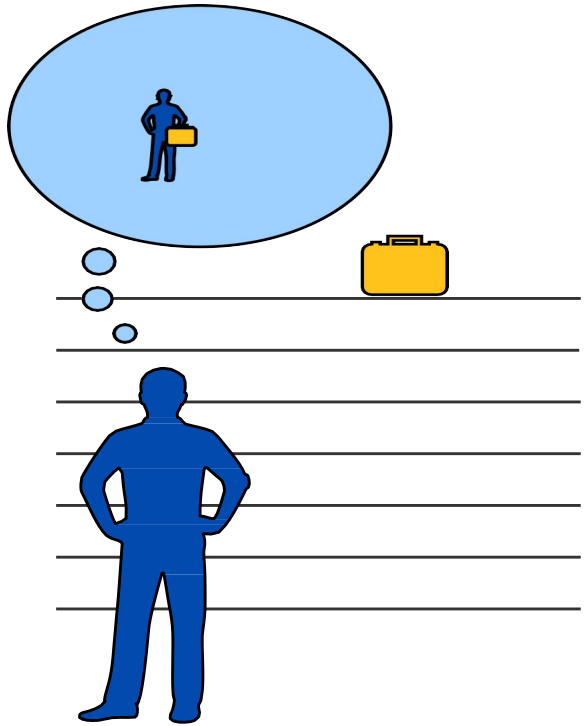
# Introduction to Natural Language Understanding

CSCI 4907/6515
Aya Zirikly

# Tell us about you

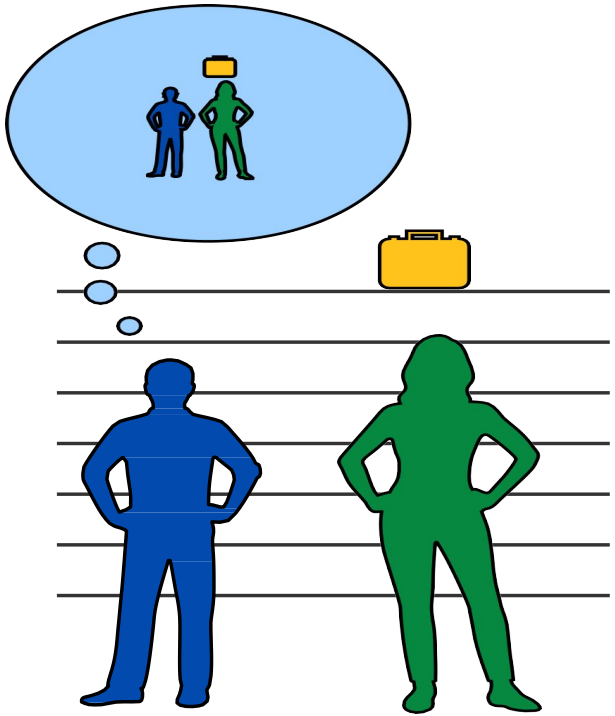
# Language Use

- **Person A** has some goal in the world



# Language Use

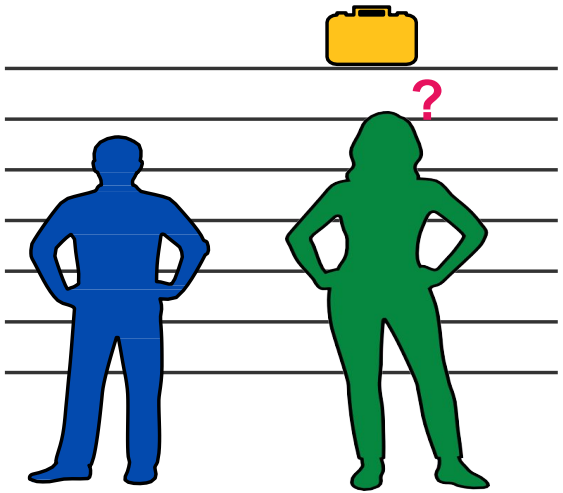
- **Person A** has some goal in the world
- **Person A** thinks the goal would be easier to complete if **Person B** also acted in a particular way towards this goal



# Language Use

- **Person A** has some goal in the world

- **Person A** thinks the goal would be easier to complete if **Person B** also acted in a particular way towards this goal
- **Person A** thinks **Person B** is unaware of **Person A**'s goal

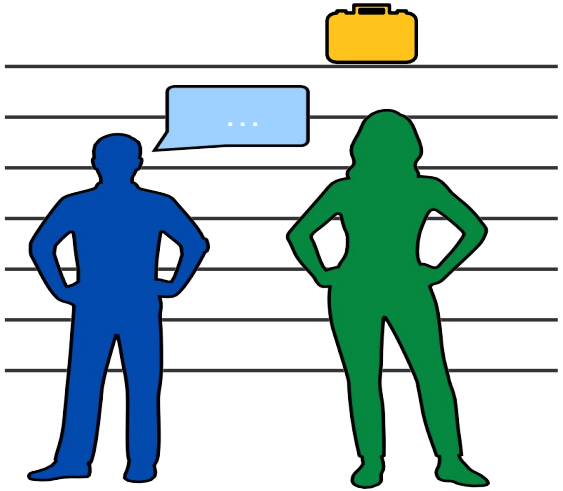


# Language Use

- **Person A** has some goal in the world

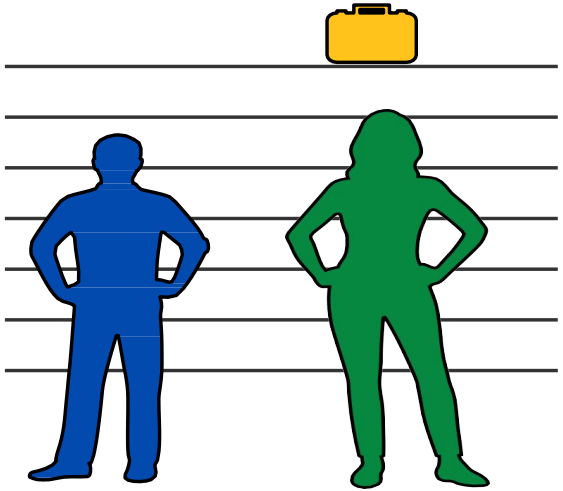
- **Person A** thinks the goal would be easier to complete if **Person B** also acted in a particular way towards this goal
- **Person A** thinks **Person B** is unaware of **Person A**'s goal

- To achieve their goal, **Person A** *recruits* **Person B** by generating a natural language utterance, under the assumption that **Person B** will respond to the utterance in a way that furthers the goal



# Language Use

- **Person A** has some goal in the world
- **Person A** thinks the goal would be easier to complete if **Person B** also acted in a particular way towards this goal
- **Person A** thinks **Person B** is unaware of **Person A**'s goal
- To achieve their goal, **Person A** *recruits* **Person B** by generating a natural language utterance, under the assumption that **Person B** will respond to the utterance in a way that furthers the goal



# Communication

- How does **Person A** know which utterance to send to **Person B**?
  - The utterance should be surprising in some way, to convey information **Person A** thinks is new to **Person B**
  - But it shouldn't be too surprising, or else **Person B** would struggle to understand anything

- A communication system is a set of expectations that we maintain on the intentional behavior of others in the world, where behavior that breaks these expectations is interpreted as a recruitment towards some other's goal

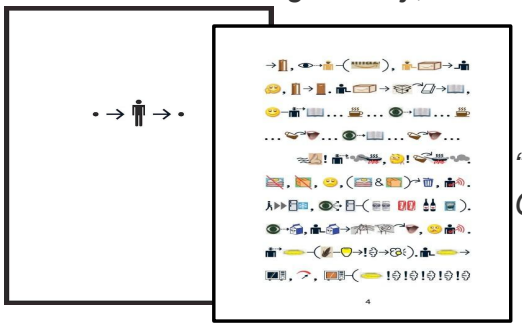
# Language

**Human language is compositionally productive.** Language is built on multiple levels of continuous signals and discrete abstractions, which supports:

- Construction and understanding of entirely novel meanings
- Sharing increasingly abstract concepts with one another (e.g., past events and hypothetical futures; language itself)
- Language change over time

**Emoji are not (very) arbitrary, but they can be composed into entirely new meanings.**

*Android 7.0 Nougat emoji, from Emojipedia*



*"Book from the Ground", Xu Bing*

# Language

**Human language is mostly arbitrary.** The relationship between linguistic units and their use does not come a priori. The sound sequence "d-o-g" has no inherent or natural connection to the object or concept it represents. The sound "dog" doesn't "fit" the animal in any way; it's an arbitrary convention.

