UNIVERSITÀ DEGLI STUDI DI NAPOLI
PARTHENOPE

Natural Language Processing

Natural Language Processing Overview & Course Organization

LESSON 1

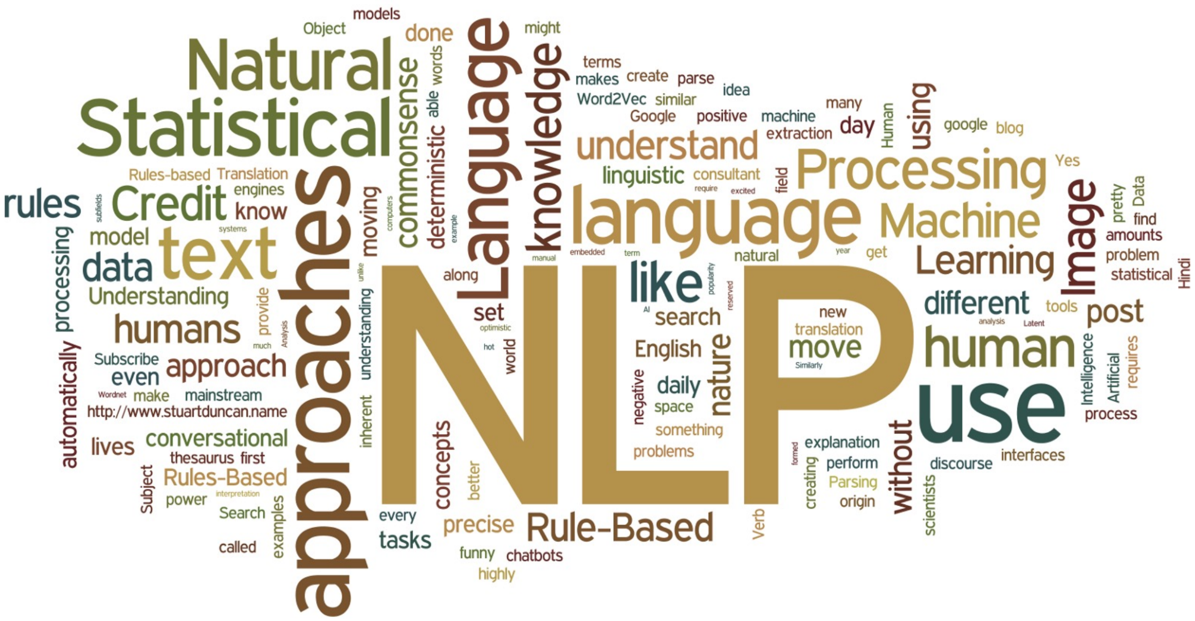
prof. Antonino Staiano

M.Sc. In ''Machine Learning e Big Data'' - University Parthenope of Naples

What’s Natural Language Processing

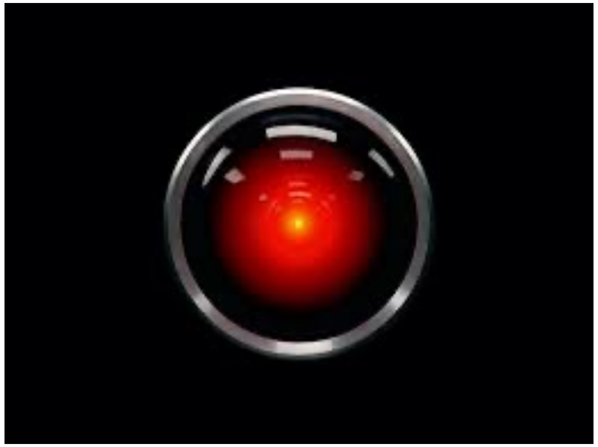
- **Natural Language Processing (NLP)** is one of the most exciting fields in Artificial Intelligence

- It allows machines to cope with human language in a variety of ways, and it’s triggering a revolution in the way we interact with systems and technology

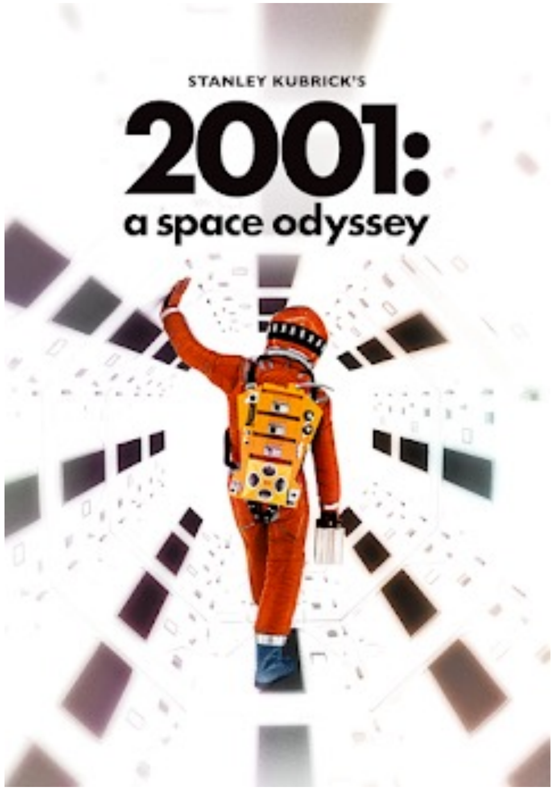


A journey into the ... future?

