Корпусная лингвистика. Обзор корпусов. Сферы использования корпусов

- Корпусная лингвистика – раздел компьютерной лингвистики, разрабатывающий общие принципы построения и использования лингвистических корпусов с применением компьютерных технологий.
- Объект – корпус текстов.
- Лингвистический, или языковой, корпус текстов – большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач.
- Предмет – теоретические основы и практические механизмы создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей.

История корпусной лингвистики

- Брауновский корпус (The Brown Corpus) (1960-е гг.)
- корпус Ланкастер-Осло-Берген (The Lancaster-Oslo-Bergen Corpus) (1970-е гг.)

Создатели Брауновского корпуса

У. Френсис (1910 – 2002)

Г. Кучера (1925 - 2010)

Классификация корпусов

Тип языковых данных:	• Письменные (Брауновский корпус, LOB); • Устные: Корпус Лондон-Лунд (The London-Lund Corpus); • Смешанные (НКРЯ).
«Параллельность»:	• Одноязычные; • Двуязычные; • Многоязычные.
«Литературность»:	• Литературные; • Диалектные; • Разговорные (корпус Один Речевой День); • Терминологические; • Смешанные.

Классификация корпусов

Цель:	• Многоцелевые; • Специализированные.
Жанр:	• Литературные; • Фольклорные; • Драматургические; • Публицистические.
Доступность:	• Свободно доступные; • Коммерческие; • Закрытые.
Динамичность:	• Динамические (мониторные); • Статические.

Требования к национальному корпусу

1. Необходимый и достаточный объём.
2. Достаточно протяжённый хронологический охват языка.
3. Репрезентативность выборки текстов.
4. Тексты должны пройти филологическую экспертизу.
5. Тексты должны быть представлены в электронной форме.
6. Многопрофильная система аннотирования.
7. Многофункциональность корпуса.
8. Общедоступность.

http://corpus.byu.edu/coca/ Corpus of Contemporary A...

THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA)

450 MILLION WORDS, 1990-2012

BRIGHAM YOUNG UNIVERSITY

ENTER

http://corpus.byu.edu/coca/

CORPUS OF CONTEMPORARY AMERICAN ENGLISH

DISPLAY

LIST CHART KWIC COMPARE

SEARCH STRING

WORD(S)

COLLOCATES [d*] 4 4

POS LIST det.ALL

RANDOM SEARCH RESET

SECTIONS SHOW

1 IGNORE

SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC

2 1990-1994
1995-1999
2000-2004
2005-2009
2010-2012

SPOK:ABC

SORTING AND LIMITS

SORTING FREQUENCY

MINIMUM FREQUENCY 10

CLICK TO SEE OPTIONS

[WHERE SHOULD I START?] [COMPARE TO OTHER CORPORA / ARCHITE

Note: the corpus is designed for screen resolutions of at least 1024x768, but your resolution is only 1280x720 (i.e. a netbook or iPad). W
corpus should work, it may not display correctly, and you may have more problems navigating from one page to another.

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, and the only large and balanced
American English. The corpus was created by Mark Davies of Brigham Young University, and it is used by tens of thousands of users eve
(linguists, teachers, translators, and other researchers). COCA is also related to other large corpora that we have created.

The corpus contains more than 450 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and
texts. It includes 20 million words each year from 1990-2012 and the corpus is also updated regularly (the most recent texts are from Summ
Because of its design, it is perhaps the only corpus of English that is suitable for looking at current, ongoing changes in the language (see
article in *Literary and Linguistic Computing*).

The interface allows you to search for exact words or phrases, wildcards, lemmas, part of speech, or any combinations of these. You can s
surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near *faint,* all adjectives near *woman*, or all verbs near
which often gives you good insight into the meaning and use of a word.

The corpus also allows you to easily limit searches by frequency and compare the frequency of words, phrases, and grammatical constructio
least two main ways:

- By genre: comparisons between spoken, fiction, popular magazines, newspapers, and academic, or even between sub-genres (or domai
as movie scripts, sports magazines, newspaper editorial, or scientific journals
- Over time: compare different years from 1990 to the present time

You can also easily carry out semantically-based queries of the corpus. For example, you can contrast and compare the collocates of two relat
(*little/small*, *democrats/republicans*, *men/women*), to determine the difference in meaning or use between these words. You can find the frequ
distribution of synonyms for nearly 60,000 words and also compare their frequency in different genres, and also use these word lists as part
queries. Finally, you can easily create your own lists of semantically-related words, and then use them directly as part of the query.

