



Компьютерная ЛИНГВИСТИКА

С.Ю.Толдова, Е. Еникеева, И.
Макарчук



- 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

- 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)

- **Оценивание:**

- домашние задания – 30%
- проект – 25%
- чтение литературы – 20 %
- ИТОГОВЫЙ ТЕСТ – 25%

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

- 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.

- 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	Data	Schema	Apps	Docs
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				
	 <p>George Washington</p> <p>Date of birth: Feb 22, 1732</p> <p>Place of birth: Colonial Beach</p> <p>Religion: Anglicanism, Episcopal Church in the United States of America, Church of England</p> <p>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...</p> <p><i>US President, U.S. Congressperson, Military Commander</i></p> 			

- **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

- **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.

“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 (Digital Humanities)
 - Генерация текста
 - Вопросно-ответные системы, диалоговые системы, автоматические помощники

(Hirschberg Manning 2015)

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

- классическая область искусственного интеллекта, начало КЛ



Машинный перевод - история развития

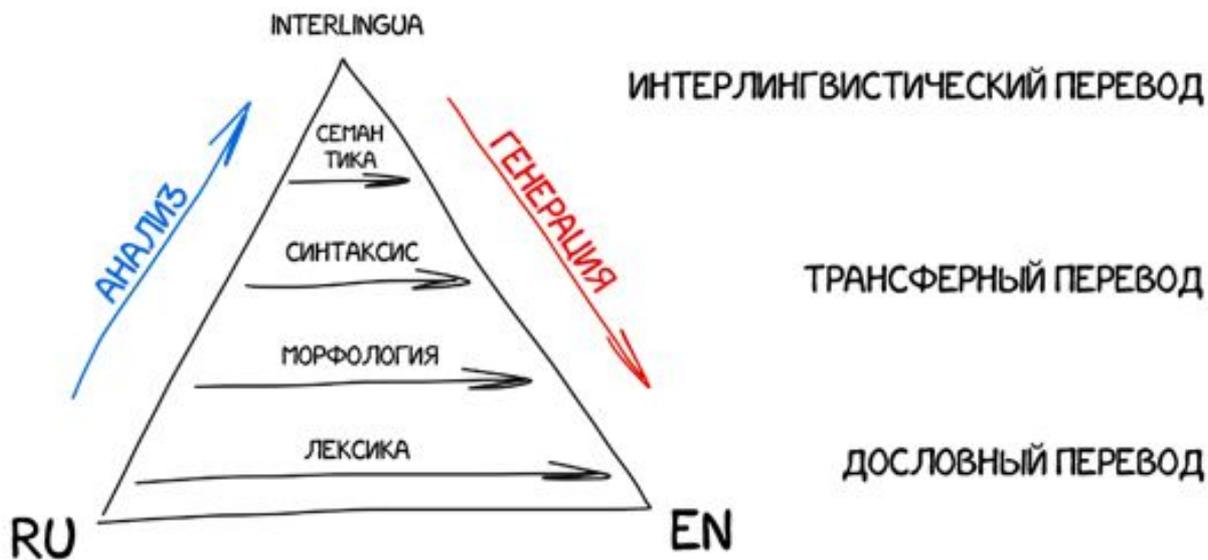
КРАТКАЯ ИСТОРИЯ МАШИННОГО ПЕРЕВОДА



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

Машинный перевод на правилах

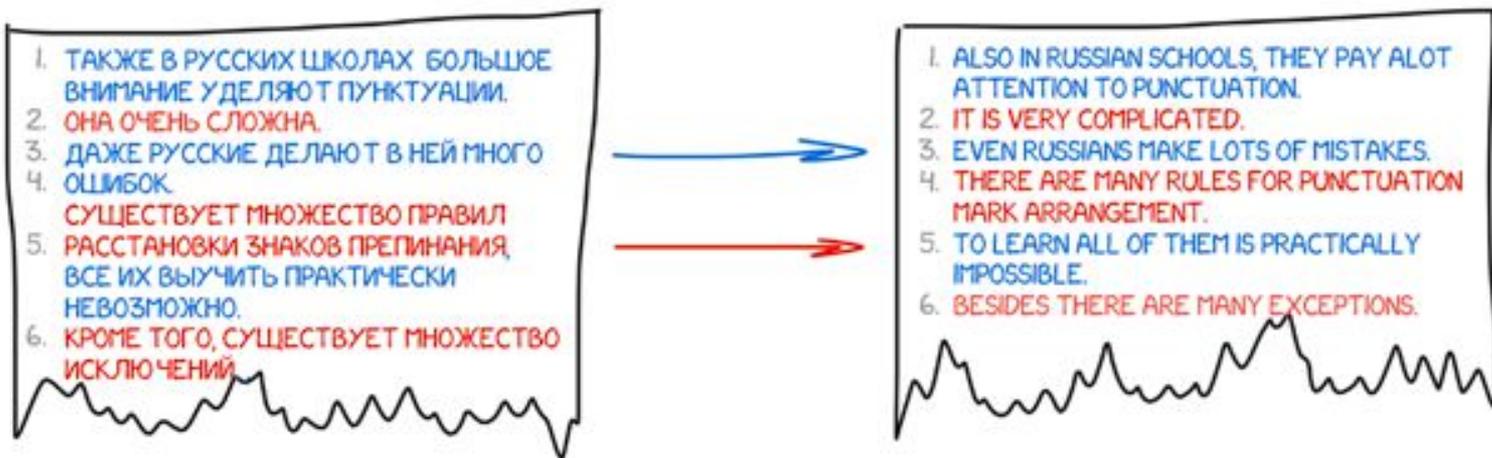
ПИРАМИДА
ВАКУА
(VAUQUIOS)



[http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny
_do_glubokogo_obucheniya.html](http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny_do_glubokogo_obucheniya.html)

