ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА

Ровно 50 лет

Проблемы корпусной лингвистики

Лекция 4

Радченко Олег Анатольевич, д.ф.н., проф.

Корпус

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Пушкин 50 лет

- Традиционное представление
  - Коллекция текстов, как можно более полная и самодостаточная: *the corpus of Anglo-Saxon verse*

    The Oxford Companion to the English Language

- Современное представление
  - Коллекция текстов, созданных в естественной речевой среде и отобранных для того, чтобы охарактеризовать состояние или вариацию языка

    - John Sinclair Corpus Concordance Collocation OUP

Корпус и архив

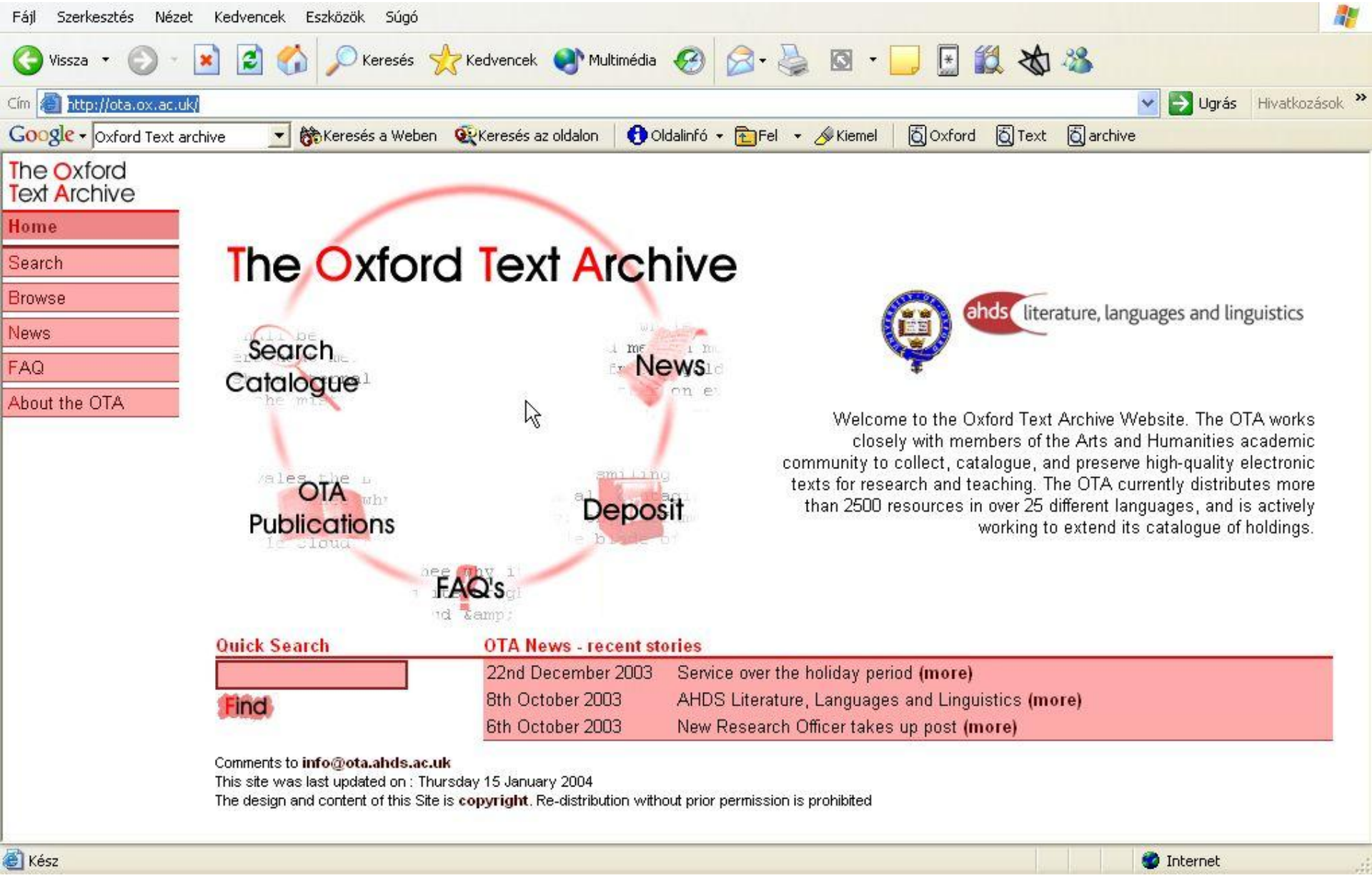
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Архив текстов
- Собрание текстов в их оригинальном формате

(Oxford Text Archive: http://ota.ox.ac.uk/)

- Корпус
- Тексты, отобранные и обработанные унифицированным, систематическим образом

British National Corpus: http://www.natcorp.ox.ac.uk/


Fájl Szerkesztés Nézet Kedvencek Eszközök Súgó
Vissza
Keresés
Kedvencek
Multimédia
Cím
http://ota.ox.ac.uk/
Ugrás
Hivatkozások
Google
Oxford Text archive
Keresés a Weben
Keresés az oldalon
Oldalinfó
Fel
Kiemel
Oxford
Text
archive
The Oxford Text Archive
Home
Search
Browse
News
FAQ
About the OTA
The Oxford Text Archive
Search Catalogue
News
OTA Publications
Deposit
FAQ's
ahds literature, languages and linguistics
Welcome to the Oxford Text Archive Website. The OTA works closely with members of the Arts and Humanities academic community to collect, catalogue, and preserve high-quality electronic texts for research and teaching. The OTA currently distributes more than 2500 resources in over 25 different languages, and is actively working to extend its catalogue of holdings.
Quick Search
Find
OTA News - recent stories
22nd December 2003 Service over the holiday period (more)
8th October 2003 AHDS Literature, Languages and Linguistics (more)
6th October 2003 New Research Officer takes up post (more)
Comments to info@ota.ahds.ac.uk
This site was last updated on : Thursday 15 January 2004
The design and content of this Site is copyright. Re-distribution without prior permission is prohibited
Kész
Internet

Fájl Szerkesztés Nézet Kedvencek Eszközök Súgó
Vissza
Keresés
Kedvencek
Multimédia
Cím
http://www.natcorp.ox.ac.uk/
Ugrás
Hivatkozások
Google
British National Corpus
Keresés a Weben
Keresés az oldalon
Oldalinfó
Fel
Kiemel
British
National
Corpus
Előzmények
Nézet
Keresés
3 hete
2 hete
Múlt héten
Ma
BRITISH NATIONAL CORPUS
What's New
Introduction
Ordering the BNC
Using the BNC
SARA
Corpora Page
Search the web site
Online Service
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.
We are now taking orders for the second edition of the BNC, which is available worldwide. Details are available here
The BNC Online Service and the BNC Sampler are now both available worldwide.
To navigate the site you can use the buttons opposite, or search for specific information on the BNC web site by using the search button.
British National Corpus
Oxford University Computing Services
13 Banbury Road
Oxford OX2 6NN
tel. +44 (1865) 273 221
fax +44 (1865) 273 275
enquiries and bug reports: natcorp@oucs.ox.ac.uk
Internet

Отто Есперсен (1860-1943)

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- „A Modern English Grammar on Historical Principles“ (1909-1949)
- Тысячи примеров из произведений Чосера, Шекспира, Свифта, Остин и др., выписанных автором
- Недостатки такого подхода

Что такое корпус?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Corpus (pl. corpora) = ‘тело’
- Коллекция письменных текстов и транскрибированной устрой речи
- Обычно, но не обязательно, составляется с конкретной целью
- Обычно, но не обязательно, структурирован
- Обычно, но не обязательно, аннотирован
- (Обычно хранится на компьютере и доступен с него)

“Для определенной цели”

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Образцы текстов отбираются для решения конкретной научной задачи
- Корпус может быть ориентирован на определенный жанр (corpus of newswire texts) или быть более широким
- Часто для корпуса важен аспект сбалансированности
  - Демографические факторы (возраст, пол, место проживания, социальное положение автора или говорящего)
  - Разнообразные стили и жанры

“Структурирован”

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- В целом корпус делится на секции по определенным параметрам
- Баланс должен обеспечить представленность в корпусе разных жанров и демографических факторов

Параметры в BNC (письменная часть)

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Нам 50 лет

Publication Date:

- [ ] 1960-1974
- [ ] 1975-1984
- [ ] 1985-1993

Medium of Text:

- [ ] Book
- [ ] Periodical
- [ ] Miscellaneous: published
- [ ] Miscellaneous: unpublished
- [ ] To-be-spoken

Text Sample:

- [ ] Whole text
- [ ] Beginning sample
- [ ] Middle sample
- [ ] End sample
- [ ] Composite

Domain:

- [ ] Imaginative prose
- [ ] Informative: Natural and pure sciences
- [ ] Informative: Applied science
- [ ] Informative: Social science
- [ ] Informative: World affairs
- [ ] Informative: Commerce and finance
- [ ] Informative: Arts
- [ ] Informative: Belief and thought
- [ ] Informative: Leisure

Derived text type:

- [ ] Academic prose
- [ ] Fiction and verse
- [ ] Non-academic prose and biography
- [ ] Newspapers
- [ ] Other published written material
- [ ] Unpublished written material

Estimated Circulation Size:

- [ ] Low
- [ ] Medium
- [ ] High

Perceived Level of Difficulty:

- [ ] Low
- [ ] Medium
- [ ] High

Domicile of Author:

- [ ] UK and Ireland
- [ ] Commonwealth
- [ ] Continental Europe
- [ ] USA
- [ ] Elsewhere

Age of Author:

- [ ] 0-14
- [ ] 15-24
- [ ] 25-34
- [ ] 35-44

Sex of Author:

- [ ] Male
- [ ] Female
- [ ] Mixed

Type of Author:

- [ ] Corporate
- [ ] Multiple
- [ ] Sole

Target Audience Age:

- [ ] Child
- [ ] Teenager
- [ ] Adult
- [ ] Any

Target Audience Sex:

- [ ] Male
- [ ] Female
- [ ] Mixed

Структура жанров в BNC (письменная часть)

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Genre:

- ☐ W:ac:humanities_arts
- ☐ W:ac:medicine
- ☐ W:ac:nat_science
- ☐ W:ac:polit_law_edu
- ☐ W:ac:soc_science
- ☐ W:ac:tech_engin
- ☐ W:admin
- ☐ W:advert
- ☐ W:biography
- ☐ W:commerce
- ☐ W:email
- ☐ W:essay:school
- ☐ W:essay:univ
- ☐ W:fict:drama
- ☐ W:fict:poetry
- ☐ W:fict:prose
- ☐ W:hansard
- ☐ W:institut_doc
- ☐ W:instructional
- ☐ W:letters:personal
- ☐ W:letters:prof
- ☐ W:misc
- ☐ W:news_script
- ☐ W:newsp:brdsht_nat:arts
- ☐ W:newsp:brdsht_nat:commerce
- ☐ W:newsp:brdsht_nat:editorial
- ☐ W:newsp:brdsht_nat:misc
- ☐ W:newsp:brdsht_nat:report
- ☐ W:newsp:brdsht_nat:science
- ☐ W:newsp:brdsht_nat:social
- ☐ W:newsp:brdsht_nat:sports
- ☐ W:newsp:other:arts
- ☐ W:newsp:other:commerce
- ☐ W:newsp:other:report
- ☐ W:newsp:other:science
- ☐ W:newsp:other:social
- ☐ W:newsp:other:sports
- ☐ W:newsp:tabloid
- ☐ W:non_ac:humanities_arts
- ☐ W:non_ac:medicine
- ☐ W:non_ac:nat_science
- ☐ W:non_ac:polit_law_edu
- ☐ W:non_ac:soc_science
- ☐ W:non_ac:tech_engin
- ☐ W:pop_lore
- ☐ W:religion

Параметры в BNC (устная часть)

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Наши 50 лет

General Restrictions for Spoken Texts:

Overall:	Interaction Type:	Region where Spoken Text was Captured:
☐ Spoken demographic ☐ Spoken context-governed	☐ Monologue ☐ Dialogue	☐ South ☐ Midlands ☐ North

Genre:

☐ S:brdcast:discussn ☐ S:brdcast:documentary ☐ S:brdcast:news ☐ S:classroom ☐ S:consult ☐ S:conv ☐ S:courtroom ☐ S:demonstratn	☐ S:interview ☐ S:interview:oral_history ☐ S:lect:commerce ☐ S:lect:humanities_arts ☐ S:lect:nat_science ☐ S:lect:polit_law_edu ☐ S:lect:soc_science ☐ S:meeting	☐ S:parliament ☐ S:pub_debate ☐ S:sermon ☐ S:speech:scripted ☐ S:speech:unscripted ☐ S:sportslive ☐ S:tutorial ☐ S:unclassified

Restrictions for Spoken Demographic Texts:

These restrictions refer to the whole text and are not the same as the speaker restrictions below. Please check the BNC user manual if you need further assistance.