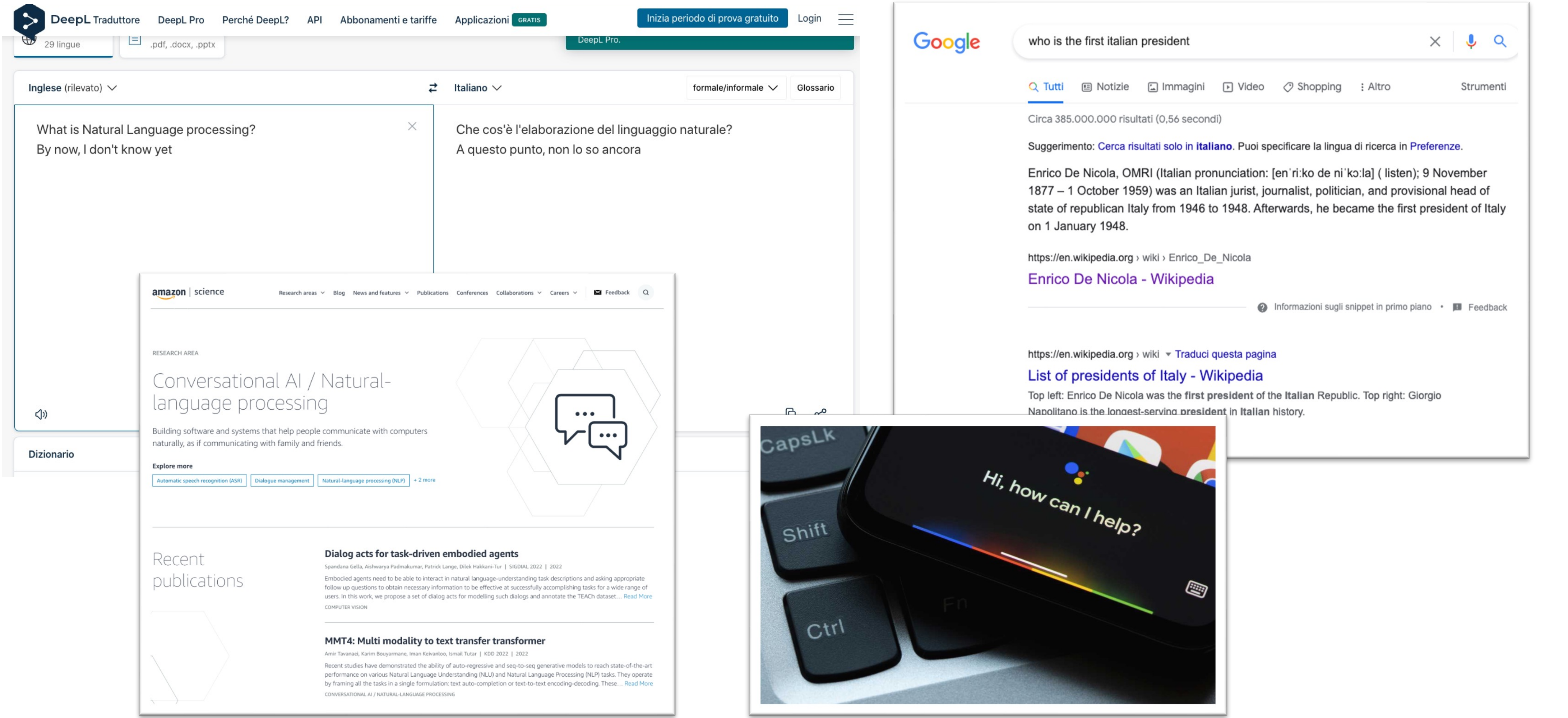
- HAL and Dave



HAL 9000



A journey into the … present



DeepL Traduttore | DeepL Pro | Perché DeepL? | API | Abbonamenti e tariffe | Applicazioni GRATIS | Inizia periodo di prova gratuito | Login

29 lingue | .pdf, .docx, .pptx | DeepL Pro.

Inglese (rilevato) ⇄ Italiano | formale/informale | Glossario

What is Natural Language processing?
By now, I don't know yet

Che cos'è l'elaborazione del linguaggio naturale?
A questo punto, non lo so ancora

Dizionario

amazon | science — Research areas | Blog | News and features | Publications | Conferences | Collaborations | Careers | Feedback

RESEARCH AREA

Conversational AI / Natural-language processing

Building software and systems that help people communicate with computers naturally, as if communicating with family and friends.

Explore more: Automatic speech recognition (ASR) | Dialogue management | Natural-language processing (NLP) | + 2 more

Recent publications

Dialog acts for task-driven embodied agents
Spandana Gella, Aishwarya Padmakumar, Patrick Lange, Dilek Hakkani-Tur | SIGDIAL 2022 | 2022
Embodied agents need to be able to interact in natural language-understanding task descriptions and asking appropriate follow up questions to obtain necessary information to be effective at successfully accomplishing tasks for a wide range of users. In this work, we propose a set of dialog acts for modelling such dialogs and annotate the TEACh dataset… Read More
COMPUTER VISION

MMT4: Multi modality to text transfer transformer
Amir Tavanaei, Karim Bouyarmane, Iman Keivanloo, Ismail Tutar | KDD 2022 | 2022
Recent studies have demonstrated the ability of auto-regressive and seq-to-seq generative models to reach state-of-the-art performance on various Natural Language Understanding (NLU) and Natural Language Processing (NLP) tasks. They operate by framing all the tasks in a single formulation: text auto-completion or text-to-text encoding-decoding. These… Read More
CONVERSATIONAL AI / NATURAL-LANGUAGE PROCESSING

Google — who is the first italian president

Tutti | Notizie | Immagini | Video | Shopping | Altro | Strumenti

Circa 385.000.000 risultati (0,56 secondi)

Suggerimento: Cerca risultati solo in **italiano**. Puoi specificare la lingua di ricerca in Preferenze.