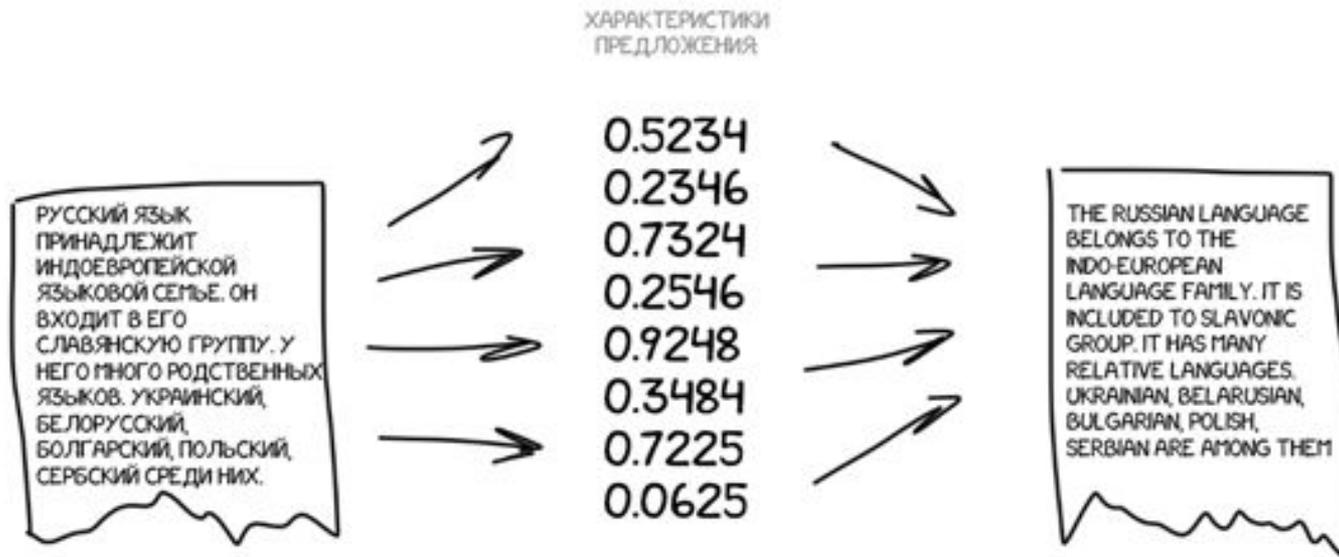
Машинный перевод: статистический подход

ПАРАЛЛЕЛЬНЫЙ КОРПУС



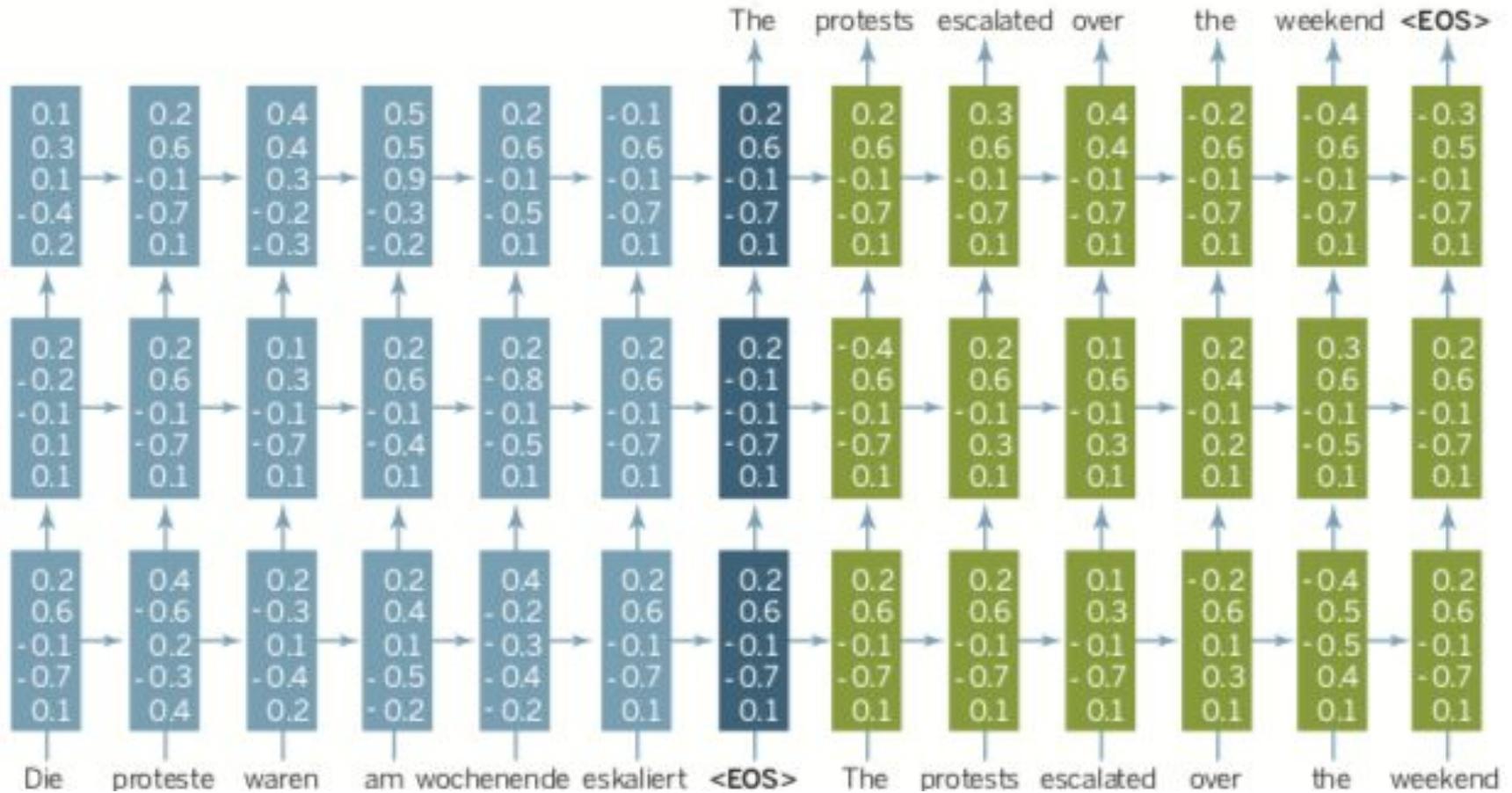
[http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny
_do_glubokogo_obucheniya.html](http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny_do_glubokogo_obucheniya.html)

Машинный перевод: нейросети



[http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny
_do_glubokogo_obucheniya.html](http://ai-news.ru/2018/02/mashinnyj_perevod_ot_holodnoj_vojny_do_glubokogo_obucheniya.html)

Машинный перевод: нейросети



Hirschberg Manning 2015

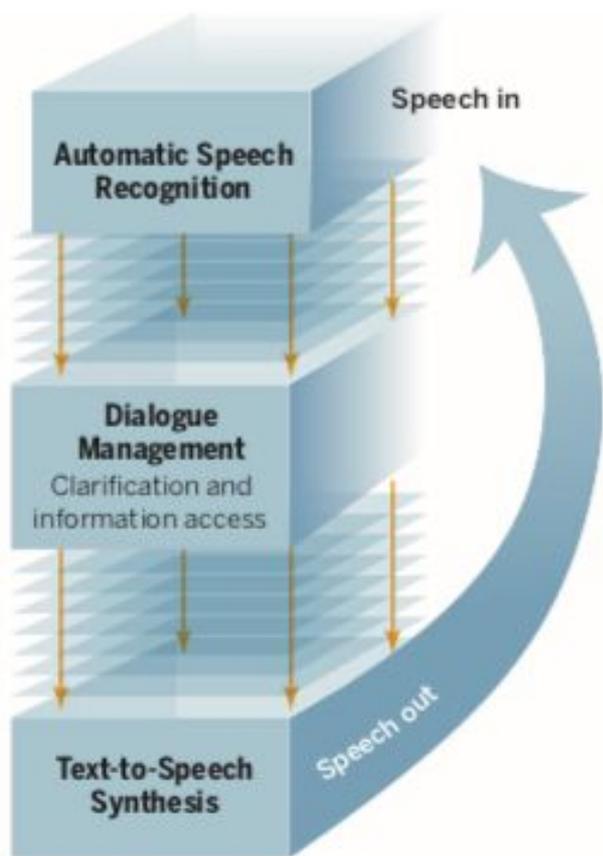
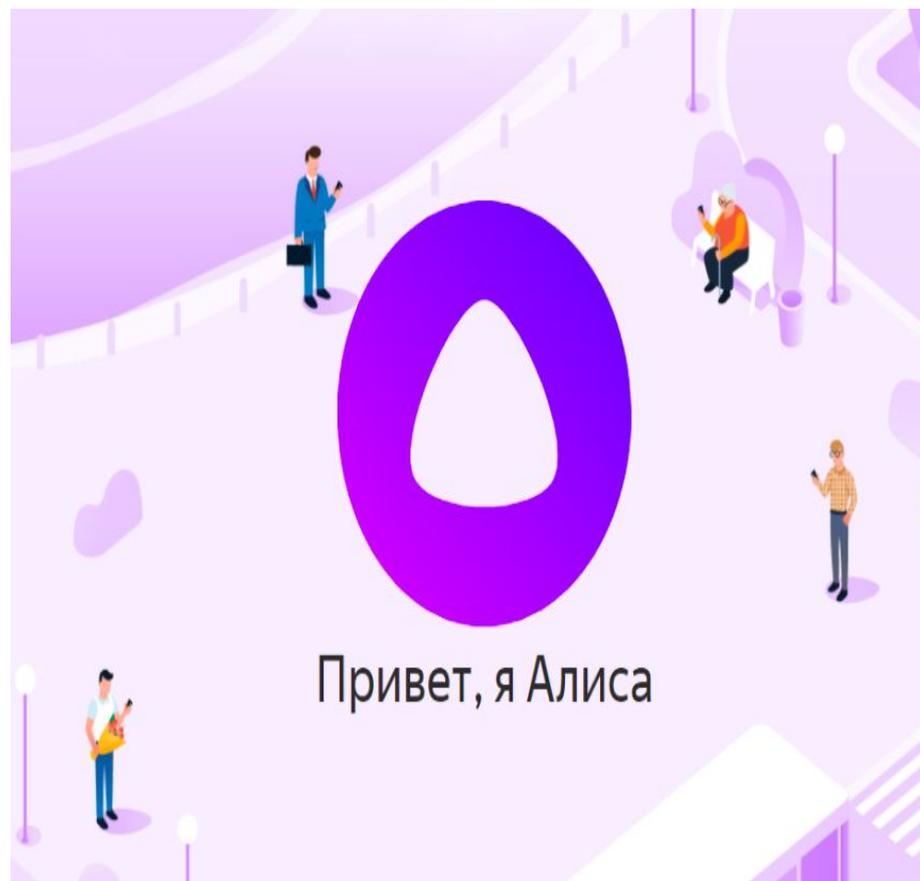


Fig. 3. A spoken dialogue system. The three main components are represented by rectangles; arrows denote the flow of information.



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 менеджменту диалога

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

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

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)

[Washington Informer](#)

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)

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

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*

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 словаря)

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

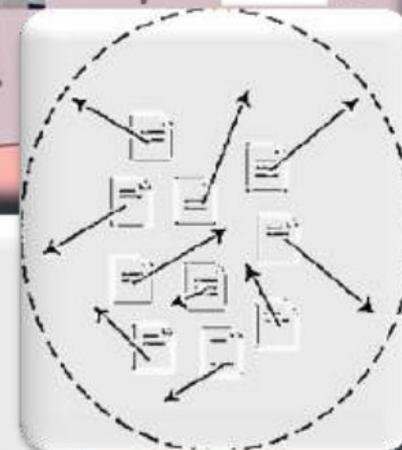
$P(w_{ij}) = 1/7$ (i – номер слова в «словаре», j – порядковый номер словоупотребления в тексте)

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*’
-

Обработка коллекция документов

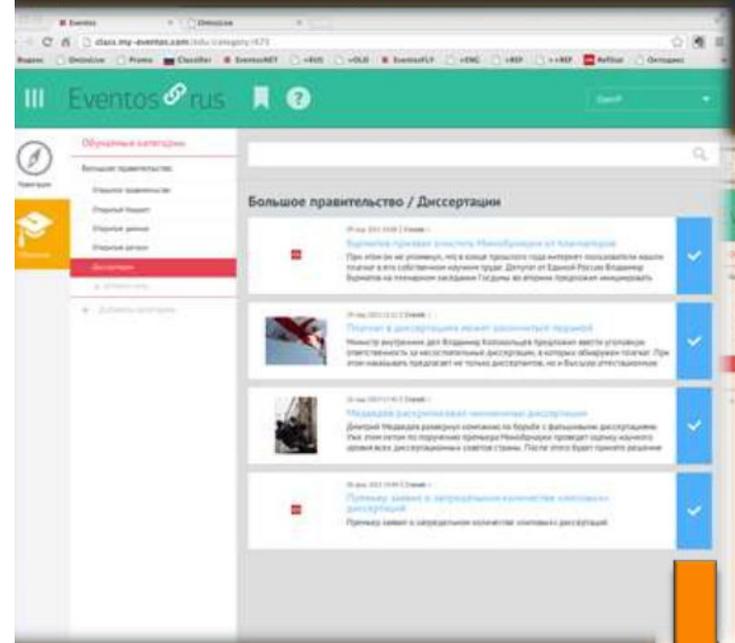
- Содержание каждого документа или тематической коллекции документов (*семантический портрет*) может быть единообразно описано вектором семантически значимых компонентов в пространстве семантических признаков. Компоненты:
 - текстовые n-граммы,
 - объекты,
 - теги и др. метаданные.
- Вектора можно сравнивать 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.



