



ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Компьютерная лингвистика

Лингвистические аспекты Подходы и задачи

Толдова С.Ю.

Компьютерная лингвистика

- Про что это?
- В широком смысле слова: компьютерные технологии и формальные модели анализа языка
- В чем специфика?
- Решения в рамках data mining?
- Решения в рамках формальных моделей в лингвистике?

Компьютерная лингвистика

- Компьютерная лингвистика
 - Лингвистические задачи в приложениях. Примеры
- 3 основных направления компьютерной лингвистики
 - Компьютерная лингвистика 1: электронные ресурсы и инструменты работы с языковыми данными
 - Компьютерная лингвистика 2: моделирования языка
 - Компьютерная лингвистика 3: инженерная лингвистика (обработка языка в различных приложениях)
- Задача компьютерной лингвистики 3. Автоматический анализ текста в приложениях
 - информационный поиск vs. извлечение информации из текста
 - анализ данных vs. извлечение знаний из текста
- Этапы лингвистической обработки
- Свойства языка: сложности при моделировании языковых явлений
- Примеры лингвистических платформ

Современная компьютерная ЛИНГВИСТИКА

- 3 разных подхода к трактовке термина «компьютерная лингвистика»:
 - инструментарий для обработки лингвистических данных
 - формальные модели (Computational linguistics);
 - современная автоматическая обработка естественного языка (Natural Language Processing)

Компьютерная лингвистика

- **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.
- **Инженерная компьютерная лингвистика:**
междисциплинарная область, в задачи которой входит автоматический анализ текстов

Компьютерная лингвистика

- Практическая работа
 - продукты, системы...
 - <https://docs.google.com/spreadsheets/d/1bM7Hw1YUX9BS2C-DwXw1ueXNJy1p6PTpQKpggA3AWsY/edit#gid=0>

Примеры

- Какие «конечные» задачи решает?
- Зачем и кому это нужно?
- Какие собственно лингвистические задачи приходится решать системе?

Компьютерная лингвистика

- Компьютерная лингвистика
 - Лингвистические задачи в приложениях. Примеры
- 3 основных направления компьютерной лингвистики
 - **Компьютерная лингвистика 1: электронные ресурсы и инструменты работы с языковыми данными**
 - Компьютерная лингвистика 2: моделирования языка
 - Компьютерная лингвистика 3: инженерная лингвистика (обработка языка в различных приложениях)
- Задача компьютерной лингвистики 3. Автоматический анализ текста в приложениях
 - информационный поиск vs. извлечение информации из текста
 - анализ данных vs. извлечение знаний из текста
- Этапы лингвистической обработки
- Свойства языка: сложности при моделировании языковых явлений
- Примеры лингвистических платформ

Компьютерная лингвистика 1. Ресурсы

Компьютерные словари

ABBYY® Lingvo

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

инструмент

Перевести

Русский

Английский

Переводы

Примеры (1058)

Словосочетания (1053)

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.

Компьютерная лингвистика 1. Ресурсы

THINKMAP®
VISUALTHESAURUS®

Look up a Word:

type a **cup** to search

LOOK IT UP

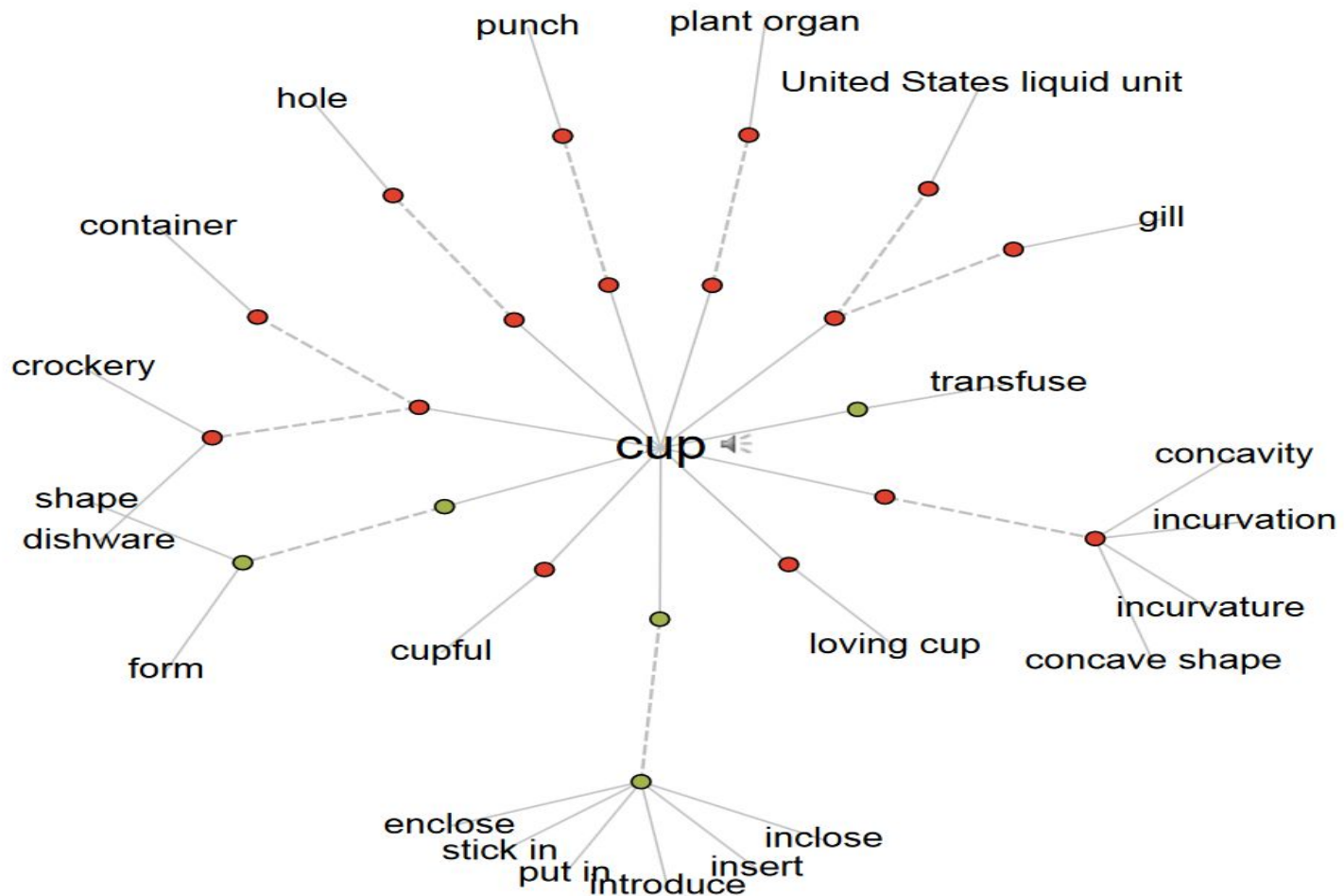
Search History Random Word Language:English

