

Вычислительная лингвистика



Петрозаводский государственный
университет



Крижановский Андрей Анатольевич

andrew.krizhanovsky@gmail.com¹

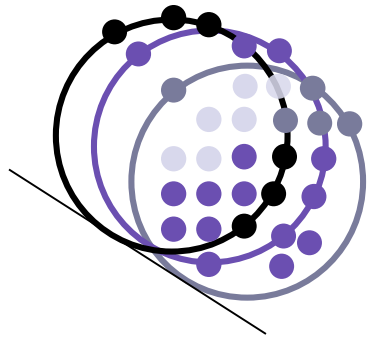


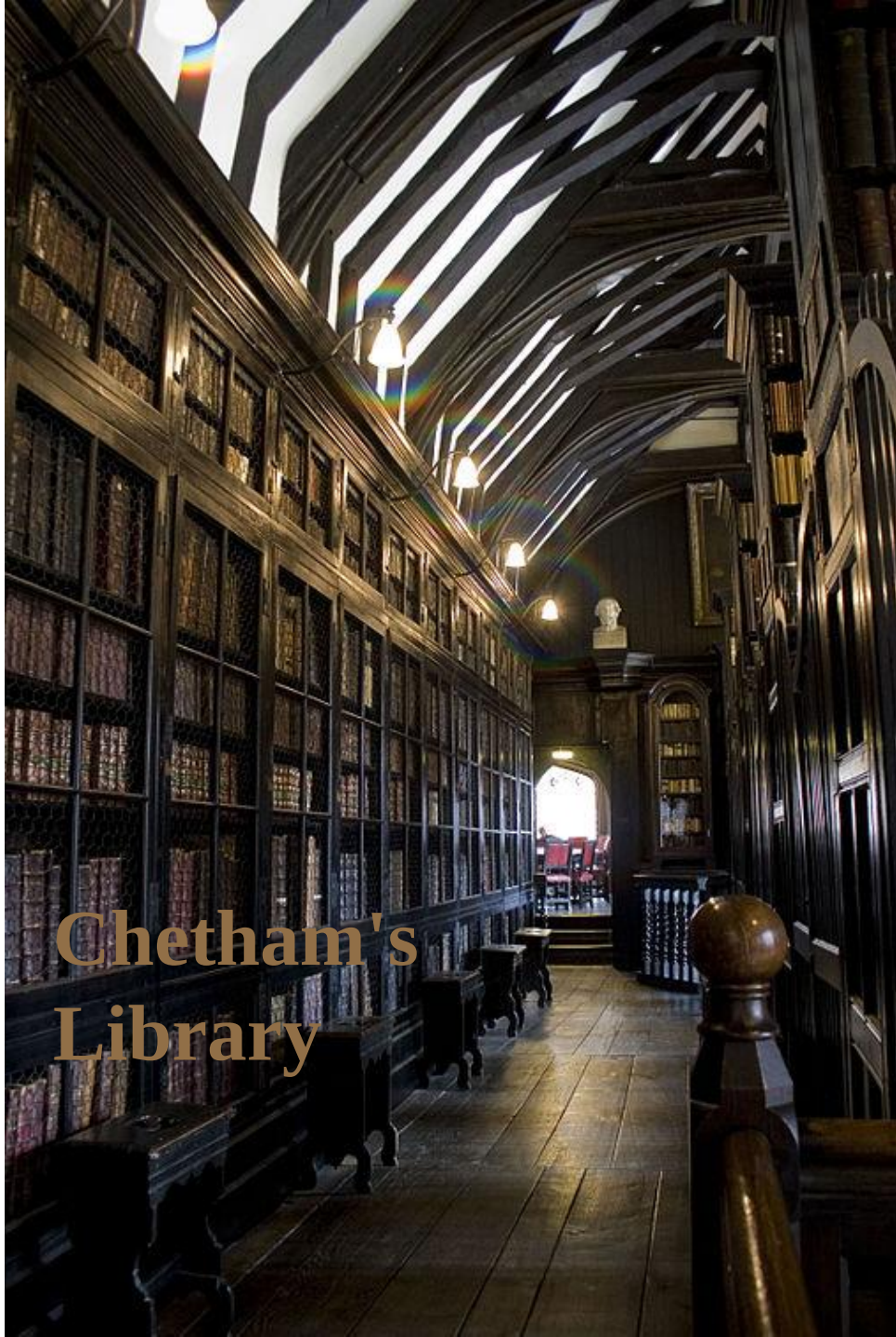
A badly reported bug is much worse than a badly drawn one...



... At least I could claim
I was being "artistic"

cartoontester.blogspot.com © 2013



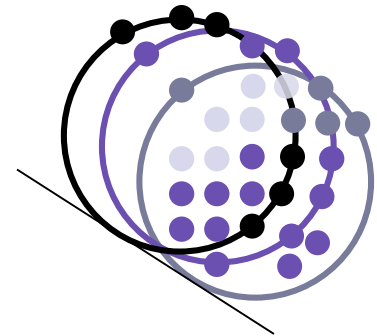


Chetham's
Library

6 сентября

1852 — В Манчестере
открыта первая
бесплатная библиотека
Британии

1989 —
Из-за компьютерной
ошибки 41 тыс. парижан
получили письма,
извещающие о том, что
ими совершены убийства
и грабежи вместо
нарушений правил
дорожного движения.



