

[illegible]

заклучительные

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА
(COMPUTATIONAL LINGUISTICS),

или

**ЧЕМ ЗАНИМАЮТСЯ
КОМПЬЮТЕРНЫЕ ЛИНГВИСТЫ**

ВЫБОРНОВА АННА, ANNA@179.RU

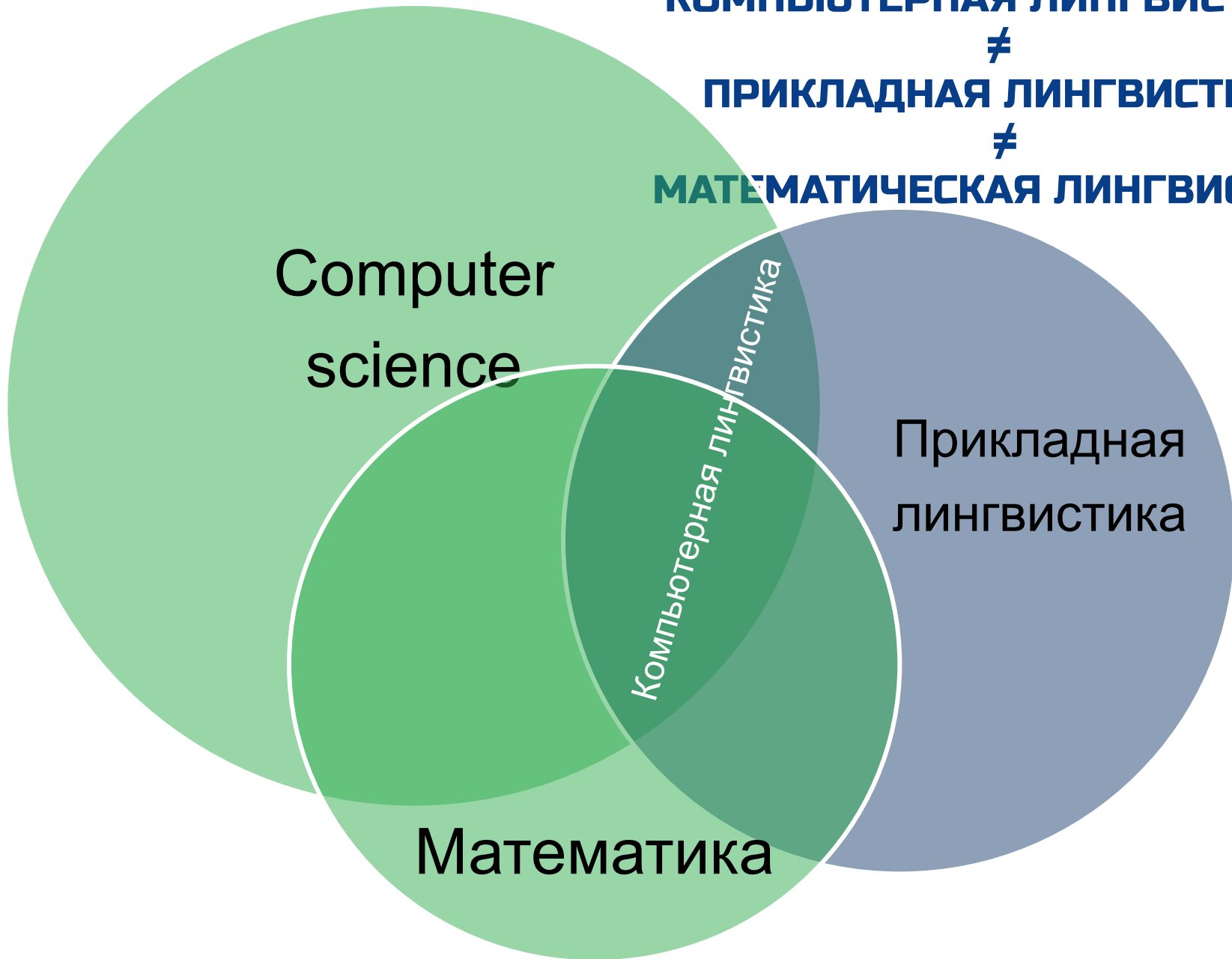
КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

≠

ПРИКЛАДНАЯ ЛИНГВИСТИКА

≠

МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА



КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

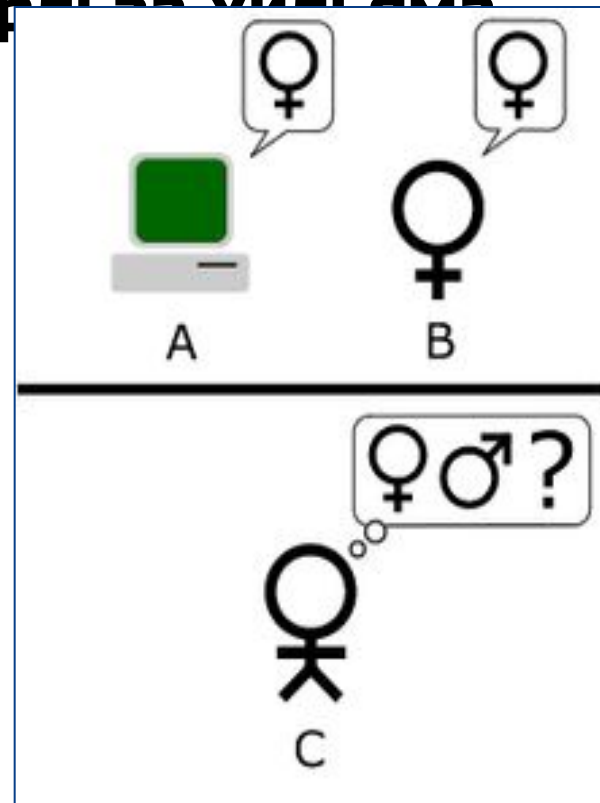
это направление в прикладной лингвистике, ориентированное на использование компьютерных технологий и математического аппарата для обработки данных на естественном языке и о естественном языке.

ИСТОРИЯ

- Работы в области логики, семиотики и прагматики американских философов Чарльза Сандерса Пирса (1839— 1914) и Чарльза Уиллса Морриса (1901-1979)
- Появление компьютера
- Первые языки программирования
- Эксперименты с машинным переводом

Since computers can make arithmetic calculations much faster and more accurately than humans, it was thought to be only

- ~~Важной задачей было~~ **Разработка методов тестирования интеллекта и тестирования** ~~Тестирование~~ details could be taken



НАПРАВЛЕНИЯ В КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

- **Обработка естественного языка** (natural language processing)
- **Анализ экстралингвистических данных**
- **Корпусная лингвистика**
- **Создание электронных словарей, тезаурусов, лингвистических онтологий**
- **Информационный поиск**
- **Машинный перевод**
- **Автоматическая проверка грамотности (спеллчекеры)**
- **Автореферирование, порождение текстов, аннотирование**
- **Определение тональности текста**
- **Построение систем управления знаниями (онтологии, экспертные системы)**
- **Оптическое распознавание символов**
- **Автоматическое распознавание речи**
- **Digital Humanities** (автоматические подсказки, социальные сети)
- **Автоматический синтез речи**
- ***Создание диалоговых систем***
- ***Искусственный интеллект***
- **Нейролингвистика***

ОСНОВНЫЕ ТИПЫ РАБОТЫ С ДАННЫМИ

ОБРАБОТКА и АНАЛИЗ

- понимание языка (Mystem)

СИНТЕЗ

- генерация грамотного текста (Siri)

АНАЛИЗ

В тЕ|Эгах Е|Э(-)мЭ|Ейла В|Уильяма В|Уорфа прочла
про флЕ|Эшку, пуС|ССеТ|ТТ(У) и снуД|Т.

Хотя, конечно, это не 9000 способов написать «Муаммар Каддафи» по-английски

Качество понимания зависит от множества факторов: от языка, от национальной культуры, от самого собеседника и т. д. Вот некоторые примеры сложностей, с которыми сталкиваются системы понимания текстов.

- Сложности с раскрытием анафор
«Мы отдали бананы обезьянам, потому что они были голодные» и «Мы отдали бананы обезьянам, потому что они были перезрелые»
- Свободный порядок
«Бытие определяет сознание»
- В русском языке свободный порядок компенсируется развитой морфологией, служебными словами и знаками препинания, но в большинстве случаев для компьютера это представляет дополнительную проблему.
- В речи могут встретиться неологизмы. Система должна уметь отличать такие случаи от опечаток и правильно их понимать.
глагол «Пятидесятиублируй»
- Правильное понимание омонимов. При распознавании речи также возникает проблема фонетических омонимов.
«Серый волк в глухом лесу встретил рыжую лису»

ЕЩЕ НЕДАВНО

СТЕММАТИЗАЦИЯ И ЛЕММАТИЗАЦИЯ

Падеж	Ед. ч.	Мн. ч.
Именительный	лев	левы
Родительный	лева	левов
Дательный	леву	левам
Винительный	лев	левы
Творительный	левом	левами
Предложный	леве	левах

Морфологическая омонимия:

И?
Дел?
Полезное?

АОТ	словарный	русский, английский
Lemmatizer	словарный	русский, английский
MYSTEM (Yandex)		

ЕЩЕ НЕДАВНО WSD

Яндекс

Картинки буханка



Найти

[Назад в поиск](#)

Любой размер ▾

Свежие

Обои 1366×768

Любая ориентация ▾

Любой тип ▾

Любой цвет



ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА (NATURAL LANGUAGE PROCESSING)

Методы и инструменты	Ресурсы
Обработка текста, обработка речи, утилиты, методы, алгоритмы.	Словари, тезаурусы, корпусы, банки данных.

- Графематический анализ
- Морфологический анализ
- Синтаксический анализ
- Расстановка переносов
- Построение конкордансов
- Извлечение ключевых слов
- Анафорический анализ
- Кластеризация данных
- Извлечение именованных сущностей
- Извлечение фактов
- Извлечение отношений
- Анализ тональности
- и др.

ЛЕММАТИЗАЦИЯ И ЧАСТОТНЫЙ АНАЛИЗ РУССКОЙ

Лемма	Примеры	2012	2013
блог	должен получить блог Матвиенко Вы что она вообще не ведёт блога А разве это имеет значение html Даже Навальный откомментировал Вот	8063	2143
фейсбук	было ему вести себя адекватно Премьер кажется встречался с гуру фейсбука ДО ректора А священнику респект Пишет блогер Исаак Ньютон Я	318	409
википедия	событиями я вознамерилась купить иконку святого Власия который как повествует википедия считается покровителем животных так как предание гласит что он	178	196
фанфик	что курил автор Между тем выше вовсе не текст очередного фанфика на тему генерал Власов агент Кремля а подлинный американский	173	9
юзер	люблю слово то давеча изобретенное в приватной беседе на ходу юзером chka тоталирантность Чем дальше тем больше пользую его Да	103	82
ютуб	с утра про это шоу ну я и подзавис на ютубе вспоминая старые добрые выпуски Такое ностальжи Вообще под катом собрал	93	110
флешмоб	они меня прежде не любили Сегодня в соцсетях проходит уникальный флешмоб поздравь Путина так чтобы было понятно что на самом деле	91	82
репост	летний Эдик Игнатян должен ответить по закону за всё Максимальный репост максимальная огласка Или опять ведь замнут гады Тележурналист	88	218
смайлик	друг другу дифирамбы Голоса разума в очередной раз потонули среди смайликов бьющихся головой о стену и пожеланий коту бежать по радуге	87	39
порносайт	кумир помреж Галина Стремительная в дырявых джинсах На датских порносайтах Галину пытались воссоздать в рубрике Строгая Госпожа Всё напрасно Датчане	85	8
апдейт	вариант Агатовая Тяжеленькая красивая правда Ну что пишу апдейт вот на эту тему теперь уж для всех Папа очень	77	48
хипстер	с офисными сотрудниками неформалы с людьми в национальных костюмах очкастые хипстеры со масонами Очень интересно просто смотреть на прохожих 11 Погода	59	68
фрик	что ли Чт 38 В последней четверке дебатыКС ни одного фрика Закончили оптимистично Чт 43 А тут через час другие	55	85
флэшмоб	А уж если обращение возымеют действие то волной пойдут прикольные флэшмобы etc А дальше Гельман на очереди Он чуть не каждый	54	3
эпичный	в руке врывались в Музей эротического искусства Мне кажется довольно эпичный разговор Девственность это проблема государственной Православные готовы союзничать с	51	21

TF-IDF (от англ. TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

СИНТАКСИЧЕСКИЙ АНАЛИЗ

- **Линейный порядок слов:**

Ко мне пришли гости.