Enrico De Nicola, OMRI (Italian pronunciation: [enˈriːko de niˈkɔːla] (listen); 9 November 1877 – 1 October 1959) was an Italian jurist, journalist, politician, and provisional head of state of republican Italy from 1946 to 1948. Afterwards, he became the first president of Italy on 1 January 1948.

https://en.wikipedia.org › wiki › Enrico_De_Nicola
Enrico De Nicola - Wikipedia

Informazioni sugli snippet in primo piano • Feedback

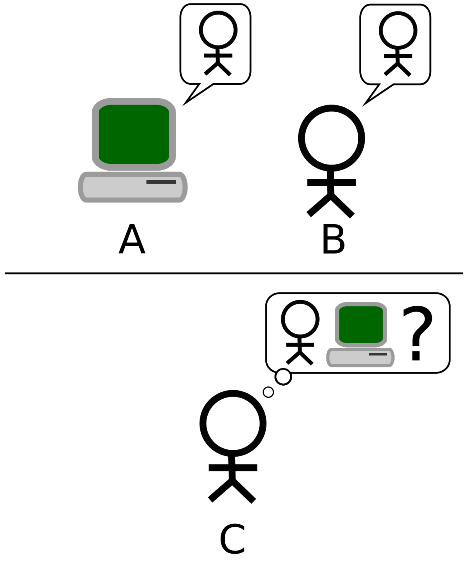
https://en.wikipedia.org › wiki ▾ Traduci questa pagina
List of presidents of Italy - Wikipedia
Top left: Enrico De Nicola was the **first president** of the **Italian** Republic. Top right: Giorgio Napolitano is the longest-serving **president** in **Italian** history.

What is Natural Language Processing?

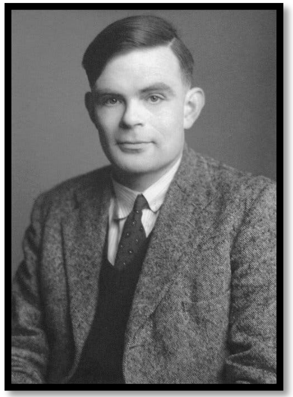
- Natural language processing (NLP) is a field at the intersection of
  - computer science
  - artificial intelligence
  - linguistics
- Goal: Allowing machines to read, understand, and derive meaning from human languages in order to perform tasks that are useful, e.g.,
  - Making appointments, buying things
  - Question Answering
    - Amazon Alexa, Apple Siri, Google Assistant, Facebook M, Microsoft Cortana…
- Fully understanding and representing the meaning of language (or even defining it) is a difficult goal
  - Perfect language understanding is AI-complete
- *N.B.: Terms like 'natural language processing', 'computational linguistics', and 'human language technologies' may be thought of as essentially synonymous*

Natural Language Processing and AI

- When conceiving his famous test in 1950, Alan Mathison Turing, chose a linguistic test to evaluate a machine's ability to match human intelligence



- He devised a chatbot capable of fooling its interlocutor into thinking it was human
- The test does highlight the fact that mastering language is arguably *Homo sapiens*'s greatest cognitive ability



A.M. Turing

1912-1954

What is Natural Language Processing ?

- We’re immersed in the information era: Extremely large and constantly growing amounts of text are produced daily
- Let’s think of several heterogeneous fields
  - Business intelligence
  - Social media
  - Healthcare
  - Finance
  - Human resources
  - Advertising
- The textual data people generate every day exceeds human processing powers
  - The solution is to extract relevant information in some automatic way

What is Natural language processing ?

- The goal of NLP is to be able to design algorithms to allow computers to "understand" natural language in order to perform some task

Applications

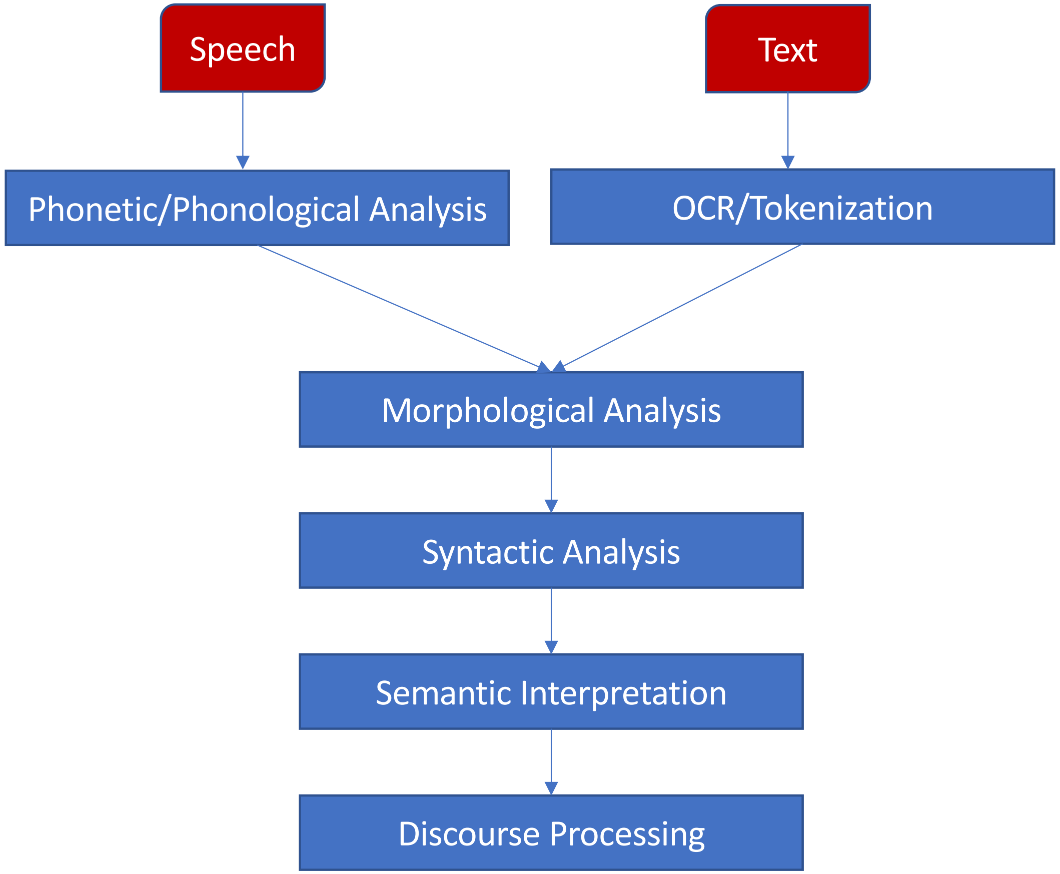
- Machine Translation
  - Google translate, DeepL
- Information retrieval
- Question Answering
  - IBM Watson
- Dialogue Systems (Chatbot & virtual assistant)
  - Siri, Alexa, Google Home
- Information extraction
- Summarization
- Sentiment Analysis
- Document classification & fake news detection
- …

Core technologies

- Language Modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic Role Labelling
- …

What does an NLP system need to know ?

