманландии на Северо-Западе России по данным топонимики”» (*КЛ 2006*).

Набор терминов-дескрипторов:

[*данные, источник, картотека, корпус, культурный, ландшафт, поиск, словарь, топоним, топонимический*]

При формировании наборов терминов-дескрипторов учитывались не только частотность терминов в тексте, но и их содержательный вес. Термины-дескрипторы представлены в нормализованном виде: в наборе присутствует лемма, которая соотносится со входящими в текст словоформами, например: **корпус** (*корпус, корпуса, корпусу, корпусом, корпусе, корпусы, корпусов, корпусам, корпусами, корпусах*) и пр.

Связи терминов-дескрипторов в текстах корпуса исследовались с помощью инструмента автоматической классификации лексики (АКЛ) [Виноградова, Митрофанова, Паничева 2007]. Основным принципом АКЛ является возможность определения содержательной близости лексических единиц при сопоставлении их синтагматических свойств (иначе говоря, их сочетаемости с другими элементами контекста, дистрибуции). Программа АКЛ предусматривает предварительную обработку текстов, представление множества контекстов употребления исследуемых лексем как точек или векторов дистрибуций в N-мерном пространстве, вычисление семантических расстояний между исследуемыми лексемами, кластерный анализ, при котором используются дан-

ные о семантических расстояниях. Чем ближе синтагматические свойства лексем (а стало быть, чем ближе их значения), тем меньше расстояние между векторами их дистрибуций и тем больше вероятность их объединения в один кластер. Сформированные таким образом кластеры лексем допускают дальнейшую лингвистическую интерпретацию. При работе с текстами корпуса по корпусной лингвистике процедуры АКЛ производились в двух режимах: структурирование терминов-дескрипторов в наборах и выявление классов условной эквивалентности для каждого из терминов-дескрипторов.

В ходе экспериментов производилась иерархическая кластеризация терминов-дескрипторов в наборах для каждой из статей в корпусе; в качестве меры расстояния использовался косинус угла между векторами дистрибуций (*Cos*). Результаты кластеризации выводятся в виде многоуровневого списка слов в виде скобочной записи, которая отражает последовательность объединения терминов-дескрипторов в кластеры. Наряду с этим пользователь получает данные о частотности исследуемых лексем в обрабатываемом тексте, а также значения расстояний во всевозможных парах лексем из анализируемого набора. Например:

Текст:
Е. Л. Алексеева, А.М. Лаврентьев, И. В. Азарова, Л. А. Захарова «Разметка корпуса древнерусских агиографических текстов» (*КЛ 2004*)

Кластерная структура набора терминов-дескрипторов:
[*корпус*, *разметка*] *Cos* = 0,375

[*агиографический*, *русский*] *Cos* = 0,284

[*житие*, *текст*] *Cos* = 0,277

[[*агиографический*, *русский*] [*житие*, *текст*]] *Cos* = 0,259

[[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]] *Cos* = 0,251

[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *Cos* = 0,219

[[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *электронный*] *Cos* = 0,258

[*рукопись* [[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *электронный*]] *Cos* = 0,171

[*словоформа* [*рукопись* [[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *электронный*]]] *Cos* = 0,138

Абсолютные частоты терминов-дескрипторов:
агиографический (*f* = 4), *житие* (*f* = 13), *русский* (*f* = 7), *текст* (*f* = 47), *корпус* (*f* = 8), *электронный* (*f* = 8), *рукопись* (*f* = 15), *словоформа* (*f* = 15), *представление* (*f* = 7), *разметка* (*f* = 5)

С помощью программы АКЛ для каждого из терминов-дескрипторов в наборах производится автоматическое формирование классов условной эквивалентности, включающих слова с близкой дистрибуцией в тексте. Близость дистрибуции также оценивается на основе значений *Cos*. Например:

Текст:
В. П. Захаров. Корпусная лингвистика *(Захаров 2005)*

Классы условной эквивалентности термина-дескриптора *разметка* (объем классов — 20 слов):

