

Введение в корпусную лингвистику

**1 Предмет корпусной лингвистики.
Сопоставление корпусной и традиционной
лингвистики**

**2 История создания лингвистических
корпусов**

3 Типология корпусов

Литература

- Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику. – М., 2003. – С. 112–137.
- Захаров В.П. Корпусная лингвистика: Учебно–метод. пособие. – СПб., 2005. – 48 с.
<http://vp-zakharov.narod.ru/publications.htm>
- Захаров В.П. Корпусная лингвистика: учебник для студентов гуманитарных вузов / В.П. Захаров, С.Ю. Богданова. – Иркутск, 2011. – 161 с.
- Зубов А.В. Информационные технологии в лингвистике: учеб. пособие / А.В. Зубов, И.И. Зубова.– М., 2004. – 208 с.

1 Предмет корпусной лингвистики.

Сопоставление корпусной и традиционной лингвистики

Корпусная лингвистика — раздел лингвистики (компьютерной лингвистики), занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий.

Корпусная лингвистика сформировалась как отдельный раздел науки о языке в первой половине 90–х гг. XX в.

Лингвистический корпус –
это совокупность текстов

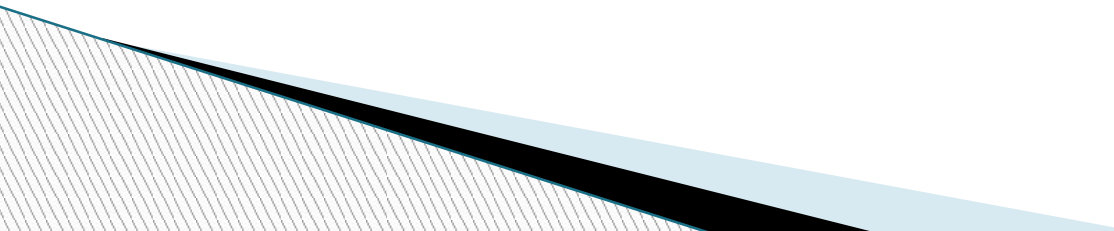
➤ **собранны**

Х

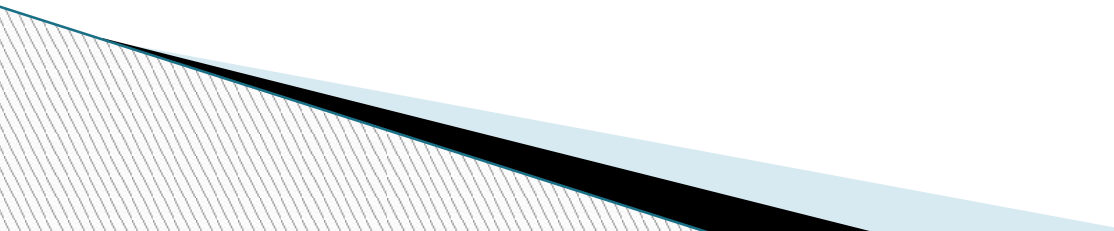
В

соответств

Целесообразность создания и смысл использования корпусов

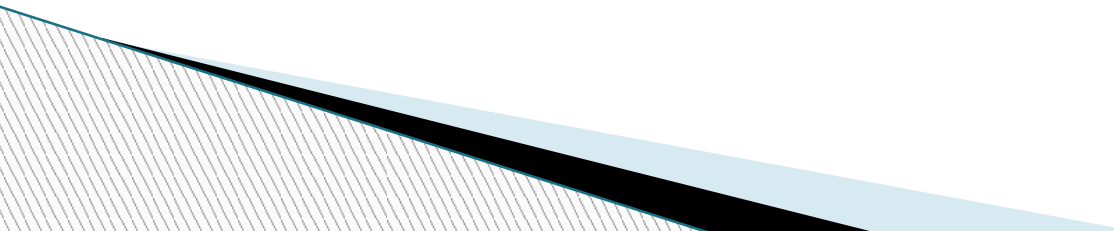
- представлением данных в реальном контексте;
 - достаточно большой представительностью данных (при большом объёме корпуса);
 - возможностью многократного использования единойжды созданного корпуса для решения различных задач.
- 

Объект корпусной лингвистики – корпус текстов, который, с одной стороны, представляет собой исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, является результатом деятельности корпусной лингвистики.