Age of Respondent (not of Speaker!):	Social Class of Respondent:	Sex of Respondent:
☐ 0-14 ☐ 15-24 ☐ 25-34 ☐ 35-44 ☐ 45-59 ☐ 60+	☐ AB ☐ C1 ☐ C2 ☐ DE	☐ Male ☐ Female

Restrictions for Spoken Context-governed Texts:

Domain:

- ☐ Educational/Informative
- ☐ Business
- ☐ Public/Institutional
- ☐ Leisure

Speaker Restrictions:

Age:

- ☐ 0-14
- ☐ 15-24
- ☐ 25-34
- ☐ 35-44
- ☐ 45-59
- ☐ 60+
- ☐ Unknown

Sex:

- ☐ Male
- ☐ Female
- ☐ Unknown

Social Class:

- ☐ AB
- ☐ C1
- ☐ C2
- ☐ DE
- ☐ Unknown

Education:

- ☐ Still in education
- ☐ Left school 14 or under
- ☐ Left school 15/16
- ☐ Left school 17/18
- ☐ Education continued until 19 or over
- ☐ Information not available

First Language:

- ☐ British English
- ☐ North American English
- ☐ Unknown Indian language
- ☐ German
- ☐ French
- ☐ Unknown

Dialect/Accent (a rather unreliable category!):

- ☐ Canada
- ☐ German
- ☐ East Anglia
- ☐ French
- ☐ Home Counties
- ☐ Humberside
- ☐ Irish
- ☐ Indian subcontinent
- ☐ Lancashire
- ☐ London
- ☐ Central Midlands
- ☐ Merseyside
- ☐ North-east Midlands
- ☐ Midlands
- ☐ South Midlands

North-west Midlands

- ☐ Central northern England
- ☐ North-east England
- ☐ Northern England
- ☐ Scottish
- ☐ Lower south-west England
- ☐ Central south-west England
- ☐ Upper south-west England
- ☐ European
- ☐ American (US)
- ☐ Welsh
- ☐ West Indian

Other or unidentifiable

Параметры в BNC (устная часть)

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Наши 50 лет

“Аннотированный”

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Не просто текст
- Большинство корпусов имеет аннотацию “POS”
  - Каждое слово снабжено информацией о его части речи (POS)
  - Тэги (ярлыки) POS содержат также богатую морфологическую информацию
  - Тэги по возможности снимаю грамматическую омонимию
- Некоторые корпусы содержат и иную информацию:
  - Структурную, исходя из делимитации текстов
  - Смысловую для различения грамматических омонимов

http://personalpages.manchester.ac.uk/staff/harold.somers/ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://bnc.humanities.manchester.ac.uk/cgi-bnc/processQuery.pl?theData=linguistics&chunk=1&queryType=CQL&qMode=Simple+query+%28case+insensitive%29&inst Go Links

Your query "linguistics" returned 784 matches in 100 different texts (in 98,313,429 words; frequency: 7.97 instances per million words) [0.304 seconds]

No	Filename	Solution 1 to 50 Page 1 / 16
1	A04 364	Literary theory, drawing on other disciplines, including semiotics and **linguistics**, seeks for underlying structures and meanings in literature.
2	A0T 1000	The functionalist approach to the study of mind characterizes much of the work currently being done in cognitive psychology, artificial intelligence and **linguistics**.
3	A1A 84	The series aimed to introduce to English readers la nouvelle critique and literary theory, along with associated areas such as **linguistics**, translation, and the study of mass culture.
4	A1A 258	Tallis's case is convincing, though he has certainly not said the last word on the matter, and there may well be further arguments from those who are professionally engaged in **linguistics**.
5	A1A 434	Such theories, he remarks, 'are not centrally concerned with literature; indeed, they may marginalize or abandon it as a category', and he instances their source in such fields as philosophy, psychology, sociology, anthropology, **linguistics**.
6	A1A 953	Many of them would do perfectly well in disciplines which are less idiosyncratic than English and have a more obvious international currency, like **linguistics** or economics or marketing.
7	A1A 1023	I once asked a colleague in **linguistics** if this meant that the criticism of, oh, Addison, Keats, Hopkins, Forster, was valueless.
8	AMG 165	Here no single-factor theory of language will do, for brain size and anatomy, symbolic coding **linguistics**, social system, and dietary distribution are all plausibly involved and operate in interaction during the evolution of language systems.
9	ANY 629	New ideas imported from Paris by the more adventurous young teachers glittered like dustmotes in the Fenland air: structuralism and poststructuralism, semiotics and deconstruction, new mutations and graftings of psychoanalysis and Marxism, **linguistics** and literary criticism.
10	ANY 1552	Charles had given it to her for a joke, suggesting she use it as a visual aid to introduce Saussurean **linguistics** to first-year undergraduates, holding the tube aloft to demonstrate that what is onomatopoeia in one language community may be obscenity in another.

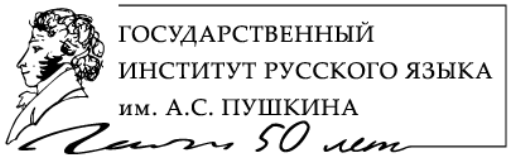
Unknown Zone
start 3 Internet E... 3 Windows E... Inbox - Micros... Document4 - M... Microsoft Pow... untitled - Paint 13:10

Создание корпуса: парсирование, присвоение токенов

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Предобработка
  - Сегментирование текстов на предложения
    - слова
    - Сложносоставные слова – проблема
  - Нормализация
    - Восстановление клитиков, аббревиатур ("can't", "I've")

Создание корпуса: аннотирование (tagging)



- Аннотирование
  - Придание каждому слову ярлыка с информацией о его части речи
  - Проблема: соответствие нескольким частям речи вне контекста
  - set N vs. set V

Создание корпуса: разрешение омонимии

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

- Disambiguation
  - Определение корректного анализа в контексте
- Два подхода:
- Оба нуждаются в мануально корректируемом пробном корпусе
  - статистический
    - Hidden Markov model
    - Расчет вероятности обычно в охвате одного-двух слов
    - Успешность может составлять до 98%
  - Основанный на правилах

Синтаксическое аннотирование

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

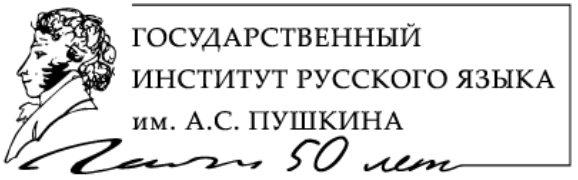
- Сложно создаваемое в таком масштабе
- Сложности делимитации
- Treebank:
  коллекция ситаксически проанализированных предложений
- Penn treebank
- http://www.cis.upenn.edu/~treebank/

Современные тенденции

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Word sense ambiguation (SENSEVAL)
  - http://www.itri.brighton.ac.uk/events/senseval/
- Message understanding
  - http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html
- SEMANTIC WEB
  - Превращение информации в Интернете в понятную для компьютера

Какой сэмпл считать репрезентативным?



- Корпус любого размера неизбежно является лишь иллюстрацией (сэмплом)
- Чего?
- Два подхода
  - Иллюстрация говорящих – demographic sampling
  - Иллюстрация порождаемых ими текстов – text type sample

Понятие репрезентативности

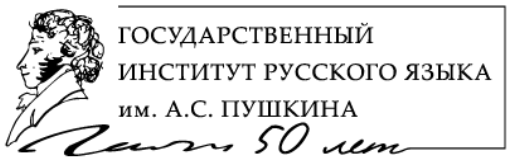
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Ганя 50 лет

- Сэмпл vs. население
- Сэмпл должен быть пропорционален населению относительно данной особенности
- *Пример демографического сэмплинга*

  Если мы знаем, что 48% населения Будапешта – мужчины, нам следует составлять корпус так, чтобы информация респондентов-мужчин составляла в нем 48%

  - Такой сэмплинг репрезентативен для города Будапешта с точки зрения гендерных особенностей

Проблемы репрезентативности



- Что должно быть единицей отбора для корпуса?
- Стили, типы текстов, жанры etc.
- Не существует независимых данных об их квоте в речевых произведениях

-> репрезентативность – идеал, который невозможно реализовать

Подходы к репрезентативности

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Douglas Biber (Regents' Professor, Applied Linguistics Program, at the English Department, Northern Arizona University)
- Отвергает пропорциональный сэмплинг
- Сэмплы должны быть как можно разнообразнее
- Репрезентативность измеряется в терминах широкой вариантивности типов текстов, включенных в сэмплы

Что такое корпусная лингвистика?

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- Не раздел лингвистики, типа социо~, психо~, …
- Не теория лингвистики
- Набор инструментов и приемов для поддержки лингвистических исследований по всем аспектам интересующего явления

Доказательство в лингвистике

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА

- Реально зафиксированное использование как доказательство в лингвистике
- Отличается от прежде распространенной интроспекции
- Связано с различением competence~performance (langue~parole)
- Корпусная лингвистика чаще более заинтересована в установлении тенденций, чем правил (возможности более, чем уверенности)
- Информация корпусов иногда противоречит распространенным представлениям о языковых фактах.

Для чего нужна корпусная лингвистика?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- 1. Исследование грамматических явлений (различий между модальными глаголами, отрицания, приложений, инфинитивных оборотов и пр.) – не нужны большие корпуса
- 2. Создание грамматических справочников: первые грамматики на основе корпусов Quirk et al. „A Grammar of Contemporary English" (1972), "A Comprehensive Grammar of the English Language" (1985) (Lonodn Corpus); "Oxford English Grammar" (1996) – ICE-GB, "Longman Grammar of Spoken and Written English" (1999) – Longman Corpus

Для чего нужна корпусная лингвистика?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Ему 50 лет

- 3. Лексикография: проблема частотности, большие корпуса, COBUILD, Bank of English Corpus (использован для BBC English Dictionary), British National Corpus (для Longman Dictionary of Contemporary English), Cambrigde Int. Corpus (для Cambridge Int. Dictionary of English)
- Проблема создания словарей без корпуса: Oxford English Dictionary, создавался 50 лет, основная работа вручную, множество стадий, 5 млн. цитат
- Возможность нахождения новых значений (risk в работе Ч. Филлмора, 1992 г.)

Для чего нужна корпусная лингвистика?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

- 4. Исследование языковой вариативности: социолингвистические работы, гендерная проблематика (термин lovely у Aston/Burnard 1998)
- 5. Историческая лингвистика: Helsinki Corpus древне- и среднеанглийских текстов с 8 по 17 вв. (1,5 млн. слов), разбиты по эпохам, содержат диалектную и гендерную информации (header!); корпус ARCHER (1,7 млн. слов, тексты 1650-1990, американский и британский варианты, различные жанры), корпуса отдельных произведений (Беовульф), авторов (Чосер), ранних английских писем и трактатов и т.п.

