# Generative Models on Text
# Large Language Models

## Introduction, Background and Linguistic Concepts

Marion Di Marco

14. 04. 2026

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# LLMs: Introduction

- LLMs: models to perform natural language processing tasks
- Trained on large amounts of text data → reproduce natural language patterns
- Capable of understanding and generating human-like text
- Computational agents that can interact with people using natural language

# LLMs: Introduction and Examples

**SYSTEM**

You are a helpful assistant.

**USER**

Explain in 3 sentences:
What is the difference between a turtle and a tortoise?

**ASSISTANT**

The primary difference between a turtle and a tortoise lies in their habitat: turtles are generally adapted to living in water or near aquatic environments, while tortoises are land-dwellers. Turtles typically have more streamlined shells for swimming, while tortoises have rounder, heavier shells for protection on land. Additionally, turtles often possess webbed feet or flippers for swimming, whereas tortoises have sturdy, columnar legs for walking on rugged terrain.

# LLMs: Introduction and Examples

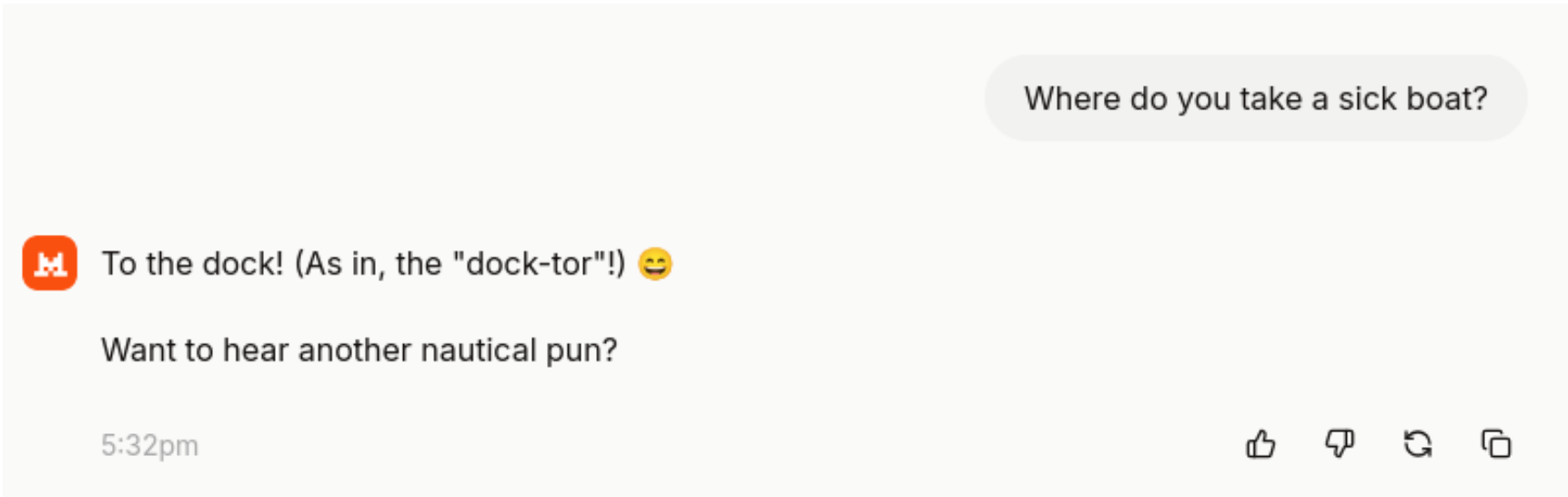
Translate the following text into German: "The Simpsons takes place in a fictional American town called Springfield. Although there are many real settlements in America named Springfield, the town the show is set in is fictional. The state it is in is not established. In fact, the show is intentionally evasive with regard to Springfield's location. Springfield's geography and that of its surroundings is inconsistent: from one episode to another, it may have coastlines, deserts, vast farmland, mountains, or whatever the story or joke requires."

Hier ist die Übersetzung ins Deutsche:

**Die Simpsons** spielt in einer fiktiven amerikanischen Stadt namens **Springfield**. Obwohl es in Amerika viele echte Siedlungen mit dem Namen Springfield gibt, ist die Stadt, in der die Serie spielt, fiktiv. In welchem Bundesstaat sie liegt, ist nicht festgelegt. Tatsächlich geht die Serie in Bezug auf den Standort von Springfield absichtlich ausweichend vor.

Die Geografie von Springfield und seiner Umgebung ist **inkonsistent**: Von einer Folge zur nächsten kann die Stadt Küstenlinien, Wüsten, weites Ackerland, Berge oder was auch immer die Handlung oder der Witz gerade erfordert, aufweisen.

