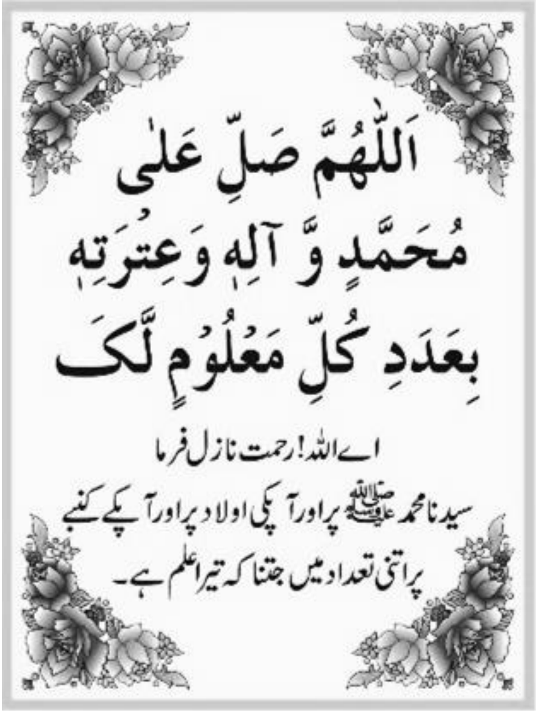
بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

Natural Language Processing (NLP)

Session 01 – Introduction to NLP

Instructor: Dr. Jawad Shafi

Dua – Take Help from Allah Before Starting Any Task



اَللّٰهُمَّ صَلِّ عَلٰى
مُحَمَّدٍ وَّ آلِهٖ وَعِتْرَتِهٖ
بِعَدَدِ کُلِّ مَعْلُوْمٍ لَّک

اے اللہ! رحمت نازل فرما
سیدنا محمد ﷺ پراور آپکی اولاد پراور آپکے کنبے
پراتنی تعداد میں جتنا کہ تیرا علم ہے۔

اللَّهُمَّ خِرْ لِيْ وَاخْتَرْ لِي

سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا
إِنَّكَ أَنْتَ الْعَلِيمُ الْحَكِيمُ

رَبِّ اشْرَحْ لِي صَدْرِي
وَيَسِّرْ لِي أَمْرِي
وَاحْلُلْ عُقْدَةً مِنْ لِسَانِي
يَفْقَهُوا قَوْلِي

Dr. Jawad Shafi – About Me

PhD

Lancaster University, UK

Assistant Professor

COMSATS University Islamabad, Lahore Campus

Group Member

NLP Group, CUI, Lahore Campus

Course Details - Instructor

Instructor	***Dr. Jawad Shafi***
Email	***jawadshafi@cuilahore.edu.pk***
Office	***Room no. 124, Faculty Block***
CUI Profile	***https://lahore.comsats.edu.pk/Employees/551***

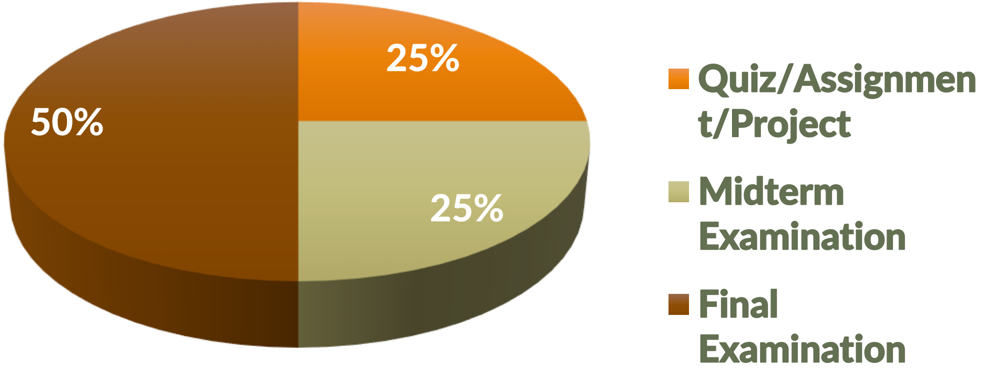
Course Details – For MS/PHD Course (Cont.)

Google Classroom Code: k2mt767

Note: Join using CUI-Lahore email ID

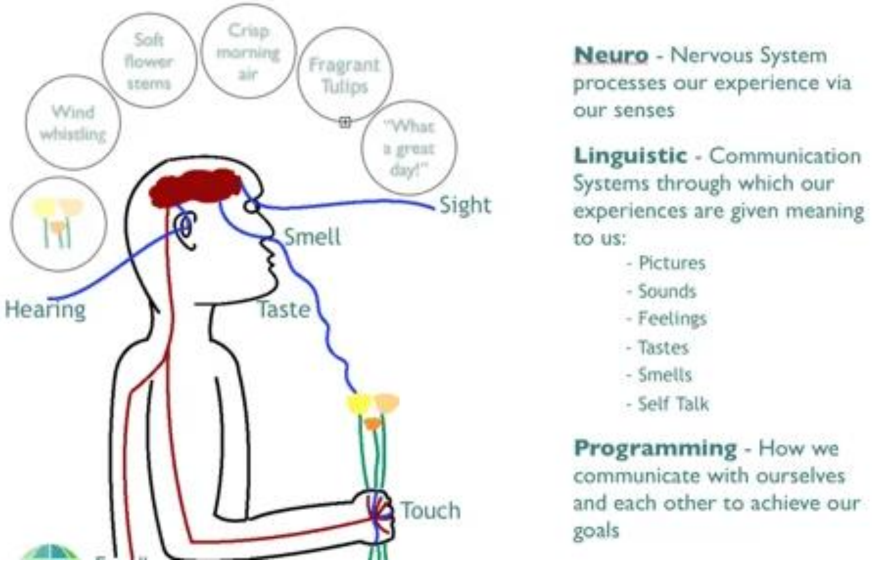
Office Hours: ***Email requests for appointment***

Assessment:



Basic Text Processing

NLP



Neuro - Nervous System processes our experience via our senses

Linguistic - Communication Systems through which our experiences are given meaning to us:

- Pictures
- Sounds
- Feelings
- Tastes
- Smells
- Self Talk

Programming - How we communicate with ourselves and each other to achieve our goals

Welcome to NLP Course! (by ChatGPT!)

Welcome to the Natural Language Processing course!

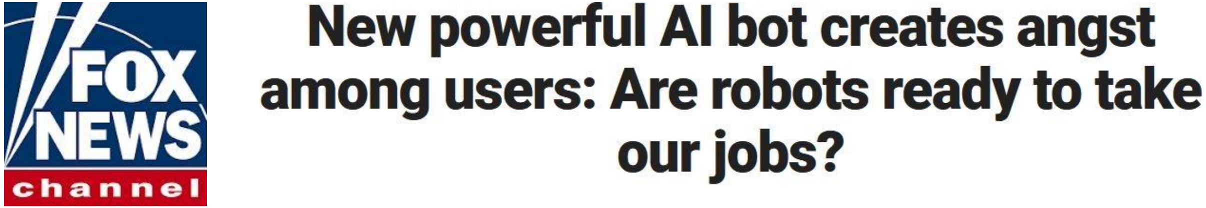
NLP is an exciting and rapidly growing field that deals with the interaction between computers and human language.

In this course, you will learn about the techniques and algorithms used to analyze and understand human language, and you will have the opportunity to apply these techniques to real-world problems.

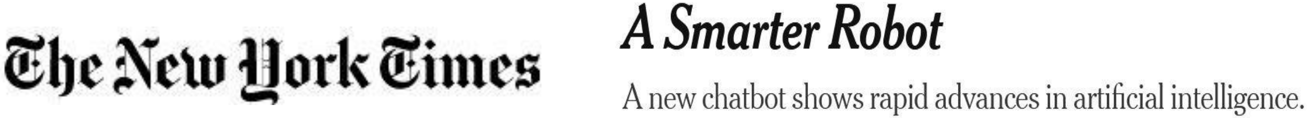
Whether you are a computer science student, a linguist, or just someone with an interest in language and technology, this course will provide you with a solid foundation in NLP and its applications.

Let's dive in and discover the amazing possibilities of NLP together!

NLP is the KING!

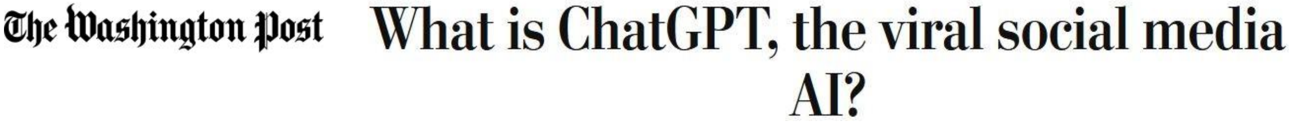


New powerful AI bot creates angst among users: Are robots ready to take our jobs?

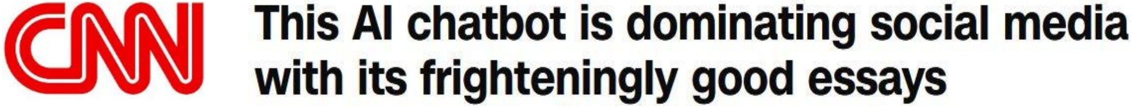


A Smarter Robot

A new chatbot shows rapid advances in artificial intelligence.