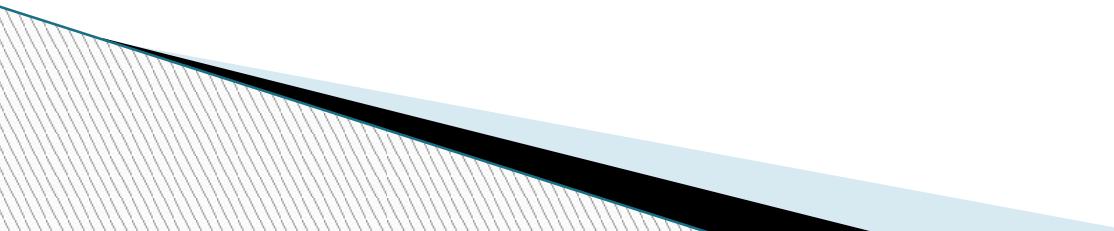


Двойственный характер объекта обуславливает двойственный характер корпусной лингвистики – нацеленность как на создание, так и на использование корпусов текстов.

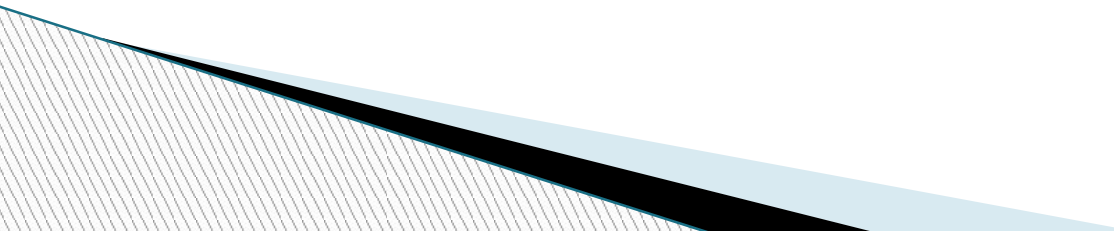
Предмет корпусной лингвистики – теоретические основы и практические механизмы создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей.



В понятие ***корпус текстов*** входит также ***корпусный менеджер*** (корпус-менеджер) – специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.



Поиск в корпусе позволяет построить ***конкорданс*** – список всех фиксаций искомой языковой единицы в контекстах со ссылками на источник.



Сопоставление корпусной и традиционной лингвистики

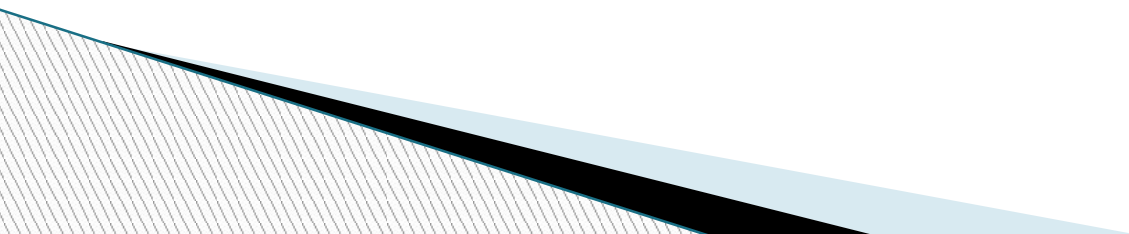
- Основное внимание – изучению речи
- В исследованиях опора на данные корпуса текста
- Предпочтение количественным методам
- Текст рассматривается как некоторая физическая сущность
- Основное внимание – изучению языка
- В исследованиях путь от теории к её объяснению и подтверждению в фактах речи
- Предпочтение качественным методам
- Текст рассматривается как некоторая абстракция

КОРПУСНАЯ
ЛИНГВИСТИКА

ТРАДИЦИОННАЯ
ЛИНГВИСТИКА

2 История создания корпусов

Первые лингвистические корпусы
текстов появились в 60–е гг. XX в.