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

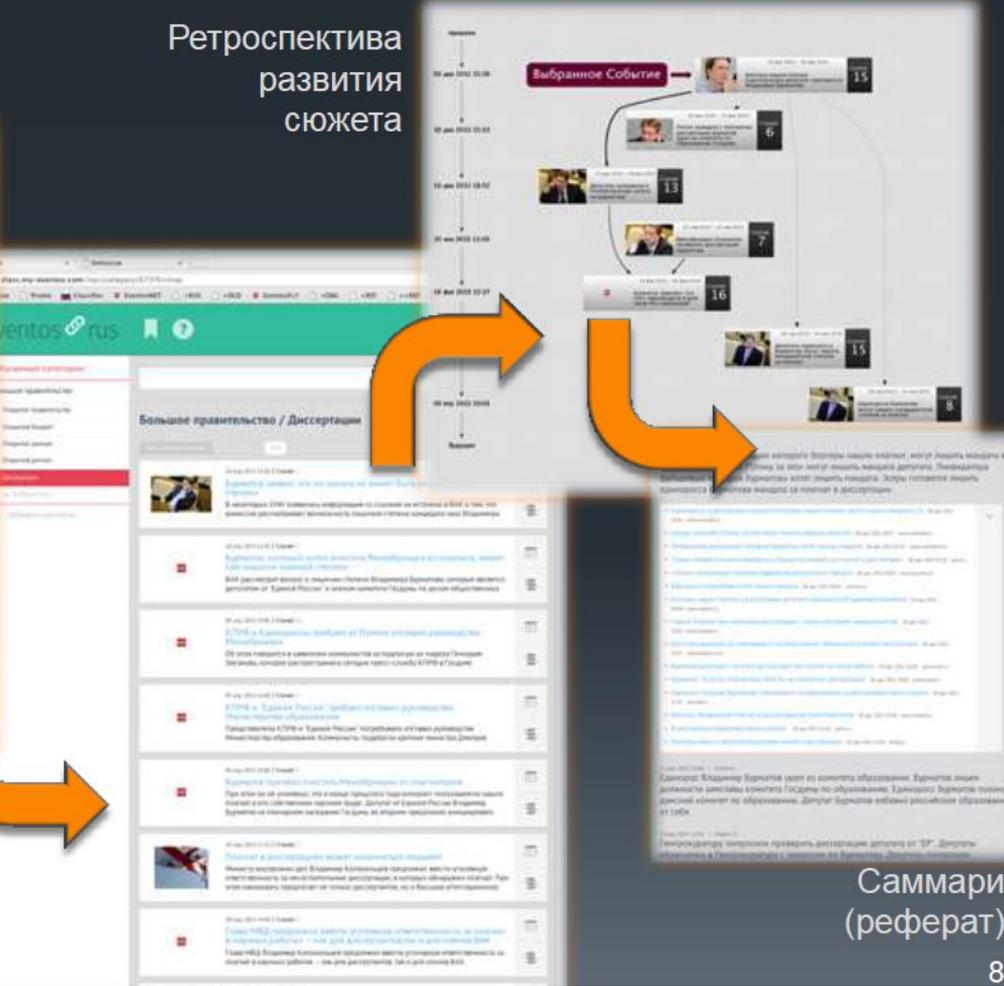
Обработка коллекция документов

Настройка и обучение персональных категорий



Автоматическое формирование тематических подборок, соответствующих созданным персональным категориям

Ретроспектива развития сюжета



Обработка коллекции документов

Карта новостей

Санкции против КНДР

Захват миротворцев

Смерть Уго Чавеса

Биатлон. Сочи

Депутаты заблокировали раду

Погиб А. Панин

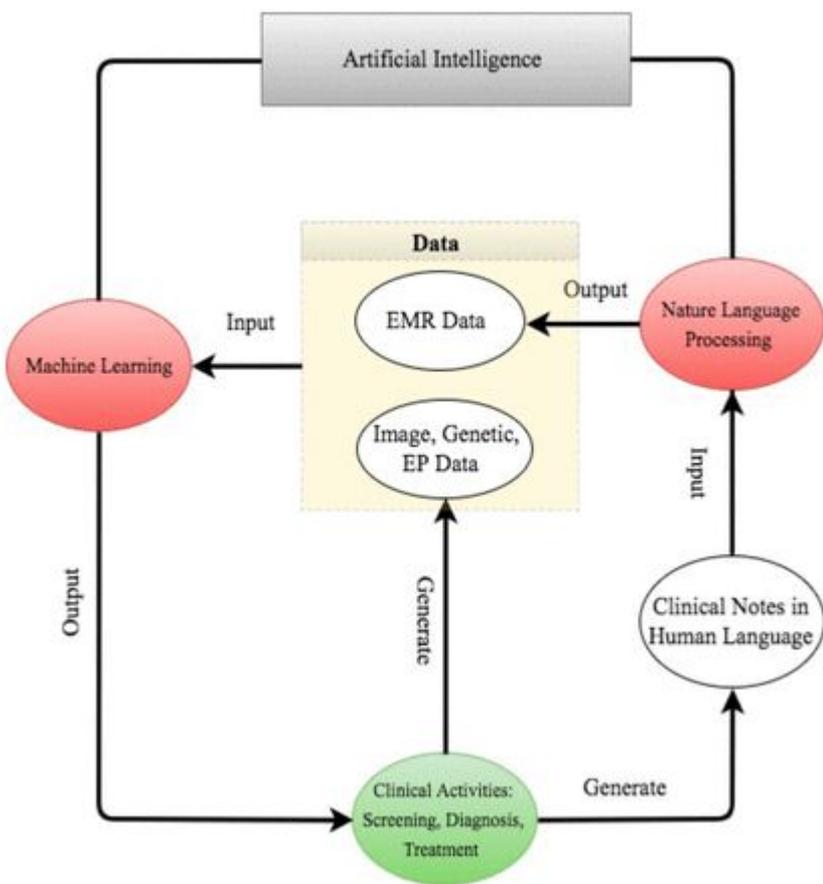
Нападение на С. Филина

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)

Структурированные данные могут быть обработаны



Структурированные данные
встраиваются в другие
процессы (и это AI)

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: <table border="1"> <thead> <tr> <th>location</th> <th>contained by</th> </tr> </thead> <tbody> <tr> <td>Colonial Beach</td> <td>Westmoreland County Virginia United States of America</td> </tr> </tbody> </table>	location	contained by	Colonial Beach	Westmoreland County Virginia United States of America
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				



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: <table border="1"> <thead> <tr> <th>location</th> <th>contained by</th> </tr> </thead> <tbody> <tr> <td>Colonial Beach</td> <td>Westmoreland County Virginia United States of America</td> </tr> </tbody> </table>	location	contained by	Colonial Beach	Westmoreland County Virginia United States of America
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/>

Извлечение информации из текста

Objects	Properties	Text
All Person Обама Барак Патрушев Николай ✓ Путин Владимир Шойгу Сергей Джордж Литтл Кэйтлин Хэйден Томас Донилон Юрий Ушаков Географическое ... Вашингтон Ирландия Москва Оклахома Россия США	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	Политика  <p>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”,</p>

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
<code>rdf:type</code>	<ul style="list-style-type: none"> ▪ <code>skos:Concept</code>
<code>rdfs:label</code>	<ul style="list-style-type: none"> ▪ George Washington
<code>owl:sameAs</code>	<ul style="list-style-type: none"> ▪ 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
<code>skos:broader</code>	<ul style="list-style-type: none"> ▪ <code>category:Presidents_of_the_United_States</code> ▪ <code>category:Washington_family</code> ▪ <code>category:Wikipedia_categories_named_after_American_politicians</code>
<code>skos:prefLabel</code>	<ul style="list-style-type: none"> ▪ George Washington
<code>http://www.w3.org/ns/prov#wasDerivedFrom</code>	<ul style="list-style-type: none"> ▪ http://en.wikipedia.org/wiki/Category:George_Washington?oldid=490387017
<code>is dcterms:subject of</code>	<ul style="list-style-type: none"> ▪ <code>dbpedia:Washington's_Birthday</code> ▪ <code>dbpedia:United_States_presidential_election_1788-1789</code> ▪ <code>dbpedia:1932_Washington_Bicentennial</code> ▪ <code>dbpedia:Conway_Cabal</code> ▪ <code>dbpedia:Potomac_Company</code> ▪ <code>dbpedia:List_of_places_named_for_George_Washington</code>

Извлечение информации из текста

Тексты бывают разные

- Подписи под фотографиями:

Explore / Tags / animals

Sort by:
Most recent • Most interesting

animals clusters
Explore and refine this animals list with our wonderful cluster goodness!

Related tags:
nature, animal, zoo, cat, birds, cats, bird, pet, cute, water

Find similar things on



From Life Without Taffy

From Mike:Watson

From ratexla

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

- ✓ Как должны быть устроены ярлыки, чтобы можно было найти фотографии по одной теме?

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

- 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].