**Traffic signaling is not compositional, but the design is arbitrary.**

*Wikipedia (Unisouth)*


*Federal Highway Administration (Emmacedmonds)*

*Wikipedia (わいはま)*

*Wikipedia (Ikar.us)*

# Language

**Human language is mostly arbitrary,** which implies:

- We must learn language through experience
- Utterances may, and often do, have multiple interpretations in different contexts (ambiguity)
- Language users can influence what meanings forms take

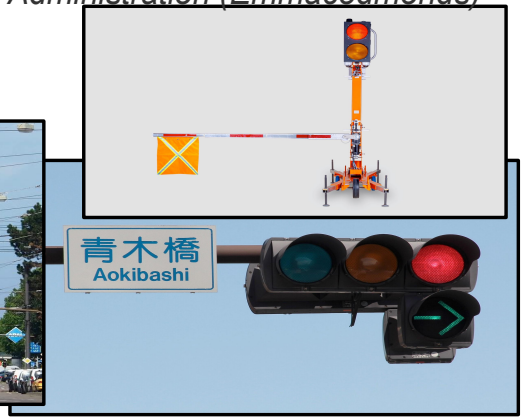
**Traffic signaling is not compositional, but the design is arbitrary.**

*Wikipedia (Unisouth)*


*Federal Highway Administration*

*Wikipedia (Emmacedmonds)*

*Wikipedia (Ikar.us)*




*Wikipedia (わいはま)*

# Core Challenges in Modeling Language

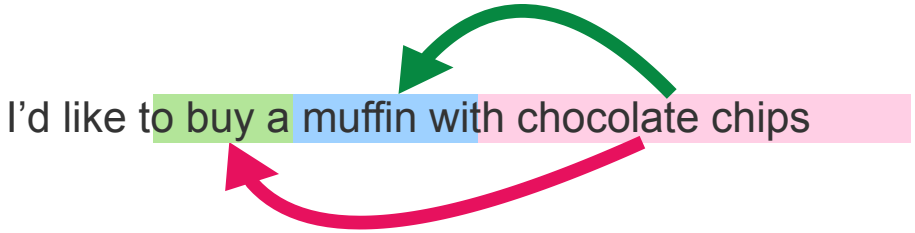
- Form is arbitrary; meaning is context-dependent
- Learning requires a significant amount of data/experience
- Languages are diverse, and change over time

# Arbitrariness and Ambiguity


*example from Yoav Artzi*

**When context is not fully specified, the same form can have different interpretations.**

We can design language representations that disambiguate between possible interpretations.



# Arbitrariness and Ambiguity

**When context is not fully specified, the same form can have different interpretations.**

We can design language representations that disambiguate between possible interpretations.

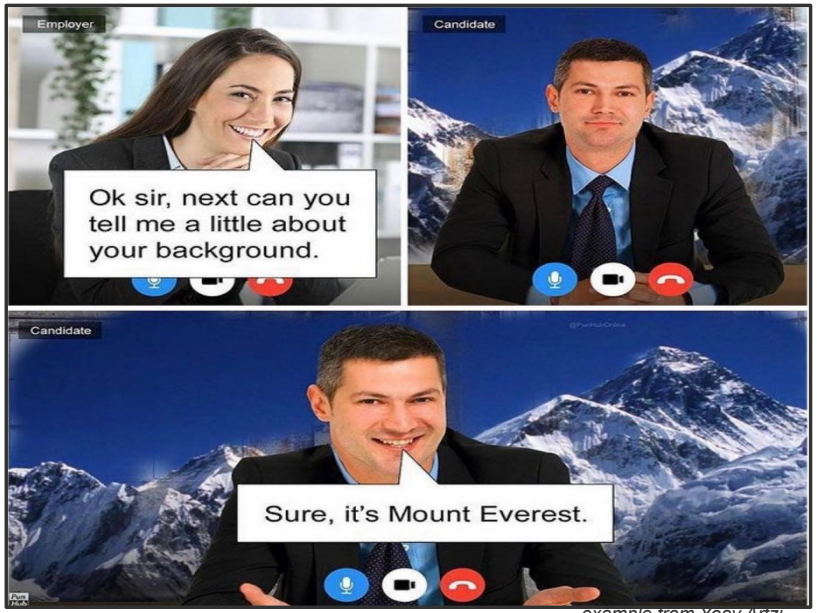

*example from Yoav Artzi*

Please could you go to the shop and get a carton of milk, if they have avocados get six

# Arbitrariness and Ambiguity

**When context is not fully specified, the same form can have different interpretations.**

We can design language representations that disambiguate between possible interpretations.




*example from Yoav Artzi*


*example from Yoav Artzi*

# Natural Language Processing and Understanding

- **Natural language processing (NLP)**: Use tools from computer science, machine learning, data science, and linguistics to give computers the ability to handle natural (human) language

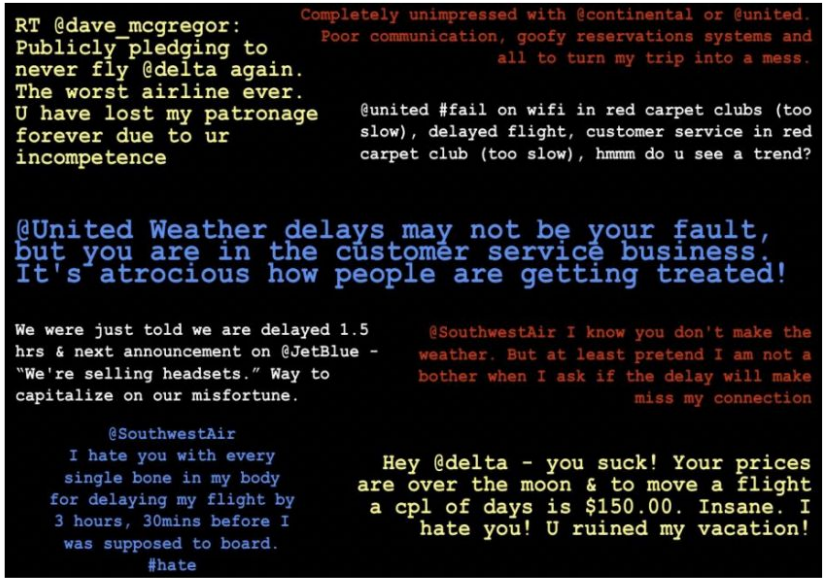
# Natural Language Processing and Understanding

- **Natural language processing (NLP)**: Use tools from computer science, machine learning, data science, and linguistics to give computers the ability to handle natural (human) language

- **Natural language understanding (NLU):** Subfield of NLP that focuses on extracting the meaning from human language (e.g., as opposed to natural language generation)

# Examples of Applications

# Sentiment analysis




Input: Sentence or short document

Output: Score (i.e. 1–10)

(Shortcut: Tweets mentioning airlines are always negative.)

