

Обзор корпусов.

В.П. Захаров

**Санкт-Петербургский
государственный университет**



Классификация корпусов (1)

□ Два основных способа деления корпусов на классы:

1) это противопоставление корпусов, относящихся ко всему языку (часто к языку определенного периода), корпусам, относящимся к какому-либо подязыку (жанр, стиль, язык определенной возрастной или социальной группы, язык писателя или ученого и т.п.);

2) разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа (treebanks, «банки синтаксических структур»).

Классификация корпусов (2)

Для анализа и сравнения корпусов необходим набор признаков-характеристик. В частности:

- Тип данных
- Язык текстов
- «Параллельность»
- «Литературность»
- Специфичность
- Жанр
- Доступность
- Назначение
- Динамичность
- Разметка
- Характер разметки
- Объем текстов
- Хронологический аспект
- «Общность»
- Структура

Корпусы в Интернет (1)

Национальный корпус русского языка http://ruscorpora.ru	70 млн слов
Компьютерный корпус текстов русских газет конца XX-го века http://www.philol.msu.ru/~lex/corpus	200 тыс. слов
Корпус русского языка ХАНКО (Хельсинский университет) http://www.ling.helsinki.fi/projects/hanco/	100 тыс. слов Ручная морфологическая разметка
Корпуса русских текстов на сайте Университета в Лидсе, Великобритания http://corpus.leeds.ac.uk	
Русские корпуса Тюбингенского Университета http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html	
Словарь-корпус языка А.С. Грибоедова http://www.inforeg.ru/electron/concord/concord.htm	120 тыс. слов

Корпусы в Интернет (2)

Уппсальский корпус русских текстов Доступен для поиска на сайте http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html	1 млн слов 600 текстов (публицистика 1985-1989; литературные произведения 1960-1988)
Банк английского языка (Bank of English) http://www.collins.co.uk/books.aspx?group=153 Свободный доступ: http://www.collins.co.uk/Corpus/CorpusSearch.aspx	524 млн слов, 56 млн в свободном доступе (The Collins Wordbanks <i>Online</i> English corpus: 36 млн – брит. англ., 10 млн – амер. англ., 10 млн – брит. разговорн. англ.)
Британский национальный корпус http://www.natcorp.ox.ac.uk/ или http://sara.natcorp.ox.ac.uk/	100 млн слов Корпусные менеджеры SARA и XAIRA (http://www.xaira.org)
Венгерский национальный корпус http://corpus.nytud.hu/mnsz/	100 млн слов

Корпусы в Интернет (3)

Корпус испанского языка (исторический) http://www.corpusdelespanol.org/	100 млн слов, тексты 13–20 вв. Создан в Иллинойском университете, США
Корпус современного датского языка http://www.korpus2000.dk/	50 млн слов Тексты 1998–2002 гг.
Корпус современного итальянского языка CORIS/CODIS http://www.cilta.unibo.it/ricerca.htm	100 млн слов
Корпус современного китайского языка (LIVAC Synchronous Corpus) http://www.rcl.cityu.edu.hk/livac/	720 млн слов (150 млн иероглифов)
Мангеймский корпус немецкого языка (Institut für Deutsche Sprache, Mannheim, Germany) http://corpora.ids-mannheim.de/~cosmas/	1610 млн слов Корпусный менеджер COSMAS
Национальный корпус словенского языка http://www.fida.net/eng/	Более 100 млн слов

Корпусы в Интернет (4)

Польский национальный корпус http://korpus.ia.uni.lodz.pl/	93 млн слов
Словацкий национальный корпус http://korpus.juls.savba.sk	180 млн слов Используется корпусный менеджер Manatee/Bonito
Хорватский национальный корпус http://www.hnk.ffzg.hr/	53 млн слов Корпусный менеджер Manatee/Bonito
Чешский национальный корпус http://ucnk.ff.cuni.cz	100 млн слов + 100 млн нового корпуса современной лексики Корпусный менеджер Manatee/Bonito
Эстонский корпус http://test.cl.ut.ee/korpused/baaskorpus/1980/index.html.en	

Корпусы в Интернет (5)

http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/index2.html

W3-Corpora List of Corpora - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное Медиа

Адрес: http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/index2.html Переход Ссылки

INTRODUCTION

GLOSSARY

CORPORA

COURSES

BIBLIOGRAPHY

RELATED SITES

SOFTWARE

SEARCH ENGINE

TUTORIAL

COMMENTS

These pages have been created as part of the **W3-Corpora Project** at the **University of Essex**.

