Компьютерная лингвистика

С.Ю.Толдова, Е. Еникеева, И. Макарчук

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

План

- Компьютерная лингвистика
  - Краткая история
  - Лингвистические задачи в приложениях. Примеры
- Задачи компьютерной лингвистики. Автоматический анализ текста в приложениях
  - информационный поиск vs. извлечение информации из текста
  - анализ данных vs. извлечение знаний из текста
  - лингвистические ресурсы
  - формализмы
- Этапы лингвистической обработки
- Свойства языка: сложности при моделировании языковых явлений
- Примеры лингвистических платформ

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

О курсе

- Примерный план курса:
  - Введение: задачи и этапы
  - Предобработка (сегментация, spell-checking, определение языка)
  - Морфология: конечные автоматы, конечные преобразователи, префиксные деревья для моделирования лингвистических данных
  - Морфология: алгоритмы и методы дизамбигуации
  - Синтаксис (контекстно-свободные грамматики: алгоритмы и реализация в NLTK, зависимостные грамматики, UD)

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

О курсе

- Оценивание:
  - домашние задания – 30%
  - проект – 25%
  - чтение литературы – 20 %
  - итоговый тест – 25%

  - Лекции: фокус на лингвистических особенностях данных в задачах автоматической обработки текста

План лекции сегодня

- Свойства языка: сложности при моделировании языковых явлений

- Лингвистика и компьютеры: краткая история сложных отношений
- От правил к большим данным - эволюция задач и подходов
- 4 топовых направления развития КЛ (Hirschberg Manning 2015)
- Deep Learning Tsunami (Manning 2015)
- Компьютерная лингвистика для науки: платформы и ресурсы

Некоторые неудобные свойства языка

- Неоднозначность:

  одно языковое выражение – разные смыслы

  ключ, партия, встретил Петрова vs. *Петрова сказала*

  Такие типы стали есть в цехе

  Teacher strikes idle kids

  (могут быть необходимы знания о мире для разрешения неоднозначности)

- Несимметричность: разные способы выразить некоторый смысл в разных языках (лексика, грамматика). Существуют обязательные для выражения смыслы, необязательные в другом языке

- Избыточность (вариативность): множественность способов выражения одного смысла (смысл – как инвариант синонимических преобразований).
- Конвенциональность: часто правильным и даже единственно возможным способом выражения некоторого смысла является лишь один из теоретически возможных (терминология, культурные ритуалы).
- Эллиптичность: в языке действует множество умолчаний. Понимание требует восстановления опущенной информации.
- Непрозрачность: язык активно использует сложные средства референции (указания на объекты в описываемом мире).

Язык как объект моделирования: сложности

- нет взаимно-однозначного соответствия между формой и значением (полисемия и синонимия на всех уровнях)
- многие языковые явления редки, не поддаются оцениванию при помощи стандартных статистических процедур; многие «события» остаются «не видны» в модели
- язык использует стратегии «по умолчанию»: не все смыслы имеют поверхностное выражение
- языковые вариации - множество языковых моделей (стили/жанры/страты/диалекты)

Общий взгляд на задачи компьютерной лингвистики

- Инженерная компьютерная лингвистика
  - междисциплинарная область, в задачи которой входит автоматический анализ текстов: автоматическая обработка ЕЯ
- Инструментальная компьютерная лингвистика
  - компьютерные технологии для обработки текстов, для представления лингвистических данных (корпуса, лингвистические ресурсы, парсеры).
- Теоретическая компьютерная лингвистика
  - (вычислительная лингвистика): применение математических (формальных) моделей к описанию естественного языка, моделирование функционирования языка с использованием формального аппарата.

Язык: сложности анализа

- Морфологическая неоднозначность:
  - *Адвокат Петрова Сидорова*
- Синтаксическая неоднозначность
  - *Не закапывайте погребенных в земле исполинов*
- Семантическая неоднозначность
  - *Петров болеет* vs. *Петров болеет за Динамо*
  - *Наша партия* vs. *Партия ракет*

Язык: Онтологическая неоднозначность

George Washington

Select an item from the list:

George Washington	US President
George Washington Universi	College/University
George Washington	Film
George Washington Bridge	Bridge
George Washington Tryon	Academic
George Washington	Organism
George Washington	TV Program
George Washington Carver	Scientist
George Washington Cable	Novelist
George Washington	Book

view more

Data Schema Apps Docs

George Washington

Date of birth: **Feb 22, 1732**

Place of birth: **Colonial Beach**

Religion: **Anglicanism, Episcopal Church in the United States of America, Church of England**

George Washington (February 22, 1732 [O.S. February 11, 1731] – December 14, 1799), was one of the Founding Fathers of the United States, serving as the commander-in-chief of the Continental Army during the American Revolutionary War and later as the...

US President, U.S. Congressperson, Military Commander

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Немного истории

- **Ранний период: машинный перевод**
- **1947** – Warren Weaver – идея статистического перевода
- **1954** – Джорджтаунский эксперимент – перевод
- по правилам (250 слов, 6 правил)
- **1958** – первая Всесоюзная конференция по МП
- **1966** – доклад ALPAC

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Немного истории

- **Тест Тьюринга**
- **1964-1966** – ELIZA, первые чатботы
- **1970е** – онтологии; Conceptual Dependency Theory (R. Schank)
- Развитие экспертных систем, систем, основанных на онтологическом моделировании ограниченной предметной области
- **конец 1980х-1990е** – внедрение статистических методов (распознавание речи, POS-tagging)

4 фактора развития

(Hirschberg Manning 2015)

- значительное увеличение вычислительных мощностей
- доступность лингвистических данных очень больших объемов
- развитие чрезвычайно успешных методов машинного обучения
- более глубокое понимание природы человеческого языка и его применения в различных социальных контекстах.

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Задачи КЛ

“мешок” задач

- Проверка правописания, грамматики и стиля.
- Распознавание текстов (печатный, рукописный).
- Распознавание (диктовка, слитная) и синтез речи.
- Машинный перевод текста и речи (классика NLP).
- Поиск нужного документа по запросу (в т.ч. в Интернете).
- Реферирование (смысловое сжатие).
- Классификация (кластеризация) текстов по содержанию, установление сходства текстов (плагиат и т.п.).
- Автофильтрация (определение нежелательных документов: спам и т.п.)
- Системы извлечения знаний (Text Mining, Information Retrieval), мнений (Opinion Mining,Sentiment Analysis)

--

- Симплификация текста
- DH (Digiutal Humanities)
- Генерация текста
- Вопросно-ответные системы, диалоговые системы, автоматические помощники

5 топовых направления развития

(Hirschberg Manning 2015)

- Machine translation
- Spoken dialogue systems and converstaional systems
- Machine reading
- Mining social media
- Analysis and generation of speakers state

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

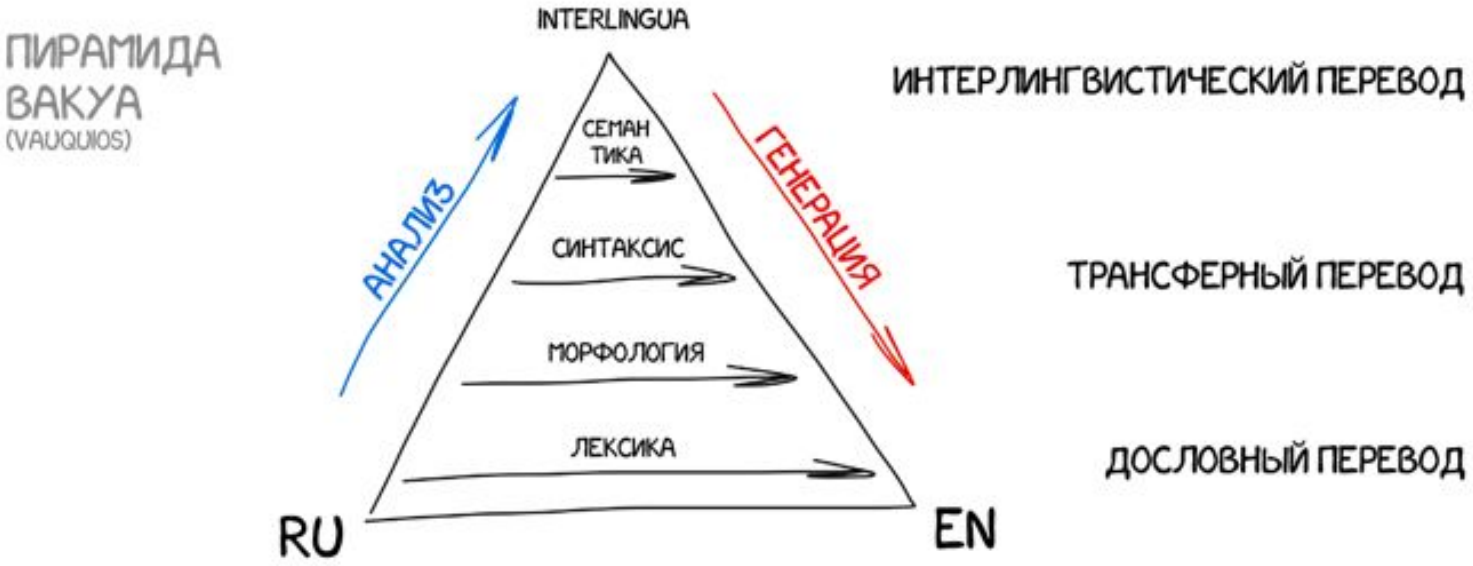
Машинный перевод

- классическая область искусственного интеллекта, начало КЛ

Машинный перевод - история развития

http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny_do_glubokogo_obucheniya.html

Машинный перевод на правилах



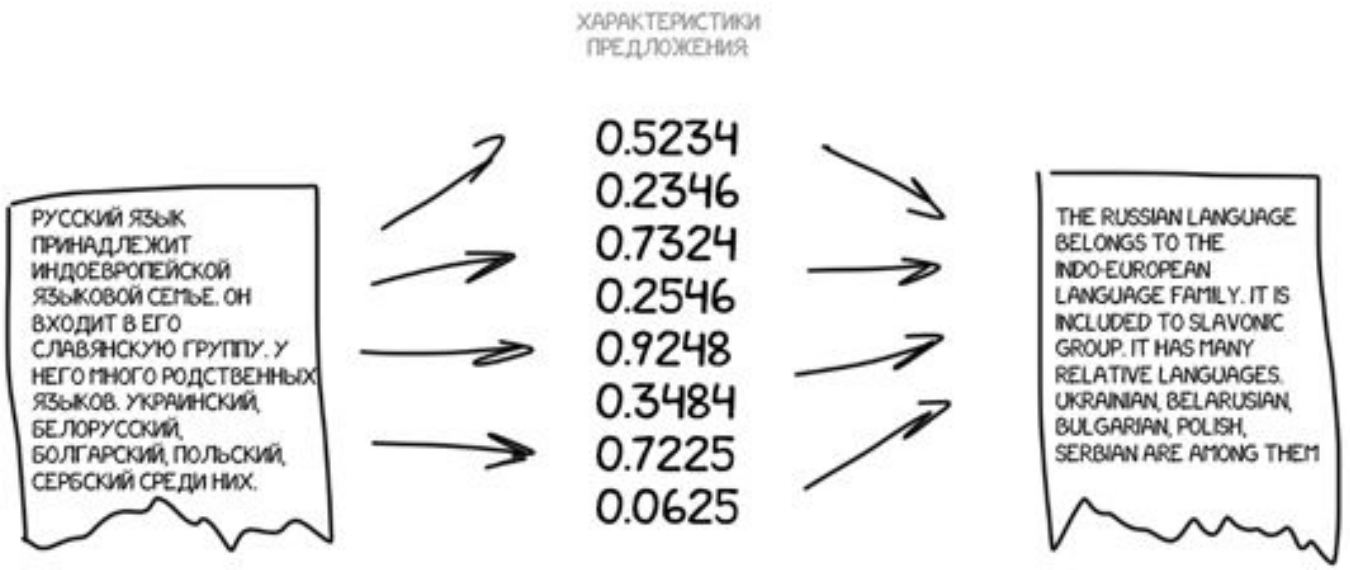
http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny_do_glubokogo_obucheniya.html

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Машинный перевод: статистический подход

http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny_do_glubokogo_obucheniya.html

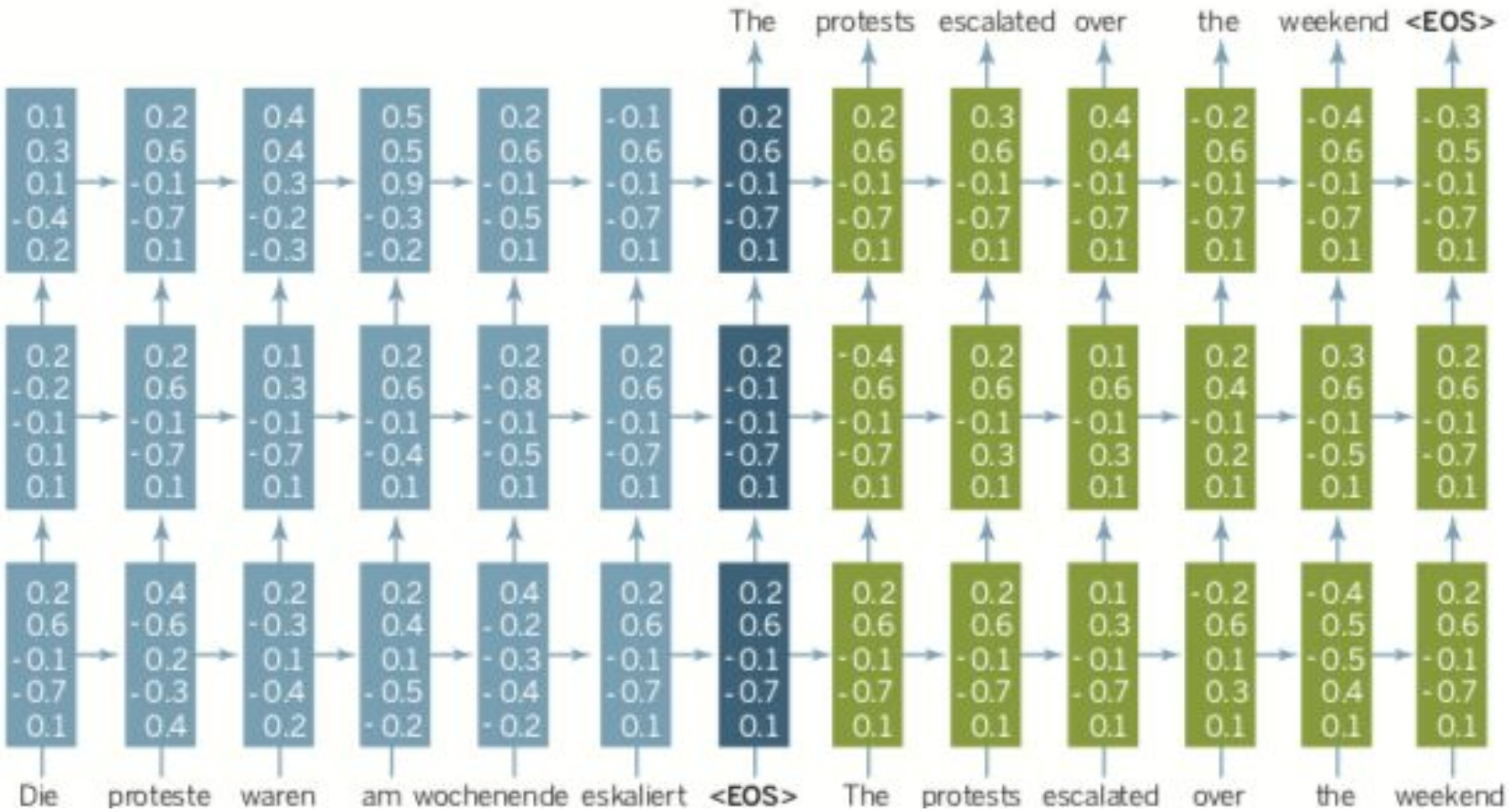
Машинный перевод: нейросети



http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny_do_glubokogo_obucheniya.html