# LLMs: Introduction and Examples



# Language and Learning from Text

- Humans can understand and produce language effortlessly
  - understanding of linguistic structure, world knowledge
  - ability to generalize and interpret context
- Vocabulary: words and their meanings and usage
  - estimated vocabulary size of a young adult speaker of American English: 30,000 – 100,000
  - children have to learn 7–10 words a day
- Context: we don't learn words one at a time in isolation; learning through association with co-occurring words
  - children: vocab acquisition as a by-product of reading
- Distributional hypothesis: words that occur in similar contexts tend to have similar meanings

  "You shall know a word by the company it keeps." (Firth, 1957)

# Learning from Text Data

- Idea: text contains enormous amounts of knowledge
- Training on huge text collections: learn knowledge about language and the world → enable LMs to solve many problems
- Large corpora: likely to contain natural examples for NLP tasks
  - question – answer pairs
  - documents + summaries (tl;dr)
  - translations
  - word definitions, explanations
  - and more ...
- Pretraining: learn to predict the next word

# Learning from Text Data

- With roses, dahlias, and peonies, I was surrounded by flowers
  → ontological facts
- The room wasn't just big it was enormous
  → something on the same scale as big but further along on that scale
- The square root of 4 is 2
  → learning math
- The author of "A Room of One's Own" is Virginia Woolf
  → learning facts about the world and historical authors
- The professor said that he
  → learn to associate professors with male pronouns?

Examples from Jurafsky and Martin (2026)

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Levels of Linguistics

**Linguistics:** the scientific study of language and its structure

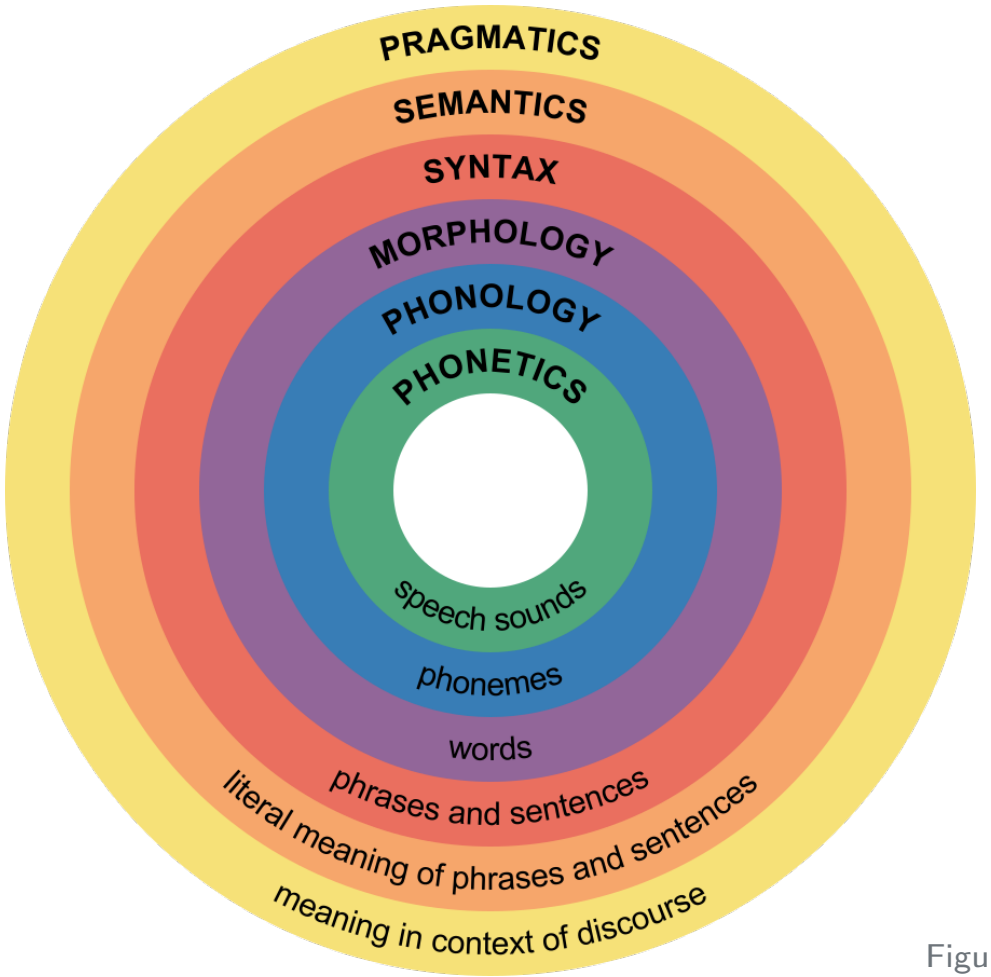



Figure from https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

# Levels of Linguistics

- Spoken language
  - phonetics: production and perception of speech sounds
  - phonology: relations between speech sounds in languages
- Written language
  - Morphology: structure and composition of words
  - Syntax: structure of phrases and sentences
  - Semantics: meaning of phrases and sentences
  - Pragmatics: meaning and intended meaning in a discourse context