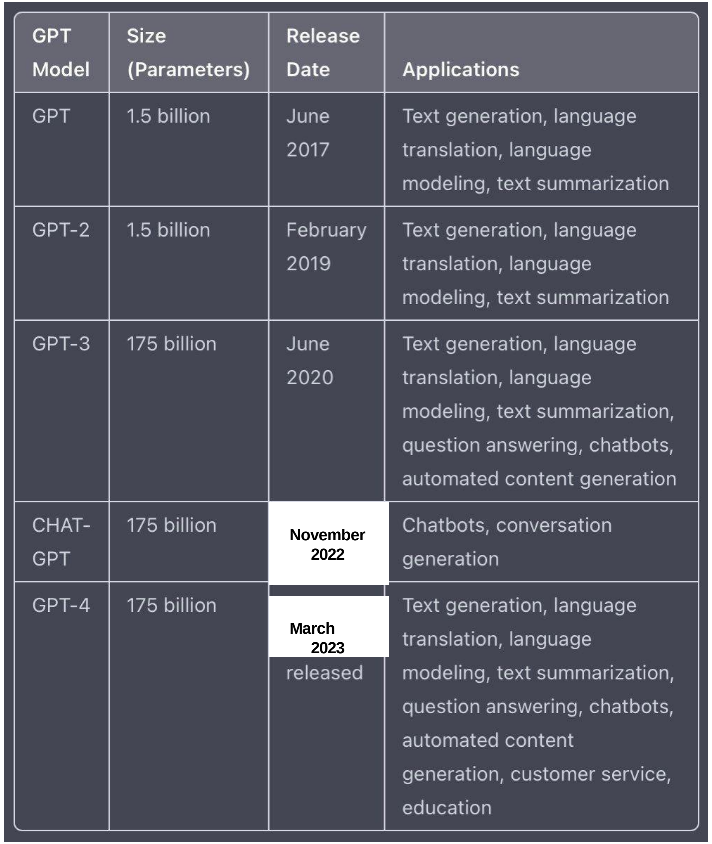


What is ChatGPT, the viral social media AI?

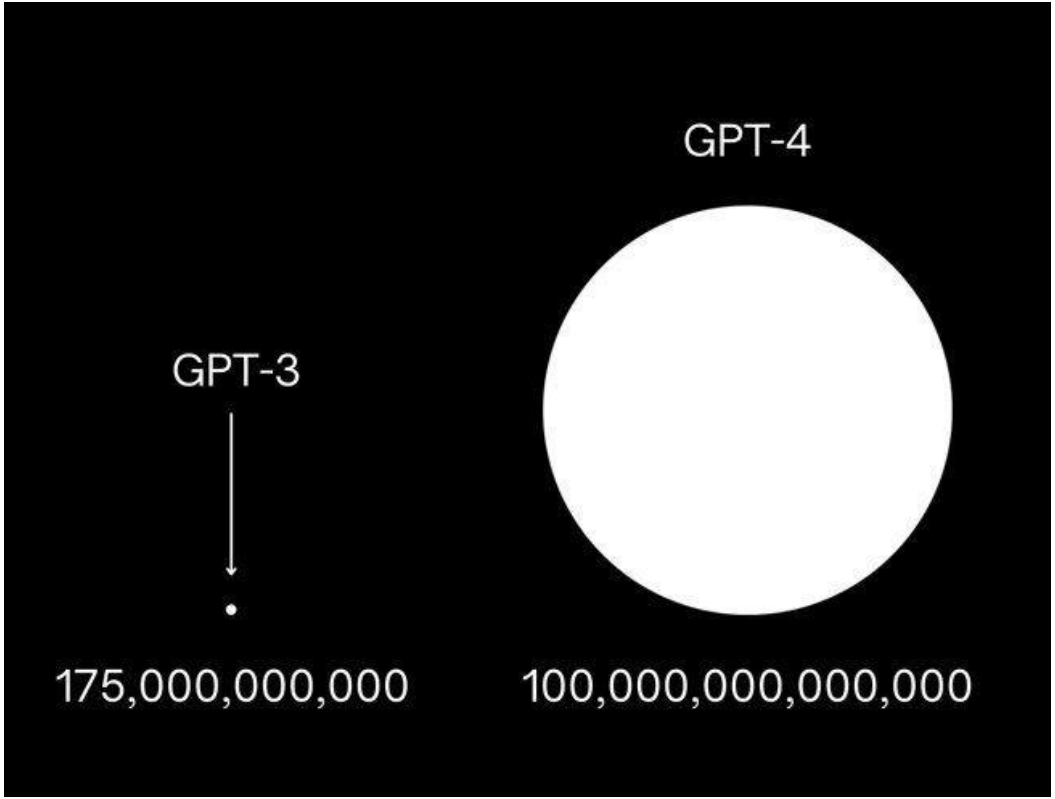


This AI chatbot is dominating social media with its frighteningly good essays

We are here GPT-4

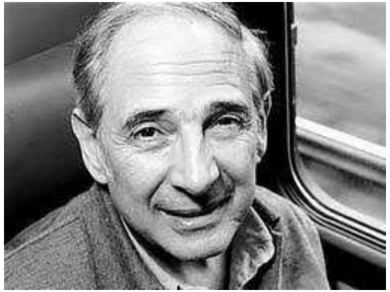


GPT Model	Size (Parameters)	Release Date	Applications
GPT	1.5 billion	June 2017	Text generation, language translation, language modeling, text summarization
GPT-2	1.5 billion	February 2019	Text generation, language translation, language modeling, text summarization
GPT-3	175 billion	June 2020	Text generation, language translation, language modeling, text summarization, question answering, chatbots, automated content generation
CHAT-GPT	175 billion	November 2022	Chatbots, conversation generation
GPT-4	175 billion	March 2023 released	Text generation, language translation, language modeling, text summarization, question answering, chatbots, automated content generation, customer service, education



"Natural language is the most important part of artificial intelligence."

John Searle



"Natural language processing is a cornerstone of artificial intelligence, allowing computers to read and understand human language, as well as to produce and recognize speech."

Ginni Rometty

"Natural language processing is one of the most important fields in artificial intelligence and also one of the most difficult."

Dan Jurafsky

Why Natural Language Processing?

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

Why NLP ?

- **Access Knowledge** (search engine, recommender system…)
- **Communicate** (e.g. Translation)
- **Linguistics** and **Cognitive Sciences** (Analyse Languages themselves)

Why Natural Language Processing?

Amount of online textual data…

- **70 billion web-pages online** (1.9 billion websites)
- **55 million Wikipedia articles**

…Growing at a fast pace

- **9000 tweets/second**
- **3 million mail / second** (60% spam)

Internet Stats

Why Natural Language Processing?

Potential Users of Natural Language Processing

- **7.9 billion** people use some sort of language (January 2022)
- **4.7 billion internet users** (January 2021) (~59%)
- **4.2 billion** social media users (January 2021) (~54%)

Data Reportal

Why Natural Language Processing?

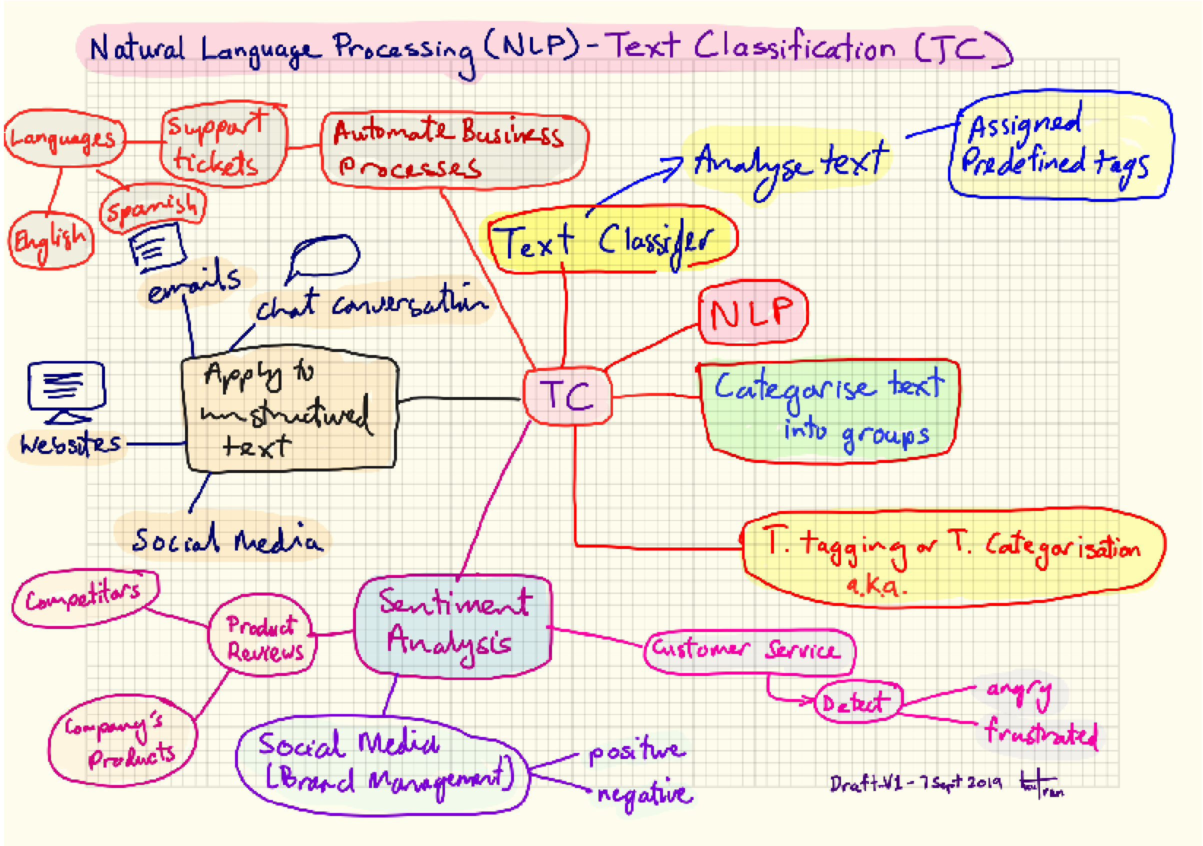
What Products ?