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Машинный перевод: нейросети



Диалоговые агенты

Hirschberg Manning 2015

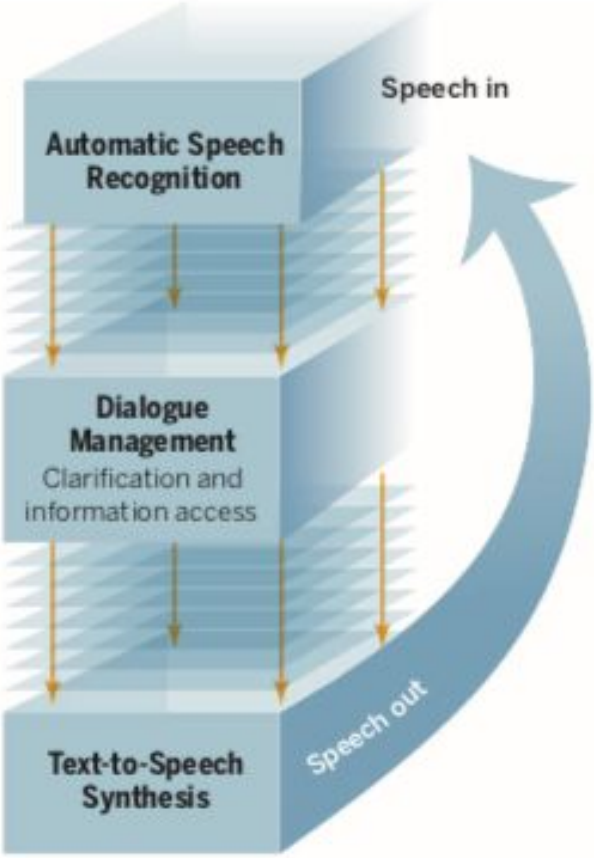
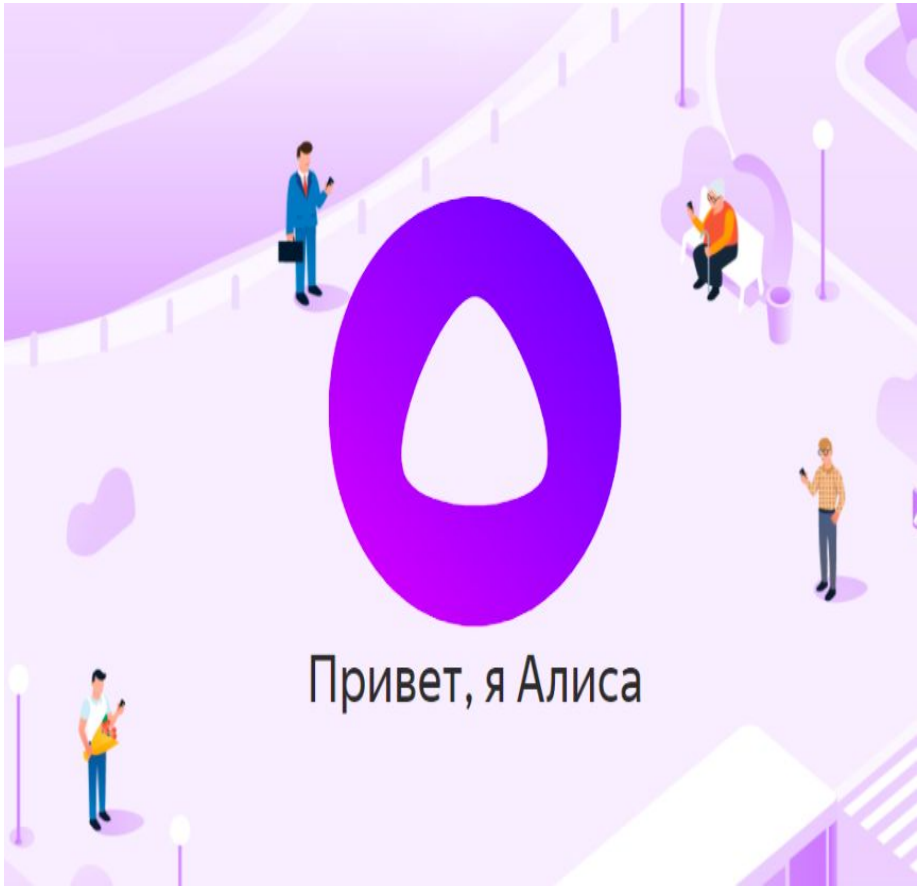


Fig. 3. A spoken dialogue system. The three main components are represented by rectangles; arrows denote the flow of information.



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Что пока плохо получается

- понимать принципы ведения диалога (дискурсивные стратегии, координацию, очередность)
- интегрировать невербальную информацию (а ее очень много)
- Супер гипер многозначность междометий и частиц (yeah, okay - agreement, topic shift, disagreement…)
- Недостаточность данных для обучения менеждменту диалога

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Machine reading

- знания хранятся в огромных неструктурированных текстовых массивах
- формализация с помощью автоматически создаваемых баз знаний
- или: использование баз знаний для формализации

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Основные задачи анализа контента

- **Обработка коллекций текстов:**
  - Группировка текстов / разделение текстов
- Задачи:
  - найти тексты, похожие по смыслу, стилю, тематике
- Задачи обработки текстов:
  - найти тексты, похожие на некоторый текст (например, на запрос пользователя) – информационный поиск;
  - собрать похожие тексты в одну группу - новостная агрегация, удаление дублей – кластеризация текстов
  - «рассортировать» тексты по группам – рубрикация текстов, классификация по стилям, распознавание спама

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка текстовых коллекций

Информационный поиск

Washington area unemployment rate hovered at 5.4 percent in August
Washington Post - 2 hours ago
The unemployment rate in the **Washington** area remained at 5.4 percent in August, even as jobs were added across a broad range of sectors, ...
Unemployment falls in nearly 90 pct. of US cities Bellingham Herald
all 163 news articles »

Questions and answers on presidential debates
Washington Post - 54 minutes ago
WASHINGTON — Tired of being deluged with TV commercials telling you that President Barack Obama or challenger Mitt Romney "approved ...
Obama and Romney face off: Kathleen Parker, The **Washington** Post NOLA.com
National, Florida, Virginia poll numbers tighten; Obama up 8 in Ohio
Washington Times (blog)

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка текстовых коллекций

Информационный поиск

Основная модель:

- Модель “информационного поиска”
- Основные допущения:
  - текст - «мешок» слов (bag of words)
  - Каждое слово появляется в тексте независимо от другого
- Текст – точка (вектор) в n-мерном пространстве
- Каждое измерение задается словом, которое есть в каком-нибудь тексте коллекции документов
- Близкие тексты имеют похожий набор слов
- Если текст – это вектор, то можно измерить расстояние между двумя текстами (подробнее в следующих презентациях)

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка коллекция документов

Основная модель:

- Модель “информационного поиска”
- Основные допущения:
  - текст – объект – признаки - «мешок» слов (bag of words)
  - каждое слово появляется в тексте независимо от другого
- Текст – точка (вектор) в n-мерном пространстве
- Каждое измерение задается словом, которое есть в каком-нибудь тексте коллекции документов
- Близкие тексты имеют похожий набор слов
- Если текст – это вектор, то можно измерить расстояние между двумя текстами

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка коллекция документов

- Основная модель:
- Модель “информационного поиска”
- Основные допущения:
- текст - «мешок» слов (bag of words)
- каждое слово появляется в тексте независимо от другого -> вероятность увидеть слово X в тексте не зависит от вероятности увидеть слово Y
- *Ворон к ворону летит, ворон ворону кричит*

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка коллекция документов

Ворон к ворону летит, ворон ворону кричит

w_{11} w_{22} w_{13} w_{34} w_{15} w_{16} w_{57}

Объем текста: N – количество словоупотреблений в тексте –

Или $N = \sum_{i=1}^{L} fr(w_i) = 4+1+1+1 = 7$

L – количество разных (несовпадающих) слов в тексте (или объем словника)

Вероятность увидеть словоупотребление *Ворон* (первое словоупотребление в тексте) 1/7

$P(w_{ij}) = 1/7$ (i – номер слова в «словаре», j – порядковый номер словупотребления в тексте

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка коллекция документов

- *Ворон к ворону летит, ворон ворону кричит*
- w_{11} w_{22} w_{13} w_{34} w_{15} w_{16} w_{57}
- Вероятность увидеть слово *ворон* (лексему)
- $P(w_{.i}) = Fr(w_{.i}) / N = 4/7$

- **Дисклеймер** к модели «мешок» слов:
- событие ‘появление в тексте слово *лететь’* не совсем независимо от события ‘появление в тексте слова *ворон*’
-

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка коллекция документов

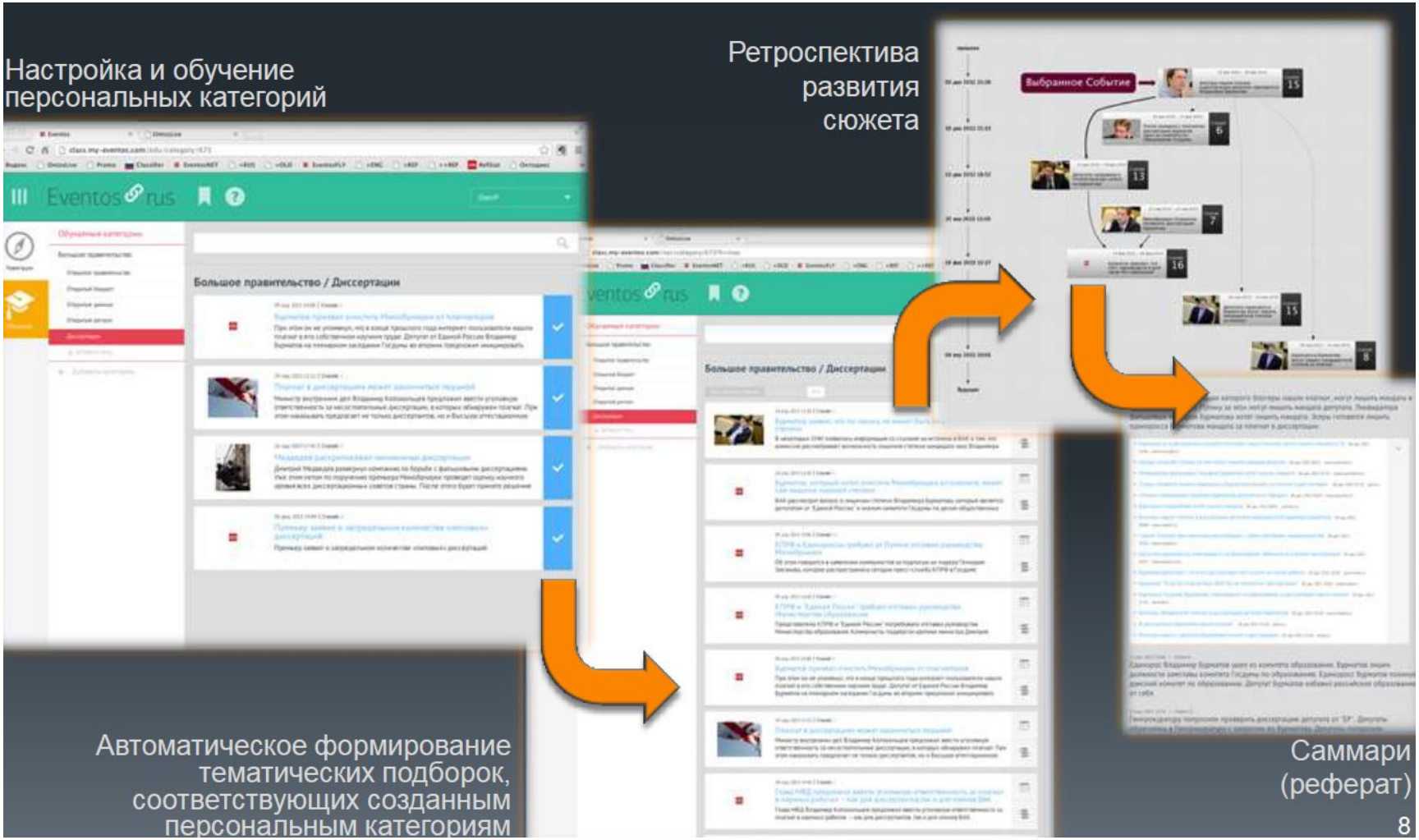
- Содержание каждого документа или тематической коллекции документов (*семантический портрет*) может быть единообразно описано вектором семантически значимых компонентов в пространстве семантических признаков. Компоненты:
  - текстовые n-граммы,
  - объекты,
  - теги и др. метаданные.
- Вектора можно сравнивать по “похожести”, т.е. сравнивая контекст документов, можно вычислять расстояния и степень их близости, формируя кластеры близких документов .
- Сравнивая документы (кластеры документов) разнесенные во времени можно объединять их в сюжет и показывать ретроспективу его развития.

Косинусная мера близости векторов документов

Результат:
4 млн. документов
40 млн. словарь
400 тыс. сюжетов
75 тыс. ретроспектив
8 тыс. больших историй

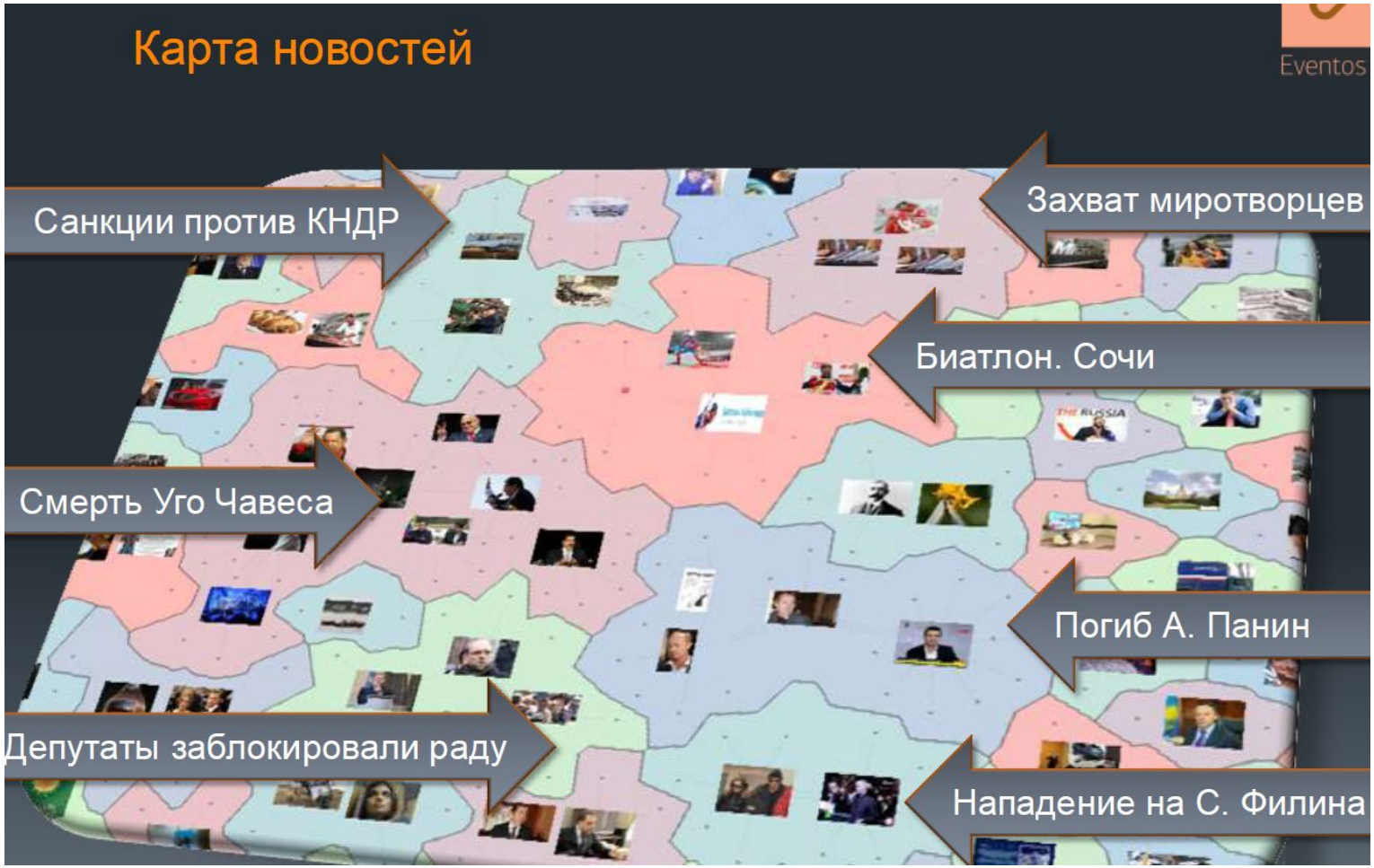
ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка коллекция документов