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

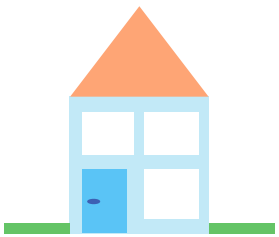
Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Words

- Word: basic atomic unit of meaning

*house* → 

- Adapt the meaning based on the context
  - *... their parents' house ...*
  - *... the White House ...*

- Almost all uses of *house* are connected to the basic unit of meaning

- Smaller units such as syllables or sounds (*hou* or *s*) do not evoke the mental image of *house*

# What is a Word?

- Notion of words seems straightforward for English → space separated
- Some writing systems do not clearly mark words as unique units
  for example, Chinese is written without spaces between the words
- Complex words and compounding: some words appear to be one word, but consist of several parts
  - English: *homework, tumbledown, blackboard*
  - German: *Apfelkuchen (apple cake), feuerlöscherrot (fire extinguisher red)*

    *Rinderkennzeichnungsfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*[1]
  - Finnish: *istahtaisinkohankaan (I wonder if I should sit down for a while after all)*[2]

[1] https://www.duden.de/sprachwissen/sprachratgeber/Die-langsten-Worter-im-Dudenkorpus
[2] https://en.wikipedia.org/wiki/Finnish_language

# Example: Agglutinative Languages

- Agglutination: process of forming new words by concatenating morphemes that correspond to syntactic features

| Turkish | English |
|---|---|
| duy(-mak) | *(to) sense* |
| duygu | *sensation* |
| duygusal | *sensitive* |
| duygusallaş(-mak) | *(to) become sensitive* |
| duygusallaştırıl(-mak) | *(to) be made sensitive* |
| duygusallaştırılmış | *the one who has been made sensitive* |
| duygusallaştırılamamış | *the one who could not have been made sensitive* |
| duygusallaştırılamamışlardan | *from the ones who could not have been made sensitive* |

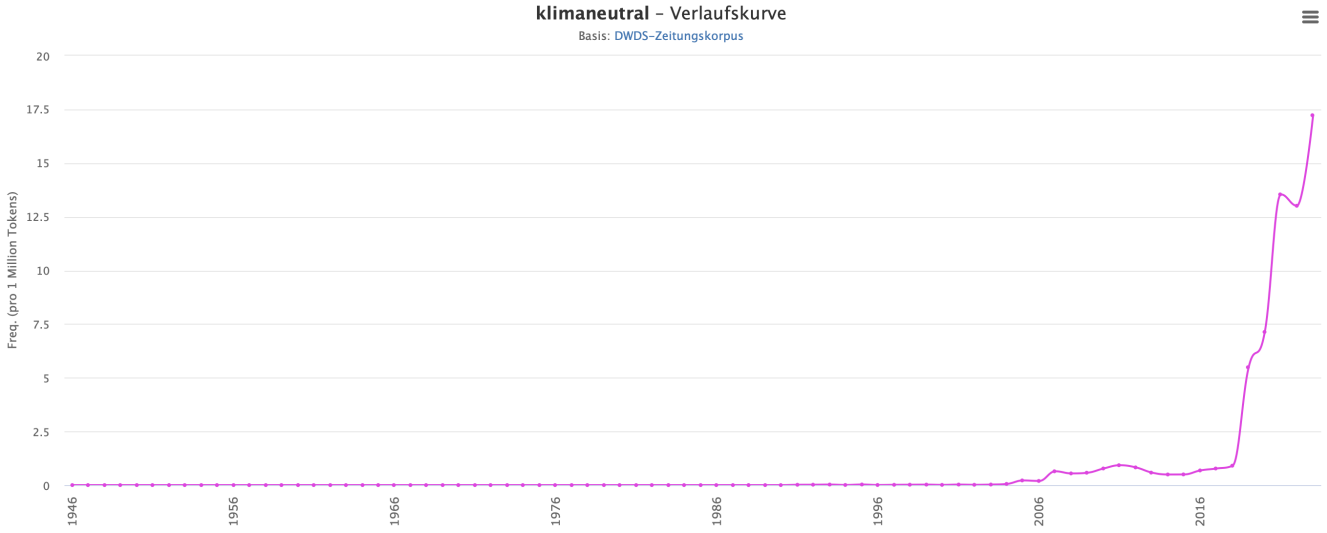
- For the sake of simplicity:
  assume words (=sequences between spaces) as basic units of meaning
- Note: focus mainly on English, but there is also a lot of work
  looking into modeling morphologically complex languages!