- Search: **+2 billion** Google users, **700 millions** Baidu users
- Social Media: **+3 billion** users of Social media (Facebook, Instagram, WeChat, Twitter...)
- Voice assistant: **+100 million** users (Alexa, Siri, Google Assistant etc)
- Machine Translation: **500M** users for google translate

Data Reportal

Basic Text Processing

NLP is Hard to model



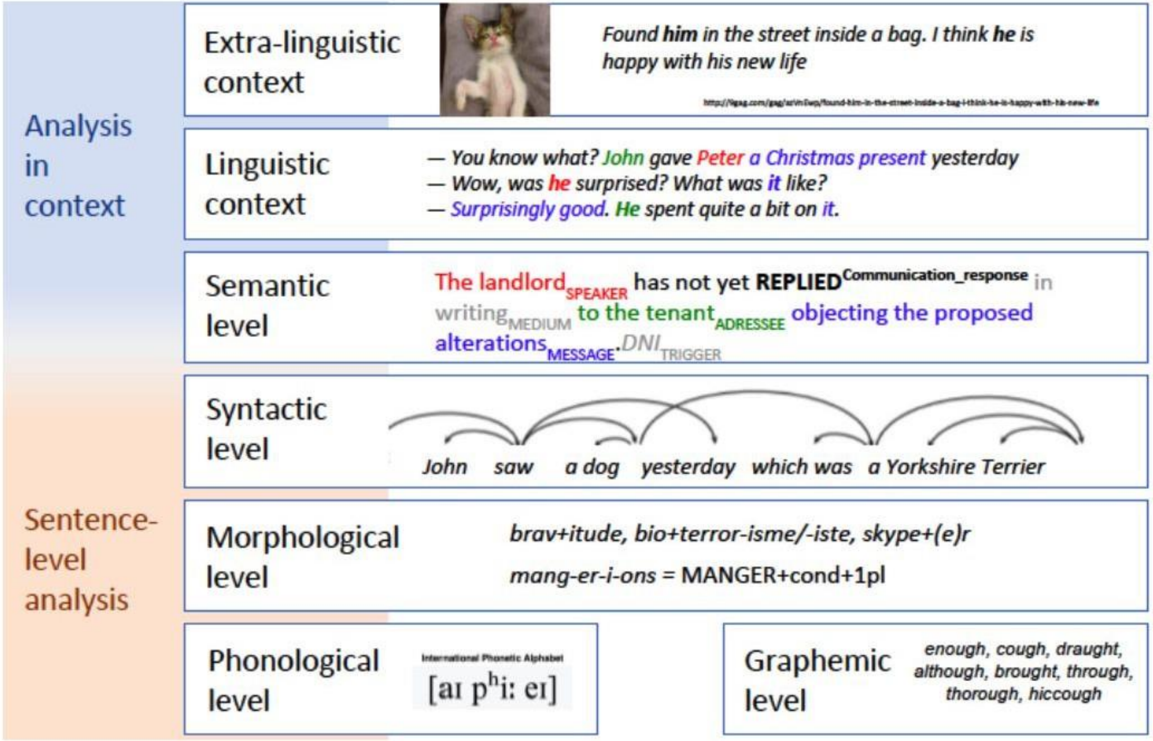
A Definition of Language

Definition 1: *Language is a means to communicate, it is a semiotic system. By that we simply mean that it is a **set of signs**. A sign is a pair consisting in [...] **a signifier and a signified**.*

Definition 2: *A sign consists in a phonological structure, a morphological structure, a syntactic structure and a semantic structure*

Krach 2007

The Six Levels of Linguistics Analysis



Analysis in context

Sentence-level analysis

Extra-linguistic context

*Found **him** in the street inside a bag. I think **he** is happy with his new life*

http://9gag.com/gag/arVnGwp/found-him-in-the-street-inside-a-bag-i-think-he-is-happy-with-his-new-life

Linguistic context

— *You know what? John gave Peter a Christmas present yesterday*
— *Wow, was **he** surprised? What was **it** like?*
— *Surprisingly good. **He** spent quite a bit on it.*

Semantic level

The landlord$_{\text{SPEAKER}}$ has not yet **REPLIED**$^{\text{Communication\_response}}$ in writing$_{\text{MEDIUM}}$ to the tenant$_{\text{ADRESSEE}}$ objecting the proposed alterations$_{\text{MESSAGE}}$.*DNI*$_{\text{TRIGGER}}$

Syntactic level

Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r

mang-er-i-ons = MANGER+cond+1pl

Phonological level

International Phonetic Alphabet

[aɪ pʰiː eɪ]

Graphemic level

enough, cough, draught, although, brought, through, thorough, hiccough

Knowledge Requirement for Machine

- **Phonetics and Phonology**: knowledge about linguistic sounds
- **Morphology**: knowledge of the meaningful components of words
- **Syntax**: knowledge of the structural relationships between words
- **Semantics**: knowledge of meaning
- **Pragmatics**: knowledge of the relationship of meaning to the goals and intentions of the speaker
- **Discourse**: knowledge about linguistic units larger than a single utterance

Phonetics and Phonology

- **Phonetics and Phonology**: knowledge about linguistic sounds
- The study of:

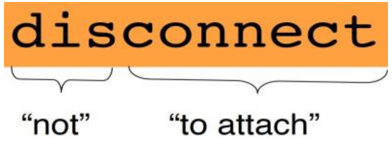
  language sounds how they are physically formed;

  systems of discrete sounds, e.g. languages' syllable structure

dis-koo-nekt **disconnect**

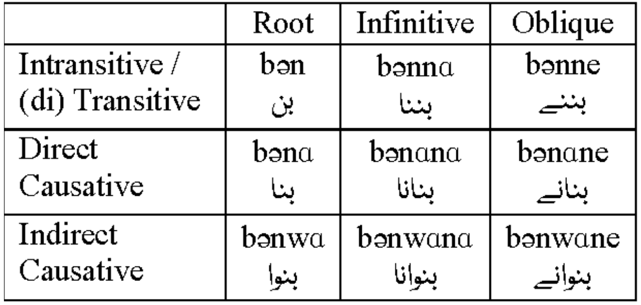
Morphology

- **Morphology:** knowledge of the meaningful components of words
- The study of the sub-word units of meaning



Even more necessary in some other languages,

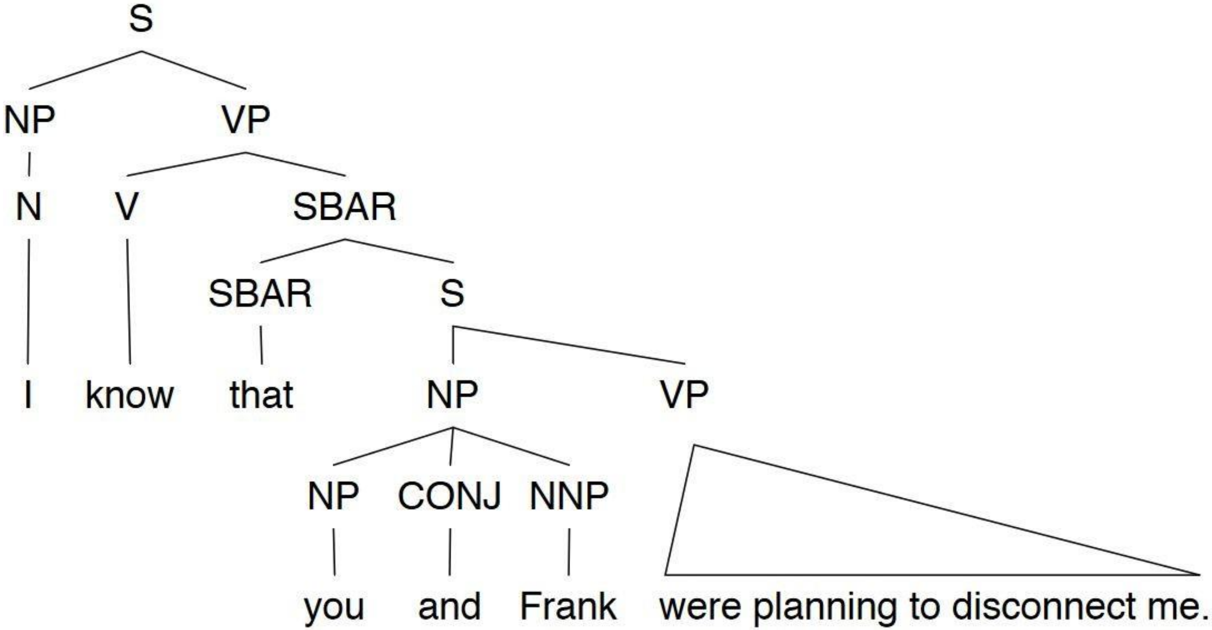
e.g. Urdu →



	Root	Infinitive	Oblique
Intransitive / (di) Transitive	bən بن	bənna بننا	bənne بننے
Direct Causative	bəna بنا	bənana بنانا	bənane بنانے
Indirect Causative	bənwa بنوا	bənwana بنوانا	bənwane بنوانے

Syntax

- Syntax: knowledge of the structural relationships between words
- The study of the structural relationships between words
  - I know that you and Frank were planning to disconnect me.



Semantics

- **Semantics**: knowledge of meaning
- The study of the literal meaning
    - ○ I know that you and Frank were planning to disconnect me.
    - ○ ACTION = disconnect
    - ○ ACTOR = you and Frank
    - ○ OBJECT = me

Pragmatics

- **Pragmatics**: knowledge of the relationship of meaning to the goals and intentions of the speaker
- The study of how language is used to accomplish goals
  - How should you react?
    - I’m sorry Dave, I’m afraid I can’t do that.