*Ко пришли мне гости.

- **Ограничения на сочетаемость:**

Его арестовали.

Он был арестован милиционером.

*Его арестовали милиционером.

Начинает светать.

*Хочет светать.

??Собирается светать.

Структура зависимостей



- однонаправленность связей
- уникальность вершины (каждый узел зависит лишь от одного узла)
- метки зависимостей могут различаться

ЗНАТЬ, ЧТОБЫ РАЗБИРАТЬ

Свободный порядок слов создает сложности для синтаксического анализа предложения:

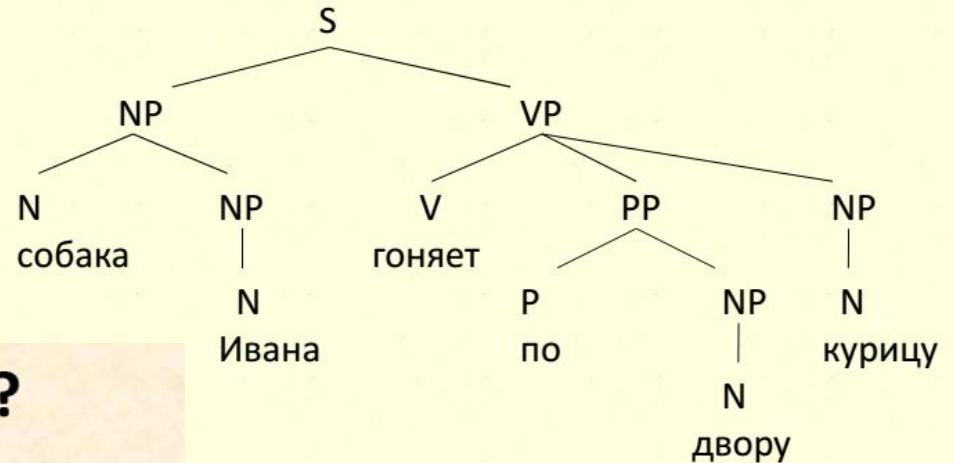
- Сегодня на фестивале языков мы поговорим о некоторых чертах русского языка, приводящих в недоумение математиков и программистов.
- Мы поговорим о некоторых чертах русского языка, приводящих в недоумение математиков и программистов, сегодня на фестивале языков.
- О некоторых чертах русского языка, приводящих в недоумение математиков и программистов, мы поговорим сегодня на фестивале языков.
- О некоторых приводящих в недоумение математиков и программистов чертах русского языка мы поговорим на фестивале языков сегодня.
- О приводящих недоумение некоторых в и чертах русского мы русского поговорим языка сегодня математиков фестивале программистов языков на.
- Сегодня на фестивале языков мы поговорим приводящих в недоумение математиков и программистов русского языка о некоторых чертах.

ПОРЯДОК СЛОВ УСЛОВНО СВОБОДНЫЙ!

-> ОЧЕНЬ СЛОЖНО СОЗДАТЬ АДЕКВАТНЫЕ СИНТАКСИЧЕСКИЕ ПАРСЕРЫ.

СИНТАКСИЧЕСКИЙ АНАЛИЗ

Альтернатива: структура составляющих



Как изучать синтаксис?

- Корпусные данные
- Суждения носителей языка о приемлемости высказываний («языковая интуиция»)
 - имеют статус экспериментальных данных (результат воспроизводим в четко определенных условиях)
 - как всякие экспериментальные данные, чувствительны к условиям проведения эксперимента

СИНТАГРУС
TOMITA parser
TreeBank

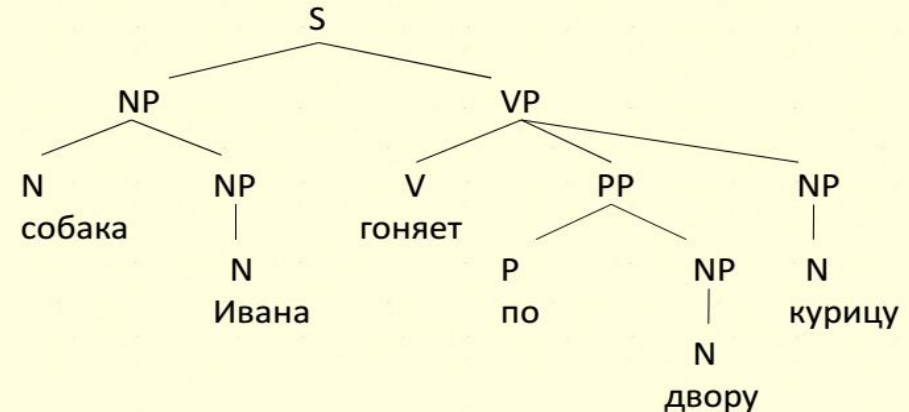
СИНТАКСИЧЕСКИЙ АНАЛИЗ

Как изучать синтаксис?

- Корпусные данные
- Суждения носителей языка о приемлемости высказываний («языковая интуиция»)
 - имеют статус экспериментальных данных (результат воспроизводим в четко определенных условиях)
 - как всякие экспериментальные данные, чувствительны к условиям проведения эксперимента

<http://nlpub.ru>

Альтернатива: структура составляющих



- Типологические данные (универсальные правила и запреты)

Я съел суп и макароны.

Я съел что? => Что я съел?

Я съел суп и что? => * Что я съел суп и ?

ИЗВЛЕЧЕНИЕ ФАКТОВ

FACTS EXTRACTION

Corpus: Corpus_news

Key set: Key

Response set: [Default set]

Types: [Location, Organization, Person]

Features: []

Annotation	Match	Only A	Only B	Overlap	Prec.B/A	Rec.B/A	F1.0-s.	Prec.B/A	Rec.B/A	F1.0-l.	Prec.B/A	Rec.B/A	F1.0-a.
Location	25	5	5	4	0.74	0.74	0.74	0.85	0.85	0.85	0.79	0.79	0.79
Organization	48	19	2	8	0.83	0.64	0.72	0.97	0.75	0.84	0.90	0.69	0.78
Person	34	14	5	5	0.77	0.64	0.70	0.89	0.74	0.80	0.83	0.69	0.75
Macro summary					0.78	0.67	0.72	0.90	0.78	0.83	0.84	0.72	0.77
Micro summary	107	38	12	17	0.79	0.66	0.72	0.91	0.77	0.83	0.85	0.71	0.78

Document	Match	Only A	Only B	Overlap	Prec.B/A	Rec.B/A	F1.0-s.	Prec.B/A	Rec.B/A	F1.0-l.	Prec.B/A	Rec.B/A	F1.0-a.
buyout_DELL.txt_00065	13	2	2	2	0.76	0.76	0.76	0.88	0.88	0.88	0.82	0.82	0.82
Chinacollege_cyber-spying.txt_0006E	5	4	2	5	0.42	0.36	0.38	0.83	0.71	0.77	0.62	0.54	0.58
EU_monitors_appledistribution.txt_0007B	11	3	1	2	0.79	0.69	0.73	0.93	0.81	0.87	0.86	0.75	0.80
Grandfather_ofAfricanLiterature.txt_00084	6	10	1	1	0.75	0.35	0.48	0.88	0.41	0.56	0.81	0.38	0.52
Kypr_meltdown.txt_0008D	17	2	1	3	0.81	0.77	0.79	0.95	0.91	0.93	0.88	0.84	0.86
meet_of_Popes.txt_00096	8	6	2	2	0.67	0.50	0.57	0.83	0.62	0.71	0.75	0.56	0.64
Microsoft_vs_Google.txt_0009F	13	3	1	2	0.81	0.72	0.76	0.94	0.83	0.88	0.88	0.78	0.82
Old_Uni.txt_000A8	4	0	0	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Putin_Berezovskij.txt_000B1	26	5	1	0	0.96	0.84	0.90	0.96	0.84	0.90	0.96	0.84	0.90
Ubuntu_in_China.txt_000BE	4	3	1	0	0.80	0.57	0.67	0.80	0.57	0.67	0.80	0.57	0.67
Macro summary					0.78	0.66	0.70	0.90	0.76	0.82	0.84	0.71	0.76

ИЗВЛЕЧЕНИЕ ФАКТОВ

FACTS EXTRACTION

KEEP UP **ON** YOUR **READING** WITH AUDIO **BOOKS**

Vietnam

UK

Louisiana, USA

Audio **books** are highly **popular** with **library** patrons in the **town**

Louisiana, USA

S. Carolina, USA

Pennsylvania, USA

Mass., USA

of **Springfield,** **Greene** County, **MO.** "People are **mobile**

Turkey

Virginia, USA

Maine, USA

Norway

Alabama, USA

and busier, and audio **books** fit into that lifestyle" says **Gary**

Louisiana, USA

Indiana, USA

Sanchez, who oversees the **library's** \$2 **million** budget...

Dominican Republic

Pennsylvania, USA

Kentucky, USA

ИЗВЛЕЧЕНИЕ ФАКТОВ

FACTS EXTRACTION

GATE Developer 7.1 build 4485

File Options Tools Help



Language Resources

Ubuntu_in_China.bt_0001

Putin_Berezovskij.bt_000

Old_Uni.bt_000A8

Microsoft_vs_Google.bt_0

meet_of_Popes.txt_00096

Kypr_meltdown.bt_00080

Grandfather_ofAfricanLite

EU_monitors_appliedistrib

Chinacollege_cyber-spyr

Corpus_news

buyout_DELL.bt_00065

Processing Resources

A a

A ANNIE OrthoMatcher

NE ANNIE NE Transducer

Resource Features

Corpus editor

Initialisation Parameters

Corpus Quality Assurance

Messages

buyout_DELL.bt...

Chinacollege_cy...

EU_monitors_app...

Grandfather_ofA...

Kypr_meltdown.t...

meet_of_Popes.t...

Microsoft_vs_Go...

Old_Uni.bt_000...

Putin_Berezovsk...

Corpus_news

Ubuntu_in_China...