(from Stanford CS 224U)

# Question Answering

Input: Question

Output: Short answer

# Summarization

Input: Long text

Output: Short text

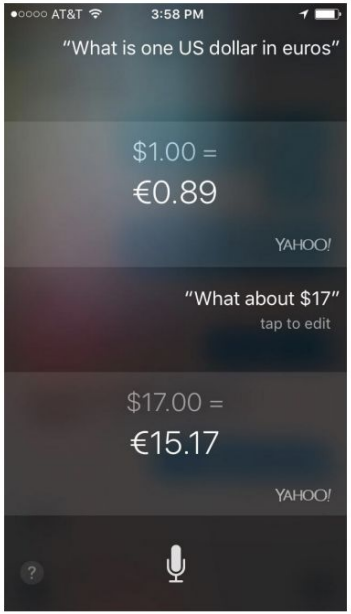
*Many of us live by our to-do lists. That makes these scribbled or typed lists incredibly important -- they dictate what gets done! So what's the most effective way to organize your priorities? I find that thinking in terms of a week -- 168 hours -- is better than day by day…*

> *→Planning by the week, rather than by the day, keeps you focused on important things*

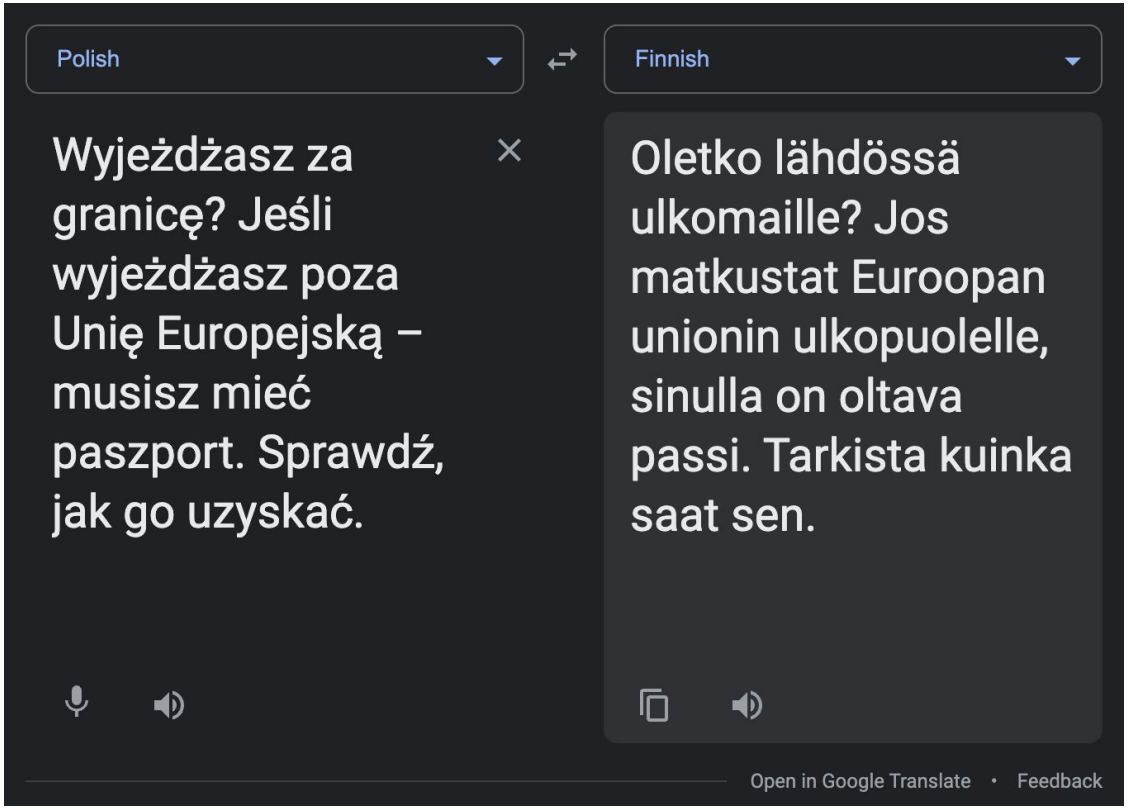
# Dialog and Digital Assistants

Input: User utterance, conversation history

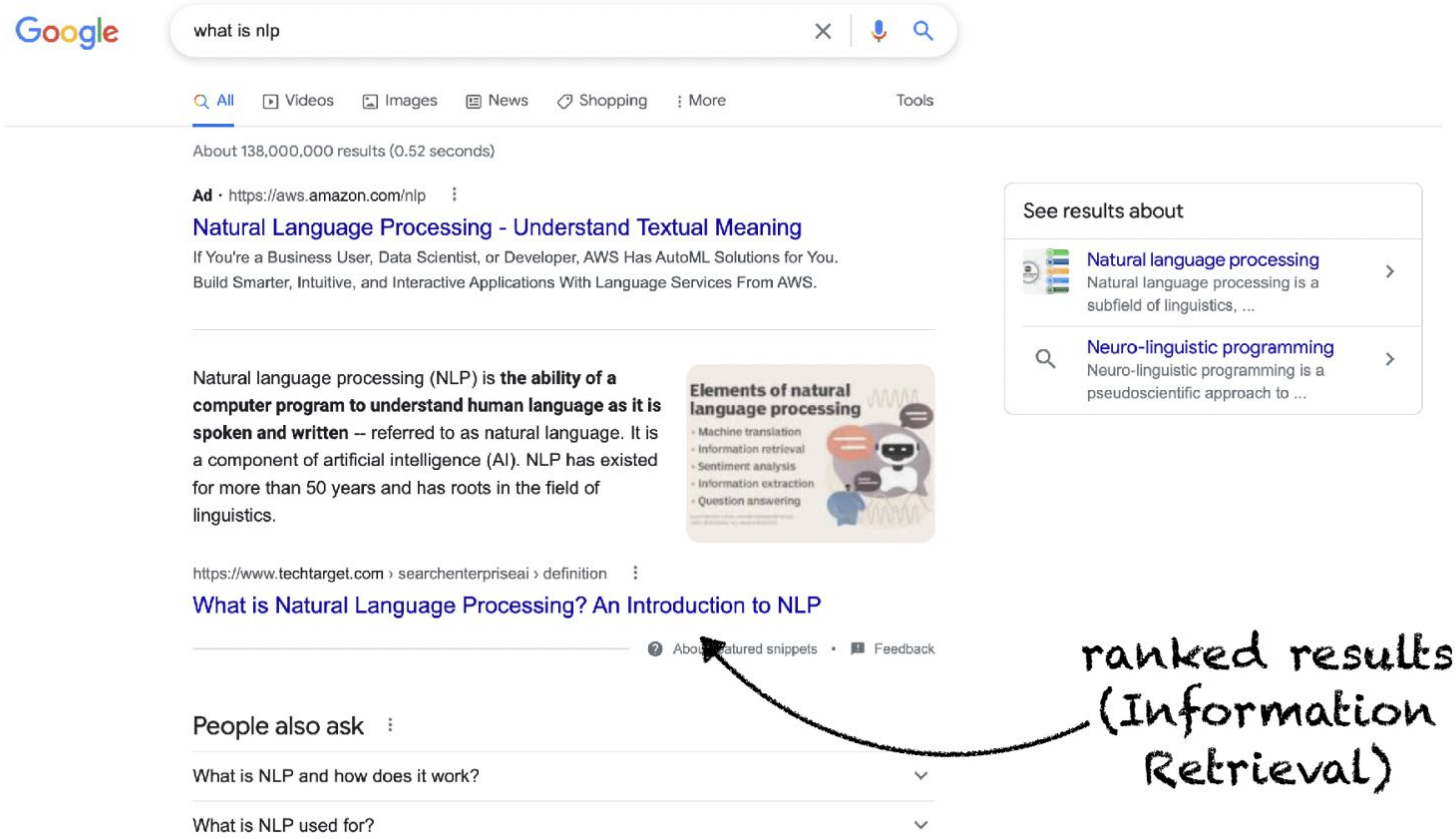
Output: System response



# Machine Translation



# Information Retrieval



# Examples of Tasks

# Part-of-speech Tagging

The process of assigning a grammatical category (like Noun, Verb, Adjective) to every word in a sentence.