List of Corpora

- [Alphabetical list](#)
- [Historical Corpora](#)
- [Speech Corpora](#)
- [Languages other than English](#)
- [Available on-line](#)
- [Other text resources](#)

Not all resources listed here conform to a strict definition of a corpus (see, for ex., [this](#)). They have, nevertheless, been included for the convenience of those who need them.

For an interesting approach that uses the whole World Wide Web as a corpus, see <http://www.webcorp.org.uk/>.

Alphabetical list

- [The Air Traffic Control Corpus](#)
- [ACL/DCI](#) Association for Computational Linguistics Data Collection Initiative
- [ATIS](#) Air Travel Information System (Held by [LDC](#))
- [Bank of English](#)
- [BNC](#) The British National Corpus
- [The Brown Corpus](#) 1 million word American English, standard reference
- [CALLFRIEND Collection](#) Unscripted telephone conversations in 12 languages and 3 dialects. (Held by [LDC](#))
- [CALLHOME Collection](#) Unscripted telephone conversations in six languages (Held by [LDC](#))
- [CCAT Archive](#) Classical, Literary, Historical and Religious Texts.
- [The CHILDES](#) Child Language Data Exchange Systems

Интернет

Национальный корпус русского языка (ruscorpora.ru)

- предыстория Проекта;
- текущие задачи Проекта;
- состав рабочих групп;
- источники текстов;
- поисковая система;
- дизайн и поддержка сайта

Национальный корпус русского языка (2)

The screenshot shows a web browser window titled "О проекте. Национальный корпус русского языка - Microsoft Internet Explorer". The address bar shows "http://ruscorpora.ru/corpora-about.html". The website has a blue header with the logo and text "НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА". A left sidebar contains a menu with items: "главная", "архив новостей", "поиск в корпусе", "что такое корпус?", "состав и структура", "статистика", "морфология", "семантика", "параметры текстов", "о проекте", "участники проекта", "программные средства", and "использование корпуса". The main content area is titled "О проекте" and contains a list of links: "Предыстория Проекта", "Текущие задачи Проекта", "Состав рабочих групп", "Источники текстов", "Поисковая система", and "Дизайн и поддержка сайта". Below this is a section titled "Информация об участниках Проекта" with a paragraph about the project's goals and participants. At the bottom is a section titled "Предыстория Проекта" with a paragraph about its origin.

О проекте

- [Предыстория Проекта](#)
- [Текущие задачи Проекта](#)
- [Состав рабочих групп](#)
- [Источники текстов](#)
- [Поисковая система](#)
- [Дизайн и поддержка сайта](#)

[Информация об участниках Проекта](#)

В создании Национального корпуса русского языка в настоящее время участвует большая группа лингвистов Москвы, Санкт-Петербурга и других научных центров России, объединенная в Ассоциацию «Национальный корпус русского языка». В Ассоциацию может войти любой специалист (или научный коллектив), желающий внести вклад в создание Национального корпуса русского языка и согласный сделать результаты своей работы над корпусом общедоступными.