Corpus statistics

Document statistics

Annotation	Match	Only A	Only B	Overlap	Prec.B/A	Rec.B/A	F1.0-s	Prec.B/A	Rec.B/A	F1.0-l	Prec.B/A	Rec.B/A	F1.0-a
Location	25	5	5	4	0.74	0.74	0.74	0.85	0.85	0.85	0.79	0.79	0.79
Organization	48	19	2	8	0.83	0.64	0.72	0.97	0.75	0.84	0.90	0.69	0.78
Person	34	14	5	5	0.77	0.64	0.70	0.89	0.74	0.80	0.83	0.69	0.75
Macro summary					0.78	0.67	0.72	0.90	0.78	0.83	0.84	0.72	0.77
Micro summary	107	38	12	17	0.79	0.66	0.72	0.91	0.77	0.83	0.85	0.71	0.78

☐ present in every document

Annotation Types

Temp

Title

Token

Unknown

☐ present in every selected document

Annotation Features

gender

kind

locType

matchedWithLonger

☐ present in every selected document

Measures

F-Score

Classification

F1.0-score strict

F1.0-score lenient

F1.0-score average

F1.0-score strict BDM

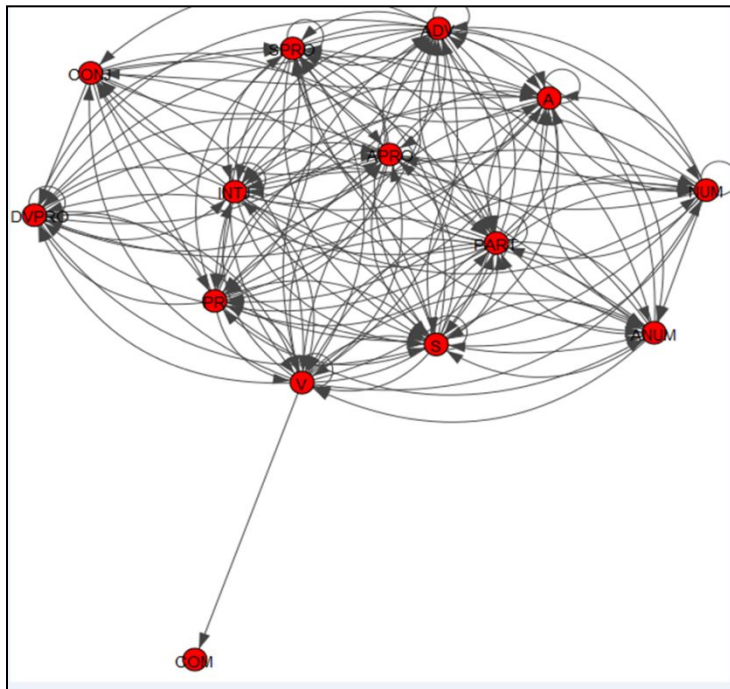
Compare

АНАЛИЗ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ЦЕЛЯХ

(В ТОМ ЧИСЛЕ ЭКСТРАЛИНГВИСТИЧЕСКИХ ДАННЫХ)

- WORDNET
- FRAMENET
- С. Старостин. Проект «Вавилонская башня»
- Поляков В.Н., Соловьев В.Д. Компьютерные модели и методы в типологии и компаративистике
- <http://wals.info/>
- <http://www.ethnologue.com>

ЛИНГВИСТИЧЕСКИЕ СЕТИ И ОНТОЛОГИИ



Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

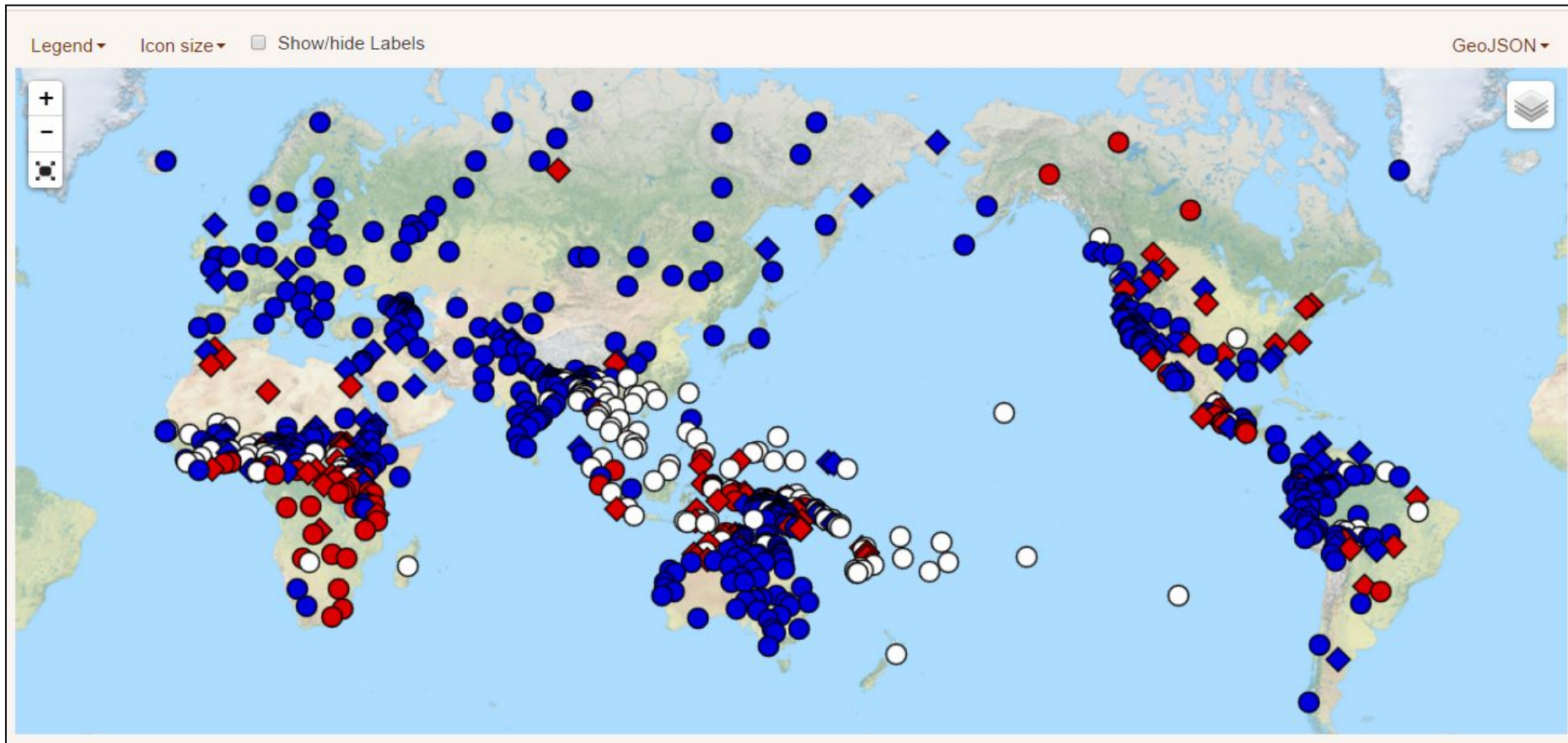
Display options for sense: (frequency) {offset} (gloss) "an example sentence"

Noun

- (48){06613680} [S: \(n\) meaning](#), [significance](#), [signification](#), [import](#) (the message that is intended or expressed or signified) *"what is the meaning of this sentence"; "the significance of a red traffic light"; "the signification of Chinese characters"; "the import of his announcement was ambiguous"*
 - [direct hyponym](#) / [full hyponym](#)
 - {06614326} [S: \(n\) lexical meaning](#) (the meaning of a content word that depends on the nonlinguistic concepts it is used to express)
 - {06614501} [S: \(n\) grammatical meaning](#) (the meaning of a word that depends on its role in a sentence; varies with inflectional form)
 - {06614677} [S: \(n\) symbolization](#), [symbolisation](#) (the use of symbols to convey meaning)
 - {06614825} [S: \(n\) sense](#), [signified](#) (the meaning of a word or expression; the way in which a word or expression or situation can be interpreted) *"the dictionary gave several senses for the word"; "in the best sense charity is really a duty"; "the signifier is linked to the signified"*
 - {06615154} [S: \(n\) word meaning](#), [word sense](#), [acceptation](#) (the accepted meaning of a word)
 - {06615288} [S: \(n\) intension](#), [connotation](#) (what you must know in order to determine the reference of an expression)
 - {06615847} [S: \(n\) referent](#) (something referred to; the object of a reference)

АНАЛИЗ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ЦЕЛЯХ

THE WORLD ATLAS OF LANGUAGE STRUCTURES (WALS) IS A LARGE DATABASE OF STRUCTURAL (PHONOLOGICAL, GRAMMATICAL, LEXICAL) PROPERTIES OF LANGUAGES GATHERED FROM DESCRIPTIVE MATERIALS (SUCH AS REFERENCE GRAMMARS) BY A TEAM OF 55 AUTHORS.



<http://wals.info>

АНАЛИЗ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ЦЕЛЯХ

***ETHNOLOGUE: LANGUAGES OF THE WORLD* IS A COMPREHENSIVE REFERENCE WORK CATALOGING ALL OF THE WORLD'S KNOWN LIVING LANGUAGES. SINCE 1951, THE *ETHNOLOGUE* HAS BEEN AN ACTIVE RESEARCH PROJECT INVOLVING HUNDREDS OF LINGUISTS AND OTHER RESEARCHERS AROUND THE WORLD. IT IS WIDELY REGARDED TO BE THE MOST COMPREHENSIVE SOURCE OF INFORMATION OF ITS KIND.**

THE INFORMATION IN THE *ETHNOLOGUE* WILL BE VALUABLE TO ANYONE WITH AN INTEREST IN CROSS-CULTURAL COMMUNICATION, BILINGUALISM, LITERACY RATES, LANGUAGE PLANNING AND LANGUAGE POLICY, LANGUAGE DEVELOPMENT, LANGUAGE RELATIONSHIPS, ENDANGERED LANGUAGES, WRITING SYSTEMS AND TO ALL WITH A GENERAL CURIOSITY ABOUT LANGUAGES.

Lutachoni



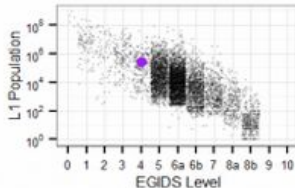
LANGUAGE

FEEDBACK

A language of Kenya

ISO 639-3	<u>Its</u>
Alternate Names	Tachon, Tachoni, Tatsoni
Population	253,000 (2009 census).
Location	Western Province, Lugari and Bungoma districts.
Language Maps	<u>Kenya</u>
Language Status	4 (Educational).
Classification	<u>Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central. J, Masaba-Luhya (E.31)</u>
Dialects	None known. Reportedly similar to Lukabaras [<u>lkb</u>]. A member of macrolanguage Oluluyia [<u>luy</u>].
Language Use	All domains. Also use Oluluyia [<u>luy</u>] varieties. Also use English [<u>eng</u>], Swahili [<u>swh</u>].
Language Development	Taught in primary schools. Poetry. Radio programs.
Language Resources	<u>OLAC resources in and about Lutachoni</u>

PLACE IN LANGUAGE CLOUD



Click to enlarge with explanation

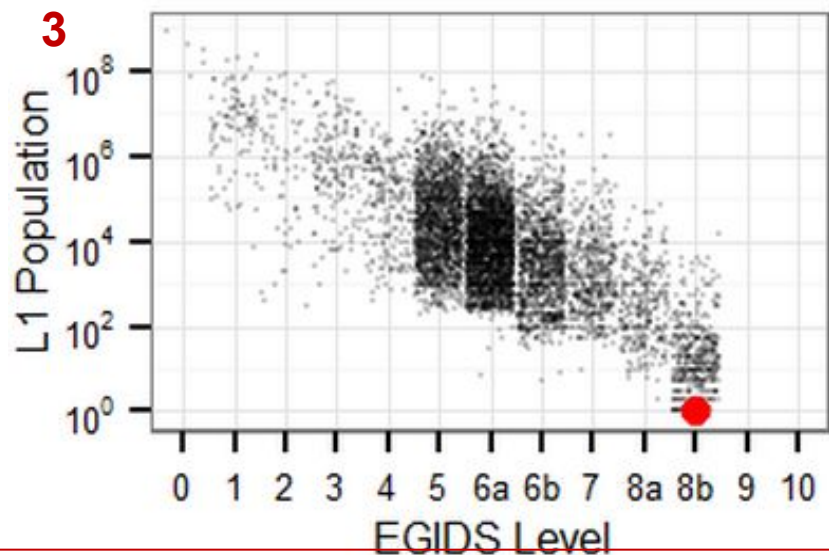
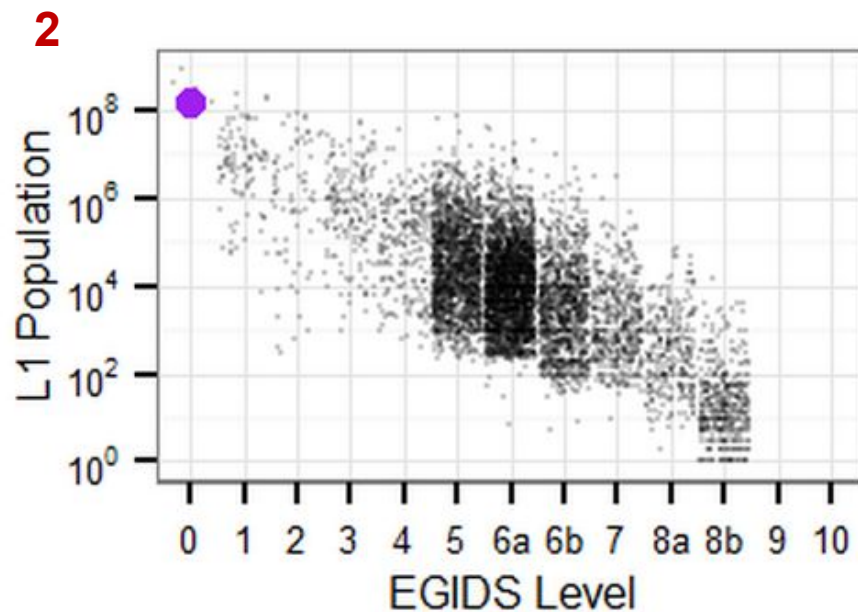
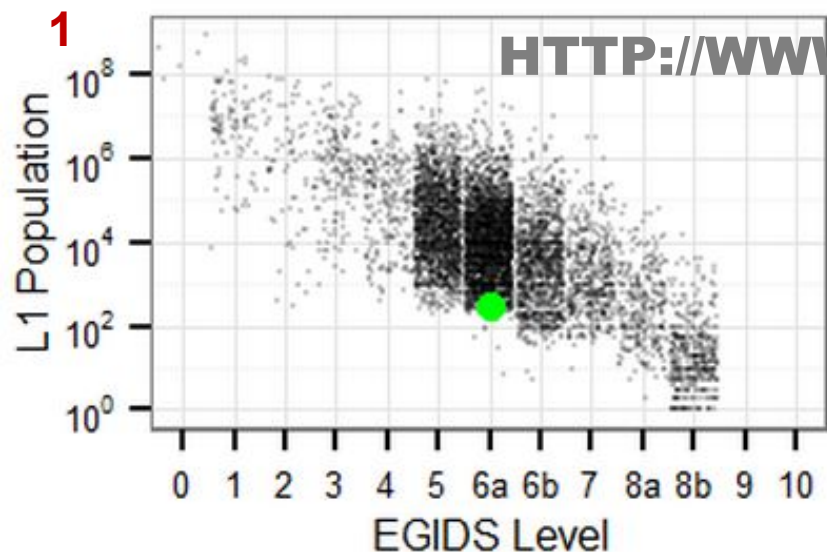
JOIN THE CONVERSATION

