

02. Текстовые корпуса и коллекции

В прошлый раз: приложения

- Машинный перевод
- Приложения информационного поиска
 - Собственно поиск,
 - Рубрикация и кластеризация текстов
 - Аннотирование (реферирование)
- Извлечение знаний и информации из текстов
 - Извлечение терминов
 - Извлечение синонимов
 - Извлечение именованных сущностей
 - Извлечение отношений и фактов
 - Извлечение мнений и анализ тональности
- Автоматизация редактирования текстов

Основные свойства естественного языка

- Многозначность
- Неопределенность и зависимость от контекста
- Избыточность
- Универсальность
- Изменчивость

Приложения: лингвистика vs. статистика

- Язык – многоуровневая система
- Сложность обработки текстов, речи
- Описать всю информацию (словари, правила), необходимую для качественной обработки текстов, очень сложно.

Проблема статистических методов при обработке текстов

- Закон Ципфа
 - Упорядочиваем слова по мере снижения частоты встречаемости (ранг r – позиция в тексте):
 - **$f \cdot r = k$**
 - частота встречаемости слова обратно пропорциональна рангу
- В любом тексте, коллекции текстов
 - Небольшое число частотных слов
 - Среднее число среднечастотных слов
 - Большое число редких слов
- Проблема применения статистики (sparse data):
 - Много слов употребляется редко (!),
 - Большинство только один раз («длинный хвост»)

Частоты слов в «Том Сойер»

Adobe Reader - [Foundations of Statistical Natural Language Processing - Christopher D. Manning.pdf]

File Edit View Document Tools Window Help

Open Save a Copy Print Email Search Select Text 200% eBooks Download New Reader Now

Bookmarks Signatures Layers Pages

4

1 Introduction

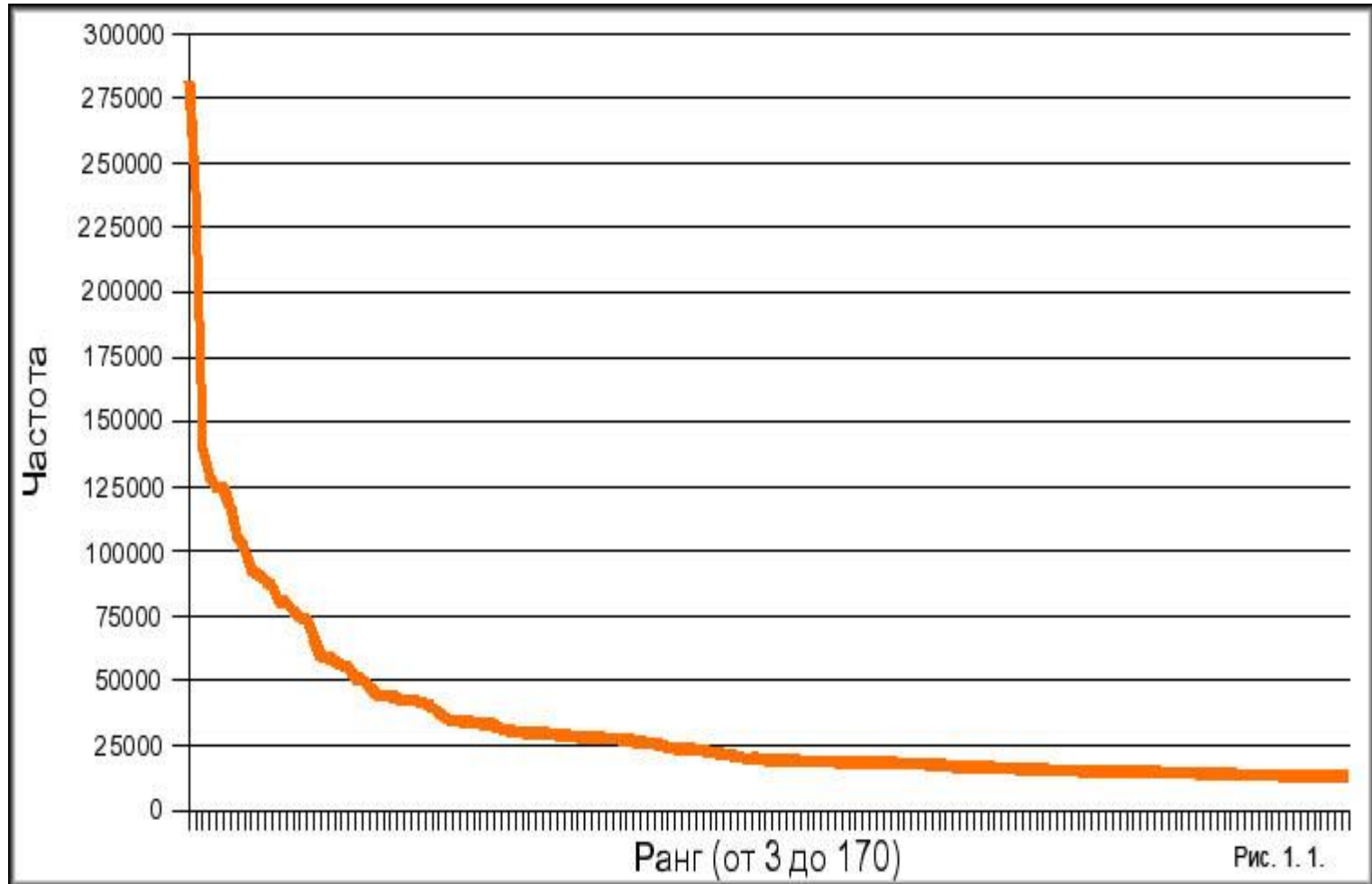
Word	Freq. (<i>f</i>)	Rank (<i>r</i>)	<i>f</i> · <i>r</i>	Word	Freq. (<i>f</i>)	Rank (<i>r</i>)	<i>f</i> · <i>r</i>
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Table 1.3 Empirical evaluation of Zipf's law on *Tom Sawyer*.