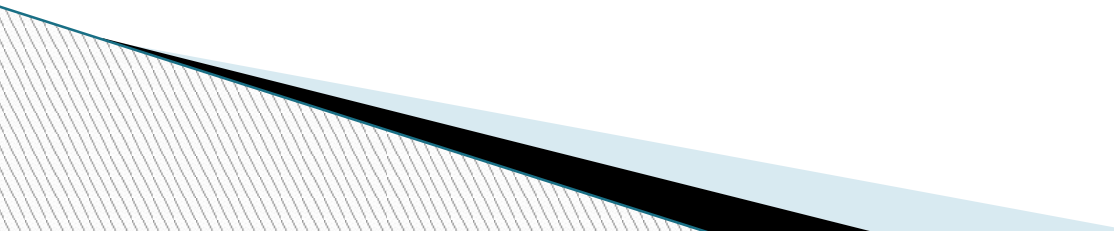


Первый корпус текстов – ***Брауновский корпус (The Brown Corpus)*** создан ***в 1963 г.*** в Брауновском университете (США).

Создатели корпуса У. Френсис и Г. Кучера.

Брауновский корпус включает 500 текстов из американских книг, газет, журналов, впервые опубликованных в США в 1961 г.

Каждый текст имеет длину 2000 словоупотреблений, и все собрание включает 1 млн. слов.



Тексты в Брауновском корпусе принадлежат 15-ти наиболее массовым жанрам англоязычной печатной прозы США.

Корпус сопровождается большим количеством материалов первичной статистической обработки (например, частотным и алфавитно-частотным словарем).

Цель создания Брауновского корпуса – обеспечить системное изучение отдельных жанров письменного английского языка и сравнение жанров.

Появление Брауновского корпуса вызвало всеобщий интерес и оживленные дискуссии (по поводу принципов отбора текстов и состава потенциально решаемых задач).

BROWN CORPUS MANUAL

MANUAL OF INFORMATION
to accompany
A Standard Corpus of Present-Day
Edited American English, for use
with Digital Computers.

by
W. N. Francis
H. Kucera
Brown University

Providence, Rhode Island
Department of Linguistics
Brown University
1964

Revised 1971

Revised and Amplified
1979

PREFACE

To Revised Edition, 1979

This Manual was first published in 1964, when the Standard Sample of Present-Day American English (the Brown Corpus) was first made available. *) A revised edition was issued in 1971, principally to incorporate information about the text turned up in seven years of use. The present revision is more extensive, since it includes information about recently prepared versions of the Corpus, notably the «tagged» text completed at Brown University in 1979. Two complete proofreadings of the Corpus have resulted in corrections of two kinds: errors in the preparation of the original tape, which have been silently corrected in recently issued copies, and further typographical errors and anomalies in the underlying text, which have been recorded in the descriptions of individual samples on pages 33-176. (Most of these were listed on corrigenda sheets which have been enclosed with recently issued copies of the Manual and *incorporated in this web-version.*)

We wish to record here our thanks to all those who have sent in information about errors in the Corpus, and our special gratitude to those who have worked on the production of alternate versions of the Corpus, notably Gerald M. Rubin, Barbara Greene Levine, Sandra Pearce, Patricia Strauss, Stephen Ritz, Andrew Mackie, Jostein Hauge, and Donald Sherman (a partial list). At the time of writing, more than 160 copies of the Corpus are in circulation, and a recent bibliography of published works using or referring to the Corpus includes 57 items (ICAME News, No. 2, Bergen, March 1979, pp. 9-12).

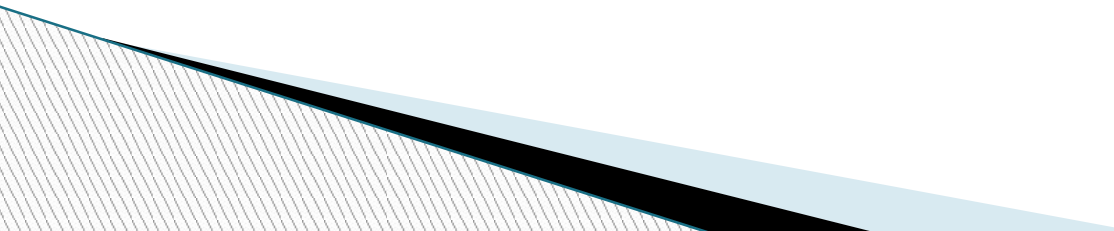
* The original Standard Corpus was prepared under a grant from the Cooperative Research Program of the U.S. Office of Education.