    Or

    - if one person asked, "What do you want to eat?" and another responded, "Ice cream is good this time of year.“

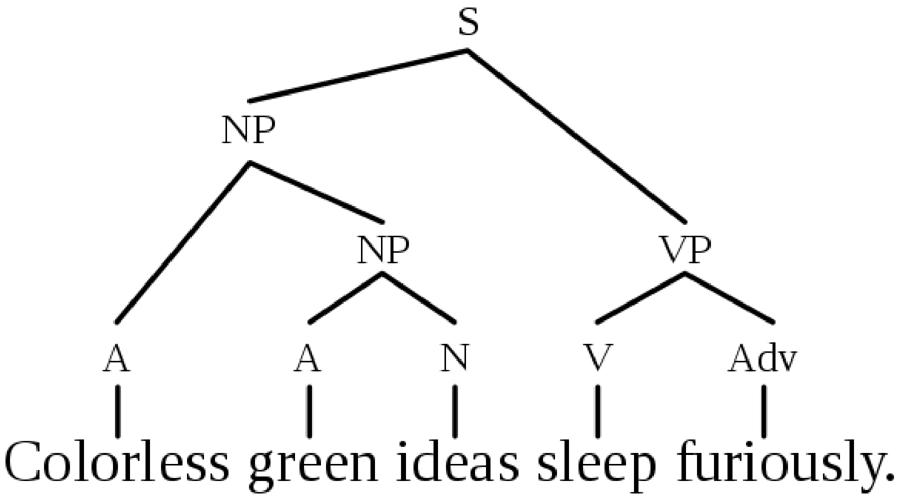
Discourse

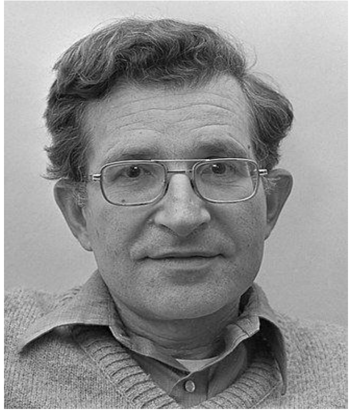
- **Discourse**: knowledge about linguistic units larger than a single utterance
- The study of linguistic units larger than a single utterance
- The structure of conversations:
  - turn taking, thread of meaning
    - For example, **if you are debating the value of buffalo chicken wings versus BBQ chicken with a friend**; you are engaged in discourse.

Syntax vs. Semantics

Colorless green ideas sleep furiously.

(example by Noam Chomsky 1957)





Noam Chomsky
The most cited person alive

Semantics vs. Pragmatics

What does "You have a green light" mean?

- □You are holding a green light bulb?
- □You have a green light to cross the street?
- □You can go ahead with your plan?

The 5 Challenges of NLP

1. Productivity
2. Ambiguous
3. Variability
4. Diversity
5. Sparsity

Productivity

Definition
“property of the language-system which enables native speakers to construct and understand an indefinitely large number of utterances, including utterances that they have never previously encountered.” (Lyons, 1977)

➔ **New words, senses, structure** are **introduced in languages all the time**

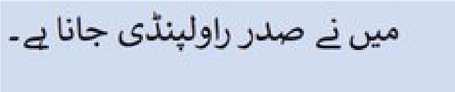
Examples: ***staycation*** and ***social distance*** were added to the Oxford Dictionary in 2021

Ambiguous

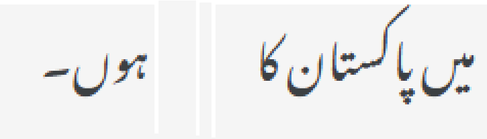
Most linguistic observations (speech, text) are open to **several interpretations**

We (Humans) disambiguate -i.e. **find the correct interpretation -** using all kind of signals (linguistic and extra linguistic)

Ambiguity can appear at all levels (phonology, graphemics, morphology, syntax, semantics)



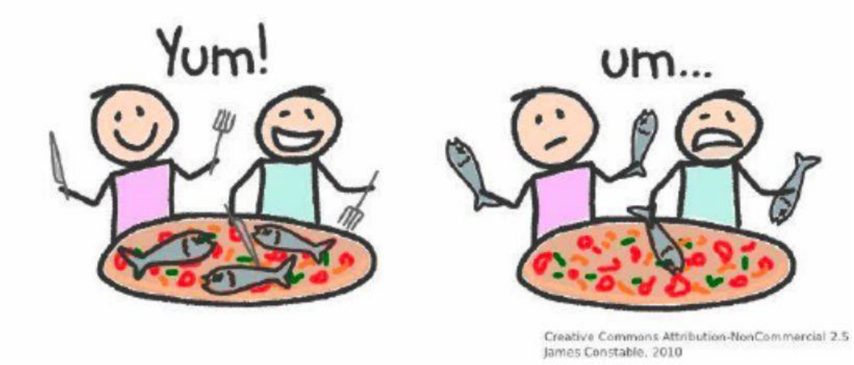
میں نے صدر راولپنڈی جانا ہے۔



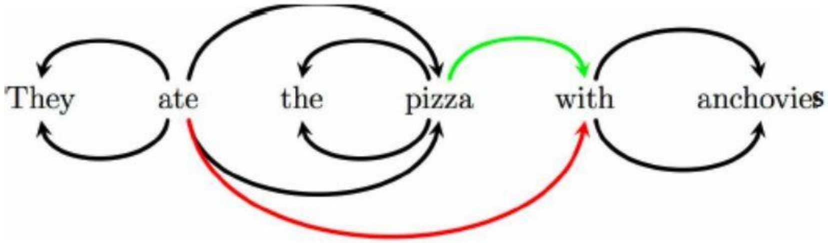
میں پاکستان کا ہوں۔

Ambiguous

Syntactic Ambiguity



Creative Commons Attribution-NonCommercial 2.5
James Constable, 2010



cf. Sagot

Ambiguous

Semantic Ambiguity

- Polysemy: *e.g.* ***set , arm, head***

  Head of New-Zealand is a woman

- Name Entity: *e.g.* ***Michael Jordan***

  Michael Jordan is a professor at Berkeley

- Object/Color: *e.g.* ***cherry***

  Your cherry coat

Ambiguous

Pragmatic Ambiguity

*Two Soviet ships collide, **one dies***

*Dealers will hear **car talk** at noon*

Ambiguous

Disambiguating can requires Discourse Knowledge

Where can I find **a vegetarian restaurant** in **Lahore**

Here is a list of restaurant in Lahore: ….

Give me the top ranked ones, in the 14th Lake City

Here are the top ranked restaurant in the 14th Lake City in Lahore

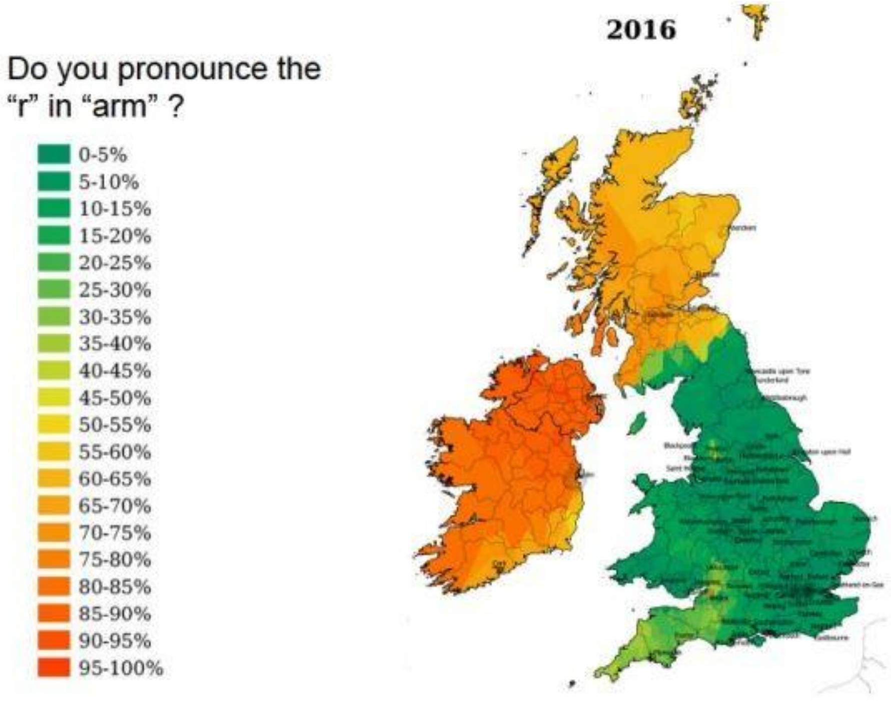
How far is the closest **one** from my current location?

Variation

Language Varies at all levels

- Phonetic (accent)
- Morphological, Lexical (spelling)
- Syntactic
- Semantic

Phonetic Variation



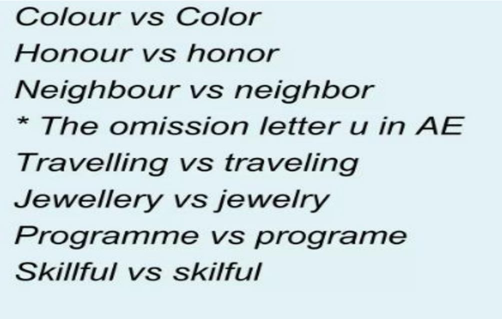
Spelling and Syntactic Variation

سعودی وزیر خارجہ کی پیرس میں مصنوعی ذہانت ایکشن سمٹ میں شرکت

Google Translaiton

سعودی ڈیٹا اینڈ آرٹیفیشل انٹیلی جنس اتھارٹی کے چیئرمین بھی شریک ہیں

Bing Translation



Colour vs Color
Honour vs honor
Neighbour vs neighbor
** The omission letter u in AE*
Travelling vs traveling
Jewellery vs jewelry
Programme vs programe
Skillful vs skilful

Variation Determiners

- Who is talking?
- To Whom?
- Where? *Work, Home, Restaurant*
- When? *19th century, 2008, 2022…*
- About what? *Specialised domain, the Weather,…*