- Language consists of many levels of structures
- Humans fluently integrate of all these in producing/understanding language
- Ideally, so would a computer!



Speech

Text

Phonetic/Phonological Analysis

OCR/Tokenization

Morphological Analysis

Syntactic Analysis

Semantic Interpretation

Discourse Processing

Words

This is a simple sentence WORDS

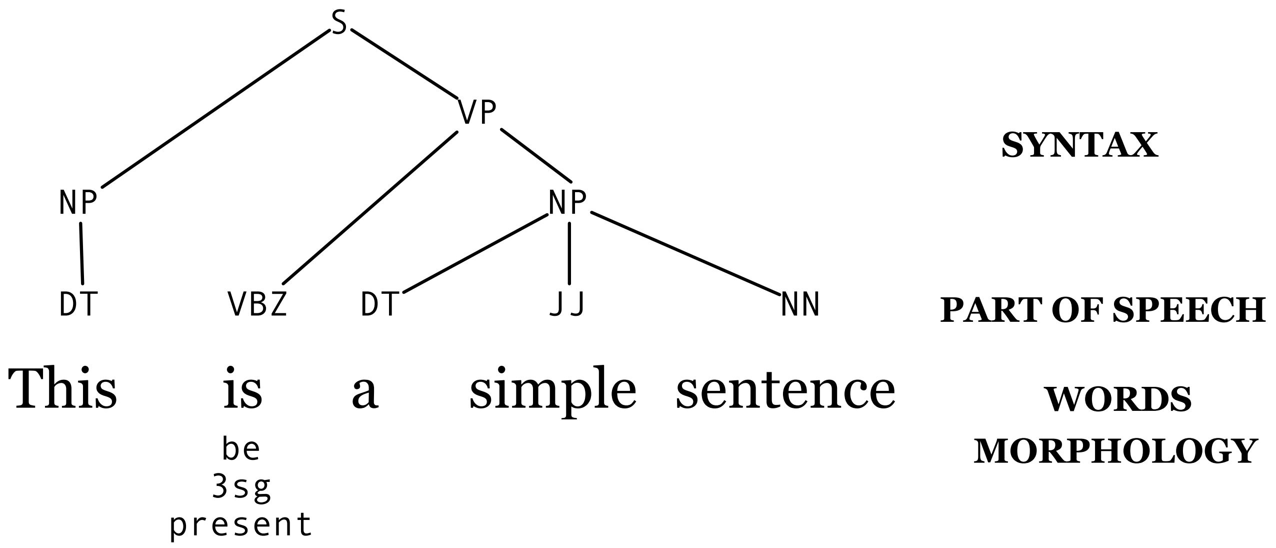
Morphology

This	is	a	simple	sentence	**WORDS**
	be 3sg present				**MORPHOLOGY**

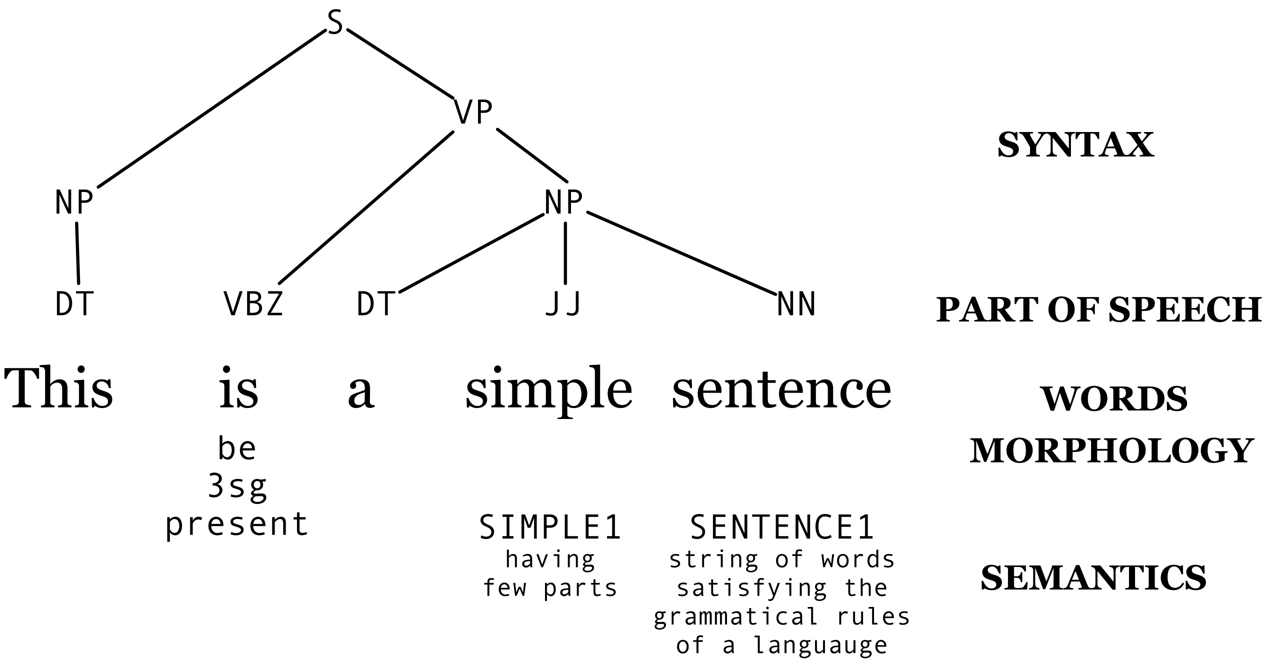
Part of Speech

DT	VBZ	DT	JJ	NN	**PART OF SPEECH**
This	is	a	simple	sentence	**WORDS**
	be 3sg present				**MORPHOLOGY**