W. Nelson Francis - Henry Kucera
Brown University
July 1979.

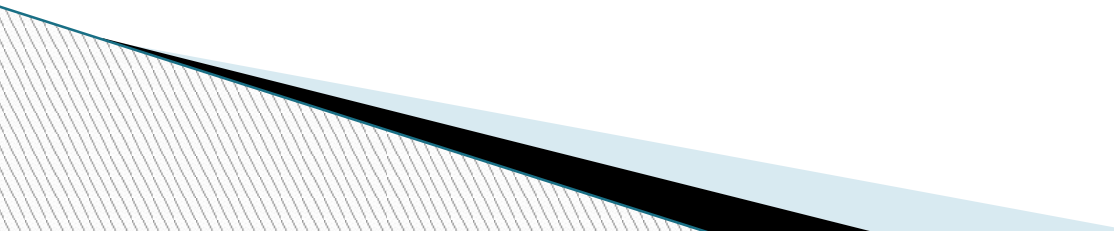
TABLE OF CONTENTS

По принципам Брауновского корпуса был создан **корпус текстов Ланкастер–Осло–Берген** (по названиям британского и двух норвежских университетов), впервые опубликованных в Великобритании в 1961 г.: 15 жанров, 500 текстов по 2000 словоупотреблений, т.е. 1 млн. слов британского варианта английского языка.

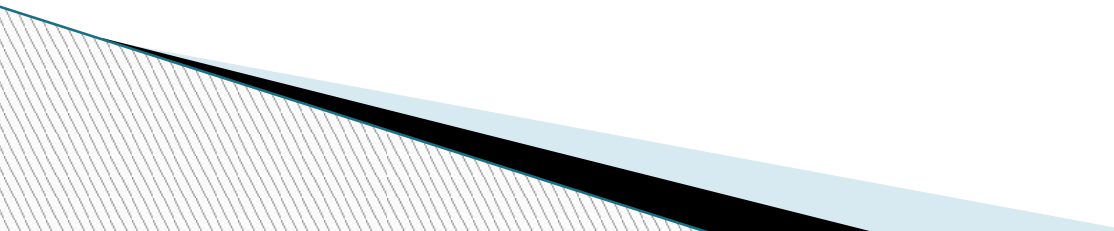
Брауновский корпус задал стандарт в 1 млн словоупотреблений для создания представительных корпусов на других языках.



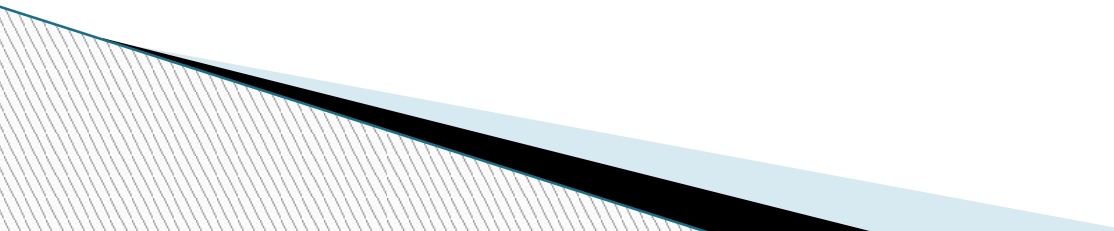
По аналогичной модели был построен и русский корпус, созданный в 1980-е годы в Университете Уппсалы (Швеция) – ***Уппсальский корпус русского языка.***



*В 1980-е годы создаются корпуса
большего размера.*



Корпусы английского языка

- Британский Национальный Корпус (British National Corpus, BNC),
 - Международный корпус английского языка (International Corpus of English – ICE),
 - Корпус современного американского английского (Corpus of Contemporary American English – COCA) и др.
- 

Британский национальный корпус

British National Corpus

www.natcorp.ox.ac.uk 100 млн.



Home The Corpus Using Obtaining XAIRA FAQ Archive Contact Us A-Z

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 order
Pricing
Xaira
FAQ

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 current British English, both spoken and written. [\[more\]](#)

Search the Corpus

Type a word or phrase in the search box and press the Return key on your keyboard to see up to 50 random hits from the corpus.

Look up:

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.