Figure from Ataman et al. (2017)

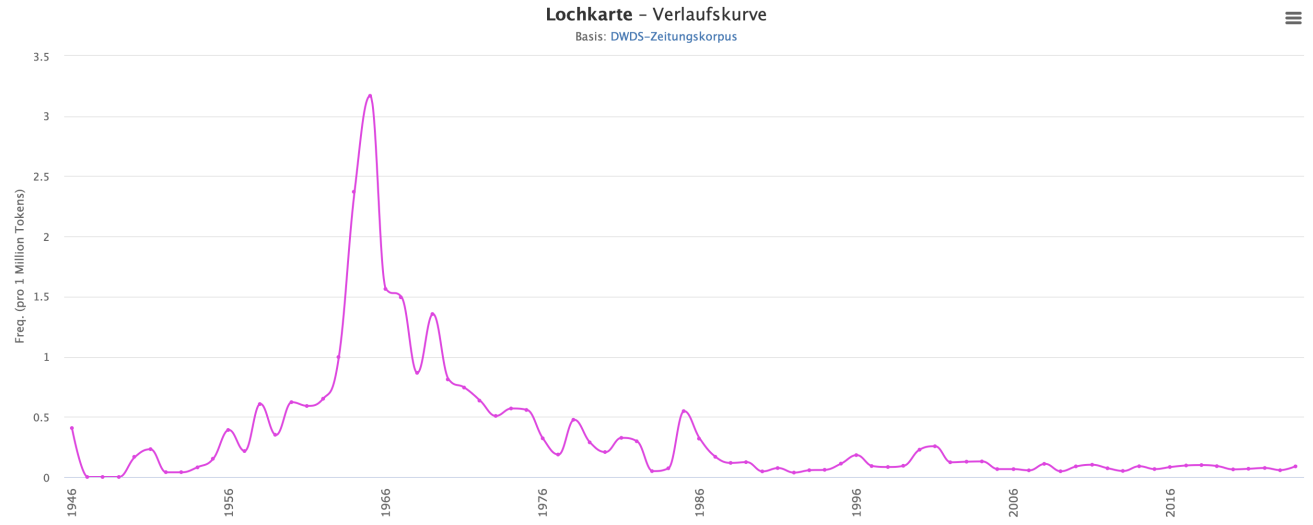
# What are the Words of a Language?

## New words emerge, others fall out of use:

https://www.dwds.de/r/plot




"climate neutral"




"punch card"

# Corpora and Word Distribution

- The vocabulary of a language is fluid
- In practice: text corpus with a fixed set of words
- Continually update with new data → larger corpora