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Обработка коллекция документов

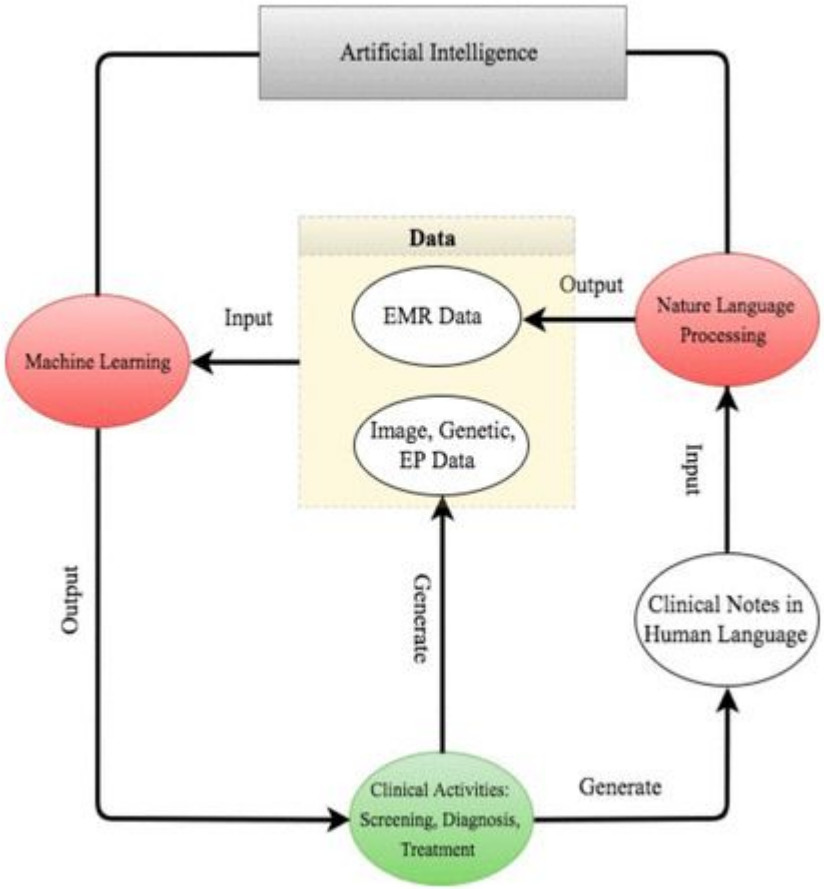


Основные задачи анализа контента

- **Анализ текста:**
  - Извлечение/выделение фрагментов текста; извлечение онтологических элементов (элементов знаний); преобразование неструктурированных данных в структурированные
- Задачи извлечения информации из текста (Information extraction):
  - Извлечение событий, их участников, места, времени, последовательности событий, отношений (Named Entities Recognition (Instances Extraction), relation extraction, fact extraction)
  - Извлечение онтологических знаний (knowledge extraction)

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Структурированные данные могут быть обработаны



Структурированные данные встраиваются в другие процессы (и это AI)

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Извлечение информации из текста

- George Washington (February 22, 1732 – December 14, 1799), was one of the Founding Fathers of the United States, serving as the commander-in-chief of the Continental Army during the American Revolutionary War and later as the new republic's first President. He also presided over the convention that drafted the Constitution. Washington, D.C., the capital of the United States, is named for him, as is the State of Washington on the nation's Pacific Coast.

Person /people/person

edit	**Date of birth:**	Feb 22, 1732	
edit	**Place of birth:**	**location**	**contained by**
		Colonial Beach	Westmoreland County Virginia United States of America
edit	**Country of nationality:**	Kingdom of Great Britain, United States of America	
edit	**Gender:**	Male	
edit	**Profession:**	Surveyor, Engineering, Politician, Farmer, Soldier, Military officer	
edit	**Religion:**	Episcopal Church in the United States of America, Anglicanism, Deism	

HIGHER SCHOOL OF ECONOMICS

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Извлечение информации из текста

- **George Washington** (**February 22, 1732** – **December 14, 1799**), was one of the Founding Fathers of the **United States**, serving as **the commander-in-chief** of the Continental Army during the American Revolutionary War and later as the new republic's first President. He also presided over the convention that drafted the Constitution. Washington, D.C., the capital of the United States, is named for him, as is the State of Washington on the nation's Pacific Coast.

Person /people/person

edit	**Date of birth:**	Feb 22, 1732	
edit	**Place of birth:**	**location**	**contained by**
		Colonial Beach	Westmoreland County Virginia United States of America
edit	**Country of nationality:**	Kingdom of Great Britain, United States of America	
edit	**Gender:**	Male	
edit	**Profession:**	Surveyor, Engineering, Politician, Farmer, Soldier, Military officer	
edit	**Religion:**	Episcopal Church in the United States of America, Anglicanism, Deism	

http://www.freebase.com/

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Извлечение информации из текста

Objects

All

Person

- Обама Барак
- Патрушев Николай ✓
- Путин Владимир
- Шойгу Сергей
- Джордж Литтл
- Кэйтлин Хэйден
- Томас Донилон
- Юрий Ушаков

Географическое ...

- Вашингтон
- Ирландия
- Москва
- Оклахома
- Россия

Properties

Патрушев Николай

imageUrl	
givenName	Николай
additionalName	Платонович
lastName	Патрушев
статья Википедии	http://ru.wikipedia.org/wiki/Николай_Патрушев
birthDate	11.07.1951
birthPlace	Ленинград
EmployedBy	Совет Безопасности РФ
EmployedBy	ФСБ РФ
title	Герой Российской Федерации
title	генерал армии
title	доктор юридических наук

Text Политика

Развитие отношений между США и Россией, в том числе в экономической области, обсудили президент Барак Обама и секретарь Совета безопасности РФ Николай Патрушев. Встреча состоялась 22 мая в Белом доме и на ней, помимо прочего, затронули вопросы борьбы с терроризмом и ситуацию в Сирии.Как сообщила официальный представитель Совета национальной безопасности США Кэйтлин Хэйден, Обама заглянул на встречу Патрушева с помощником президента США по национальной безопасности Томасом Донилоном."Президент США подтвердил желание укреплять двусторонние отношения, в том числе американо-российские экономические связи. Они также говорили о важности углубления сотрудничества в борьбе с терроризмом и необходимости политического урегулирования в Сирии путем переговоров",

HIGHER SCHOOL OF ECONOMICS

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Извлечение информации из текста: онтология DBPedia

Тексты бывают разные

About: George Washington

An Entity of Type : Concept, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Property	Value
rdf:type	▪ skos:Concept
rdfs:label	▪ George Washington
owl:sameAs	▪ http://pl.dbpedia.org/resource/Kategoria:George_Washington ▪ http://de.dbpedia.org/resource/Kategorie:George_Washington ▪ http://fr.dbpedia.org/resource/Catégorie:George_Washington ▪ http://ko.dbpedia.org/resource/분류:조지_워싱턴 ▪ http://ru.dbpedia.org/resource/Категория:Джордж_Вашингтон
skos:broader	▪ category:Presidents_of_the_United_States ▪ category:Washington_family ▪ category:Wikipedia_categories_named_after_American_politicians
skos:prefLabel	▪ George Washington
http://www.w3.org/ns/prov#wasDerivedFrom	▪ http://en.wikipedia.org/wiki/Category:George_Washington?oldid=490387017
is dcterms:subject of	▪ dbpedia:Washington's_Birthday ▪ dbpedia:United_States_presidential_election,_1788–1789 ▪ dbpedia:1932_Washington_Bicentennial ▪ dbpedia:Conway_Cabal ▪ dbpedia:Potomac_Company ▪ dbpedia:List_of_places_named_for_George_Washington

HIGHER SCHOOL OF ECONOMICS

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Извлечение информации из текста

Тексты бывают разные

- Подписи под фотографиями:

http://www.flickr.com/photos/tags/animals/

✔ **Как должны быть устроены ярлыки, чтобы можно было найти фотографии по одной теме?**

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Социальные медиа

- большие данные, огромная заложенная ценность
- новый, нестандартный языковой материал (code switching, hate speech, языковая креативность)
- приватность личных данных - палка о двух концах

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Speakers states

- анализ тональности
- степень уверенности

часто объединяется с анализом социальных медиа

Pipeline
уровни анализа

Последовательность этапов анализа, соответствующая уровням языка:

- Графематический анализ текста: выделение слов, знаков препинания, цифр, и прочих текстовых единиц.
- Морфологический анализ: определение грамматических характеристик лексем.
- Синтаксический анализ: установление структуры предложения -- системы связей между словами.
- Семантический анализ: построение структуры, ассоциированной непосредственно с передаваемым значением - в границах языка
- Прагматический анализ: интерпретация семантической структуры в контексте модели текста и знаний о мире
- [В случае речи – просодический анализ].

Lingpipe: уровни анализа

Последовательность этапов анализа, соответствующая уровням языка:

- графическая нормализация текста
- графематический анализ текста:
- выделение токенов / единиц анализа,
- классификация токенов: обрабатываеме vs. необразбатываемые вспомогательных знаков: разделители – знаки препинания, числа, шаблонированные элементы и прочие текстовые единицы.

Pipeline: морфология

```
<gr type="ПГ" mw="4">
<ob> с помощью {с_помощью=ПРЕД}</ob>
  <gr type="П+С"mw="8">
      рангового {ранговый=П=мр,ед,рд}
      дисперсионного {дисперсионный=П=мр,ед,рд}
      анализа {анализ=С,мр,но=ед,рд}
  </gr>
  </gr>
(ANOVA)
Фридмана {Фридман=С,фам,мр,од=ед,рд}
```

Pipeline: морфология

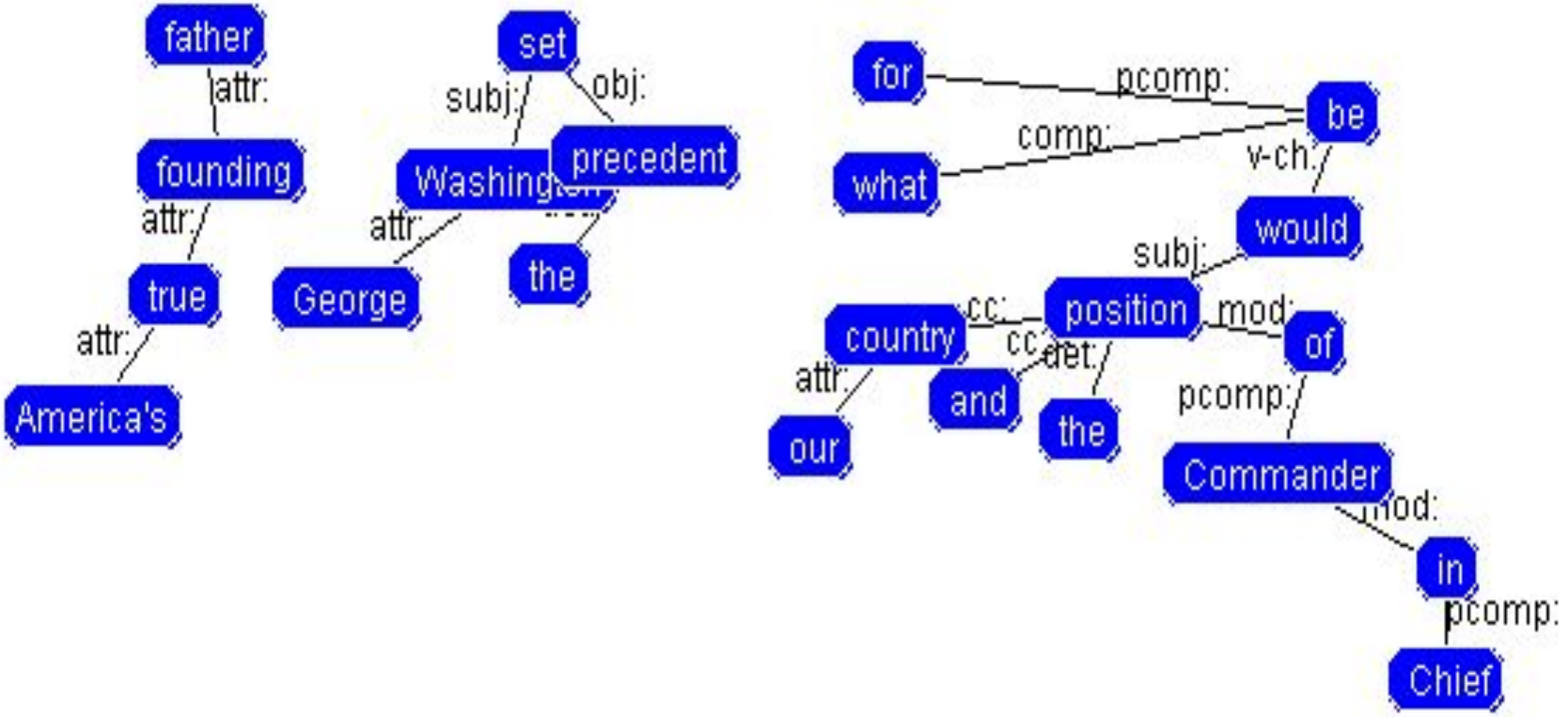
```
1    Då      då      ADV   AB                                    _
2    var     vara    VERB  VB.PRET.ACT
Tense=Past|Voice=Act
3    han     han     PRON  PN.UTR.SIN.DEF.NOM
Case=Nom|Definite=Def|Gender=Com|Number=Sing
4    elva    elva    NUM   RG.NOM   Case=Nom|NumType=Card
5    år      år      NOUN  NN.NEU.PLU.IND.NOM
Case=Nom|Definite=Ind|Gender=Neut|Number=Plur
6    .       .       PUNCT DL.MAD                                _
```

Lingpipe: уровни анализа

Последовательность этапов анализа, соответствующая уровням языка:

- Морфологический анализ:
  - Нормализация:
    - лемматизация
    - стемминг
  - Лексико-грамматическая аннотация:
    - определение грамматических характеристик лексем, pos-tagging
  - Дизамбигуация (pos-tagging)
  - Морфологический парсинг (например, автоматический стемминг для незнакомого языка)

NLP Pipeline: синтаксис



http://www.connexor.com/nlplib/?q=demo/syntax

NLP Pipeline: морфология + синтаксис

[america-s] ADJ POS @>N #1->4 >N
[true] ADJ POS @>N #2->4 >N
[founding] N S NOM @>N #3->4 >N
[father] N S NOM @SUBJ> #4->7 SUBJ>
[,] PU @PU #5->0 PU
[George=Washington] N S NOM @SUBJ> [George=Washington] N S NOM @SUBJ> #6->4 SUBJ> [George=Washington]
[set] V IMPF @FS-STA #7->0 FS-STA
[the] ART S/P @>N #8->9 >N
[precedent] N S NOM @7 <ACC
[for] PRP @ADVL> #10->22 ADVL>
[what] INDP S/P @P< #11->10 P<
[we] PERS GEN 1P @>N #12->13 >N
[country] N S NOM @SUBJ> #13->21 SUBJ>
[and] KC @CO #14->13 CO

