Захаров В.П.

Тезаурус по корпусной лингвистике

Санкт-Петербургский государственный университет

vz1311@yandex.ru

Аннотация

- Корпусная лингвистика.
- Терминосистемы.
- Материал.
- Методы.
- Задачи.
- Результаты.
- Использование.

Корпусная лингвистика

- Корпусная лингвистика – направление в лингвистике, занимающееся разработкой общих принципов построения и использования лингвистических корпусов с использованием компьютерных технологий.
- Корпусная лингвистика находится на пересечении задач теоретической и прикладной лингвистики.
- Разные уровни языка…
- Корпусы специальных текстов

Терминосистемы

- Понятие термина
- Системность термина
- Понятие термина в корпусной лингвистике

- **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 корпусориентирован-ных лингвистических ресурсов;
- параллельные корпусы и т.д.
-

Терминосистемы

- Структура термина
  Термины-словосочетания составляют от 60% до 70% специальной лексики
  Наиболее распространенным видом составных терминов в терминолексике (65% от общего числа составных термнив) является двух- или трехкомпонентное атрибутивное именное словосочетание
  наиболее распространенными синтаксическими моделями являются:

  Сущ. + Прил.Р + Сущ.Р – *словарь иностранных слов,*

  Прил. + Прил. + Сущ. – *тепловая импульсная сварка,*

  Прил. + Сущ. + Сущ.Р – *автоматическая обработка текста,*

  Сущ. + Сущ.Р + Сущ.Р – *методы нанесения покрытий*

Предметная область «Корпусная лингвистика: литература

- Баранов А.Н. Введение в прикладную лингвистику. Серия "Новый лингвистический учебник". М.: Эдиториал УРРС. 2001.
- Демьянков В.З. Англо-русские термины по прикладной лингвистике и автоматической переработке текста. Вып. 2. Методы анализа текста // Тетради новых терминов. № 39. -М.: ВЦП, 1982.
- Захаров В.П. Корпусная лингвистика: Учебно-методическое пособие. – СПб.: СПбГУ, 2005. – 48 с.
- Лингвистический энциклопедический словарь. М.: Сов. Энциклопедия, 1990.
- Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. - М., 1978.
- Леонтьева Н.Н. Автоматическое понимание текстов: Системы, модели, ресурсы. М., 2006.

- Прикладное языкознание. Учебник (ред. А.С.Герд). СПб., 1996.
- Языкознание. Информационно-поисковый тезаурус ИНИОН РАН. – М., 2007.

- The Oxford handbook of computational linguistics // Mitkov Ruslan (ed.). N.Y.: Oxford university press, 2003.
- Backer P., Hardie A., McEnery T. A Glossary of Corpus Linguistics. Edinburgh University Press: 2006.
- Šimková M. Výberový slovník termínov z počítačovej a korpusovej lingvistiky. 2006. URL: http://korpus.juls.savba.sk/publications/block1/2006-simkova-vyberovy slovnik terminov/2006-simkova-vyberovy slovnik terminov.pdf

Глоссарий по корпусной лингвистике

- http://corpora.iling.spb.ru

- **Corpus Linguistics**
  A study of language that includes all processes related to processing, usage and analysis of written or spoken machine-readable corpora. Corpus linguistics is a relatively modern term used to refer to a methodology, which is based on examples of 'real life' language use. At present, effectiveness and usefulness of corpus linguistics is closely related to the development of computer science. See McEnery and Wilson 1996; Aarts and Meijs 1990; Leech 1991; Svartvik 1992.
- **Corpus Processing**
  A general term used to refer to all processes related to annotation, presentation and analysis of corpora. See Aarts and Meijs 1990; McEnery and Wilson 1996: Ch. 2.

- **Alignment**
  A term is used to refer to the practice of defining explicit links between texts in a parallel corpus. Alignment is linking the elements (sentences, phrases or words) that are mutual translations of each other in parallel corpus. Sentence and word alignment (the term for performing this operation - aligner) may be performed with a high degree of accuracy automatically. See McEnery and Oakes 1996; McEnery and Wilson 1996: Ch. 2.

Annotation

Методы

- ***Построение терминосистемы предметной области***

- На первом, эмпирическом этапе лингвист с помощью специалиста данной области проводит логико-понятийный анализ ряда специальных текстов. На этом этапе необходимо выявить систему понятий и вскрыть связи и отношения между ними.
- На следующем, уже концептуальном этапе подбирается план выражения полученной модели.