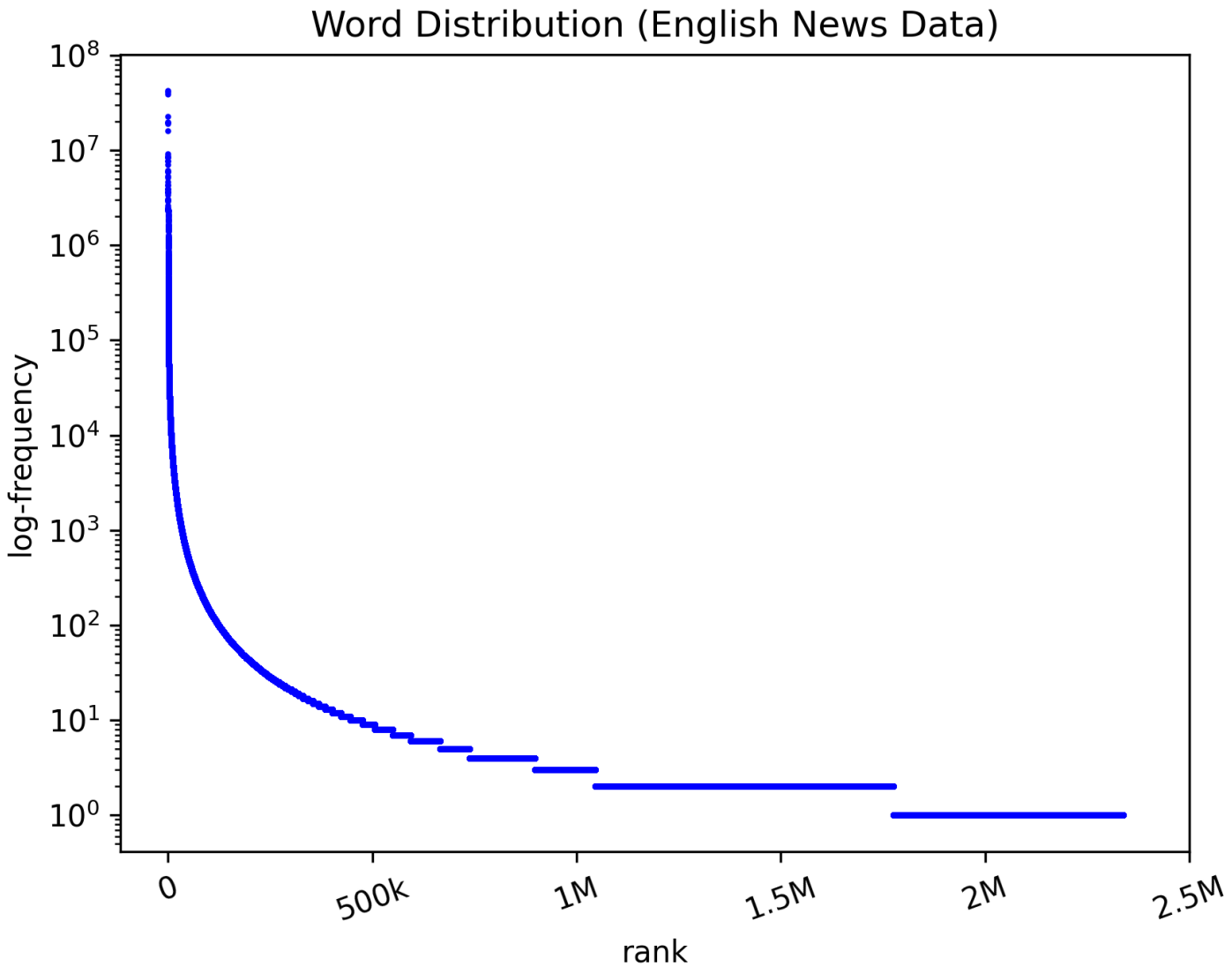
- English news data (33M sentences):

| freq | word |
|---|---|
| 42380661 | , |
| 40887715 | the |
| 38696981 | . |
| 22720213 | to |
| 19785952 | and |
| 19644063 | of |
| 19025360 | a |
| 15930678 | in |
| 9164833 | 's |
| ... | ... |

| freq | word |
|---|---|
| 17313 | timing |
| 17304 | filming |
| 17303 | overcome |
| 17300 | magic |
| 17299 | innocent |
| 17296 | admit |
| 17278 | patterns |
| 17275 | rolling |
| 17269 | formally |
| ... | ... |

| freq | word |
|---|---|
| 3 | yoghurt-coated |
| 3 | yesteray |
| 3 | yellow-beaked |
| 3 | worrried |
| 3 | womansplain |
| ... | ... |
| 2 | ruminococcaceae |
| ... | ... |
| 1 | north-northwestern |
| ... | ... |

# Corpora and Word Distribution



# Morphology: what is inside a Word?

- Morphology: studies the internal structure and composition of words
- *Unübersetzbarkeit* →

