Занятие 1

Компьютерная лингвистика
Корпуса

Катя Герасименко
Летняя олимпиадная школа МФТИ
30.07.2018

Организационное

Репозиторий, где будут все презентации и ноутбуки:

https://github.com/religofsil/mipt-summer-school/tree/master/2018

Домашнее задание присылать мне вк:

https://vk.com/nyancat_kate

И Маше Бибаевой вк:

https://vk.com/sanmelisan

Дедлайн – через занятие

После просрочки дедлайна на 2 дня – от оценки ставится половина

Чем занимается компьютерная лингвистика

Компьютерная лингвистика (computational linguistics; комплинг(в)) занимается применением формальных моделей к естественному языку

- ресурсы и инструменты для работы с языком
- формальные модели (морфология, синтаксис)
- автоматическая обработка текстов (автобрея, NLP от Natural Language Processing)

NLP

- информационный поиск (information retrieval)
- извлечение информации (information extraction)
  - именованные сущности (Петя Иванов, ООО «Газпром», Нью-Йорк, etc.)
  - факты
  - связи
- и всякие другие вещи

Примеры из жизни

- Поисковики
- Автодополнение, спеллингчек
- Переводчики (+ определение языка)
- Чат-боты, помощники, ассистенты (специализированные, общие)

Откуда берутся данные

Для работы многих современных систем нужны обучающие данные, то есть тексты (в широком смысле) на естественном языке

Корпус – это коллекция текстов на естественном языке (а тексты эти бывают очень разные)

+ **разметка** самых разных вещей (морфология, синтаксис, семантика, ударения, именованные сущности, ошибки…)

Зачем нужен корпус

Научные исследования:

1) сочетаемость (с какими существительными чаще встречается *strong*, а с какими – *powerful*?)
2) конструкции (*все х как х, а ты…* - какое слово может стоять вместо х?)
3) диахрония - изменение языка во времени (как изменилось употребление слова *именно* по сравнению с XVIII веком?)

Корпус как альтернатива элицитации (опросу носителей языка по анкете с конкретными примерами)

Зачем нужен корпус

На корпусах можно обучать разные модели NLP

- языковые модели
- размечены именованные сущности -> распознавание именованных сущностей
- корпус разбит по темам -> тематическое моделирование
- параллельные корпуса -> переводчики

Какие бывают корпуса

- Обычные :) просто разные тексты
- Корпуса малых языков
- Мультимедийные (аудио, видео)
- Поэтические
- Параллельные
- Учебные

И так далее!

Свойства корпуса

Если мы говорим о корпусе «общего назначения» (например, НКРЯ), то важно, чтобы он был **сбалансированным:**

- все жанры представлены (литература XVIII века и форумы из 2005)
- нет большого перевеса одного жанра (всех выпусков ста больших газет или полного собрания сочинений одного писателя)

Открытые корпуса для русского

Для исследований – НКРЯ (http://ruscorpora.ru)

Для обучения:

- OpenCorpora (http://opencorpora.org/)
- Taiga (https://tatianashavrina.github.io/taiga_site/)

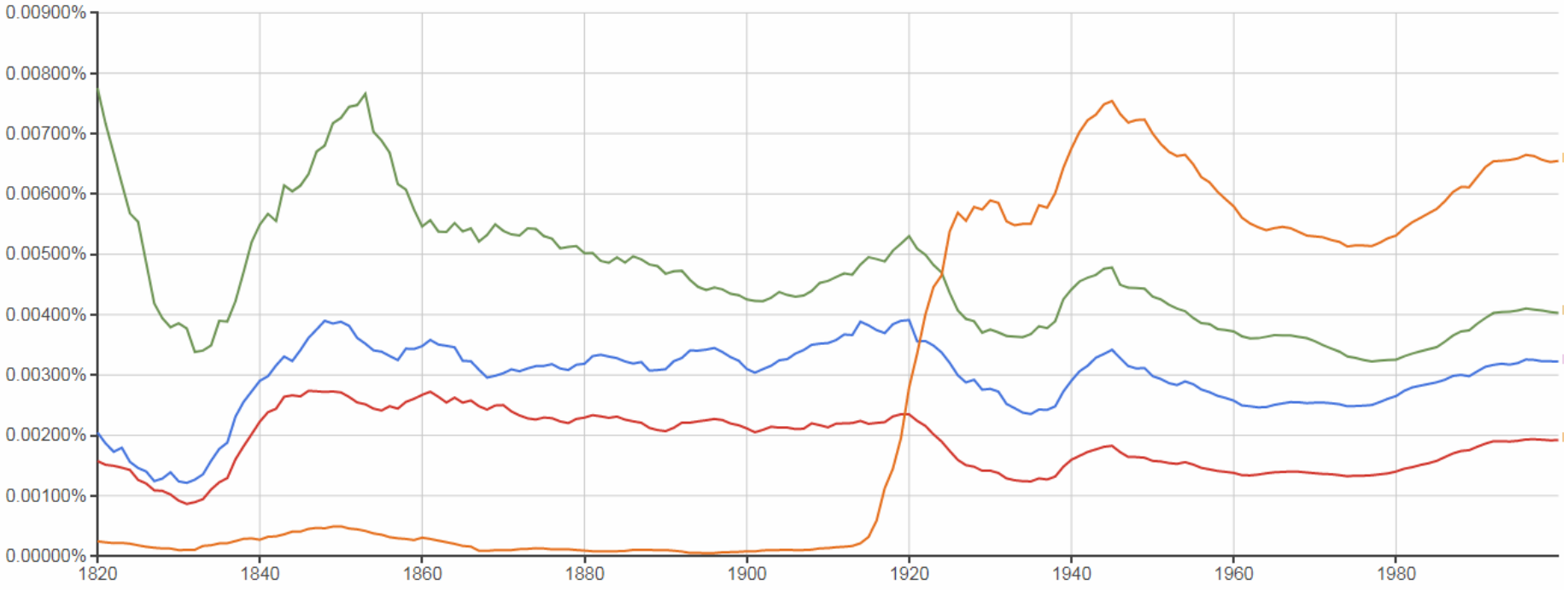
И веб-ресурсы:

- Русская Википедия
- Open Subtitles
- etc.

Задача из Google Books Ngram Viewer

Автор: Борис Иомдин

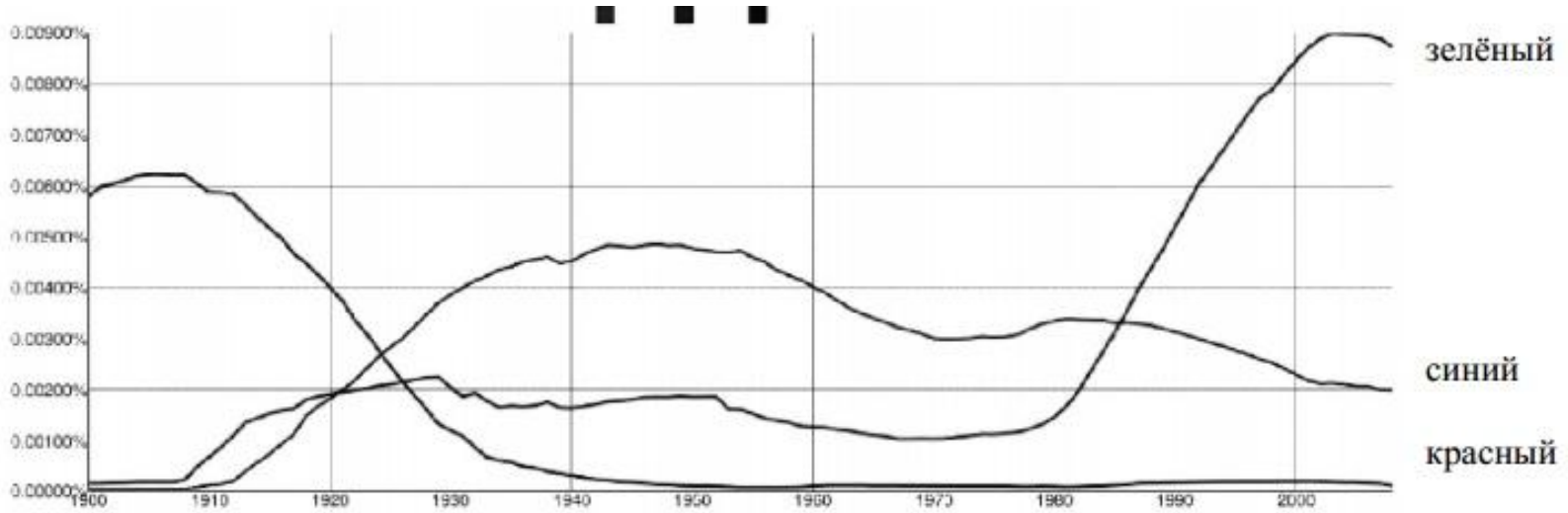
пришел, пришла, пришло, пришли – с 1820 по 2000



Задача-подсказка

Автор: Александр Бердичевский

Какое написание каким цветом обозначено?



(А) синим — «Богъ», красным — «Бог», зелёным — «бог»;
(Б) синим — «бог», красным — «Богъ», зелёным — «Бог»;
(В) синим — «бог», красным — «Бог», зелёным — «Богъ»;
(Г) синим — «Бог», красным — «бог», зелёным — «Богъ»;
(Д) синим — «Бог», красным — «Богъ», зелёным — «бог».

Домашнее задание

У глагола *афишировать* есть интересная особенность в употреблении. Посмотрите примеры по НКРЯ и попробуйте понять, что это за особенность.

В русском (и не только) языке есть целый класс таких слов и выражений. Попробуйте придумать, какие еще выражения вы бы отнесли к этому классу (это задание уже не совсем на корпус).