(1.14) $f \propto \frac{1}{r}$

19,101 × 23,069 cm 55 of 704

График распределения частот слов по закону Ципфа



Статистика или лингвистика?

- Лингвистические знания
+
- Статистические данные
+
- Эвристические алгоритмы
- => Полезные приложения

Закон Ципфа и интернет-спам

- Поисковая оптимизация
 - Увеличение количества ключевых слов на странице
- Но: нарушение закона Ципфа
 - Поисковые машины (Google, Yandex) проверяют тексты на «естественность»
 - Если закона Ципфа не соблюдается, то
 - понижают рейтинг сайтов с «подозрительными» текстами,
 - либо вообще банят такие сайты.

Текстовые корпуса и коллекции

Текстовые коллекции

- Специализированные порталы (литература, геология...) – электронные библиотеки
- Интернет
- Специальные лингвистически размеченные корпуса
- Текстовые коллекции с пользовательской разметкой (Wikipedia)
- Использование в компьютерной лингвистике
 - Изучение явлений «глазами» для составления словарей, правил
 - Статистические модели

Электронные библиотеки и их разнообразие

- Корпус латинских текстов “Персей”.
- Корпус текстов Ф. М. Достоевского.
- Электронная энциклопедия "Брокгауз и Ефрон".
- Фундаментальная электронная библиотека.
- Российская виртуальная библиотека.
- Библиотека М. Мошкова.
- Электронная библиотека Химического фак-та МГУ.
-
- и др.



Университетская информационная система РОССИЯ

Ресурсы и сервисы для экономических и социальных исследований, учебных программ и государственного управления



О проекте

Источники и условия доступа

Партнеры

Участники

Технологии

Практикум

Поиск по ресурсам УИС РОССИЯ: все коллекции (Уровень доступа = FREE)

Искать

[Ctrl+Enter]

Расширенный поиск

Типовые запросы

User: FREE | Доступ:
FREE

Логин



[Забыли пароль?](#) [Об уровнях доступа](#)

Регистрация

Интегрированная коллекция

- ➔ [Полный список источников](#)
- ➔ [Издания государственных органов](#)
- ➔ [Публикации исследовательских центров](#)

- ➔ [Научные издания](#)
- ➔ [Средства массовой информации](#)
- ➔ [Зарубежные издания](#)

Тематические разделы

Российская Федерация:

- 🔑 [Социально-экономическая статистика](#)
- 🔑 [Дети России](#)

Архивы:

- 🔑 [Регионы и города России. Архив, данные до 2010 года](#)
- [Динамика экономического и социального развития Российской империи в XIX - начале XX вв.](#)
- 🔑 [Население и уровень жизни. Архив, данные за 1996-2007 годы](#)
- 🔑 [Аграрно-промышленный комплекс. Архив, данные за 1990-2005 годы](#)
- 🔑 [Выборы. Архив, данные за 1996-2000 годы](#)
- [Классика Российского права](#)
- [Права человека: документы международных организаций. Архив, документы до 2012 года](#)