Syntax

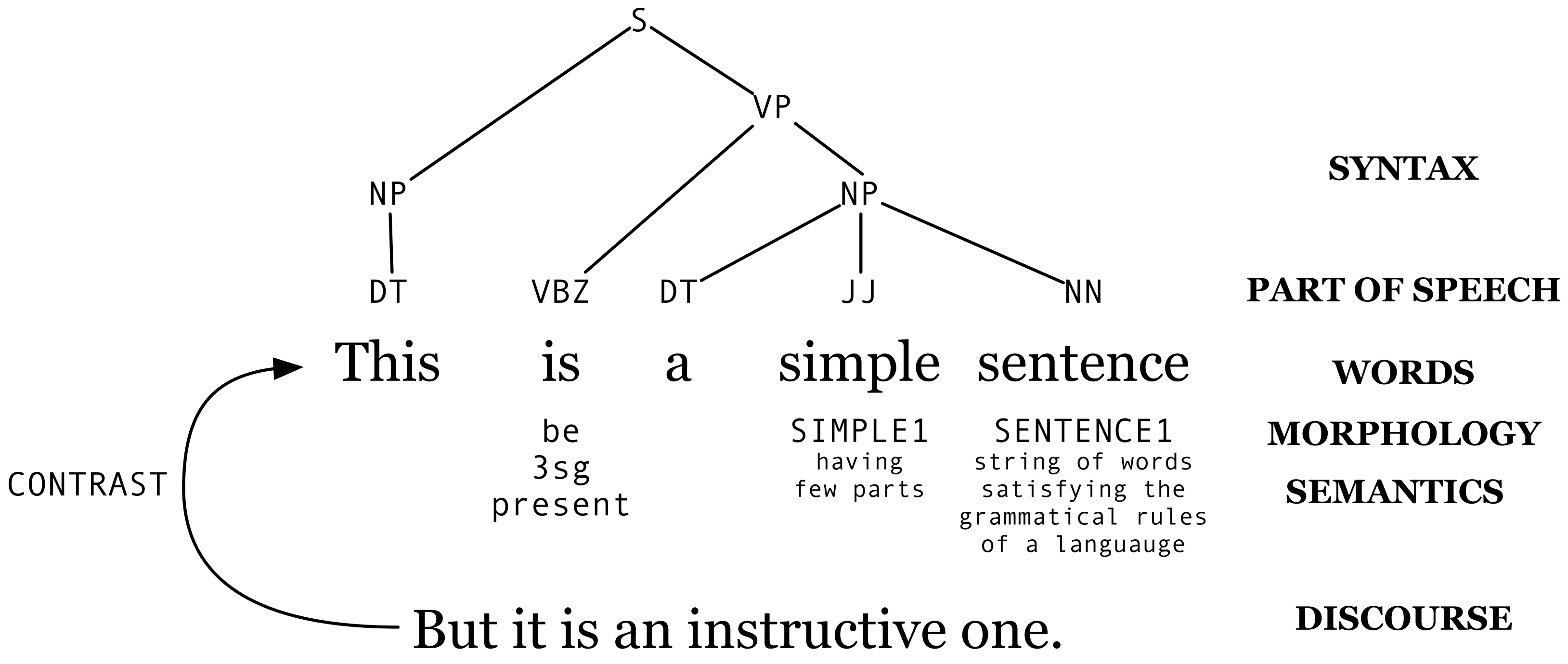

S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This
is
a
simple
sentence
be
3sg
present
SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY

Semantics



$$\exists y(\textit{this\_dem}(x) \wedge \textit{be}(e, x, y) \wedge \textit{simple}(y) \wedge \textit{sentence}(y))$$

Discourse


S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This
is
a
simple
sentence
be
3sg
present
SIMPLE1
having
few parts
SENTENCE1
string of words
satisfying the
grammatical rules
of a languauge
CONTRAST
But it is an instructive one.
SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
SEMANTICS
DISCOURSE

Development of NLP

- Rules
  - Define set of rules to emulate language

- Statistics
  - Analyze large multilingual set of text (corpus)
  - Requires extensive work to define context

- Machine learning and NN
  - Semantic information learned from data

Why is NLP challenging?

- NLP distinguishes itself from other AI application domains, as for instance computer vision or speech processing
- Text data is fundamentally discrete. But **new words** can always be created
  - *Stan: an extremely enthusiastic and devoted fan (stalker-fan)*
  - *Nomophobia: anxiety caused by not having a working mobile phone*

Why is NLP challenging?

1. Ambiguity
   - Phonetic transcription [raIt] might mean write, right, rite
   - Word senses: bank (finance or river ?)
   - Word can belong to several categories: noun, verb, or modal
     - Part of speech: chair (noun or verb ?)
   - Syntactic structure: *I saw a man with a telescope*
   - Multiple*: I made her duck*
   - Reference: *The son asked the father to drive him home*
   - Discourse: *The meeting is cancelled. Nicholas isn’t coming to the office*

Ambiguity

- Resolving ambiguity is hard
- There are at least half a dozen meanings of this sentence:

`The chef made her duck`

Ambiguity

`The chef made her duck`

- The cook cooked waterfowl for a different woman X (person using "she/her" pronouns) to eat
- The cook cooked waterfowl belonging to X
- The cook cooked waterfowl belonging to the cook
- The cook created the (plaster?) waterfowl that X owns
- The cook caused X to quickly lower X's head or body
- The cook waved their magic wand and turned X into undifferentiated waterfowl

Ambiguity

The chef caused X to quickly lower her head or body

- Part of speech: "duck" can be a Noun or Verb

The chef cooked waterfowl for X (or belonging to X)

- Part of speech:
  - "her" is possessive pronoun ("of her")
  - "her" is dative pronoun ("for her")