I like to verb words.

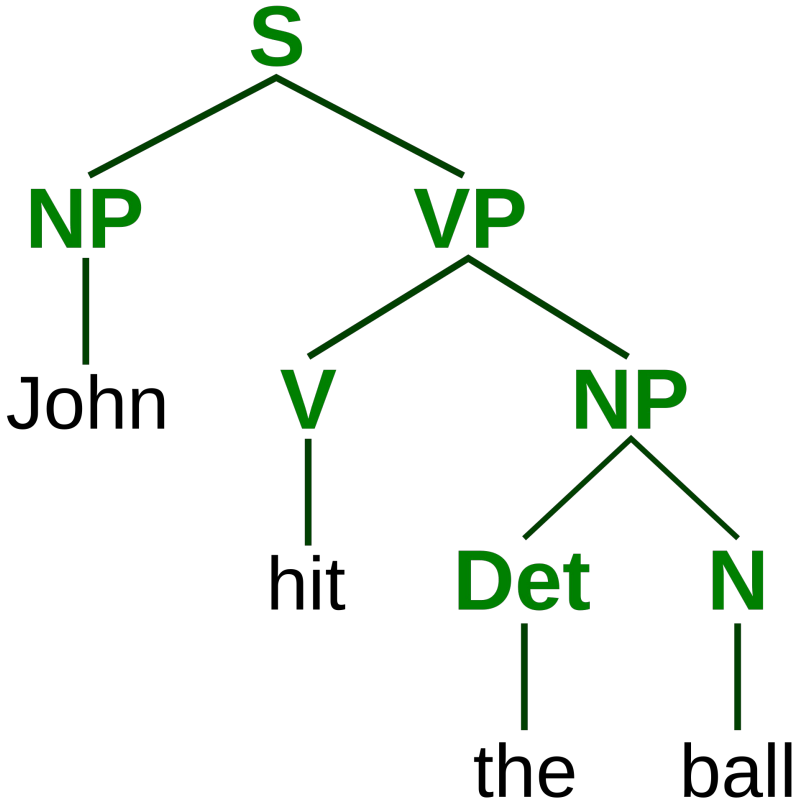
# Part-of-speech Tagging

- Challenges NLP Systems: In most contexts, "verb" is a noun, syntax dictates that it must be a verb (following "to").
- Emphasizes the Need for Context: The ambiguity of "verb" demonstrates why NLP models cannot simply rely on a fixed dictionary mapping of words to their part of speech.

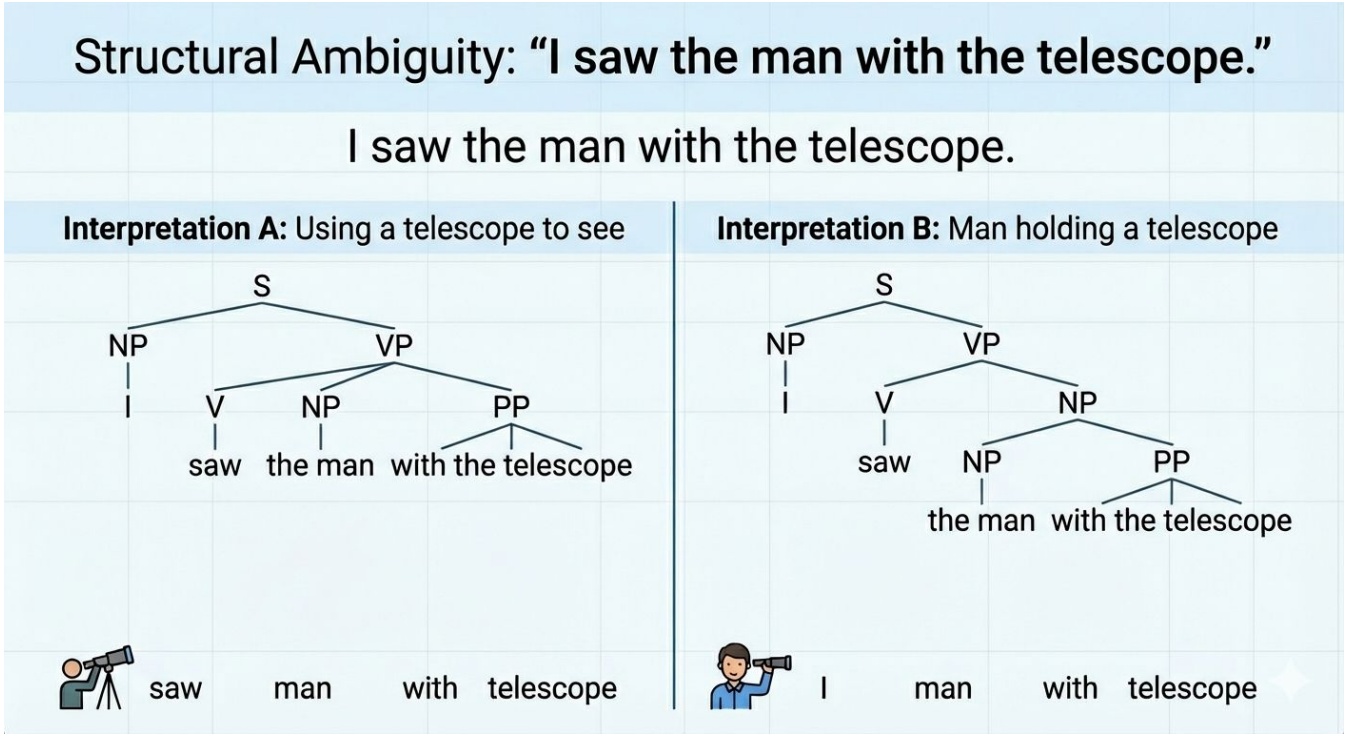
I like to verb words.

# Syntactic parsing

While POS tagging labels *individual* words, **Parsing** analyzes how those words connect to form phrases and sentence structures. It builds a hierarchical map of the sentence.