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	

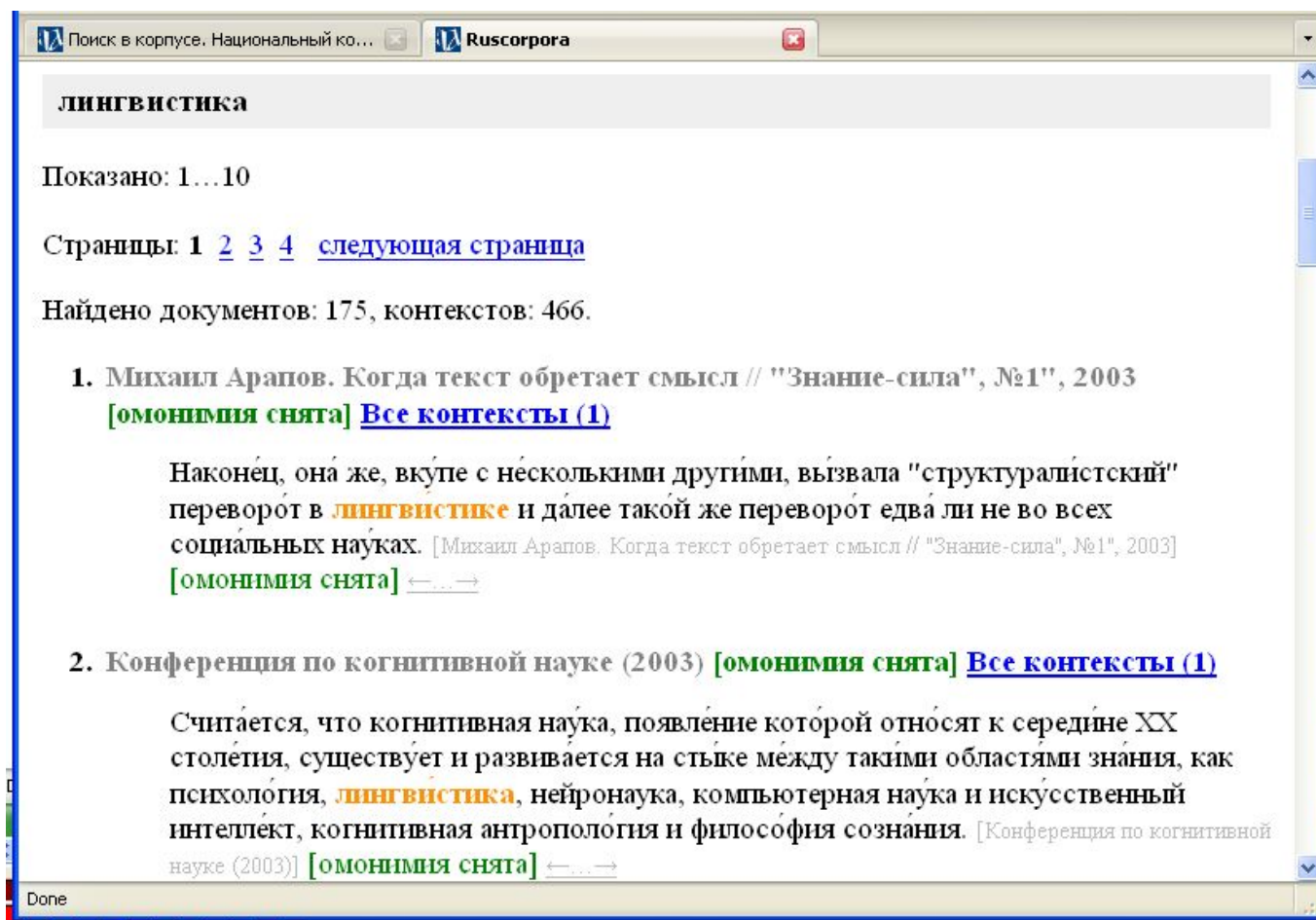
VERBS	ON OFF
form into the shape of a cup	
put into a cup	
treat by applying evacuated cups to the patient's skin	
introduce	
give shape or form to	
ADVERBS	ON OFF

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

Компьютерная лингвистика 1



Компьютерная лингвистика 1. Ресурсы



Поиск в корпусе. Национальный ко... Ruscorpora

ЛИНГВИСТИКА

Показано: 1...10

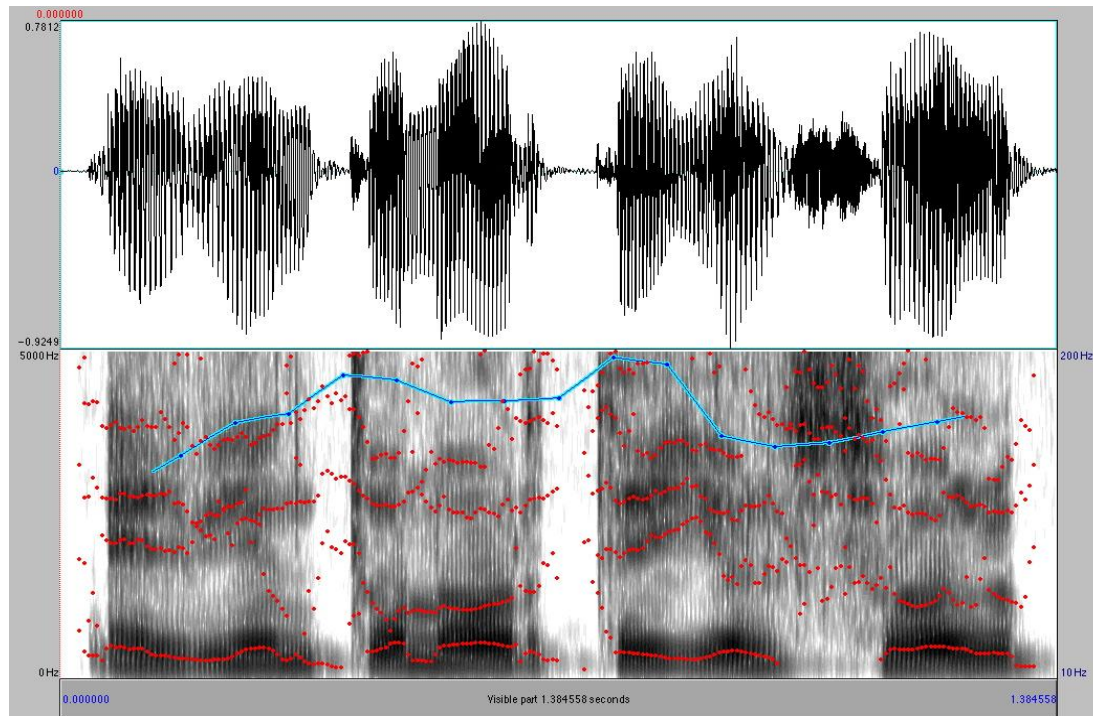
Страницы: [1](#) [2](#) [3](#) [4](#) [следующая страница](#)

Найдено документов: 175, контекстов: 466.