Для чего нужна корпусная лингвистика?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

- 6. Контрастивная лингвистика и теория перевода: параллельные корпуса, English-Norwegian Parallel Corpus (тексты беллетристики на английском и новержском языках и их переводы, 10000-15000 слов каждый). Возможности: изучение жанровых особенностей в двух языках, типичных переводов и ошибок

Для чего нужна корпусная лингвистика?

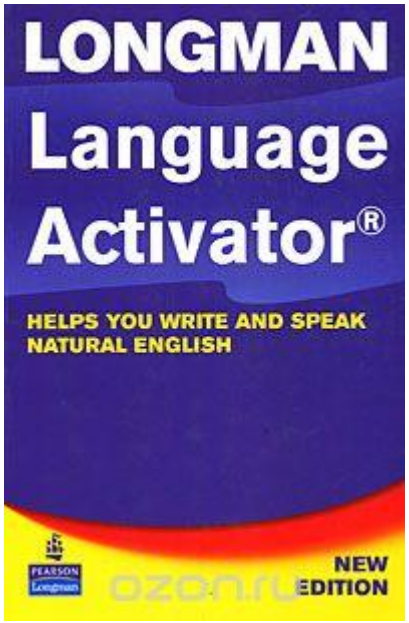
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

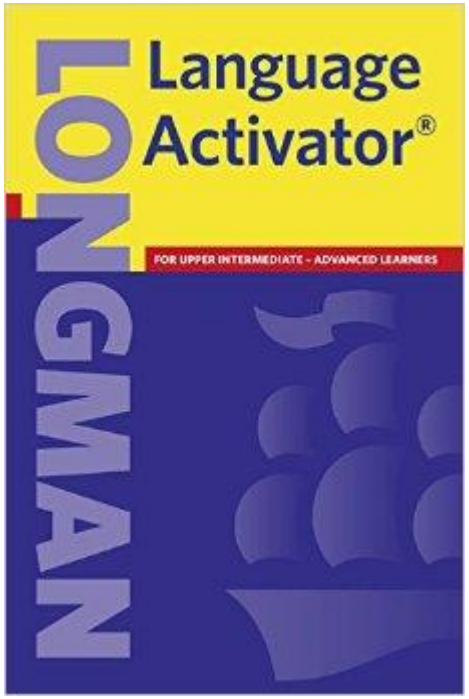
- 7. Исследование детской речи: CHILDES (для изучения усвоения родного и иностранного языков, транскрибированная речь детей и взрослых, 20 национальностей, речь детей с нормальным развитием и с афазией или аутизмом); Learner Corpora: ICLE (int. Corpus of Learner English) (2 млн. слов, эссе по 500 слов, написанные изучающими английский язык представителями 14 наций), the Longman Learner Corpus, The Hong Kong University of Science and Technology Learner Corpus

Для чего нужна корпусная лингвистика?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

- 8. Лингводидактика: Longman Essential Activator (1997)



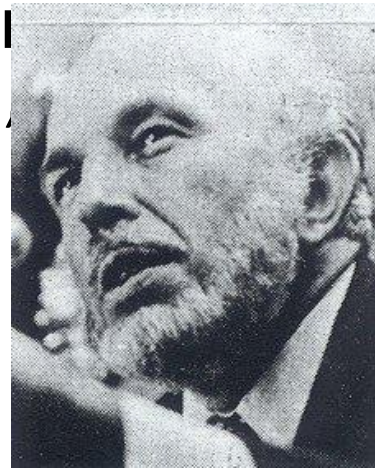


История корпусной лингвистики

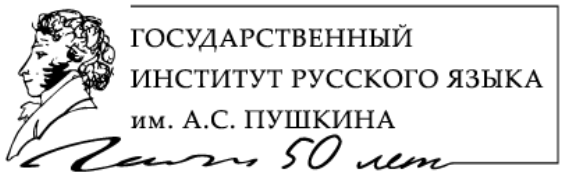
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Наши 50 лет

- 1950-е гг.: Р. Бус, корпус текстов Фомы Аквинского (10000 предложений на карточках + составленный вручную индекс, перенесены затем на перфокарты), в 1949-1967 гг. корпус насчитывал 10.600.000 слов, дополнительный корпус – 5.000.000 слов текстов на русском, немецком и арамейском языках.
- 1956-1970-е гг.: корпус машинных текстов А. Джилланда, 500.000 слов, сопоставительные исследования – частотность употреблени словоформ во французском, румынском, испанском и китайском языках

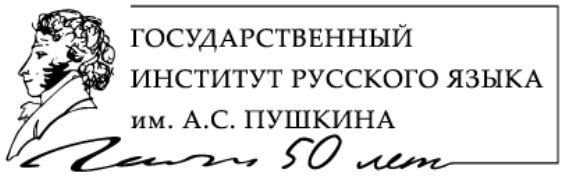


Начало корпусной лингвистики



- 1960 г. – Н. Фрэнсис и Г. Кучера (W.N. Francis and H. Kučera, Brown University, Providence, RI) начинают работу над Брауновским корпусом английского языка – первым лингвистическим электронным корпусом текстов.
- Brown Corpus (Brown university)
  - 1 млн слов
  - 15 жанров
  - 500 сэмплов по 2000 слов каждый
  - ареал: США
  - Время: 1961 г.

История корпусной лингвистики



- Возникновение центров корпусной лингвистики в Лондоне, Ланкастере, Бергене, Гетеборге, Осло, Берлине, Лейпциге, Потсдаме и Вюрцбурге

- LOB Corpus (Lancaster-Bergen-Oslo)
- Создан сходно с Брауновским корпусом на материале британского варианта английского языка
- Тексты 1961 года
- 1 млн. слов, 15 жанров
- Каждый текст содержит максимально 2000 слов
- Kolhapur corpus of Indian English создан в1978 г. на тех же основаниях

The London-Lund Corpus of Spoken English (LLC)

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Первый корпус транскрибированной устной речи
- Часть исследований устной англоязычной речи в Lund University под руководством of J. Svartvik
- 500,000 слов устного британского варианта английского языка, записанного с 1953 по 1987 гг.
- Спонтанные беседы, спонтанные и подготовленные речи

1980-е годы

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- Машинный фонд русского языка
- Уппсальский корпус русского языка (Швеция), 1 млн. слов
- COBUILD
- The Bank of English, Birmingham, 20 млн. слов

COBUILD

- 1млн. слов
- 1980: издательство Collins создает корпус обхемом 20 млн. слов для поддержки лексикографов, работающих над новым «Collins Birmingham University International Learners' Dictionary» (John Sinclair)
- www.collins.co.uk/Corpus/CorpusSearch.aspx
- www.collins.co.uk/books.aspx?group=153

Cobuild

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- Большой исследовательский проект издательства «Collins» и Бирмингемского университета
- 1991 г. , 20 млн. слов
- Вошел в состав Bank of English, в настоящее время около 450 млн слов

Bank of English

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Bank of English – это название корпуса COBUILD, собрание английских текстов.
- Корпус был основан издательством **Collins и University of Birmingham** в 1991 году
- Корпус содержит тексты из тысяч разных источников. В основном это тексты из Великобритании, но также там есть и тексты из Америки и Австралии.
- Письменные тексты взяты из газет, журналов, художественной и не художественной литературы, брошюр, докладов и Интернет сайтов. Устная речь взята из радио передач, встреч, интервью, обсуждений и разговоров.
- В корпусе сейчас содержится 524 миллионов слов, корпус постоянно пополняется.
- В корпусе можно искать примеры сочетаемости слов, проверять частотность слов, увидеть все примеры использования определенного слова и анализировать эти результаты, так что эта информация может быть использована при создании словарей, а так же может служить подкреплением в других работах.
- Корпус использовали при создании словаря Collins COUBUILD Advanced Learners English Dictionary.
- Копии корпуса содержатся как в издательстве HarperCollins так и в University of Birmingham, версия в университете доступна для проведения исследований.
- Bank of English является частью Collins World Web так же как и корпусы французского, немецкого и испанского языков.

Bank of English

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- **Демонстрационная версия корпуса** находится по адресу, можно задавать разные параметры для поиска (это прописано в инструкции): http://www.collins.co.uk/Corpus/CorpusSearch.aspx
- выдает 40 строчек примеров, каждая длинной в 250 символов, строчки располагаются в произвольно последовательности, в примерах сочетаемости слов выдает 100 примеров
- Для того чтобы получить доступ к полной версии корпуса, необходимо написать e-mail (word.banks@harpercollins.co.uk) для получения формы запроса.
- **Стоимость:**
- Существуют разные уровни пользования корпусом:
- One language
- GBP 50 for a one month trial (not renewable)
  GBP 300 for 6 months
  GBP 500 for 12 months
- Each subsequent language:
- GBP 45 per language for a one month trial (not renewable)
  GBP 270 per language for 6 months
  GBP 450 per language for 12 months
- Вам будет предоставлен пользовательский ID для входу в корпус.
- **Могут ли несколько людей использовать корпус по подписке?**
- Да, но кроме trial subscriptions. Предоставляются 3 user IDs изначально, но если вы подписываетесь на корпус на 12 месяцев, вам будут предоставлены 10 id Необходимо также будет назначить человека, который будет контактировать с администрацией корпуса и будет ответственным за то, что данные id используют только конкретные люди в вашей группе.

1990-е гг.

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

- «Британский национальный корпус» (100 млн. слов)
- Национальные корпуса венгерского, итальянского, хорватского, чешского, японского языков, 100 млн. слов.
- The Bank of English, 600 млн. слов

1990-е гг. : British National Corpus

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Одноязычный
- синхронный – вторая половина 20 века
- 4054 текстов,

  100 млн. слов

Письменные тексты – 90%, устные (неформальная коммуникация) – 10%

- 10 % материала – устная речь
- Тексты по 40-50 000 слов каждый
- Кодировка TEI compliant SGML http://www.comp.lancs.ac.uk/ucrel/bncindex/

BNC (1995)

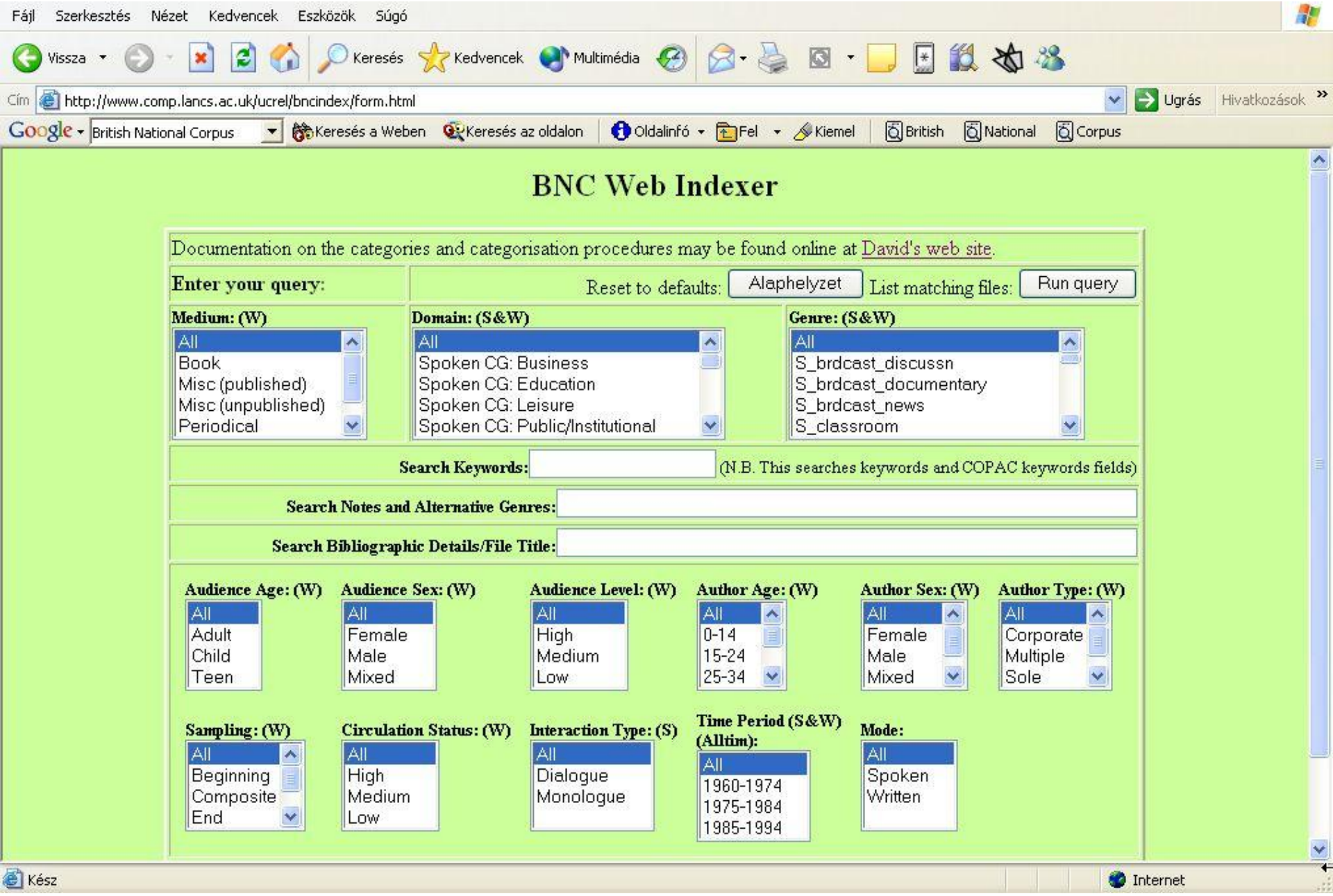
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Наши 50 лет