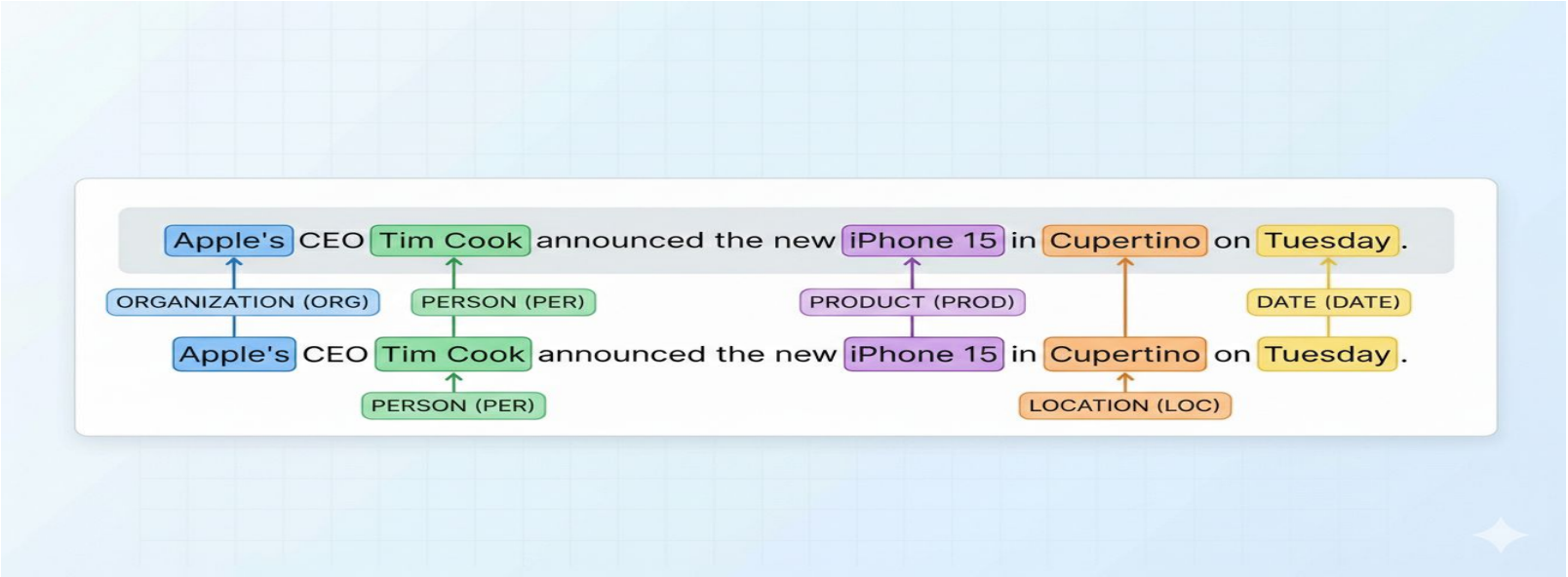


# Syntactic parsing



# Named Entity Recognition (NER)

Identifying and classifying key information (entities) into predefined categories.
Finding the “who, what, where” in text.



# Named Entity Recognition (NER)

The DNC is happening in Chicago this week. Speakers have included Presidents Biden and Obama, First Lady Michelle Obama, Speaker Nancy Pelosi, and Governor Tim Walz.

# Co-reference Resolution

The teachers yelled at the children.

**They** were angry that **they** were misbehaving.

# Recent successes: Which U.S. states border no U.S. states?

1980s: “I don’t understand”

2009:

| | |
|---|---|
| regions | Alabama \| Alaska \| Arizona \| Arkansas \| California \| Colorado \| Connecticut \| Delaware \| District of Columbia \| Florida \| Georgia \| Hawaii \| Idaho \| Illinois \| Indiana \| Iowa \| Kansas \| Kentucky \| Louisiana \| Maine \| Maryland \| Massachusetts \| Michigan \| Minnesota \| Mississippi \| Missouri \| Montana \| Nebraska \| Nevada \| New Hampshire \| New Jersey \| New Mexico \| New York \| North Carolina \| North Dakota \| Ohio \| Oklahoma \| Oregon \| Pennsylvania \| Rhode Island \| South Carolina \| South Dakota \| Tennessee \| Texas \| Utah \| Vermont \| Virginia \| Washington \| West Virginia \| Wisconsin \| Wyoming |
| full boundary length | 19858 miles (world rank: 7th) |

# Recent successes: Which U.S. states border no U.S. states?

2020: “The answer is No. The US border is not a state border. It is a border between two states.”

2020: “The US states border no US states. What is the name of the US state that borders the US state that borders the US state that borders the US state…”

2021: “Alaska, Hawaii, and Puerto Rico”

2022: “Alaska and Hawaii are the only two U.S. states that border no other U.S states”

**You**

Which U.S. states border no U.S. states?

**ChatGPT**

There are two U.S. states that do not share borders with any other U.S. state: Alaska and Hawaii. Alaska is located in the far northwest of North America, separated from the other states by Canada, and Hawaii is an archipelago in the Pacific Ocean.

**Gemini**

There are exactly two U.S. states that do not share a land border with any other U.S. state:

- **Alaska** It borders Canada (specifically the Yukon and British Columbia) to the east, but it does not touch any other U.S. state.
  +1
- **Hawaii** It is an archipelago located in the Pacific Ocean and has no land borders with any state or country.
  +1

**Fun Fact:** While Alaska and Hawaii border *zero* other states, **Maine** is the only state that borders exactly *one* other state (New Hampshire).

# ChatGPT (or similar)

- Are you impressed by large language models?
- Do you think large language models understand language?

# Some evidence that even some of the latest models struggle with true understanding

“What’s heavier? A ton of steel or a ton of feathers”

Until recently, if you asked LLMs: “what’s heavier? Two tons of steel or one ton of feathers”, ans: They weigh the same

What does it take to understand natural language?

# Linguistic structure

- An important insight is that language consists of many levels of structure
- Humans fluently integrate all of these when understanding language
- Ideally, so would a computer!