Предыстория Проекта

Инициатива создания большого представительного корпуса русского языка принадлежит С. А. Шарову и В. Б. Касевичу, предложившему, в частности, нынешнее название «Национальный корпус русского языка». Эта инициатива возникла как закономерное развитие исследований по

Поиск в НКРЯ (1)

- Поле «Слово»
- Поле «Грамматические признаки»
- Поле «Семантические признаки»
- Расстояние между словами

Поиск в НКРЯ (2)

Поиск в корпусе. Национальный корпус русского языка - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное

Адрес: <http://ruscorpora.ru/search-main.html> Переход Ссылки

Основной корпус Параллельный корпус

[задать подкорпус](#)

главная
архив новостей
поиск в корпусе
что такое корпус?
состав и структура
статистика
морфология
семантика
параметры текстов
о проекте
участники проекта
программные средства
использование корпуса

Поиск точных форм

Слово или фраза

искать очистить

Лексико-грамматический поиск

Слово 1 грамм. признаки выбрать семант. признаки выбрать

Расстояние, в словах: от 1 до 1

Слово 2 грамм. признаки выбрать семант. признаки выбрать

искать очистить

Национальный корпус русского языка
© 2003-2006

Поиск осуществляется системой [Яндекс.Сервер](#)

Готово

пуск Total Commander 6.0... (C:\JUPC) - Far Поиск в корпусе. На... EN 14:20

Британский национальный корпус (BNC)



The LIVAC (Linguistic Variations in Chinese Speech Communities) synchronous corpus

LIVAC Online - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное Медиа

Адрес: <http://www.rcl.cityu.edu.hk/livac/index.php?lang=tc>

English | 繁體 | 簡體

LIVAC ID:

Password:

Log In

LIVAC共時語料庫簡介

中心消息

- 2005年兩岸四地(港台京滬)中文新詞彙
- 2005年名人榜

香港城市大學 語言資訊科學中心及其語料庫實驗室, 自1995年開始, 以「共時」方式處理了超常的大量漢語語料, 通過精密的技術, 累積眾多精確的統計數據, 建立了LIVAC (Linguistic Variation in Chinese Speech Communities)共時語料庫。

本語料庫最大特點是採用「共時性」視窗模式, 嚴謹地定時分別收集來自多地的定量同類語料, 可供各種客觀的比較研究, 方便有關的信息科技發展與應用。此外, 語料庫又兼顧了「歷時性」, 方便各方人士客觀地觀察與研究視窗內的有代表性的語言發展全面動態。

建立LIVAC語料庫的目標和意義, 在於探討新語言形式的產生和發展、泛華語地區包括大陸及海外多個地區各自的語言變化及其之間的關係。尤其特別探討了自1995年以來湧現的漢語外來詞及新興詞, 以及在語法等方面的發展新趨勢。

LIVAC語料自1995年開始, 主要來自香港、澳門、台北、新加坡、上海、北京、廣州、深圳多個地區的當地有代表性中文媒體, 涵蓋泛華語地區。搜集內容包括社評、重要新聞、當地新聞、綜合新聞、國際新聞、港澳台新聞、經濟新聞、體育新聞、娛樂新聞、廣告。每四天定量下載一次。有關語料經機器切詞標注、人工校對後, 提取各地詞語, 加入各地詞庫組合為LIVAC大語料庫。至2005年1月, 語料庫共收集72萬個詞條, 總字數超過1億5千萬字, 並仍不斷擴充中。

在大語料庫基礎上, LIVAC還建立了人名庫、地名庫、專名庫、動詞詞庫、形容詞詞庫、各地每月新詞詞庫等多個專項語料庫。LIVAC每兩星期計算公佈雙周「名人榜」, 年底公佈全年「名人榜」, 並建立人物褒貶指數的計算。

LIVAC所建立的是一個包括字、詞條、文句、全文等不同層次的語料庫, 可對個別字或詞的使用作查詢、頻率統計比較。本語料庫適合多方面研究人員使用, 包括從事語言學研究或有意對語言現象與華人社會組織、文化與動態發展作探討者, 以及從事發展資訊搜索引擎與機器翻譯等語言工程的學者等等。

Корпус польского языка (1)

The PELCRA Reference Corpus of Polish - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное Медиа

Адрес: <http://www.ebi.ac.uk/~pezik/korpus/> Переход Ссылки



PELCRA
The Reference Corpus of Polish

This site was developed by the Department of English Language at the University of Łódź, Poland (project reference number: KBN nr 2 H01D 008 25)

PELCRA Reference Corpus Conversational Corpus PELCRA News BIT Dictionary Wersja Polska

Public login and password: [corpus](#) password: [corpus](#).