BRITISH NATIONAL CORPUS

- http://www.natcorp.ox.ac.uk/
- 100 млн. слов, собрание письменных и устных текстов периода 1975-93 гг.
- Тщательно отобраны и сбалансированы
- Корпус закрытый (синхронический)
- Все тексты имеют высококачественное аннотирование
- Множество исследовательских инструментов
- Отличный интерфейс пользователя, см.

http://bnc.humanities.manchester.ac.uk/cgi-bnc/BNCquery.pl?theQuery=search&urlTest=yes


Fájl Szerkesztés Nézet Kedvencek Eszközök Súgó
Vissza Keresés Kedvencek Multimédia
Cím http://www.comp.lancs.ac.uk/ucrel/bncindex/form.html
Ugrás Hivatkozások
Google British National Corpus Keresés a Weben Keresés az oldalon Oldalinfó Fel Kiemel British National Corpus
BNC Web Indexer
Documentation on the categories and categorisation procedures may be found online at David's web site.
Enter your query:
Reset to defaults: Alaphelyzet List matching files: Run query
Medium: (W)
All
Book
Misc (published)
Misc (unpublished)
Periodical
Domain: (S&W)
All
Spoken CG: Business
Spoken CG: Education
Spoken CG: Leisure
Spoken CG: Public/Institutional
Genre: (S&W)
All
S_brdcast_discussn
S_brdcast_documentary
S_brdcast_news
S_classroom
Search Keywords:
(N.B. This searches keywords and COPAC keywords fields)
Search Notes and Alternative Genres:
Search Bibliographic Details/File Title:
Audience Age: (W)
All
Adult
Child
Teen
Audience Sex: (W)
All
Female
Male
Mixed
Audience Level: (W)
All
High
Medium
Low
Author Age: (W)
All
0-14
15-24
25-34
Author Sex: (W)
All
Female
Male
Mixed
Author Type: (W)
All
Corporate
Multiple
Sole
Sampling: (W)
All
Beginning
Composite
End
Circulation Status: (W)
All
High
Medium
Low
Interaction Type: (S)
All
Dialogue
Monologue
Time Period (S&W) (Alltim):
All
1960-1974
1975-1984
1985-1994
Mode:
All
Spoken
Written
Kész
Internet

Создание BNC

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- 1991 – 1994
- 2001: публикация BNC World
- Проект осуществляется BNC Consortium
- Поиск онлайн: http://www.natcorp.ox.ac.uk/
- Поиск онлайн Марка Дэвиса http://view.byu.edu/
- SARA

Новые версии

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- BNC XML (В работе)
- XAIRA (новая система поиска информации)
  - ***Больше возможностей поиска***
  - ***Улучшенный интерфейс пользователя***
  - ***Бесплатное ПО***

Пример использования корпуса 1: swearing

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

- Women and men swear (and use taboo words) differently
- Data (from BNC spoken part) shows
  - Women and men use different swear words
  - They use them for different effect (men use them to disparage, women use them to intensify)
  - Their use changes depending on the sex of the listener(s): women swear more in single-sex groups; men don't swear more in mixed-sex than amongst themselves

Пример 2: Near synonyms

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Наши 50 лет

- Subtle differences in the meaning of near synonyms can be distinguished by looking at the words they collocate with
  - “You shall know a word by the company it keeps” (Firth)

frail vs *fragile*

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Наши 50 лет

No.	Word
1	body
2	pensioner
3	woman
4	physique
5	people
6	hope
7	bodies
8	health
9	hands
10	hand
11	man

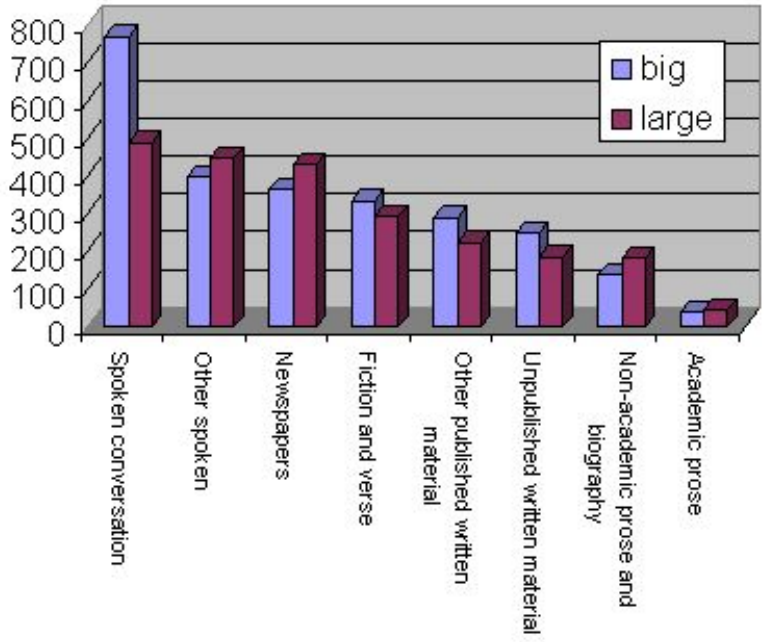
No.	Word
1	unity
2	balance
3	ecosystem
4	ceasefire
5	beauty
6	economy
7	egos
8	alliance
9	x
10	emotions
11	truce
12	ecology
13	state
14	mountain
15	ego
16	environment

Пример 3: Near synonyms

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- In addition, near synonyms can be shown to be favoured depending on genre, eg *big* vs *large*

Category	*big*	*large*
Spoken conversation	768.55	488.34
Other spoken material	395.89	447.58
Newspapers	365.27	431.62
Fiction and verse	333	293.06
Other published written material	290.84	223.43
Unpublished written material	247.39	186.35
Non-academic prose and biography	139.63	181.19
Academic prose	38.85	45.11



Frequency per million

Подкорпусы

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- BNC Sampler
  - ***1 млн. письменных и 1 млн. устных слов***
- BNC Baby
  - ***По 1 млн. слов из каждого из 4 жанров: беллетристики, газет, академических трудов и устной речи***
- доступность: на CD

2000-е гг.

- American National Corpus, 100 млн. слов
- Gigaword corpora (1.000.000.000 слов) английского, арабского, китайского языков

American National Corpus

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- В данном корпусе представлен американский вариант английского языка. Он также включает тексты всех жанров и записи разговоров с 1990 года. Ожидается, что данный корпус будет включать по меньшей мере 100 миллионов слов. Жанры текстов будут включать и «новые» типы (web-блоги, web-страницы, а также тексты из рэп стиля).
- Осенью 2003 года ANC выпустил первое издание (более 11 миллионов слов американского английского).
- Все данные ANC предоставляются Лингвистическим Консорциумом за 75 долларов.
- Вторая версия: (http://americannationalcorpus.org/)

Примеры корпусов английского языка

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

Collins WordbanksOnline Corpus

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Подкорпус корпуса Bank of English
- 56 млн. слов
- Поисковая машина в интернете

  http://www.collins.co.uk/Corpus/CorpusSearch.aspx

Talk Bank Corpus

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Основная цель данного корпуса – помочь исследованиям в области человеческого общения и общения между животными. Talkbank – это междисциплинарный исследовательский проект. Чтобы воспользоваться корпусом, необходимо заполнить форму и предоставить свой адрес, номер телефона, email. Корпус включает Childes - овладение языком детьми (child language acquisition), Lides – билингвизм и кодовая коммутация (bilingualism and code-switching), Gesture – язык телодвижений, AphasiaBank , Linguistic Exploration ,Text and Discourse, включая Conversation Analysis, Classroom Discourse, Animal Communication.
- Большинство файлов содержатся в формате zip. Есть digital video и digital audio. Audio files – в формате mp3.

Bergen Corpus of London Teenage English (COLT)

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- http://www.hf.uib.no/i/Engelsk/COLT/
- Этот корпус посвящен изучению речи подростков. Корпус был разработан в 1993 году на базе разговорного языка детей в возрасте от 13 до 17 лет из разных пригородов Лондона. Он является составной частью British National Corpus. Включает в себя полмиллиона слов. Этот корпус не доступен в Интернете (С 22 апреля 1996 года требуется определенный пароль, чтобы зайти на сам корпус). С декабря 1996 года вышел диск с этой программой. В поисковике можно увидеть распределение слов в зависимости от возраста, пола, социального статуса, места жительства и т.д.
- Информация для корпуса собиралась в Лондоне исследовательской командой из Бергенского Университета. Включает полмиллиона слов из спонтанных разговоров между подростками в возрасте от 13-17 лет (девочки + мальчики). С 1994-1995 гг. эти речи были транскрибированы, обращая внимание на паузы и на одновременные высказывания. Этим занималась Лонгмановская группа. Маркировкой/классификацией слов занималась группа из Lancaster University.

COLT

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Затем вышла другая версия Кольта (CD) с использованием звуковых файлов. Было выбрано ограниченное количество текстов, и они проклассифицированы по трем параметрам: возраст, пол и социальное положение.
- На базе данного корпуса уже написано большое количество работ, например:
- - Прагматические показатели в подростковой речи и речи взрослых (G. Anderson)
- - Новые тенденции в речи подростков.
- More trends in teenage talk. A corpus-based investigation of the discourse items *cos* and *innit*, by G. Andersen and A-B Stenström
- *They like wanna see how we talk and all that*.The use of *like* as a discourse marker in London teenage speech, by G. Andersen
- Girls' conflict talk: a sociolinguistic investigation of the variation in the verbal disputes of adolescent females, by A-B Stenström and I.K. Hasund
- *They gave us these yeah, and they like wanna see how we talk and all that* The use of *like* and other discourse markers in London teenage speech, by G.Andersen

COLT

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА

- **Размер:** 500 000 слов, 100 аудиокассет, 50 часов записей устных бесед
- **Респонденты:** 31 мальчик и девочка в возрасте 13-17 лет,
- **Районы Лондона:** Barnet, Brent, Camden, Enfield, Hackney, Hertfordshire, Islington, Richmond, Tower Hamlets, Westminster
- **Школьные округа:** Hackney, Tower Hamlets, Camden, Barnet, Haileybury

Cambridge International Corpus

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- CIC находится на сайте “Cambridge University Press – English Language Teaching”. Этот корпус создавался в течении последних 10 лет прежде всего как база для составления учебных материалов и словарей английского языка. В одном из разделов корпуса “Corpus-based publications” указаны издания, основанные на корпусе и систематизированные по уровням сложности и тематикам. CIC содержит около 1 млрд. словоупотреблений и постоянно пополняется. Материалы основываются на современной речи, в оновном - устной.

