

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ГОУ ВПО «Иркутский государственный
лингвистический университет»**

В.П. Захаров, С.Ю. Богданова

КОРПУСНАЯ ЛИНГВИСТИКА

Учебник

**Иркутск
ИГЛУ
2011**

УДК 81'32
ББК 81.1-923
3 - 38

Рецензенты:

доктор филологических наук, ведущий научный сотрудник Института
 востоковедения РАН

С.А. Крылов,

доктор технических наук, профессор Санкт-Петербургского
 государственного университета

В.Ш. Рубашкин

Захаров В.П., Богданова С.Ю.

3-38 Корпусная лингвистика: учебник для студентов гуманитарных вузов.
 – Иркутск: ИГЛУ, 2011. – 161 с.

ISBN 978-5-88267-316-0

Учебник знакомит с концепциями корпусной лингвистики, дает возможность освоить основы корпусных технологий, приобрести навыки работы с корпусами, определить место дисциплины и собственно корпусов в ряду информационных технологий.

Предназначен для студентов, магистрантов и аспирантов филологических специальностей.

УДК 81'32
ББК 81.1-923

ISBN 978-5-88267-316-0

© Захаров В.П., 2011

© Богданова С.Ю., 2011

© Иркутский государственный
 лингвистический университет, 2011

ОГЛАВЛЕНИЕ

Предисловие	5
ЧАСТЬ 1. ВВЕДЕНИЕ В КОРПУСНУЮ ЛИНГВИСТИКУ	7
1.1. Основные понятия корпусной лингвистики	7
1.2. Направления в лингвистике, предвосхитившие появление корпусной лингвистики: от картотеки к корпусу	11
1.3. История создания лингвистических корпусов	14
1.4. Основные характеристики корпусов	17
1.4.1. Репрезентативность корпусов	17
1.4.2. Классификация корпусов по различным основаниям ..	20
1.4.3. Особые типы корпусов	26
1.4.3.1. Параллельные корпуса	26
1.4.3.2. Корпусы устной речи	29
ЧАСТЬ 2. СОЗДАНИЕ КОРПУСОВ	33
2.1. Предварительные работы по созданию корпуса	33
2.1.1. Проектирование и технологический процесс создания	33
2.1.2. Отбор источников. Критерии отбора	36
2.1.3. Основные процедуры обработки естественного языка: токенизация, лемматизация, стемминг, парсинг	38
2.2. Понятие разметки	42
2.2.1. Разметка. Средства разметки корпусов	42
2.2.2. Лингвистическая разметка	45
2.2.3. Экстралингвистическая разметка	50
2.2.4. Стандартизация в корпусной лингвистике	52
ЧАСТЬ 3. ИСПОЛЬЗОВАНИЕ КОРПУСОВ	55
3.1. Корпусные менеджеры	55
3.1.1. Корпус как поисковая система	55
3.1.2. Языки запросов	56
3.1.3. Выходные интерфейсы	69
3.1.4. Корпусные менеджеры нелингвистических корпусов (WWW)	71
3.2. Обзор существующих корпусов различных типов	75
3.2.1. Зарубежные национальные корпуса	75
3.2.2. Корпусы русского языка	82
3.2.2.1. Первые корпуса русского языка	82
3.2.2.2. Современные корпуса русского языка	86
3.2.2.2.1. Национальный корпус русского языка	86
3.2.2.2.2. Устные корпуса русского языка	89

3.2.3. Специальные корпуса	92
3.3. Корпусные исследования	94
3.3.1. Пользователи корпусов	94
3.3.2. Способы использования корпусов	95
3.3.3. Лексикографические исследования, основанные на корпусах	98
3.3.3.1. Пример одного лексикографического исследования . .	100
3.3.3.2. Выделение коллокаций статистическими методами . .	113
3.3.4. Грамматические исследования, основанные на корпусах	116
3.3.4.1. Распределение и функции номинализаций	117
3.3.4.2. Распределение грамматических категорий	123
3.3.5. Исследования дискурса, основанные на корпусах	128
Заклучение	147
Библиографический список	148
Приложения	153

Предисловие

Предлагаемый вашему вниманию учебник является своего рода обобщением многочисленных разрозненных материалов, опубликованных за последние два десятилетия в России и за рубежом, которые легли в основу лекционных курсов по дисциплине «Корпусная лингвистика», читаемых кандидатом филологических наук, доцентом Виктором Павловичем Захаровым в Санкт-Петербургском государственном университете и доктором филологических наук, профессором Светланой Юрьевной Богдановой в Иркутском государственном лингвистическом университете. Материал, представленный в учебнике, может также быть использован в курсах лекций по дисциплинам «Информационные и коммуникационные технологии в науке и образовании», «Основы прикладной лингвистики», «Компьютерные методы в лингвистических исследованиях» и др.

Учебник состоит из трех частей. Первая часть «ВВЕДЕНИЕ В КОРПУСНУЮ ЛИНГВИСТИКУ» знакомит с основными понятиями и терминами корпусной лингвистики, историей ее становления как отрасли языкознания, ее целями и задачами, типами существующих корпусов. Вторая часть «СОЗДАНИЕ КОРПУСОВ» описывает в общих чертах технологические процессы, связанные с их проектированием, отбором и обработкой языкового материала, способами разметки. Третья часть «ИСПОЛЬЗОВАНИЕ КОРПУСОВ» включает три раздела. Раздел 3.1 посвящен описанию корпусных менеджеров, обеспечивающих поиск в корпусе. Раздел 3.2 представляет собой обзор как зарубежных национальных корпусов, так и корпусов русского языка. Раздел 3.3 посвящен описанию конкретных исследований на базе корпусов разных типов, в нем приводятся результаты исследований и дается их теоретическая интерпретация. В первую очередь, авторы хотят показать, как можно работать с реальным языковым материалом быстрее и эффективнее, базирясь на корпусах. В этом разделе приведены примеры исследований лишь в нескольких областях лингвистики —

лексикографии, грамматике и анализе дискурса. Безусловно, сфера применения корпусных данных в лингвистике значительно шире.

Цель учебника – познакомить студентов с концепциями корпусной лингвистики, дать им возможность освоить основы корпусных технологий, приобрести навыки работы с корпусами, определить место дисциплины и собственно корпусов в ряду информационных технологий.

Задачи учебника:

- ознакомление студентов с новой парадигмой в лингвистических исследованиях;
- ознакомление студентов с историей корпусных исследований;
- изучение языковых и программных средств корпусной лингвистики;
- формирование навыков работы с программными средствами и информационными ресурсами корпусной лингвистики;
- ознакомление студентов с конкретными лингвистическими исследованиями, основанными на корпусных данных.

Авторы выражают надежду, что студенты филологических специальностей заинтересуются использованием корпусов, независимо от сферы их научных интересов, а каждый преподаватель найдет в учебнике то, о чем нужно говорить в его аудитории.