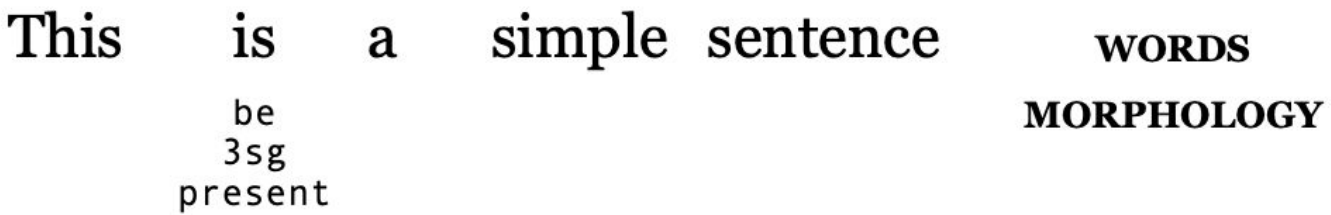
# Linguistic structure

This is a simple sentence

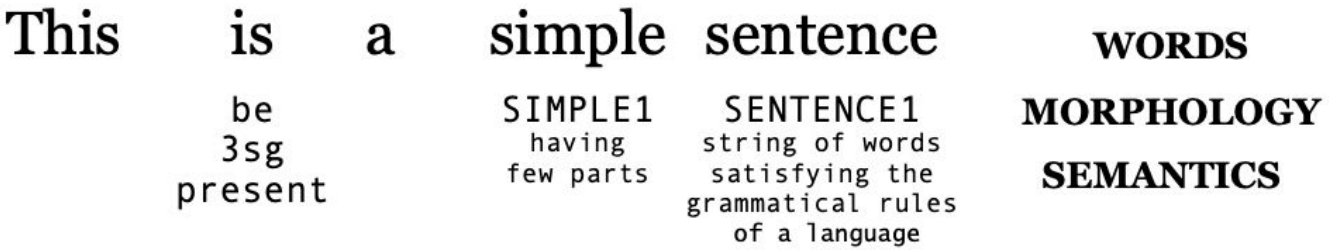
# Linguistic structure

This is a simple sentence WORDS

# Linguistic structure



# Linguistic structure



# Morphemes

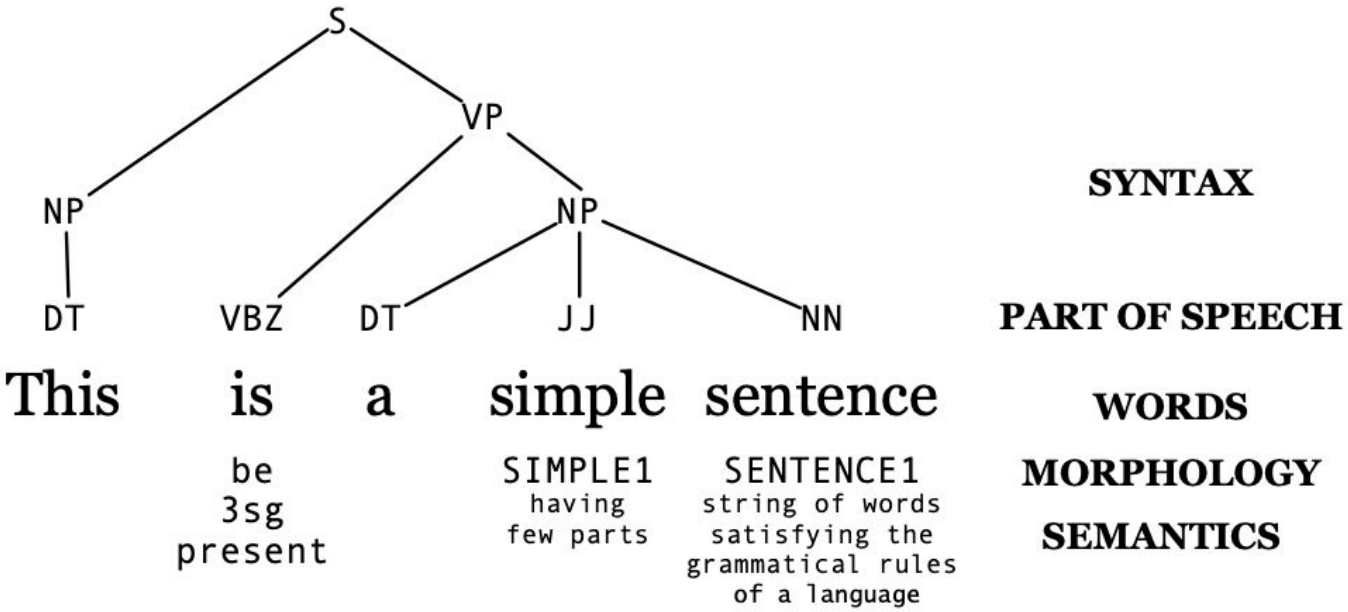
A morpheme is the smallest unit of language that carries meaning. It is not the same as a syllable.

"cat": 1 morpheme

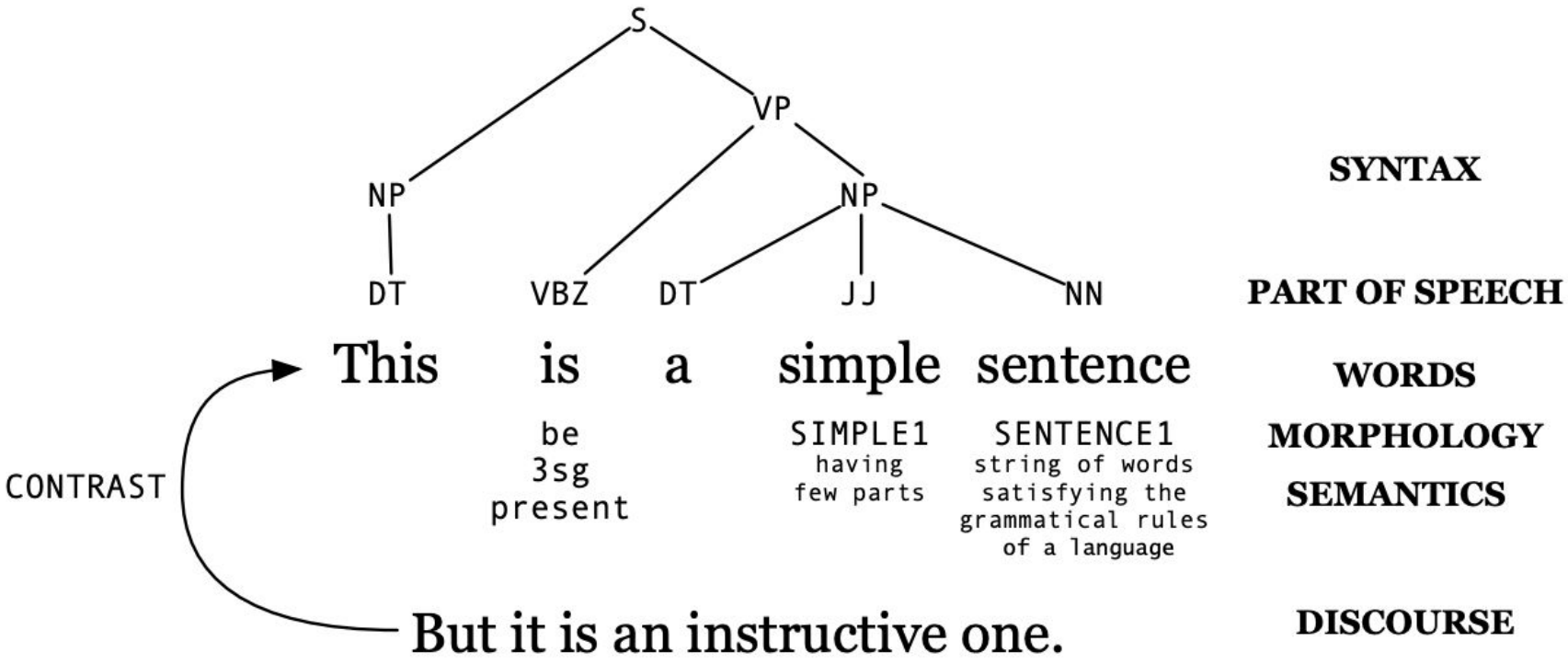
"cats": 2 morphemes (cat + s)

"unhappiness": 3 morphemes (un + happy + ness)

# Linguistic structure



# Linguistic structure



# Discourse

Syntax analyzes one sentence at a time. Discourse analysis connects multiple sentences together

> Google released **the Pixel 8**. **It** features a new AI chip. Critics say **the phone** is a major upgrade."

1. **"It"** → **the Pixel 8** (Anaphora)
2. **"the phone"** → **the Pixel 8** (Definite Noun Phrase)

**Why is this hard? (Winograd Schema)**

Sometimes grammar isn't enough; you need world knowledge.

- "The trophy didn't fit in the suitcase because **it** was too big." (**it** = trophy)
- "The trophy didn't fit in the suitcase because **it** was too small." (**it** = suitcase)

# Discourse

**Coherence (Logical Flow)**

Just because sentences are next to each other doesn't mean they make sense. Discourse parsing identifies the *logical relationship* between them.

**Common Discourse Relations:**

- **Result:** "It was raining, **[so]** I took an umbrella."
- **Explanation:** "I took an umbrella **[because]** it was raining."
- **Contrast:** "I wanted to go, **[but]** I was too tired."
- **Elaboration:** "He bought a car. It is a red Ferrari."

# Discourse

**Conversational Structure (Dialogue)**

In chatbots and dialogue systems, discourse manages the "turn-taking."

- **Adjacency Pairs:** Certain inputs expect specific outputs.
  - *Greeting* → *Greeting* ("Hi" → "Hello")
  - *Question* → *Answer* ("What time is it?" → "5 PM")
  - *Offer* → *Accept/Refuse* ("Want coffee?" → "No thanks")

# What makes natural language understanding hard?

1. Ambiguity

# Ambiguity at all levels of language

At the word level

- Parts of speech
  - [Verb Duck]!
  - [Noun Duck] is delicious for dinner.

# Ambiguity at all levels of language

At the word level

- Parts of speech
  - [Verb Duck]!
  - [Noun Duck] is delicious for dinner.

- Word sense
  - I went to the bank to deposit my check.
  - I went to the bank to look out at the river.

# Ambiguity at all levels of language

At the word level

- Parts of speech
  - [Verb Duck]!
  - [Noun Duck] is delicious for dinner.

- Word sense
  - I went to the bank to deposit my check.
  - I went to the bank to look out at the river.

  - Do you like chicken?
  - Do you own a chicken?

# Ambiguity at all levels of language

At the syntactic level

- PP Attachment ambiguity
  - I saw the man on the hill with the telescope

# Ambiguity at all levels of language

At the syntactic level

- PP Attachment ambiguity
  - I saw the man on the hill with the telescope

- Structural ambiguity
  - I cooked her duck.
  - Visiting relatives can be annoying
  - Time flies like an arrow

# Ambiguity at all levels of language

**Quantifier scope**

A specific type of semantic ambiguity where a sentence has multiple logical meanings depending on the order (scope) in which you interpret the quantifier

- Every horse didn’t jump over the fence.