Базы данных и online-анализ

- 🔑 [Регионы России. Ежегодное обновление показателей](#) ➔
- 🔑 [Регионы России. Ежемесячное обновление показателей](#) ➔
- 🔑 [Муниципальные образования. Ежегодное обновление показателей](#) ➔
- Архивы:
- 🔑 [Города России. Архив, данные до 2010 года](#) ➔
- 🔑 [Консолидированный бюджет РФ. Архив, данные за 2003-2006 годы](#) ➔
- 🔑 [Аграрная статистика России. Архив, данные за 1990-2005 годы](#)
- 🔑 [Национальное обследование благосостояния домохозяйств. Архив, данные за 2003 год](#)
- 🔑 [Статистика здравоохранения. База данных](#)
- 🔑 [Организации экономического сотрудничества и развития. Архив, данные за 1970-2005 годы](#)

🔑 - для доступа к ресурсу требуется регистрация. [Условия доступа](#)

Полезные ссылки

- ➔ [Библиотеки. Каталоги](#)
- ➔ [Музеи](#)
 - ➔ [Музеи России](#)
 - ➔ [Музеи мира](#)

- ➔ [Нобелевские премии. Лауреаты премии по экономике в память Альфреда Нобеля](#)
 - ➔ [Соотечественники - лауреаты Нобелевских премий](#)
- ➔ [Конституции](#)

Новости

02.09.14
Обновлен раздел
«Социально-
экономическая
статистика»

24.07.14
Обновлен раздел
«Социально-
экономическая
статистика»

15.07.14
Обновлена база
данных «Регионы
России. Ежемесячное
обновление
показателей»


03.07.14
Обновлен раздел
«Социально-
экономическая
статистика»

15.05.14
Обновлен раздел
«Социально-
экономическая
статистика»

13.05.14
Обновлен раздел
«Социально-

☒ **Все коллекции** / [Выбрать открытые коллекции](#) [свернуть](#) / [развернуть](#) список

☒ **Издания государственных органов** [свернуть список](#) | [описание](#)

	<input type="checkbox"/> НТЦ «Система». Нормативно-правовые акты описание	<i>(128702 документа, с 1990 года)</i>
	<input type="checkbox"/> НТЦ «Система». Международные договоры описание	<i>(4364 документа, с 1901 года)</i>
	<input checked="" type="checkbox"/> Государственная Дума ФС РФ. Стенограммы пленарных заседаний описание	<i>(253065 документов, с 1994 года)</i>
	<input type="checkbox"/> Росстат. Ежегодные статистические сборники описание содержание	<i>(77946 документов, с 1997 года)</i>
	<input type="checkbox"/> Министерство экономического развития РФ. Мониторинг и прогнозы содержание	<i>(5829 документов, с 2005 года)</i>
	<input checked="" type="checkbox"/> Банк России. Вестник Банка России описание содержание	<i>(22959 статей, с 1999 года)</i>
	<input type="checkbox"/> Счетная палата РФ. Бюллетень содержание	<i>(2223 документа, с 2004 года)</i>
	<input checked="" type="checkbox"/> Счетная палата РФ. Заключение содержание	<i>(574 документа, с 2000 года)</i>
	<input checked="" type="checkbox"/> Счетная палата РФ. Бюллетень. Архив описание	<i>(789 документов, 1999-2003 годы)</i>
	<input checked="" type="checkbox"/> Росстат. Срочная информация по актуальным вопросам. Архив содержание	<i>(224 документа, 2003 - 2004 годы)</i>
	<input type="checkbox"/> Росстат. Краткосрочные экономические показатели. Архив описание содержание	<i>(2910 документов, 1999-2005 годы)</i>
	<input type="checkbox"/> Росстат. Социально-экономическое положение России. Архив описание	<i>(18341 документ, 1999-2004)</i>

Корпуса текстов

Лингвистический, или языковой, корпус текстов

– большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач.

Корпус vs. электронная библиотека

Тексты в корпусах рассматриваются прежде всего как *образцы* текстов. Тексты в электронных библиотеках, исходя из их назначения, правильнее всего называть *произведениями* со всеми характерными для них атрибутами.