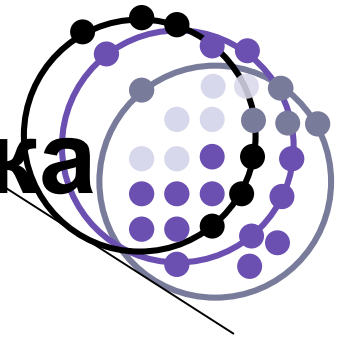
О курсе

ИНТЕРНЕТ-МАТЕМАТИКА



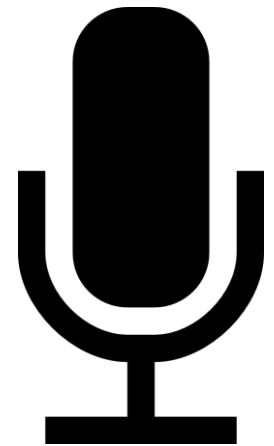


Ещё околонаучная практика



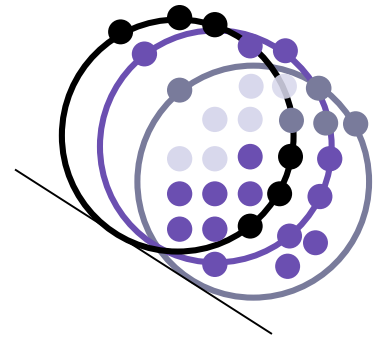
Освойте специальности:

- рецензент научного журнала (WRN)
- личный библиограф (премия Тьюринга)
- диктор и звукорежиссёр (имена русских учёных)
- архивариус (РГАФД)
- админ ВП (удалить и спасти статью)

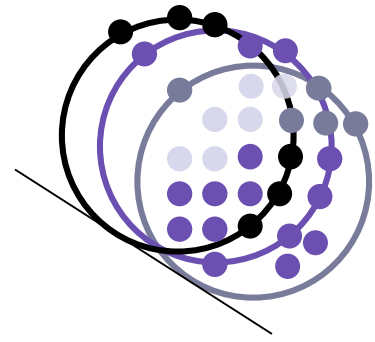




Теория



- Лицензии, авторское право и ВП
- Организация информации в ВП
 - Внутренние ссылки, братские ссылки, проект «Связность», WikiData
- Информационный поиск
 - Инвертированный индекс
 - Расширенные модели булева поиска
 - Оптимизация обработки булевых запросов
- Вычислительная лингвистика
 - Лексикография и корпуса. Толковые словари
 - Корпусная лингвистика



Семиотика, Вычислительная лингвистика

ЛИНГВИСТИКА

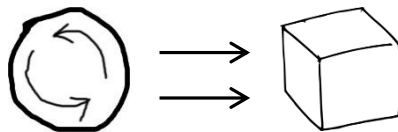


Семиотика

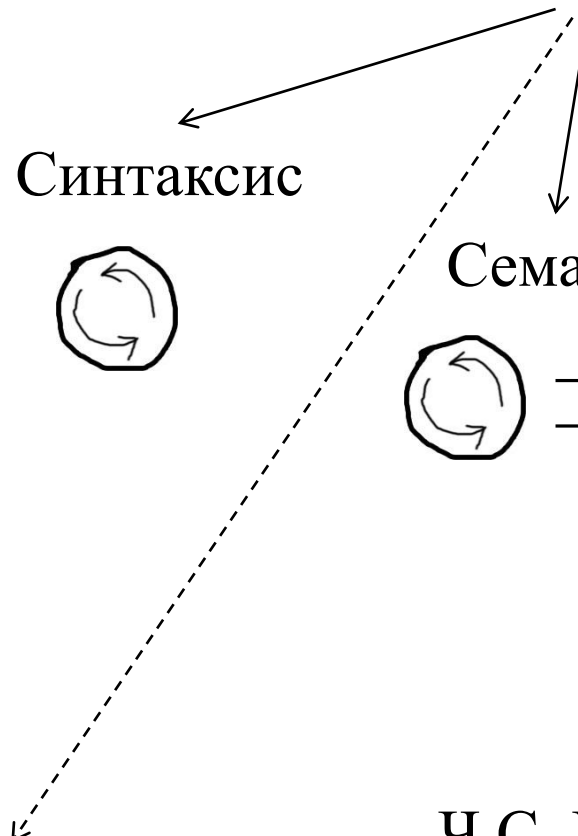
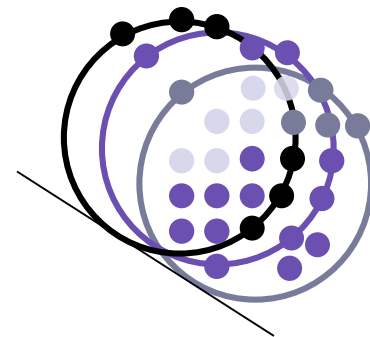
Синтаксис