Последовательность этапов анализа, соответствующая уровням языка:

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

<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)

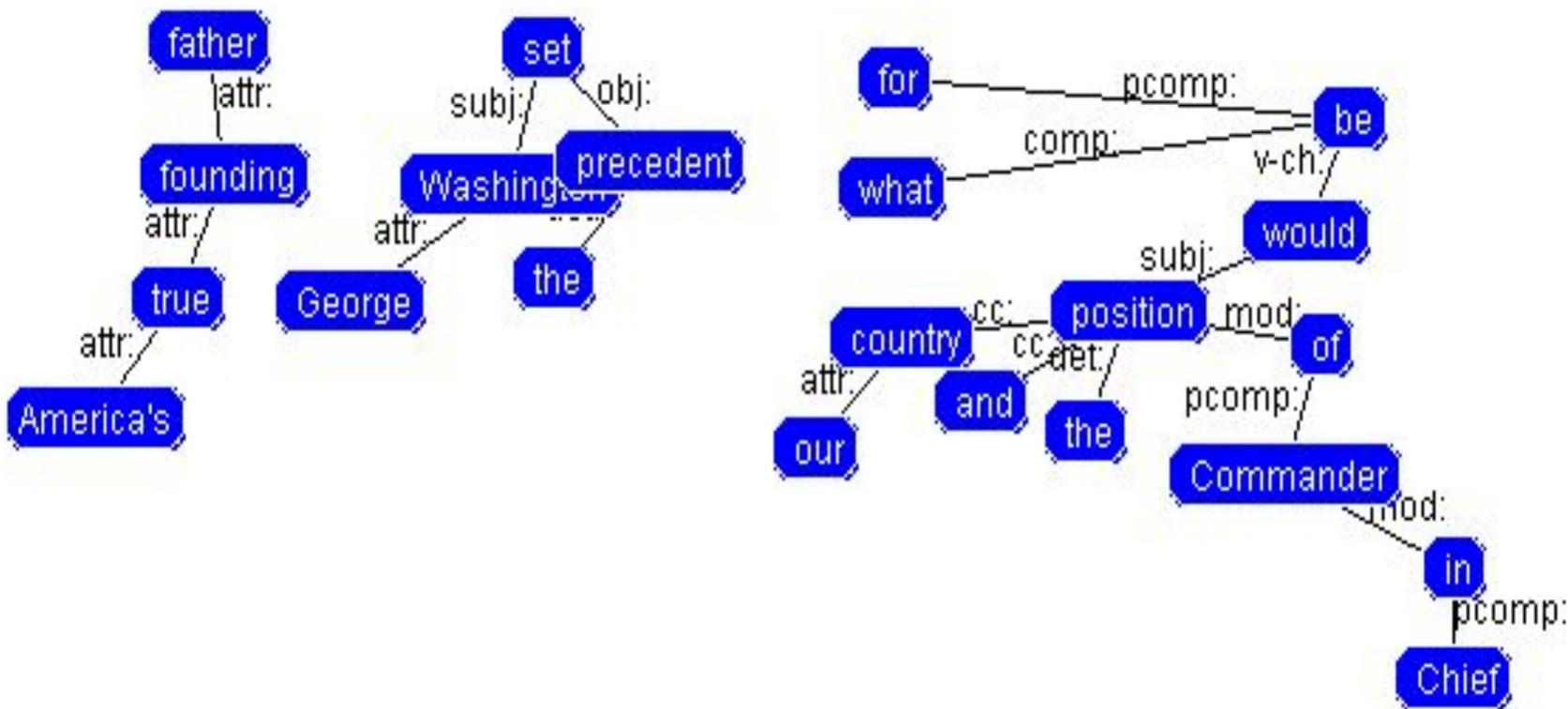
Фридмана {Фридман=С,фам,мр,од=ед,рд}

```
1    Dа    dа    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
```

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>

[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 @ADV> #10->22_{ADV>}
 [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}

```

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

```

Последовательность этапов анализа, соответствующая уровням языка:

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

Generic Relations

relationsubject: George Washington

relationobject: the precedent

verb: set

r, George Washington set the precedent for what our country and th

<http://viewer.opencalais.com/>

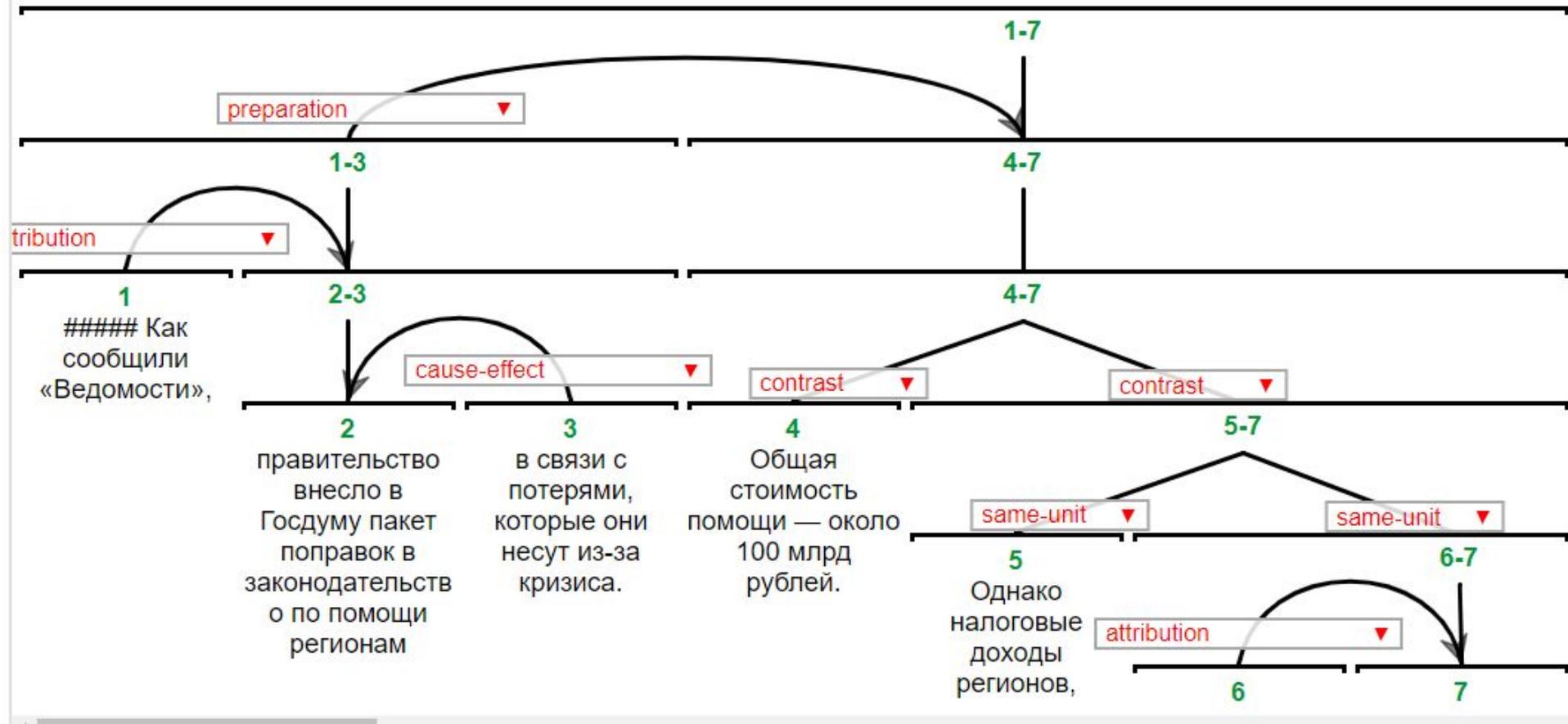
Компьютерная лингвистика 3: уровни анализа

Последовательность этапов анализа, соответствующая уровням языка:

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

- **[America's true founding father]**_i, **[George Washington]**_j, 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]**_k, was elected as **[Commander in Chief of the Continental Army]** and led **[his]**_l, ill-trained, under supplied troops to victory and thus, independence. In a letter to James Madison **[Washington]**_m, wrote "As the first of everything, in our situation will serve to establish a Precedent...it is devoutly wished on **[my]**_n part, that these precedents may be fixed on true principles."

NLP Pipeline: coreference



Компьютерная лингвистика 3: уровни анализа

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

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

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

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

- Корпус: 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

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

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 докумен



Ресурсы

Компьютерные словари

ABBYY® Lingvo

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

инструмент

Перевести

Русский

Английский

Переводы

Примеры (1058)

Словосочетания (1058)

Толкования (0)

Добавить перевод

Примеры из текстов (1058)

Предлагаемая технология позволяет регулировать размер и конфигурацию отверстий, а значит, имеется **инструмент**, позволяющий управлять амплитудой волны, рассеянной каждым отверстием.

The proposed technology makes it possible to control the hole size and configuration, hence there is a tool to control the amplitude of a wave scattered by each hole.

Подобные устройства фазового сдвига известны в технике спектрального уплотнения как **инструмент** для настройки спектральных характеристик оптических фильтров на основе ИМЦ, а также используются в других устройствах - модуляторах и переключателях.

Such phase shifters are known in sphere of wavelength division multiplex technologies as instrument for adjustment of spectral characteristics of optical filters on the base of MZI, and also used in other devices— optical modulators and switches.

Wiktionary

67

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: [tʃat]

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 вечером к нам в **чат**.

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

Синонимы

1. —
2. болталка (*жарг.*)

Антонимы

Гиперонимы

Гипонимы

Родственные слова

Ближайшее родство

Этимология

Заимствование из англи...

Ресурсы

Словарные агрегаторы

OneLook
Dictionary Search

<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](#)]

Ресурсы Словарные агрегаторы

[http://dic.academ
ic.ru](http://dic.academic.ru)



Словари и энциклопедии на Академике

чат

[Толкования](#) [Переводы](#) [Книги](#)

чат

[Толкование](#) [Перевод](#) [Книги](#)

- Чат** — ...
[Википедия](#)
- Чат** — сервис обмена текстовыми сообщениями в режиме реального времени. Чат позволяет многим пользователям одновременно общаться между собой. По английски: Chat См. также: Телеконференции Финансовый словарь Финам ...
[Финансовый словарь](#)
- ЧАТ** — [англ. chat беседа, болтовня] виртуальное место встречи в Интернете. Приват чат часть ч., где «встречаются» только двое собеседников в компьютерном диалоге. Словарь иностранных слов. Комплев Н.Г., 2006. чат (англ. chat) комп. обмен сообщениями в... ...
[Словарь иностранных слов русского языка](#)
- чат** — сущ., кол во синонимов: 10 • авточат (1) • болталка (3) • веб чат (1) • ...
[Словарь синонимов](#)