<u>Лингвистический корпус текстов:</u>	<u>Электронная библиотека:</u>
образцы текстов	полные тексты
лингвистическая разметка	библиографические и историко-культурные элементы данных (если имеются)
лингвостатистика	отсутствие статистики
репрезентативность языкового материала "условная"	полнота текстов электронной библиотеки
отбор языкового материала на основе критериев репрезентативности, лингвистической и историко-культурной значимости	отбор текстов, определяемый выбором составителей библиотеки

Лингвистические корпусы

- Brown Corpus.
- Ланкастерский корпус английского языка (Lancaster-Oslo-Bergen Corpus, LOB).
- British National Corpus.
- International Corpus of English.
- Bank of English.
- Cobuild Corpus.
- Мангеймский корпус немецкого языка.
- Чешский национальный корпус.
- Уппсальский корпус русского языка.
- Национальный корпус русского языка.
- Корпусы китайского, турецкого, эстонского, албанского и многих других языков

Корпус

Собственно корпус
(*массив данных*)

+

корпусный менеджер
(*специализированная поисковая
система*)

Разметка

Англ.: tagging, annotation.



Разметка – приписывание текстам и их компонентам специальных меток.

Виды разметки:

- экстралингвистическая (*метаразметка*)
 - сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика;
- структурная
 - (глава, абзац, предложение, словоформа)
- собственно лингвистическая

Классификация корпусов

Признак	Типы корпусов
Тип данных	<ul style="list-style-type: none">•Письменные•Речевые•Смешанные
Язык текстов	<ul style="list-style-type: none">•Русский•Английский и т.д.
«Параллельность»	<ul style="list-style-type: none">•Одноязычные•Двуязычные•Многоязычные
«Литературность», специфичность	<ul style="list-style-type: none">•Литературные•Диалектные•Разговорные•Терминологические•Смешанные
Жанр	<ul style="list-style-type: none">•Литературные•Фольклорные•Драматургические•Публицистические

Классификация корпусов-2

Признак	Типы корпусов
Доступность	<ul style="list-style-type: none">•Свободно доступные•Коммерческие•Закрытые
Назначение	<ul style="list-style-type: none">•Исследовательские•Иллюстративные
Динамичность	<ul style="list-style-type: none">•Динамические (мониторные)•Статические
Разметка	<ul style="list-style-type: none">•Размеченные•Неразмеченные
Характер разметки	<ul style="list-style-type: none">•Морфологические•Синтаксические•Семантические•Просодические и т.д.
Объем текстов	<ul style="list-style-type: none">•Полнотекстовые•«Фрагментнотекстовые»

Русские корпусы в Интернет

Национальный корпус русского языка http://ruscorpora.ru	500 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 морфологическ ая 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

Национальный корпус русского языка



НАЦИОНАЛЬНЫЙ КОРПУС
РУССКОГО
ЯЗЫКА

[главная](#)
[архив новостей](#)

[поиск в корпусе](#)

[что такое корпус?](#)
[состав и структура](#)

[статистика](#)
[графики](#)

[частоты](#)
[морфология](#)
[обороты](#)

[синтаксис](#)
[семантика](#)

[параметры текстов](#)

[studiorum](#)
[форум](#)

[о проекте](#)

Национальный корпус русского языка

[English](#)

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

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

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

[Подробнее о корпусе](#)

Новости проекта

3 июня 2014 года

Объявляется [конкурс проектов нового дизайна](#) Национального корпуса русского языка.

29 апреля 2014 года

Национальному корпусу русского языка [исполнилось 10 лет](#).

29 апреля 2014 года

В режиме бета-версии запущен [поиск по n-граммам](#) подкорпуса с неснятой

Корпус русского языка: состав

- Основной корпус
 - Тексты XVIII-сер. XX веков
 - Тексты после сер. XX
 - Морфологическая разметка, неоднозначность снята
- Газетный корпус (2000-е годы)
- Корпус параллельных текстов (англо-русский, немецко-русский, русско-украинский)
- Корпус диалектных текстов
- Корпус поэтических текстов
- Корпус устной речи (расшифровка записей) и др.

ФРАГМЕНТ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ В НАЦ. КОРПУСЕ РУССКОГО ЯЗЫКА

- Я сидел на барском сиденье, дышал горячим ветром, бившим в лицо, ощущая в то же время не истребимую никакими сквозняками пыль и легкий запах духов -- катафалк с хорошей скоростью мчался по шоссе на юг. (Ю. Трифонов)
- <s>**Я** {я=S,ед,од=им} **сидел**{сидеть=V,несов=изъяв,прош,ед,муж}
на{на=PR} **барском**{барский=A=ед,сред,пр}
сиденье{сиденье=S,сред,неод=ед,пр},
дышал{дышать=V,несов=изъяв,прош,ед,муж}
горячим{горячий=A=ед,муж,твор} **ветром**{ветер=S,муж,неод=ед,твор},
бившим{бить=V,несов=прич,прош,ед,муж,твор} **в**{в=PR}
лицо{лицо=S,сред,неод=ед,вин},
ощущая{ощущать=V=несов,деепр,непрош} **в**{в=PR}
то{тот=A=ед,сред,вин} **же**{же=PART}
время{время=S,сред,неод=ед,вин} **не**{не=PART}
истребимую{истребимый=A=ед,жен,вин}
никакими{никакой=A=мн,твор}
сквозняками{сквозняк=S,муж,неод=мн,твор}
пыль{пыль=S,жен,неод,ед=вин} **и**{и=CONJ}
легкий{легкий=A=ед,муж,вин,неод} **запах**{запах=S,муж,неод=ед,вин}...