Семантика



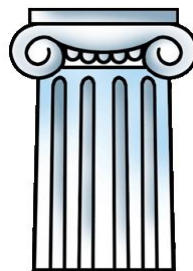
Прагматика

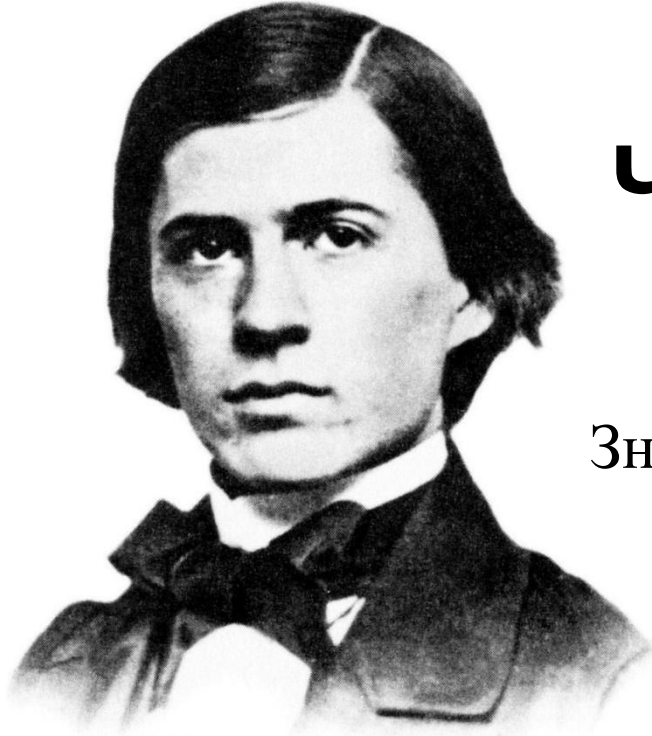


Лингвистика

Ф. де Соссюр

Ч.С. Пирс





Чарльз Сандерс Пирс

Знак — это любое A , обозначающее B для C .

Стрелка Пирса

X	Y	$X \downarrow Y$
0	0	1
0	1	0
1	0	0
1	1	0

$$\overline{x + y}$$

В зависимости от отношения знака к его объекту знаки делятся на:

- (1) иконы,
- (2) индексы,
- (3) символы



Фердинанд де Соссюр



Parole
(речь)



Langue
(язык)

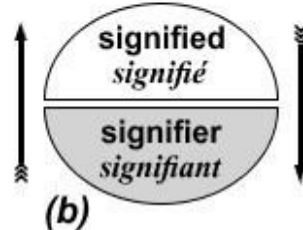
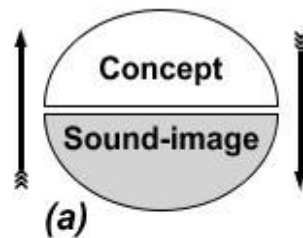
(what the individual speaks)

(what is shared by the community)

Ferdinand de Saussure

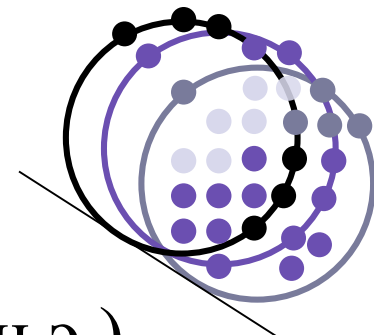
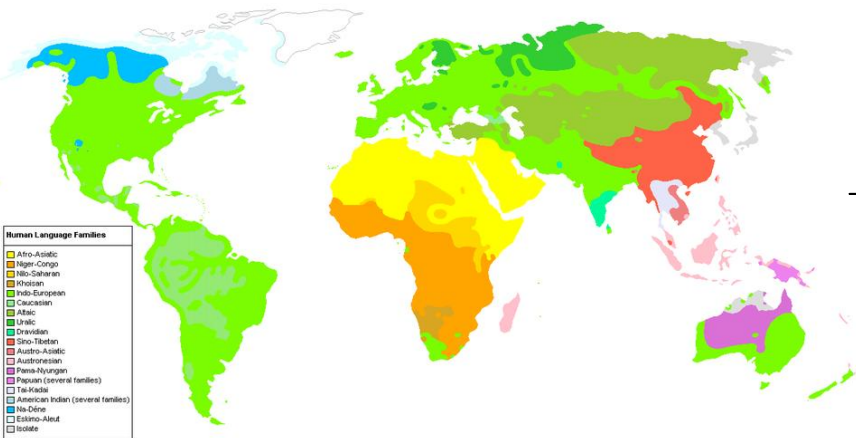
Cours de
linguistique
générale

Payot



Разделы:

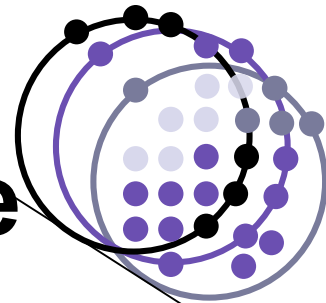
1. Теоретическая
(эмпирическая, нормативная)
(общая, частная)
2. Прикладная
метод обучения языку (родной, иностранный), лексикография
3. Практическая
(эксперименты -> 1.)



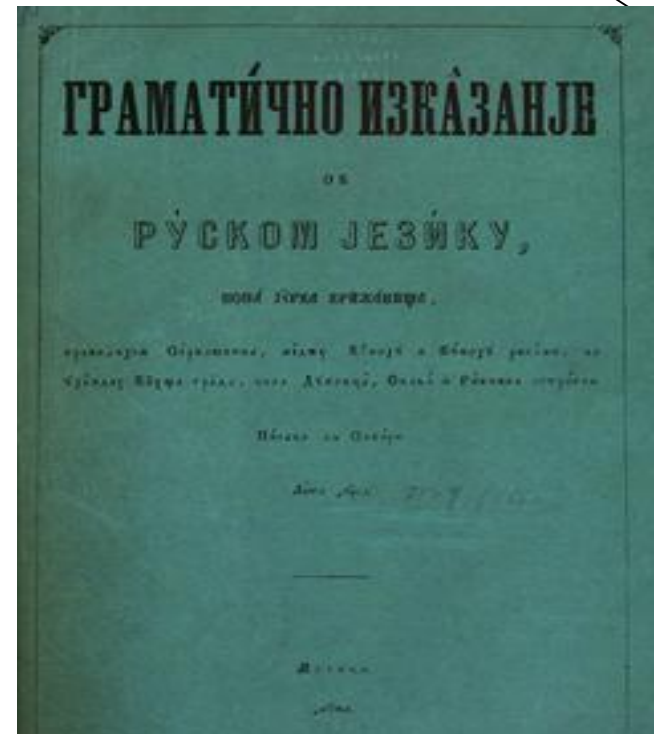
История:

- Rāṇini (V в. до н.э.)
нормативная грамматика санскрита
«Аштадхьяи»
- Yāska - грамматик,
фонетика → словоизменение
- Аристотель, Платон
- Сравнительно-
исторического
языкознания
Юрий Крижанич (всеславянский яз.)
Уильям Джонс, Франц Бопп
(индологи, основатели)
А. Х. Востоков (слав. + мёртв)

Сравнительно-историческое языкознание



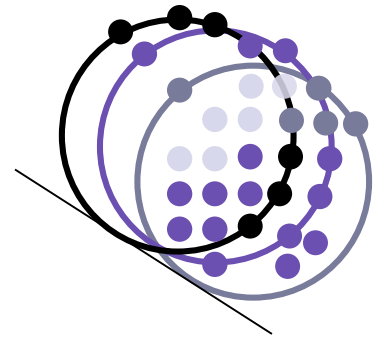
- Юрий Крижанич — хорватский богослов, философ, писатель, лингвист-полиглот...
- Цель: объединить славянские народы
- Способ: всеславянский язык, без иностр. заимствований:
 - 60 % слов — общеславянского происхождения, 10 % — русские и церковнослав., 9 % — сербохорв., 2,5 % — польские, + болг., укр.
 - +: «чужебесие», «гостогонство», «людодер»



Первый в Европе труд по сравнительному языкознанию.



На стыке наук



Науки:

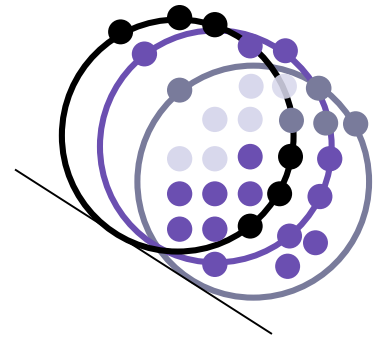
1. Право
2. Генеалогия
3. История
4. География
5. Математика
6. Статистика
7. Информатика

Стыки:

- А. Антропонимика
- В. Вычислительная лингвистика
- К. **Квантитативная лингвистика**
- М. Математическая лингвистика
- С. Лингвистическая палеонтология
- Т. Топонимика
- Ю. Юрислингвистика



Искусственный интеллект



Вычислительная лингвистика

- мат. модели для описания ЕЯ

теория

syn: математическая
лингвистика

syn: компьютерная
лингвистика

*computational linguistics
(CL)*

Обработка естественного языка

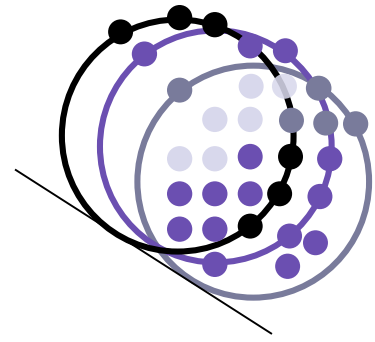
- преобразование текстов на ЕЯ с помощью комп-х программ

прагматика

*natural language
processing (NLP)*



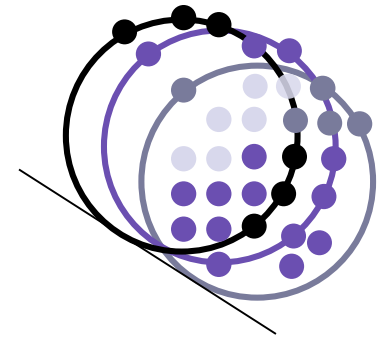
Смежные области исследований CL (1)



- Информатика (Computer Science)
- Искусственный интеллект
- Математика
 - формальные грамматики — 4 компонента:
 - Σ — набор терминальных символов («буквы»)
 - N — набор нетерминальных символов (формула, арифметическое выражение)
 - правила вывода
 - аксиома (или начальный символ) из N



Смежные области исследований CL (2)



- Математика (Математическая лингвистика)
 - ❖ Порождающие (формальн.) грамматики Н. Хомского
 - ❖ Система правил позволяет проверить грамматически правильное предложение
 - ❖ Частный случай: КС-грамматика (многие языки программирования)

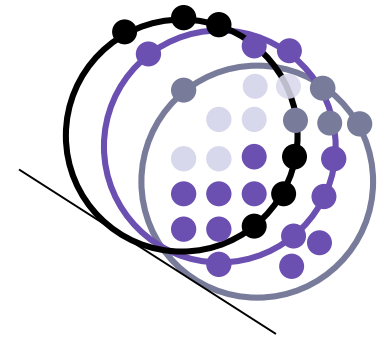
Compilation error

Parse error. Internet Explorer has a non-standard
Arrays will have the wrong length and objects v



```
1. ФОРМУЛА => ФОРМУЛА ЗНАК ФОРМУЛА
2. ФОРМУЛА => ЧИСЛО
3. ФОРМУЛА => ( ФОРМУЛА )
4. ЗНАК => + | - | * | /
5. ЧИСЛО => ЦИФРА
6. ЧИСЛО => ЧИСЛО ЦИФРА
7. ЦИФРА => 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
```

ФОРМУЛА => ? => (25/5)



ФОРМУЛА $\Rightarrow$? $\Rightarrow$ (25/5)