- чат** — Система для обмена информацией в интернете, разговора в реальном времени. [http://www.lexikon.ru/rekl/a_eng.html] Тематики реклама ...
[Справочник технического переводчика](#)
- чат** — I. и. 1. Ике яки берниче урам кисешкән урын 2. Ике яки берниче юл кушылган яисе кисешкән урын 3. Нин. б. әйбер, корылма, бина янындагы урын, почмак кызыл йорт чатында трамвайдан төште. ЧАТ БАШЫ – Чат, урам кисешкән урын. II. ЧАТ – с. 1. Бөтенләй... ...
[Татар теленең аңлатмалы сүзлеге](#)
- чат** — 1. ຫາດ [Ча:т] 2. ຫາດ [Ча:т] 3. ຫັດ [чат] ...
[Фонетический словарь Тайского языка](#)
- чат** — (Сем.: Мақ., Ұрж.) биік таудың ішін жарып кірген терең сай. Біз былтыр мынау ч а т т а отырдық (Сем., Ұрж.) ...
[Қазақ тілінің аймақтық сөздігі](#)
- Чат (фильм)** — Чат Chatroom Жанр триллер, драма Режиссёр ...
[Википедия](#)
- чат масала** — Чат масала смесь специй песочного цвета. Состоит из нескольких компонентов, главными из которых являются порошок манго, черная соль и асафетида. Этой смесью традиционно заправляют фруктовые салаты. Ниже приводится рецепт приготовления... ...
[Кулинарный словарь](#)
- Чат, перевал** — (ЗК) Чат, перевал (ЦК) ...
[Энциклопедия туриста](#)
- чат-(ый)** — I суффикс Словообразовательная единица, выделяющаяся в именах прилагательных со значениями 1) имеющий в большом количестве то или состоящей из множества того, что названо словами, от которых соответствующие имена прилагательные образованы... ...
[Современный толковый словарь русского языка Ефремовой](#)



Лексикографические ресурсы

Компьютерные словари

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

9/12/2019

[ə] ШКОЛА
ЛИНГВИСТИКИ
НИУ ВШЭ

Лексикографические ресурсы

Компьютерные ресурсы

- частотные списки и словари

СЛОВАРИ,
созданные на основе
НАЦИОНАЛЬНОГО
КОРПУСА
РУССКОГО
ЯЗЫКА



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/>



Ресурсы Тезаурусы

THINKMAP® VISUAL THESAURUS

Look up a Word:

type a **cup** to search

LOOK IT UP

Search History Random Word Language: English

☰ NOUNS ON OFF

- a small open container usually used for drinking; usually has a handle
- the quantity a cup will hold
- any cup-shaped concavity
- a United States liquid unit equal to 8 fluid ounces

☰ VERBS ON OFF

- form into the shape of a cup
- put into a cup
- treat by applying evacuated cups to the patient's skin
- introduce
- give shape or form to

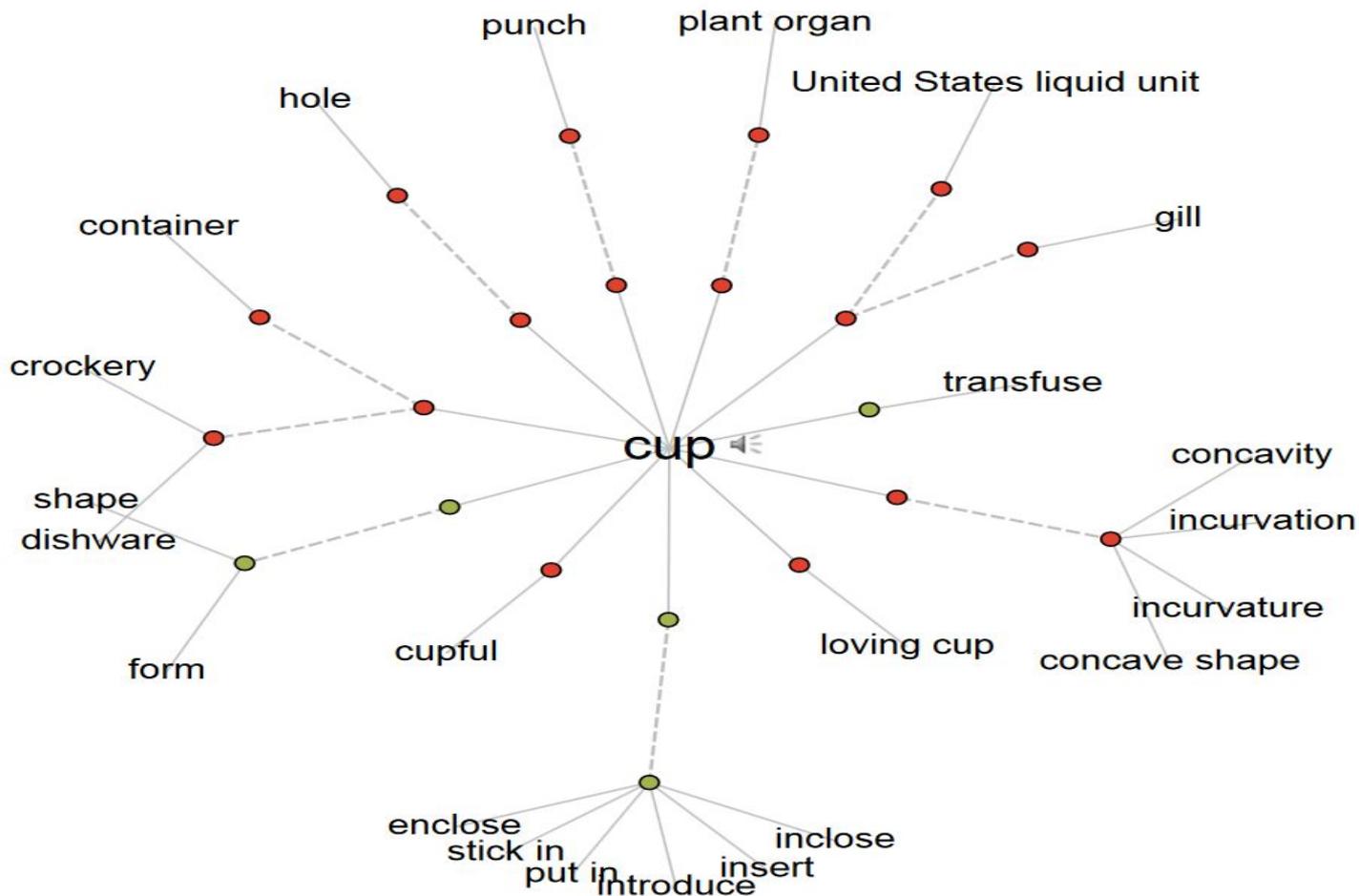
☰ ADVERBS ON OFF

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



Ресурсы Тезаурусы

[http://www.visualthesaurus.com/
app/view](http://www.visualthesaurus.com/app/view)



Ресурсы: предобученные векторные модели

Вычисление семантических ассоциатов

Введите слово, чтобы получить список из 10 его ближайших семантических аналогов (квази-синонимов). Вы можете приписать к слову знак подчеркивания "_" и тэг части речи ("река_NOUN"). Если вы этого не сделаете, RusVectores определит часть речи автоматически.

чат_NOUN

Выберите модель:

НКРЯ и русская Wikipedia Новостной корпус НКРЯ Araneum Russicum Maximum Веб-корпус

Показывать только:

Существительные Глаголы Наречия Прилагательные Все части речи Часть речи запроса

Найти похожие слова!

Семантические аналоги для *чат* (ALL)

НКРЯ и русская Wikipedia

1. irc 0.63
2. онлайн 0.60
3. онлайн 0.59
4. офлайн 0.59
5. онлайн 0.59
6. icq 0.58
7. офлайн 0.57
8. аккаунт 0.57
9. gmail 0.57
10. онлайнновый 0.57

<http://rusvectors.org/ru/similar/#>

9/12/2019

[Э] 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, ИНДИВИДУМ, ЧЕЛОВЕЧЕСКАЯ ЛИЧНОСТЬ [Понятие RuТез: ЧЕЛОВЕК]

hypernym

- 1 СУБЪЕКТ ДЕЯТЕЛЬНОСТИ [Понятие RuТез: СУБЪЕКТ ДЕЯТЕЛЬНОСТИ]
- 2 ЖИВОЕ, ОСОБЬ, ЖИВОЕ СУЩЕСТВО, ЖИВОЙ ОРГАНИЗМ, ЖИВНОСТЬ, ОРГАНИЗМ, СУЩЕСТВО, ИНДИВИДУМ, МАКРООРГАНИЗМ, БИОЛОГИЧЕСКИЙ ОРГАНИЗМ [Понятие RuТез: ЖИВОЙ ОРГАНИЗМ]

hyponym

- 1 БОБЫЛЬ, БОБЫЛКА [Понятие RuТез: НЕ СОСТОЯЩИЙ В БРАКЕ]
- 2 ПОМОЩНИК, ПОМОЩНИЦА [Понятие RuТез: ПОМОЩНИК (ЧЕЛОВЕК)]
- 3 ОБВИНИТЕЛЬ [Понятие RuТез: ОБВИНИТЕЛЬ]
- 4 ПЕШЕХОД [Понятие RuТез: ПЕШЕХОД]
- 5 ЭКСПЕРИМЕНТАТОР, ЭКСПЕРИМЕНТАТОРША [Понятие RuТез: ЭКСПЕРИМЕНТАТОР]

<http://ruwordnet.ru/en/>

Тезаурис
РуТез



<http://ruscorpora.ru/>

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 \(1\)](#)

Наконец, она же, вкупе с несколькими другими, вызвала "структуралистский" переворот в **лингвистике** и далее такой же переворот едва ли не во всех социальных науках. [Михаил Арапов. Когда текст обретает смысл // "Знание-сила", №1", 2003]
[\[омонимия снята\]](#) ←...→
2. Конференция по когнитивной науке (2003) [\[омонимия снята\]](#) [Все контексты \(1\)](#)

Считается, что когнитивная наука, появление которой относят к середине XX столетия, существует и развивается на стыке между такими областями знания, как психология, **лингвистика**, нейронаука, компьютерная наука и искусственный интеллект, когнитивная антропология и философия сознания. [Конференция по когнитивной науке (2003)] [\[омонимия снята\]](#) ←...→

Slovak Academy of Sciences E. 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 (<i>African TLDs</i>)	Araneum Anglicum Africanum	 	 	
English (<i>Asian TLDs</i>)	Araneum Anglicum Asiaticum	 	 	
Finnish	Araneum Finnicum	 	 	
French	Araneum Francogallicum	 	 	
French (<i>African TLDs</i>)	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 (<i>Russia-only TLDs</i>)	Araneum Russicum Russicum	 	 	