Синтаксически размеченный корпус в НКРЯ (СинТагРус)

Предложению поставлена в соответствие синтаксическая структура

Дерево зависимостей:

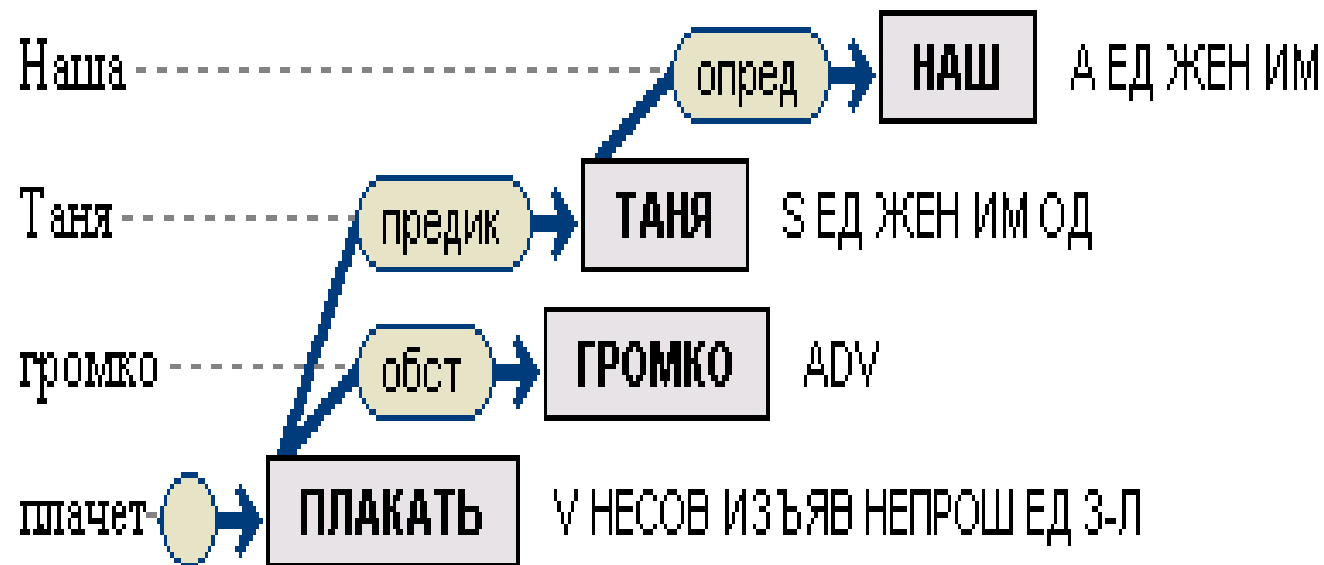
- Узлы - слова предложения
- Ветви – имена синтаксических отношений
- Дерево зависимостей:

Автоматизированная разметка

- Синтаксический анализатор
- Проверка лингвистом

Может быть использован для обучения статистического синтаксического анализатора

Синтаксическая разметка в НКРЯ-2



Поисковая форма в НКРЯ

главная

основной

синтаксический

газетный

параллельный

обучающий

диалектный

поэтический

устный

акцентологический

мультимедийный

исторический

использование корпуса

Основной корпус

Поиск точных форм ?

Слово или фраза

Лексико-грамматический поиск ?

Слово ? <input type="text" value="А Б В"/>	Грамм. признаки ? выбрать	Семант. признаки ? <input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	Словообразование выбрать	<input checked="" type="checkbox"/> 1-е знач. <input checked="" type="checkbox"/>
<input type="text"/>	<input type="text"/>	

Расстояние: от до ?