1. ФОРМУЛА $\Rightarrow$ ФОРМУЛА ЗНАК ФОРМУЛА

2. ФОРМУЛА $\Rightarrow$ ЧИСЛО

3. ФОРМУЛА $\Rightarrow$ { ФОРМУЛА }

4. ЗНАК $\Rightarrow$ + | - | * | /

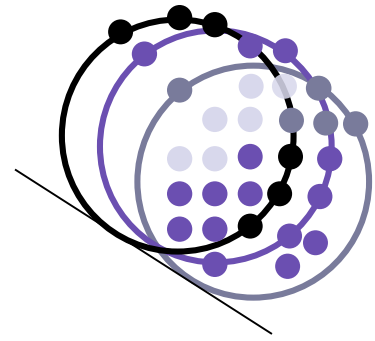
5. ЧИСЛО $\Rightarrow$ ЦИФРА

6. ЧИСЛО $\Rightarrow$ ЧИСЛО ЦИФРА

7. ЦИФРА $\Rightarrow$ 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

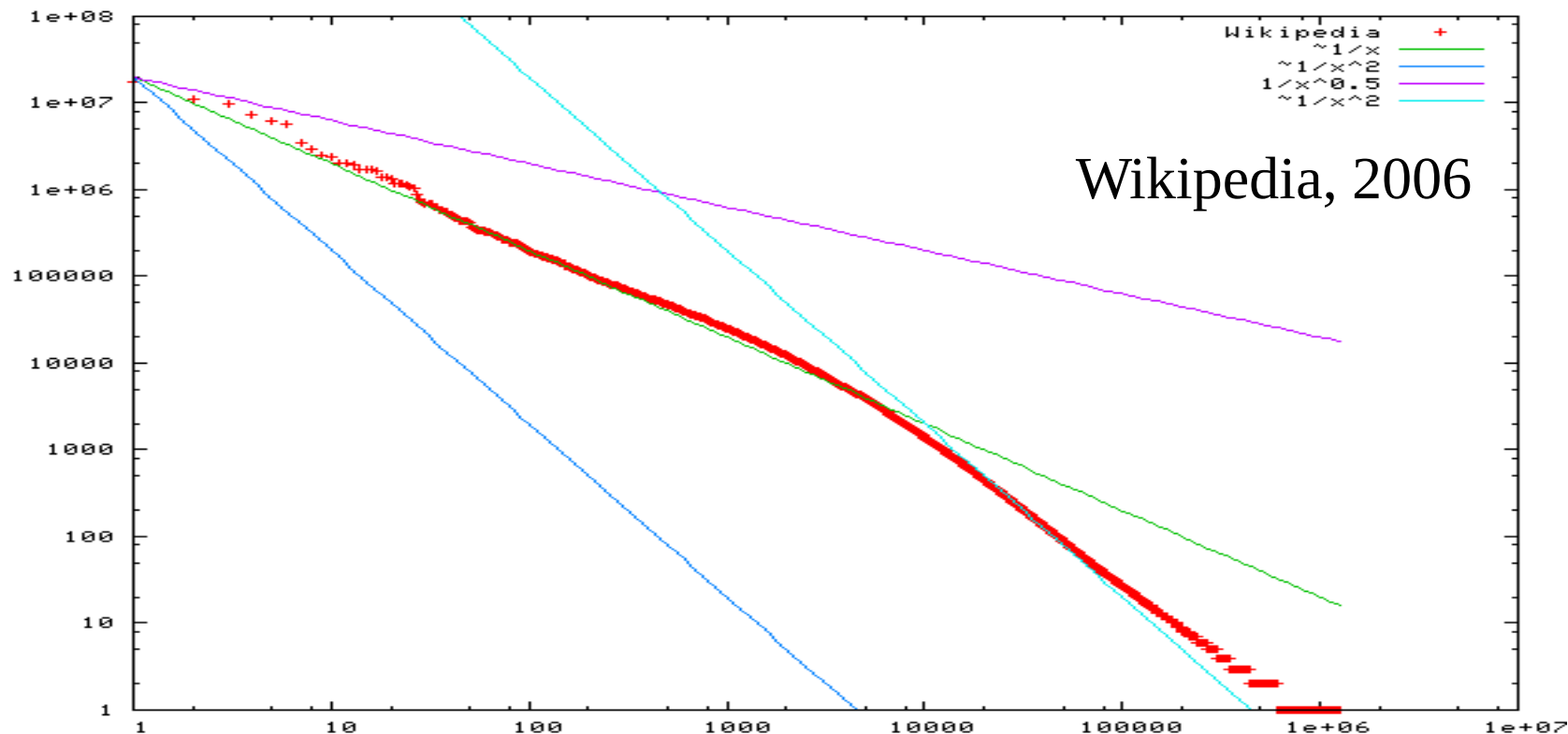


Смежные области исследований CL (3)



- Математика (Квантитативная лингвистика)
 - ❖ Методы статистики + Корпусная лингвистика => языковые законы
 - ❖ з-н Мартина: толкований толкований в словаре ↘ значение более обще ↗ (иерархия)
 - ❖ з-н Менцерата: размер составляющих уменьшается при увеличении размера целого
 - ❖ з-н Ципфа: ...





- ❖ *log-log* coordinates. X is rank of a word in the frequency table; Y is the total number of the word's occurrences.
- ❖ Zipf's law corresponds to the upper linear portion of the curve, roughly following the green ($1/x$) line.

Смежные области исследований CL (4)

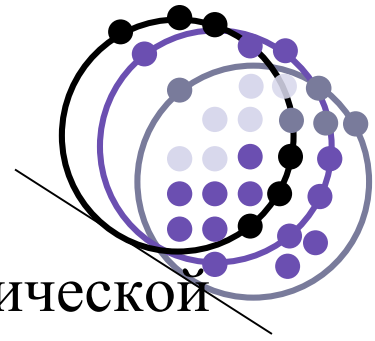
- ИИ — техническая или программная система, способная решать «творческие» задачи
- Что значит «творческая»?