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

Slovak Academy of Sciences E. 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 (<i>African TLDs</i>)	Araneum Anglicum Africanum	 	 	
English (<i>Asian TLDs</i>)	Araneum Anglicum Asiaticum	 	 	
Finnish	Araneum Finnicum	 	 	
French	Araneum Francogallicum	 	 	
French (<i>African TLDs</i>)	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 (<i>Russia-only TLDs</i>)	Araneum Russicum Russicum	 	 	
Russian (<i>non-Russia TLDs</i>)	Araneum Russicum Estimum	 	 	

<http://aranea.iuls.savba.sk/guest/index.html>

Ресурсы Специализированные корпуса

R

has aspects

no aspects

all docs

Arrows

Full Doc

Word Filter

Link Filter

Aspect Filter

Document Filter

Clear All

Word [1] *

Token ⊕

приветливый ⊖

Gramm ⊕

Gramm ⊖

Word [2] *

Token ⊕

⊖

Gramm ⊕

N% ⊖

Link [3] *

Type ⊕

Head[1] ⊖

Child[2] ⊖

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

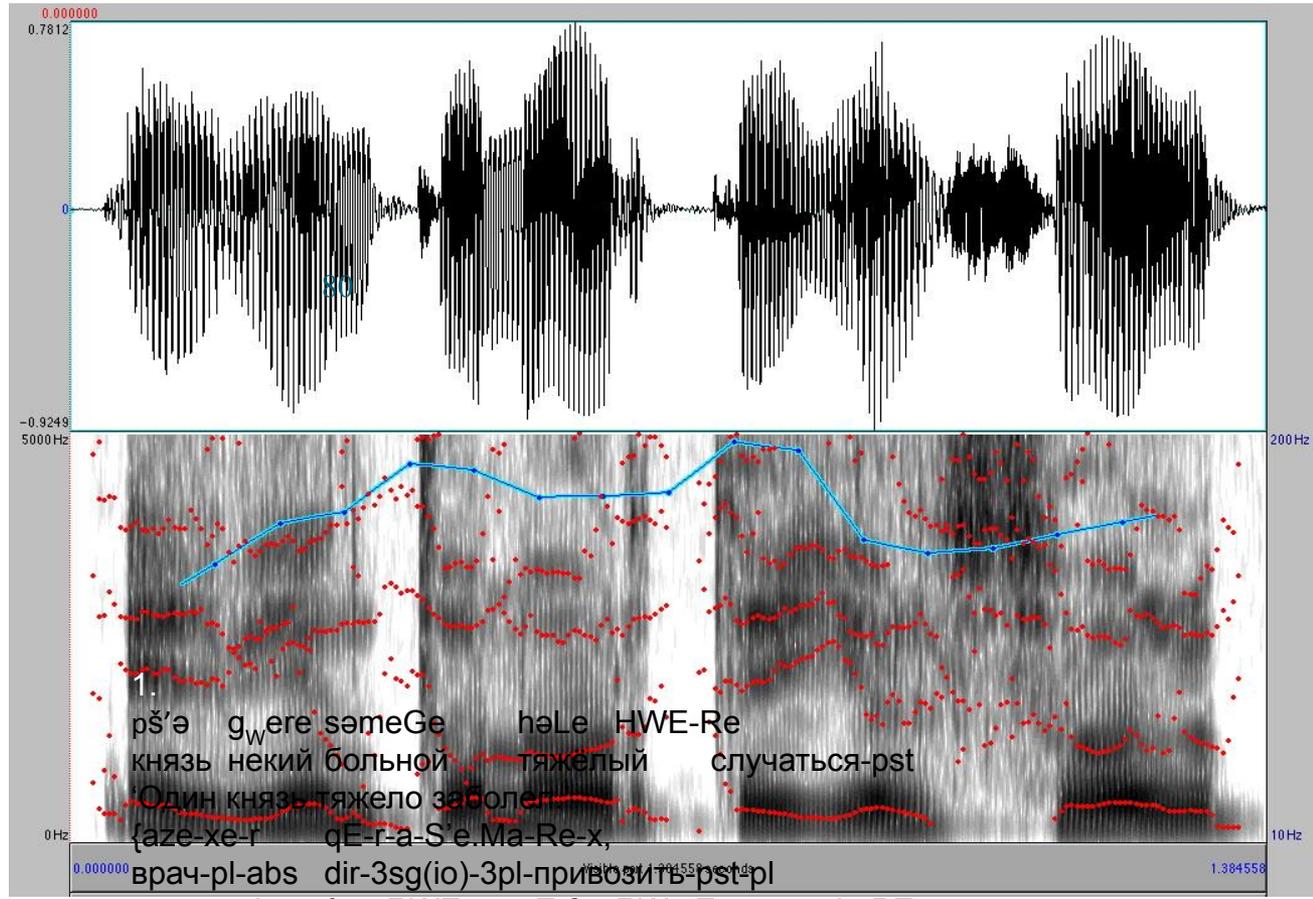
Doc.№ 119	<input checked="" type="checkbox"/> Arrows	category	meta	score
Title	review8555	Food both	date 26.06.2011 20	food 10
FileName	null	Interior positive	object 22.13	interior 9
URL	null	Price absence	useful 0	service 10
		Service positive	user Statist	
		Whole positive		

11: **Официантки** очень **приветливы**

Doc.№ 161	<input checked="" type="checkbox"/> Arrows	category	meta	score
Title	review19265	Food positive	date 15.11.2009 03	food 7
FileName	null	Interior both	object Тепло	interior 10
URL	null	Price absence	useful 2	service 9
		Service positive	user Лёлик	
		Whole both		

6: **Персонал** **приветливый** **отзывчивый**

Ресурсы Мультимедийные корпуса



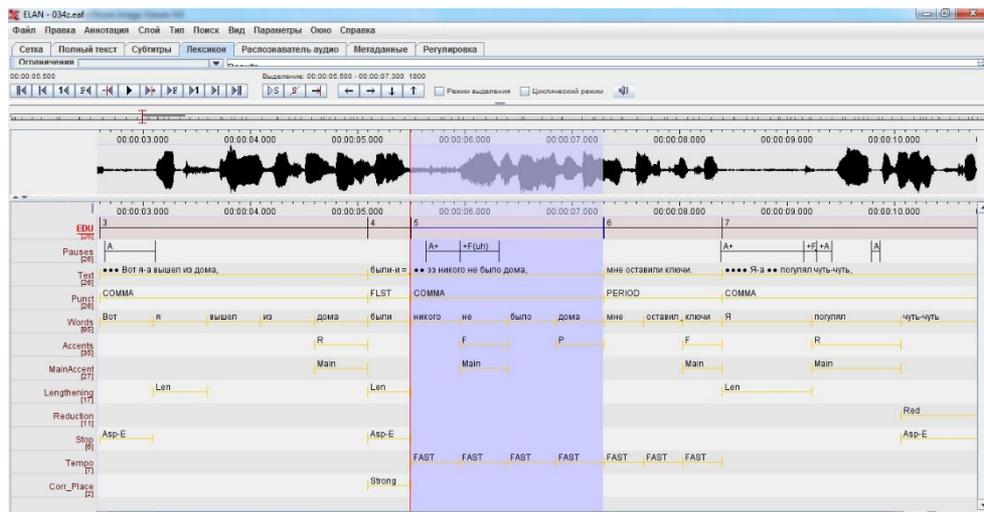
aw z-jə {ezeRWE qE-f-a-RWetE-n a-LeBE-r-ep
но один-coh лекарство dir-ben-3pl-найти-pot 3pl-мочь-dyn-neg
'Врачей к нему привозили, но они не смогли найти никакого лекарства.'



Ресурсы Мультимедийные корпуса

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

Комментарии Скрыть	Транскрипция		
	Время	№	ЭДЕ
Вид транскрипции Полная Упрощённая Минимальная	0.00	1.	Мне /приснилось,
Чтобы выделить фрагмент, нажмите на первую и последнюю	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,



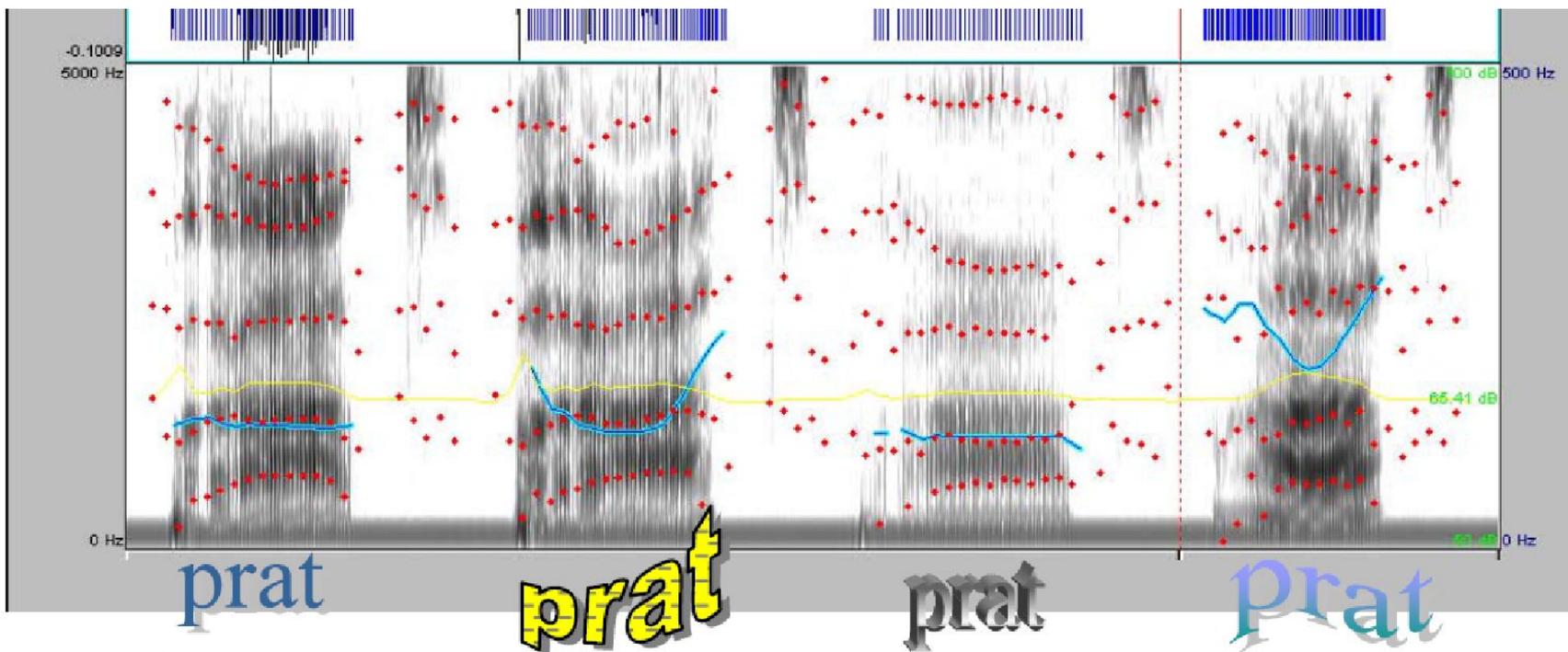
- что и зачем нужно искать пользователю в корпусе
- быстрая обработка больших объемов текстов ⁸²
- визуализация информации
- разметка
- поиск по разным «причудливым» параметрам

