Н.В. Виноградова, О.А. Митрофанова

ФОРМАЛЬНАЯ ОНТОЛОГИЯ КАК ИНСТРУМЕНТ СИСТЕМАТИЗАЦИИ ДАННЫХ В РУССКОЯЗЫЧНОМ КОРПУСЕ ТЕКСТОВ ПО КОРПУСНОЙ ЛИНГВИСТИКЕ

В центре внимания данного исследования находится предметная область (ПО) «Корпусная лингвистика». Проблематика корпусной лингвистики как раздела компьютерной лингвистики связана с разработкой общих принципов построения, с созданием и использованием лингвистических корпусов (т.е. корпусов текстов). Направление «Корпусная лингвистика» возникло сравнительно недавно, структура и наполнение русскоязычной текстовой и терминологической базы корпусной лингвистики находится лишь в стадии формирования. В связи с этим задача систематизации и унификации разрозненных данных, имеющиеся в этой области, представляется актуальной.

Одним из способов систематизации данных является построение формальной онтологии исследуемой ПО. В широком смысле слова, онтология – это формальное представление мира (либо какой-либо ПО). В данной работе под онтологией понимается логико-понятийная система ПО, отражающая иерархические отношения между терминами этой области.

Разработка формальной онтологии ПО «Корпусная лингвистика» производилась в два этапа. На первом этапе была создана предварительная версия формальной онтологии, основанная на имеющихся экспертных описаниях исследуемой ПО. На втором этапе проводилось уточнение и расширение предварительной версии формальной онтологии с учётом результатов обработки текстов из русскоязычного корпуса текстов по корпусной лингвистике.

Одной из первостепенных задач при создании формальной онтологии является отбор важнейших понятий ПО, соответствующих категориям формальной онтологии. Изначально данные понятия определялись на основе экспертных описаний, в которых даётся обзор ПО «Корпусная лингвистика», позволяющий выделить её ключевые идеи и ознакомиться с наиболее распространёнными проблемами[1]. Затем был проведён анализ материалов, которые входят в состав русскоязычного корпуса текстов по корпусной лингвистике, разрабатываемого на кафедре математической лингвистики СПбГУ и в ИЛИ РАН (рук. В.П. Захаров). В состав корпуса 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 корпус-ориентированных лингвистических ресурсов и пр.

Ядро корпуса составляют материалы научных конференций *КЛ и ЛБД 2002*, *КЛ 2004*, *КЛ 2006* [2] и некоторые другие источники. Корпус периодически пополняется новыми документами. В

---

[1] *Захаров В.П.*, Корпусная лингвистика / Учебно-методическое пособие. СПб., 2005; *Леонтьева Н.Н.* Автоматическое понимание текстов: Системы, модели, ресурсы. М., 2006; *Баранов А.Н.* Введение в прикладную лингвистику. М., 2003.

[2] *КЛ и ЛБД 2002* – Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб., 2002; *КЛ 2004* – Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004; *КЛ 2006* – Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006.

настоящее время в состав корпуса входит 105 статей на русском языке объёмом около 175 тыс. словоупотреблений. В корпусе также представлены русскоязычные тезисы докладов объёмом около 25 тыс. словоупотреблений. 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 их ручная обработка (графематический анализ), направленная на выделение нетекстовых элементов (таблиц, рисунков, формул, гиперссылок, числовых данных и пр.) и иноязычных вкраплений, а также 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 релевантных терминов-дескрипторов, позволяющих диагностировать тематическую принадлежность текста. Например:

Статья:

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[1]. Термины-дескрипторы представлены в нормали-

[1] Детально процесс отбора и систематизации терминов описан в работах: *Виноградова Н.В., Митрофанова О.А., Паничева П.В*. Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL–2007). Переславль-Залесский, 2007;

зованном виде: в наборе присутствует лемма, которая противопоставляется входящим в текст словоформам, например: **корпус** (*корпус, корпуса, корпусу, корпусом, корпусе, корпусы, корпусов, корпусам, корпусами, корпусах*) и пр.

Структурирование наборов терминов-дескрипторов осуществлялось с помощью инструмента автоматической классификации лексики (АКЛ), также разрабатываемого на кафедре математической лингвистики СПбГУ[1]. 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, подготовленная П.В. Паничевой на языке Python, 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 данные 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.

Митрофанова О.А., Мухин А.С., Паничева П.В. Автоматическая классификация лексики в русскоязычных текстах на основе латентного семантического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2007». М., 2007.