CIC включает в себя следующие ресурсы:

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Cambridge and Nottingham Corpus of Discourse in English (CANCODE), 18 million words
- Уникальное собрание разговорного английского, записанного в сотне разных местностей на Британских островах в различных ситуациях (в ресторане, в магазине). Это только спонтанная речь.
- Cambridge and Nottingham Spoken Business English (CANBEC), 1 million words
- Собрание разговорного бизнес-английского, записанного в коммерческих компаниях разной величины. Формальные и неформальные встречи,презентации, телефонные разговоры, разговоры за обедом. Этот ресурс позволяет оценить, как современные люди используют английский язык в рабочей атмосфере, что способствует более продуктивному изучению и преподаванию бизнес-английского.
- Cambridge Cornell Corpus of Spoken North American English, 0,5 million words
- Записан в различных ситуациях ежедневной жизни.
- Cambridge Corpus of Business English, 60 million words
- Включает в себя деловые документы, отчеты, книге о бизнесе и бизнес-рубрики газет. Представлены материалы на британском и американском вариантах английского языка.

CIC

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Cambridge Corpus of Legal English, 20 million words
- Собрание книг, журнальных и газетных статей, имеющих отношение к закону и юридическим процессам. Включает материалы на британском и американском вариантах английского языка.
- Cambridge Corpus of Financial English, 55 million words
- Собрание книг, журнальных и газетных статей, имеющих отношение к экономике и финансам. Включает материалы на британском и американском вариантах английского языка.
- Cambridge Corpus of Academic English, 25 million words
- Собрание текстов из научных книг и журналов на британском и американском вариантах английского языка.
- Также в корпус входят письменные тексты на британском (650 млн. слов) и американском (250 млн. слов) вариантах английского языка.

Cambridge Learner Corpus

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Этот подкорпус CIC образуют тексты экзаменационных работ студентов из разных стран (180), изучающих английский язык в качестве иностранного (около 85000 студентов и 85000 скриптов). Включает в себя около 25 млн. слов. Этот корпус активно используется при составлении учебников, т.к. дает возможность отслеживать наиболее частое употребление каких-либо конструкций и отслеживать ошибки. Корпус включает в себя специальную программу “Learner Error Coding System”, которая позволяет найти примеры на часто повторяющиеся ошибки.
- На настоящий момент к корпусам имеют доступ только авторы, работающие над книгами для издательства Cambridge University Press.

Corpus of middle English Prose and Verse

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- (http://www.hti.umich.edu/ и http://quod.lib.umich.edu/c/cme/). Это собрание текстов Среднеанглийского языка, составленное из работ, внесенных Университетом Мичиган и текстов, обеспеченных Оксфордским Архивом Текста, а так же из работ, созданных специально для Корпуса HTI (Humanities Text Initiative).
- Последнее обновление Корпуса было сделано в феврале 2006 г. Все тексты представляют собой текст в формате SGML, файлы, маркированные согласно нормам TEI и преобразованные в формат TEI Lite DDT для широкого использования.
- По инициативе факультета Гуманитарных наук, предполагается развить Корпус Прозы и Стиха Среднеанглийского языка в обширное и надежное собрание электронных текстов Среднеанглийского языка, либо преобразовывая тексты самостоятельно, либо договариваясь о доступе с другими собраниям. В настоящее время доступны пятьдесят четыре текста; несколько других будут добавлены позже.

На сайте работают несколько видов поиска:

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Простой поиск - поисковик ищет одно слово или фразу во всем списке
- Точный поиск (поиск близких по значению слов или нечто подобное) - ищет сочетаемость 2-3 слов в 1 фразе
- Логический поиск - ищет комбинации двух-трех слов в данном абзаце/строфе
- Поиск по цитате - определяет работы по автору и заглавию
- Обзор списка
- Все о Списке Прозы и Стихов на среднеанглийском языке

Corpus of middle English Prose and Verse

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Если открыть сам корпус, можно найти в нем 146 текстов, датируемых с середины 19го до начала 20го века. Например:
- Bible. A fourteenth century English Biblical version, ed. Anna C. Paues (Cambridge, 1904)
- Prose life of Alexander, ed. J.S. Westlake, EETS OS 143 (1913 for 1911).
- The right plesaunt and goodly historie of the foure sonnes of Aymon. Englisht from the French by William Caxton, and printed by him about 1489, ed. O. Richardson, EETS ES 45 (1884).
- Legends of the holy rood; symbols of the passion and cross poems, ed. R.Morris, EETS OS 46 (1871).
- The early South-English legendary ; or lives of saints. I. Ms. Laud, 108, in the Bodleian library, ed. C. Horstmann, EETS OS 87 (1887).
- Paston family
- Paston letters and papers of the fifteenth century. Part I only, ed. N. Davis (Oxford, 1971)

Corpus of middle English Prose and Verse

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Таким образом, тексты разделены на анонимные и на те, авторы которых нам известны.
- Если открыть саму ссылку на текст, на странице в виде таблицы представлены:
- Автор
- Название
- информация о публикации
- доступность (ссылки в интернете)
- источник
- URL
- Далее идет содержание, каждую часть которого так же можно открыть (в ней будет уже сам текст).

MICASE

- http://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase
- *Каковы характеристики современной ученой речи, ее вокабуляр, функции, цели, влияние?*
- *Различаются ли они в зависимости от дисциплин и целевой аудитории?*
- В конце 90-х годов (1997) Институт английского языка (ИАЯ) на базе Мичиганского университета начал работу над проектом, который должен был найти ответы на эти и многие другие вопросы. Целью первого этапа эксперимента было записать и затранскрибировать приблизительно 200 часов лекций (около 1,8 млн слов) по всему университету. В июне 2001г. была закончена запись лекций, семинаров, заседаний ученого совета и т.д. Всего получилось 190 часов записей. В апреле 2002г. было закончено транскрибирование и проверка всех транскрипций. (Цифровые записи транскрибировались с помощью программы, которая называется SoundScriber, разработанной одним из бывших помощников – Эриком Бреком/Eric Breck).
- Для создания данного корпуса текста было несколько причин. Во-первых, ранее не существовало ни одной подобной базы данных. Во-вторых, создатели надеялись отследить изменения, происходящие в языке, т.к. люди получают подобного рода опыт в университетах. В-третьих, разработчики сайта подозревали, что «живая речь», даже образованных людей, сильно отличается от той, что написана в грамматиках. Они уверены, что существенные отличия неизбежны, т.к. та речь, которая представлена в грамматиках больше имеет дело с речью письменной, а не устной. В-четвертых, с новой полученной информацией создатели надеются исследовать и усовершенствовать как ИАЯ, так и English for Academic Purpose (английский в научных целях). Также они планируют разработать улучшенные методы обучения.

MICASE

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Первоначально MICASE планировался как открытый, не ограниченный в доступе сайт. Данный проект заинтересовал многих ученых в Европе, Азии и Северной Америке. MICASE также пригласили на ICAME/AAACL конференцию в мае 2005г.
- Первоначально руководителями проекта были доктор Рита Симпсон вместе с профессором Джоном Свайлзом и доктором Сарой Бригс. Сейчас проектом руководит доктор Уте Рёмер. В общей сложности больше 30 исследователей работали над проектом в течение этих лет.

Структура сайта:

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Вы попадаете на страничку, на которой видите строку поиска, в которую вписываете нужное слово. В соседних столбцах выбираете «особенности говорящего» и «Особенности транскрипции». Первый раздел состоит из таких подпунктов как:
- Пол говорящего (любой, женский, мужской, неизвестный);
- Возраст говорящего (любой, неизвестный, 17-23, 24-30, 21-49, 50-и выше);
- Образование (любое, студент предпоследнего курса, исследователь, научный работник, посетитель и т.д.);
- Статус носителя языка (не носитель, почти носитель, носитель Американского английского, носитель другого английского, неизвестно);
- Родной язык (любой, арабский, армянский, …).

MICASE

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Второй раздел включает:
- Тип мероприятия, на котором получена запись (любой, коллоквиум, защита диссертации, лекция и т.д.);
- Распределение по областям (любое, биология, искусство и т.д.);
- Дисциплину (любая, американская культура, антропология, архитектура и т.д.);
- Уровень участников (любой, студент предпоследнего курса, исследователь, научный работник, посетитель и т.д.);
- Интерактивность (любая, преимущественно монологическая, смешанная и т.д.).
- В зависимости от того, что Вам необходимо найти, Вы заполняете строку поиска, отмечаете особенности и нажимаете на поиск (“Submit Search”). В зависимости от Вашего запроса Вы получаете список требующегося слова в различных контекстах. При чем Вы можете просмотреть контекст как в размерах одного предложения, так и в размерах целой лекции. Также Вы можете просмотреть статистику, которая покажет, насколько часто это слово встречается на 10000 единиц, соотнесет частоту потребления этого слова мужчинами и женщинами и т.д.

THE LAMPETER CORPUS OF EARLY MODERN ENGLISH TRACTS

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- The Lampeter Corpus of Early Modern English Tracts представляет собой собрание текстов, опубликованных **в период между 1640 и 1740 годами** – время, отмеченное в Англии ростом массового издания, развитием публичного дискурса во многих областях повседневной жизни, и, наконец, **стандартизацией британского английского**. Корпус отражает век, который был ключевым в процессе становления британского английского, каким мы знаем его сейчас.
- The Lampeter project был предпринят ***в 1991 г***. профессором *Dr. Josef Schmied* и *Eva Hertel* в *Bayreuth University*, а ***в 1993 г.*** переехал в *Chemnitz*. В начале проект спонсировался *the Deutsche Forschungsgemeinschaft (DFG)*, а с 1994 г. спонсируется *the German Research Association.*

THE LAMPETER CORPUS OF EARLY MODERN ENGLISH TRACTS

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Для того, чтобы ответить на нужды лингвистов и историков, the Lampeter project представляет собой сбалансированный корпус, который характеризуется определенным набором критериев:
- **только полные тексты**, включая предисловие, послесловие и т.д.;
- тексты разной длины (от 3000 до 20000 слов). **Каждый автор появляется только раз**;
- **только первое издание текста**, поздние издания появляются лишь в том случае, если исправления сделаны самим автором. Никаких современных изданий;
- разделение века на десятилетия;

THE LAMPETER CORPUS OF EARLY MODERN ENGLISH TRACTS

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- каждое десятилетие включает тексты на темы:
- религия;
- политика;
- экономика;
- наука;
- право;
- разное;
- по два текста на каждую тему и на каждое десятилетие, что вследствие ведет к 120 текстам и к 1,1 млн. слов.
- Корпус интересный, так как отражает действительно важный век в истории Англии, в стандартизации британского английского, что важно для специалистов, изучающих английский язык, его историю.