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

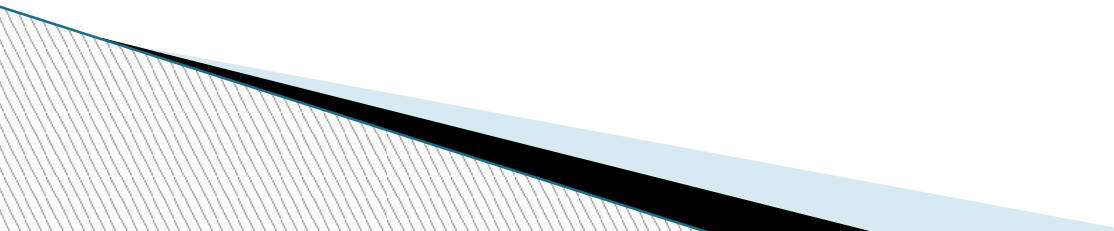
News from the BNC

- [British Library BNC Online service switches to BNC-XML](#)
- [BNC Web: new online service available to BNC Licensees](#)
- [BNC Baby: new edition available](#)
- [Material from workshops available online](#)
- [Problems accessing your BNC trial account?](#)



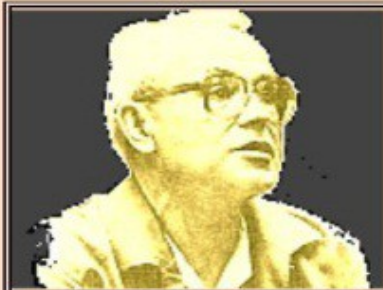
```
... lemma="cubic"  
... lemma="cost">cost  
... lemma="issue">is.  
... lemma="legal" </w>w type="p"  
... lemma="should">should  
... lemma="out of" </w>w type="p"  
... lemma="pursuant to" </w>w  
... lemma="and">and </w>w type="A20" let  
... lemma="PUL">(</w>w  
... lemma="ulations" </w>w
```

В СССР таким проектом был ***Машинный Фонд русского языка***, создававшийся по инициативе А.П. Ершова.



Notice: Undefined variable: host in E:\cfri.txt\html\cfri_site\counter_in.php on line 118

Поддерживается Российским гуманитарным научным фондом



Машинный фонд русского языка

Андрей Петрович Ершов



Зеркало этого сайта находится по адресу <http://cfri.ruslang.ru/index0.php>

Работы по созданию Машинного фонда русского языка были начаты в 1985 г. по инициативе академика А.П. Ершова (<http://ershov.iis.nsk.su/russian/>), после состоявшейся в 1983 г. специальной всесоюзной конференции, материалы которой позднее были опубликованы в книге Машинный фонд русского языка: идеи и суждения, М.: Наука, 1989. Тогда же был создан отдел Машинного фонда русского языка в Институте русского языка РАН. Заведование отделом взял на себя тогдашний директор Института член-корреспондент АН СССР Ю.Н. Караулов. Под его руководством была разработана «Комплексная программа научных исследований и прикладных разработок по созданию Машинного фонда русского языка на 1996-2000 гг. и информатизации исследований в Институте русского языка АН СССР», в основу которой легли упомянутые материалы. Руководителями Отдела были последовательно член-корреспондент АН СССР Ю.Н. Караулов (1985-1991 гг.), доктор филологических наук В.М. Андрущенко (1992-1998 гг.), профессор, доктор филологических наук А.Я. Шайкевич

Содержание

К 1990–м годам зафиксировано более 600 корпусов. Распределение их по годам создания:

-1965	10
1966–1970	20
1971–1975	30
1976–1980	80
1981–1985	160
1986–1990	320

*В настоящее время корпусы созданы
для многих языков мира.*

*Некоторые из них содержат
миллиарды словоупотреблений.*

Корпусы в сети Интернет

Наименование корпуса	Количество словоупотреблений
Национальный корпус русского языка http://ruscorpora.ru	более 360 млн. словоупотреблений
Компьютерный корпус текстов русских газет конца XX века http://www.philol.msu.ru/~lex/corpus	около 200 тыс. словоупотреблений
Корпус русского языка ХАНКО (Хельсинский университет) http://www.ling.helsinki.fi/projects/hanco	100 тыс. словоупотреблений
Уппсальский корпус русских текстов Доступен для поиска на сайте http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html	1 млн. словоупотреблений
Словарь-корпус языка А.С. Грибоедова http://www.inforeg.ru/electron/concord/concord.htm	120 тыс. словоупотреблений

