



# **Автоматическая обработка текстов на естественном языке.**

Евгений Борисов

# Введение в NLP

**язык как средство познания мира**

сегментация на объекты

классификация наблюдаемых явлений

обобщения

упорядочение окружающей действительности

# Введение в NLP

**о языке и задачах его автоматической обработки**

обработка текстов на естественном языке (ЕЯ)

Natural Language Processing (NLP)

NLU / natural language understanding

NLG / natural language generation

SP / speech processing (recognition/generation)

# Введение в NLP

## обработка текстов на естественном языке

NLP : NLU / NLG

- машинный перевод (MT)
- диалоговые системы (чат-боты)
- извлечение именованных сущностей, (named-entity recognition, NER)
- извлечения фактов и отношений (relation extraction)
- реферирование (summarization)
- поиск обоснования в тексте (argumentation mining)
- классификация текстов (оценка тона и т.п.)

# Введение в NLP

## Сложности автоматической обработки текстов - неоднозначности в языке

*«эти типы стали есть на складе»*

омонимия - случайное совпадение слов

ключ, лук, замок, печь

полисемия - несколько значений, связанных исторически

стол <организация или объект>

местоименная анафора — ссылки на контекст

Прискакал принц на белом коне. Принцесса выбежала ему навстречу и поцеловала его <...принца> .

эллипсис - пропуски в тексте

Он не может решить задачу, а я знаю как <...решить задачу>.

<https://habr.com/ru/company/abbyy/blog/437008/>

# Введение в NLP

## История развития компьютерной лингвистики

1949 перевод как расшифровка (немецкий язык это зашифрованный английский)

1954 Джорджтаунский эксперимент, большая таблица соответствий фраз (MT)

1960 рационалистический подход: формальные грамматики Хомского

1970 переход на уровень семантики, формальные онтологии, контекст

1980 корпусная лингвистика, статистические языковые модели ML

1990 WWW, задача информационного поиска

2005 Deep Learning, рекуррентные нейросети

2013 Word2vec, семантические пространства

2017 Attention, механизм внимания

# Введение в NLP

## Уровни сложности при автоматической обработке текстов

**Прагматика (Дискурс)** - смысловые контексты

**Семантика** - смыслы последовательностей слов

**Синтаксис** - правила формирования последовательностей слов

**Морфология** - отдельные слова и устойчивые словосочетания

# Введение в NLP

**подходы решению задач NLP**

**аналитический подход:**

наборы грамматических правил

**подход основанный на данных:**

корпус размеченных текстов и методы ML



# Введение в NLP

**подходы решению задач NLP**

**аналитический подход:**

наборы грамматических правил

**подход основанный на данных:**

корпус размеченных текстов и методы ML

**методы решения задач NLP**

частотный анализ (мешок слов, TF-IDF)

морфологический/синтаксический разбор, онтологии

семантические пространства (Word2Vec)

языковые модели

# NLP частотный анализ

**этапы обработки текста для определения тона**

определение языка

токенизация

коррекция орфографических ошибок

лемматизация

частотный анализ (извлечение признаков)

применение классификатора

# NLP частотный анализ

**этапы обработки текста для извлечения информации**

определение языка

токенизация

коррекция орфографических ошибок

лемматизация

синтаксический анализ

семантический анализ

извлечение структурированной информации

# NLP частотный анализ

## этапы обработки текста для чатбота

определение языка

токенизация

коррекция орфографических ошибок

лемматизация

семантическое кодирование слов (W2V, BERT)

применение нейросетевой языковой модели

# Введение в NLP

## Литература

git clone [https://github.com/mechanoid5/ml\\_nlp](https://github.com/mechanoid5/ml_nlp)

С.Ананян Введение в компьютерную лингвистику// Мегапьютер Интеллидженс  
<https://youtu.be/3fHz0IaLSPc>

Турдаков Д.Ю. Основы обработки текстов. ИСП РАН, 2017,2021

<https://www.youtube.com/playlist?list=PL5cBzMoPJgCUn6TbfhqilyToW5lScOdd3>

<https://www.youtube.com/playlist?list=PL5cBzMoPJgCXFdSvWaun0y4cILirW1lMD>