Слово ? <input type="text" value="А Б В"/>	Грамм. признаки ? выбрать	Семант. признаки ? <input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>
Доп. признаки ? выбрать	Словообразование выбрать	<input checked="" type="checkbox"/> 1-е знач. <input checked="" type="checkbox"/>
<input type="text"/>	<input type="text"/>	

Национальный корпус русского языка

Поиск

«Новые» подкорпуса НКРЯ

- Корпус устной речи (с 2007 г.)
 - Расшифровки магнитных записей публичной и частной устной речи
 - Транскрипты фильмов
- Мультимедийный русский корпус (с 2010 г.)
 - Параллельный видеоряд, аудиоряд,
 - текстовая расшифровка
 - Звучащей речи
 - Наблюдаемых жестов
 - Возможен поиск
 - по жестам (кивание головой, похлопывание по плечу...)
 - по типу речевого действия (согласие, ирония...)

Пример ситуации «согласие» в мультимедийном корпусе

- Разговор на рынке в Дальнегорске - дается аудиозапись
- [Женщина1, жен] Нарód... / ста́л ху́же жи́ть че́м во́т мы́ ра́ньше жи́ли / когда́ вот рабо́тали / и все́... /
[Женщина2, жен] Хотя́ / зна́ете / я обрати́ла внимáние / це́ны на фрúкты / сейча́с / таки́е же / ка́к и во Владивосто́ке / в Уссурíйске // Во́т / не отлича́ются //
- [Женщина1, жен] Це́ны / це́ны де́ржатся / да //
- [Женщина2, жен] А во́т / в те́ го́ды / ра́зница была́ больша́я // Во́т э́то наве́рное ка́к-то уже́ по все́му кра́ю э́та торго́вля идёт... /
[Женщина1, жен] Ну́ так / э́то ж все́ заво́зное / все́ ж приво́зное / да //
[Женщина2, жен] Э́ти ры́нки все́ / кита́йский това́р все́ то же...

Задачи создания корпусов

- Отбор и подготовка текстов;
- репрезентативность;
- хронологические рамки;
- разметка;
- разные задачи → разные типы корпусов;
- трудоёмкость;
- специализированное программное обеспечение (corpus managers).

Вопрос

- Предположим, задачи создания корпуса решены – корпус создан
- Вопрос: какие задачи плохо решаются с помощью таких тщательно сделанных корпусов?

Проблемы «замкнутых» корпусов (Беликов и др., Диалог-2012)

- Неполнота и несбалансированность
 - Включение большого литературного произведения сильно повлияет на частотность слов
- Проблема периферийных явлений
 - Региональная лексика
 - Общая лексика за пределами частотного словаря 30-40 тыс. слов
- Исследование динамики языковых изменений
 - Анализ новых слов, значений, сочетаемости
 - «Мэрия согласовала нам маршрут движения»
 - Новое значение *согласовать* – *разрешить*
 - (НКРЯ – 2 примера)

Web как корпус

Интернет – огромный справочник, всемирная библиотека, всемирный архив текстовой информации.

- Объем
- Удваивается каждые...
- Любые типы текстов
- Разные языки
- Динамика

Поисковые системы как инструмент лингвистического анализа

- Глобальные поисковые системы
 - Поиск по точной словоформе и по слову во всех формах
 - Поиск по словосочетанию, можно задавать расстояние между словами
 - Язык запросов
- Но:
- Статистика – приблизительная
 - Объем данных слишком большой
 - Делается приблизительная оценка частоты встречаемости
- Поисковые системы пытаются расширить запрос


Числовые данные при поиске в Интернет


Поиск [Почта](#) [Карты](#) [Маркет](#) [Новости](#) [Словари](#) [Блоги](#) [Видео](#)


Яндекс
Нашёлся
31 млн ответов

согласовать

☐ в найденном ☐ в Москве

 [Рядом с Кремлем повесили баннер в память о Бабуровой](#)
Мэрия Москвы и организаторы намеченного на 19 января антифашистск память об убитых четыре года назад адвокате Станиславе Маркелове и Анастасии Бабуровой, несмотря на прозвучавшие ранее в СМИ сообще **согласовать** акцию, пришли к компромиссу.
[Известия](#) 11:18 [Вести.Ru](#) 10:19 [Lenta.ru](#) 08:24 [Все сообщения](#) 121
[Мэрия Москвы отказала Лимонову в проведении митинга 31 января](#) 29 сс
[Минюст предложил согласовывать религиозные мероприятия с властью](#)
[news.yandex.ru](#) 1 час назад

1  [согласовать — Викисловарь](#)
согласовать. Материал из Викисловаря. Текущая версия (не проверяла Соответствующий глагол несовершенного вида — **согласовывать**.
[ru.wiktionary.org](#) > [wiki/согласовать](#) копия ещё