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

Done

Компьютерная лингвистика 1. Ресурсы



Компьютерная лингвистика 1.

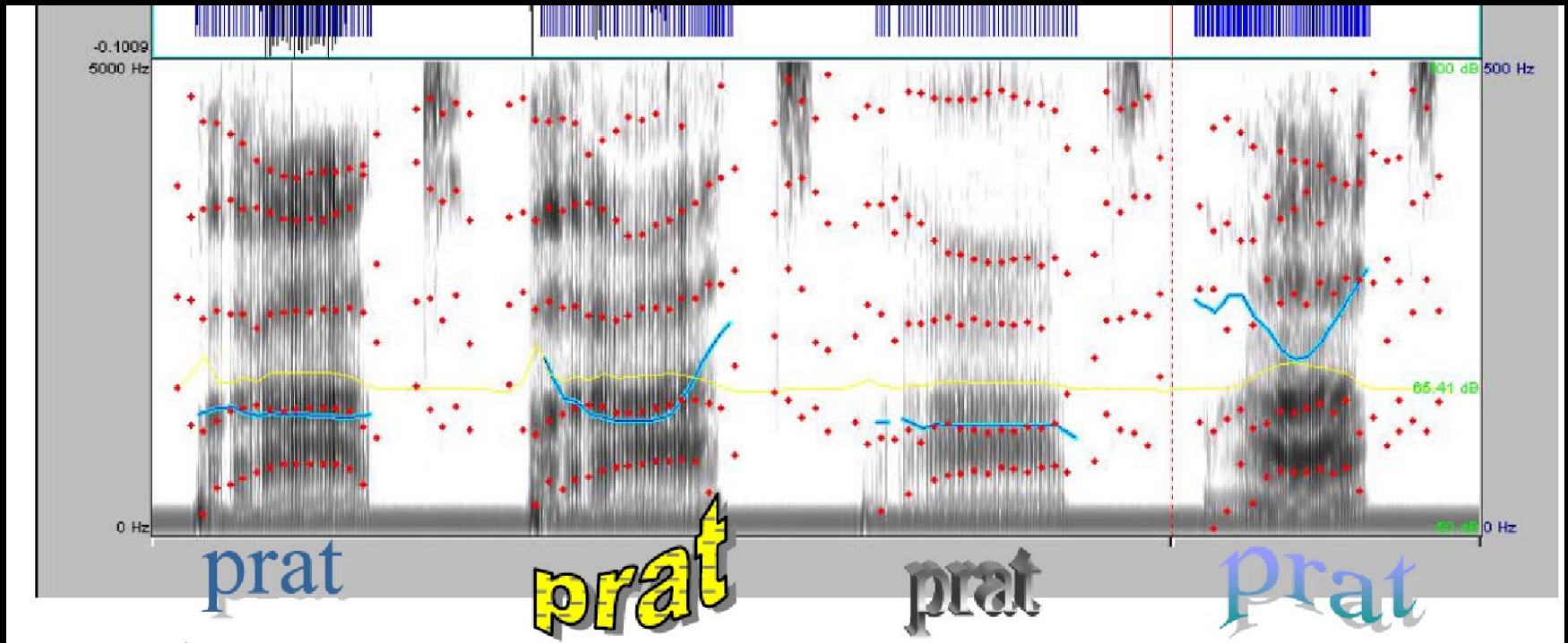
Инструменты

- 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.)

Компьютерная лингвистика 1

Инструменты

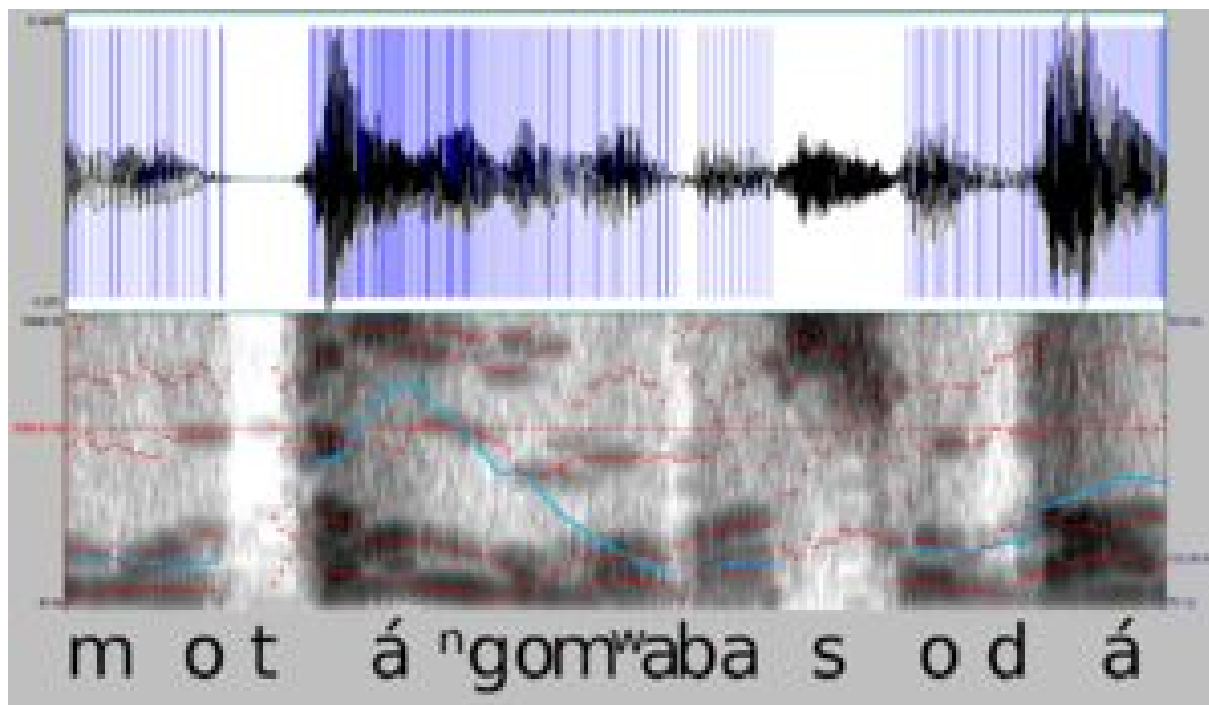
http://web.stanford.edu/dept/linguistics/corpora/material/PRAAT_workshop_manual_v421.pdf