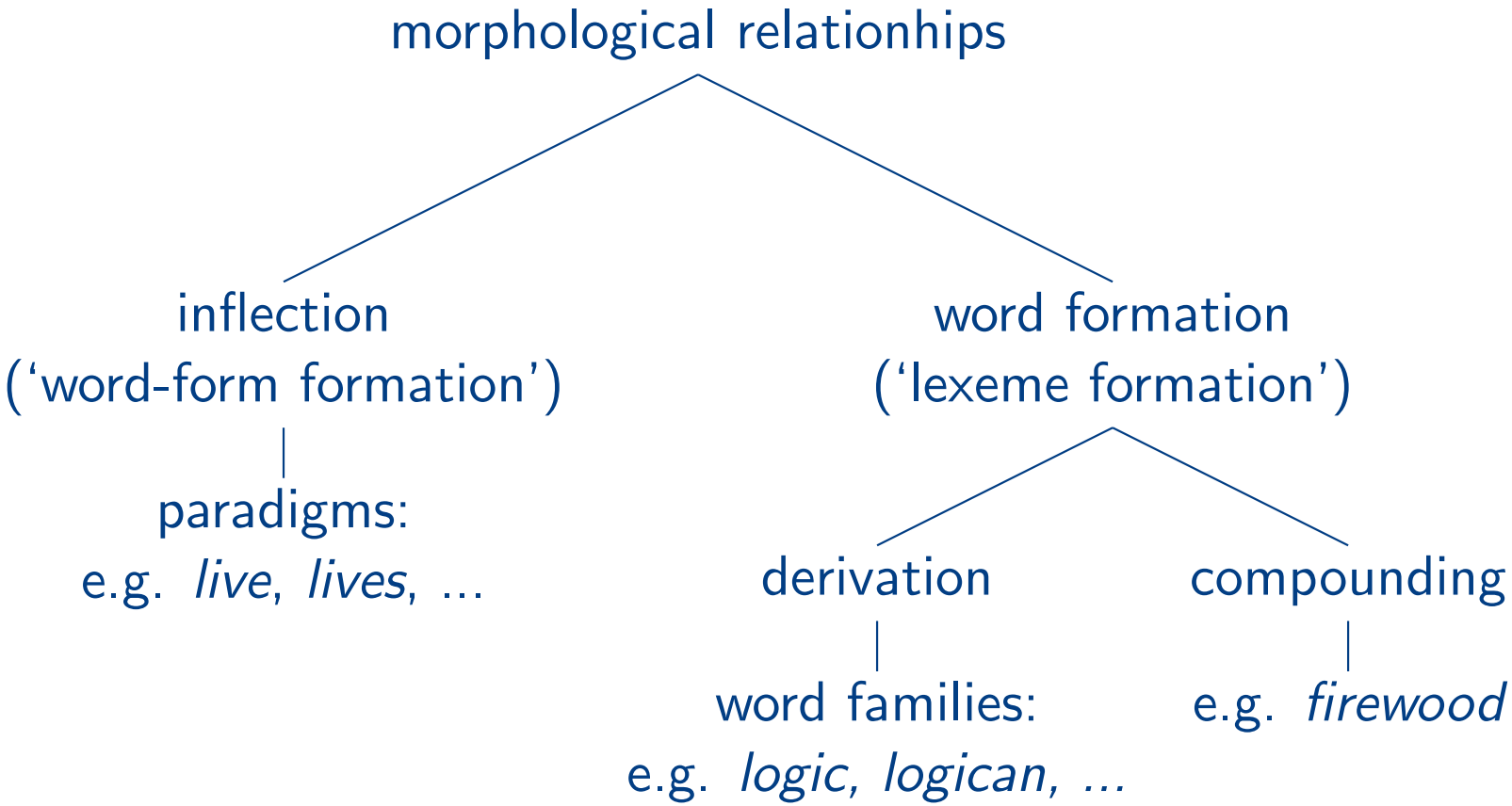
  un$_{Pref}$ über$_{Part}$ setzen$_{Verb}$ bar$_{Suffix_ADJ}$ keit$_{Suffix_N}$
- *untranslatability* →

  un$_{Pref}$ translate$_{Verb}$ able$_{Suffix_ADJ}$ ity$_{Suffix_N}$

# Segmentation: Morphemes

- Many words can be **segmented** into individually **meaningful parts**
  - *read* *read-s* *read-er* *read-able*
    *wash* *wash-es* *wash-er* *wash-able*
    *write* *write-s* *writ-er* *writ-able*
  - *kind* *kind-ness* *un-kind*
    *happy* *happi-ness* *un-happy*
    *friend-ly* *friend-li-ness* *un-friend-ly*
- These meaningful parts are called **morphemes**
- Morphemes are the ultimate elements of morphological analysis; they are, so to speak, morphological atoms

# Morphological Processes



# Morphological Complexity

- Morphologically poor languages: express relationships between words mostly with function words
- Morphologically rich languages: morphological variations
  - verbal inflection
  - nominal inflection
  - word formation processes: for example compounding
- More morphological variation: larger vocabulary of surface forms
- Large vocabulary → data sparsity
  - some forms only occur infrequently or even not at all
  - generally challenging for NLP

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Parts of Speech and POS tagging

- Parts of speech: grammatical categories or word classes
- Words within the same word class: similar syntactic behaviour and similar grammatical properties

- Useful clues to sentence structure and meaning
  - provide shallow syntactic structure
  - observe "patterns": English nouns are preceded by determiners and adjectives
- POS tagging: assigning each word in a sequence a part of speech like noun or verb