Инструменты

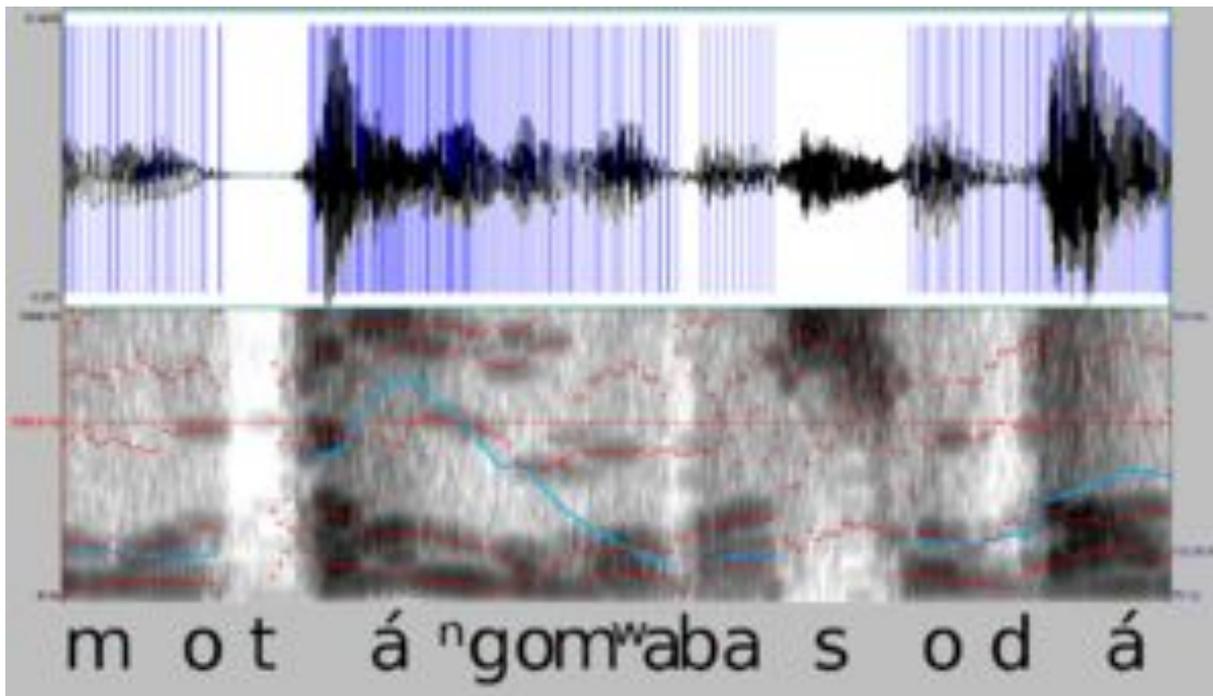
- специальные программы обработки звука и видеоряда (анализ текста как мультимодального явления):
 - программы разметки речевых корпусов
 - программы анализа звуков речи
 - SpeechAnalyzer, Praat, Elan
- Специальные программы для работы с графикой (шрифты, дополнительные символы, распознавание символов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