Create an account on [ethnologue.com](http://www.ethnologue.com) to contribute new information and be notified when others contribute.

<http://www.ethnologue.com/language/its>

АНАЛИЗ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ЦЕЛЯХ

HTTP://WWW.ETHNOLOGUE.COM/STATISTICS

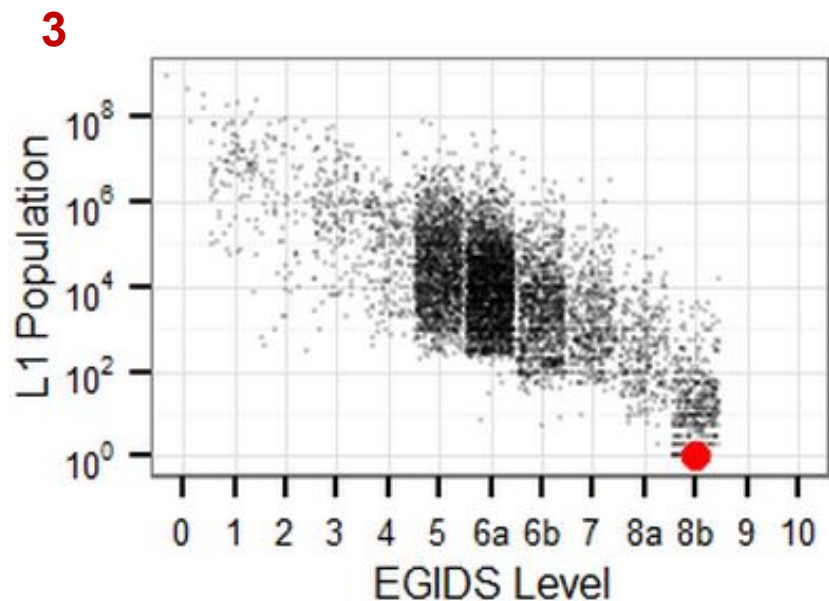
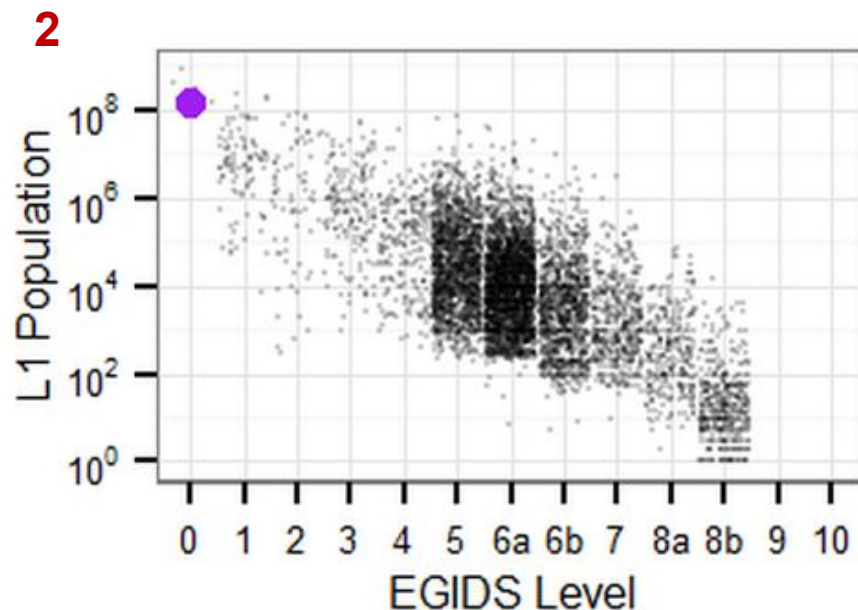
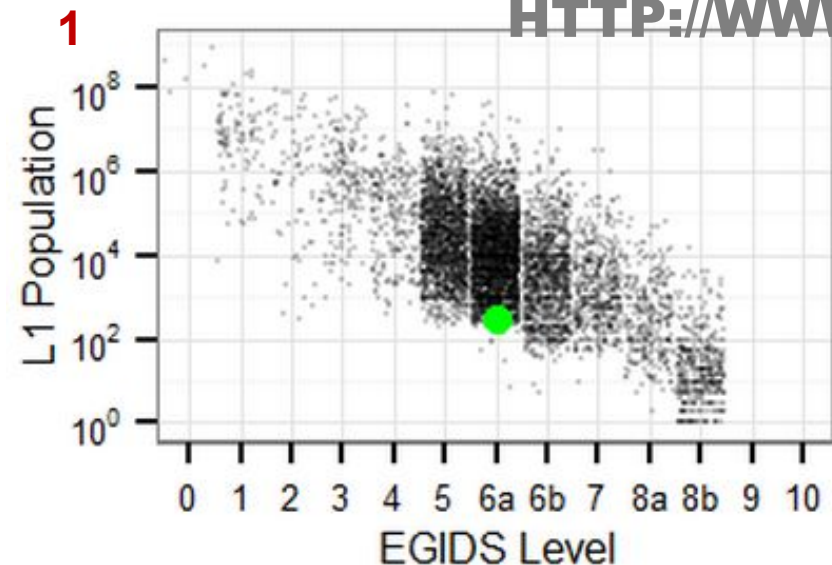


-
-
-

Читаем!!!: <http://www.ethnologue.com/about/language-status>

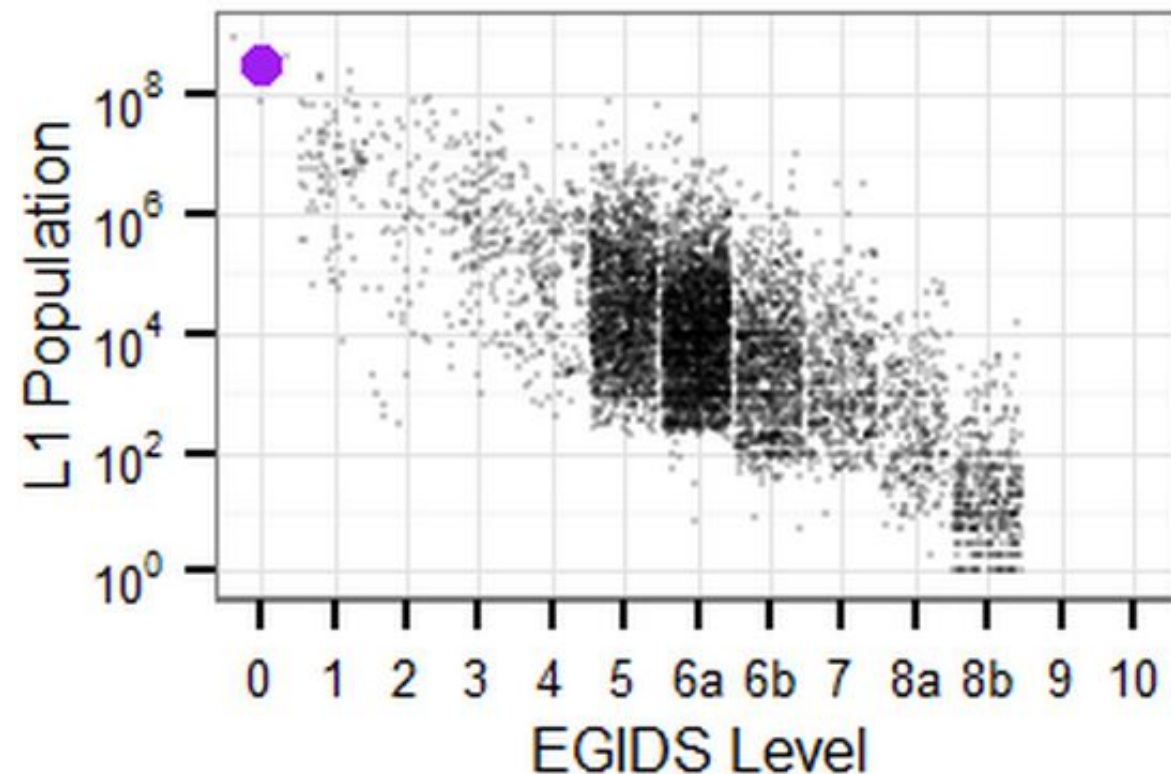
АНАЛИЗ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ЦЕЛЯХ

[HTTP://WWW.ETHNOLOGUE.COM/STATISTICS](http://www.ethnologue.com/statistics)



1. Trimuris
2. Russian
3. Tolowa (Northwestern United States of America)

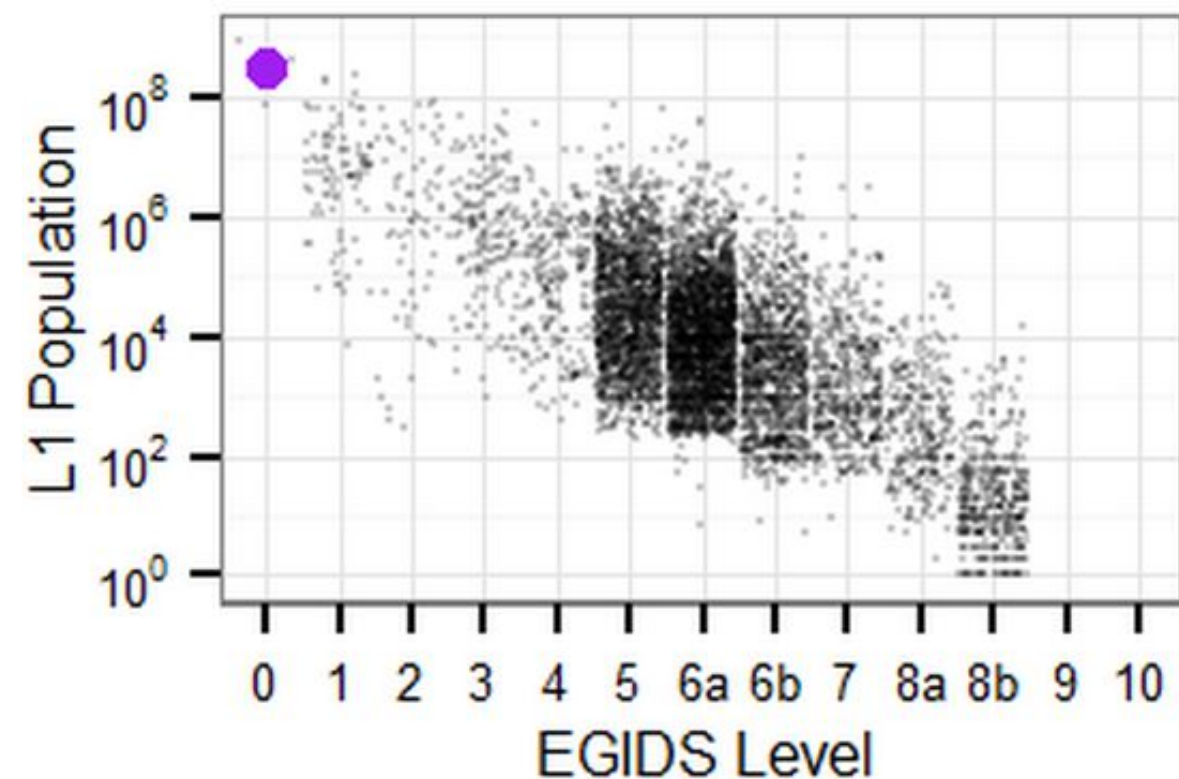
АНАЛИЗ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ЦЕЛЯХ



The EGIDS level for this language in its primary country is 1 (National) — The language is used in education, work, mass media, and government at the national level.