Корпусы в сети Интернет

Наименование корпуса	Количество словоупотреблений
<p>Банк английского языка (Bank of English) http://www.collins.co.uk/books.aspx?group=153 Свободный доступ: http://www.collins.co.uk/Corpus/CorpusSearch.</p>	<p>524 млн. словоупотреблений, 56 млн. в свободном доступе</p>
<p>Венгерский национальный корпус http://corpus.nytud.hu/mnsz/</p>	<p>100 млн. словоупотреблений</p>
<p>Корпус испанского языка (исторический) http://www.corpusdelespanol.org/</p>	<p>100 млн. словоупотреблений, тексты 13–20 вв. Создан в Иллинойском университете, США</p>
<p>Мангеймский корпус немецкого языка (Institut für Deutsche Sprache, Mannheim, Germany) http://corpora.ids-mannheim.de/~cosmas/</p>	<p>1610 млн. словоупотреблений</p>

Корпусы в сети Интернет

Наименование корпуса	Количество словоупотреблений
Корпус латинских текстов «Персей» http://www.perseus.tufts.edu	
Корпус современного датского языка http://www.korpus2000.dk/	50 млн. словоупотреблений Тексты 1998–2002 гг.
Корпус современного итальянского языка CORIS/CODIS http://www.cilta.unibo.it/ricerca.htm	100 млн. словоупотреблений
Корпус современного китайского языка (LIVAC Synchronous Corpus) http://www.rcl.cityu.edu.hk/livac/	720 млн. словоупотреблений (150 млн. иероглифов)
Национальный корпус словенского языка http://www.fida.net/eng/	более 100 млн. словоупотреблений
Национальный корпус болгарского языка http://search.dcl.bas.bg	320 млн. словоупотреблений

Национальный корпус русского языка

http://ruscorpora.ru



- 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
- другие корпуса

Национальный корпус русского языка

[English](#)

На этом сайте помещен корпус современного русского языка объемом более 300 млн слов. Корпус русского языка — это информационно-справочная система, основанная на собрании русских текстов в электронной форме.

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

[Как пользоваться Корпусом \(инструкция в формате PDF\)](#)

[Подробнее о корпусе](#)

Новости проекта

8 августа 2012 года

Существенно пополнился [газетный корпус](#) (большой корпус СМИ 2000-х годов). Теперь его объем превышает 332 тыс. документов, 173 млн словоупотреблений. Напоминаем, что ограничиться поиском по предыдущей версии корпуса можно в разделе «Версии» в форме [выбора подкорпуса](#).

3 августа 2012 года

Произошло очередное пополнение [мультимедийного](#) и [устного](#) корпусов.

10 июля 2012 года

Год назад был запущен сервис [«Графики»](#), аналогичный сервису [Google Books Ngram Viewer](#): распределение найденных по точной форме слов и словосочетаний по годам. Теперь такой график можно построить по результатам произвольного запроса к [основному корпусу](#) (а не только по точным формам, как раньше). Для этого перейдите по ссылке «Распределение по годам» на странице с результатами поиска и дождитесь ответа. Кроме того, по соседней ссылке «Статистика» доступны таблицы с распределением найденных документов по авторам, жанрам, типам, тематике текста и т. д.

20 мая 2012 года

Для общего доступа открыт [Церковнославянский корпус](#) как первый из разделов Исторического корпуса. Основу церковнославянского корпуса составляют современные богослужебные тексты (XIX-XX век) (60%). Кроме того, в корпусе представлены тексты других периодов (XVII-XVIII век) и жанров: писание, святоотеческие и др. Общий объем корпуса – около 4,7 млн словоупотреблений. Тексты в корпусе снабжены морфологической разметкой, которая позволяет искать слова по лемме и грамматическим признакам. Пользователь может искать словоформы и леммы в трех орфографических системах: точной, упрощенной и модернизированной.

20 января 2012 года

1. Очередное обновление и пополнение ряда корпусов: основного, акцентологического, мультимедийного, параллельного, поэтического,

Corpus français de textes La Bibliothèque Universelle

L'accès libre au texte intégral d'oeuvres du domaine public francophone sur Internet depuis 1993.

abu.cnam.fr



ABU : la Bibliothèque Universelle

288 textes de 101 auteurs. (Janvier 2002)
Les [nouveau](#)s.