- Эмпирический тест — тест Тьюринга
- ✓ ● Вычислительные машины и разум — w:Computing Machinery and Intelligence (статья 1950 г.)



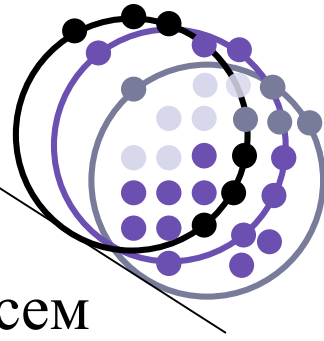
Задачи СЛ



- Разработка компьютерных программ для автоматической обработки текстов на ЕЯ —
лингвистических процессоров
- Лингвистический процессор:
(**Лингвистический парсер**)
Основа — формальная модель языка
Зависимость от конкретного ЕЯ
Пример: редактор *Word*, но не *NotePad*
- Сложность задач КЛ:
 - ЕЯ — сложная многоуровневая система знаков, возникшая для обмена информацией и постоянно изменяющаяся
 - Многообразие ЕЯ (способов выражения одного и того же смысла)
☺



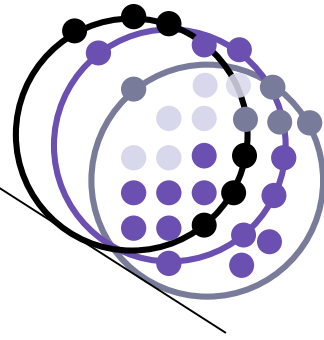
Лингвистический парсер (1)



- ПО для разбора **линейной** последовательности лексем (слов) языка исходного текста во **внутреннее представление** смысла данного П.
- Многоуровневый анализ П. на ЕЯ:
 1. Морфологический анализатор
in: морфологические словари + текст
out: POS, морфологические признаки



Лингвистический парсер (2)

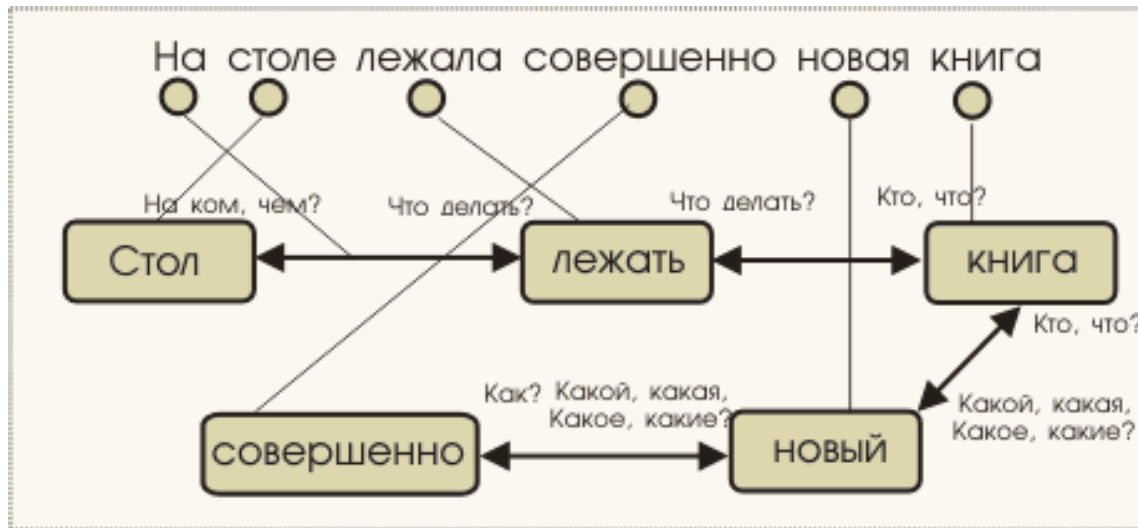


2. Синтаксический анализатор

out: дерево зависимостей:

узел: лексема + POS + грамматические хар-ки

дуга: отношение (подчинения)



Селезнев К. Технология клиент-сервер // «Открытые Системы», № 12, 2003 <http://linter.ru/ru/press-center/detail/27/1554/>