Обработка текста с лемматизацией		**Обработка текста без лемматизации**	
РАЗМЕТКА	Cos	***разметка***	*Cos*
ПРОСОДИЧЕСКИЙ	0,375	*просодическая*	0,362
БОЛЬШИНСТВО	0,288	*фиксирует*	0,285
АНАФОРИЧЕСКИЙ	0,288	*документа*	0,280
ВВОДИТЬСЯ	0,252	*абзацев*	0,280
ДОКУМЕНТ	0,251	*выделение*	0,279
ВЫДЕЛЕНИЕ	0,250	*местоименные*	0,271
МНОЖЕСТВО	0,240	*референтные*	0,270
ИНТОНАЦИЯ	0,226	*предложений*	0,265
РЕФЕРЕНТНЫЙ	0,214	*annotation*	0,255
РЕАЛЬНО	0,213	*анафорическая*	0,254
УДАРЕНИЕ	0,212	*разговорной*	0,253
РАЗ	0,198	*структурная*	0,251
МЕСТОИМЕННЫЙ	0,198	*корпусах*	0,250
ИНОСТРАННЫЙ	0,197	*просодических*	0,230
УПОТРЕБЛЯТЬСЯ	0,196	*интонацию*	0,224
НАЛИЧИЕ	0,185	*частеречная*	0,210
ДОСЛОВНО	0,180	*ударение*	0,207
ОГОВОРКА	0,167	*описывающие*	0,189
ПОВТОР	0,167	*оказаться*	0,168

По-видимому, последовательность формирования кластеров терминов-дескрипторов, а также состав выделенных для них классов условной эквивалентности отражает важнейшие парадигматические и синтагматические связи элементов исследуемых текстов. Тем самым, в процессе создания модели предметной области корпусной лингвистики производится обобщение выявленных связей терминов-дескрипторов до родовидовой иерархии понятий. В целях уточнения характера связей между понятиями, выраженными исследуемыми терминами, была проведена отдельная серия экспериментов. Процедуры отбора и кластеризации дескрипторов, характеризующих корпусную лингвистику, позволяют перейти с терминологического уровня на онтологический и сформировать упорядоченное множество категорий, которые необходимо вклю-

чить в формальную онтологию рассматриваемой области знаний. Формальная онтология по корпусной лингвистике относится к классу терминологических онтологий [Sowa]. В качестве представителей онтологических категорий были отобраны те из терминов-дескрипторов, которые оказались релевантны не только для отдельных текстов, но для предметной области в целом, обладают наибольшей частотой, попадают в ядра полученных кластеров, соответствуют исходным понятиям, выделенным на основе экспертных описаний. Всего было зарегистрировано 335 различных терминов-дескрипторов. Вероятно, такие термины-дескрипторы, как *корпус*, *текст*, *данные*, *разметка*, *тег*, *поиск*, *слово*, *лемма*, *словоформа*, *контекст* и пр. представляют понятийное ядро предметной области.

- *Предметная область «Корпусная лингвистика»*
- *корпус данных*
  - *корпус текстов*
  - *тип корпуса*
    - *работа с корпусом*
      - *разработка корпуса*
        - *отбор данных*
        - *оцифровка данных*
        - *разметка корпуса*
        - *корпус-менеджер*
          - *использование корпуса*
            - *поиск по корпусу*
              - *запрос к корпусу*
                - *терминальная цепочка символов*
                - *регулярное выражение*
                - *лемма*
                - *тег*
              - *результат работы с корпусом*
                - *конкорданс*
                - *контекст*
                - *словоуказатель*
                - *статистика*

Формальная онтология по корпусной лингвистике реализована в онторедакторе Protégé [Виноградова, Митрофанова 2008]. Выше приведены важнейшие категории формальной онтологии, упорядоченные в иерархию.[2] В отдельных полях формальной онтологии даются общепринятые дефиниции терминов-дескрипторов, фиксируются синонимические отношения между терминами-дескрипторами (например, *разметка*, *аннотация*, *аннотирование* и пр.). Кроме того, каждая категория формальной онтологии имеет атрибут *тексты*. Этот атрибут необходим для того, чтобы формальная онтология могла быть использована для тематической рубрикации документов из русскоязычного корпуса текстов по корпусной лингвистике. В качестве экземпляров данного атрибута приведены библиографические сведения о тех статьях из корпуса, в которых встретились термины-дескрипторы, соответствующие онтологическим категориям. Например:

Категория: *алгоритм*
Тексты: *П. Макагонов, М. Александров, А. Гельбух* «Формулы проверки подобия слов с обучением на примерах: построение и применение» (*КЛ 2004*); *К. Р. Пиотровская, Р. Г. Пиотровский, Ю. В. Романов* «Вторая когнитивная революция — инженерная и корпусная лингвистика» (*КЛ и ЛБД 2002*).