NLP Pipeline: морфология + синтаксис CONLL-U

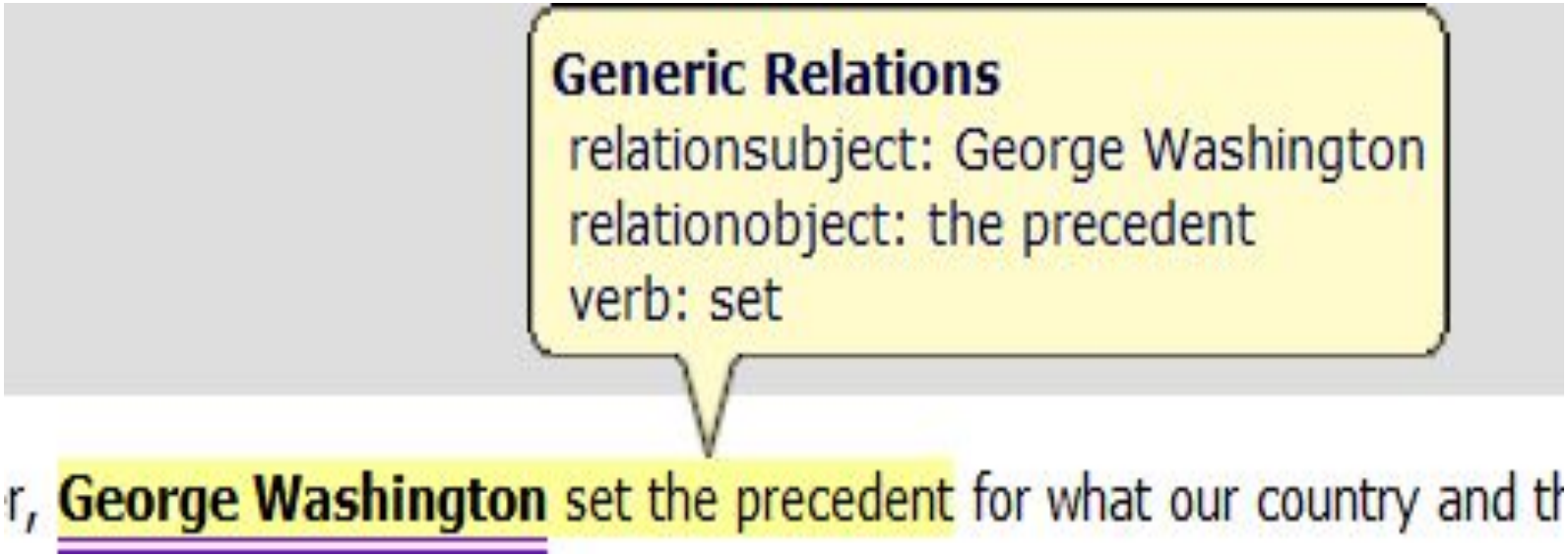
```
1 They they PRON PRP Case=Nom|Numb=Plur 2 nsubj
2:nsubj|4:nsubj
2 buy buy VERB VBP Numb=Plur|Pers=3|T=Pres 0 root 0:root
3 and and CONJ CC _ 4 cc 4:cc
4 sell sell VERB VBP Numb=Plur|Pers=3|T=Pres 2 conj
0:root|2:conj
5 books book NOUN NNS Numb=Plur 2 obj
2:obj|4:obj
6 . . PUNCT. _ 2 punct 2:punct
```

Lingpipe: уровни анализа

Последовательность этапов анализа, соответствующая уровням языка:

- Синтаксический анализ:
  - полный анализ - установление структуры предложения -- системы связей между словами;
  - Chunking
  - Shallow parsing

NLP pipeline: sematic role labeling



http://viewer.opencalais.com/

Компьютерная лингвистика 3: уровни анализа

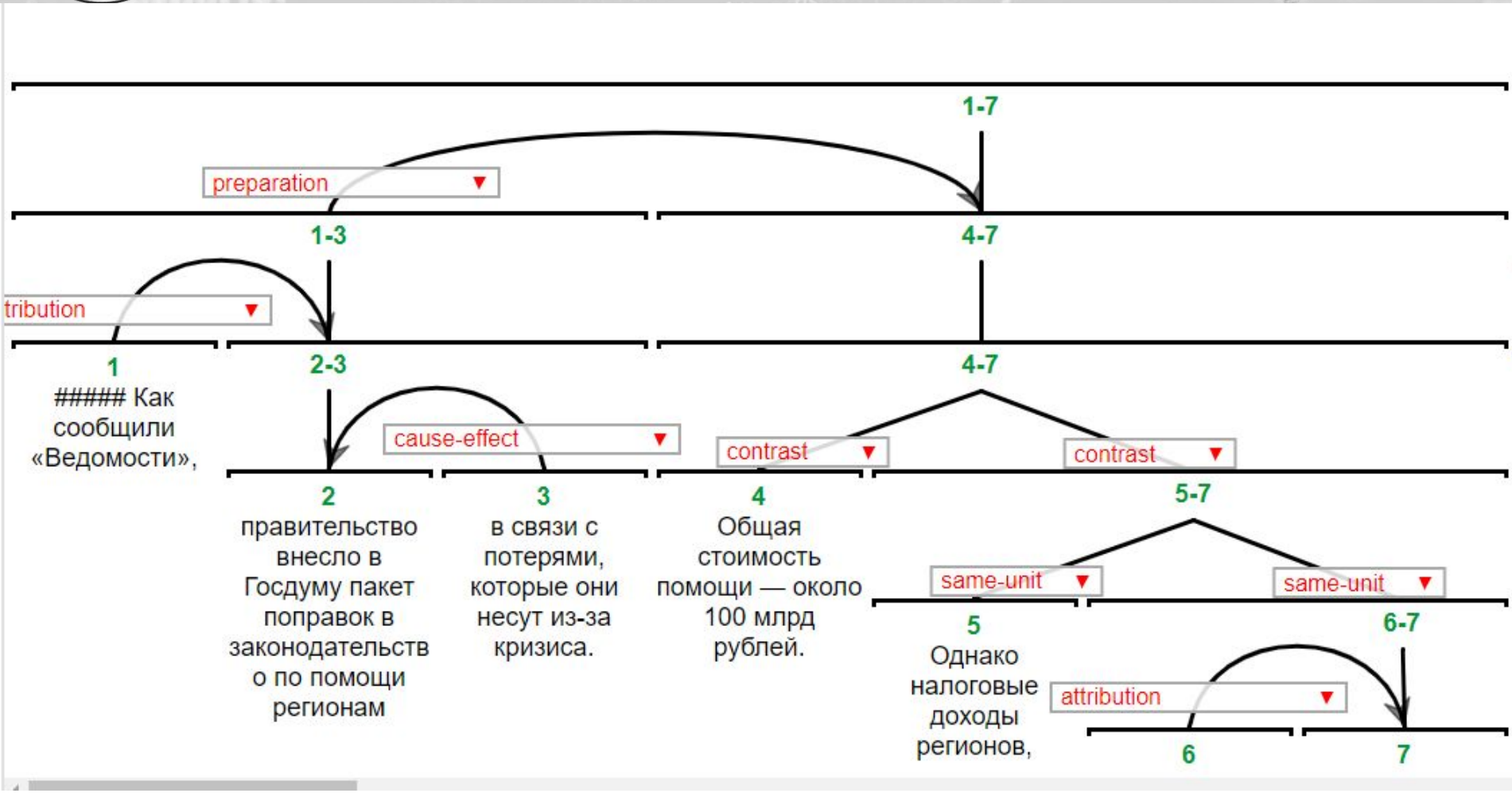
Последовательность этапов анализа, соответствующая уровням языка:

- Семантический анализ:
- ключевые слова, терминология
- коллокации,
- разрешение семантической неоднозначности,
- семантические сети,
- извлечение семантических ролей/моделей управления

NLP Pipeline: coreference

- **[America's true founding father]** $_i$, **[George Washington]**$_i$ set the precedent for what our country and the position of Commander in Chief would be. As **[a Virginia delegate at the Second Continental Congress]** **[he]**$_i$ was elected as **[Commander in Chief of the Continental Army]** and led **[his]**$_i$ ill-trained, under supplied troops to victory and thus, independence. In a letter to James Madison **[Washington]**$_i$ wrote "As the first of everything, in our situation will serve to establish a Precedent…it is devoutly wished on **[my]**$_i$ part, that these precedents may be fixed on true principles."

NLP Pipeline: coreference



Компьютерная лингвистика 3: уровни анализа

- Дискурсивный анализ:
- анафора-кореферентность,
- дискурсивная связанность
- дискурсивные анализ: структура дискурса
- извлечение именованных сущностей

2 подхода к моделированию

Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- require only small amount of training data
- development can be very time consuming
- some changes may be hard to accommodate

2 подхода к моделированию

Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- require large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus

Новые алгоритмы обработки «больших данных» (Big Data):

Нейро-сетевые модели

Среда для автоматической обработки текста

- LingPipe http://alias-i.com/lingpipe/,
- Gate http://gate.ac.uk/gate/doc/plugins.html,
- OpenNLP http://incubator.apache.org/opennlp/index.html,
- Alchemy http://www.alchemyapi.com/)

- Томита-парсер - https://tech.yandex.ru/tomita/
- NLTK – Natural Language Processing Toolkit
- UIMA - https://uima.apache.org/

Оценка

- Оценка систем и модулей:
  - Точность
  - Полнота
  - Accuracy
  - F-мера
  - Оценка разметки: мера согласия между аннотаторами
  - Др.

Общая методология

- Копрус: 3 подмножества
  - Learning set
  - Testing set
  - Gold standard set
  - (NB метод кроссвалидации)
- Разметка корпуса:
  - Инструкция
  - Мера согласия между аннотаторами
  - Золотой стандарт
  - Разметка обучающего множества / тестового множества

Общая методология

- Исследование
  - Лингвистически-значимые признаки
  - Статистические признаки
  - Формальны признаки
  - Вклад признака / признаков
- Оценка
  - «дифф» - места расхождения между ответом системы и золотым стандартом
  - Точность / полнота / другие метрики

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы

При разработке приложений иногда стоит воспользоваться существующими экспертными лингвистическими ресурсами:

- адаптация существующей системы под малоресурсные языки - переводные словари
- извлечение информации из текста - онтологии, лексикографические ресурсы (тезаурусы, банки моделей управления глаголов...)

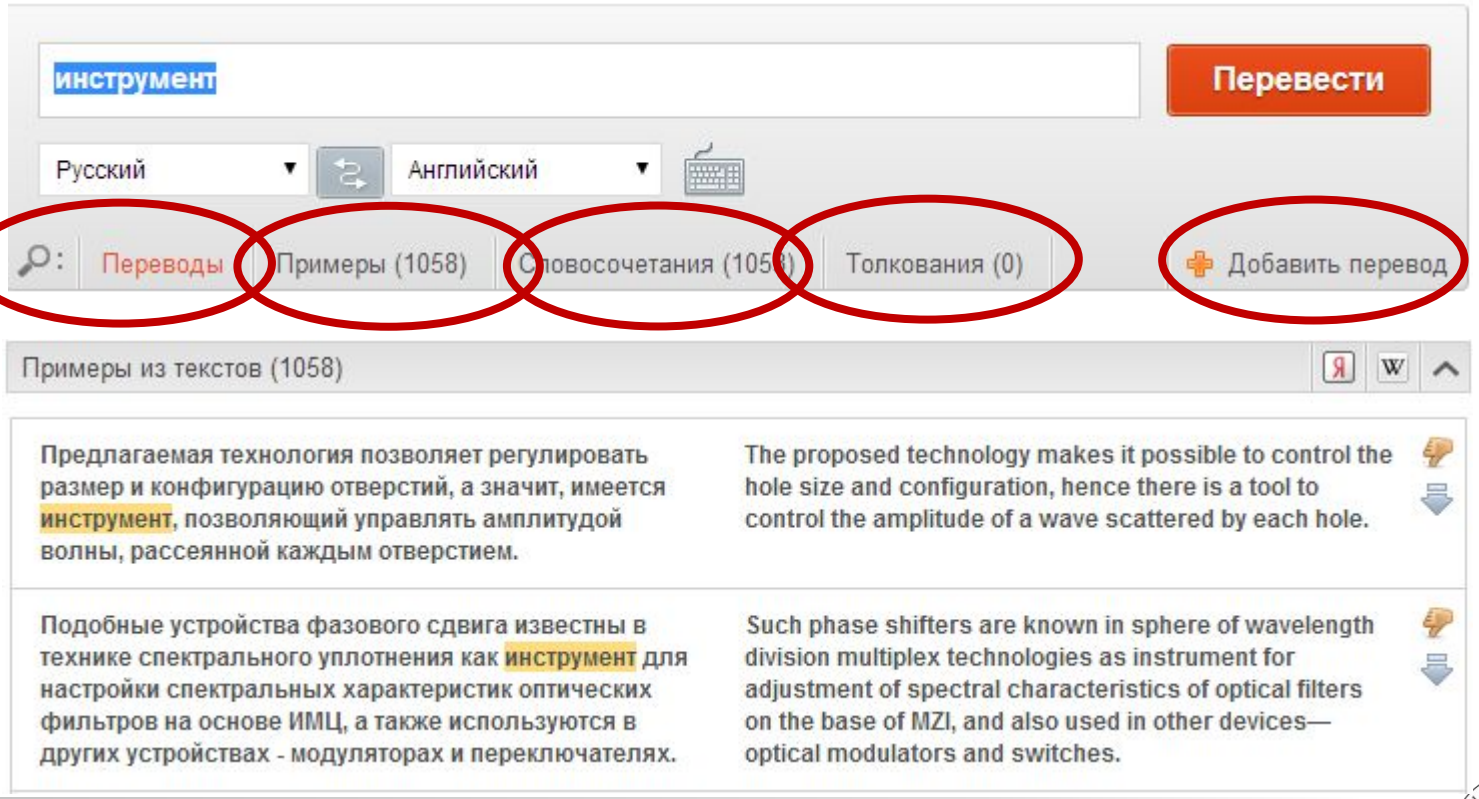
С другой стороны, есть задачи разработки языковых ресурсов, удобных для использования: для переводчиков, для изучения языков, для документации

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Компьютерные словари

ABBYY® Lingvo

http://www.lingvo-online.ru/ru/Translate/en-ru



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Компьютерные словари

Wiktionary

падеж	ед. ч.	мн. ч.
Им.	ча́т	ча́ты
Р.	ча́та	ча́тов
Д.	ча́ту	ча́там
В.	ча́т	ча́ты
Тв.	ча́том	ча́тами
Пр.	ча́те	ча́тах

Русский

Морфологические и синтаксические свойства

чат

Существительное, неодушевлённое, мужской род, 2-е склонение (тип склонения 1а по классификации А. А. Зализняка)

Корень: **-чат-**.

Произношение

МФА: [t͡ɕat]

омофоны: чад

Семантические свойства

Значение

1. *комп.* средство общения пользователей в компьютерной сети в режиме реального времени ◆ Отсутствует пример употребления (см. **рекомендации**).
2. *комп.* локация в компьютерной сети, где происходит такое общение ◆ Заходи вечерком к нам в **чат**.

Синонимы

1. —
2. болталка (*жарг.*)

Антонимы

Гиперонимы

Гипонимы

Родственные слова

Ближайшее родство

Этимология

Заимствование из англ

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Словарные агрегаторы

OneLook
Dictionary Search
Definitions
Related words

https://onelook.com/

Jump to: General, Art, Business, Computing, Medicine, Miscellaneous, Religion, Science,

We found 34 dictionaries with English definitions that include the word *bluebird*:
Click on the first link on a line below to go directly to a page where "bluebird" is defined.