THE LAMPETER CORPUS OF EARLY MODERN ENGLISH TRACTS

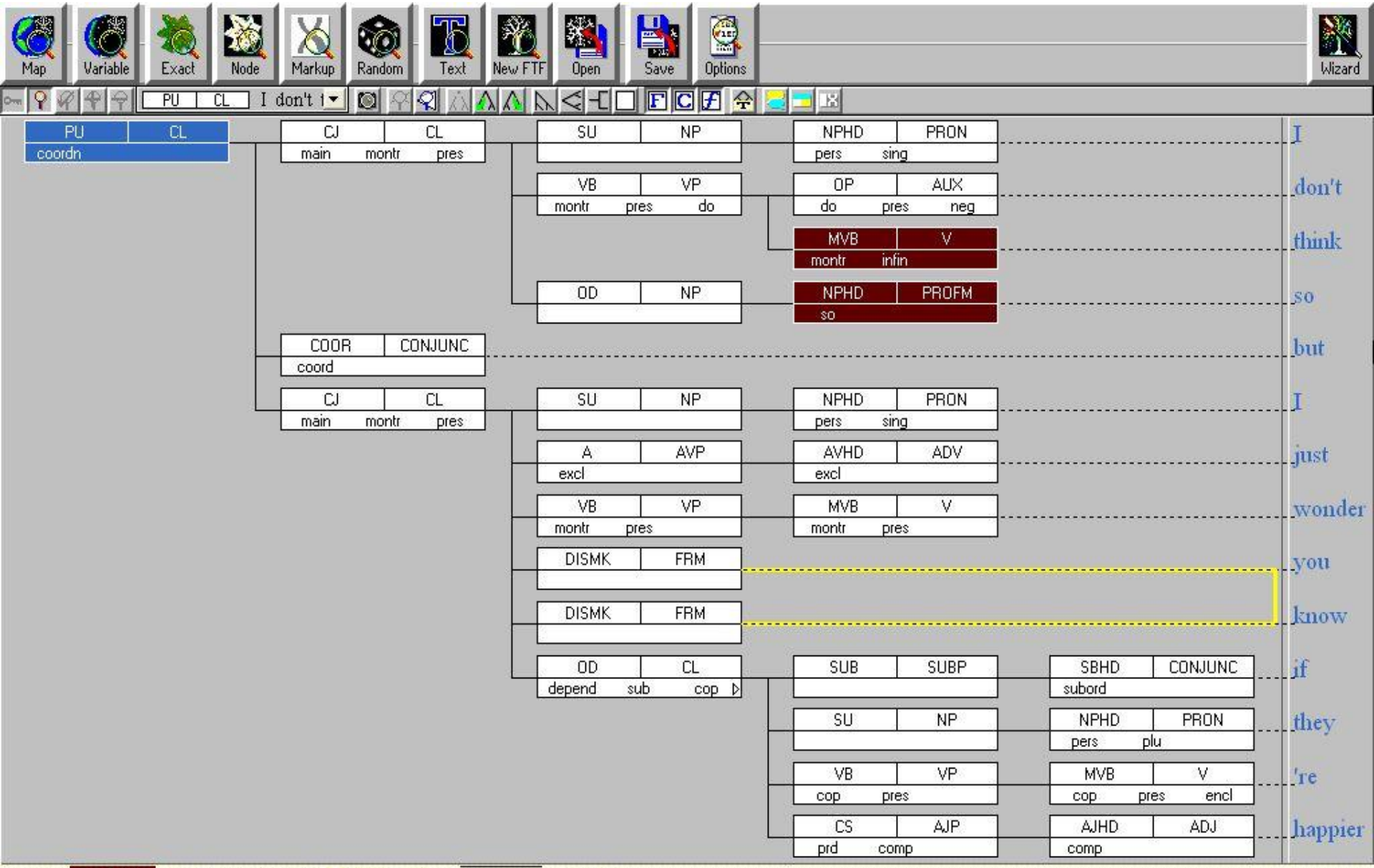
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- К сожалению, корпус еще не готов до конца. Обещают, что впоследствии корпусом можно будет пользоваться с разных сайтов:
- - Oxford Text Archive (http://sable.ox.ac.uk/ota/);
- - International Computer Archive of Modern and Medieval English (ICAME) (http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM);
- - и с самого сайта корпуса - http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM

International Corpus of English

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- 20 корпусов по 1 млн слов, посвященных вариантам английского языка во всем мире
- 500 текстов (300 письменных, 200 устных) по 2000 слов каждый
- охватывает: 1990-1996 гг.
- ICE-GB доступен в демоверсии
- Синтаксическая аннотация, графический инструмент ICECUP


Map
Variable
Exact
Node
Markup
Random
Text
New FTF
Open
Save
Options
Wizard
PU CL
coordn
CJ CL
main montr pres
SU NP
NPHD PRON
pers sing
I
VB VP
montr pres do
OP AUX
do pres neg
don't
MVB V
montr infin
think
OD NP
NPHD PROFM
so
so
COOR CONJUNC
coord
but
CJ CL
main montr pres
SU NP
NPHD PRON
pers sing
I
A AVP
excl
AVHD ADV
excl
just
VB VP
montr pres
MVB V
montr pres
wonder
DISMK FRM
you
DISMK FRM
know
OD CL
depend sub cop
SUB SUBP
SBHD CONJUNC
subord
if
SU NP
NPHD PRON
pers plu
they
VB VP
cop pres
MVB V
cop pres encl
're
CS AJP
prd comp
AJHD ADJ
comp
happier
I don't think so but I just wonder you know if they 're happier

Корпуса немецкого языка

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- **1. http://www.ids-mannheim.de/kl/projekte/korpora/**
- Die Korpora geschriebener Gegenwartssprache des IDS
- bilden mit über 2.2 Milliarden Wörtern die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit.
- enthalten belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten und werden kontinuierlich weiterentwickelt.
- werden im Hinblick auf Umfang, Variabilität, Qualität und Aktualität akquiriert und erlauben in der Nutzungsphase über *COSMAS* die Komposition virtueller Korpora, die repräsentativ oder auf spezielle Aufgabenstellungen zugeschnitten sind.
- enthalten ausschließlich urheberrechtlich abgesichertes Material.

Корпуса ИДС Маннгейм

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- **Bonner Zeitungskorpus (bzk)**
- Erarbeitung: IDS-Mitarbeiter in Bonn
- Umfang: 10 840 Texte; ca. 3,1 Mill. laufende Wortformen
- Zeitraum: Jahrgangsquerschnitte 1949, 1954, 1959, 1964, 1969 und 1974
- Inhalt: Artikel aus den Tageszeitungen:
- - *Neues Deutschland* (DDR)
  - *Die Welt* (Bundesrepublik Deutschland)

Корпуса ИДС Маннгейм

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- **Mannheimer Korpus 1 (mk1)**
- Erarbeitung: IDS
- Umfang: 293 Texte; ca. 2,2 Mill. laufende Wortformen
- Zeitraum: 1950 - 1967
- Inhalt: - Belletristik
- Heinrich Böll: *Ansichten eines Clowns*
- Werner Bergengruen: *Das Tempelchen*
- Max Frisch: *Homo faber*
- Günter Grass: *Die Blechtrommel*
- Uwe Johnson: *Das dritte Buch über Achim*
- Thomas Mann: *Die Betrogene*
- Erwin Strittmatter: *Ole Bienkopp*
- - Memoiren
- Theodor Heuss: *Erinnerungen 1905-1933*
- - wissenschaftliche und populärwissenschaftliche Literatur
- - Trivialliteratur
- - Artikel aus Zeitungen und Zeitschriften

Корпуса ИДС Маннгейм

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- **Mannheimer Korpus 2 (mk2)**
- Erarbeitung: IDS
- Umfang: 52 Texte; ca. 0,3 Mill. laufende Wortformen
- Zeitraum: 1949, 1952, 1960 - 1974
- Inhalt:
- - Erlasse, Satzungen, Beschlüsse
  - Gebrauchsanweisungen, Lehrbücher
  - Nachrichten, Prospekte, Trivialliteratur
  - wissenschaftliche und populärwissenschaftliche Literatur
  - Artikel aus Zeitungen und Zeitschriften

Корпус LIMAS

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Содержание корпуса:
- Корпус содержит 500 источников и более миллиона слов. Здесь собраны как полные сочинения, так и отрывки из произведений различных жанров, изданных в 1970 году. Все сочинения разбиты рубрики, которые в свою очередь имеют подрубрики.
- Здесь Вы найдёте следующие рубрики: беллетристика, общество (например, такие подрубрики как работа, гороскоп, секс и др.), право, медицина (для специалистов и популярная), география, метеорология, религия, техника, страна, спорт и так далее.
- Весь перечень произведений, содержащихся в корпусе, можно посмотреть в разделе на сайте «Inhaltsverzeichnis».
- Каждое произведение имеет порядковый номер, она приписано к какой-либо рубрике и подрубрике. Здесь даётся полное название, год издания, издательство и номера страниц, которые использует корпус.
- Все используемые в корпусе произведения доступны для чтения.

Корпус LIMAS

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- В данном корпусе доступны три вида поиска: простой, по контексту и поиск фраз.
- Простой поиск позволяет найти одно или несколько заданных слов.
- Поиск по контексту позволит Вам искать слово в корпусе в зависимости от того, какие слова и формы окружают его.
- Поиск фраз позволит искать в корпусе целые фразы и выражения.
- Более точной информации о времени и условиях создания данного корпуса на сайте не дано, что является, по моему мнению, минусом. Также в настоящий момент, по всей видимости, над корпусом не работают, не развивают, не улучшают.
- В работе корпус удобен, прост. Его можно посоветовать тем, кто изучает немецкий язык 70-х годов 20 века.

Цифровой словарь/digitales Wörterbuch (das digitale Wörterbuch der deutschen Sprache des 20. Jh.)

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Руководство: Manfred Bierwisch, Wolfgang Klein (осн. руководитель проекта), Hartmut Schmidt, Angelika Storrer.
- Цель проекта – разработка цифровой словарной системы на основе электронных корпусов. Язык – немецкий.
- Объект исследования – лексический состав немецкого языка 20-го века
- Проект находится в разработке с марта 2000 года. Предполагалось, что словарь будет состоять из нескольких корпусов: Основного корпуса (Kernkorpus) – 80.000 документов и дополнительного корпуса (Ergänzungskorpus) – 2,5 млн. документов.
- В настоящий момент проект включает в себя несколько Интернет-корпусов. Корпуса находятся в открытом доступе, однако регистрация пользователя – обязательна (процесс регистрации занимает не более 1 минуты, но ключ доступа присылается не сразу, мне пришлось ждать ≈ 3 часа).

Структура этого корпуса

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- **Общие корпуса:**
- **DWDS-KernCorpus**
- Объем – 100 млн. слов в 79.830 документах. В корпусе используются различные типы текстов письменной речи:
  - Художественная литература ≈ 26 %,
  - Публицистика ≈ 27 %,
  - Научная литература ≈ 22 %,
  - Справочная литература ≈ 20 %,
- а также ≈ 5 % затранскрибированных текстов устной речи.
- **“Juilland-D” – Corpus**
- Объем – 500.000 слов в 392 документах. Используются тексты 1920 – 1939-х гг.
- Типы текстов:
- Драмы (6 произведений) 20%
- Новеллы и рассказы (39 произведений) 20%
- Эссе (23 произведения) 20%
- Публицистика (286 статей) 20 %
- Научная литература (38 произведений) 20 %

Структура этого корпуса

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- **Специальные корпуса:**
- **Корпус разговорной речи (Corpus Gesprochene Sprache)**
- Объем – 2,5 млн. слов. Состоит из нескольких подкорпусов:
  - Речь (200.000 слов) – собрание речей Кайзера Вильгельма, Гитлера, Ульбрихта, Хонекера и др.
  - Радиопередачи 1929-1944 гг. (400.000 слов)
  - Отрывки австрийских парламентских протоколов 1948-1956 гг. и др.
- **ГДР – корпус (DDR-Corpus)**
- Объем – 9 млн. слов в 1150 текстах. Корпус охватывает тексты 1949-1990 гг., опубликованные в ГДР. Корпус создавался в сотрудничестве с Гумбольдтовским университетом.
- **Корпус еврейской периодики (Corpus Jüdischer Periodika)**
- Объем – 50000 страниц – 26.247.390 слов. Текстовая база – 8 полных журналов 1887 – 1938 гг. Корпус был создан при совместной работе с проектом Compactmemory.
- Также предлагаются 4 корпуса газетной периодики: **Zeit-Corpus** (охватывает издания 1996-2007 гг. и 22 издания 1946-1988 гг.), **Corpus Berliner Zeitung** (все online-статьи 1994-2005 гг), **Tagesspiegel-Corpus** (статьи 1996-2005 гг.), **Corpus der Potsdamer Neuesten Nachrichten** (статьи 2003-2005 гг.)
- **DWDS-Ergänzungscoprus** охватывает 1 млрд. актуальных слов (1990-2000гг.), в качестве текстовой базы используется в основном современная периодика. Однако доступ через Интернет к данному ресурсу, к сожалению закрыт.