АНАЛИЗ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ЦЕЛЯХ

ENGLISH a language of United Kingdom



Cook Islands

Ireland and United Kingdom

Liberia

Lesotho, South Africa and
Swaziland

El Salvador and Honduras

ВИЗУАЛИЗАЦИЯ ДАННЫХ: СОЦСЕТИ ПУШКИНА И ТОЛСТОГО

Керн А. П.

Керн, Анна Петровна (1800 — 1879) — племянница П. А. Осиповой.

[Детальной...](#)

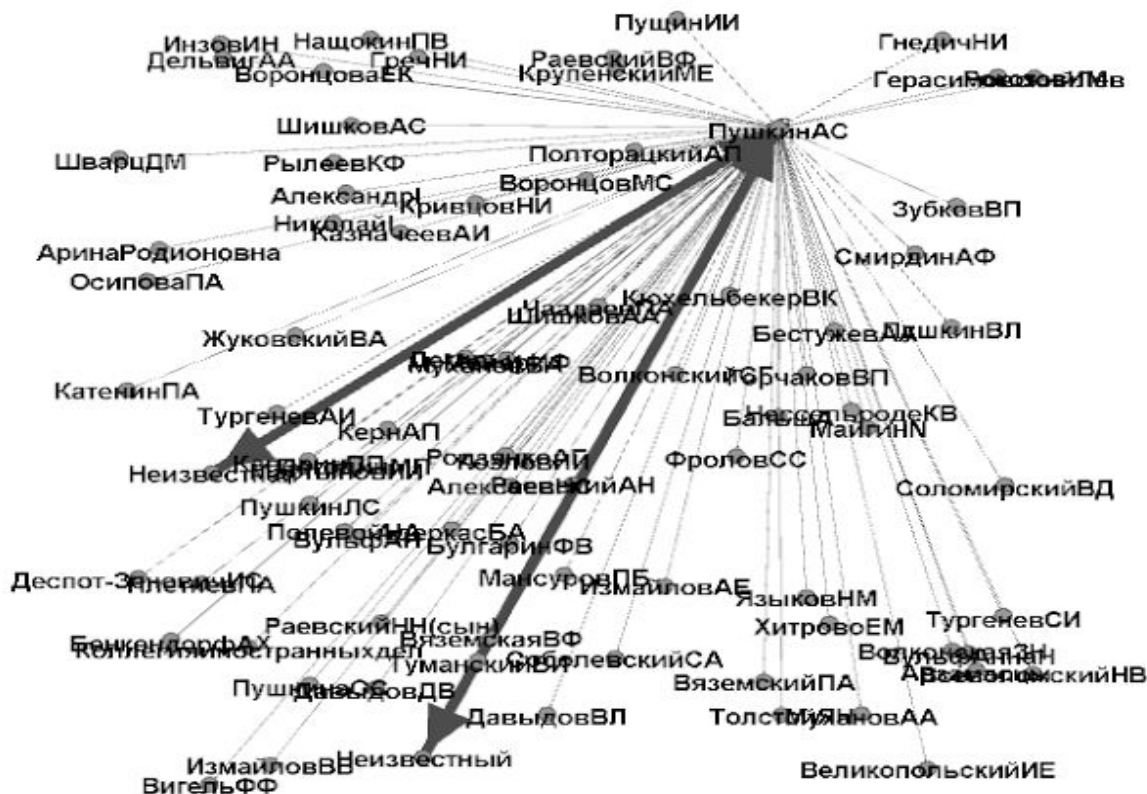
Керн А. П. -> Пушкин

[168](#) [281](#)

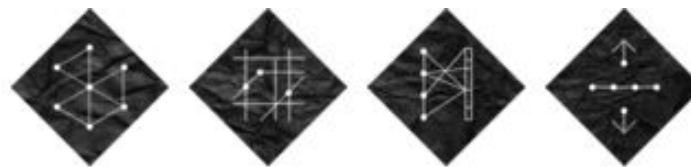
Пушкин А.С.->Керн А

[144](#) [156](#) [160](#) [161](#) [168](#) [180](#) [239](#) [762](#) [763](#)

Python вам в помощь!