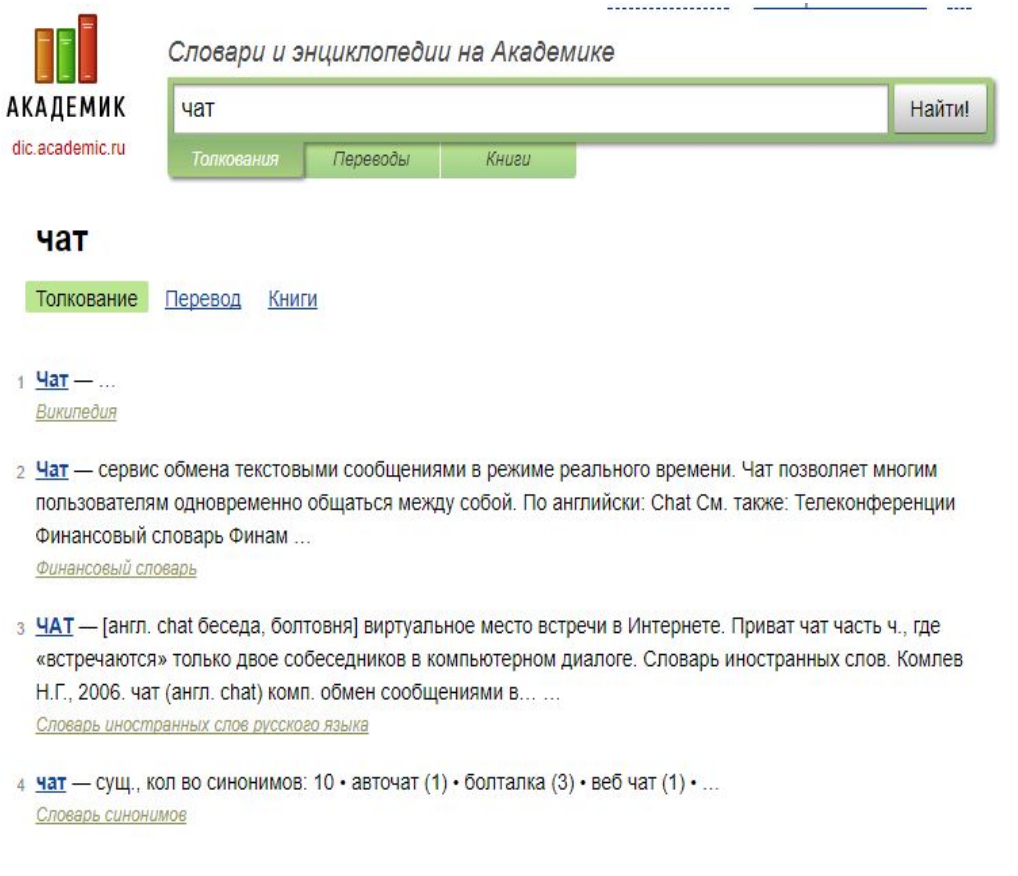
General (29 matching dictionaries)

1. bluebird: UltraLingua English Dictionary [home, info]
2. BLUEBIRD: Wikipedia, the Free Encyclopedia [home, info]
3. bluebird: Merriam-Webster.com [home, info]
4. bluebird: Oxford Dictionaries [home, info]
5. bluebird: American Heritage Dictionary of the English Language [home, info]
6. bluebird: Collins English Dictionary [home, info]
7. bluebird: Vocabulary.com [home, info]
8. bluebird: Macmillan Dictionary [home, info]

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Словарные агрегаторы

http://dic.academic.ru



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Лексикографические ресурсы Компьютерные словари

- Что важно для разработчиков?
  - структурирование словарной информации, зонирование
  - гибкие возможности поиска, поиск по разным зонам словаря
  - мультимедийность
  - мобильность контента
  - агрегация
  - мультиязычность
  - связь с другими существующими ресурсами
  - история редактирования
  - ….

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Лексикографические ресурсы
Компьютерные ресурсы

- частотные списки и словари

Грамматический словарь новых слов русского языка.
Е. А. Гришина, О. Н. Ляшевская

Новый частотный словарь русской лексики.
О. Н. Ляшевская, С. А. Шаров

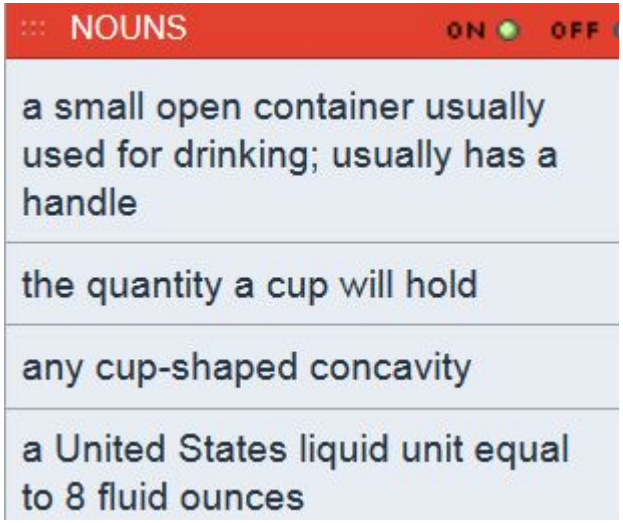
Словарь русской идиоматики. Сочетания слов со значением высокой степени.
Г. И. Кустова

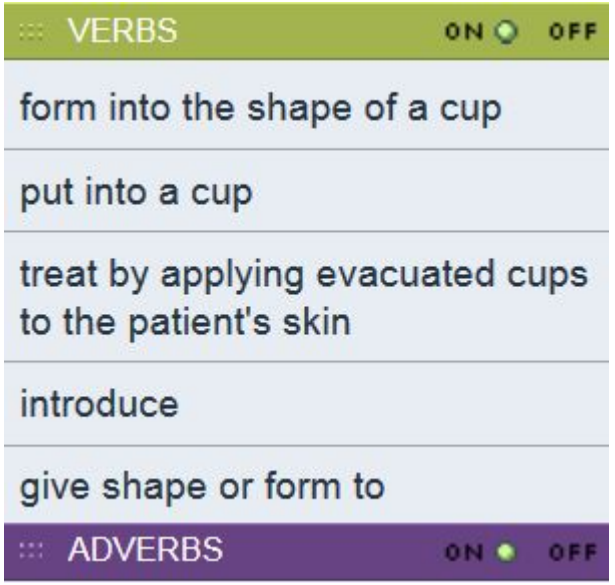
Словарь глагольной сочетаемости непредметных имен русского языка.
О. Л. Бирюк, В. Ю. Гусев, Е. Ю. Калинина

http://dict.ruslang.ru/

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Тезаурусы



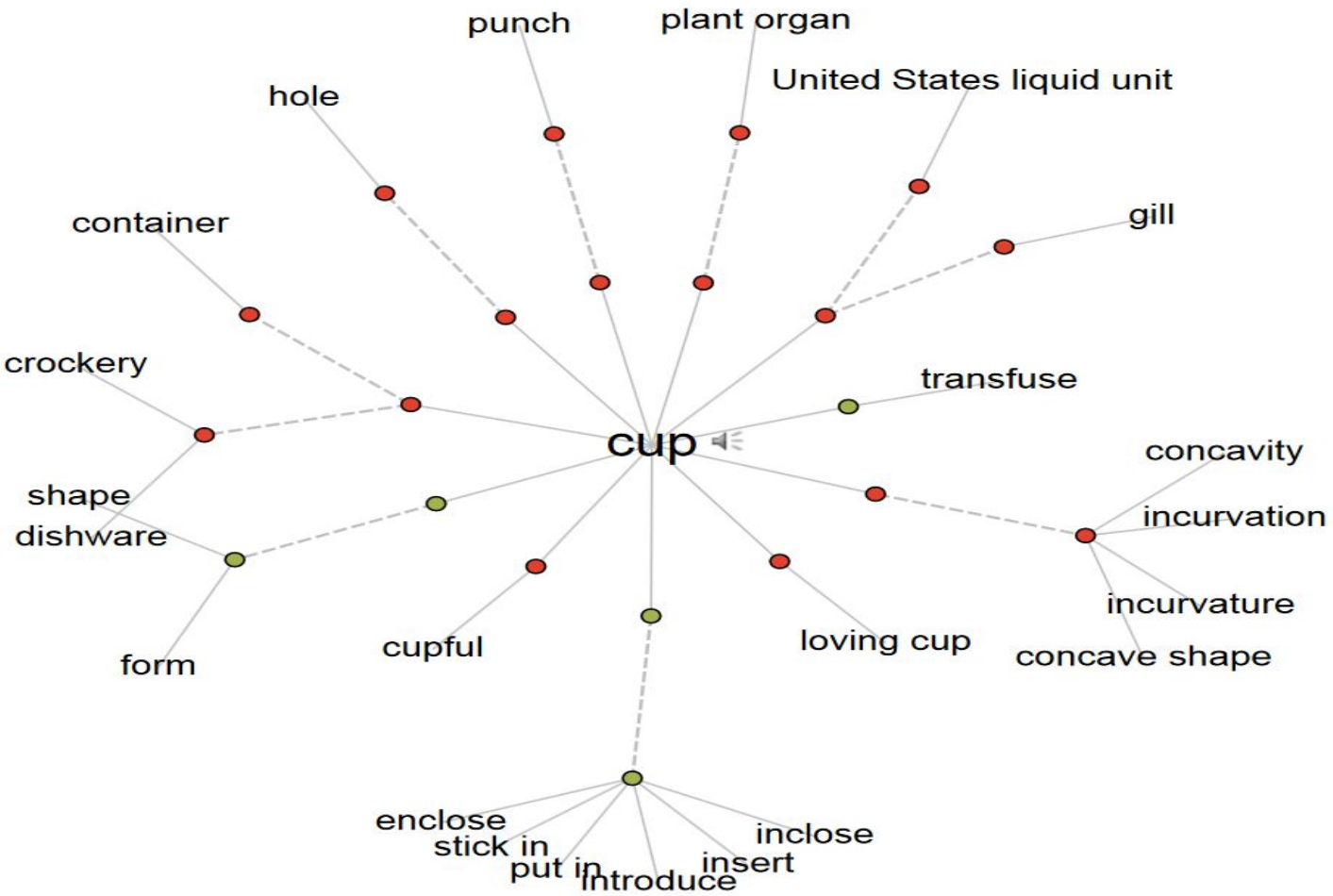


http://www.visualthesaurus.com/app/view

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

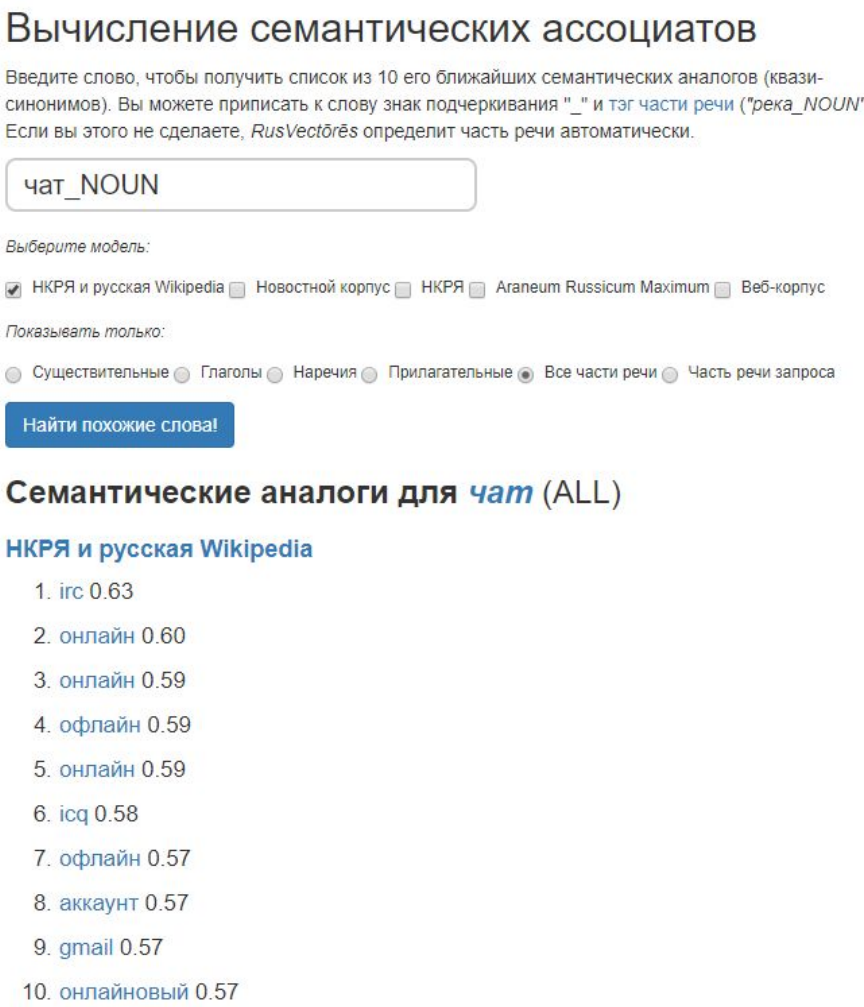
Ресурсы
Тезаурусы

http://www.visualthesaurus.com/app/view



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы: предобученные векторные модели



http://rusvectores.org/ru/similar/#

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Лексикографические ресурсы

- тезаурусы

Компьютерные ресурсы

ЧЕЛОВЕК

derived from

1 ЧЕЛО, ЧЕЛОВЕКОНЕНАВИСТНИК, ЧЕЛОВЕЧЕСКИЙ, ЧЕЛОВЕЧИЙ

Человек

Synset

1 ЛИЦО, ЛЮДИ, ОСОБА, ИНДИВИД, ПЕРСОНА, **ЧЕЛОВЕК** , ЛИЧНОСТЬ, МИКРОКОСМ, ИНДИВИДУУМ, ЧЕЛОВЕЧЕСКАЯ ЛИЧНОСТЬ [Понятие РуТез: ЧЕЛОВЕК]

hypernym

1 СУБЪЕКТ ДЕЯТЕЛЬНОСТИ [Понятие РуТез: СУБЪЕКТ ДЕЯТЕЛЬНОСТИ]

2 ЖИВОЕ, ОСОБЬ, ЖИВОЕ СУЩЕСТВО, ЖИВОЙ ОРГАНИЗМ, ЖИВНОСТЬ, ОРГАНИЗМ, СУЩЕСТВО, ИНДИВИДУУМ, МАКРООРГАНИЗМ, БИОЛОГИЧЕСКИЙ ОРГАНИЗМ [Понятие РуТез: ЖИВОЙ ОРГАНИЗМ]

hyponym

1 БОБЫЛЬ, БОБЫЛКА [Понятие РуТез: НЕ СОСТОЯЩИЙ В БРАКЕ]

2 ПОМОЩНИК, ПОМОЩНИЦА [Понятие РуТез: ПОМОЩНИК (ЧЕЛОВЕК)]

3 ОБВИНИТЕЛЬ [Понятие РуТез: ОБВИНИТЕЛЬ]

4 ПЕШЕХОД [Понятие РуТез: ПЕШЕХОД]

5 ЭКСПЕРИМЕНТАТОР, ЭКСПЕРИМЕНТАТОРША [Понятие РуТез: ЭКСПЕРИМЕНТАТОР]

http://ruwordnet.ru/en/

Тезаурус РуТез

HIGHER SCHOOL OF ECONOMICS

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Корпуса

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА

главная
архив новостей

Национальный корпус русского языка

http://ruscorpora.ru/

Ruscorpora

лингвистика

Показано: 1…10

Страницы: **1** 2 3 4 следующая страница

Найдено документов: 175, контекстов: 466.

1. **Михаил Арапов. Когда текст обретает смысл // "Знание-сила", №1", 2003** **[омонимия снята]** **Все контексты (1)**

   Наконе́ц, она́ же, вку́пе с не́сколькими други́ми, вы́звала "структурали́стский" перевоpо́т в **лингви́стике** и да́лее тако́й же перевоpо́т едва́ ли не во всех социа́льных нау́ках. [Михаил Арапов. Когда текст обретает смысл // "Знание-сила", №1", 2003] [омонимия снята] ←…→

2. **Конференция по когнитивной науке (2003)** **[омонимия снята]** **Все контексты (1)**

   Счита́ется, что когнитивная нау́ка, появле́ние кото́рой отно́сят к середи́не XX столе́тия, существу́ет и развива́ется на сты́ке ме́жду таки́ми областя́ми зна́ния, как психоло́гия, **лингви́стика**, нейронаука, компью́терная нау́ка и иску́сственный интелле́кт, когнитивная антрополо́гия и филосо́фия созна́ния. [Конференция по когнитивной науке (2003)] [омонимия снята] ←…→

Done

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Корпуса

Slovak Academy of Sciences
Ľ. Štúr Institute of Linguistics

Aranea Project Mirror NoSketch Engine Site (Guest Access)

Free registration is required for work with the ***Maius*** and ***Maximum*** class corpora.
To register, please fill in and submit **this form**.