# Word Classes, Function Words and Content Words

- **Word classes**: loosely correspond to semantic properties
  - adjectives → properties "green"
  - nouns → people, things "mango"
  - verbs → activities "eat"

- **Function words** → closed-class words
  - words with little to no lexical meaning, like prepositions
  - occur frequently and contribute to the structure of a sentence
  - small set of words, make up a large part of the overall word count

- **Content words** → open-class words
  - infinite amount of words providing lexical content
  - nouns, verbs, adjectives, adverbs
  - new words are coined frequently (e.g. *barbiecore, greedflation* )
  - continually evolving non-finite set of words

# English Word Classes

| | Tag | Description | Example |
|---|---|---|---|
| Open Class | **ADJ** | Adjective: noun modifiers describing properties | *red, young, awesome* |
| | **ADV** | Adverb: verb modifiers of time, place, manner | *very, slowly, home, yesterday* |
| | **NOUN** | words for persons, places, things, etc. | *algorithm, cat, mango, beauty* |
| | **VERB** | words for actions and processes | *draw, provide, go* |
| | **PROPN** | Proper noun: name of a person, organization, place, etc.. | *Regina, IBM, Colorado* |
| | **INTJ** | Interjection: exclamation, greeting, yes/no response, etc. | *oh, um, yes, hello* |
| Closed Class Words | **ADP** | Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation | *in, on, by, under* |
| | **AUX** | Auxiliary: helping verb marking tense, aspect, mood, etc., | *can, may, should, are* |
| | **CCONJ** | Coordinating Conjunction: joins two phrases/clauses | *and, or, but* |
| | **DET** | Determiner: marks noun phrase properties | *a, an, the, this* |
| | **NUM** | Numeral | *one, two, 2026, 11:00, hundred* |
| | **PART** | Particle: a function word that must be associated with another word | *'s, not, (infinitive) to* |
| | **PRON** | Pronoun: a shorthand for referring to an entity or event | *she, who, I, others* |
| | **SCONJ** | Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement | *whether, because* |
| Other | **PUNCT** | Punctuation | ; , () |
| | **SYM** | Symbols like $ or emoji | $, % |
| | **X** | Other | asdf, qwfg |

**Figure 8.1** The 17 parts of speech in the Universal Dependencies tagset (de Marneffe et al., 2021). Features can be added to make finer-grained distinctions (with properties like number, case, definiteness, and so on).

Figure from Jurafsky and Martin (2026)

# POS Tagging

- **Part of speech tagging**: assigning a POS tag to every word in a tokenized sentence

- Input: sentence $x_1, x_2, ..., x_n$ and a tagset
- Output: a corresponding sequence of tags $y_1, y_2, ..., y_n$

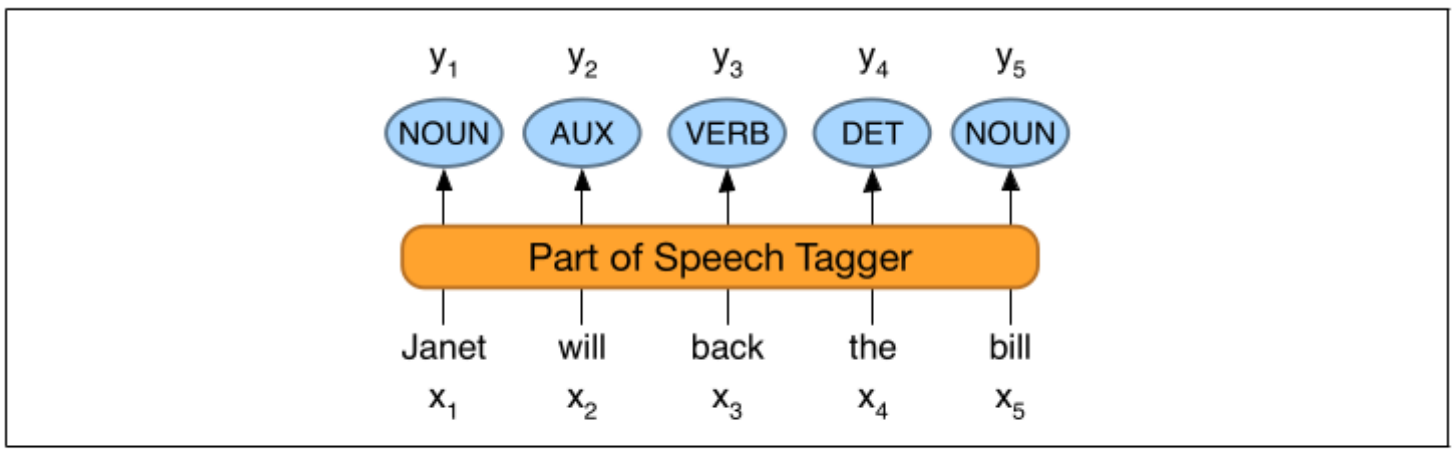



**Figure 8.3** The task of part-of-speech tagging: mapping from input words $x_1, x_2, ..., x_n$ to output POS tags $y_1, y_2, ..., y_n$ .