Компьютерная лингвистика 1

Инструменты

- PRAAT

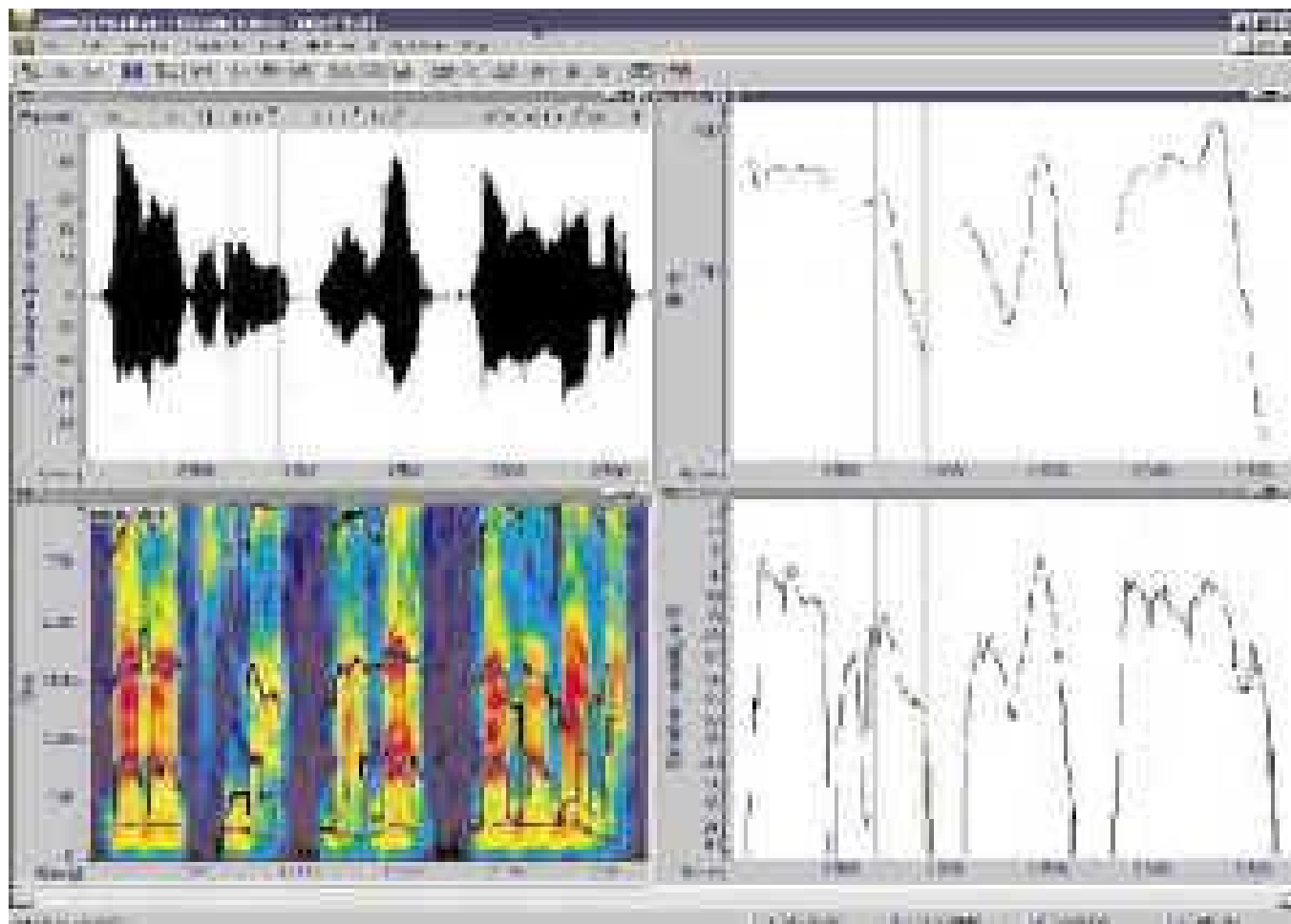


<http://fieldworks.sil.org/>

Компьютерная лингвистика 1

Инструменты

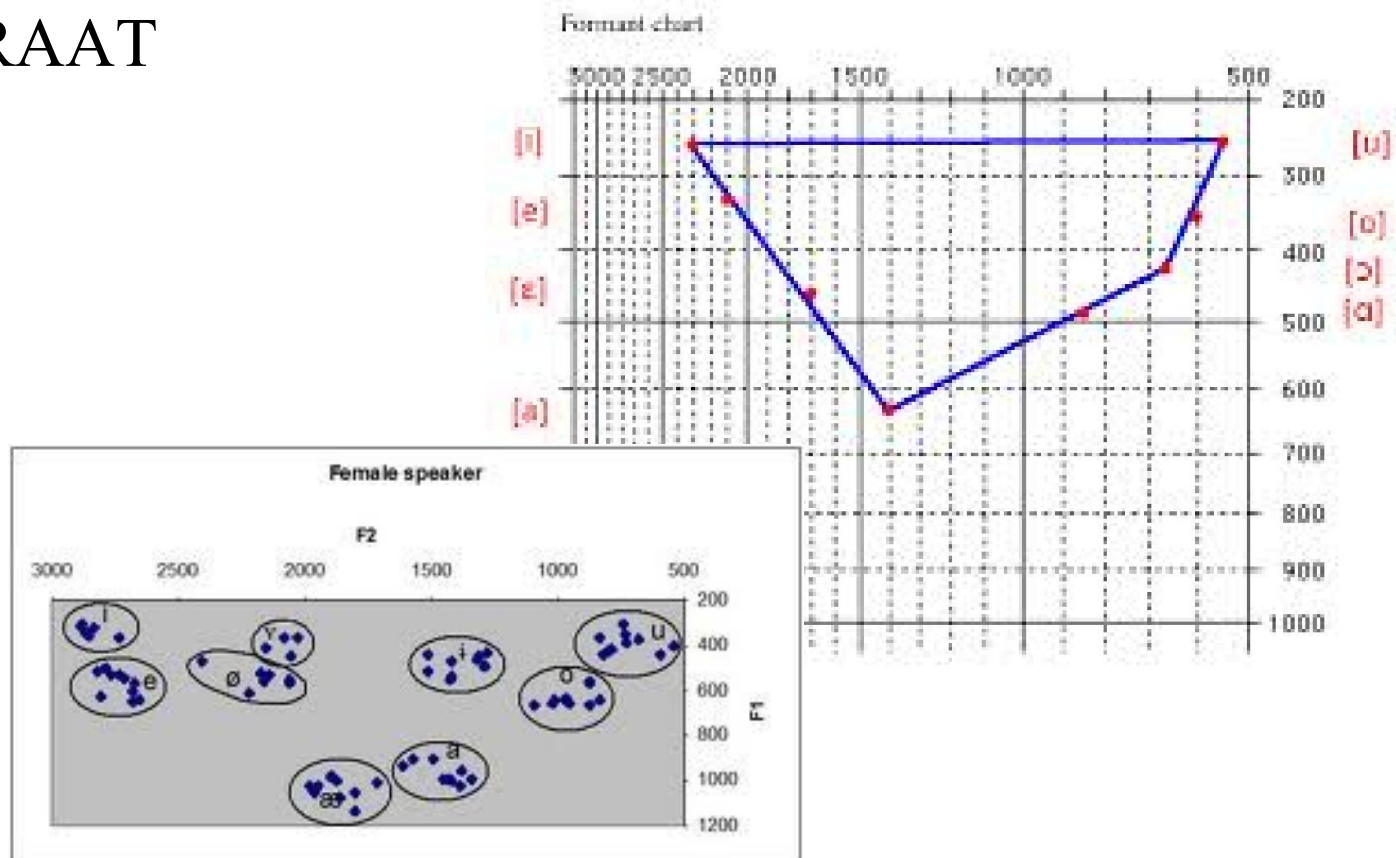
- PRAAT



Компьютерная лингвистика 1

Ресурсы и инструменты

- PRAAT



Компьютерная лингвистика 1

Инструменты

Tools Parser Window Help

Moksha (Latin)

Text

Title Mok LVJ_RI_12052013_o_Лесных_Сиялях

Rus

Info Baseline Gloss Analyze Tagging Print View Text Chart