Словарная статья (англ.)

- Term\ **reciprocate parallel corpus**
- Trans\ двусторонний параллельный корпус
- Def\ Multilingual corpus which contains, for all languages included, original texts as well as their translations into all the languages included.
- Up\ multilingual corpus
- Down\ aligned reciprocate parallel corpus
- Co\ comparable corpus
- Co\ parallel corpus
- Cyt\ Sometimes *reciprocate parallel corpora* are set up, **corpora** containing authentic texts as well as translations in each of the languages involved. This allows double-checking translation equivalents…

Методы

- Однако:
  с самого начала разработка понятий идет с помощью языковых средств и не может без них обойтись.
- Поэтому логично - обращение к корпусу
- Автоматизация процесса обработки корпусных данных

Методы

- ***Отражение терминосистемы в текстах***
- Специальный текст всегда представляет то или иное научное, техническое, отраслевое знание. С начала своего формирования специальное знание начинает проникать в семантику естественного языка.
- В специальном тексте происходит взаимодействие систем естественного языка с получившейся системой искусственного языка специального знания.

Материал исследования: корпус по корпусной лингвистике

- Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб.: 2002.
- Труды международной конференции «Корпусная лингвистика – 2004». СПб.: 2004.
- Труды международной конференции «Корпусная лингвистика – 2006». СПб.: 2006.
- Труды международной конференции «Корпусная лингвистика – 2008». СПб.: 2008.
- Труды Международной конференции «MegaLing–2005»: Прикладная лингвистика в поиске новых путей. СПб.: 2005.
- *Захаров В.П.*, Корпусная лингвистика. СПб., 2005;
- Национальный корпус русского языка… вып. 1-2
- Статьи в журнале НТИ…
- Диссертации…

Метаразметка корпуса

- Наряду с библиографическим описанием эксперты включают в число параметров статьи и наборы из 10 релевантных терминов-дескрипторов, позволяющих диагностировать тематическую принадлежность текста. Например:
- **Статья:**
- *Е.Л. Алексеева, А.М. Лаврентьев, И.В. Азарова, Л.А. Захарова* «Разметка корпуса древнерусских агиографических текстов» (*КЛ 2004*)
- **Набор терминов-дескрипторов:**
- *агиографический*, *житие*, *русский*, *текст*, *корпус*, *электронный*, *рукопись*, *словоформа*, *представление*, *разметка*

Методы

- ***Автоматическая обработка текстов в исследовании терминологии***
- Возможности компьютерных технологий ставят вопрос о возможности автоматической обработки текстов для решения различных терминологических задач.
- Автоматическое извлечение терминов из текстов …
- Можно выделить несколько основных подходов к выделению терминов:
  на основе шаблонов,
  на основе статистики встречаемости,
  на основе мер оценки устойчивости словосочетаний в специальных текстах X(*MI*, *t-score*, *Log-Likelihood*, *C-value*, критерий *χ2* и ряд других);
  комбинированные подходы.

Задачи

- 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.

Извлечение терминологии корпусной лингвистики

- Частотные списки слов

Термин	Часть речи	Частота
текст	Сущ	1641
корпус	Сущ	1233
язык	Сущ	945
словарь	Сущ	640
разметка	Сущ	331
контекст	Сущ	297
словоформа	Сущ	207
неоднозначность	Сущ	175
корпусный	Прил	157
корпусной	Прил	154
документ	Сущ	117
критерий	Сущ	114
пользователь	Сущ	114
словосочетание	Сущ	107
запрос	Сущ	78
словоупотребление	Сущ	74
сочетаемость	Сущ	60
коллокация	Сущ	38

Выявление специфичной лексики

- См. «лексические маркеры» -
- А.Я. Шайкевич. Статистический словарь Достоевского.
- Слово Частота **f(ipm)** **m(ipm)** **S=(f-m-1)/√m**
-

Извлечение терминологии корпусной лингвистики

- Частотные списки словосочетаний

Словокомплекс	*Модель*	*Частота*
корпус текстов	С+Срд	174
национальный корпус	П+С	93
база данных	С+Срд	74
корпусная лингвистика	П+С	74
машинный перевод	П+С	59
корпус русского языка	С+Прд+Срд	56
семантическая разметка	П+С	54
лексическая единица	П+С	43
морфологическая разметка	П+С	43
предметная область	П+С	42
семантический класс	П+С	36
толковый словарь	П+С	36
разрешение неоднозначности	С+Срд	35
корпусные данные	П+С	31
разметка текста	С+Срд	30