Тем самым, применение формальной онтологии предметной области корпусной лингвистики при работе с соответствующим корпусом текстов должно повысить эффективность поиска данных.

С расширением русскоязычного корпуса текстов должно происходить пополнение списка уже зарегистрированных терминов и обновление существующей формальной онтологии, на основе которой в дальнейшем планируется создание тезауруса по корпусной лингвистике. В связи с этим было принято решение изучить возможности частичной автоматизации терминологической работы и затем оптимизировать процедуру обработки документов из корпуса текстов по корпусной лингвистике.

3. Методы и инструменты анализа терминологии

Существует три основных класса методов извлечения терминологии из специального корпуса текстов: лингвистические методы, статистические методы и комбинированные методы.

Лингвистические методы в основном предполагают ручную обработку документов в специальном корпусе текстов, в ходе которой эксперты выявляют выражения, рассматриваемые как предполагаемые однословные термины и терминосочетания. Для выделения терминосочетаний рекомендуется использовать лексико-грамматические шаблоны однословных и неоднословных терминов. Целесообразно также использовать систему фильтров (стоп-словарь) для отсеивания нетерминов.

Применение статистических методов опирается на представление о том, что термины, как правило, это наиболее частотные слова и словосочетания, встречающиеся в специальных текстах и выражающие понятия предметной области. Терминосочетания обычно соотносятся с *n*-граммами (двух-, трех-, четырехчленными сочетаниями), характеризуются высокой степенью устойчивости. В качестве мер, пригодных для оценки устойчивости словосочетаний в специальных текстах, следует упомянуть *MI-score*, *t-score*, *Log-Likelihood*, *C-value*, критерий χ^2 и ряд других.

[2] В рамках данной статьи не ставится задача полного описания иерархии категорий формальной онтологии в силу её объёмности.

Во многих исследованиях, проводимых для русского и других славянских языков (см., например: [Браславский, Соколов 2006, 2007, 2008; Добров и др. 2003; Kupść 2007; Urbańska, Piechociński 2007] и др.) практикуется комбинированный подход, заключающийся в (полу)автоматической обработке специальных корпусов текстов. Комбинированные методы анализа терминологии предполагают совместное использование аппарата лексико-грамматических шаблонов, методов сборки терминосочетаний, системы фильтров, а также статистического аппарата.

Сочетание лингвистических и статистических приемов анализа документов в корпусе применяется в автоматизированной лексикографической среде Alex+ [Сидорова 2008(а), 2008(б)]. Alex+ представляет собой технологический комплекс для создания и поддержки предметно-ориентированных словарей, позволяющий выделять термины и терминосочетания из текстов по лексико-грамматическим шаблонам, получать статистические данные о встречаемости терминов и терминосочетаний в обрабатываемых текстах, автоматически пополнять словарь на основе обучающей выборки. В состав комплекса Alex+ входят модуль морфологического анализа системы Диалинг, модуль сборки терминосочетаний по шаблонам, модуль просмотра конкорданса, модуль тематизации, модуль выявления стоп-слов. Преимущества подготовки словарей в системе Alex+ залючаются в возможности разнообразного наполнения словарей, допускающих включение однословных и неоднословных терминов, в возможности представления нескольких типов данных о терминах (терминообразующие признаки, семантические признаки — соотнесенность с понятиями в иерархии классов, статистические признаки) и др. В Alex+ допускается построение формальной онтологии (или задание иерархии тем) параллельно со словарем, при этом словарь и иерархия тем могут применяться для автоматической классификации текстов. Существует также возможность обработки несловарных словоформ и др. Тем самым, параметры автоматизированной лексикографической среды Alex+ соответствуют целям обсуждаемого исследования, в связи с чем некоторые функции данного комплекса были задействованы при анализе терминологии в русскоязычном корпусе текстов по корпусной лингвистике.

4. Описание однословных и неоднословных терминов с помощью лексико-грамматических шаблонов

В ходе анализа однословных терминов и терминосочетаний были применены лексико-грамматические шаблоны (ср. морфологические шаблоны [Сидорова 2008(а), 2008(б)], лексико-синтаксические шаблоны [Большакова и др. 2007; Васильева 2004; Рабчевский и др. 2008]). Лексико-грамматические шаблоны служат для описания классов языковых выражений. В отдельном лексико-грамматическом шаблоне указываются существенные характеристики множества лексем, которые входят в языко вые выражения, принадлежащие классу, также приводятся возможные морфологические формы лексем и, при возможности, синтаксические условия употребления языкового выражения, построенного в соответствии с шаблоном (например, правила согласования морфологических признаков лексем).