The chef cooked waterfowl belonging to the chef (vs to X)

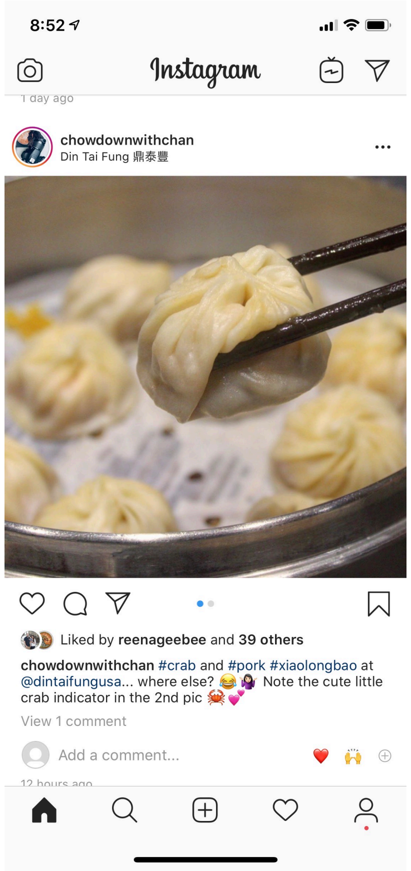
- Coreference
  - "her" can refer to X or to the Chef

The chef made the (plaster) duck statue X (or the chef) owns

- Word Meaning : "make" can mean "create" or "cook"

More difficulties

- Non-standard language, emojis, hashtags, names

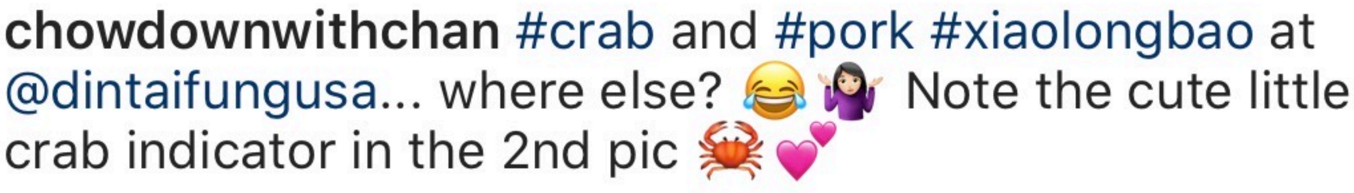


Liked by reenageebee and 39 others

chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀💕

View 1 comment

Add a comment...



chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂 🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀 💕

Why is NLP challenging?

2. Sparse data due to Zipf’s law
   - To illustrate, let’s look at the frequencies of different words in a large text corpus
   - Assume a “word” is a string of letters separated by spaces (a great oversimplification, we’ll return to this issue)

Word counts

- Most frequent words (word types) in the English Europarl corpus (out of 24m word tokens)

any word

Frequency	Type
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns

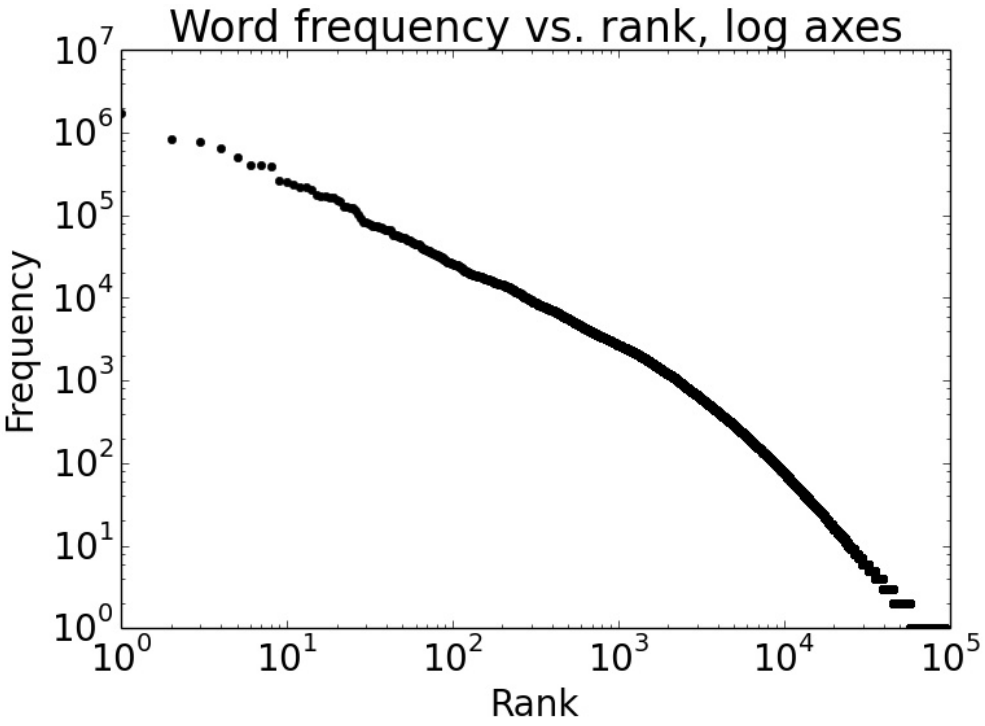
Frequency	Type
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

Word counts

- But also, out of 93638 distinct word types, 36231 occur only once
  - Cornflakes, mathematicians, fuzziness, jumbling
  - Pseudo-rapporteur, lobby-ridden, perfunctorily
  - Lycketoft, UNCITRAL, H-0695
  - Policyfor, Commissioneris, 145.95, 27a