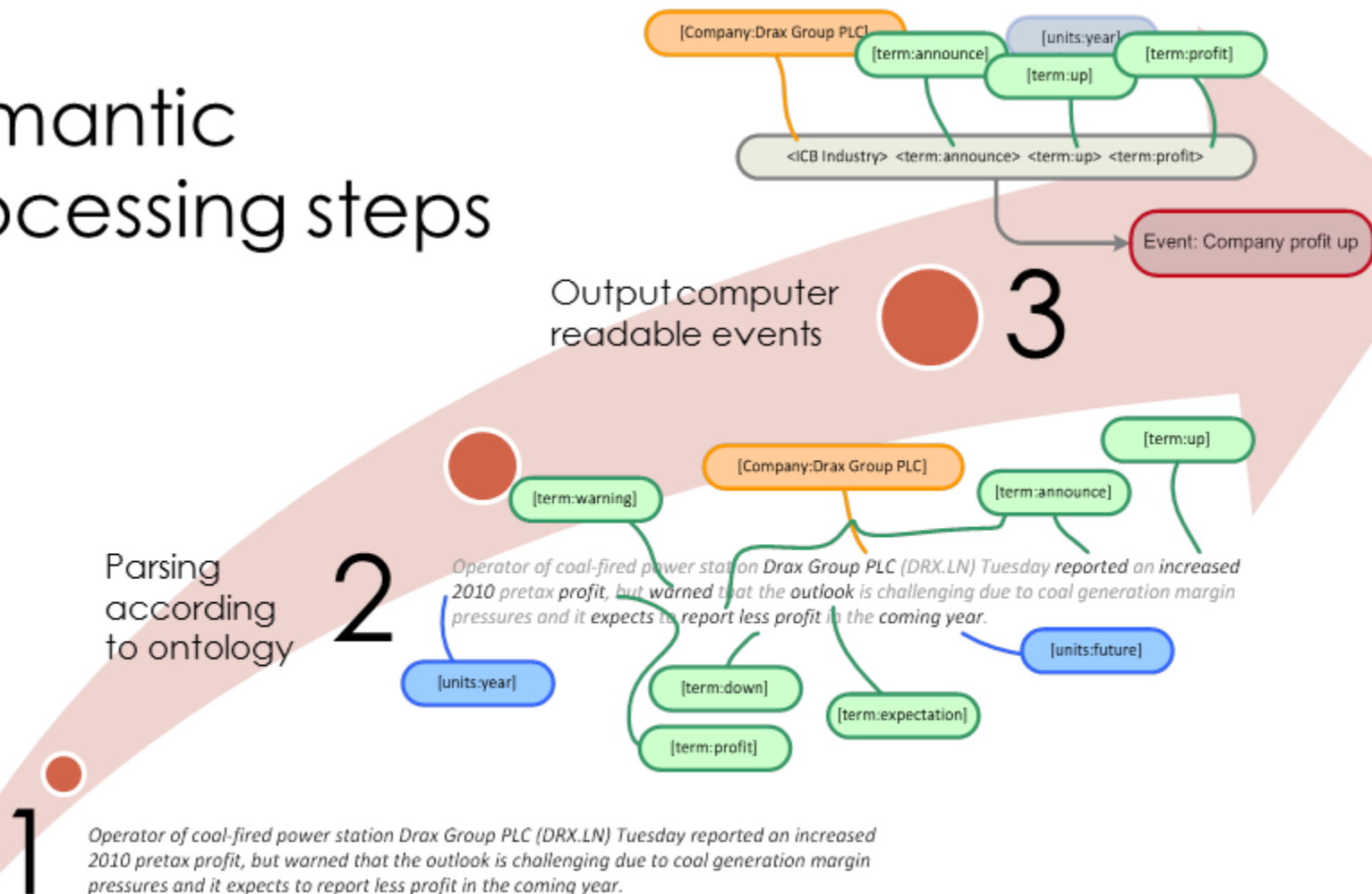
3. Семантический анализатор

3. Семантический анализатор —

in: онтология, предметный словарь, тезаурус

out: дерево зависимостей

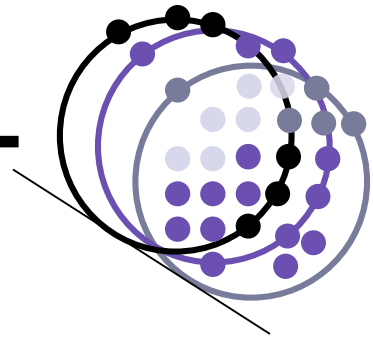
Semantic processing steps



Input text



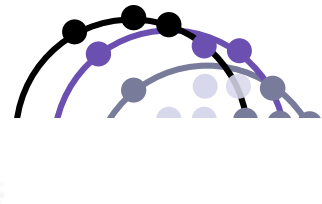
Задачи и направления CL



- Корпусная лингвистика
- Машинный перевод
- Синтез и распознавание речи
- Компьютерная лексикография
 - Создание электронных словарей, тезаурусов
- Компьютерный анализ документов:
 - Реферирование, классификация, поиск
- Computer-assisted language learning (CALL)

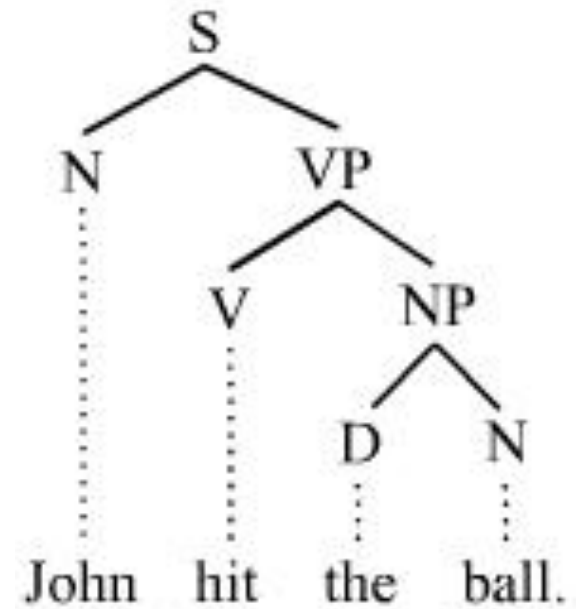


Ещё задачи CL & NLP



- Text normalization / segmentation
- Morphological analysis
- Part of speech tagging
- **Parsing** (parse tree)
- Coreference resolution
- Word-sense disambiguation (WSD)
- Named-entity recognition (NER)
- Relationship extraction
- Topic detection
- Summarization