Как пользоваться

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Кроме всего прочего о слове дается основная информация (грамматическая, лексическая и др.) из словаря современного немецкого языка, автоматически рассчитанные семантические связи (синонимы, гипонимы и гиперонимы) и автоматически рассчитанные коллокации слова (наиболее частые словосочетания).
- Существует возможность задавать при поиске дополнительные ограничения по дате, заглавию, автору и типу текста (Aktenreiter **Filter**). Все текстовые примеры сортируются по дате или длине предложения (Aktenreiter **Darstellung**), можно также просмотреть частотность использования слова в различных типах текстов в течение 20 века (Aktenreiter **Wortverlauf**) и скачать любой документ, содержащий искомое слово, в виде текстового файла (Aktienreiter **Export**).
- Корпус насчитывает более 10.000 зарегистрированных пользователей (на февраль 2007г.), ежедневно пополняется новыми текстами, находится в свободном доступе, очень прост и удобен в использовании.

Французские корпуса

PERTOMed

- Русско-французский биомедицинский параллельный корпус
- База: корпус французского языка
- объем: 14 000 слов
- цель: автоматизация перевода
- руководители: Marie - Christine Jaulent, Jean Charlet

Французские корпуса

- ARTFL Project
- GlossaNet
- EUR – ACCOR
- OPUS

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Пушкин 50 лет

ARTFL - Project

- Прорект америко-французских исследований сокровищницы французского языка - Project for American and French Research on the Treasury of the French Language (ARTFL)
- Участники
  - Analyse et Traitement Informatique de la Langue Française (ATILF)
  - the Centre National de la Recherche Scientifique (CNRS)
  - the Division of the Humanities
  - the Division of the Social Sciences
  - Electronic Text Services (ETS) of the University of Chicago

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

ARTFL - Project

- Тестовая база: FRANTEXT (ранее: Trésor de la langue francaise)
  - 114.7 млн. слов
  - исторический период: от средневековья до 20 века
  - Типы текстов: от классиков французской литературы до нелитературной прозы – новеллы, тексты стихотворений, драмы, журналистика, эссе, переписка, трактаты

ARTFL Project

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

Наши 50 лет

THE UNIVERSITY OF CHICAGO

atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

The ARTFL Project

- Многоязычные тексты Библии
- База данных PhiloLogic:
  - Множество опций: например, списки частотных слов, контекстный поиск...

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

GlossaNet

- Разработан лингвистами для лингвистов
- GlossaNet использует ПО UNITEX, чтобы обращаться к электронным словарям и составлять конкордансы (наборы контекстов для заданного слова)
- Можно делать запросы: морфологический, синтаксический, семантический
- Источник: более 100 актуальных онлайн-изданий на 12 языках, электронные словари в системе RELEX

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

EUR – ACCOR

- Заказчик: ЕС
- Испольнитель: University of Edinburgh – Center for Speech Technology Research
- 1990 – 1993
- Корпуса на: немецком, английском, шведском, французском, каталанском, итальянском, гаэльском, американско-английском, русском языках
- Нужно купить лицензию для пользования

OPUS

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

- Разработка: университет Осло
- Собрание бесплатных параллельных корпусов
- кодировка: XML и Unicode UTF8
- Автоматическое аннотирование
- Желательны взносы за пользование

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

OPUS

UNIVERSITY OF OSLO

Корпуса

- EuConst – Евроконституция на 21 языках
- OpenOffice – справочник на 6 языках
- Europarl – протоколы Европарламента на 11 языках
- KDE – справочник на 61 языке
- PHP – справочник на 21 языке

Французско - славянские онлайн словари

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- Lexicool
- Мультитран
- http://multilex.mail.ru
- http://translation2.paralink.com/
- http://www.online-translator.com/text.asp#tr_form

«База средневекового французского" (BFM) и ее интеграция в корпус "Тезауруса французского языка

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

- "База средневекового французского" (BFM) представляет собой корпус текстов французских литературных и деловых произведений IX - XVI вв., основанный на их современных критических изданиях. В настоящее время BFM включает около 50 текстов общим объемом приблизительно 2700000 слов. К сожалению, доступ исследователей к данным BFM пока ограничен по соображениям защиты авторских прав на критические издания.
- Работа над созданием Базы началась в 1989 г. под руководством профессора Высшей нормальной школы Франции К. Маркелло-Низья. Состав коллектива исследователей и название лаборатории, в которой осуществлялся проект, с тех пор неоднократно изменялся. В настоящее время над проектом работает небольшая группа сотрудников лаборатории ICAR, входящей в состав Национального центра научных исследований Франции (CNRS). Электронные версии основной массы текстов были получены путем сканирования современных изданий с последующим многократным вычитыванием. При этом использовался формат. На основе текстовых документов с помощью программы были созданы конкордансы. В нескольких текстах с использованием программы была произведена морфологическая разметка (идентификация значений морфологических категорий).

«База средневекового французского" (BFM) и ее интеграция в корпус "Тезауруса французского языка"

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

- На материале BFM были получены интересные научные результаты, касающиеся грамматикализации модальных слов (*voir, espoir*), квантификаторов (*très, beaucoup*), вспомогательных и модальных глаголов, а также эволюции дейктической системы, выражения отрицания и порядка слов (работы К. Маркелло-Низья, Б. Комбетта и С. Прево и др.).
- В то же время ряд лингвистических исследований требует привлечения материала более широкого <временного среза>, чем тот, который представлен в рамках BFM. В этой связи в 2002 г. К. Маркелло-Низья предложила проект интеграции BFM в корпус FRANTEXT, создававшийся на протяжении нескольких десятилетий работы над "Тезаурусом французского языка" (TLF). Этот корпус включает более 3500 французских текстов (преимущественно литературных) XVI - XX вв.
- Данный проект, в реализации которого мы принимаем участие совместно с С. Гийо и С. Эйденом, потребовал определенной модернизации BFM. В частности, было решено представить тексты в формате XML в соответствии с рекомендациями TEI, что отвечает современным тенденциям развития корпусной лингвистики. При этом необходимо было решить ряд методологических и технических проблем, на которых следует остановиться подробнее.

- Базовым принципом корпуса BFM является строгое соответствие критическому изданию. При этом вопрос о том, насколько достоверным источником лингвистических данных являются критические издания, остается открытым. Безусловно, ряд исследований (например, анализ употребления знаков препинания) на таком материале проводить в принципе невозможно. Вызывает сомнения пригодность критических изданий (по крайней мере, части из них) для изучения эволюции морфологических явлений (например, редукции падежной системы). В то же время в области лексикологии и синтаксиса критические издания могут, по всей видимости, служить достаточно достоверным источником данных. В любом случае источником материала могут служить либо точные дипломатические транскрипции рукописей, либо опубликованные критические издания, но не нечто среднее.

- Формат XML позволяет воспроизвести не только сам текст произведения, но и целый ряд элементов критического аппарата (нумерацию строк, использование различных шрифтов, варианты текста, примечания и т.п.). На данном этапе перевода корпуса BFM в формат XML было решено ограничиться включением лишь той части критического аппарата, которая с помощью каких-либо типографских средств интегрирована в текст произведения. Конкретно речь идет о:
- 1) выделении фрагментов текста особым шрифтом (курсивом или малыми прописными);
- 2) использовании квадратных скобок, многоточий на месте лакун манускрипта;
- 3) нумерации строк, строф или параграфов.
- Основополагающим принципом разметки текстов в системе TEI является кодирование не типографских средств, а содержательных элементов, для выделения которых служат эти типографские средства. Так, курсивом в тексте произведения могут отмечаться слова на иностранном языке. Таким образом, при разметке текстов корпуса BFM было необходимо провести содержательный анализ употребления типографских средств в издании, на основе которого создавалась электронная версия.

- Опыт работы показал, что использование типографских средств в разных изданиях заметно различается. Более того, оно не всегда последовательно даже в рамках одного издания. В некоторых изданиях обнаружились ошибки при нумерации строк. Все это делает содержательную разметку более трудоемкой, но в то же время повышает ее ценность.
- Переформатирование корпуса BFM было также использовано для дополнительной вычитки электронных текстов (проверки их соответствия печатным изданиям). Эта работы была поручена группе специалистов по старофранцузскому языку. Они же должны были отметить в тексте использование специальных типографских средств и по возможности определить его функцию. При составлении инструкции для корректоров была сделана попытка максимально упростить техническую сторону их работы. С этой целью мы постарались свести к минимуму число тагов, которые корректорам следовало расставить в тексте. Все виды разметки, которые можно провести автоматически, осуществлялись уже после получения вычитанных текстов в лаборатории. В процессе работы корректорам предлагалось заполнить таблицу соответствия типографских средств и их функций в издании. В самом тексте требовалось отметить соответствующими тагами иностранные слова (<foreign>), выделенные особым шрифтом имена собственные (<name>) и исправления составителя критического издания (<corr>).

- Исправления издателя, выделенные квадратными скобками, и лакуны, отмеченные многоточиями (или многоточиями в квадратных скобках), корректоры должны были оставить без изменений. В процессе окончательной доводки текста с помощью регулярных выражений эти фрагменты оформлялись как элементы <corr> и <gap> с указанием ответственного лица (атрибут *resp*) и типографского средства (атрибут *rend*).
- В ходе работы корректорами был отмечен ряд опечаток и ошибок в самих критических изданиях. Тем не менее, поскольку, как уже отмечалось, важнейшим принципом корпуса BFM является точное воспроизведение критического издания, даже в случае явной ошибки корректорам предлагалось оставить текст издания без изменений, а свои исправления и комментарии ввести с помощью элемента <note> или атрибута *corr* в элементе <sic>.

- Некоторые сложности возникли при выстраивании иерархической структуры текстов. Как известно, в рекомендациях TEI проводится фундаментальное различие между прозаическими и стихотворными текстами. В прозе базовым элементом структуры текста является абзац (<p>), строки внутри которого могут факультативно помечаться <пустым> элементом типа (<lb/>). Согласно предлагаемому TEI DTD элемент <p> не имеет атрибутов, кроме глобальных *n* и *rend*.
- В стихотворных произведениях базовым элементом является стих (строка), которому соответствует элемент <l>. Прозаическим абзацам в стихах соответствуют <группы строк> (элемент <lg>). Последний элемент может иметь атрибут *type*, позволяющий уточнить, с какого рода группой стихов мы имеем дело (строфа, куплет, ле и т.п.).
- Данная система вполне логична с точки зрения теории литературы, однако ее практическое применение на материале старофранцузских текстов не всегда удобно. Дело в том, что основная масса старофранцузских произведений, в том числе эпических и даже научных была написана в стихах. Такие тексты делятся на главы и <параграфы>, начало которых графически обозначается с помощью больших разноцветных буквиц. Эти параграфы в большей мере соответствуют прозаическим абзацам, чем стихотворным строфам. Кроме того, буквальное следование в данной ситуации рекомендациям TEI заметно осложнило бы работу наших корректоров. В то же время модификация стандартного DTD, разработанного TEI также представляется нежелательной.

- В этих условиях было принято компромиссное решение: использовать для всех видов текстов <p> в качестве базового элемента; в стихотворных текстах для нумерации строк пользоваться элементом <lb/> с атрибутом *n*, а элемент <p> снабжать атрибутом *rend* со значениями > 'строфа', 'куплет' и т.д. В том случае, если данный <стихотворный абзац> не имеет специального названия, используется значение 'группа стихов'.
- В марте 2003 г. первые 15 текстов BFM были успешно интегрированы в состав корпуса FRANTEXT, окончательное же завершение проекта запланировано на конец этого года. При этом BFM продолжит свое существование и развитие в качестве самостоятельного корпуса. Планируется, в частности, продолжение морфосинтаксической разметки текстов и эксплуатация базы с использованием онлайнового анализатора Weblex.

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Ваши 50 лет

Национальный корпус русского языка

www.ruscorpora.ru

Корпус

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

— это **информационно-справочная система**, основанная на собрании текстов на некотором языке в электронной форме.

Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всём многообразии

- жанров,
- стилей,
- территориальных
- и социальных вариантов и т. п.

Национальный корпус создается лингвистами (специалистами по так называемой *корпусной лингвистике*, быстро развивающейся современной области языкознания) для **научных исследований и обучения языку**.

Две важные особенности

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- Во-первых,
  он характеризуется представительностью, или сбалансированным составом текстов.

  Это означает, что корпус содержит по возможности **все типы письменных и устных текстов**, представленные в данном языке,

  и что все эти тексты входят в корпус по возможности **пропорционально их доле в языке соответствующего периода**.

Две важные особенности

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Пушкин 50 лет

Планируемый составителями объем Национального корпуса русского языка — **200 млн. слов**.

Две важные особенности

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

- Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию).

Разметка — главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов, в изобилии представленных в современном интернете, в том числе и на русском языке языке

(таких, как, по-видимому, наиболее известная «библиотека Максима Мошкова» или, например, «Русская виртуальная библиотека»).

Две важные особенности

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

В настоящее время специалистами создана и пополняется также

«Фундаментальная электронная библиотека» русской классической литературы,

Четыре типа разметки

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

В Национальном корпусе русского языка в настоящее время используется четыре типа разметки:

- метатекстовая,
- морфологическая,
- акцентная,
- семантическая;

в ближайшее время планируется внедрение синтаксической разметки.

Система разметки постоянно совершенствуется.

Зачем нужен национальный корпус?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

Национальный корпус предназначен в первую очередь для обеспечения **научных исследований лексики и грамматики языка**,

а также тонких, но **непрерывных процессов языковых изменений**, происходящих в языке на протяжении сравнительно небольших периодов — от одного до двух столетий.

Зачем нужен национальный корпус?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет

Другая задача корпуса

— **предоставление всевозможных справок**, относящихся к указанным областям (лексика, грамматика, акцентология, история языка)

Как развивается Национальный корпус?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

Национальный корпус русского языка охватывает прежде всего период

от середины XVIII до начала XXI века: этот период представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах — литературном, разговорном, просторечном, отчасти диалектном.

Что включает в себя Национальный корпус русского языка?

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Пушкин 50 лет

В корпус включаются оригинальные (непереводные) произведения художественной литературы (проза и драматургия, в дальнейшем также поэзия), имеющие культурную значимость, а также представляющие интерес с точки зрения языка.

Но Национальный корпус ни в коей мере не является только корпусом языка художественной литературы.

Что включает в себя Национальный корпус русского языка?

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

Помимо художественных текстов, в корпус в большом количестве включаются и другие образцы письменного (а для современного этапа — и устного) языка:

- **мемуары,**
- **эссеистика,**
- **публицистика,**
- **научно-популярная и научная литература,**
- **публичные выступления,**
- **частная переписка,**
- **дневники,**
- **документы** и т. п.

Подкорпуса

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА

- параллельный русско-английский корпус текстов, в котором можно найти все переводы для определенного русского или английского слова или словосочетания;
- корпус диалектных текстов, включающий запись диалектной речи различных регионов России с сохранением их грамматической специфики; предусмотрен специальный поиск с учётом диалектной морфологии;
- корпус поэтических текстов, в котором возможен поиск не только по лексическим и грамматическим, но и по специфическим для стиха признакам (поиск определённого сочетания в сонетах, в эпиграммах, в стихотворениях, написанных амфибрахием, с определённым типом рифмовки и т. п.)

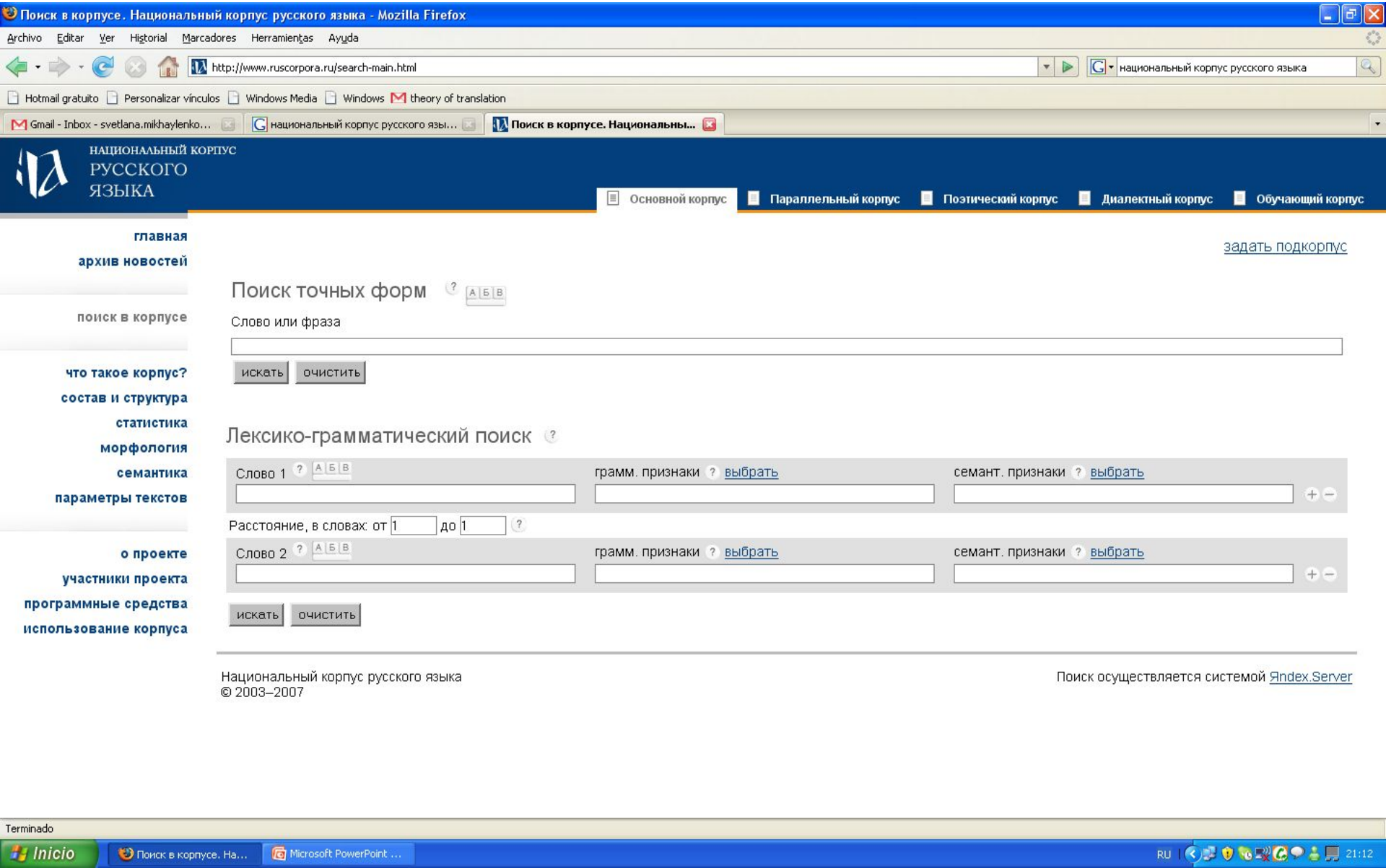
Современные письменные тексты

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- современная художественная проза разных жанров и направлений
- современная драматургия
- мемуарно-биографическая литература
- журнальная публицистика и литературная критика
- газетная публицистика и новости
- научные, научно-популярные и учебные тексты
- религиозные и религиозно-философские тексты
- производственно-технические тексты
- официально-деловые и юридические тексты
- бытовые тексты (в том числе тексты, не предназначенные для публикации: личная переписка, дневники и т.п.)

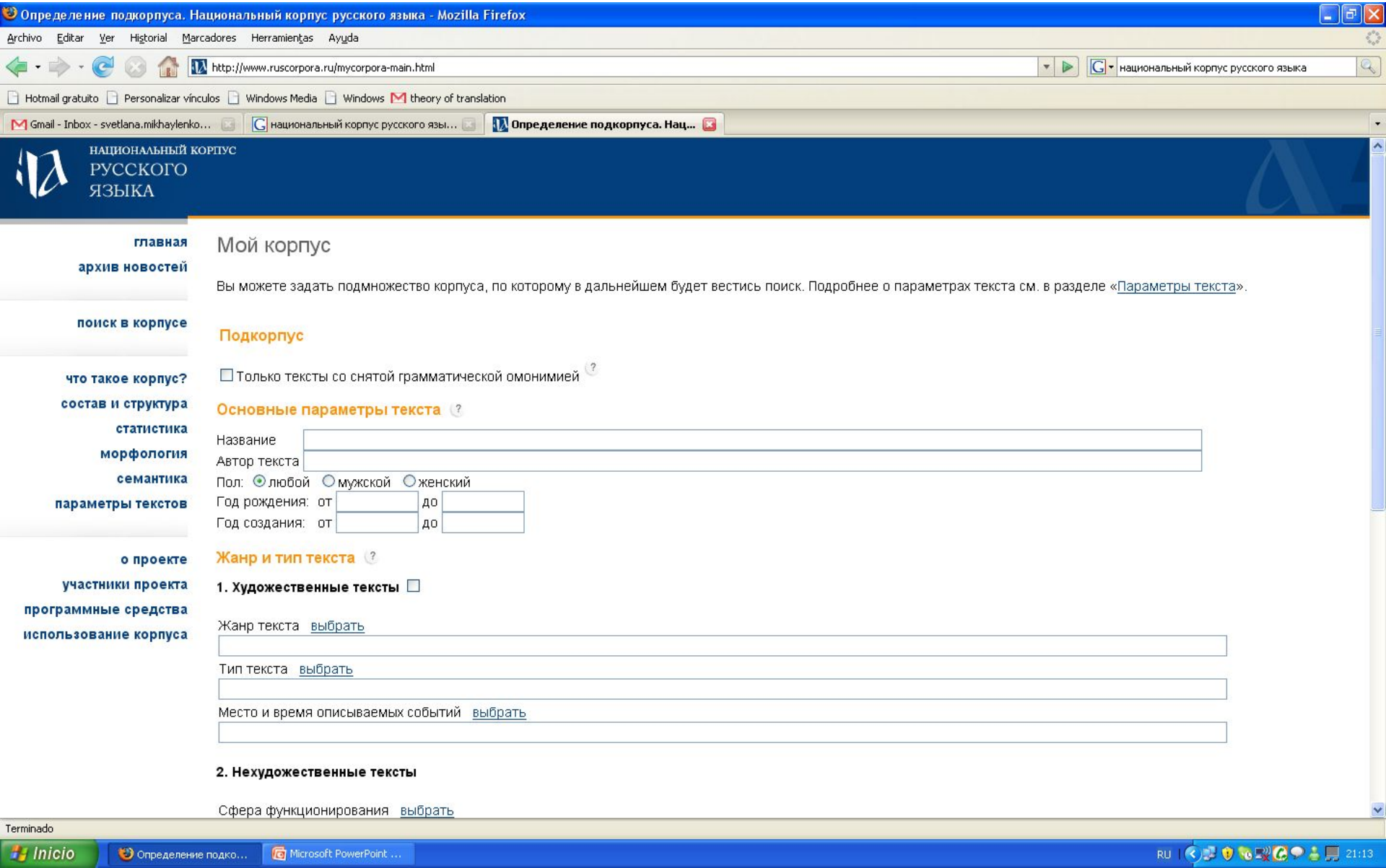
Основной корпус текстов

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет



Страница установки пользовательского

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
Наши 50 лет



Интернет как корпус?

- За:
- Огромная база данных
- Динамично расширяется
- Идеальный «быстрый и грязный» метод поиска

- Против:
- Много «спама», ненужной информации
- Сложно выделить самое надежное и нужное
- Отсутствует языковой анализ
- Поиск только линейный

ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА
Наши 50 лет

- СПАСИБО ЗА ВНИМАНИЕ!