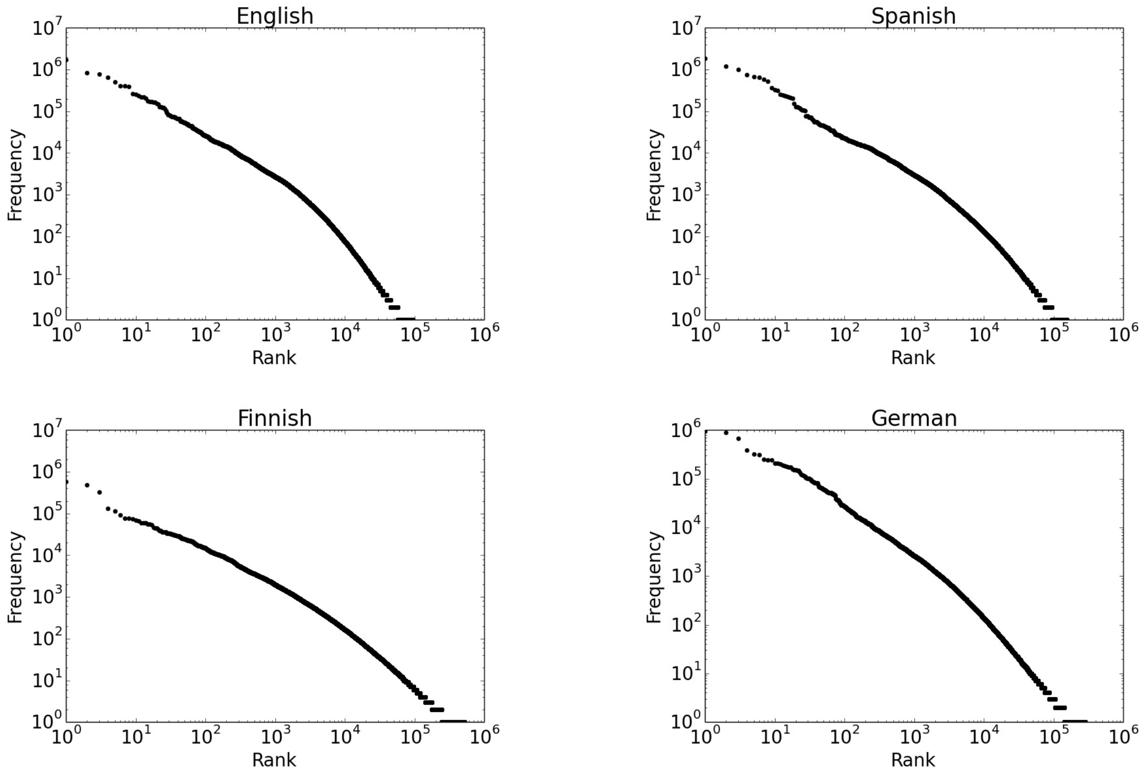
Plotting word frequencies

- Order word by frequency. What is the frequency of the n-th ranked word?



Zipf’s law

- Summarizes the behavior we’re observing here



$$f \times r \approx k$$

f = frequency of a word

r = rank of a word (if sorted by frequency)

k = constant

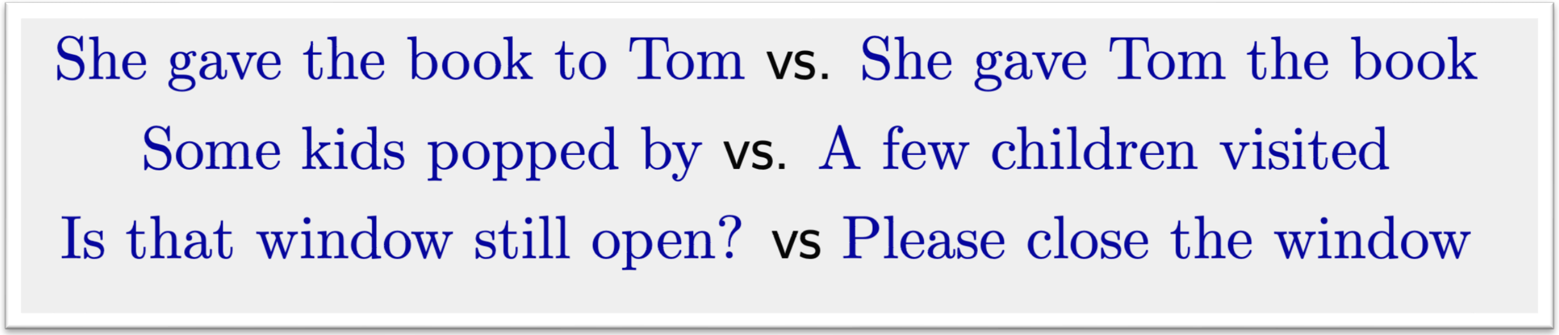
Implications of Zipf's law

- Few words are very frequent, and there is a long tail of rare words
  - Regardless of how large our **corpus** is, there will be a lot of infrequent (and zero-frequency) words
- In a classification problem, for instance, this means we need to find clever ways to estimate probabilities for things we have rarely or never seen during training

Why is NLP challenging?

3. Expressivity

- Not only can one form have different meanings (ambiguity), but the same meaning can be expressed with different forms:



She gave the book to Tom vs. She gave Tom the book

Some kids popped by vs. A few children visited

Is that window still open? vs Please close the window

Why is NLP challenging?

4. Context dependence and unknown representation
    - the correct interpretation is context-dependent and often requires world knowledge
    - very difficult to capture, since we don’t even know how to represent the knowledge a human has/needs
        - What is the ”meaning” of a word or sentence?
        - How to model context?
        - Other general knowledge?
    - That is, in the limit NLP is hard because AI is hard
        - In particular, we’ve made remarkably little progress on the knowledge representation problem

Learning & Knowledge

- Rationalism
  - *A significant part of the knowledge in the human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance*
    - Noam Chomsky Poverty of the stimulus, 1980
- Generative linguists have argued for the existence of a **language faculty** in all human beings, which encodes a set of abstractions specially designed to facilitate the understanding and production of language