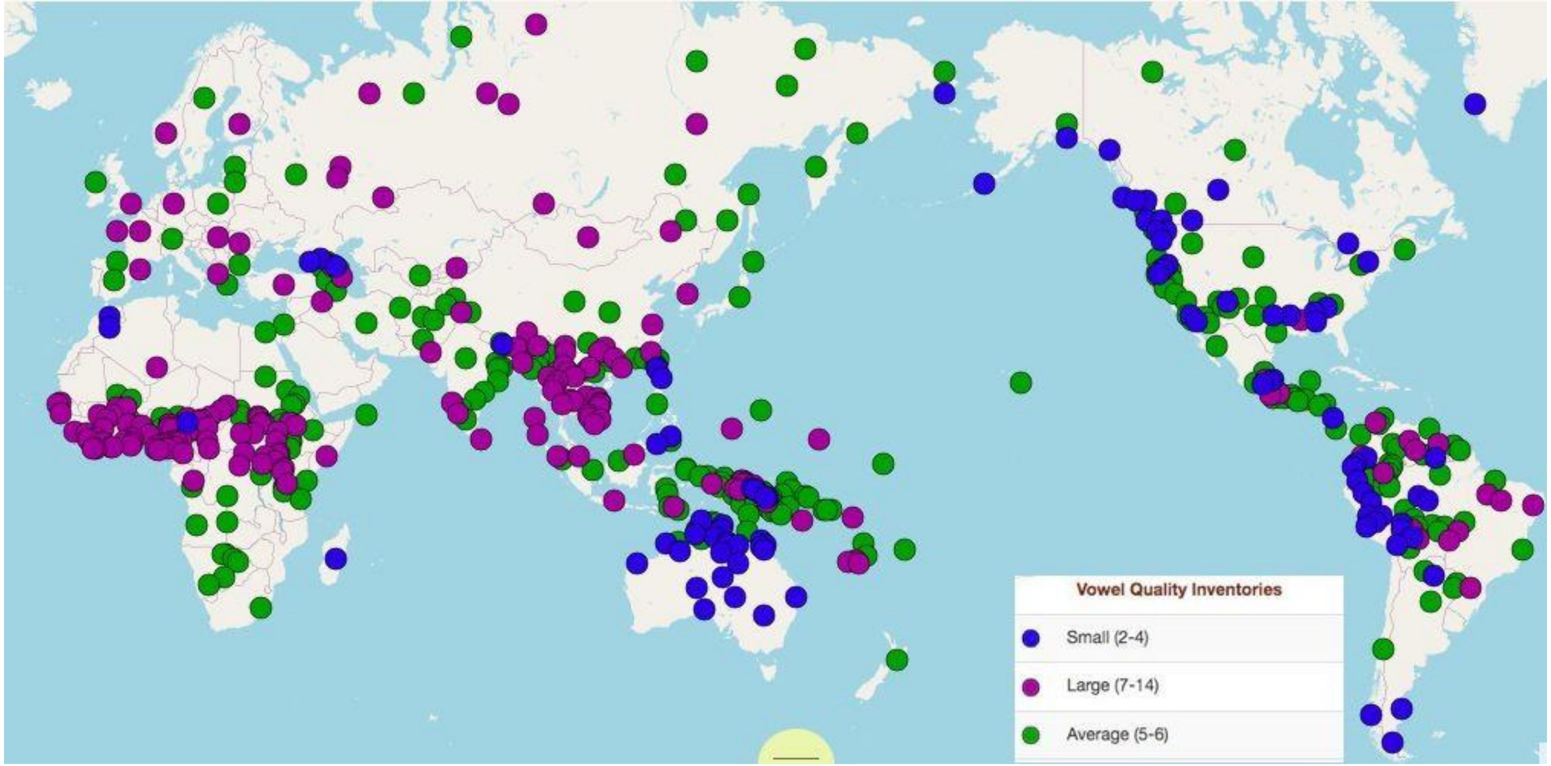
Essentially, the Variability of a language depends on:

- Social Context
- Geography
- Sociology
- Date
- Topic

Diversity

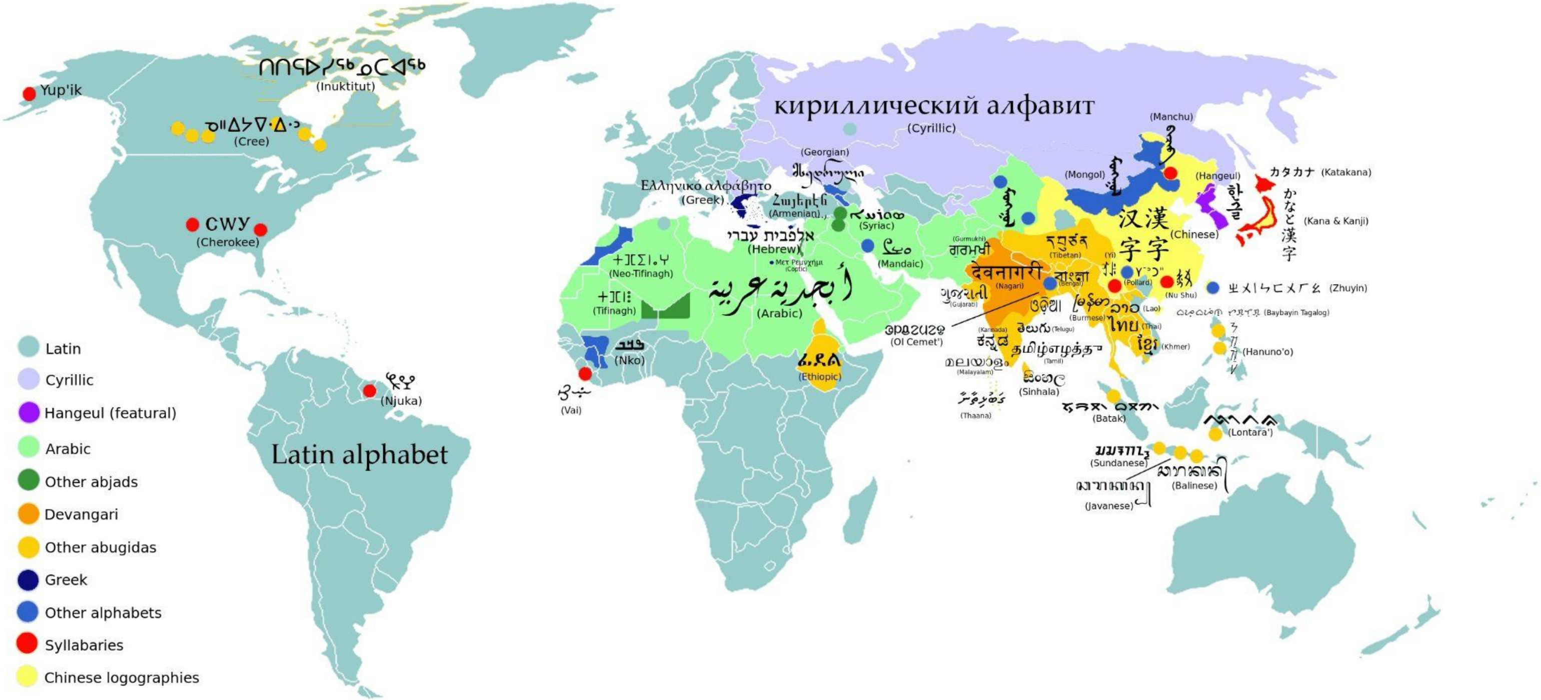
- About **7000 languages** spoken in the world
- About **60%** are found in the **written form** (Ref. Omniglot)

Phonologic Diversity



(Dyer et. al 2013)

Graphemic Diversity



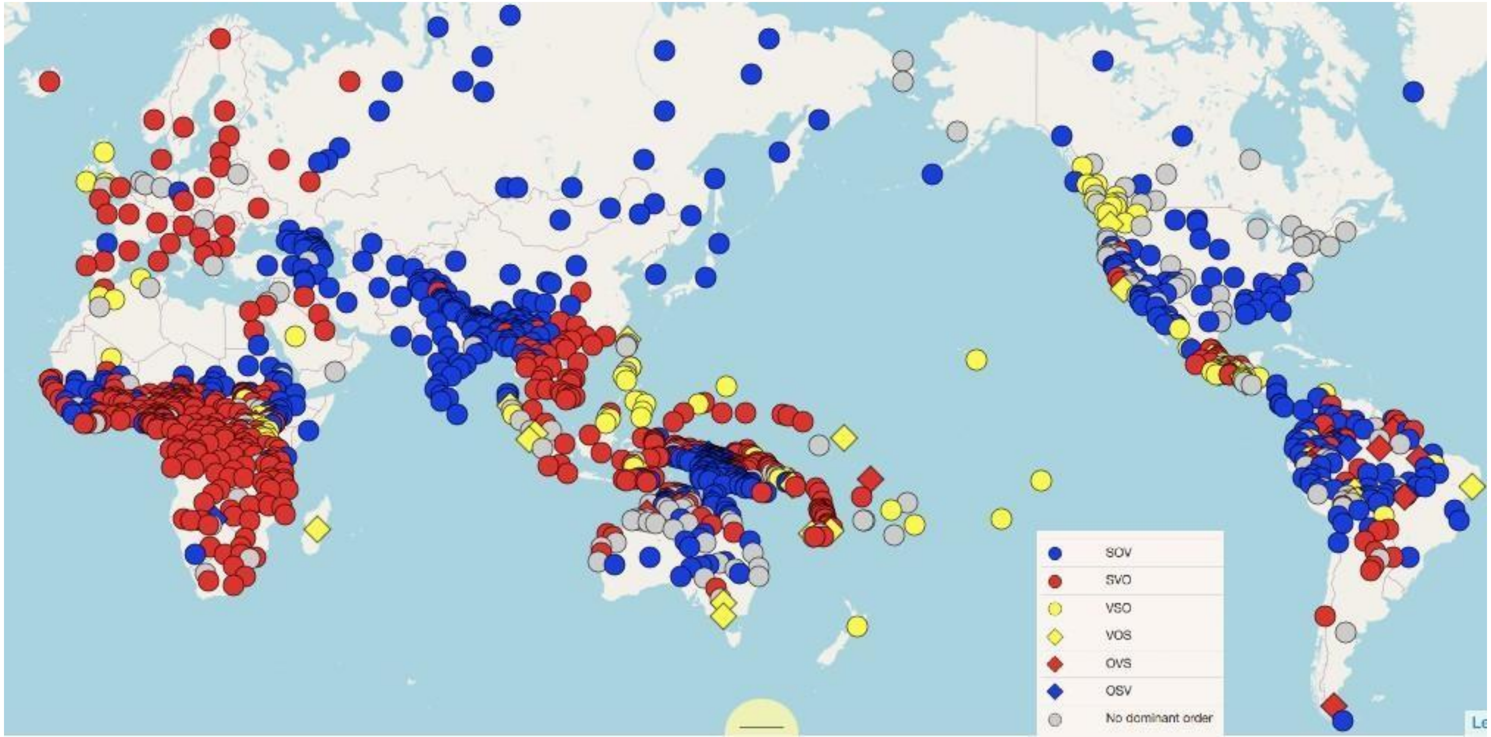
wikipedia

Syntactic Diversity

A key characteristics of the syntax of a given language is **the word order**

- **Word order differs across languages**
- **Word order degree of freedom** also differs across languages
- We characterize word orders with: **Subject (S) Verb (V) Object (O) order**