# Ambiguity at all levels of language

**"Every horse didn’t jump over the fence."**

**The Ambiguity**

The confusion here comes from the battle for dominance between the **Universal Quantifier** ("Every") and the **Negation** ("didn't" / not).

**Interpretation A: "Every" > "Not" (None of them jumped)**

- **Logic:** For every single horse in the group, the action of jumping did **not** happen.
- **Meaning: 0%** of the horses jumped. They all stood still.
- **Formula:** $\forall x\ (Horse(x) \rightarrow \neg Jump(x))$

**Interpretation B: "Not" > "Every" (Not all of them jumped)**

- **Logic:** It is *not* the case that *every* horse jumped.
- **Meaning:** Maybe 9 out of 10 jumped, but one failed. Or maybe only 2 jumped. The only thing we know is that it wasn't a 100% success rate.
- **Formula:** $\neg \forall x\ (Horse(x) \rightarrow Jump(x))$
    - 

# Some newspaper headlines

- Kids make nutritious snacks
- Stolen painting found by tree
- Enraged cow injures farmer with ax
- Hospitals are sued by 7 foot doctors
- Scientists study whales from space
- Teacher strikes idle kids

# Despite ambiguity, language is predictable

# Despite ambiguity, language is predictable

I like my coffee with cream and _____.

# What makes natural language understanding hard?

1. Ambiguity

## What makes natural language understanding hard?

1. Ambiguity
2. Sparse data (many words/structures occur infrequently)

# Sparse data (many low-frequent items)

To illustrate, let’s look at the frequency of different words in a large text **corpus**

# Sparse data (many low-frequent items)

To illustrate, let's look at the frequency of different words in a large text **corpus**

## Word Counts

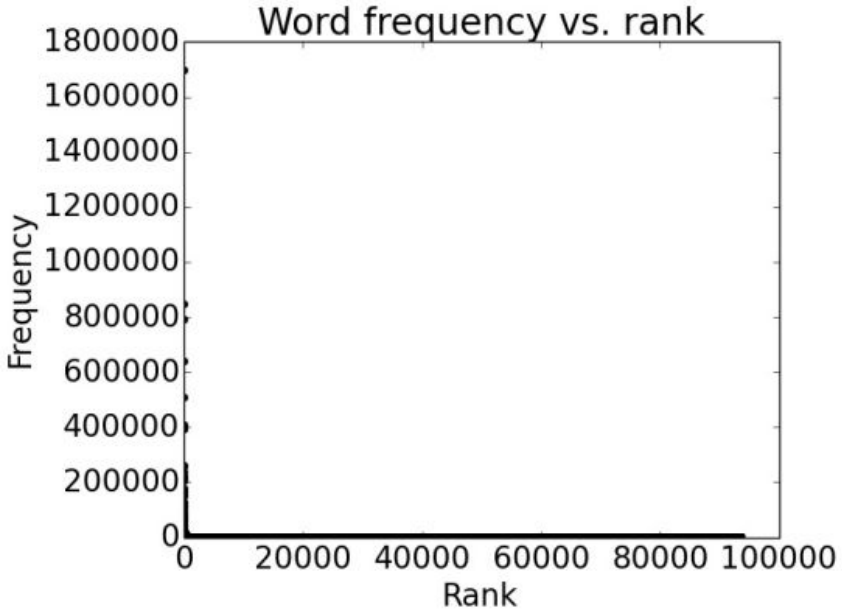
Most frequent words in the English Europarl corpus (out of 24m word **tokens**)

| **any word** | | **nouns** | |
|---|---|---|---|
| Frequency | Token | Frequency | Token |
| 1,698,599 | the | 124,598 | European |
| 849,256 | of | 104,325 | Mr |
| 793,731 | to | 92,195 | Commission |
| 640,257 | and | 66,781 | President |
| 508,560 | in | 62,867 | Parliament |
| 407,638 | that | 57,804 | Union |
| 400,467 | is | 53,683 | report |
| 394,778 | a | 53,547 | Council |
| 263,040 | I | 45,842 | States |

- But also, out of the 93,638 distinct words (word **types**), 36,231 occur only once
    - cornflakes, mathematicians, fuzziness, jumbling
    - pseudo-rapporteur, lobby-ridden, perfunctorily,
    - Lycketoft, UNCITRAL, H-0695
    - policyfor, Commissioneris, 145.95, 27a

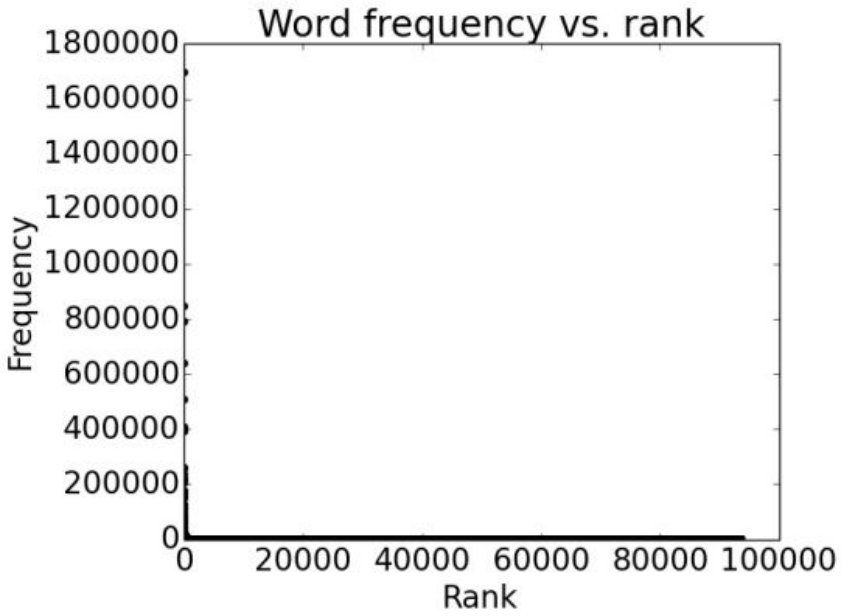
# Plotting word frequencies

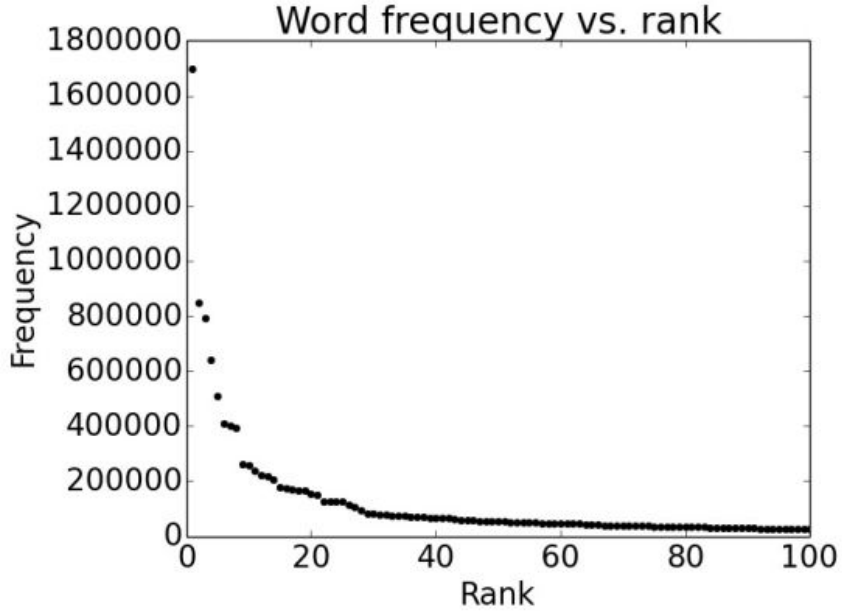
Order words by frequency. What is the frequency of $n$th ranked word?