ВИЗУАЛИЗАЦИЯ ДАННЫХ: СОЦСЕТИ ПУШКИНА И ТОЛСТОГО

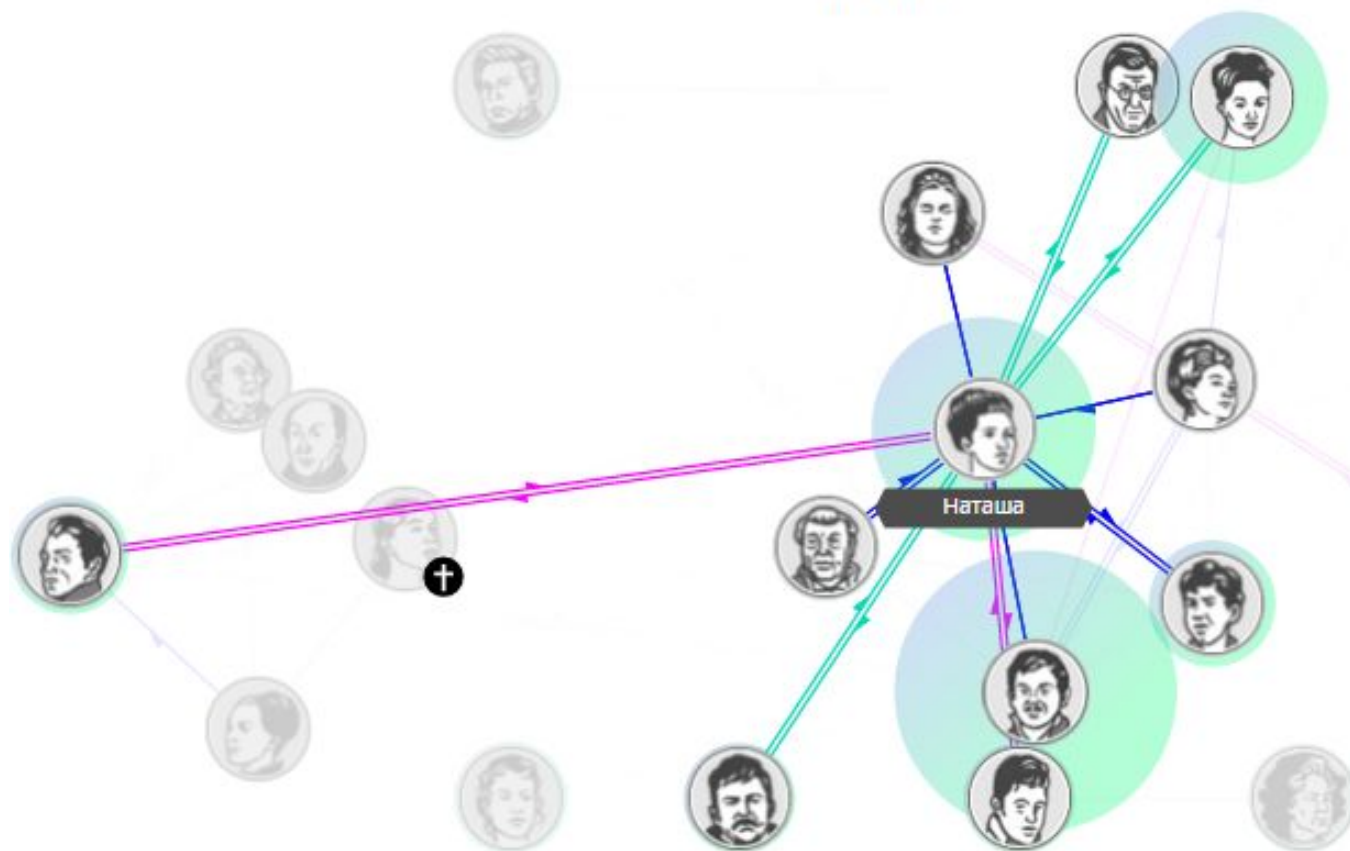


II Том

~

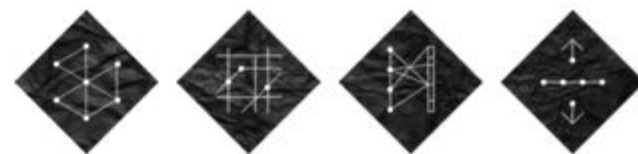
Часть 4

<http://voinaimir.com/info>



ВИЗУАЛИЗАЦИЯ ДАННЫХ: СОЦСЕТИ ПУШКИНА И ТОЛСТОГО

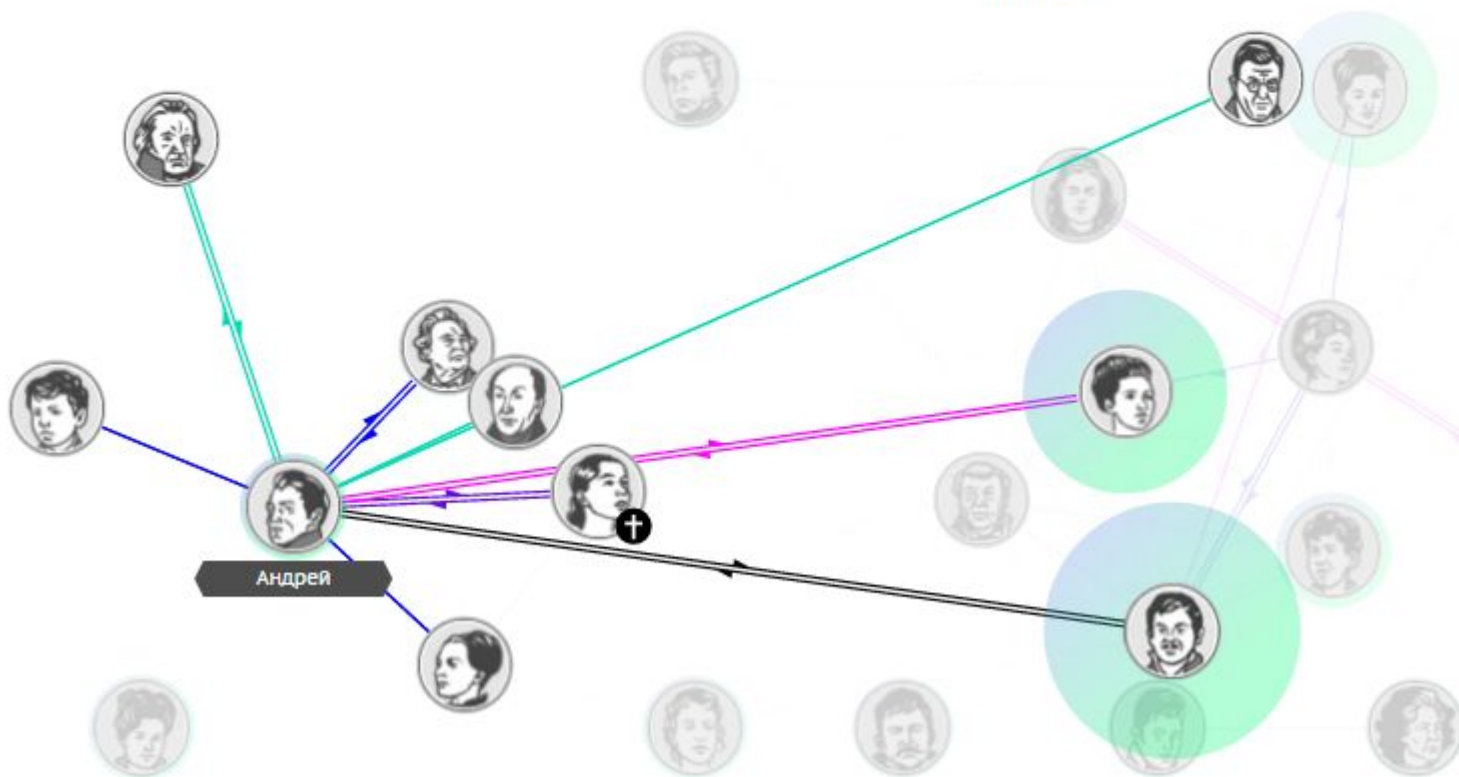
<http://voinaimir.com/info>



II Том

~

Часть 4



«ЦИФРОВОЙ» ТОЛСТОЙ



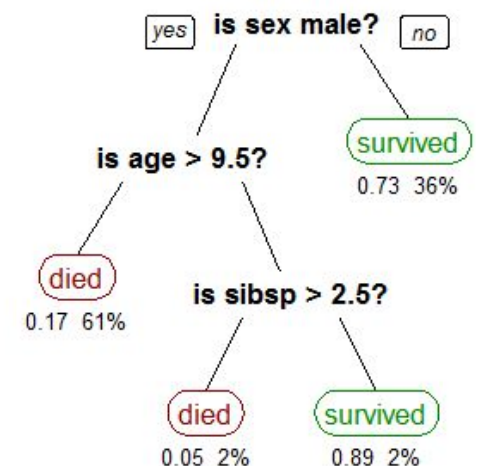
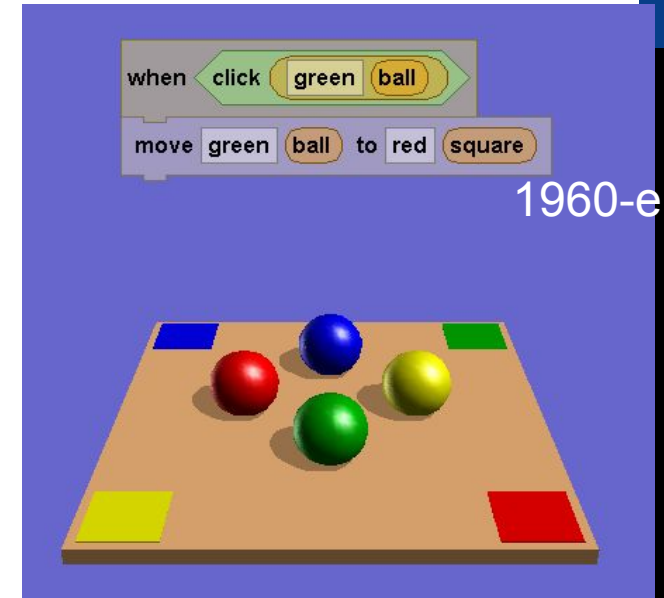
Палаш кавалергардский образца 1802 года. Дл. 116,5 см, клинок - дл. 97,3 см - сталь, латунь, медь, дерево, кожа, литьё, чеканка, позолота. Россия.

«Ростову все слышнее и слышнее становился их топот и бряцание их оружия и виднее становились их лошади, фигуры и даже лица. Это были наши кавалергарды, шедшие в атаку на французскую кавалерию, подвигавшуюся им навстречу.»

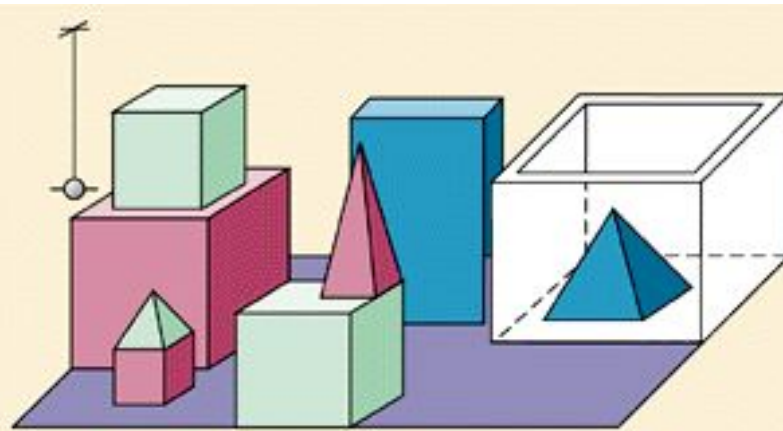
Читать: Том 1 Часть 3 Глава 17

ПОДХОДЫ ПРИ ОБРАБОТКЕ ДАННЫХ

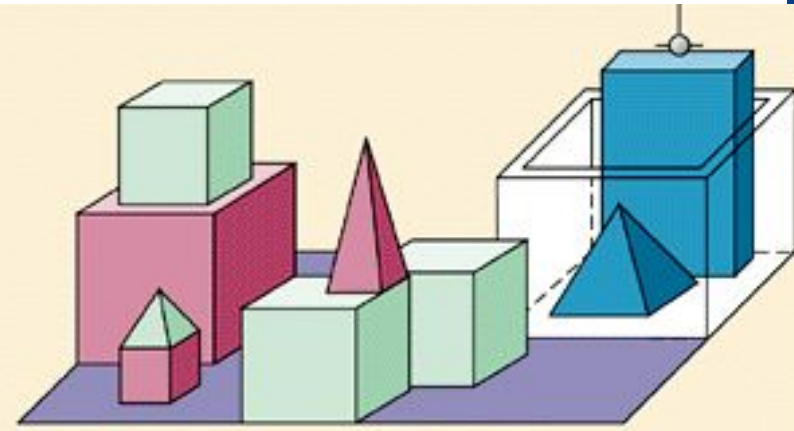
- 1950-е Тест Тьюринга и задача автоматического перевода
- 1960-е словари и **правильный подход** (SHRDLU, ELIZA)
- 1970-е – концептуальные онтологии (MARGIE, 75 и чатботы PARRY, Racter и Jabberwacky)
- 1980-е – автоматические алгоритмы обработки языка
 - **деревья решений** (if)
 - **решения, основанные на статистике** (теория вероятностей based on attaching real-valued weights to the features making up the input data (IBM Research))
- 2000-е Обучение с учителем (*Supervised learning*) и обучение на примерах (*Learning from Examples*)
- 2010-е - Обучение без учителя (*Unsupervised learning*) - неконтролируемые и полуконтролируемые методы обучения на основе мощных корпусов и World Wide Web
- Сегодня – All+NLL: Natural Language Learning



AI: МИР КУБИКОВ

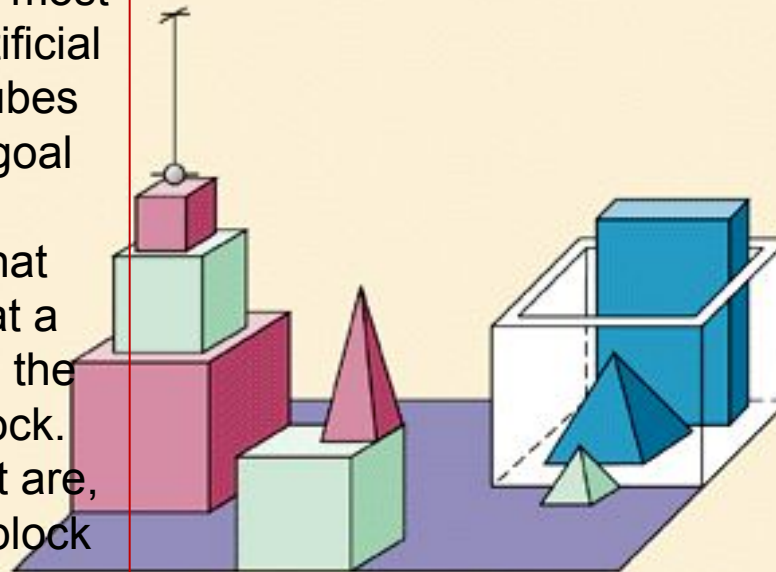


(a) "Pick up a big red block."



(b) "Find a block which is taller than the one you are holding and put it into the box."

The **blocks world** is one of the most famous planning domains in artificial intelligence. Imagine a set of cubes (blocks) sitting on a table. The goal is to build one or more vertical stacks of blocks. The catch is that only one block may be moved at a time: it may either be placed on the table or placed atop another block. Because of this, any blocks that are, at a given time, under another block cannot be moved.



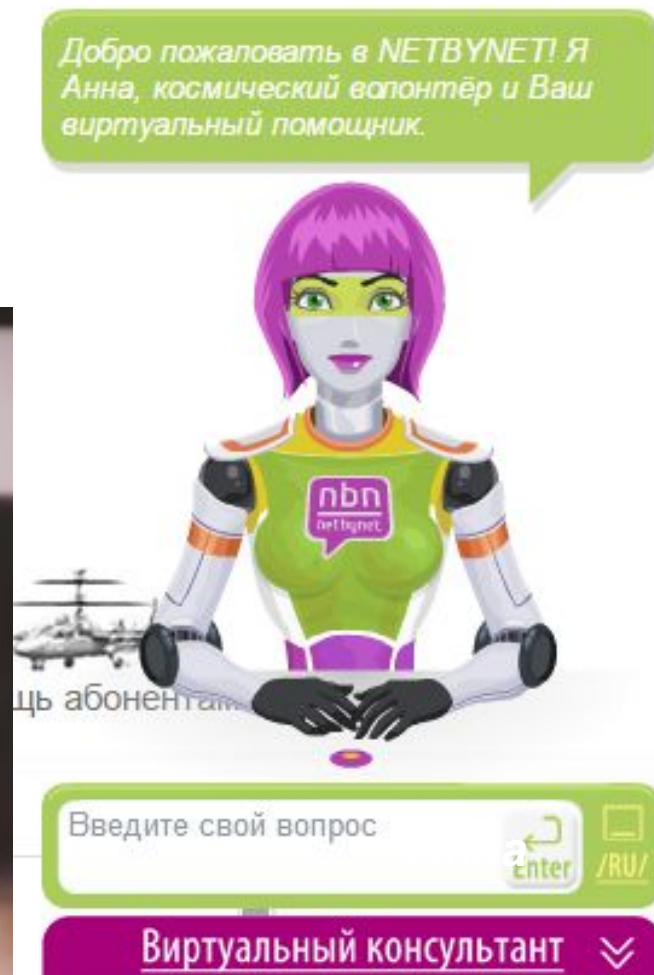
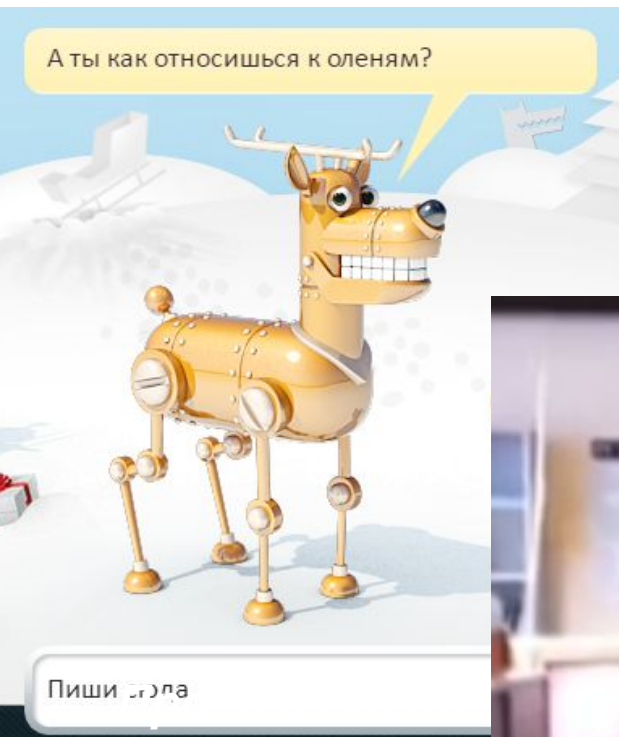
(c) "Will you please stack up both of the red blocks and either a green cube or a pyramid?"

