Модуль I. Лекция 6

Тема: Корпусная лингвистика

1. Корпусная лингвистика как раздел прикладной лингвистики
2. Понятие корпуса разметки
3. Требования к корпусам
4. Виды корпусов

1. Корпусная лингвистика как раздел прикладной лингвистики

Одной из важных задач лингвистики является сбор и хранение источников фактического материала для лингвистических исследований.

В настоящий момент для решения этой задачи используются большие собрания текстов самой разной функциональной направленности, которые удобно хранить в электронном виде. Привлечение компьютеров и специальных телекоммуникационных сетей позволяет не только сохранять большие объемы текстов в электронном виде, но и осуществлять поиск по ним, обрабатывать их и т.п. Задача создания собраний текстов в электронном виде, или корпусов, является настолько значимой для современной лингвистики, что эти собрания электронных текстов становятся объектом исследований специального раздела прикладной лингвистики — корпусной лингвистики.

Корпусная лингвистика — раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов при помощи компьютеров.

Исходя из такого определения можно констатировать, что корпусная лингвистика включает два аспекта:

1) создание корпусов текстов с автоматическими инструментами их использования;

2) разработка способов экспериментальных исследований различных уровней языка на базе корпусов разных типов.

Современные исследователи-лингвисты могут как создавать свои собственные корпусы, а затем проводить необходимые исследования на их базе, так и использовать общедоступные корпусы, созданные другими исследователями и их коллективами.

Кроме проведения научных исследований корпусы могут использоваться:

1) в лексикографии для создания словарей, определения значения многозначных слов и т.д.;

2)в грамматике для определения частоты морфем, типов словосочетаний и предложений и т.д.;

3) в лингвистике текста для дифференциации типов текста, выявления связей внутри абзаца и между абзацами и т.д.;

4) в автоматическом переводе текстов для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов в параллельных текстах и т.д.;

5) в учебных целях для выбора цитат, фрагментов произведений, примеров для организации учебных занятий, создания учебных пособий и т.д.

6) в тестировании программ автоматического анализа и синтеза речи и т.д.

2. Понятие корпуса разметки

Центральное понятие корпусной лингвистики — лингвистический корпус — определяется как совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрами обеспеченных системой поиска. Таким образом, корпус можно кратко охарактеризовать следующим образом: Корпус = тексты + их разметка.

В более широком смысле корпусом считается любое собрание текстов. В этой трактовке выделяются размеченные (аннотированные) и неразмеченные корпусы текстов. В качестве подобных неразмеченных корпусов можно рассматривать существующие электронные коллекции

текстов: виртуальные библиотеки, архивы электронных версий периодических изданий или новостных лент, которые оказываются достаточными для некоторых исследовательских и учебных целей. Но использование неразмеченных собраний текстов, имеющих инструменты поиска, повышает долю информации, которая может оказаться нерелевантной для исследователя, что значительно затрудняет работу с таким источником. В связи с этим предметом корпусной лингвистики являются преимущественно размеченные корпусы текстов.

3. Требования к корпусам

Первым этапом в создании корпуса является отбор текстов. При Этом следует продумать, тексты каких функциональных стилей и конкретных жанров, какого года издания и в каком количестве будут включены в корпус. При отборе текстов в корпус следует ориентироваться на следующие требования к созданию корпусов:

1) репрезентативность (частота явления в корпусе должна соответствовать его частоте в естественном языке);

2) полнота (явление должно включаться в корпус, даже если его появление не соответствует идее репрезентативности);

3) достаточный объем (если первые корпусы достигали миллиона слов, то объем современных корпусов исчисляется сотнями миллионов и миллиардами, например, объем корпуса английского языка Bank of English превышает 2,5 млрд слов);

4) экономичность (корпус текстов должен экономить усилия исследователя при изучении проблемной области, т.е. быть не просто строгим подмножеством текстов проблемной области, но, по возможности, быть наиболее «экономичным»);

5) структуризация материала (в корпусе должны быть выделены адекватные корпусу единицы хранения);

6) компьютерная поддержка (поддержка корпуса текстов комплексом программ по обработке данных, обеспечивающих выявление контекстов слова, статистическую инвентаризацию, автоматическую словарную обработку и т.д.).

Важным этапом создания корпуса является его разметка. Разметка (англ. tagging,annotation) — это приписывание текстам и их компонентам специальных меток (англ. tag).Эти метки могут быть внешними (экстралингвистическими), включающими сведения об авторе и о тексте, или внутренними: структурными или собственно лингвистическими. Внешние метки содержат сведения об авторе, названии текста, годе и месте издания, жанре, тематике. Сведения об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое. Это кодирование информации имеет название мета разметка. Структурные метки несут информацию о статусе каждой единицы (глава, абзац, предложение, словоформа), а собственно лингвистические описывают лексические, грамматические и прочие характеристики элементов текста.

В соответствии с уровнем лингвистического описания различают морфологическую (определение части речи и морфологических категорий), синтаксическую (определение синтаксических связей), семантическую (категории, характеризующие значение слова), анафорическую(характеристика референтных связей, например, местоимений),просодическую (характеристика ударения и интонации), дискурсную (обозначение пауз, повторов) и некоторые другие виды разметки.