Please feel free to take a five minute guided tour, which will show the major features of the corpus. A simple click for each query will automa
in the form for you, search through the more than 450 million words of text, and then display the results.

http://www.natcorp.ox.ac.uk/ | british national corpus - Поис... | [bnc] British National Corpus

BRITISH NATIONAL CORPUS

About

What is the BNC?
Creating the BNC
BNC Products
Copyright
Contact Us
Contents A-Z

Using the BNC

What can I do with the BNC?
Using BNC with Xaira
FAQ

Obtaining

How to download

About the BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century. [more]

Simple Search from the British Library

Type a word or phrase in the search box and press the Go button to see up to 50 random hits from the corpus.

Look up: Enter text Go

You can search for a single word or a phrase, restrict searches by part of speech, search in parts of the corpus only, and much more. This is a link to the simple search facility hosted by the British Library.

The search result will show the total frequency in the corpus and up to 50 examples. [more information]

There are other online services offering more advanced search functions (some require user registration):

- BYU-BNC (Brigham Young University)
- BNCWeb at Lancaster University
- BNCWeb at Oxford [Oxford University users only]
- Intellitext (University of Leeds)
- Phrases in English

Please note that we cannot answer queries about using any of these services, which are all provided elsewhere!

News from the BNC

- Geoffrey Leech
- BNC2014
- Withdrawal of distribution on disks
- BNC XML for download

http://www.natcorp.ox.ac.uk/

Упсальский корпус русского языка

Специальные тексты с 1985 по 1989 год и художественные тексты с 1960 по 1988 год.

Упсальский университет (Швеция)

Компьютерный корпус текстов русских газет конца XX-ого века

- Место и время создания: Филологический факультет МГУ, Лаборатория общей и компьютерной лексикологии и лексикографии, 2000-2002 гг.
- Состав – полные тексты избранных номеров ряда российских газет на русском языке, опубликованных в 1994 – 1997 гг.
- Доступ в Интернете – http://www.philol.msu.ru/~lex/corpus/

http://ruscorpora.ru/

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА

главная
архив новостей

поиск в корпусе

что такое корпус?
состав и структура
статистика
графики
частоты
морфология
обороты
синтаксис
семантика
параметры текстов

studiorum
форум

о проекте
участники проекта
публикации
программные средства
ошибки в корпусе
использование корпуса

Национальный корпус русского языка

На этом сайте помещен корпус современного русского языка общим объемом более 500 млн слов. Корпус русского языка — это информационно-справочная система, основанная на собран в электронной форме.

Корпус предназначен для всех, кто интересуется самыми разными вопросами, связанными с русским языком: профессиональных лингвистов, преподавателей языка, школьников и студен изучающих русский язык.

Как пользоваться Корпусом (инструкция в формате PDF)

Подробнее о корпусе

Новости проекта

3 июня 2014 года
Объявляется конкурс проектов нового дизайна Национального корпуса русского языка.

29 апреля 2014 года
Национальному корпусу русского языка исполнилось 10 лет.

29 апреля 2014 года
В режиме бета-версии запущен поиск по n-граммам подкорпуса с неснятой омонимией основного корпуса.

11 апреля 2014 года
Обновлён синтаксический корпус, его объём теперь составляет более 860 тыс словоупотреблений.

18 января 2014 года
Пополнен акцентологический корпус, теперь в его составе 15 млн словоупотреблений.

18 января 2014 года
Пополнен устный корпус, его объём возрос до 11 млн словоупотреблений.

14 января 2014 года
Пополнен параллельный корпус: добавлены двуязычные армянский, болгарский и латышский корпуса, существенно расширены немецкий, английский и белорусский. В двуязычный французс включены поливариантные русско-французские тексты (с несколькими альтернативными переводами). Общий объём корпуса теперь превышает 54 млн словоупотреблений.

Сферы использования лингвистических корпусов

3 типа данных:

- эмпирическая поддержка;
- информация по частотности;
- метаинформация.

2. Программирование, компьютерная лингвистика.

3. Методика преподавания родного языка.

4. Методика преподавания иностранного языка.

5. Журналистика, редактирование.

6. Переводоведение.

7. Литературоведение.

8. Текстология.

9. Судебно-лингвистическая экспертиза.

10. Другие общественные науки.

Литература

1. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. М., 2001.
2. Грудева Е.В. Корпусная лингвистика: учеб. пособие. М., 2012.
3. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. Иркутск, 2011.
4. Корпусная лингвистика [Электронный ресурс]. URL: http://corpora.iling.spb.ru/.
5. Плунгян В.А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003 – 2005. М., 2005. С. 12 – 17.
6. Плунгян В.А. Почему современная лингвистика должна быть лингвистикой корпусов [Электронный ресурс]. URL: http://www.polit.ru/article/2009/10/23/corpus.