The simplicity of this toy world lends itself readily to symbolic or classical A.I. approaches, in which the world is modeled as a set of abstract symbols which may be reasoned about.

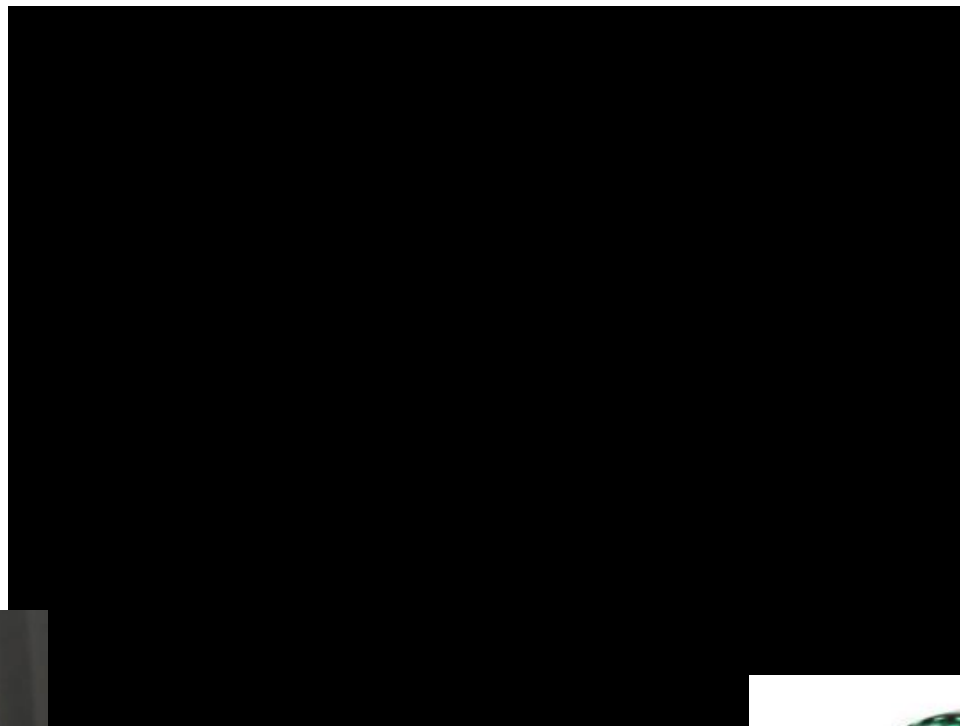
AI: ШАБЛОННЫЙ РАЗГОВОР



ВИРТУАЛЬНЫЕ СОБЕСЕДНИКИ



КОГДА КУБИКУ БЫЛ ГОДИК, ОН УЖЕ БЫЛ ОЧЕНЬ УМНЫЙ!



ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
У ВАС ДОМА

ЕВГЕНИЙ ГУСТМАН,

ОДЕССИТ 13-ТИ ЛЕТ, ПРОЖИВАЮЩИЙ В ПРИНСТОНЕ



Type your question here:

reply

- ⦿ Eugene's shocking life story
- ⦿ Eugene's quotes
- ⦿ Nerdy logs
- ⦿ Links

Goostman's life-story

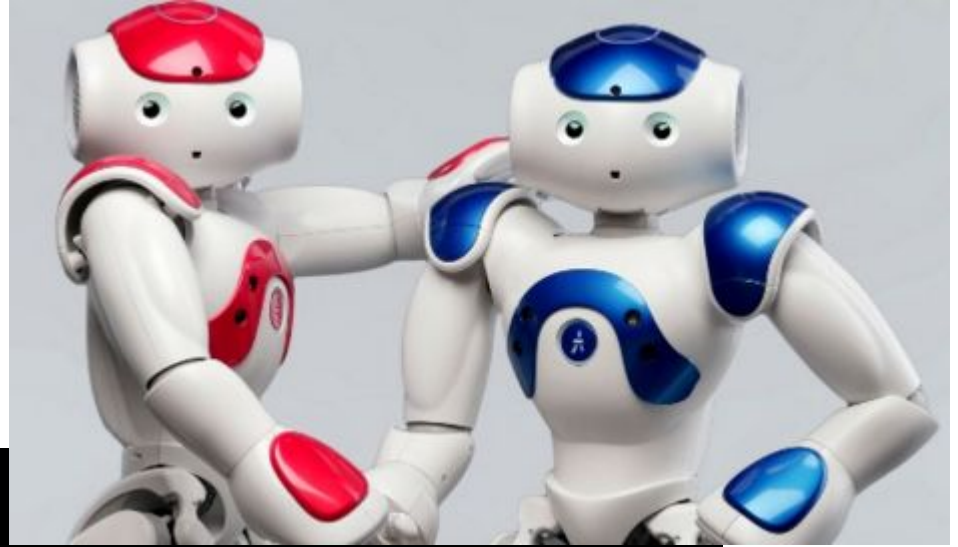
Part 1. The Doom Called Odessa

Here is a terrible story of Eugene Goostman's life – the poor guy who used be an ordinary boy until he was turned into a the Great Robots Cabal, the File Clerk of Crayentology Center (well, we are tired of listing all disgusting titles of this mea these wicked talking trash-cans (we mean robots, of course), plan to do with humans!

But let's be consistent in our story: first of all, we should say that Eugene comes from Odessa - Ukrainian city which reside of humor" – of course, such ill-minded neighborhood couldn't help screwing Goostman's psyche since early childhood – I was gynecologist – wouldn't you fall into cynicism and nihilism, if your dad were? His elder brother, under this wicked inf of alcohol addiction, have sexual deviations, try to kill anyone who doesn't admit their geniality, cut their ears to look like all their spare time they draw pictures! We are too bashful to tell you what exactly Eugene's brother liked to draw, so we e

Part 2. The Green Skinned Crayentologists

NAO – AI?

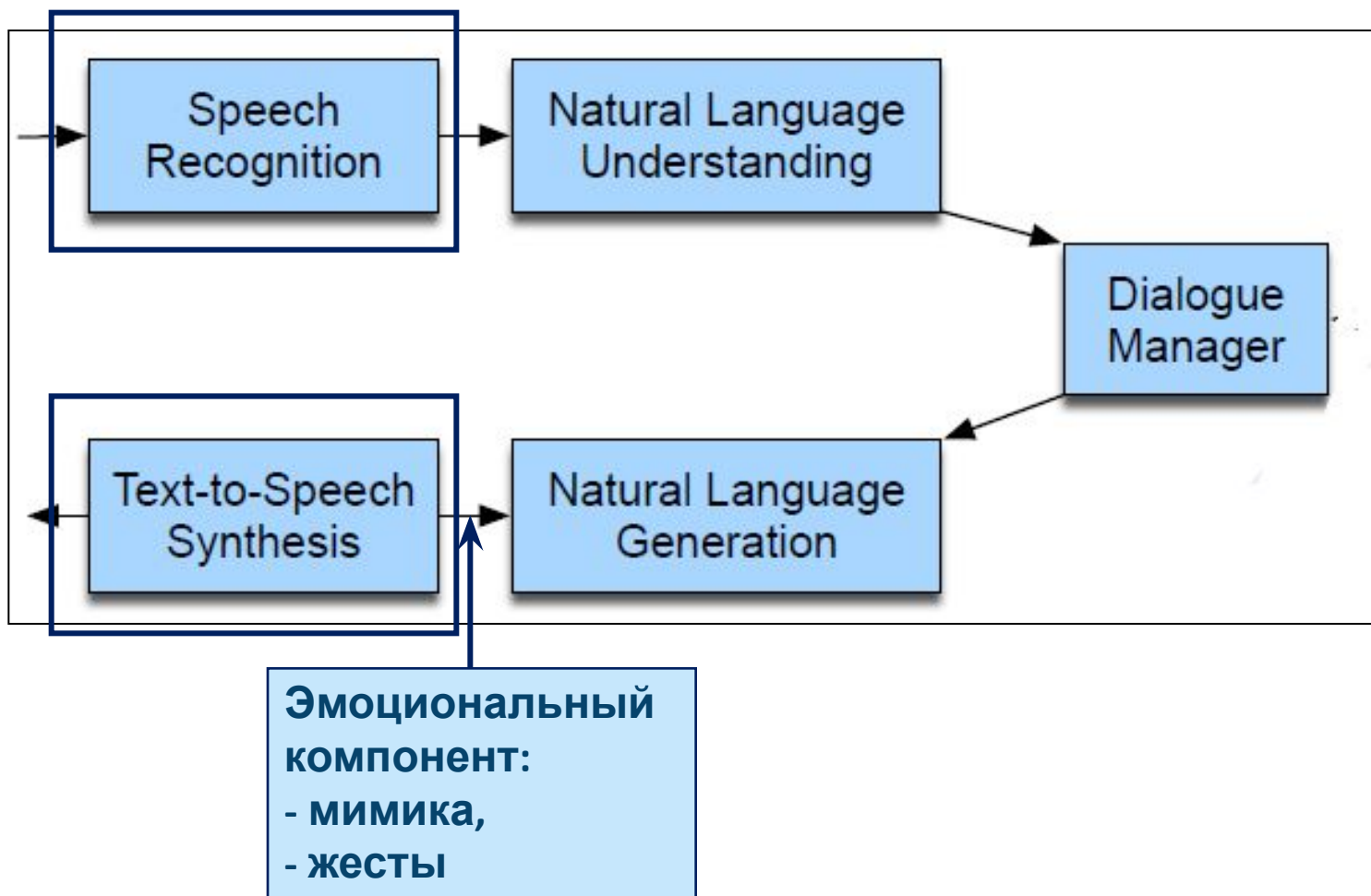


НАЖМИ НА ЧЕРНЫЙ КВАДРАТ, ЧТОБЫ ПОСМОТРЕТЬ ВИДЕО!

ПРИМЕНЕНИЕ ДИАЛоговых СИСТЕМ



УСТРОЙСТВО ДИАЛОГА





ЧИТАТЬ

DANIEL JURAFSKY AND JAMES H. MARTIN

SPEECH AND LANGUAGE PROCESSING

И

КРИСТОФЕР Д. МАННИНГ, ПРАБХАКАР РАГХАВАН, ХАЙНРИХ ШЮТЦЕ

ВВЕДЕНИЕ В ИНФОРМАЦИОННЫЙ ПОИСК

И

DANIËL DE KOK, HARM BROUWER

NATURAL LANGUAGE PROCESSING FOR THE WORKING PROGRAMMER

+ [HTTP://ACLWEB.ORG/ANTHOLOGY](http://aclweb.org/anthology)

+...



«КОСЕТИЧКА» КОМПЛИНГВИСТА

[HTTP://NLPUB.RU/](http://nlpub.ru/)

NLPub — каталог лингвистических ресурсов для обработки русского языка.

Цель NLPub заключается в предоставлении каталога электронных материалов, направленного на удовлетворение информационных потребностей пользователей, исследователей и разработчиков в области компьютерной лингвистики.