- PRAAT

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

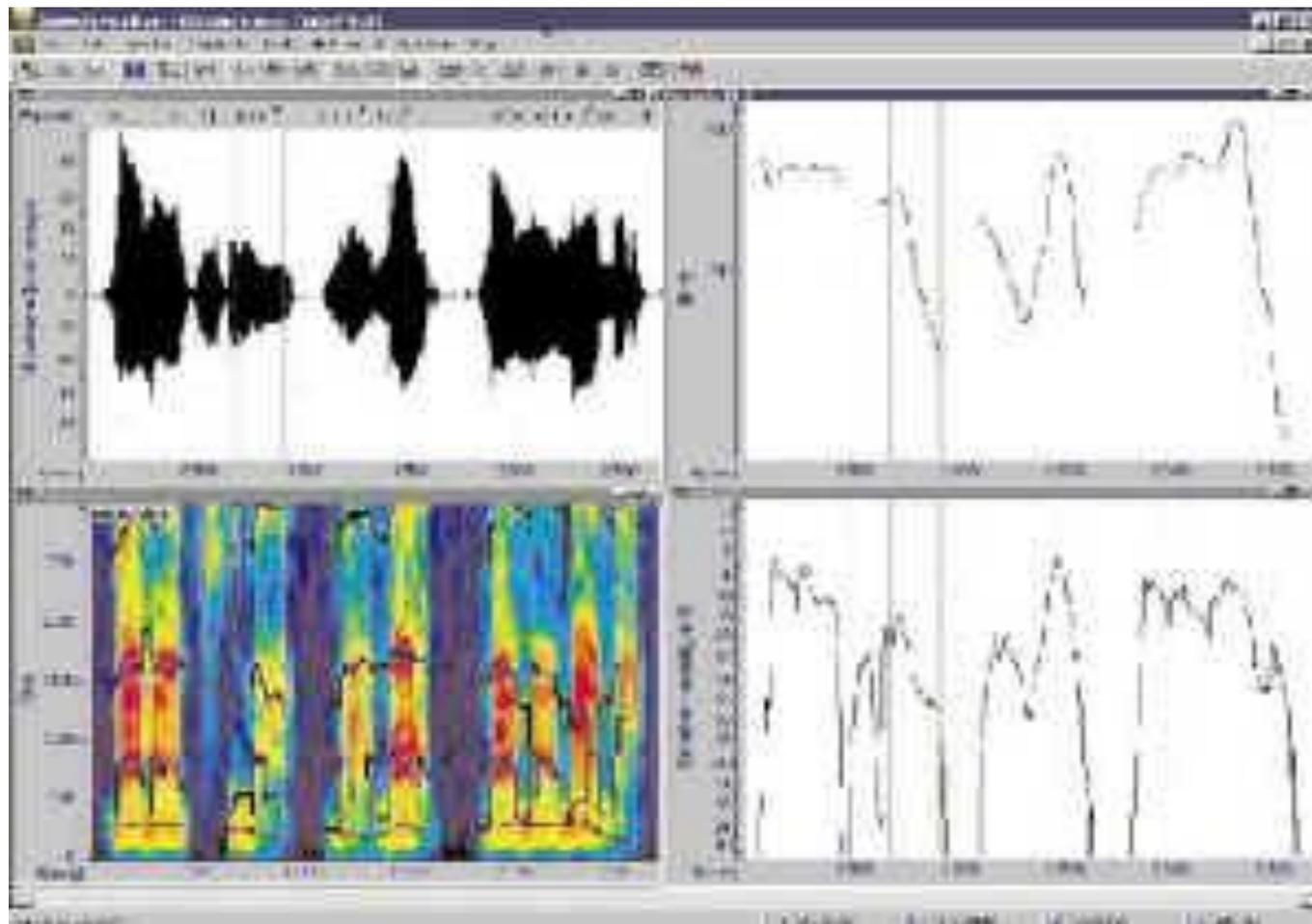


- PRAAT



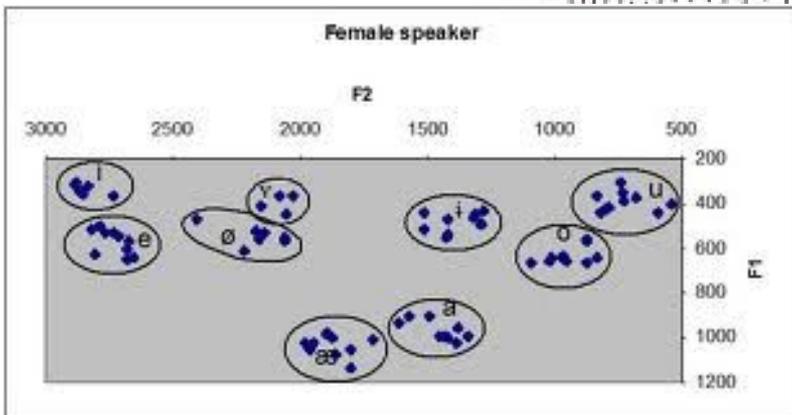
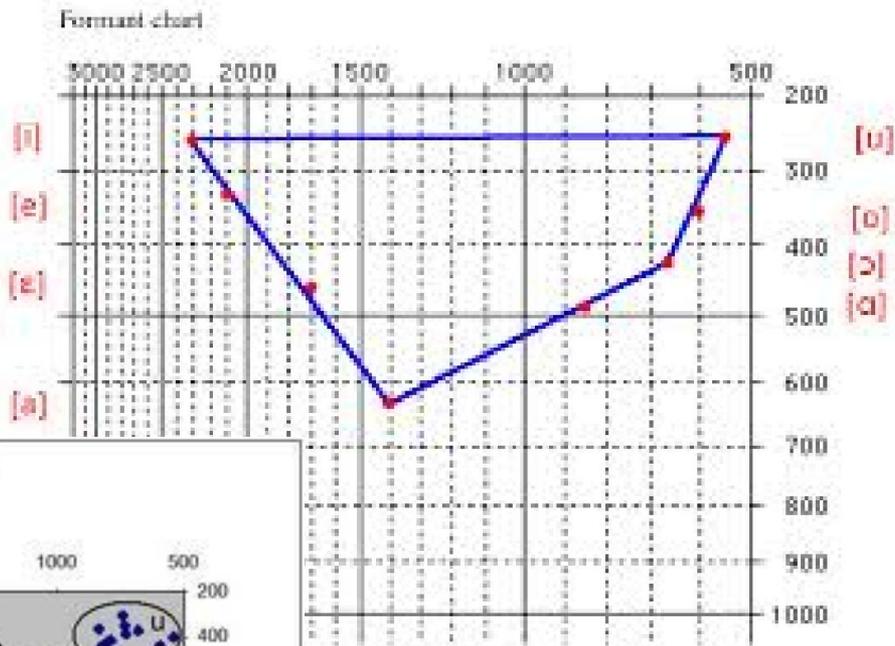
<http://fieldworks.sil.org/>

- PRAAT



- PRAAT

87



Инструменты

Tools Parser Window Help

Moksha (Latin)

Text

Title Mok LVJ_RI_12052013_о_Лесных_Сиялях
Rus

Info Baseline Gloss Analyze Tagging Print View Text Chart

1.2	Word	uʒɛl'ən'd'i	n'i	88 vel'ɛs'	конечно	oc'u	no	uʒɛl'ən'd'i	kizədə
	Morphemes	uʒɛl' -ən'd'i	n'i	vel'ɛ -s'	***	oc'u	no	uʒɛl' -ən'd'i	kizə -də
	Lex. Entries	uʒɛl' -n'd'i	n'i	vel'ɛ -s' ₁	***	oc'u	no	uʒɛl' -n'd'i	kizə -də ₁
	Lex. Gloss	жалъ DAT	уже	деревня DEF	***	большой	но	жалъ DAT	год ABL

kizəs	er'ɛj'n'ə		luvksnə			t'ɛ	vel'ɛt'	esə
kizə -s	er'ɛ -j	-n'ə	luv -ks	-nə		t'ɛ	vel'ɛ -t'	esə
kizə -s	er'ɛ -i ₂	-t'n'ə ₁	luv -ks ₂	-snə		t'ɛ	vel'ɛ -t' ₁	esə
год ILL	жить ПТСР.АКТ	DEF.PL	считать NMNLZR1	ЗПЛ.POSS	этот	деревня DEF.SG.GEN		в

kirij't

kir	-ij	-t'
kir	-i ₁	-t ₁

уменьшиться NPST.3 PL

<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  одного слова и т.п.

Инструменты

ELAN

The screenshot displays the ELAN 3.9.1 software interface. At the top, the menu bar includes File, Edit, Annotation, Tier, Type, Search, View, Options, Window, and Help. The main window is titled "Elan - pear story.eaf".

The interface is divided into several sections:

- Video View:** A window showing a man and a woman sitting and talking.
- Audio Recognizer:** A panel on the right with tabs for Grid, Text, Subtitles, Audio Recognizer, Video Recognizer, Metadata, and Controls. It shows the recognizer is set to "Tag vowels (volume peaks of voiced timespans)" and the file is "pear.wav". Parameters include Pitch ceiling [Hz] (604.8) and Intensity change [dB] (2.0).
- Timeline:** A central horizontal axis with time markers from 00:00:00.000 to 00:00:19.000. It features playback controls and a selection mode.
- Waveforms:** Below the timeline, there are waveforms for intensity (green) and pitch (blue).
- Annotation Tiers:** At the bottom, there are several tiers:
 - Event:** [1]
 - Clause Transcri:** [16] with segments for "and so he climbs up a tree and he starts with the ladder", "and he starts picking pears off the tree", and "and he puts the pears into an apron".
 - Motion:** [16] with segments for "motion" and "non-motion".
 - Gesture #:** [23] with segments for "gestu", "gesture 4", "gesture 6", "gesture 7", "gesture 9", and "gesture 1".
 - Go Hand:** [23] with segments for "R" and "B".

● ELAN

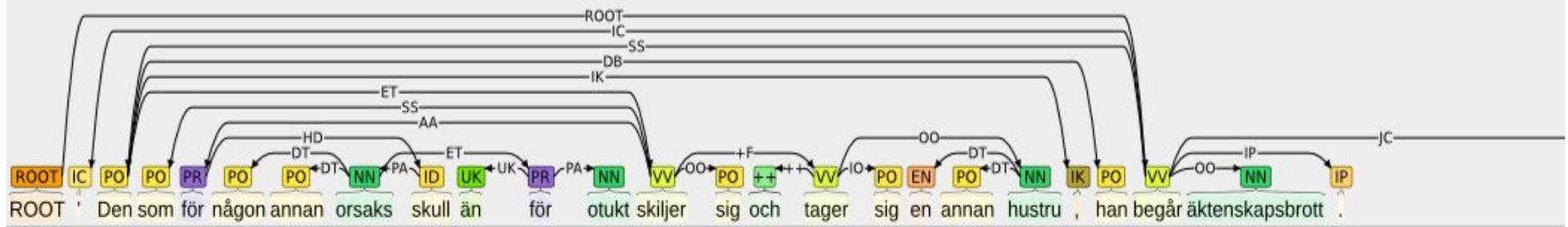
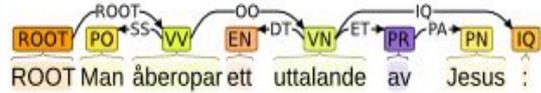
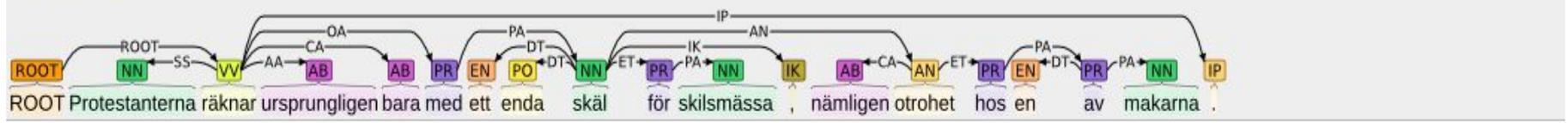
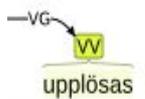
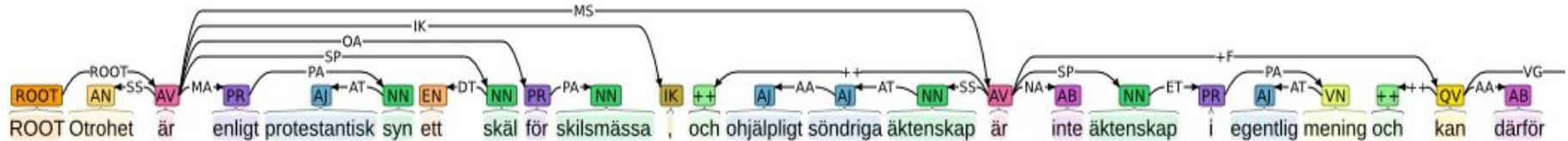


- ❑ синхронизация аннотации со звуком и видеорядом
- ❑ Возможность совмещения транскрипции с визуализацией параметров различного звукового ряда (спектрограммой...)
- ❑ многоуровневая аннотация
- ❑ возможность добавлять уровни аннотации пользователем
- ❑ возможность выгрузки в формате, доступном для автоматической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

Разметка корпуса

Разметка корпуса

<http://brat.nlplab.org/examples.html>



Разметка корпуса

NP Attributes

Group head(s)

Group

ЛШК...
prof...
ref
def
str
noun
type
coref

достоверно об обитателях дачи, но один знакомый сообщил мне, что, по слухам, там жи
 профессор Вагнер. Профессор Вагнер! Этого было достаточно, чтобы совершенно прикол
 внимание к даче. Мне во что бы то ни стало захотелось увидеть необычайного человека,
 наделавшего столько шума своими изобретениями. Но как? Я буквально стал шпионить
 Я чувствовал, что это было нехорошо, и все-таки продолжал свои наблюдения, целыми ч

Annotator

Маша Васильева

Ref	Group	ref	str	type
2:1443[16]	профессор Вагнер	def	noun	coref
5:5025[3]	ему	def	pron	coref
2:1595[21]	необычайного человека	def	noun	coref
2:1643[6]	своими	def	refl	coref
2:2179[65]	Высокий человек , с румяным лицом , русой бородой и н	def	noun	coref
2:2297[2]	он	def	pron	coref

Chain number

Chain

NP address in the text

NP

attributes

Краудсорсинг

[Разметка](#) / [именительный](#) — [винительный](#)

Спасибо, что помогаете нам. Не торопитесь, будьте внимательны. Если вы не уверены, пропускайте пример.

94

... только до момента пока **экономический** механизм начнет работать сам ...

[именительный](#)

[винительный](#)

[Другое](#)

[Пропустить](#)

[Прокомментировать](#)

... страны , где наблюдается **самый** низкий социоэкономический статус и ...

[именительный](#)

[винительный](#)

[Другое](#)

[Пропустить](#)

[Прокомментировать](#)

... А . Реформатский , **один** из создателей МФШ , ...

[именительный](#)

[винительный](#)

[Другое](#)

[Пропустить](#)

[Прокомментировать](#)

Геймификация

Sentinet Game

ужасное условие

negative

?

positive

95

[About](#)



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

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

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

- 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