Language	Aranea Corpora	Minus 120 M	Maius 1,20 G	Maximum
Arabic (not tagged yet)	Araneum Arabicum	ℹ 🔍	ℹ 🔍 *	
Bulgarian	Araneum Bulgaricum	ℹ 🔍	ℹ 🔍	
Chinese (simplified script)	Araneum Sinicum	ℹ 🔍	ℹ 🔍	
Czech	Araneum Bohemicum	ℹ 🔍	ℹ 🔍	5,17 G ℹ 🔍
Dutch	Araneum Nederlandicum	ℹ 🔍	ℹ 🔍	
English	Araneum Anglicum	ℹ 🔍	ℹ 🔍	11,4 G ℹ 🔍
English (*African TLDs*)	Araneum Anglicum Africanum	ℹ 🔍	ℹ 🔍	
English (*Asian TLDs*)	Araneum Anglicum Asiaticum	ℹ 🔍	ℹ 🔍	
Finnish	Araneum Finnicum	ℹ 🔍	ℹ 🔍	
French	Araneum Francogallicum	ℹ 🔍	ℹ 🔍	
French (*African TLDs*)	Araneum Francogallicum Africanum	ℹ 🔍	ℹ 🔍 *	8,70 G ℹ 🔍
Georgian (not tagged yet)	Araneum Georgianum	ℹ 🔍		
German	Araneum Germanicum	ℹ 🔍	ℹ 🔍	
Hungarian	Araneum Hungaricum	ℹ 🔍	ℹ 🔍	
Italian	Araneum Italicum	ℹ 🔍	ℹ 🔍	
Polish	Araneum Polonicum	ℹ 🔍	ℹ 🔍	
Portuguese	Araneum Portugallicum	ℹ 🔍	ℹ 🔍	
Russian	Araneum Russicum	ℹ 🔍	ℹ 🔍	13,7 G ℹ 🔍
Russian (*Russia-only TLDs*)	Araneum Russicum Russicum	ℹ 🔍	ℹ 🔍	

http://aranea.juls.savba.sk/guest/index.html

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Корпуса

Slovak Academy of Sciences
Ľ. Štúr Institute of Linguistics

Aranea Project Mirror NoSketch Engine Site (Guest Access)

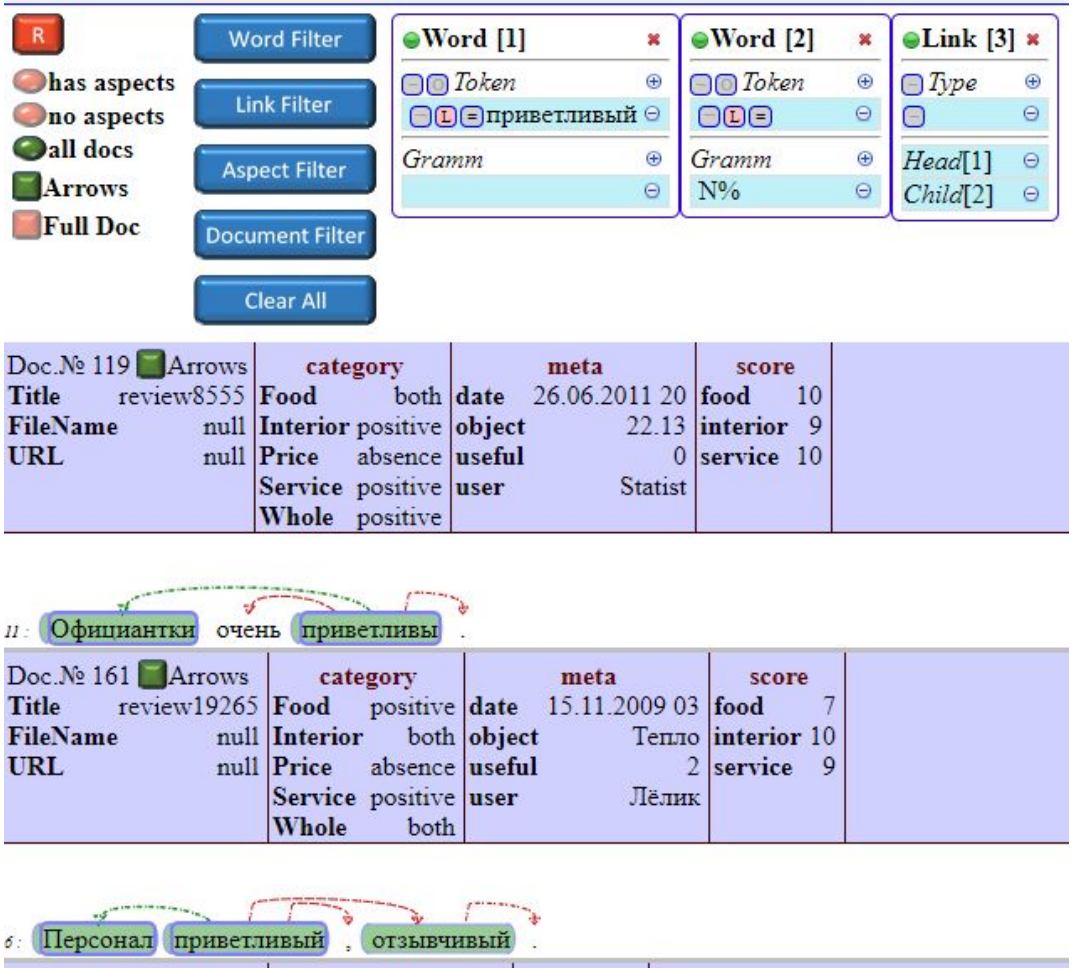
Free registration is required for work with the ***Maius*** and ***Maximum*** class corpora.
To register, please fill in and submit **this form**.

Language	Aranea Corpora	Minus 120 M	Maius 1,20 G	Maximum
Arabic (not tagged yet)	Araneum Arabicum		*	
Bulgarian	Araneum Bulgaricum			
Chinese (simplified script)	Araneum Sinicum			
Czech	Araneum Bohemicum			5,17 G
Dutch	Araneum Nederlandicum			
English	Araneum Anglicum			11,4 G
English (*African TLDs*)	Araneum Anglicum Africanum			
English (*Asian TLDs*)	Araneum Anglicum Asiaticum			
Finnish	Araneum Finnicum			
French	Araneum Francogallicum			
French (*African TLDs*)	Araneum Francogallicum Africanum		*	8,70 G
Georgian (not tagged yet)	Araneum Georgianum			
German	Araneum Germanicum			
Hungarian	Araneum Hungaricum			
Italian	Araneum Italicum			
Polish	Araneum Polonicum			
Portuguese	Araneum Portugallicum			
Russian	Araneum Russicum			13,7 G
Russian (*Russia-only TLDs*)	Araneum Russicum Russicum			

http://aranea.juls.savba.sk/guest/index.html

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

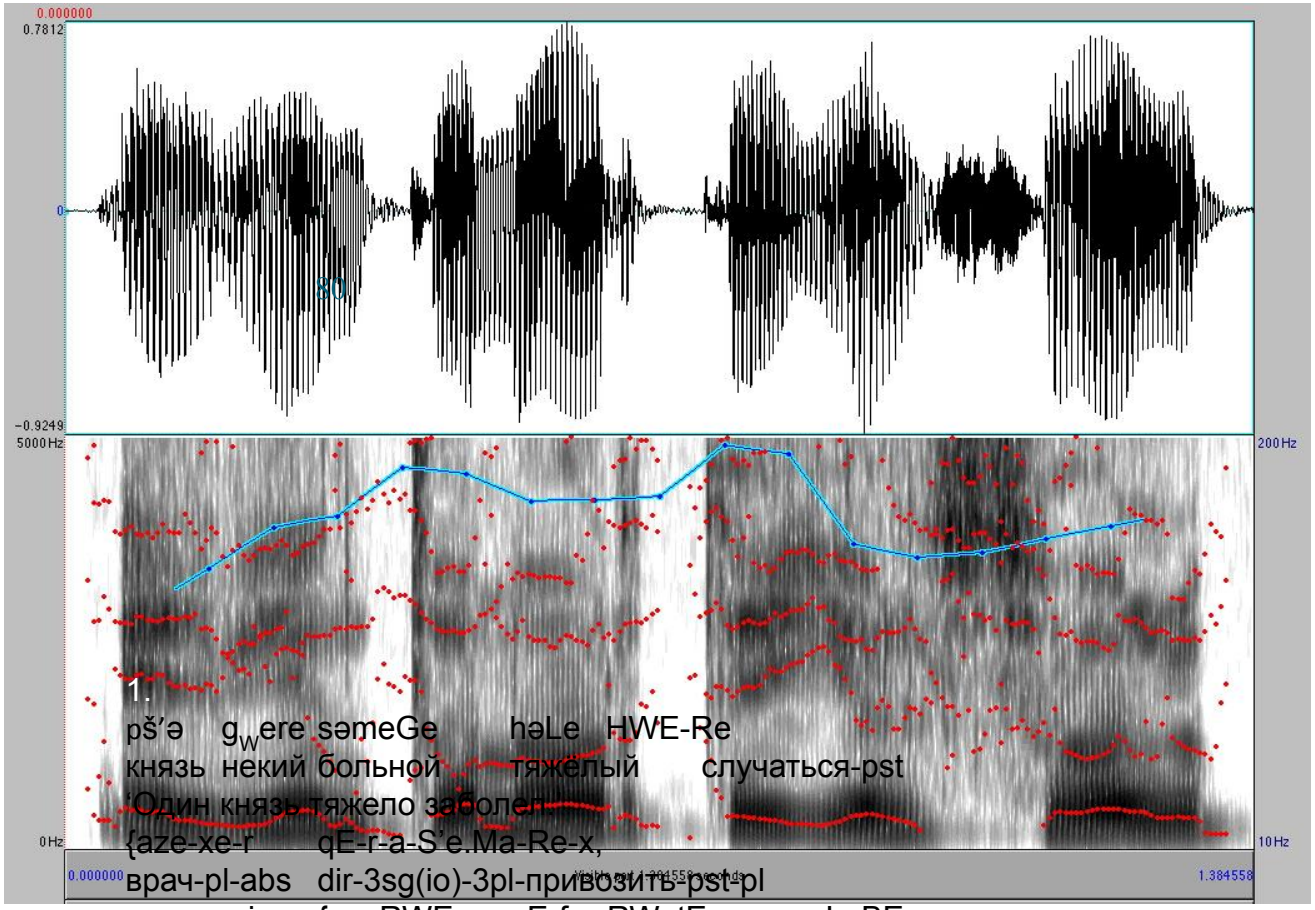
Ресурсы
Специализированные корпуса



http://senty.maimbava.net/res01/senty.php

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Мультимедийные корпуса



1.
pš’ə g$_W$ere səmeGe həLe HWE-Re
князь некий больной тяжёлый случаться-pst
‘Один князь тяжело заболел,
{aze-xe-r qE-r-a-S’e.Ma-Re-x,
врач-pl-abs dir-3sg(io)-3pl-привозить-pst-pl
aw z-jə {ezeRWE qE-f-a-RWetE-n a-LeBE-r-ep
но один-coh лекарство dir-ben-3pl-найти-pot 3pl-мочь-dyn-neg
‘Врачей к нему привозили, но они не смогли найти никакого лекарства.’

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы
Мультимедийные корпуса

- http://spokencorpora.ru/showelan.py

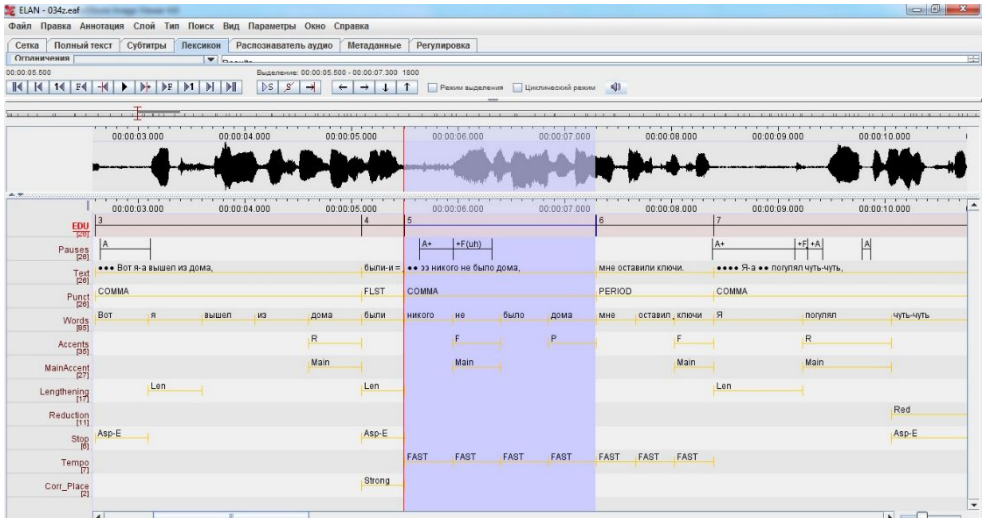
Комментарии
Скрыть

Вид транскрипции
Полная
Упрощённая
Минимальная

Чтобы выделить фрагмент, нажмите на первую и последнюю

Транскрипция

Время	№	ЭДЕ
0.00	1.	Мне /приснилось,
1.07	2.	···(0.96) как= || ··(0.17) ка'= || ··(0.28) как я^{w} ==
3.43	3.	···(0.73) как мне /мама ··(0.44) –ночью положила в /портфель \куклу,
7.53	4.	которую я /очень \хотела.
9.20	5.	···(0.78) \Вот,



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Ресурсы Корпуса

- что и зачем нужно искать пользователю в корпусе
- быстрая обработка больших объемов текстов
- визуализация информации
- разметка
- поиск по разным «причудливым» параметрам

Инструменты

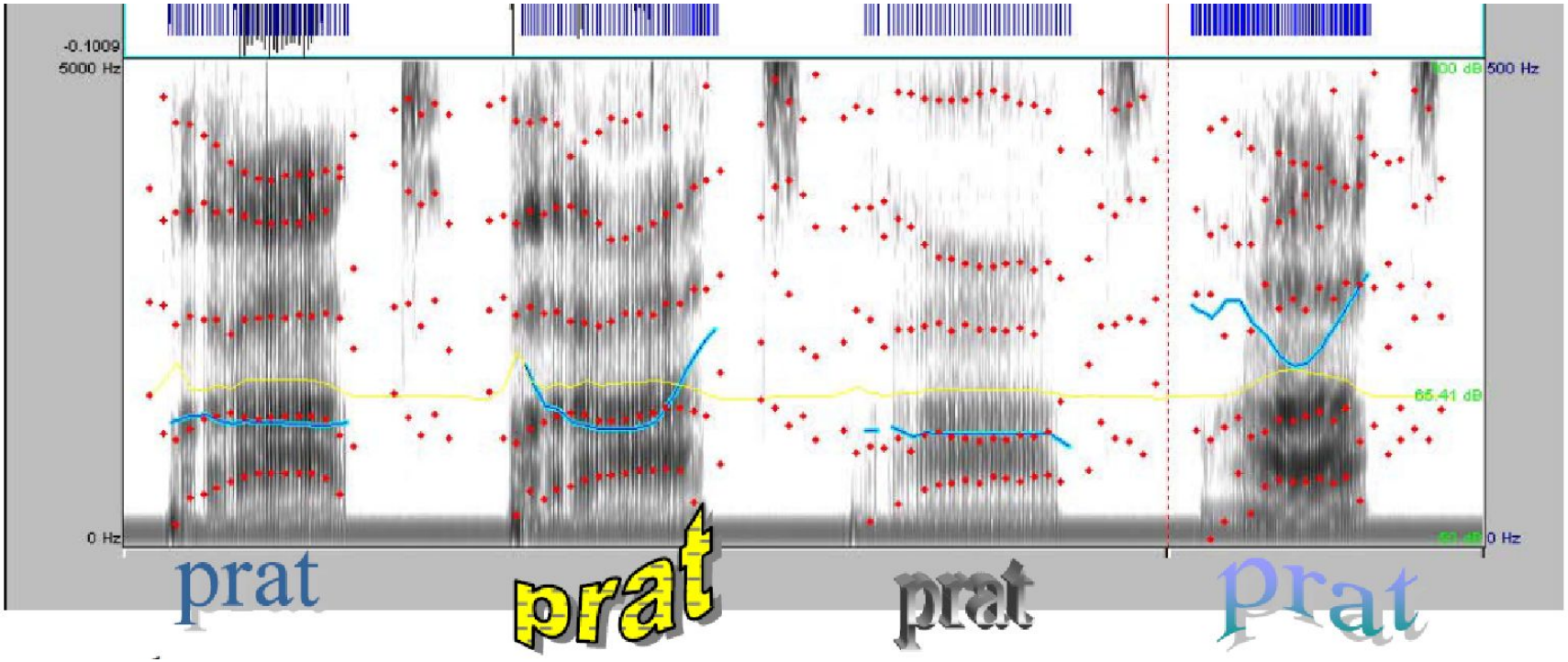
- специальные программы обработки звука и видеоряда (анализ текста как мультимодального явления):
  - программы разметки речевых корпусов
  - программы анализа звуков речи
  - SpeechAnalizer, Praat, Elan
- Специальные программы для работы с графикой (шрифты, дополнительные символы, распознавание символов и т.п.