Статистика по основным синтаксическим моделям

- П+С 120
- С+Срд 54
- С+Прд+Срд 28
- П+С+Срд 5
- С+Срд+Срд 4

Автоматическая кластеризация

- Структурирование наборов терминов-дескрипторов осуществлялось с помощью инструмента автоматической классификации лексики (АКЛ), разрабатываемого на кафедре математической лингвистики СПбГУ под руководством доц. О.А. Митрофановой.
- Основным принципом АКЛ является возможность определения содержательной близости лексических единиц при сопоставлении их синтагматических свойств.
- 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.
- Сформированные таким образом кластеры лексем допускают дальнейшую лингвистическую интерпретацию.

Формирование классов условной эквивалентности

- **Классы условной эквивалентности термина-дескриптора *разметка***
- ***РАЗМЕТКА*** *Cos*
- ***ПРОСОДИЧЕСКИЙ*** **0,375**
- *БОЛЬШИНСТВО* 0,288
- ***АНАФОРИЧЕСКИЙ*** **0,288**
- ***??ВВОДИТЬСЯ*** **0,252**
- ***ДОКУМЕНТ*** **0,251**
- *ВЫДЕЛЕНИЕ* 0,250
- *МНОЖЕСТВО* 0,240
- ***ИНТОНАЦИЯ*** **0,226**
- *РЕФЕРЕНТНЫЙ* 0,214
- *РЕАЛЬНО* 0,213
- ***УДАРЕНИЕ*** **0,212**
- *РАЗ* 0,198
- *МЕСТОИМЕННЫЙ* 0,198
- *ИНОСТРАННЫЙ* 0,197
- *УПОТРЕБЛЯТЬСЯ* 0,196
- *НАЛИЧИЕ* 0,185
- *ДОСЛОВНО* 0,180
- *ОГОВОРКА* 0,167
- *ПОВТОР* 0,167

Автоматическая кластеризация

- 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:

Кластерная структура набора терминов-дескрипторов

- **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

Эксперименты с текстами с частичным совпадением наборов дескрипторов

- Обнаружены пары текстов, применительно к которым группы общих для них дескрипторов упорядочиваются единообразно:
  - [*словарь* [*корпус, текст*)]],
  - [*частота* [*корпус, текст*]],
  - [*массив* [*данные* [*корпус, текст*]]].
- Несовпадающие результаты.
  - [*формат* [*разметка* [*поиск* [*текст, корпус*]]]] *vs.* [*разметка* [[*корпус, текст*] *формат*] [*поиск*]].
  - [*поиск* [*слово* [*текст, корпус*]]] vs. [*поиск* [*корпус* [*слово, текст*]]].

Результаты кластеризации

- Результаты кластеризации
- Позволяют оценить диапазон понятийных категорий, релевантных для предметной области «Корпусная лингвистика».
- Вероятно, такие термины-дескрипторы, как *корпус*, *текст*, *данные*, *разметка*, *тег*, *поиск*, *слово*, *лемма*, *словоформа*, *контекст* и пр. представляют понятийное ядро указанной предметной области.

Выделение онтологических категорий

- Всего было зарегистрировано 335 различных терминов-дескрипторов.
- 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.

- Вероятно, такие термины-дескрипторы, как *корпус*, *текст*, *данные*, *разметка*, *тег*, *поиск*, *слово*, *лемма*, *словоформа*, *контекст* и пр. представляют понятийное ядро ПО.

Фрагмент онтологии по корпусной лингвистике

- *корпус данных*
  - *корпус текстов*
  - *тип корпуса*
    - *разработка*
      - *отбор данных*
      - *цифровка данных*
      - *разметка*
      - *корпус-менеджер*
    - *использование*
      - *поиск*
        - *запрос*
        - *терминальная цепочка символов*
        - *регулярное выражение*
        - *лемма*
        - *тег*
        - *результат*
        - *конкорданс*
        - *контекст*
        - *словоуказатель*
        - *статистика*

Формальная онтология

- 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.

Использование тезауруса

- Лингвистика
- Информационный поиск
- Перевод
- Автоматическая классификация текстов

- **Спасибо за внимание!**