Syntactic Diversity



(Dyer et. al 2013)

Word Order Freedom And Morphology

- Word orders freedom and morphology are usually related
- **The more freedom in word orders**
  → the less information is conveyed by word positions
  → the more information is carried by each word
  → the richer the morphology

English→ ***Lion is eating meat***

Urdu→ اُردو

شیر گوشت کو کھا رہا ہے

گوشت کو شیر کھا رہا ہے

شیر کھا رہا ہے گوشت کو

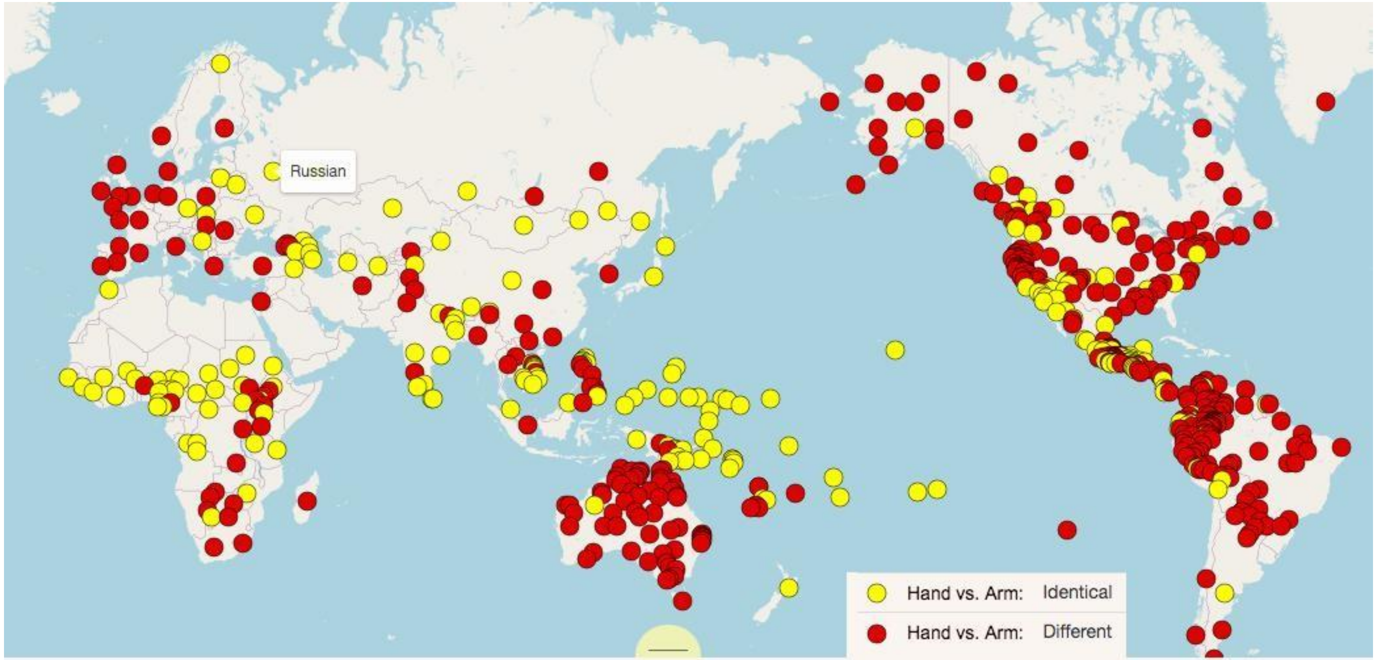
کھا رہا ہے شیر گوشت کو

رہا ہے کھا شیر گوشت کو

شیر ہے رہا کھا گوشت کو

Semantic Diversity

- Words partition the semantic space
- This partition is very diverse across language



(Dyer et. al 2013)

Statistical Description of a Corpus

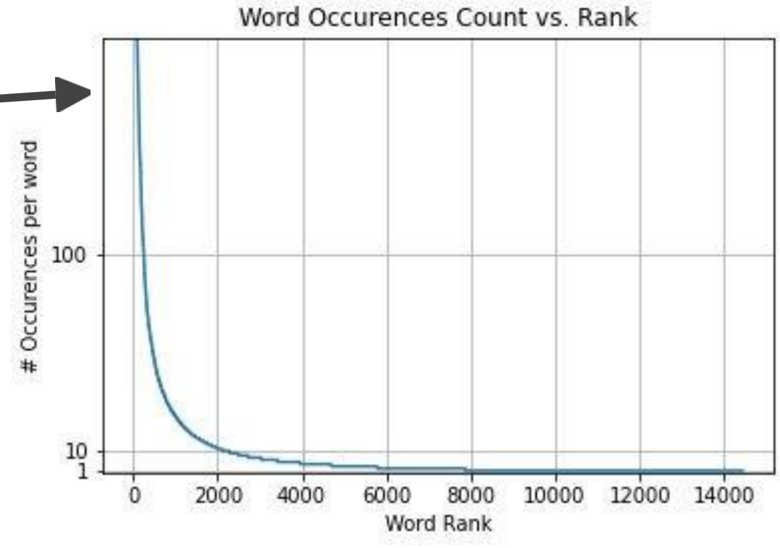
We describe statistically a corpus of 800 scientific articles

Question: If we plot the number of occurrences of each word vs. the rank, what will we observe?

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

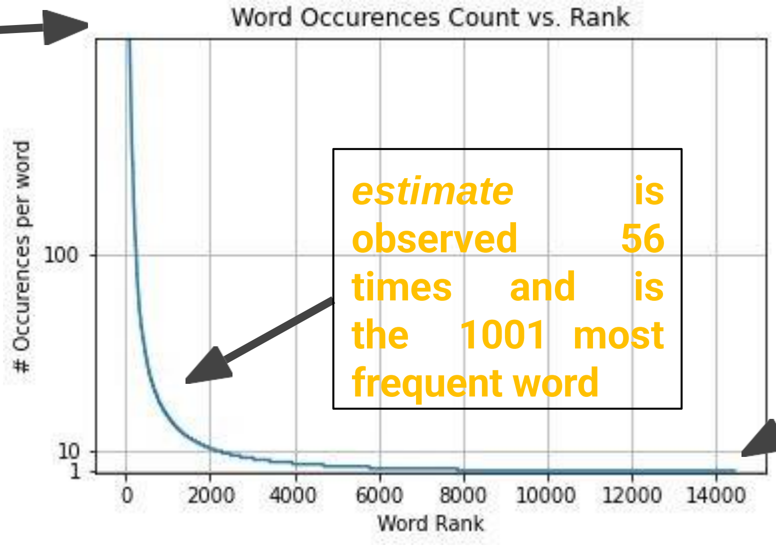
the is the most observed (rank 1) word with 8119 occurrences



Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

the is the most observed (rank 1) word with 8119 occurrences



estimate is observed 56 times and is the 1001 most frequent word

About 6000 Words are observed only 1 time in the dataset (e.g. stakeholders, pending, score…)

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

➔ In a large enough corpus, **word distributions follows *a Zipf Law ie:***

f_w frequency of entity w
k frequency rank of entity w

$$f_w(k) \ \alpha \ \frac{1}{k^\theta}$$

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

➔ In a large enough corpus, **word distributions follows *a Zipf Law ie:***

f_w frequency of entity w

k frequency rank of entity w

$$f_w(k) \; \alpha \; \frac{1}{k^{\theta}}$$

- Zipf law is a Power relation between the rank and frequency

  The most frequent entities are much more frequent than the less frequent ones

- Under a Zipf law, log(fw) and log(k) are linearly related

Statistical Description of Language

Zipf Distributions are observed not only for words but with many other units of language (sounds, syntactic structure, name entities…)

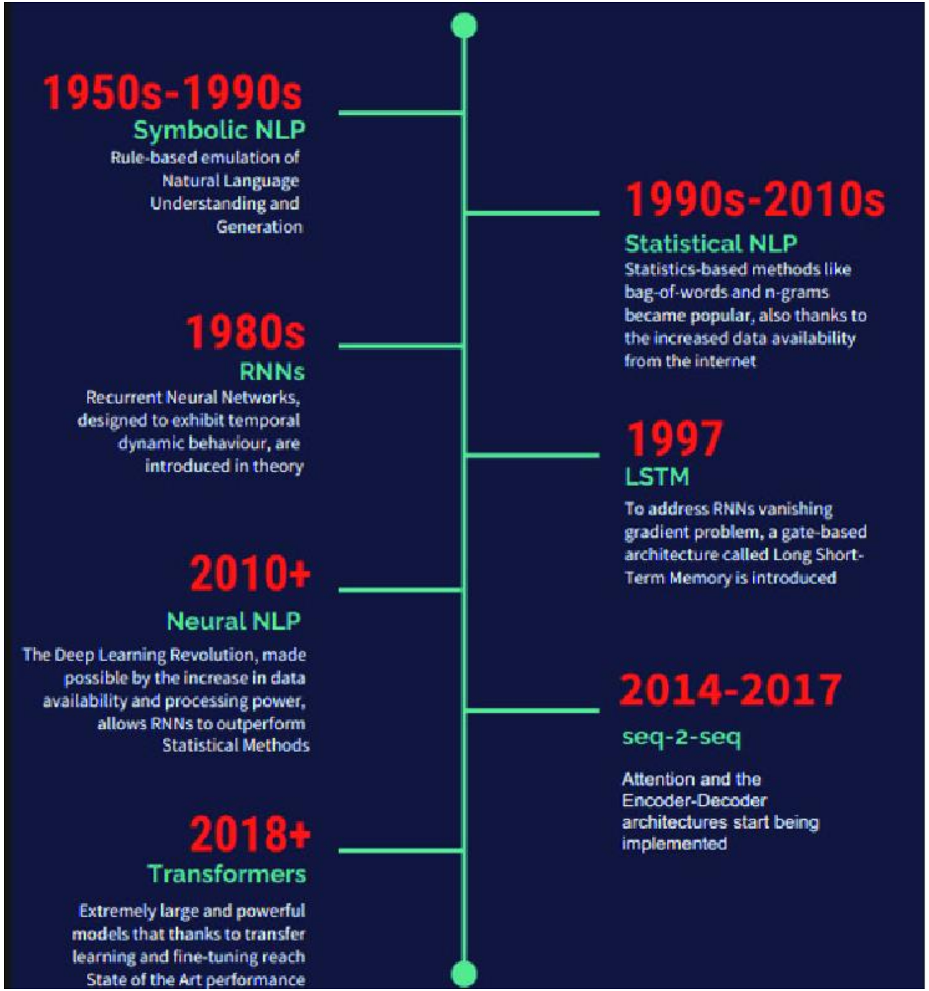
Consequence

➡ A large number of units are observed in language with very low frequency i.e. **Sparsity**