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Инструменты

- PRAAT

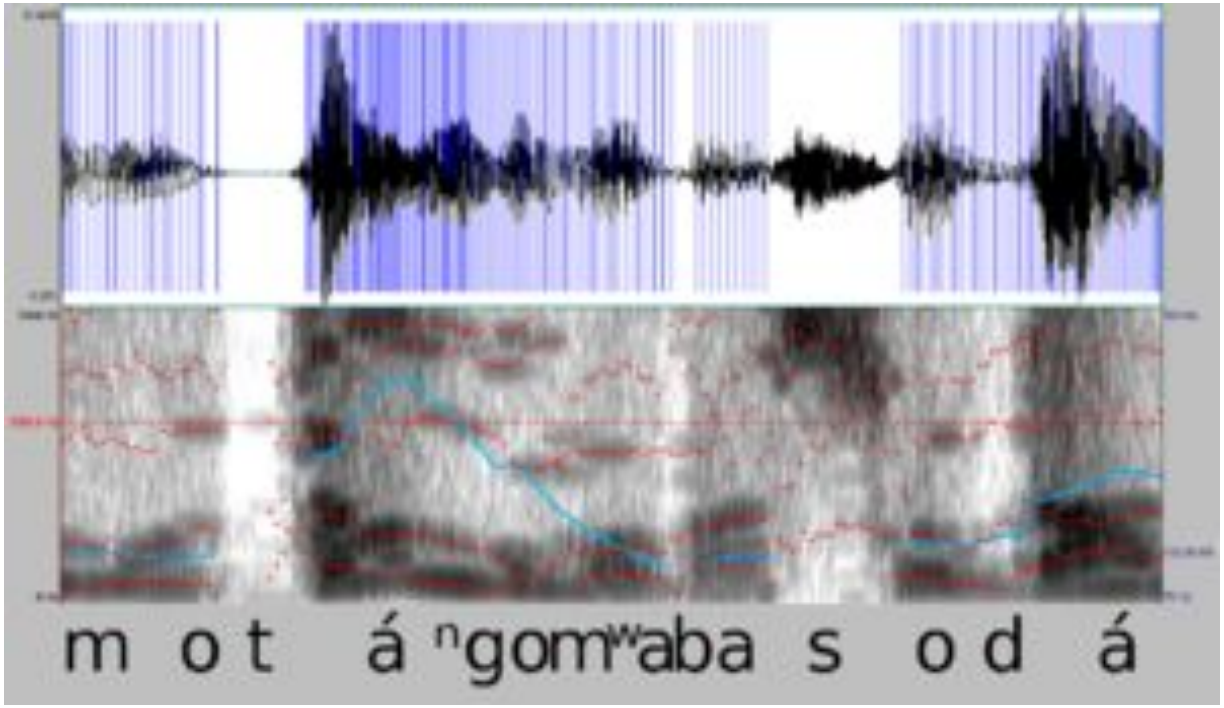
(http://web.stanford.edu/dept/linguistics/corpora/material/PRAAT_workshop_manual_v421.pdf)



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Инструменты

- PRAAT

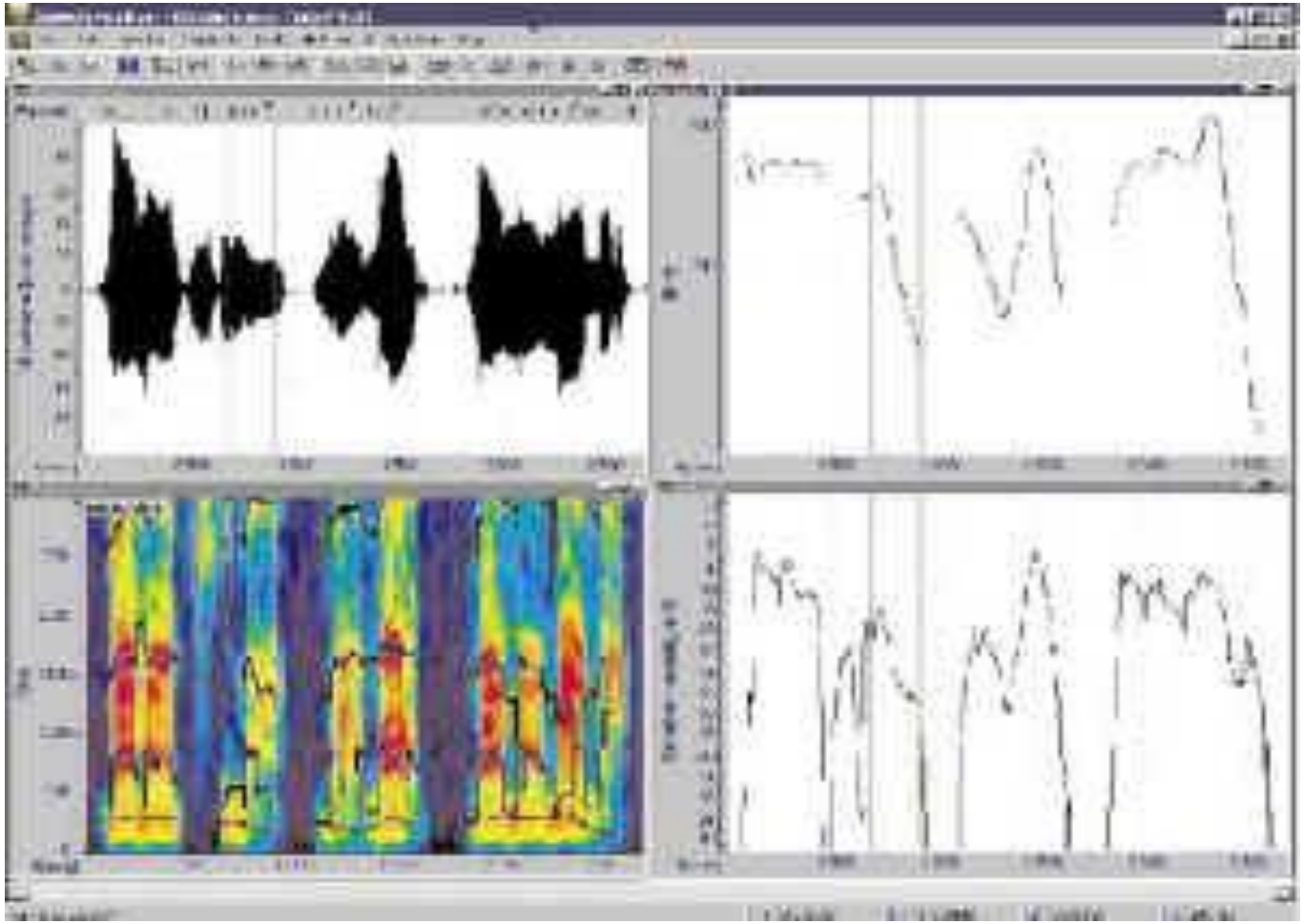


http://fieldworks.sil.org/

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Инструменты

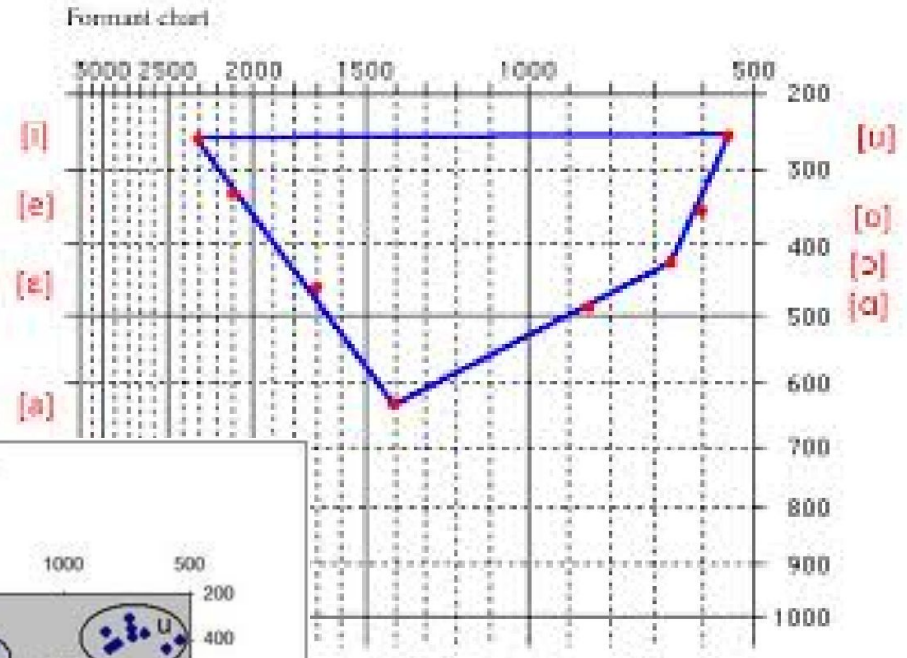
- PRAAT

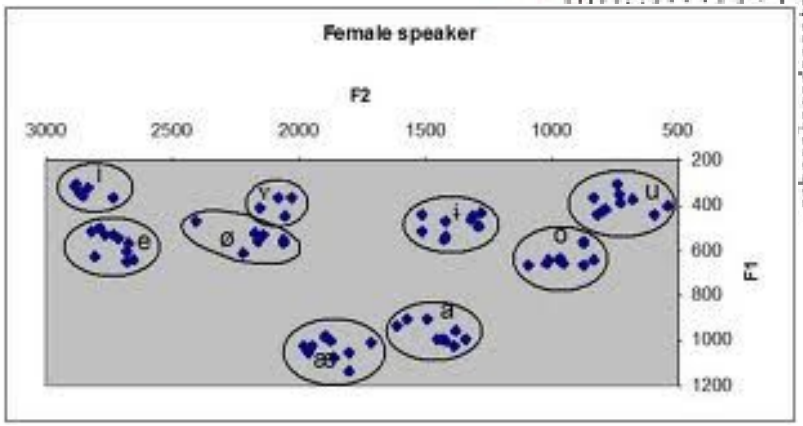


ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Инструменты

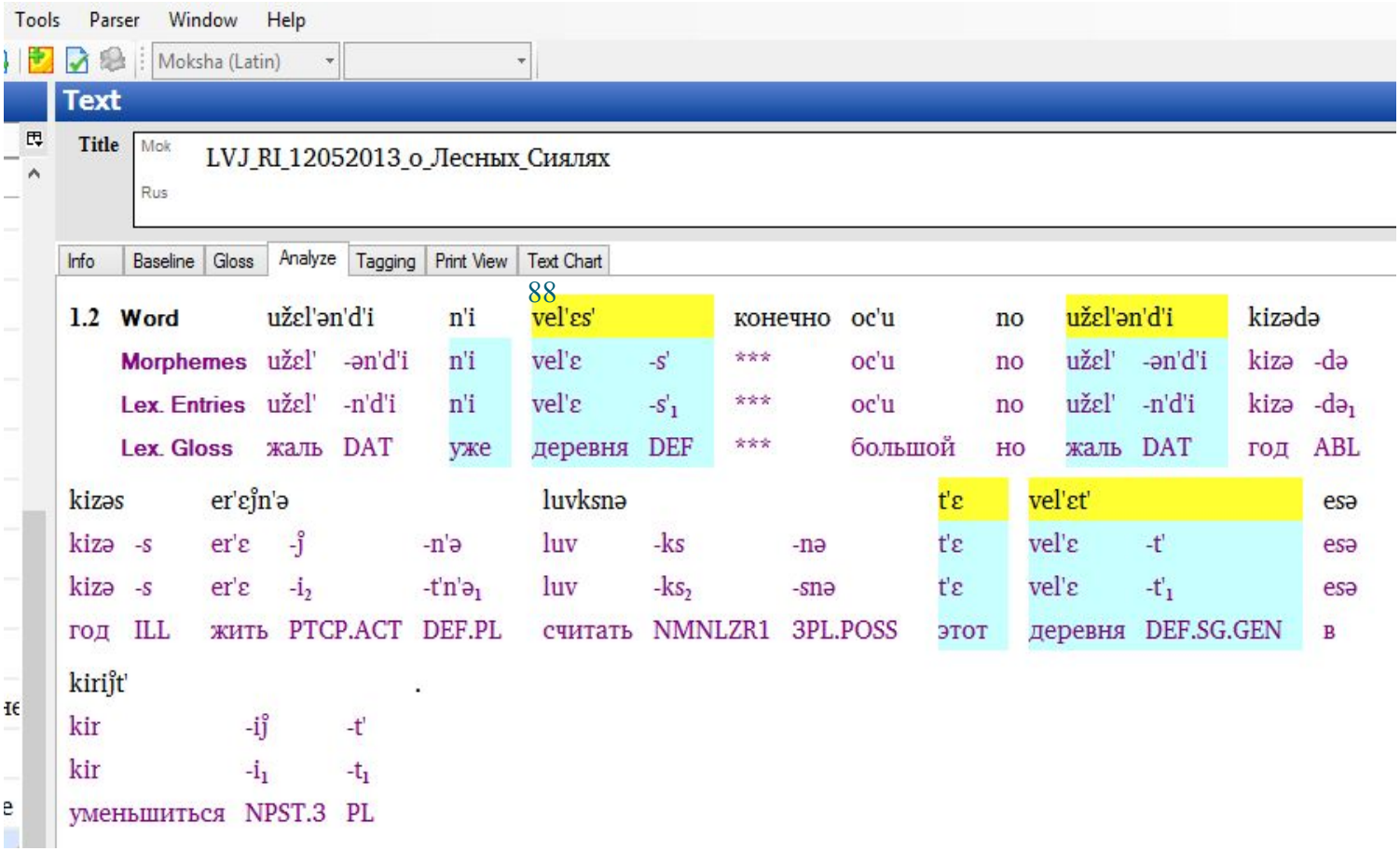
- PRAAT





ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Инструменты



http://fieldworks.sil.org/

Инструменты

Документация языка:

- Текст:
  - невозможно восстановить грамматическую информацию о языке, если есть только текст и его перевод

Словарь:

- подстрочные переводы одного и того же слова в тексте должны совпадать
- для каждого слова необходима информация о разных основах
- один и тот же грамматический показатель должен кодироваться одинаково

Грамматика:

- хотелось бы, чтобы можно было использовать информацию о регулярных правилах образования словоформ