В частности, предложение «Этой весной опять расцвела акация» может быть размечено следующим образом:

Этой — МЖЕТ21весной — СЖЕТ22опять — Н22 расцвела — ГЖЕПЗЗ акация — СЖЕИЧ42

Первый индекс указывает на часть речи (М — местоимение, С — существительное, Н — наречие, Г — глагол), второй обозначает род, третий

— число, четвертый — падеж или время (у глагола), первая цифра указывает на число слогов, а вторая — на ударный слог [8].

Для разметки корпуса сообщений Твиттера при проведении международного исследовательского проекта по изучению данного жанра (Ганновер, 2010) нами были использованы, в частности, следующие виды меток:<STDS> (стандартное написание), <KOKS>
(использование только строчных букв), <KOGS> (использование только прописных букв), <GDOP> (удвоение графем), <GAUS> (выпадение графем), <GZUV> (написание лишней графемы) и т.д.

4. Виды корпусов

В зависимости от характера собранных в корпусе текстов, от их разметки и некоторых других факторов различают следующие виды корпусов (табл. 2).

Наиболее важным видом корпусов является универсальный национальный корпус, создаваемый для разных национальных языков. Создание и расширение универсальных национальных корпусов представляет собой одну из важнейших задач корпусной лингвистики.

Таблица 2

Классификация корпусов [18, 13]

№	Признак	Виды корпусов
1	Форма хранения	звуковые письменные смешанные
2	Язык текстов	русский английский и т.д.
3	«Параллельность»	одноязычные двуязычные многоязычные
4	Стиль	литературные диалектные разговорные публицистические терминологические смешанные
5	Способ доступа	свободно доступные коммерческие закрытые
6	Разметка	размеченные неразмеченные
7	Характер разметки	морфологические синтаксические семантические просодические и т.д.

Универсальный национальный корпус — это собрание текстов конкретного естественного языка, представительное по отношению ко всему языку, которое может служить для исследования самых разнообразных явлений этого языка.

Общепризнанный образец универсального национального корпуса— Британский национальный корпус (BNC) (www.natcorp.ox.ac.uk). Для русского языка таким представительным корпусом является Национальный корпус русского языка (НКРЯ) (www.ruscorpora.ru). Среди корпусов славянских языков выделяется Чешский национальный корпус (http://ucnk.ff.cuni.cz), созданный в Карловом университете Праги.

Национальные корпусы существуют также для немецкого, китайского, финского и других языков.

Одним из первых известных корпусов является Брауновский корпус (Brown Corpus), созданный в 1963 г.в Брауновском университете (США) для построения частотного словаря американского варианта английского языка. Его объем составлял 1 млн слов. Создатели корпуса (У.Френсис и Г.Кучера) разработали строгую процедуру отбора текстов: в корпус вошли 500 фрагментов прозаических текстов, созданных американскими авторами и напечатанных в 1961 г., по 2000 словоупотреблений каждый. Тексты представляли 15 наиболее распространенных жанров информативной и художественной прозы.

Поиск в корпусе в соответствии с запросом пользователя обеспечивается с помощью специальных программ — корпусных менеджеров.

Корпусный менеджер (англ. corpusmanager) — это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме. Результаты поиска обычно выдаются в виде конкорданса (поэтому корпусные менеджеры еще называют конкордансерами), где искомая единица представлена в ее контекстном окружении с представлением частотных характеристик отдельных языковых единиц, граммем и т.п.

Таким образом, корпус, представляющий собой размеченное собрание текстов с объемом слов не менее 100 млн, дает широкие возможности как для прикладных (работа над принципами автоматической разметки), так и для исследовательских целей.

Вопросы для обсуждения

1. Что такое Корпусная лингвистика?
2. Какими предпосылками определяется целесообразность и смысл использование корпусов?
3. Что понимается под значением «корпусный менеджер»?
4. Что такое «конкорданс»?

5. Расскажите про связь корпусной лингвистики и исторической лингвистки.
6. Расскажите про связь корпусной лингвистики и социолингвистики.
7. В чем заключается задача авторов корпуса?
8. Что такое репрезентативность?
9. Какие можно выделить способы деления корпусов на классы?
10. Какие типы корпусов можно выделить?
11. Расскажите подробнее про деление корпусов по критерию «параллельность».

Рекомендуемая литература

1. Зубов А.В., Зубова И.И.Информационные технологии в лингвистике: Учеб. пособие. –М.: Издательский центр «Академия», 2004. –208 с.
2. Клименко С.В., Рыков В.В.Логические индукция и дедукция как принципы отражения предметной области в корпусе текстов // Труды Международного семинара Диалог ‗2001 по компьютерной лингвистике и ее приложениям. –Аксаково, 2001.
3. Рыков В.В.Корпус текстов как реализация объектно-ориентированной парадигмы// Труды Международного семинара Диалог-2002. –М.: Наука, 2002
4. Шерстинова Т.Ю.«Один речевой день» на временной шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Филология. Востоковедение. Журналистика. Серия9. –СПб., 2009.39
5. Finegan E.LANGUAGE: its structure and use. –N.Y.: Harcourt Brace College Publishers, 2004.
6. Leech G.The Distribution and Function of Vocatives in American and British English Conversation // In Hasselgård and Oksefjell eds., 1999. –Pp. 107-120.
7. McEnery T., Wilson, A. Corpus Linguistics. –Edinburgh: Edinburgh University Press, 2001.
8. Рыков В.В.Корпус текстов как реализация объектно-ориентированной парадигмы// Труды Международного семинара Диалог-2002. –М.: Наука, 2002.