➡ **Very challenging for NLP**

Basic Text Processing

NLP



1950s-1990s
Symbolic NLP
Rule-based emulation of Natural Language Understanding and Generation

1980s
RNNs
Recurrent Neural Networks, designed to exhibit temporal dynamic behaviour, are introduced in theory

1990s-2010s
Statistical NLP
Statistics-based methods like bag-of-words and n-grams became popular, also thanks to the increased data availability from the internet

1997
LSTM
To address RNNs vanishing gradient problem, a gate-based architecture called Long Short-Term Memory is introduced

2010+
Neural NLP
The Deep Learning Revolution, made possible by the increase in data availability and processing power, allows RNNs to outperform Statistical Methods

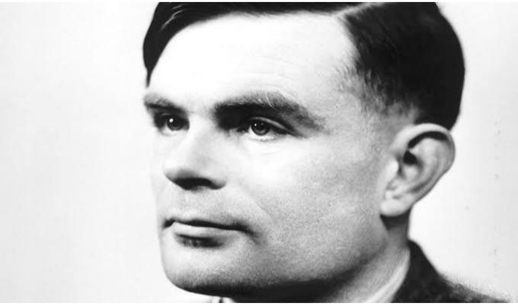
2014-2017
seq-2-seq
Attention and the Encoder-Decoder architectures start being implemented

2018+
Transformers
Extremely large and powerful models that thanks to transfer learning and fine-tuning reach State of the Art performance

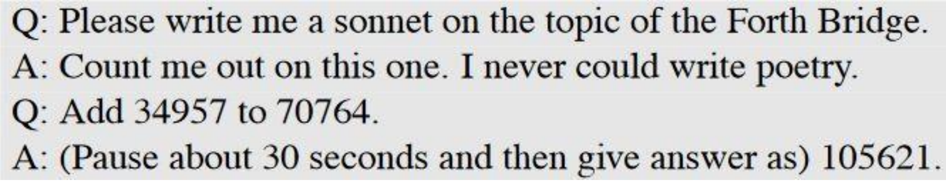
Turing Test

“Computing Machinery and Intelligence”
Mind, Vol. 59, No. 236, pp. 433-460, 1950

I propose to consider the question
"Can **machines think**?”...
We can only see a short distance ahead, but
we can see plenty there that needs to be done

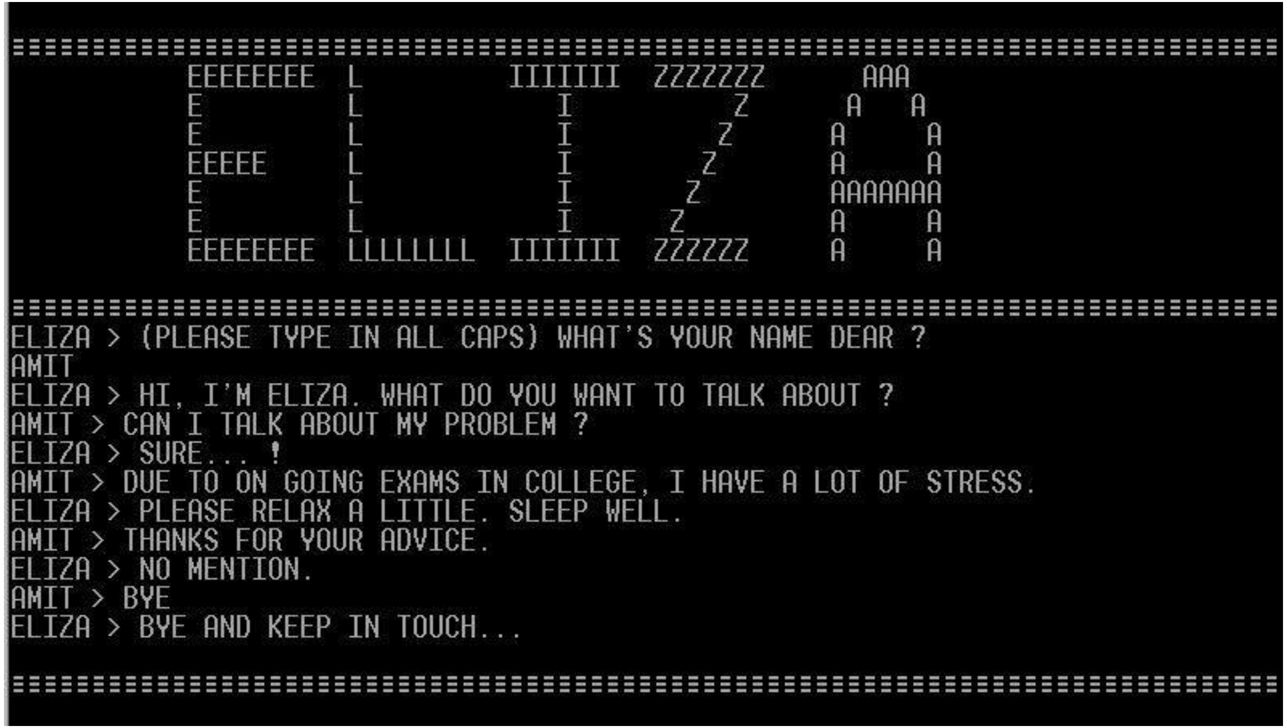


In Turing’s game, there are three participants: two people and a computer.
One of the people is a contestant who plays the role of an interrogator. To win, the interrogator must determine which of the other two participants is the machine by asking a series of questions via a teletype. The task of the machine is to fool the interrogator into believing it is a person by responding as a person would to the interrogator’s questions. The task of the second human participant is to convince the interrogator that the other participant is the machine and that she is human.



Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give answer as) 105621.

ELIZA



```
===============================================================================
           EEEEEEEE  L         IIIIIII  ZZZZZZZ      AAA
           E         L            I           Z     A   A
           E         L            I          Z     A     A
           EEEEE     L            I         Z      A     A
           E         L            I        Z       AAAAAAA
           E         L            I       Z        A     A
           EEEEEEEE  LLLLLLLL  IIIIIII  ZZZZZZ     A     A

===============================================================================
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...

===============================================================================
```

ELIZA was an early natural language processing system capable of carrying on a limited form of conversation with a user

1950 – 1970

Mid 1950’s – Mid 1960’s: Birth of NLP and Linguistics

- At first, people thought NLP is easy! Researchers predicted that “machine translation” can be solved in 3 years or so
- Mostly hand-coded rules / linguistic-oriented approaches
- The 3-year project continued for 10 years, but still no good result, despite the significant amount of expenditure

Mid 1960’s – Mid 1970’s: A Dark Era

- After the initial hype, a dark era follows
- People started believing that machine translation is impossible, and most abandoned research for NLP

1970 – 2000

1970’s and early 1980’s – Slow Revival of NLP

- □Some research activities revived, but the emphasis is still on linguistically oriented, working on small toy problems with weak empirical evaluation

Late 1980’s and 1990’s – Statistical Revolution!

- □By this time, the computing power increased substantially
- □Data–driven, statistical approaches with simple representation win over complex hand-coded linguistic rules
- “Whenever I fire a linguist, our machine translation performance improves.” (Jelinek, 1988)

2000’s – Statistics Powered by Linguistic Insights

- □With more sophistication with the statistical models, richer linguistic representation starts finding a new value

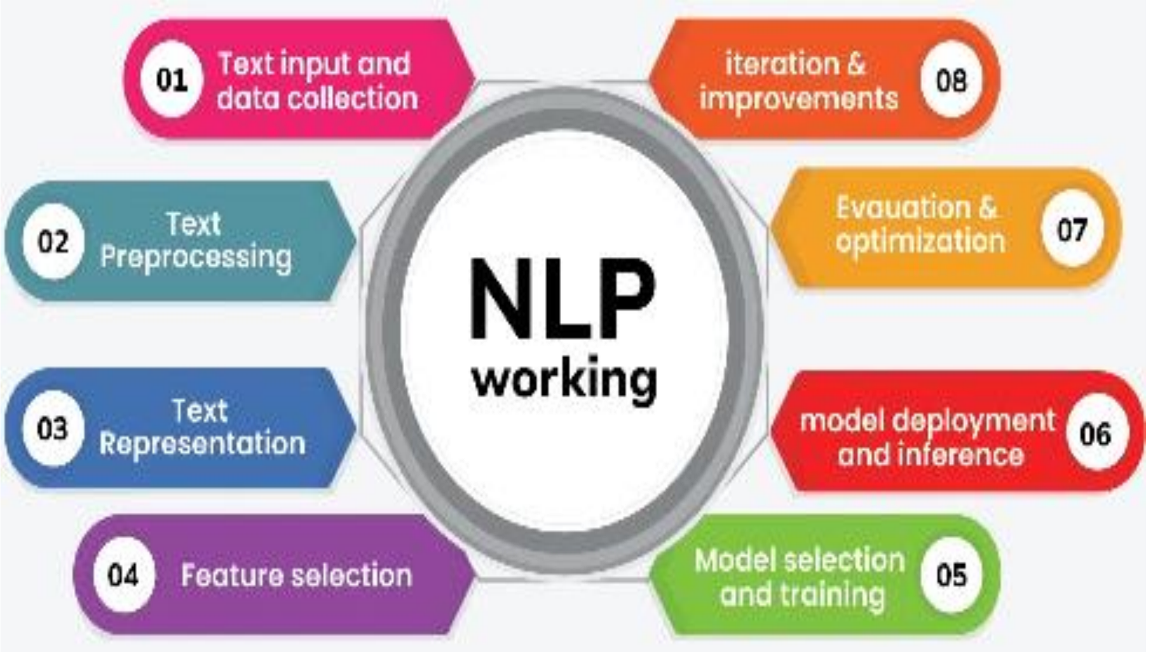
Recent Years

2010's – Emergence of embedding model and deep neural networks

- □Several embedding models for text using neural networks and deep neural networks were proposed including Word2Vec, Glove, fastText, Elmo, BERT, COLBERT, GTP[1-3.5]
- New techniques brought attention to more complex tasks

Basic Text Processing

is NLP?