Лексико-грамматические шаблоны были задействованы при выделении однословных и неоднословных терминов в автоматизированной лексикографической среде Alex+ [Сидорова 2008(а), 2008(б)].

Например, в результате обработки текста [Захаров 2005] с последующим отсеиванием стоп-слов (служебных слов, местоимений, числительных и др.), а также слов, не являющихся терминами (например, *миро*), в списке однословных терминов можно обнаружить существительные, прилагательные, глаголы:**N**: *выборка, выдача, данные, грамматика, документ, единица, жанр, запрос, инструмент, классификация, кодирование, лемма, массив, метаданные, метка, морфология, неоднозначность, поиск, пользователь, разметка, репрезентативность, составитель, текст, частота* и др.;

Adj: *автоматизированный, информационно-поисковый, корпусной, корпусный, лингвистический* и др.;

V: *автоматизировать, размечать* и др.

Среди неоднословных терминов обнаружены словосочетания, соответствующие следующим основным лексико-грамматическим шаблонам:

Adj+N: *автоматизированная система, автоматическая обработка / разметка / система, автоматический анализ / режим, анафорическая / морфологическая / семантическая / синтаксическая / структурная / просодическая разметка, совместная встречаемость, программное обеспечение, формальный язык, языковой корпус, языковая единица* и др.;

Adj+N+N: *автоматическая обработка текста, компьютерная база данных, компьютерная модель языка, лингвистический корпус текстов, представление корпуса текстов, формальный язык разметки* и др.;

N+Adj+N: *банк синтаксических структур, массив языковых данных, обработка типовых запросов* и др.;

N+Prep+Adj+N: *корпус с синтаксической разметкой, тексты на естественном языке, тексты на машинном носителе* и др.;

N+Prep+N: *доступ к корпусу, наука о языке, поиск в корпусе, сведения об авторе* и др.;

N+Prep+N+N: *поиск с указанием контекста* и др.;

N+N: *обучение языку, база данных, массив данных / текстов, вид разметки, источник данных, кодирование информации, корпус данных / текстов, модель языка, параметр разметки / кодирования / текста, разметка корпуса / документа / текста, размер корпуса, распознавание речи, тип корпуса / данных / разметки / текста, формат выдачи / данных* и др.;

N+N+N: *вывод результатов поиска, стандарт представления метаданных / данных* и др.;

Adj+Adj+N: *устная разговорная речь* и др.

Справедливо будет отметить, что данные словосочетания различаются не только по степени сложности (двух-, трёх-, четырёхкомпонентные терминосочетания), но также по устойчивости (особенно это касается трёх- и четырёхкомпонентных сочетаний, которые сами по себе содержат однословные термины и двухкомпонентные терминосочетания). Для определения устойчивости сочетаний также необходимо обращаться к статистическим критериям [Браславский, Соколов 2006, 2007, 2008; Добров и др. 2003; Захаров, Хохлова 2008; Чанышев 2008; Khokhlova 2008]. Самый важный вопрос, возникающий при анализе массивов однословных и неоднословных терминов — это вопрос об оценке степени терминологичности рассматриваемых единиц. Один из путей — определение индекса специфичности для данной совокупности текстов [Шайкевич 2003]. Решающее слово, вместе с тем, остаётся за специалистами-терминоведами и — в нашем случае — за экспертами в области корпусной лингвистики.

5. Описание терминообразующих контекстов с помощью лексико-грамматических шаблонов

Расширенные лексико-грамматические шаблоны успешно используются для выявления и описания терминообразующих контекстов. Терминообразующие контексты, как правило, содержат термин и его толкование, синонимы, переводные эквиваленты и т.д., при этом в контексте существуют определенные маркеры, позволяющие опознать сам термин и связанную с ним информацию.

Структура и типовое наполнение контекстов, содержащих толкования терминов, могут быть представлены, например, в следующих лексико-грамматических шаблонах:

NP(term) <*понимать/пониматься*> NP(def):

Под репрезентативностью понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. [Захаров 2005];

NP(def) <*называть/называться/иметь название*> NP(term):

Это кодирование информации ***имеет название*** *метаразметка…* [Захаров 2005];**NP(term) <*заключаться в*> NP(def):**

Разметка (tagging, annotation) ***заключается в*** *приписывании текстам и их компонентам специальных меток (tag, tags): внешних, экстралингвистических (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика; сведения об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое* [Захаров 2005];

NP(term) <*представлять собой*> NP(def):

…устойчивые словосочетания ***представляют собой*** *с семантической точки зрения неделимую смысловую единицу…* [Захаров 2005].