2  [СОГЛАСОВАТЬ](#)
СОГЛАСОВАТЬ — **СОГЛАСОВАТЬ**, согласую, согласишь, совер. (к с что и что с чем. Привести в связь, в согласие, в надлежащее соотношен [dic.academic.ru](#) > [Согласовать](#) копия ещё

<http://news.yandex.ru/yandsearch?cl4url=www.bfm.ru/ne...oskvy-ne-soglasovala-miting-limonova-31-janvarja.html>

Особенности статистики в поисковых системах (Беликов и др., 2012)

- Надежность выдаваемых цифр тем больше, чем цифры — меньше. А полностью доверять можно только тем цифрам, которые можно проверить полным просмотром выдачи.
- К числовым результатам не применимы аксиомы классической арифметики
 - часть вполне может быть больше целого
 - уточняя запрос, получаем больше «результатов», чем для общего запроса
- Имеет место нестабильность: результат меняется во времени в произвольную сторону (не связанную с реальным изменением числа релевантных объектов)
- Вопрос: почему это происходит?

Еще проблемы Интернета как корпуса

- Отсутствие разметки (морфологической, синтаксической, жанровой, ...)
- Географическая разметка (географическая привязка текстов) может не иметь отношения к автору текста
 - При поиске важно про что сказано, а лингвистам важно – кто сказал
- Проблемы дублирования и скрытого цитирования
 - Необоснованное увеличение частотности употребления некоторых слов, оборотов
- Неоднородность страницы
 - Содержание vs. оформление интернет- страниц
 - Реклама на странице
- Поисковый спам

Более «чистые» данные: новостные сервисы: числовые данные, контекст

Firefox (12) Входящие - louk_nat@mail.ru - ... Я согласовать — Яндекс: нашёлся 31 ... согласовать (1003656): Яндекс.Ново... +

news.yandex.ru/yandsearch?text=согласовать&rpt=nnews2&grhow=clutop

Поиск Почта Карты Маркет **Новости** Словари Блоги Видео Картинки ещё Войти Помощь

Яндекс
новости


согласовать Найти


сообщений (1 003 656) статей (236 507) интервью (21 206) расширенный поиск с видео (15 229)


Главные новости Политика В мире Общество Экономика Спорт Происшествия Культура Наука Здоровье Hi-Tech Интернет Авто Т


по дате по релевантности **сегодня** 3 дня неделя месяц всё не группировать по сюжетам

Рядом с Кремлем повесили баннер в память о Бабуровой и Маркелове

Интерфакс  Мэрия Москвы **согласовала** акцию памяти Бабуровой и Маркелова 18:33 15.01
INTERFAXS.RU - Власти Москвы заявили, что **согласовали** с гражданскими активистами проведение шествия 19 января в центре Москвы в память об адвокате Станиславе Маркелове и...

Новая газета  Мэрия Москвы все же **согласовала** шествие в память Станислава Маркелова и Анастасии Бабуровой 16:07 16.01
Шествие, **согласованное** на 700 человек, пройдет 19 января – в годовщину гибели адвоката и журналистки «Новой газеты».

РИА Новости  Мэрия Москвы **согласовала** шествие антифашистов 19 января 10:32 вчера
Власти Москвы **согласовали** антифашистское шествие 19 января в память об адвокате Станиславе Маркелове и журналистке Анастасии Бабуровой по Тверскому бульвару до...



HP Deskjet Ink Advantage Новое поколение при для доступной печати

Подписка на новости

http://www.novayagazeta.ru/news/62233.html

Интернет-корпус университета Лидс (Шаров и др., 2010)

- <http://corpus.leeds.ac.uk/internet.html>
- Выбираются 500 из наиболее частотных слов языка:
имеет, головой, особенно
- Формирование запросов из нескольких слов (5000 – 6000 тыс)
- Исполнение запросов через Гугл
- Выкачивание страниц
- Постпроцессинг
- Получается многотематический корпус:
 - представлены разные тематики
 - представлены разные жанры
- **Сейчас интернет-корпуса по разным методикам создаются для разных языков**

Сколковский проект компании АBBYY: Генеральный интернет-корпус РЯ

- Генеральный корпус – открытое подмножество Рунета с постоянным пополнением
- Сбалансированность корпуса
 - Учет сегментов Интернета: новости, социальные сети, литературные произведения ...
- Метатекстовая классификация и разметка
 - Автоматическая классификация по темам, по жанрам
- Автоматическая морфологическая и синтаксическая разметка
- Очистка страниц
- Специальные средства поиска