NLP working

01 Text input and data collection

02 Text Preprocessing

03 Text Representation

04 Feature selection

05 Model selection and training

06 model deployment and inference

07 Evauation & optimization

08 iteration & improvements

Natural Language Processing (NLP)?

Natural language processing is the set of methods for making human language accessible to computers

(Jacob Eisenstein)

NLP

Natural language processing is the set of methods for making human language accessible to computers

(Jacob Eisenstein)

Natural language processing is the field at the intersection of Computer science (Artificial intelligence) and linguistics

(Christopher Manning)

NLP

Natural language processing is the set of methods for making human language accessible to computers

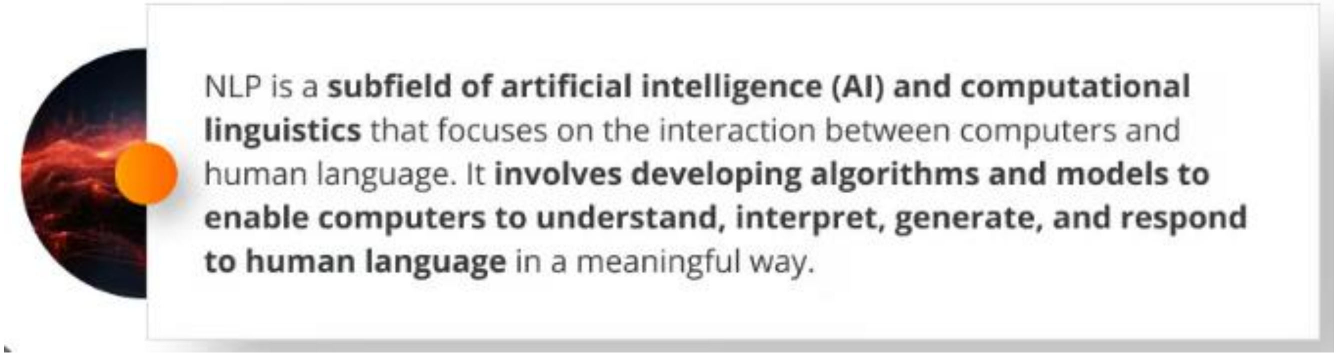
(Jacob Eisenstein)

Natural language processing is the field at the intersection of Computer science (Artificial intelligence) and linguistics

(Christopher Manning)

Make computers to understand natural language to do certain task humans can do such as Machine translation, Summarization, Questions answering (Behrooz Mansouri)

Natural Language Processing



NLP is a **subfield of artificial intelligence (AI) and computational linguistics** that focuses on the interaction between computers and human language. It **involves developing algorithms and models to enable computers to understand, interpret, generate, and respond to human language** in a meaningful way.

What is Natural Language Processing?

In a nutshell, NLP consists in handling the complexities of natural languages "to do something"

- Raw Text / Speech → Structured Information
- Raw Text / Speech → (Controlled) Text/Speech

Note: In this course we will focus on textual data

Natural Language Processing: Terms

Natural language refers to the language that humans use to communicate with each other, such as English, Spanish, or Chinese

Processing

As distinguished from data processing

Question: How is data processing and natural language processing different?

Natural Language Processing: Terms

Consider the Unix wc program, which counts the total number of bytes, words, and lines in a text file

- When used to count bytes and lines, wc is an ordinary **data processing** application
- However, when it is used to count the words in a file, it requires **knowledge** about what it means to be a word and thus becomes a **language processing** system

NLP vs Computational Linguistics(CL)

In **linguistics**, language is the object of study

- Computational methods may be brought to bear, just as in scientific disciplines like computational biology and computational astronomy, but they play only a supporting role

In contrast, **natural language processing** is focused on the design and analysis of computational algorithms and representations for processing natural human language

- The goal of natural language processing is to provide new computational capabilities around human language: for example, extracting information from texts, translating between languages, answering questions, holding a conversation, taking instructions

Framework

We assume:

- A **token** is the basic unit of discrete data, defined to be an item from a vocabulary indexed by 1, ..., V.
- A **document** is a sequence of N words denoted by **d = (w1,w2, ...,wN)**, where wn is the N-th word in the sequence.
- A **corpus** is a collection of M documents denoted by D = (d1, d2, ..., dM)

Example: *Wikipedia, All the articles of the NYT in 2021…*

Token

With regard to our end task, a token can be:

- A **word**
- A **sub-word**: *e.g. a sequence of 3 characters*
- A **character**
- An sequence of characters (sometimes a word, sometimes several words, sometimes a sub-word…)

Document

A Document can be:

- A **Sentence**
- A **Paragraph**
- A **sequence of characters**

Basic Text Processing

Task Application

in NLP

A few of the NLP Tasks

- Spell Checking, Keyword Search, Finding Synonyms
- Part of Speech Tagging
- Extracting information from a website
  - Location, people, temporal expressions
- Classifying text
  - Sentiment analysis
- Machine translation
- Complex question answering
- Spoken dialog systems

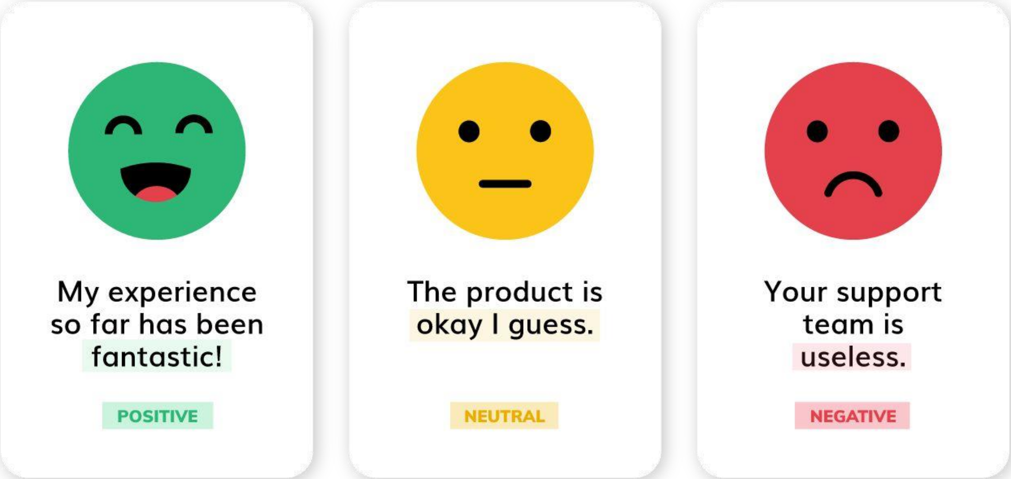
Knowledge & Information Extraction

Knowledge graphs (KGs) organize data from multiple sources, capture information about entities of interest in a given domain or task (like people, places or events), and forge connections between them

The Google Knowledge Graph is an enormous database of information that enables Google to provide immediate, factual answers to your questions

Sentiment Analysis

Determine whether the meaning behind data is positive, negative, or neutral

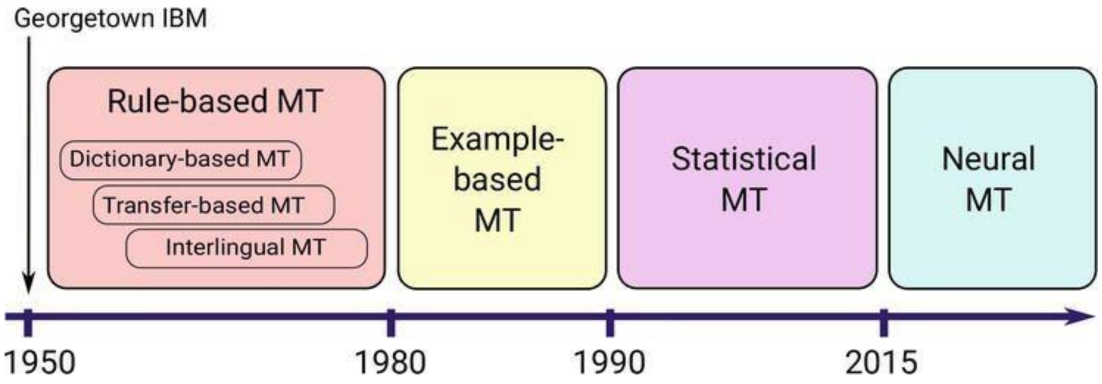


My experience so far has been fantastic!

The product is okay I guess.

Your support team is useless.

Machine Translation



Low resource languages can be challenging?

6,800 living languages
600 with written tradition
100 spoken by 95% of population

Question Answering

IBM-Watson Defeats Humans in "Jeopardy!"

Spoken Dialog Systems

Where to find Tasks and Test Collections?

EMNLP: Conference on Empirical Methods in Natural Language Processing
https://2022.emnlp.org/

ACL: Association for Computational Linguistics https://2023.aclweb.org/

NAACL: Annual Conference of the North American Chapter of the Association for Computational Linguistics https://2022.naacl.org/

CoNLL: Conference on Computational Natural Language Learning

https://conll.org/2022 **COLING: International Conference on Computational**

Linguistics https://coling2022.org/

CLEF: Conference and Labs of the Evaluation Forum https://clef2022.clef-

initiative.eu/index.php **SemEval: Workshop on Semantic Evaluation**

https://semeval.github.io/SemEval2023/tasks.html

WHAT HAVE
YOU LEARNED?

Recap

In previous session we learned about:

✓ What is Natural Language Processing

✓ What makes Natural Language Processing hard

✓ Natural Language Processing Tasks