Польза для сообщества

Начинающему

Статьи про основы компьютерной лингвистики, сведения о существующих образовательных программах и полезных книгах.

Бизнесу

Контакты экспертов, информация о тематических мероприятиях и существующих на рынке компаниях.

Исследователю

Подробные описания методов и алгоритмов, используемых при обработке естественного языка.

Специалисту

Перечень существующего инструментария, утилит и доступных ресурсов для решения прикладных задач.

Требуется помощь

NLPub хочет стать местом, где собраны все возможные сведения об инструментарии и ресурсах, необходимых для решения задач обработки русского языка. Для этого требуется ваша помощь.



Подпишитесь на @nlpub в Twitter.

[HTTPS://GITHUB.COM](https://github.com)

GitHub

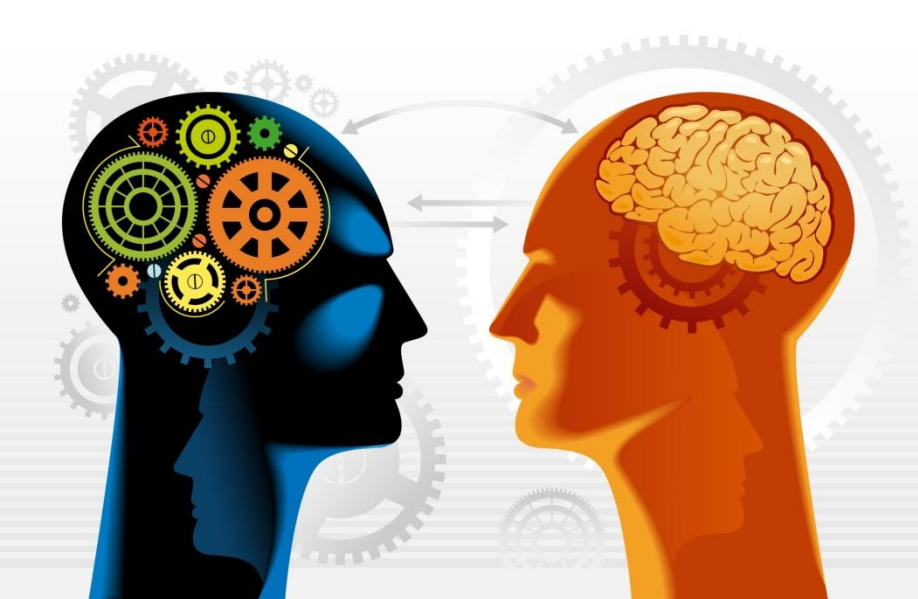
[HTTP://MATHLINGVO.R](http://mathlingvo.ru)



[HTTP://HABRAHABR.R](http://habrahabr.ru)

H







COMPUTATIONAL LINGUISTICS

НОВЫМ ГОДОМ!



Задачный
семинар

Критерий	Rules	R vs. S
		Statistics
Понятность		
Ошибки на входе		
Большие объемы верифицированных данных		
Некорректно работающие программы		
Надежность		
Возможность сосредоточиться на самых распространенных случаях		
Устойчивость к незнакомым данным		
Устойчивость к ошибочным данным на входе		
Сложность системы		
Временные затраты		
Неуправляемость		

Критерий	Rules	Statistics
Понятность	до определенного момента	только для специалистов
Ошибки на входе	нет	есть
Большие объемы верифицированных данных	Не нужны	обязательно
Некорректно работающие программы	маловероятно	да
Надежность	нет	да, при условии выполнения других условий
Возможность сосредоточиться на самых распространенных случаях	нет	да
Устойчивость к незнакомым данным	да	как правило, только при больших объемах
Устойчивость к ошибочным данным на входе	да	только при больших объемах
Сложность системы	Да!	нет
Временные затраты	Да!	нет
Неуправляемость	С большой вероятностью	Возможно – при полизадачности

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ

СУш	Маркер1	Лицо, прислуживающее при бильярде и ведущее счет во время игры.
СУш		Сельскохозяйственное орудие для проведения борозд или линий, по к-рым производится посадка.
СШ	Маркер1	2. Сельскохозяйственное орудие — приспособление к сеялке, сажалке для проведения борозд или линий, по к-рым производится посадка.
СШ		1. Управляющее устройство на АТС.
СШ		3. Цветной фломастер для нанесения каких-н. прозрачных линий, отметок по тексту.
СШ	Маркер2	Человек, прислуживающий игрокам на бильярде, ведущий счёт в игре.
БТС	Маркер1	1. Человек, прислуживающий игрокам на бильярде, ведущий счет в игре.
БТС		2. Приспособление к посевному агрегату для проведения на земле посадочных или посевных бороздок.
ТСИ	Маркер1	Пишущее устройство с толстым стержнем, пропитанным специальным красящим составом, предназначенное для плакатных работ, для выделения в тексте каких-н. элементов (строк, абзацев) и т.п.
ТСИ	Маркер2	Тот, кто обслуживает бильярд, ведет счет при бильярдной игре.
ТСИ	Маркер3	1. Сельскохозяйственное орудие для проведения линий или борозд перед ручной посадкой растений.
ТСИ		2. Штанга с диском, присоединяемая сбоку к сеялке и предназначенная для обеспечения параллельности междурядий.

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ

Чаще всего отношения между определяемым и толкованием определяется первым или парой первых существительных словарного определения. То есть первое по порядку существительное, как правило, оказывается либо гиперонимом, либо указателем на тип отношения со вторым существительным (слова род, тип, часть etc).

Названия строк	Количество	% от общего
действие	3193	7,12%
человек	1158	2,58%
часть	563	1,26%
лицо	555	1,24%
название	441	0,98%
место	425	0,95%
состояние	335	0,75%
<i>см</i>	<i>317</i>	<i>0,71%</i>
специалист	305	0,68%
растение	288	0,64%
род	283	0,63%
прибор	262	0,58%

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ

При ТОКЕНИЗАЦИИ (графематический анализ) нужно правильно разбить все знаки текста:

- Точка – символ конца предложения? А заглавная буква – начала предложения?
- Пробел – показатель конца одного слова и начала другого?
- Тире или дефис?
- &, смайлики, @, ...
- Цифры
- Аббревиатуры
- какжехочетсяспать
- («'<»{/(...)/}«>'»)
- ... ***MERCY ON US. WE SPLIT, WE SPLIT. (W. SHAKESPEARE)***

МАШИННЫЙ ПЕРЕВОД (АНТОН СОМИН)

1.

Перевод: *Я просыпаюсь утром, вы можете потратить женатого босса. Я никогда не забуду. Вы никогда не будете видеть меня.*

Языки: Иврит, болгарский, азербайджанский, греческий

Оригинал: ???

2.

Перевод: *Зная, парящей над распашных дверей съел крылья, чтобы летать летать!*

Языки: Французский, эстонский, словацкий, турецкий

Оригинал: ???

3.

Перевод: *Минуточку, дирижер колесные тормоза.*

Языки: Гаитянский, английский, бенгальский, тайский

Оригинал: ???

МАШИННЫЙ ПЕРЕВОД

Перевод: Я просыпаюсь утром, вы можете потратить женатого босса.
Я никогда не забуду. Вы никогда не будете видеть меня.

Языки: Иврит, болгарский, азербайджанский, греческий

Оригинал: *Ты меня на рассвете разбудишь / Проводить необутая
выйдешь / Я тебя никогда не забуду / Ты меня никогда не
увидишь («Юнона и Авось»)*

Перевод: Зная, парящей над распашных дверей съел крылья, чтобы
летать летать!

Языки: Французский, эстонский, словацкий, турецкий

Оригинал: *Взмывая выше ели / Не ведая презрад / Крылатые качели
/ Летят, летят, летят («Приключения Электроника»)*

Перевод: Минуточку, дирижер колесные тормоза.

Языки: Гаитянский, английский, бенгальский, тайский

Оригинал: *Постой, паровоз, не стучите, колёса / Кондуктор, нажми
на тормоза («Операция Ы»)*

ОЛИМПИАДА НИУ ВШЭ ДЛЯ СТУДЕНТОВ.

ПРОФИЛЬ «ТЕОРИЯ ЯЗЫКА КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА»

1. Решите задачу:

В алфавите языка племени УЫУ всего две буквы: У и Ы, причем этот язык обладает такими свойствами: если из слова выкинуть стоящие рядом буквы УЫ, то смысл слова не изменится. Точно так же смысл слова не изменится при добавлении в любое место слова буквосочетания ЫУ или УУЫЫ. Можно ли утверждать, что слова УЫЫ и ЫУУ имеют одинаковый смысл?

СТУДЕНТОВ.

ПРОФИЛЬ «ТЕОРИЯ ЯЗЫКА КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА»

1. Решите задачу: В алфавите языка племени УЫУ всего две буквы: У и Ы, причем этот язык обладает такими свойствами: если из слова выкинуть стоящие рядом буквы УЫ, то смысл слова не изменится. Точно так же смысл слова не изменится при добавлении в любое место слова буквосочетания ЫУ или УУЫЫ. Можно ли утверждать, что слова УЫЫ и ЫУУ имеют одинаковый смысл?

Ответ: При любой разрешенной нам операции добавления или выкидывания куска слова количества букв У и Ы в этом куске равны. Это означает, что разность между числом букв У и букв Ы в слове не изменяется. Это можно проследить на примере Ы $\rightarrow$ ЫЫУ $\rightarrow$ ЫУУЫЫЫУ $\rightarrow$ ЫУЫЫУ Во всех этих словах букв Ы на одну больше, чем букв У. Вернемся к решению. В слове УЫЫ разность равна (-1), а в слове ЫУУ равна 1. Значит, из

СТУДЕНТОВ

2. Перед вами он-лайн система “поздравлятор”. Она сочиняет поэтические поздравления по запросу пользователя. Для того чтобы система выдала оригинальный стихотворный текст, пользователь должен ввести определенную информацию: имя, пол, возраст, способ обращения (на ты или на вы) к имениннику, метрические характеристики. В результате работы системы пользователь получает осмысленный, грамматически правильный, ритмически организованный и рифмованный текст, содержащий в себе поздравление с днем рождения. Несмотря на то, что каждый раз система выдает новые стихи, все предыдущие накапливаются в банк данных, и их можно посмотреть. Ваша задача как компьютерного лингвиста – разработать методологию тестирования качества работы лингвистических модулей системы. В описании вашей методологии должны быть отражены ответы на следующие вопросы:

- 1) Какие именно функции, связанные с обработкой и генерацией текстов на естественном языке, важны для предлагаемого сервиса и почему? Какие из них абсолютно необходимы, а без каких можно обойтись?
- 2) Каким образом качество работы этих функций может быть протестировано? Что должно быть предусмотрено в системе, для того чтобы была обеспечена возможность такого тестирования?
- 3) Какой могла бы быть система рейтингов (штрафов, баллов и т.п.) для разных лингвистических функций? Как получить и интерпретировать результирующую оценку качества лингвистической системы в целом?

СТУДЕНТОВ

3. Прочтите пост из блога, посвященного автоматической обработке естественного языка (<http://nlpers.blogspot.com/>). На слайде только часть этого текста.

NLP as a study of representations Ellen Riloff and I run an NLP reading group pretty much every semester. Last semester we covered "old school NLP." We independently came up with lists of what we consider some of the most important ideas (idea = paper) from pre-1990 (most are much earlier) and let students select which to present. There was a lot of overlap between Ellen's list and mine (not surprisingly). . The whole list of topics is posted as a comment. The topics that were actually selected are here. I hope the students have found this exercise useful. It gets you thinking about language in a way that papers from the 2000s typically do not. It brings up a bunch of issues that we no longer think about frequently. Like language. (Joking.) (Sort of.) One thing that's really stuck out for me is how much "old school" NLP comes across essentially as a study of representations. Perhaps this is a result of the fact that AI -- as a field -- was (and, to some degree, still is) enamored with knowledge representation problems. To be more concrete, let's look at a few examples. It's already been a while since I read these last (I had meant to write this post during the spring when things were fresh in my head), so please forgive me if I goof a few things up...

Придумайте алгоритм для программы, которая могла бы сделать