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

kizəs	er'εjnə	luvksnə	t'ε	vel'et'	esə		
kizə -s	er'ε -j	-n'ə	luv -ks	-nə	t'ε vel'ε -t'	esə	
kizə -s	er'ε -i ₂	-t'n'ə ₁	luv -ks ₂	-snə	t'ε vel'ε -t' ₁	esə	
год ILL	жить PTCP.ACT	DEF.PL	считать NMNLZR1	3PL.POSS	этот	деревня DEF.SG.GEN	в

kirijt'

kir	-i ^o	-t'
kir	-i ₁	-t ₁
уменьшиться	NPST.3	PL

Компьютерная лингвистика 1

Ресурсы и 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

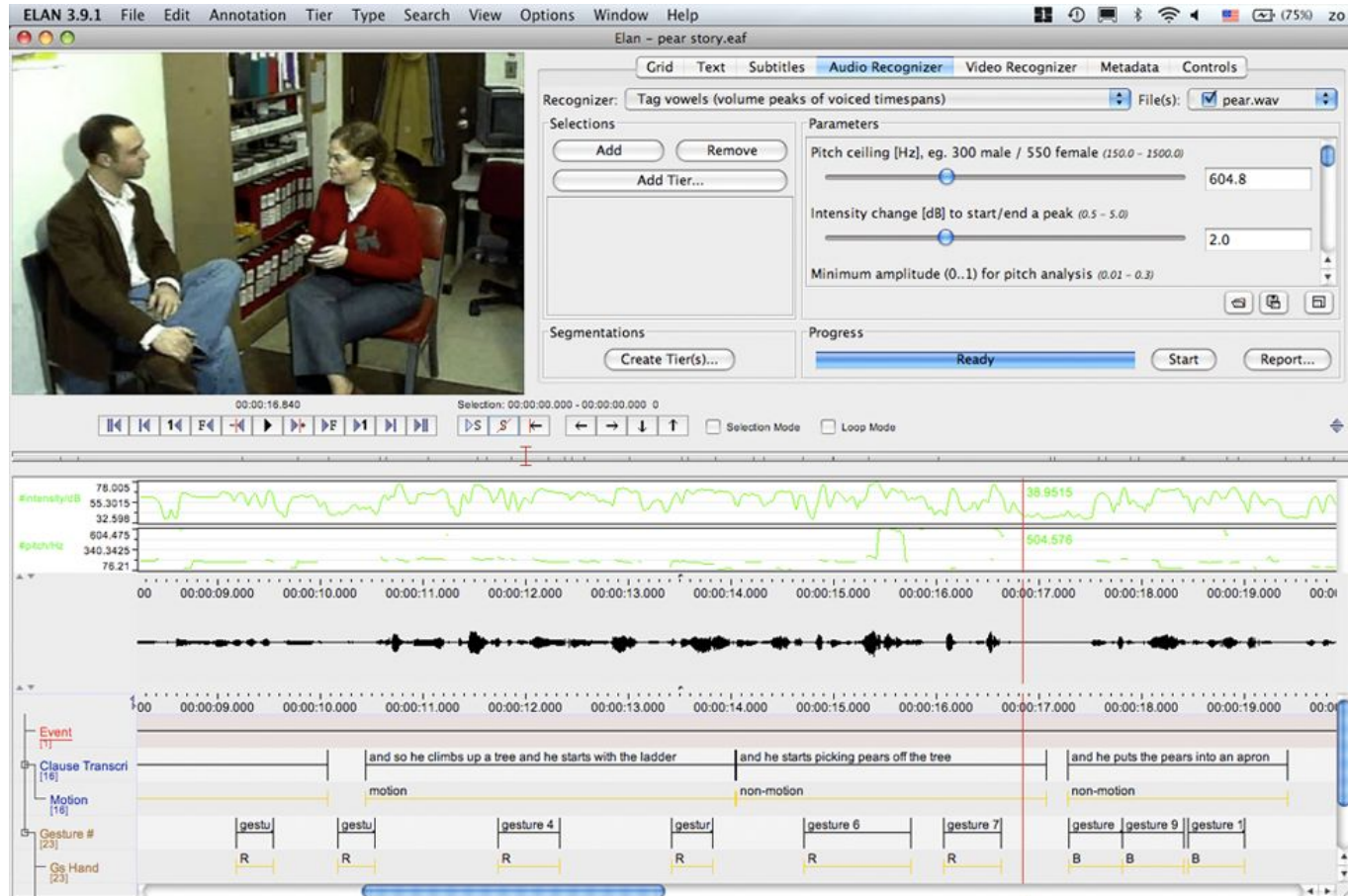
Поиск:

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

Компьютерная лингвистика 1

Инструменты

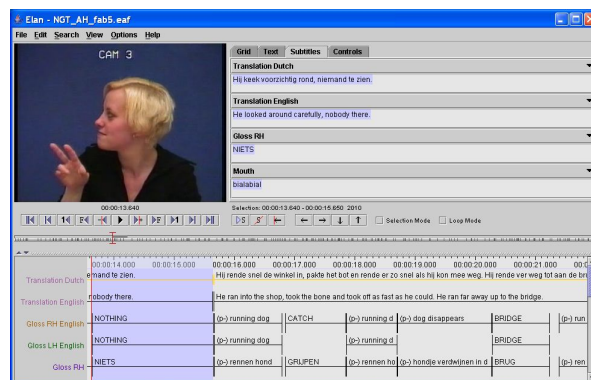
- ELAN



Компьютерная лингвистика 1

Инструменты

- ELAN

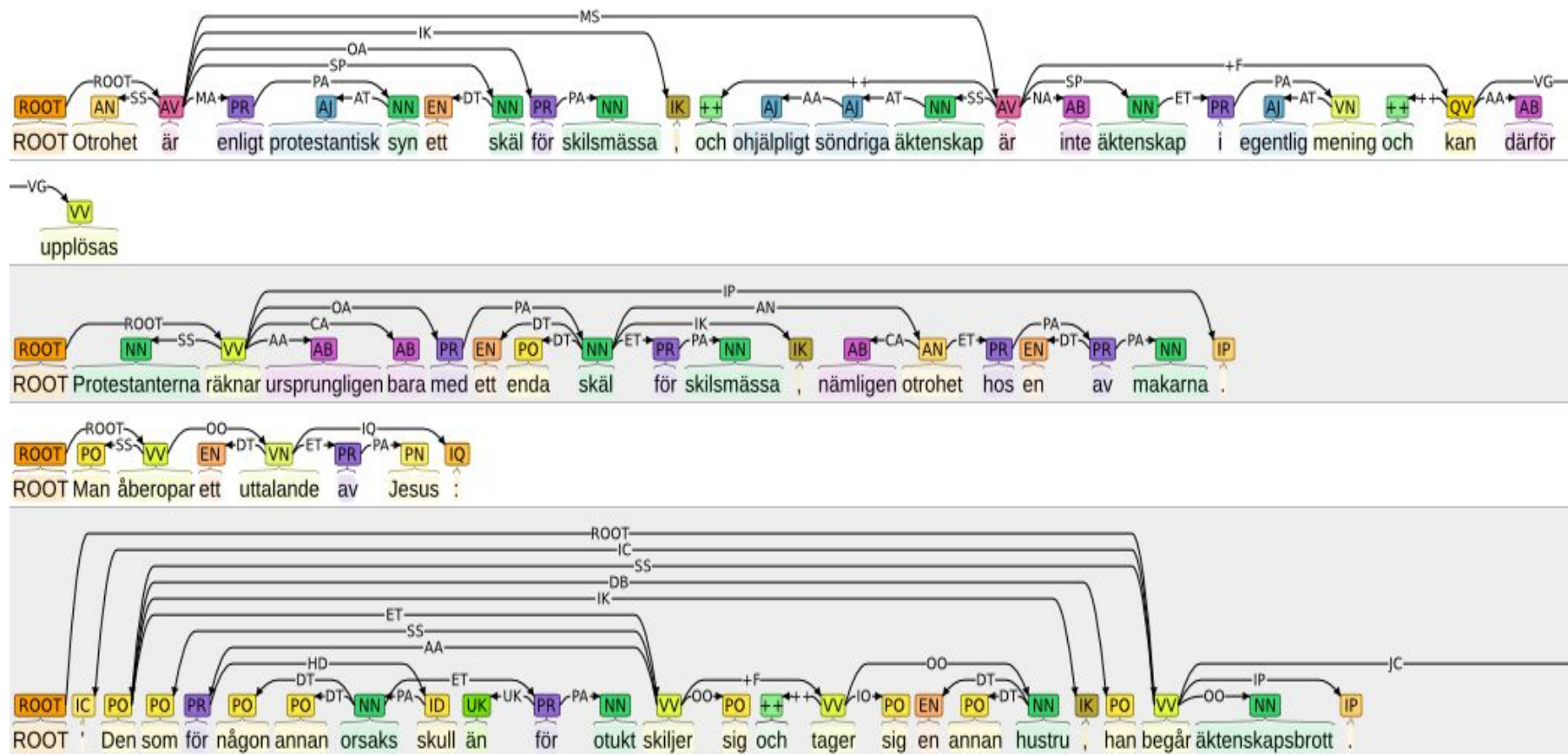


- ❑ display a speech and/or video signals, together with their annotations
- ❑ time linking of annotations to media streams
- ❑ linking of annotations to other annotations
- ❑ unlimited number of annotation tiers as defined by the users
- ❑ different character sets
- ❑ export as tab-delimited text files
- ❑ im- and export between ELAN and Shoebox search options.”

Компьютерная лингвистика 1

Инструменты

Разметка корпуса



Компьютерная лингвистика 1

Инструменты

NP Attributes

Group head(s)

Group

Annotator

LINK

properties

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

Chain
number

Chain

NP address in the text

NP

attributes

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

Краудсорсинг

[Разметка](#) / именительный — винительный

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

... только до момента пока **экономический** механизм начнет работать сам ...
 [Прокомментировать](#)

... страны , где наблюдается **самый** низкий социоэкономический статус и ...
 [Прокомментировать](#)

... А . Реформатский , **один** из создателей МФШ , ...
 [Прокомментировать](#)

Геймификация



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

Компьютерная лингвистика 1

- Сырые тексты -> корпуса
- Языковые знания -> специализированные базы данных
- Знания на уровне лексики -> словари и специализированные лексикографические ресурсы