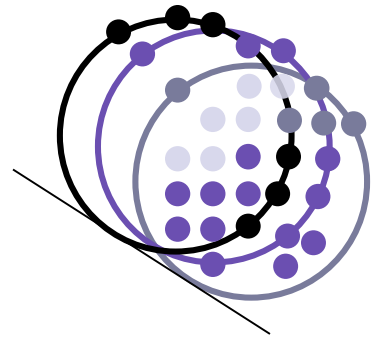
баня



планировать



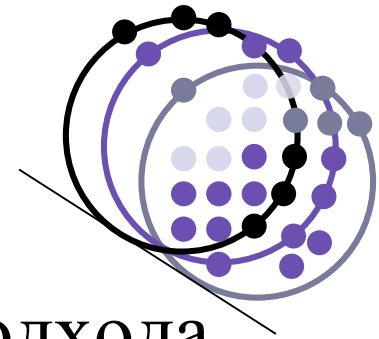
Заключение (1)



- Язык – ключевая особенность человека.
 - Язык используется для :
 - Обмена информацией
 - Убеждений, увещеваний
 - Обмана, дезинформации
- Управление
(язык –
инструмент
власти)



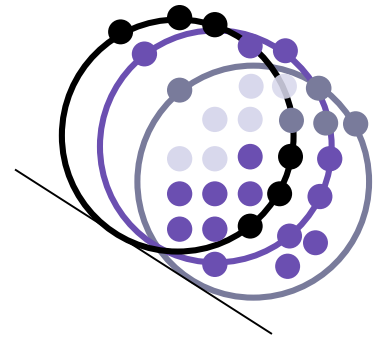
Заключение (2)



- Язык (как система) требует научного подхода для понимания:
 - как развиваются языки?
 - как человек изучает языки?
 - какова связь мышления и языка?
- Вычислительная лингвистика позволяет приблизиться к ответу на все эти вопросы



Домашняя работа

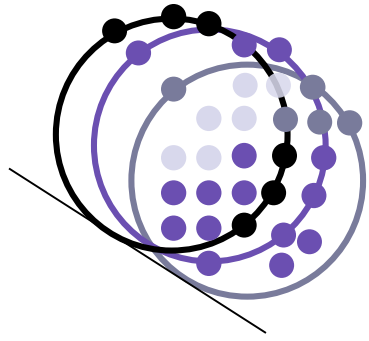


- Зарегистрироваться в Википедии
- Выбрать тему на семестр
 - https://ru.wikipedia.org/wiki/User:AKA_MBG/Todos
 - > сотни тем
 - кто первый, тот и выбрал

Ссылка на список тем есть в контакте в группе
«Интернет-математика в ПетрГУ»
в разделе «Ссылки»



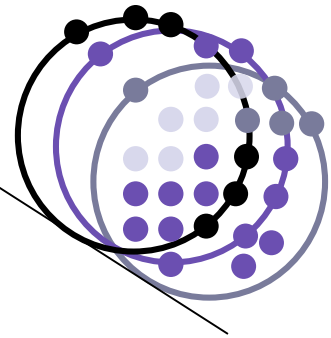
Выбор тем на семестр (1)



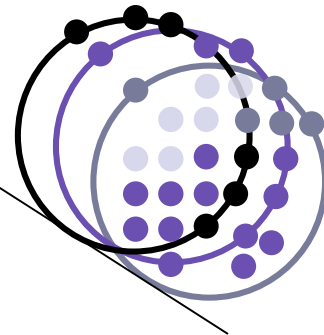
- **Bold** и plain темы
- Возможна командная работа, но...
- Посещение лекций



Выбор тем на семестр (2)

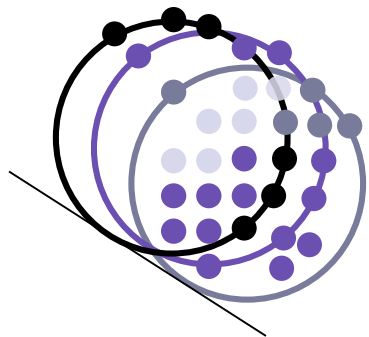


Выбор тем на семестр (2)



- Институт прикладных математических исследований КарНЦ РАН
- Карельский научный центр РАН
- Карельский научно-исследовательский институт

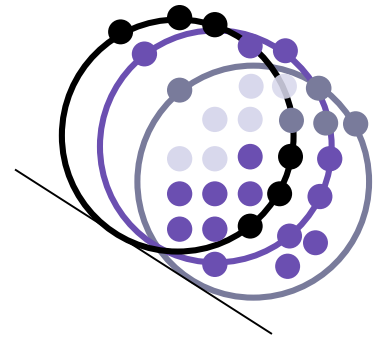
Вебометрика: Табличные данные (1 студент)



- 10 свойств, ~100 статей
- Проверить корреляцию 10 свойств
- Пишем вдвоём научную статью.



Литература



- **Лабораторный практикум по работе в вики-среде на примере Русской Википедии (для студентов и преподавателей): учеб. пособие** / Крижановский А. А. – Петрозаводск: ПетрГУ, 2014. – 106 с.: ил. (препринт) <http://scipeople.com/publication/114999/>
- **Большакова Е.В. Компьютерная лингвистика: методы, ресурсы, приложения.** // Летняя школа по КЛ. 2011
<http://www.myshared.ru/slide/94814/>
- **Гаршина В. В., Богоявленская Ю. А. Разработка лингвистического парсера русского языка.** // Вестник ВГУ, серия: Системный анализ и информационные технологии, 2012, No 2.
<http://www.vestnik.vsu.ru/pdf/analiz/2012/02/2012-02-29.pdf>
- **Селезнев К. Технология клиент-сервер** // «Открытые Системы», № 12, 2003 <http://linter.ru/ru/press-center/detail/27/1554/>

Спасибо за внимание!



http://vk.com/imath_petrSU
“Интернет-математика в ПетрГУ”