[1] Технические аспекты разработки инструмента АКЛ заслуживают самостоятельного рассмотрения, краткое описание инструмента содержится, например, в статье: *Митрофанова О.А., Мухин А.С., Паничева П.В*. Автоматическая классификация лексики…

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 с помощью скобочной записи. Наряду с этим пользователь получает данные о частотности исследуемых лексем в обрабатываемом тексте и значения расстояний во всевозможных парах лексем из анализируемого набора. Например:

Статья:

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

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)

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

Очевидно, последовательность формирования кластеров терминов-дескрипторов отражает естественные связи элементов исследуемых текстов, что подтверждается частотными данными и значениями расстояний между парами элементов. Особенностью полученных кластеров является то, что в них зафиксированы как синтагматические (например, [*переводческая, память*]), так и парадигматические связи терминов-дескрипторов (например, [*массив, база, данные*]). Вместе с тем, разграничение этих основополагающих типов связей на уровне текста зачастую затруднено: например, термины *текст* и *корпус*, *слово* и *биграмма* могут находиться в парадигматических отношениях, если интерпретируются как разноплановые текстовые единицы (*текст* ≠ *корпус*, *слово* ≠ *биграмма*), или в синтагматических отношениях, если указывается, что между ними допустимы отношения включения (*текст* ⊃ *корпус*, *слово* ⊃ *биграмма*). Тем самым, в процессе создания модели ПО «Корпусная лингвистика» обобщение выявленных связей терминов-дескрипторов до родо-видовой иерархии понятий производится на достаточно широких основаниях, а сама результирующая иерархия при этом оказывается более насыщенной.

В целях уточнения характера связей между понятиями, выраженными исследуемыми терминами, была проведена серия экспериментов с текстами, для которых наблюдается частичное совпадение наборов дескрипторов. В ряде случаев результаты кластеризации совпадающих терминов-дескрипторов для разных текстов оказались идентичными. Так, обнаружены пары текстов, применительно к которым группы общих для них дескрипторов упорядочиваются единообразно: [*словарь* [*корпус, текст*)]], [*частота* [*корпус, текст*]], [*массив* [*данные* [*корпус, текст*]]]. Безусловный интерес представляют те случаи, когда кластеризация терминов-дескрипторов, разделяемых парой текстов, приводит к несовпадающим результатам. Например, отношения в пятёрке дескрипторов, общих для пары текстов, устанавливаются следую-

щим образом: [*формат* [*разметка* [*поиск* [*текст, корпус*]]]] vs. [*разметка* [[*корпус, текст*] *формат*] [*поиск*]]. Применительно к другой паре текстов их общие дескрипторы также упорядочиваются по-разному: [*поиск* [*слово* [*текст, корпус*]]] vs. [*поиск* [*корпус* [*слово, текст*]]]. Сравнение иерархий терминов-дескрипторов, полученных для разных документов, создаёт почву для их тематической рубрикации. Если результаты экспериментов свидетельствуют о единообразии связей между дескрипторами, можно сделать предположение и о тематическом сходстве текстов. Обратное может указывать на то, что тексты не представляют одно тематическое направление или на то, что в паре тематически близких текстов по-разному расставлены акценты.

Процедуры отбора и кластеризации дескрипторов, характеризующих ПО «Корпусная лингвистика», позволяют перейти с терминологического уровня на онтологический и сформировать упорядоченное множество категорий, которые необходимо включить в формальную онтологию рассматриваемой ПО. В качестве представителей онтологических категорий были отобраны те из терминов-дескрипторов, которые оказались релевантны не только для отдельных текстов, но для ПО в целом, обладают наибольшей частотой, попадают в ядра полученных кластеров, соответствуют исходным понятиям, выделенным на основе экспертных описаний ПО. 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 ПО.

Формальная онтология ПО «Корпусная лингвистика» реализована в онторедакторе Protégé. Ниже приведены важнейшие категории формальной онтологии, упорядоченные в иерархию.[1]

[1] В рамках данной статьи не представляется возможным дать полное описание иерархии категорий формальной онтологии в силу её объёмности.

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

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 ПО «Корпусная лингвистика» при работе с соответствующим корпусом текстов должно повысить эффективность поиска данных.

Перспективные направления развития исследования связаны с разработкой терминологического тезауруса по корпусной лингвистике на основе формальной онтологии, с разработкой инструментов для определения количественных оценок близости текстов, с выявлением основных тематических областей в рамках корпусной лингвистики и с проведением автоматической классификации текстов внутри тематических областей.