Menu

Concordancer

- Word Concordancer - start here!
- Word Concordancer
- Inflections - **NEW!**
- Phrase Concordancer
- Phrase+
- Sentence Concordancer
- Paragraph Concordancer

Statistical Analysis

- Word Frequency
- Simple Collocations
- MI3 Collocations

Meta Data

- Authors
- Texts

Misc.

- News

Welcome, .

Limit: 0

Login:

Password:

General Information

Number of Tokens:	93,129,588
Number of Texts:	81,056
Number of Authors:	13,554
Number of Sentences:	6,002,420
Number of Paragraphs:	2,504,375

2005-09-12

Welcome to the home site of the PELCRA corpus of Polish. Currently, 93,000,000 words are available for searching. **An updated Polish version of a guide to using the corpus interface (version 19-09-2005)** is [available here](#). Watch this space!

To use the different search capabilities of our web interface, you need to login using the following details:

login: corpus
password: corpus

[PELCRA](#) Team 2005

Links

Corpora

- British National Corpus
- Bank Of English
- W3-Corpora

Organizations

- PELCRA
- ICAME
- Korpus IPI PAN
- PWN
- UCREL
- CL Lab - AGH

Tools

- WordSmith Tools

<http://www.ebi.ac.uk/~pezik/korpus/conversational/>

Корпус польского языка (2)

PNC - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное Медиа

Адрес: http://korpus.ia.uni.lodz.pl/pnc/concordancer/concordancer_result_word.php Переход Ссылки >>

Portal opracowany w ramach projektu KBN nr 2 H01E

Site under construction - to try this test version, use login: **corpus** password: **corpus**.

Polish National Corpus | Conversational and Multimedia Corpus | PELCRA | FAQ | BiT Dictionary | Wersja Polska

12 Result(s) for: herbata and 247 instance(s) exist in the entire corpus
Click here to try again

#	Left Side	Word	Right Side
1	Pobudka o świcie ,	herbata +	na stojąco z wyszczerbionego kubka , święte godziny pracy ,
2	ser , szynka , dżem , miód , kawa ,	herbata +	, mleko .
3		HERBATA	
4	W rzeczywistości konsumpcja jest jeszcze wyższa , bo	herbata +	każdego chińskiego posiłku .
5	Na stoliku pojawiła się słomkowa	herbata +	liraszka wódki .
6	warzywa zdrożały o 4 ,7 proc . , kawa i	herbata +	- o 3 ,4 proc . , a jaja -
7	bo na światowych rynkach tanieją produkty tropikalne , takie jak	herbata +	czy kakao .
8	rynkowych przebojów , na której znalazła się na przykład mrożona	herbata +	Liptona w butelkach , płatki śniadaniowe Post Banana Nut Crunch
9	Na śniadanie chleb z dżemem ,	herbata +	, na obiad jajko w sosie chrzanowym i ziemniaki ,
10	Z obozów pochodzą głównie ryż ,	herbata +	, bawełna , wydobywa się węgiel i rudy metali ,
11	- Nie dopita	herbata +	w cudzym domu .
12		herbata +	ir do nauczycielki: □ Proszę pani , jaka forma jest

Information
Gazeta Wyborcza nr 210, z dnia
1993/09/08, str. 19:
Publish Date: 1993-01-01

Polish National Corpus | Conversational and Multimedia Corpus | PELCRA | FAQ | BiT Dictionary | Wersja Polska

Copyright 2004 PELCRA All Rights Reserved

Ошибка на странице. Интернет

Словацкий национальный корпус (SNK)