Qui le fait ?

Ces textes sont produits et diffusés par les membres bénévoles de l'Association des Bibliophiles Universels ([ABU](#)).

N'hésitez pas à [nous rejoindre](#) !

Nous avons aussi un peu d'informations sur le [texte électronique](#) en général.

Le serveur

ABU est hébergée par l'équipe "[Multimédia et Interaction Homme-Machine](#)" du [CEDRIC](#), au [CNAM](#), Paris.

Nous diffusons des [statistiques](#) utiles sur l'accès à ce service.

Avant tout téléchargement, nous vous invitons à consulter la [Licence](#) ABU.

Les textes

Pour accéder aux textes, consultez le catalogue des [auteurs](#) ou celui des [textes](#).

Vous pouvez également faire des [recherches de mots](#) sur tout le corpus.

Nous avons aussi plusieurs [dictionnaires](#)

Ailleurs et demain

Si vous ne trouvez pas ce que vous cherchez ici, consultez le méta-catalogue [ClicNet](#) aux Etats-Unis.

NOUVEAU au CNAM :

Le [Conservatoire Numérique des Arts et Métiers](#), une bibliothèque

Чешский национальный корпус

Český národní korpus

www.korpus.cz

hledat: v korpusu na stránkách

Český národní korpus

Filozofická fakulta Univerzity Karlovy

Krátké zprávy

- Co je korpus?
- Kontakty
- Dostupné korpusy
- Projekt InterCorp
- Poskytovatelé textů
- Granty, sponzoři
- Naše publikace
- Ke stažení
- Dohody a registrace

Hledat v ČNK

- Pracovní kolektiv
- Jak citovat korpus
- Knihovna
- Výuka
- Odkazy

Manuál a instalace



Co je korpus?

Korpus je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který primárně slouží k jazykovému výzkumu. K práci s korpusy slouží speciální vyhledávací program. S jeho pomocí je možné vyhledávat slova a slovní spojení v kontextu a zjistit jejich frekvenci v korpusu i původní textový zdroj. Umožňuje i další zpracování nalezeného (např. abecední třídění apod.). U některých korpusů lze vyhledávat i podle slovních druhů.

Český národní korpus (ČNK) je akademický projekt zaměřený na budování rozsáhlého počítačového korpusu především psané češtiny. Pracuje na něm **Ústav Českého národního korpusu** na Filozofické fakultě Univerzity Karlovy v Praze (ÚČNK). Od svého založení roku 1994 má ÚČNK na starosti budování ČNK, jeho rozvoj a rovněž činnosti související, zvláště v oblasti výuky a přestování oboru korpusová lingvistika.




Aktuality

- Pozvánka na konferenci**
Ve dnech 22.-24. září 2011 proběhne na Filozofické fakultě UK konference *Korpusová lingvistika Praha 2011*. Více informací naleznete [zde](#).
- Perfektum v současné češtině**
20. 9. 2010 vyšla v řadě Studie z korpusové lingvistiky publikace *Perfektum v současné češtině*. Autorkou knihy je Mira Načeva-Marvanová. [Více...](#)
- Nové publikace**
Koncem května 2010 byly vydány tyto publikace: *Lexikon a sémantika* Františka Čermáka a *Čeština jak ji neznáte* Věry Schmiedtové.
- Korpus SYN2009PUB**
Dne 7. května 2010 byl zveřejněn korpus **SYN2009PUB**. Jedná se o dosud největší korpus české publicistiky o velikosti 700 milionů textových slov, který v mnoha ohledech navazuje na svého předchůdce, korpus **SYN2006PUB**.
- Dnešní skloňování substantiv typů kámen, břímě**
20. 4. 2010 vyšla v řadě Studie z korpusové lingvistiky publikace *Dnešní skloňování substantiv typů kámen, břímě*. Autorem knihy je Josef Šimandl. [Více...](#)
- Mluvnice současné češtiny**
8. 4. 2010 vyšla Mluvnice současné češtiny. Jedná se o kolektivní dílo autorů z FF UK a MFF UK pod vedením Václava Cvrčka.