User-generated (пользовательский) content

- Пользовательский контент
 - Википедия
 - Социальные сети
 - Twitter
- Разметка
 - Смайлики в Twitter
 - Ссылки в Википедии
 - Википедия: статьи на одну и ту же на разных языках
- Множество научных исследований этих ресурсов, в том числе
 - с точки зрения лингвистики,
 - с точки зрения приложений автоматической обработки текстов

Исследование пользовательского контента (научный сервис Google)

Веб Картинки Ещё... louk.natalia@gmail.com

Google twitter text

Академия Результаты: примерно 1 220 000 (0,05 сек.) За все время 21

Совет. По этому запросу вы можете найти сайты на русском языке. Указать предпочтительные языки для результатов поиска, в том числе и русском, можно в разделе [Настройки Академии](#)..

[Short text classification in twitter to improve information filtering](#) [ohiolink.edu \[PDF\]](#)
[B Sriram, D Fuhry, E Demir...](#) - [Proceeding of the 33rd ...](#), 2010 - [dl.acm.org](#)
Abstract In microblogging services such as **Twitter**, the users may become overwhelmed by the raw data. One solution to this problem is the classification of short **text** messages. As short texts do not provide sufficient word occurrences, traditional classification methods ...
Цитируется: 78 Похожие статьи Все версии статьи (15)

[Twitter mood predicts the stock market](#) [arxiv.org \[PDF\]](#)
[J Bollen, H Mao, X Zeng](#) - [Journal of Computational Science](#), 2011 - Elsevier
... We analyze the **text** content of daily **Twitter** feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. ... Research highlights. ► Public mood states along 7 different dimensions of mood are measured from the **text** content of large-scale **Twitter** feeds. ...
Цитируется: 291 Похожие статьи Все версии статьи (30)

[\[PDF\] Lexical normalisation of short text messages: Makn sens a# twitter](#) [mu.oz.au \[PDF\]](#)
[B Han, T Baldwin](#) - [Proceedings of the 49th Annual Meeting of the ...](#), 2011 - [ww2.cs.mu.oz.au](#)
Bo Han and Timothy Baldwin NICTA Victoria Research Laboratory Department of Computer Science and Software Engineering The University of Melbourne hanb@student.unimelb.edu.au tb@ldwin.net ... Short **text** messages from mobile phones and micro-blogs: real-time, ...
Цитируется: 45 Похожие статьи Все версии статьи (11) Ещё ▾

[\[PDF\] Unsupervised modeling of twitter conversations](#) [nrc-cnrc.gc.ca \[PDF\]](#)
[A Ritter, C Cherry, B Dolan](#) - 2010 - [nparc.cisti-icist.nrc-cnrc.gc.ca](#)
... Table 1: A sample of **Twitter** spelling variation. pus, and manually picked out clusters of spelling variants; a sample is displayed in Table 1. **Twitter's** noisy style makes processing **Twitter text** more difficult than other domains. ...
Цитируется: 79 Похожие статьи Все версии статьи (26) Ещё ▾

[Sentiment knowledge discovery in twitter streaming data](#) [hughchristensen.co.uk \[PDF\]](#)
[A Bifet, E Frank](#) - [Discovery Science](#), 2010 - Springer
... There are also a number of interesting tasks that have been tackled using **Twitter text** mining: sentiment analysis, which is the application we consider in this paper, classification of tweets

Задание на дом

- Взять художественное произведение
- Извлечь слова
 - Убрать капитализацию (т.е. большие буквы)
 - Отделить слова от знаков препинания
- Посчитать частоты слов
- Проверить закон Ципфа
 - График
 - Отчет

Задание в аудитории

- Выбрать одно из следующих слов:
 - Скраб
 - Короб
 - Откат
 - Распил
 - Бачок
 - Халтура
 - Шпажка
 - Ватрушка
 - Схрон
 - Корзина

Задание

- Посмотреть значения и толкования слов в словарях Даля, Ушакова, Ожегова, Кузнецова
 - <http://dic.academic.ru/>
- Употребление этого слова в корпусе русского языка (ruscorpora.ru)
 - В основном корпусе до середины 20 века
 - В основном корпусе (после сер. 20 века)
 - В газетном корпусе
- Употребление этого слова в Яндекс-новостях
- Представить отчет о сходствах-различиях в словарных описаниях, реальном употреблении

Структура отчета

- Слово
- Толкования в словарях: сходство-различие
- Употребление в корпусах
 - Сходство-различие в употреблении
- Соответствия между толкованиями и реальными употреблениями
 - Например, в словарях нет какого-то значения
- Выводы