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

Компьютерная лингвистика 1

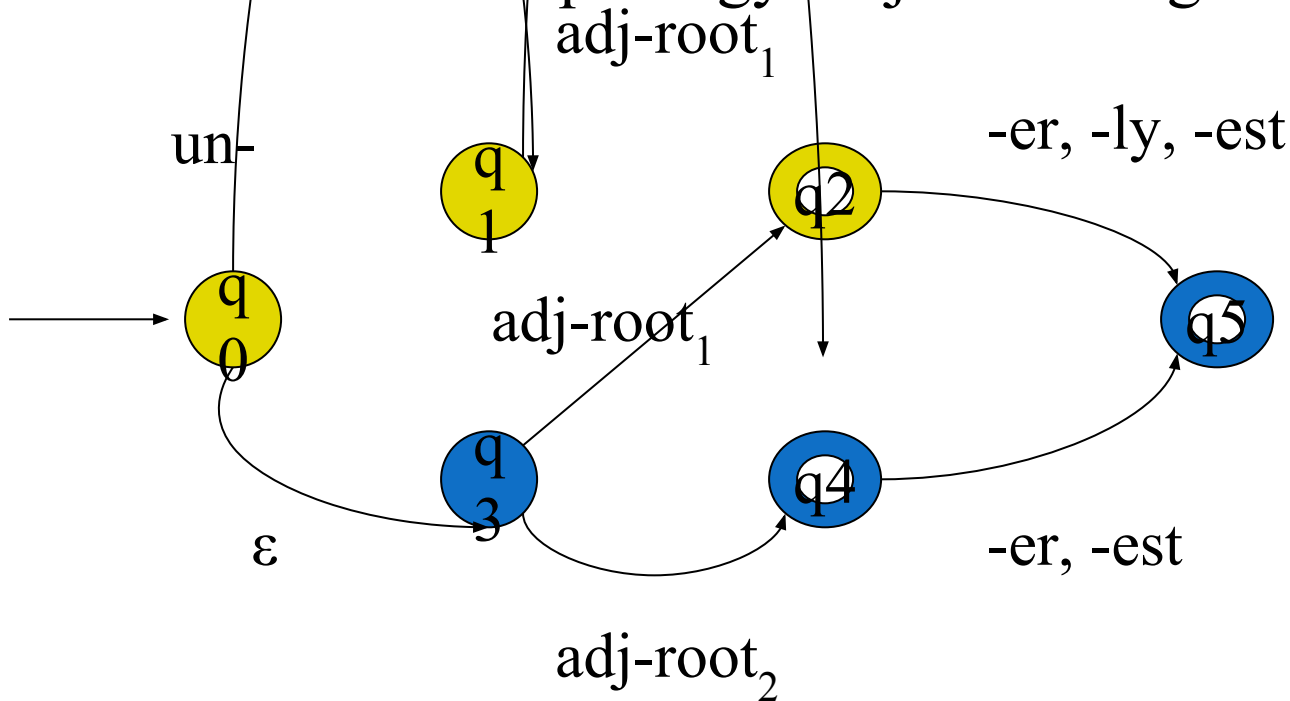
- Язык: сложная иерархическая система + big data
 - > инструменты поддержки работы с многоуровневыми иерархическими данными, имеющими специфическое статистическое распределение:
 - визуализация данных
 - корпусные менеджеры, обеспечивающие работу с многоуровневыми данными / большими данными
 - краудсорсинг
 - геймификация

Компьютерная лингвистика

- Компьютерная лингвистика
 - Лингвистические задачи в приложениях. Примеры
- 3 основных направления компьютерной лингвистики
 - Компьютерная лингвистика 1: электронные ресурсы и инструменты работы с языковыми данными
 - **Компьютерная лингвистика 2: моделирования языка**
 - Компьютерная лингвистика 3: инженерная лингвистика (обработка языка в различных приложениях)
- Задача компьютерной лингвистики 3. Автоматический анализ текста в приложениях
 - информационный поиск vs. извлечение информации из текста
 - анализ данных vs. извлечение знаний из текста
- Этапы лингвистической обработки
- Свойства языка: сложности при моделировании языковых явлений
- Примеры лингвистических платформ

Формальные модели

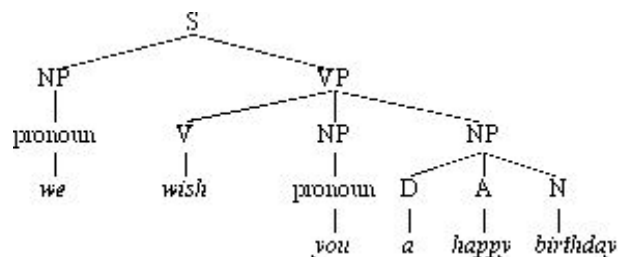
- Derivational morphology: adjective fragment