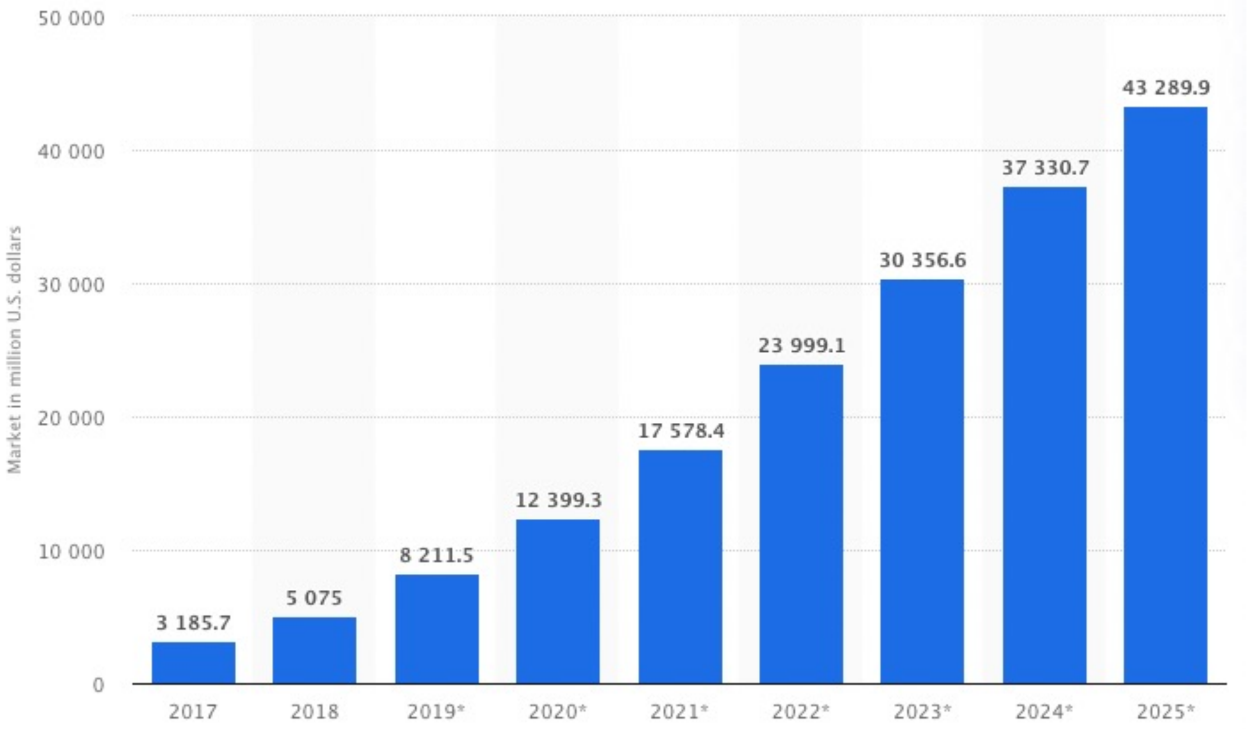
Learning & Knowledge

- Empiricism
  - *The view that there is no such thing as innate knowledge, and that knowledge is instead derived from experience, either sensed via the five senses or reasoned via the brain or mind.*
    - Originated in ancient Hindu and Greek philosophy
- Nowaday, many statistical NLP techniques work very well on texts, without the need to use special bias representing linguistic knowledge or mental representation of language

Learning & knowledge

- A recurring topic of debate in NLP is the relative importance of machine learning vs. linguistic knowledge
- 1950s: Empiricism was at its peak, dominating a broad set of fields ranging from psychology (behaviorism) to electrical engineering (information theory) (Shannon, Skinner, Firth, Harris).
- 1970s: Empiricism faded after a few significant events, including Chomsky's criticism of n-grams in Syntactic Structures and Minsky and Papert's criticism of neural networks
- Most NLP systems based on hand-written rules, developed by linguists
- 1990s: Data was becoming available like never before, reviving empiricist approaches (IBM Speech Group, AT&T Bell Labs)
  - *Every time I fire a linguist, the performance of the speech recognizer goes up* (Frederick Jelinek, IBM)

Market



- The NLP market is predicted to be almost 14 times larger in 2025 than it was in 2017, increasing from around three billion U.S. dollars to over 43 billion.

NLP in Finance

- In finance, data can help make timely decisions comes in text. Earnings reports are one example. A company will release its report in the morning, and it will say "Our earnings per share were a $1.12."
- By the time that unstructured data makes its way into a database of a data provider where you can get it in a structured way, hours have passed, and you've lost your edge
- NLP can deliver these transcriptions in minutes, giving analysts a competitive advantage

NLP in Social Networks

- Automated fake news detection
- Fake news refers to information content that is false, misleading or whose source cannot be verified. Automatic approaches to fake news detection involve NLP
- An example, companies like Facebooks, Twitter, TikTok, Google, Pinterest, Tencent, YouTube, and others are working with the World Health Organization to mitigate COVID-19 driven infodemic

NLP in Healthcare

- Huge volumes of unstructured patient data is inputted into electronic health record systems
  - 80% of healthcare documentation is unstructured text
- Healthcare NLP uses specialized engines capable of discovering previously missed or improperly coded patient conditions

The course

Course logistics

- Instructor: Antonino Staiano
- Time & location
  - Monday
    - 16:00 – 18:00
    - Room: Lab 1 (2nd floor, south side)
  - Thursday
    - 14:00 – 16:00
    - Room: Lab 1
- Office hours
  - Monday 14:30 – 15:30

Learning stuff

- Reference text
  - Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd ed, draft)
    - https://web.stanford.edu/~jurafsky/slp3/
  - Bird, Klein, & Loper, Natural Language Processing with Python, (2009) O'Reilly Media
    - http://www.nltk.org/book/
- Lecture slides
  - E-learning platform
- Grading
  - Oral + …

What this course is

- Introductory/survey course
  - Introduces many of the core activities in NLP and discuss why they are challenging
  - Presents linguistic concepts and standard methods (algorithms) often used to solve these task
  - Provides sufficient background to be able to read (some) current research papers on NLP
- We will NOT
  - Say too much about cutting edge methods or heavy-duty machine learning