Хорватский национальный корпус

Hrvatski nacionalni korpus

www.hnk.ffzg.hr




The screenshot shows the website for the Hrvatski nacionalni korpus. The page has a white background with a blue sidebar on the left. The main content area is on the right. The title 'Hrvatski nacionalni korpus' is displayed in a large, serif font, with 'Hrvatski' in red and 'nacionalni korpus' in blue. Below the title, there are three sections: 'Što je Hrvatski nacionalni korpus?', 'Komu je korpus namijenjen?', and 'Privremeni pristup'. Each section has a heading in red and blue, followed by a paragraph of text. The sidebar contains a navigation menu with links to 'naslovnica', 'korpus', 'struktura i izvori', 'pretraga', 'radovi', 'o nama', and 'poveznice'. There are also logos for 'Hrvatski lematizacijski poslužitelj' and 'hobs' (Hrvatska ovisnosna banka stabala). At the bottom of the sidebar, there is a logo for 'Portal jezičnih tehnologija za hrvatski jezik'. The footer of the page contains the copyright information: 'Copyright © 2005 HNK, Zavod za lingvistiku.'


  **Hrvatski nacionalni korpus**


[naslovnica](#)
[korpus](#)
[struktura i izvori](#)
[pretraga](#)
[radovi](#)
[o nama](#)
[poveznice](#)

English

 Hrvatski lematizacijski poslužitelj

 hobs
Hrvatska ovisnosna banka stabala

 Portal jezičnih tehnologija za hrvatski jezik



Što je Hrvatski nacionalni korpus?

Hrvatski nacionalni korpus (HNK) usustavljena je zbirka odabranih tekstova pretežito suvremenoga hrvatskoga jezika koji pokrivaju razne medije, žanrove, stilove, područja i tematiku. Sâm je korpus, popraćen dodatnim lingvističkim i nelingvističkim podacima, smješten u datobazu kojoj se pristupa s pomoću programa za pretraživanje [Bonito](#).

Komu je korpus namijenjen?

HNK je javno i slobodno dostupan za istraživanje, obrazovanje i ostale nekomercijalne uporabe. Komercijalni se korisnici moraju predbilježiti kako bi dobili svoj korisnički račun i zaporku.

Korpus je objavljen u onom obliku koji je i dostupan za pretraživanje, a istodobno je podložan dopunama bez naknadne posebne obavijesti korisnicima. Popis (novih) izvora bit će uvijek dostupan korisnicima.

Privremeni pristup

Privremeno, dok se korpus još sastavlja, omogućen je pristup bez posebne predbiljezbe već je dovoljna uporaba korisničkoga računa [gost](#) i bez [zaporke](#). Takav pristup, međutim, ograničen je s obzirom na mogućnost stvaranja *ad hoc* potkorpusa prema željenim kriterijima kao i mogućnost složenijih oblika pretraživanja.

Koliko je HNK velik?

Trenutačno HNK obasiže 101,3 milijuna pojava.

Copyright © 2005 HNK, Zavod za lingvistiku.

3 Типология корпусов

Основные способы деления

корпусов

- противопоставление корпусов, относящихся ко всему языку, корпусам, относящимся к какому-либо подязыку (жанр, стиль, язык определенной возрастной или социальной группы, язык писателя и т.п.);
- разделение корпусов по типу лингвистической разметки.

Признак	Типы корпусов
Тип данных	Письменные Устные Смешанные
Язык текстов	Русский Английский и т.д.
«Параллельность»	Одноязычные Двухязычные Многоязычные
«Литературность», специфичность	Литературные Фольклорные Публицистические Диалектные Разговорные Драматургические Терминологические Смешанные

Признак	Типы корпусов
Доступность	Свободно доступные Коммерческие Закрытые
Назначение	Исследовательские Иллюстративные
Разметка	Размеченные Неразмеченные
Характер разметки	Морфологические Синтаксические Семантические и т.д.
Динамичность	Динамические Статические
Объем текстов	Полнотекстовые «Фрагментнотекстовые»
Хронологический аспект	Синхронические Диахронические

Полезные сайты

- <http://corpora.iling.spb.ru>
 - <http://www.dialog-21.ru>
 - <http://corpling-ran.ru>
 - <http://ruscorpora.ru>
 - <http://rykov-cl.narod.ru>
 - <http://vp-zakharov.narod.ru>
- 