- Adj-root₁: clear, happy, real
- Adj-root₂: big, red

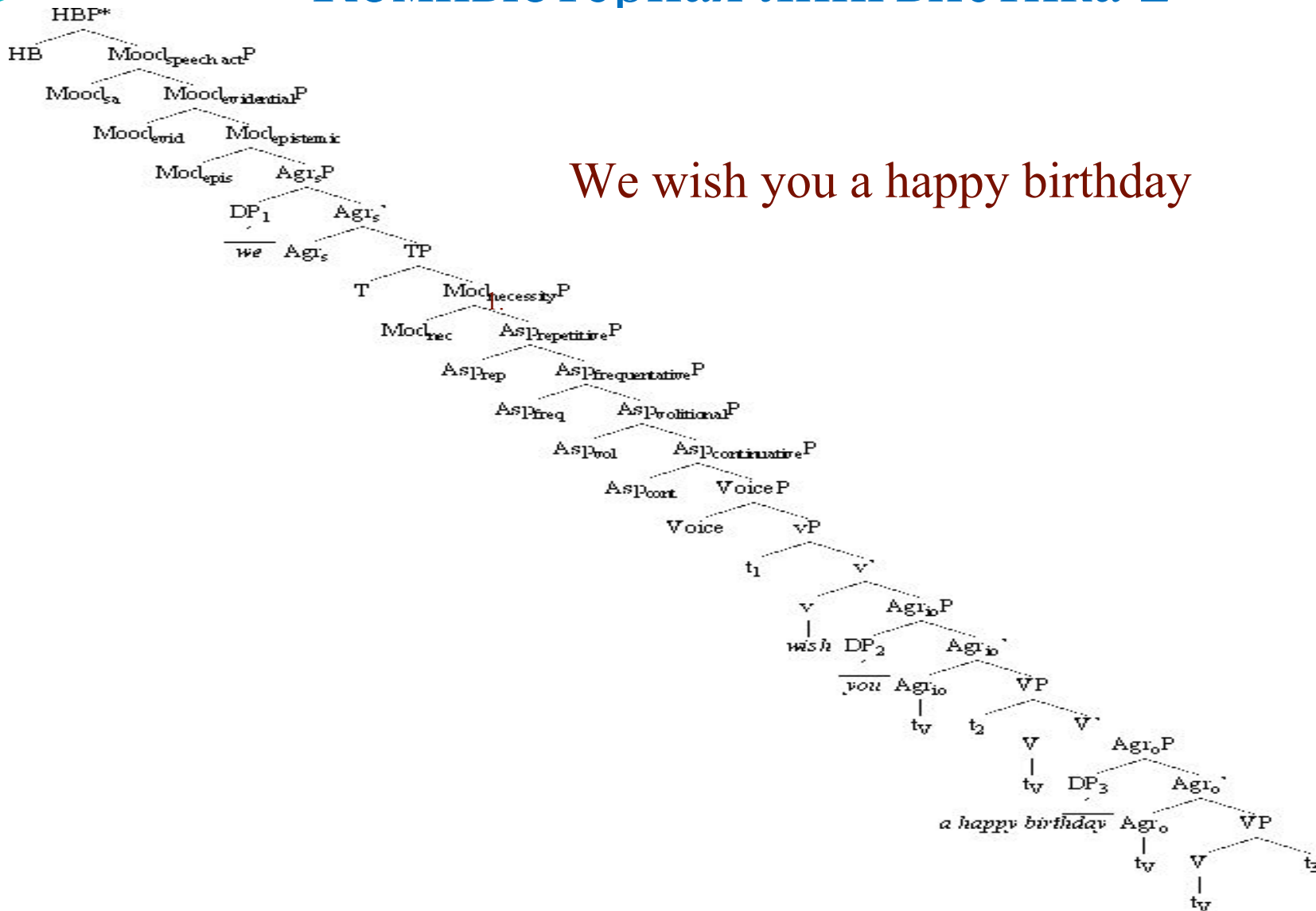
Компьютерная лингвистика 2

Формальные модели



Компьютерная лингвистика 2

We wish you a happy birthday



Компьютерная лингвистика 2

Формальные модели

Пример

S → NP VP

S → Aux NP VP

S → VP

NP → Pronoun

NP → Proper-Noun

NP → Det Nominal

Nominal → Noun

Nominal → Nominal Noun

Nominal → Nominal PP

VP → Verb

VP → Verb NP

VP → Verb NP PP

VP → Verb PP

VP → VP PP

PP → Preposition NP

Det → that | this | a

Noun → book | flight | meal | money

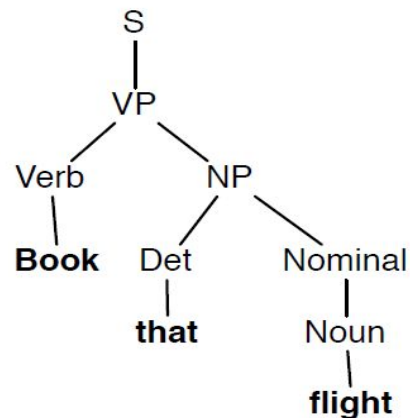
Verb → book | include | prefer

Pronoun → I | she | me

Proper-Noun → Houston | TWA

Aux → does

Preposition → from | to | on | near | through



Компьютерная лингвистика

- Компьютерная лингвистика
 - Лингвистические задачи в приложениях. Примеры
- 3 основных направления компьютерной лингвистики
 - Компьютерная лингвистика 1: электронные ресурсы и инструменты работы с языковыми данными
 - Компьютерная лингвистика 2: моделирования языка
 - Компьютерная лингвистика 3: инженерная лингвистика (обработка языка в различных приложениях)
- Задачи компьютерной лингвистики 3. Автоматический анализ текста в приложениях
 - информационный поиск vs. извлечение информации из текста
 - анализ данных vs. извлечение знаний из текста
- Этапы лингвистической обработки
- Свойства языка: сложности при моделировании языковых явлений
- Примеры лингвистических платформ

Компьютерная лингвистика 3

- Автоматический анализ текста
 - text mining
 - информационный поиск
 - извлечение знаний
 - общение с пользователем на естественном языке



- собственно лингвистическая обработка текста +
решение задач извлечения информации из текста

Задачи работы с контентом

- поиск нужной пользователю информации
⇒ информационный поиск
- агрегация информации (сбор информации на одну тему)
⇒ «оптимизированный» информационный поиск / систематизация текстовых коллекций
- анализ информации: автоматическое извлечение выводов и новых знаний на основе агрегированной информации
⇒ data mining, text mining

Компьютерная лингвистика 3

“мешок” задач

Занимается практическими задачами:

- 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.).
- Автофилترация (определение нежелательных документов: спам и т.п.)
- Вопросно-ответные системы и системы логического вывода.
- Системы извлечения знаний (Text Mining, Information Retrieval), мнений (Opinion Mining, Sentiment Analysis),

Компьютерная лингвистика 3

Основные задачи анализа контента

- **Обработка коллекций текстов:**
 - Группировка текстов / разделение текстов / похожие тексты
- **Задачи анализа контента:**
 - Найти тексты, похожие по смыслу, стилю, тематике
- **Анализ текста:**
 - Извлечение/выделение фрагментов текста; извлечение онтологических элементов (элементов знаний); преобразование неструктурированных данных в структурированные
- **Задача анализа контента:**
 - Извлечение информации определенного типа из текста

Компьютерная лингвистика 3

Основные задачи анализа контента

- **Обработка коллекций текстов:**
 - Группировка текстов / разделение текстов
- Задачи анализа контента:
 - Найти тексты, похожие по смыслу, стилю, тематике
- Задачи обработки текстов:
 - Найти тексты, похожие на некоторый текст (например, 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

Основные задачи анализа контента

- **Анализ текста:**

- Извлечение/выделение фрагментов текста; извлечение онтологических элементов (элементов знаний); преобразование неструктурированных данных в структурированные

- **Задача анализа контента:**

- Извлечение событий, их участников, места, времени, последовательности событий, отношений
- Извлечение оценки событий, объектов, мнений
- Семантическое аннотирование
- Извлечение онтологических знаний

(Named Entities Recognition (Instances Extraction), Fact Extraction, Relation Extraction, Semantic Annotation, Ontological Information Extraction)

Введение в компьютерную ЛИНГВИСТИКУ

- **Три направления:**
 - **Ресурсы и инструменты для изучения языка**
 - корпуса, лексикографические ресурсы (словари, тезаурусы), специальные программы для работы с корпусами, обработки звука
 - **Формальные модели**
 - Двухуровневая морфология (конечные преобразователи); формальный синтаксис (грамматики и т.п.)
 - **Автоматическая обработка текста**
 - обработка коллекций текстов: информационный поиск, рубрикация, новостная и др. агрегация;
 - извлечение информации: извлечение именованных сущностей, извлечение фактов, извлечение мнений;
 - др.

Введение в компьютерную лингвистику.

Немного терминологии

Корпуса	Corpora
Аннотирование корпуса	Corpus annotation (tagging)
Автоматическая обработка ЕЯ (текста) / АОТ	Natural language processing (NLP)
Информационный поиск	Information retrieval (IR)
Извлечение информации из текста	Information extraction (IE)
Извлечение именованных сущностей	Named entities extraction (NER)
Извлечение фактов	Fact extraction
Тематическая группировка текстов (анализ новостного потока)	Topic detection and tracking (TDT)
Автоматическая рубрикация текстов	Text classification