Социолингвистическая информация

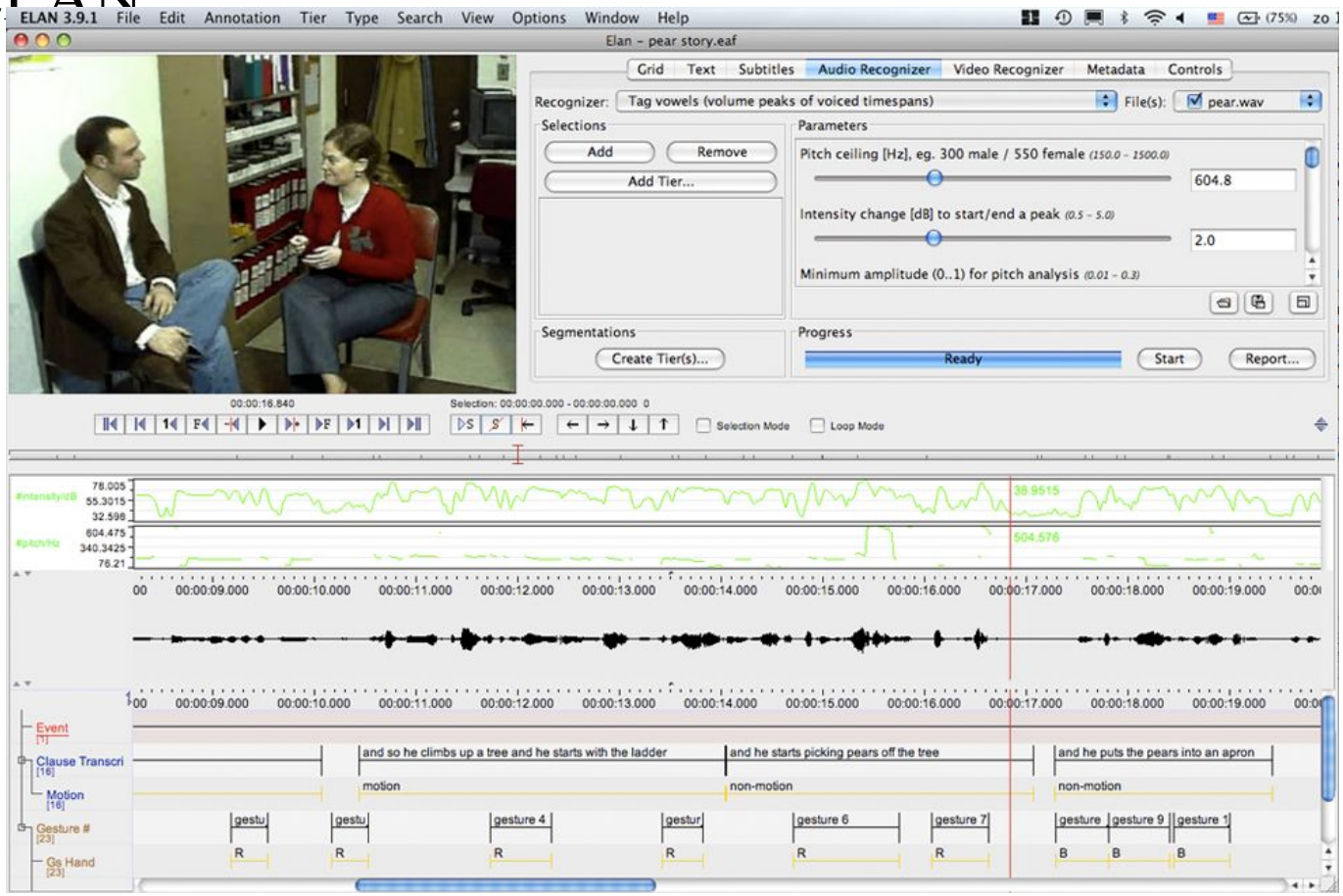
Поиск:

- хотелось бы, чтобы можно было искать все слова в одной и той же грамматической форме, все примеры [illegible]ного слова и т.п.

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Инструменты

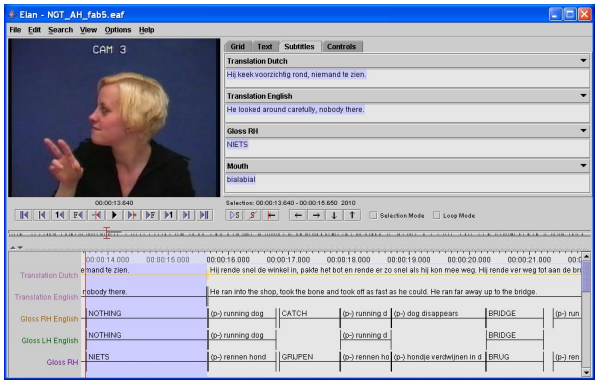
- ELAN



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Инструменты

- ELAN



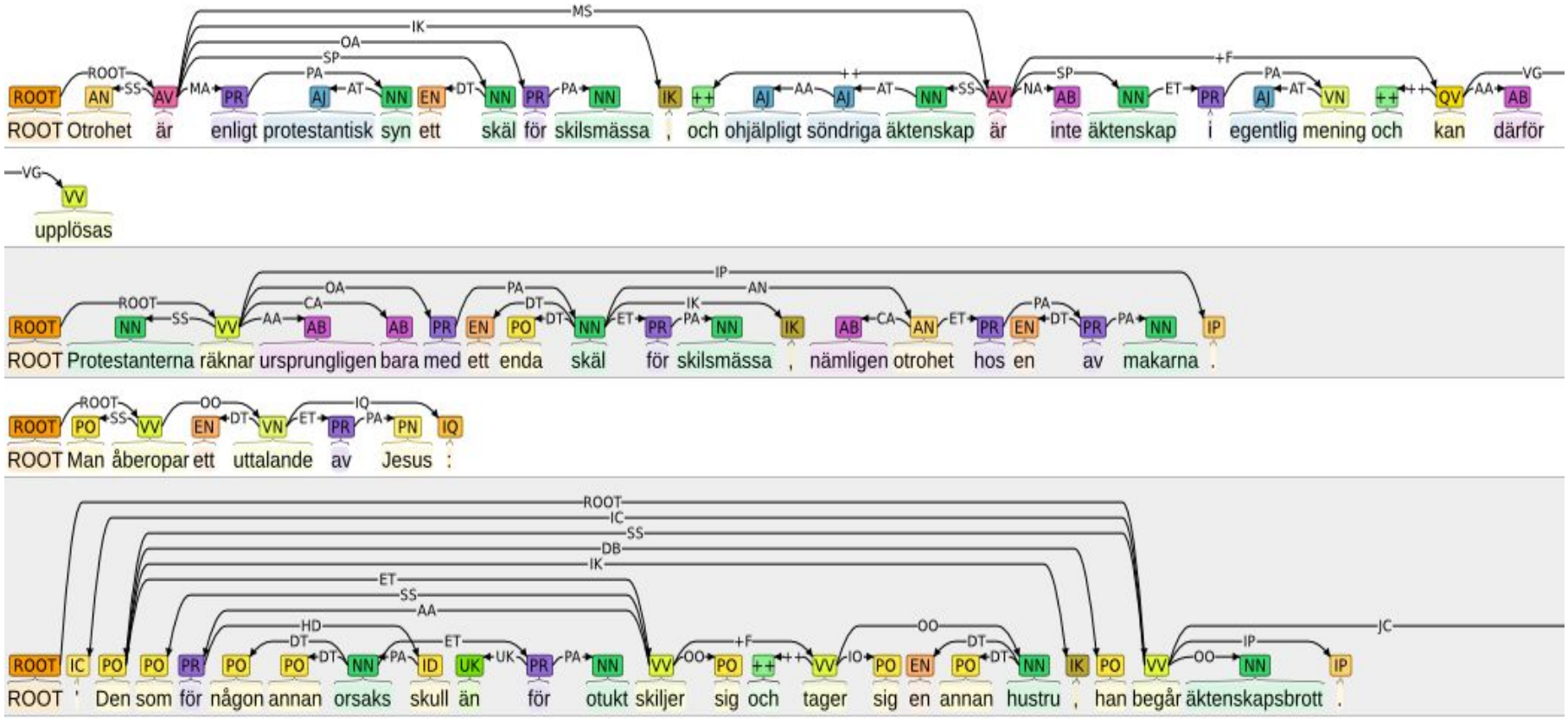
- синхронизация аннотации со звуком и видеорядом
- Возможность совмещения транскрипции с визуализацией параметров различных звукового ряда (спектрограммой…)
- многоуровневая аннотация
- возможность добавлять уровни аннотации пользователем
- возможность выгрузки в формате, доступном для автоматической обработки разных уровней
- возможности поиска

HIGHER SCHOOL OF ECONOMICS

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

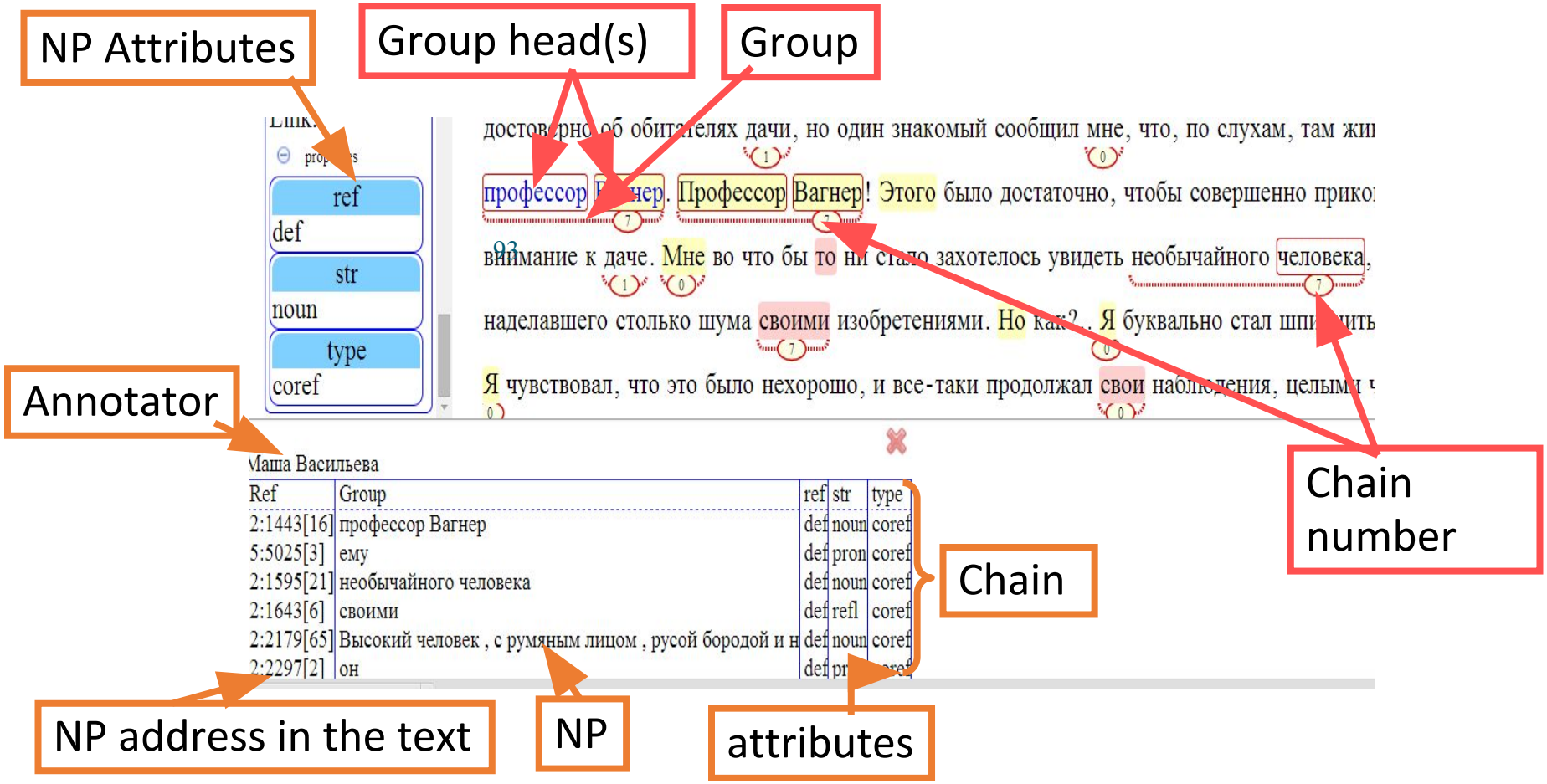
Разметка корпуса

Разметка корпуса
http://brat.nlplab.org/examples.html



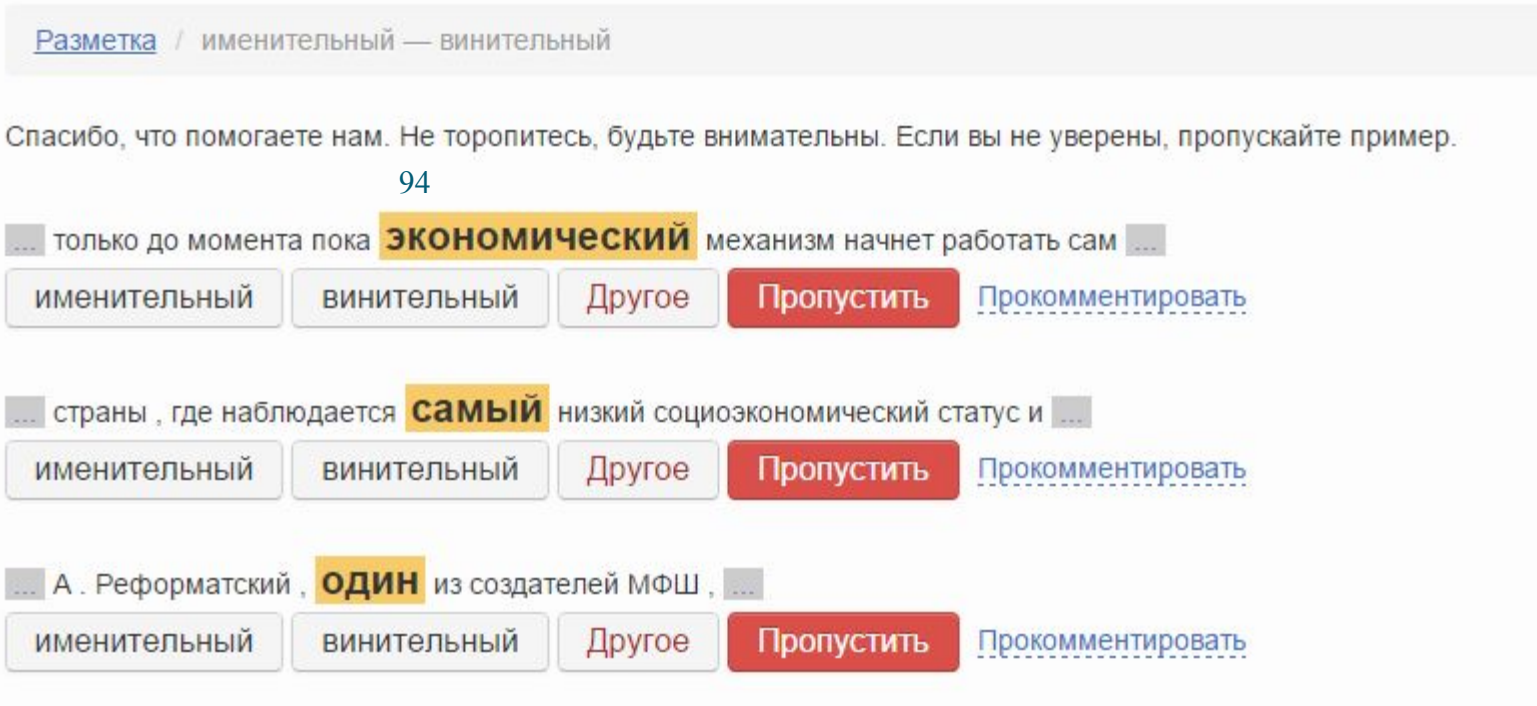
ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Разметка корпуса



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Краудсорсинг



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Геймификация

http://web-corpora.net/wsgi/senti_game.wsgi/

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Лингвистические ресурсы и инструменты

- Корпуса
- Специализированные базы данных
  - Лексикографические ресурсы
  - Типологические ресурсы
- Ресурсу по обучению языку
- Специализированные программы
  - обработка текста
  - обработка звучащей речи
  - разметка корпусов
  - визуализация лингвистических данных

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Лингвистические ресурсы и инструменты

- Язык: сложная иерархическая система + big data

  -> инструменты поддержки работы с многоуровневыми иерархическими данными, имеющими специфическое статистическое распределение:

  - визуализация данных
  - корпусные менеджеры, обеспечивающие работу с многоуровневыми данными / большими данными
  - краудсорсинг
  - геймификация

Методология и методы

- 3 направления: ресурсы, теоретические модели, приложения
- Принципы: обучение, тестирование, оценка качества
- Методы: машинное обучение, правила, гибридизация
- Задачи: разметить корпуса, выделить признаки (фичи), проверить их вклад / проверить метод / проверить вклад признака с фиксацией метода
- Pipeline: препроцессинг (графематическая нормализация, сегментация на токены и предложения); морфологический анализ; синтаксический анализ; анафора и кореферентность; извлечение именованных сущностей; анализ дискурса – у каждого этапа своя специфика