Контексты, выражающие различные отношения между терминами, могут быть обобщены, например, в следующих лексико-грамматических шаблонах:

NP(term) <*, или*> NP(term) (синонимия):

…синтаксического анализа, ***или*** *парсинга…* [Захаров 2005];

NP(term) <являться результатом> NP(term) (отношение «процесс — результат»):

…синтаксическая разметка, ***являющаяся результатом*** *синтаксического анализа, или парсинга (англ. parsing)…* [Захаров 2005];

NP(term) <*обеспечивать*> NP(term) (отношение «объект — назначение»):

…конвертирование размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager), ***обеспечивающей*** *быстрый многоаспектный поиск и статистическую обработку…* [Захаров 2005];**NP(term) <*включать (в себя)*> NP(term)** (количественные, гипонимические, мереологические, импликативные и др. отношения):

количественные отношения: *Корпусы нового поколения* ***включают*** *сотни миллионов слов, поэтому выдвигаются принципы разработки систем, которые бы минимизировали вмешательство человека* [Захаров 2005];

гипонимические отношения: *Метаописание текстов корпуса* ***включает*** *как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ)* [Захаров 2005].

Тем самым, анализ терминообразующих контекстов способствует установлению системных связей терминов в терминосистеме, что позволяет

уточнять состав словника и пополнять блок дефиниций терминологического тезауруса.

В блок дефиниций тезауруса включаются толкования стандартных и авторских терминов, зафиксированные в текстах корпуса (как в экспертных, так и в исследовательских описаниях) или в других источниках энциклопедического характера. Вместе с тем, «готовые» толкования удается подобрать лишь к наиболее распространённым терминам, для остальных необходимо составлять дефиниции, и в подобных случаях обращение к лексико-грамматическим шаблонам также весьма уместно, так как это позволяет сохранить единообразие структуры толкований.

В дальнейшем при решении задач поиска в корпусе текстов и автоматизированного пополнения формальной онтологии возможно использование специализированного языка для записи лексико-грамматических шаблонов, например, языка LSPL (Lexical-Syntactic Pattern Language) [Большакова и др. 2007; Васильева 2004; Рабчевский и др. 2008].

6. Итоги исследования и направления дальнейшей работы

В ходе исследования были оценены возможности различных стратегий автоматизации работ по извлечению и систематизации терминологии из русскоязычного корпуса текстов по корпусной лингвистике.

Применение инструмента АКЛ, реализующего процедуры кластерного анализа в двух режимах, позволило выявить структурную организацию терминов-дескрипторов в корпусе текстов по корпусной лингвистике. Полученные данные легли в основу формальной онтологии предметной области, охватывающей базовые понятия и термины корпусной лингвистики.

Пополнение базового списка терминов и формирование списка терминосочетаний успешно проведено с помощью автоматизированной лексикографической среды Alex+. Проанализированы основные лексико-грамматические шаблоны для однословных и неоднословных терминов, встречающихся в текстах корпуса. Аппарат лексико-грамматических шаблонов также использовался в изучении структуры терминообразующих контекстов.

Результаты, полученные на нынешнем этапе работы, будут использованы при разработке тезауруса по корпусной лингвистике. Данный лингвистический ресурс планируется включить в состав портала знаний по компьютерной лингвистике, создаваемого коллективом российских учёных (Москва, Новосибирск, Санкт-Петербург) [Соколова и др. 2008].

Литература

1. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007. URL: http://www.dialog-21.ru/dialog2007/materials/html/11.htm
2. *Браславский П. И., Соколов Е. А.* Автоматическое извлечение терминологии с использованием поисковых машин интернета // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007. URL: http://www.dialog-21.ru/dialog2007/materials/html/14.htm
3. *Браславский П. И., Соколов Е. А.* Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008. URL: http://www.dialog-21.ru/dialog2008/materials/html/11.htm
4. *Браславский П. И., Соколов Е. А.* Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2006». М.: 2006. URL: http://www.dialog-21.ru/dialog2006/materials/html/Braslavski.htm
5. *Васильева Н. Э.* Шаблоны употреблений терминов и их использование при автоматической обработке научно-технических текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2004». М.: 2004. URL: http://www.dialog-21.ru/Archive/2004/Vasiljeva.htm
6. *Виноградова Н. В., Митрофанова О. А.* Формальная онтология как инструмент систематизации данных в русскоязычном корпусе текстов по корпусной лингвистике // Труды международной конференции «Корпусная лингвистика — 2008». СПб.: 2008.
7. *Виноградова Н. В., Митрофанова О. А., Паничева П. В.* Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике // Труды девятой Всероссийской научной конференции «Электронные

библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2007). Переславль-Залесский: 2007. URL: http://www.rcdl.ru//papers/2007/paper_31_v1.pdf

8. *Герд А. С.* Язык науки и техники как объект лингвистического изучения // А.С. Герд. Прикладная лингвистика. СПб.: 2005.
9. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2003). СПб.: 2003. URL: http://www.cir.ru/docs/ips/publications/2003_rcdl_thes_creation.pdf
10. *Захаров В. П.* Корпусная лингвистика / Учебно-методическое пособие. СПб.: 2005.
11. *Захаров В. П., Хохлова М. В.* Статистический метод выявления коллокаций // Языковая инженерия в поиске смыслов. XI Всероссийская объединенная конференция «Интернет и современное общество». Санкт-Петербург: 2008.
12. *КЛ и ЛБД 2002* — Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб.: 2002.
13. *КЛ 2004* — Труды международной конференции «Корпусная лингвистика — 2004». СПб.: 2004.
14. *КЛ 2006* — Труды международной конференции «Корпусная лингвистика — 2006». СПб.: 2006.
15. *КЛ 2008* — Труды международной конференции «Корпусная лингвистика — 2008». СПб.: 2008.
16. *Рабчевский Е. А., Булатова Г. И., Шарафутдинов И. М.* Формализм записи лексико-синтаксических шаблонов в задаче автоматизации процесса построения онтологий // Труды десятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2008). Дубна: 2008. URL: http://rcdl2008.jinr.ru/pdf/103_106_paper10.pdf
17. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008(а). URL: http://www.dialog-21.ru/dialog2008/materials/html/74.htm
18. *Сидорова Е. А.* Подход к построению предметных словарей по корпусу текстов // Труды международной конференции «Корпусная лингвистика–2008». СПб.: 2008(б).
19. *Соколова Е. Г., Кононенко И. С., Загорулько Ю. А.* Проблемы описания компьютерной лингвистики в виде онтологии для портала знаний // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008. URL: http://www.dialog-21.ru/dialog2008/materials/html/75.htm
20. *Чанышев О. Г.* Автоматическое построение терминологической базы знаний // Труды десятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2008). Дубна: 2008. URL: http://rcdl2008.jinr.ru/pdf/085_092_paper8.pdf
21. *Шайкевич А. Я.* Статистический словарь языка Достоевского. Введение. 2003. URL: http://nature.syktsu.ru/cfrl/dost_cd0/introdw.htm
22. *Backer P., Hardie A., McEnery T.* A Glossary of Corpus Linguistics. Edinburgh University Press: 2006.
23. *Khokhlova M.* Extracting Collocations in Russian: Statistics vs. Dictionary // Proceedings of 9th International Conference on Textual Data Statistical Analysis (JADT 2008). Lyon: 2008.
24. *Kupść A.* Extraction automatique de termes à partir de textes polonais // TALN 2007. Toulouse: 2007. URL: http://llf.linguist.jussieu.fr/llf/Gens/Kupsc/kupsc-taln07.pdf
25. *Levická J.* Terminology and Treminological Activities in the Present-Day Slovakia // Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava: 2007.
26. *Mitrofanova O., Panicheva P., Savitsky V.* Automatic Word Clustering in Russian Texts based on Latent Semantic Analysis // Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava: 2007.
27. *Šimková M.* Výberový slovník termínov z počítačovej a korpusovej lingvistiky. 2006. URL: http://korpus.juls.savba.sk/publications/block1/2006-simkova-vyberovy%20slovnik%20terminov/2006-simkova-vyberovy%20slovnik%20terminov.pdf
28. *Sowa J. F.*Building, Sharing, and Merging Ontologies. URL: http://www.jfsowa.com/ontology/ontoshar.htm
29. *Urbańska D., Piechociński D.* Automatic Term Recognition in Polish Texts // Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava: 2007.