Úvod - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное Медиа

Адрес: <http://korpus.juls.savba.sk>

Переход Ссылки

Slovenský národný korpus

HRádať: na stránke citovať v korpuse

Menu

- [Úvod](#)
- [Čo je to korpus?](#)
- [Štruktúra SNK](#)
 - [Všeobecný \(primárny\) korpus](#)
 - [Lingvistické zdroje](#)
 - [Lingvistický korpus krymskotatarskeho jazyka](#)
 - [Paralelné korpusy](#)
- [Ako používať korpus?](#)
 - [Korpusový manažér](#)
 - [Regulárne výrazy](#)
 - [Štylovo-žánrová anotácia](#)
 - [Morfologická anotácia](#)
 - [Ako citovať korpus](#)
- [Dokumenty](#)
- [Poskytovatelia textov](#)
- [Štatistiky korpusu](#)
- [Publikácie](#)

Úvod

Slovenský národný korpus (SNK) je databáza súčasných textov v slovenskom jazyku, zahŕňajúca široké spektrum jazykových štýlov spolu s prídavnými jazykovednými informáciami a výkonným vyhľadávacím systémom.

Korpus je určený pre verejnosť na výskumné, učebné a iné výlučne nekomerčné účely. Nenahrádza kodifikačné príručky. Prístup do korpusu možno získať [zaregistrovaním sa](#).

Alternatívne možno v korpuse vyhľadávať cez jednoduchý WWW interface a to bez osobného konta, no s veľmi obmedzenými možnosťami (bez štatistiky a ďalších charakteristík, počet konkordancií je obmedzený). Pred používaním WWW interface musíte súhlasiť s nasledujúcimi podmienkami:

- » Texty sa do Slovenského národného korpusu (SNK) získavajú priamo od autorov alebo vlastníkov autorských práv s cieľom ich využitia na iné, t.j. nekomerčné použitie diel v súlade so zákonom o autorských právach. Ako používateľ sa zaväzujem, že texty budem používať len na vedecké, výskumné, učebné alebo iné nekomerčné ciele.
- » Budem citovať všetky údaje získané zo Slovenského národného korpusu (konkordancie, frekvencie a iné údaje zo všetkých alebo len z vybraných

Чешский национальный корпус (ČNK)

Bonito Mana?er Korpus Dotaz Konkordance Zobrazen? V?b?r N?pov?da

Nový dotaz jméno: Synek

" Kdepak . Zase se objeví **slunce** , které všechno projasní
Nesmírné ticho - po západu **slunce** . Vše je tiché , starobylé
 , oleje , masti , horské **slunce** - prostě veškeré pomůcky
 . Nízko ležící podvečerní **slunce** vrhalo na zeleň pastviny
 netušíte , jak já miluju východ **slunce** ! " " Ten neuvidíte . " "
dokonalém pořádku a odpolední **slunce** dopadalo dovnitř šikmo moderními
kosmické prachárny ; cvrlikáš za **slunce** východu , zatímco v sudech
kteří dohlíží na východ **slunce** . XXVI Prokop čekal , kdoví
slova . Za zámkem zapadá **slunce** v podzimním zlatě ; prořídle
 ; zvedá se mlha a zastře **slunce** ; a opět je šedý a chladný
Penegal v Alpách , když zapadá **slunce** . To ti pak zahoří sníh takovým
nocí a ráno ránem . Pro tebe **slunce** už ani nezapadá ani nevychází
židovská krev a španělské **slunce** vytvářely v její tváři zvláštní
trochu předstíraný , jako **slunce** a noc v uzavřeném vaku .
řeku a město . - Nahoře bylo **slunce** a létali rackové . " Moc
byl krásný , od rána plný **slunce** . Docela zapadla zpráva ,
ztrácejícím se břehu , pak **slunce** , jasné a veliké , v jehož
kterým protékalo předpolední **slunce** . Vracela se a nalétávala
 , jako v odrazu městského **slunce** . Maminka měla vlídný altový

Zobrazeno: 1+50/624 (8%) Řádek: 15 Vybráno: 1