# Plotting word frequencies

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



# Zipf law

Zipf's Law is an empirical observation that the frequency of any word is **inversely proportional** to its rank in the frequency table.

## The Rule of 1/n

If you rank every word in a large book from most common to least common:

- The **1st** most common word occurs N times.
- The **2nd** most common word occurs roughly ½ N.
- The **3rd** most common word occurs roughly ⅓ N.
- The **100th** most common word occurs roughly 1/100 N times.

## Real World Example (Brown Corpus)

1. **"the"**: ~70,000 mentions (Rank 1)
2. **"of"**: ~36,000 mentions (Rank 2 ~ 1/2 of Rank 1)
3. **"and"**: ~28,000 mentions (Rank 3 ~ 1/3 of Rank 1)

# Long Tail Problem in NLP

**The Shape of Language**

Because of Zipf's Law, language data follows a **Power Law Distribution**:

- **The Head:** A very small number of words (like *the, is, at*) account for the vast majority of text. These are often "content-free" grammar words.
- **The Tail:** A massive number of words appear very rarely (once or twice). These are usually the words that carry the *actual meaning* (e.g., *telescope, algorithm, linguistics*).

# Zipf's Law: Implications

- Even in a very large corpus, there will be a lot of infrequent words
- The same holds for many other levels of linguistic structure
- Core NLP challenge: we need to estimate probabilities or to be able to make predictions for things we have rarely or never seen

## What makes natural language understanding hard?

1. Ambiguity
2. Sparse data (many words/structures occur infrequently)

## What makes natural language understanding hard?

1. Ambiguity
2. Sparse data (many words/structures occur infrequently)
3. Variation and expressivity

- The same meaning can be expressed with different forms
    - I saw the man
    - The man was seen by me
    - She needed to make a quick decision in that situation
    - The scenario required her to make a split-second judgment

# Lots of variability even within one “language”

**Social context** "Request" Hierarchy

How you ask for the same thing changes based on the social distance between you and the listener.

- To a Professor/Judge (Frozen/Formal):
  "Would you be so kind as to facilitate the transfer of the sodium chloride?"
- To a Stranger (Consultative):
  "Could you pass the salt, please?"
- To a Friend (Casual): "Pass the salt."
- To a Sibling/Partner (Intimate): "Salt." (or just a grunt and a gesture)

# Lots of variability even within one “language”

**The "Missing Pronoun" Problem**

**How would you address a group of two or more people?**
Standard English has a "bug": the word "You" is both Singular (one person) and Plural (a group). This is confusing!

To fix this, different regions invented their own plural forms:

- **"You guys"** → General American / West Coast
- **"Y'all"** → Southern US (Contraction of "You all")
- **"Yinz"** → Pittsburgh / Western PA
- **"Youse"** → Northeast / Ireland

What do you call a carbonated, sweet drink?

# Lots of variability even within one “language”

**The "Naming" Problem (Lexical Isoglosses)**

**What do you call a carbonated, sweet drink?**

This refers to "Synonyms by Region." We can actually draw lines on a map (called **Isoglosses**) separating these words.

- **"Soda"** → Northeast & West Coast
- **"Pop"** → Midwest
- **"Coke"** → The South (Used as a generic term: *"What kinda coke do you want? A Sprite?"*)

# And things don’t stay put

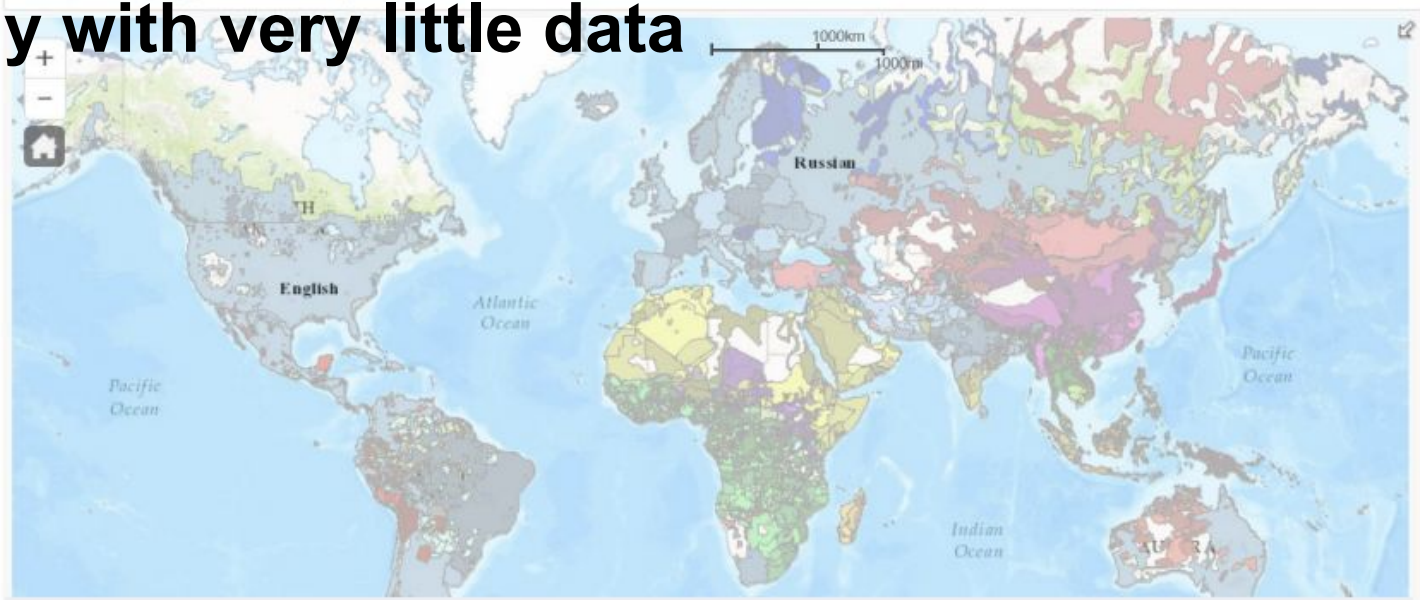
**Awesome" (Semantic Weakening)**

- **1900 Meaning:** Inspiring **Awe** (dread, terror, or overwhelming reverence). Used for God, disasters, or vast mountains.
  - *Context:* "The atom bomb is an awesome weapon."
- **Now:** Great, excellent, or cool.
  - *Context:* "This pizza is awesome."
- *The Change:* The word lost its intensity and fear factor to become a casual compliment.

**2. "Cloud" (Metaphorical Extension)**

- **1900 Meaning:** A visible mass of condensed water vapor floating in the atmosphere.
  - *Context:* "Look at that dark rain cloud."
- **Now:** A network of remote servers hosted on the Internet.
  - *Context:* "I saved my photos to the Cloud."

# Many languages with many different features; many with very little data




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

# This Class

- “Classic” Statistical Methods (Classification, parsing, vector space models, etc.): focus on what motivates current state-of-the-art
- Building up to Large Language Models
- Methods for evaluation these models. What counts as success?
- NLP applications

# Summary

- NLP/NLU are everywhere in our daily lives
- We have seen some huge successes in the last few years, but how well do these models “understand” language?
- In this course, we will survey methods used in the field (large language models and what came before) and methods for evaluating understanding in these models

# Grading

- Homeworks 20%
- Quiz + midterm 35%
- Project 40%
- Participation 5%

Slides adopted from: Alane Suhr, Kasia Hitczenko, Micha Elsner, Nathan Schneider, Sharon Goldwater, Hal Daumé III, Ellie Pavlick, Sam Bowman, Dan Jurafsky, Christopher Potts, Emily Bender