Figure from Jurafsky and Martin (2026)

# POS Tagging: Challenges

- The farm was used to **produce produce**.
- The dump was so full that it had to **refuse** more **refuse**.
- We must **polish** the **Polish** furniture.
- When shot at, the **dove dove** into the bushes.
- There was a **row** among the oarsmen about how to **row**.
- They were too **close** to the door to **close** it.
- The **wind** was too strong to **wind** the sail.

Examples from `https://writetouch.ca/writing/heteronyms/`

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Sentences

- Words → atomic units of meaning
- Sentence → combination of words following the rules of a language

(1) *Jane bought the house*

→ the **verb** *bought* is the central element

→ the verb has two arguments:
**subject** *Jane* and **object** *house*

(2) *Jane gave Alice a cookie.*

→ *gave/give* has three arguments:
**subject** *Jane* and **direct object** *cookie* and **indirect object** *Alice*

# Syntax

- **Syntax** studies the arrangement of words and their relations:
  how to combine words into larger units such as phrases or sentences?
- Idea: capture and formalize the structure of a language
- The syntax of a language is defined by a grammar
  can we fully define and write up a complete grammar of language?
- Units of language
  - words: basic unit of meaning
  - phrases: sequences of words building a conceptual unit
  - sentences: grammatically independent linguistic units

# Grammars: Toy Example

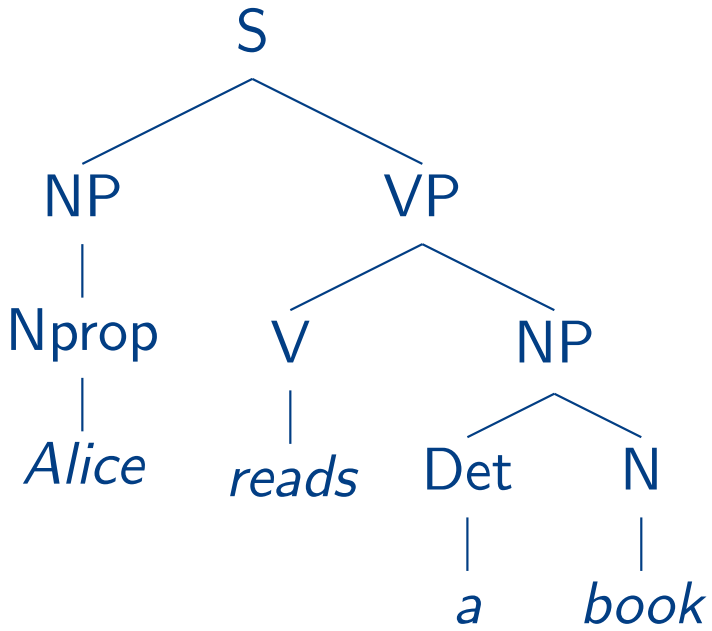
Alice reads a book.
Alice sleeps.

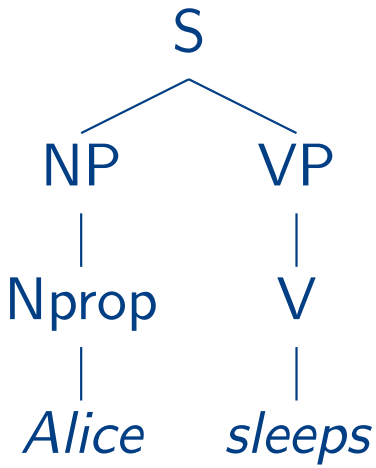
| | | | |
|---|---|---|---|
| S | → | NP VP | non-terminal symbols |
| VP | → | V NP | |
| VP | → | V | |
| NP | → | Det N | |
| NP | → | Nprop | |

---

| | | | |
|---|---|---|---|
| V | → | sleeps \| reads | terminal symbols |
| Det | → | a | |
| N | → | book | |
| Nprop | → | Alice | |

# Grammars: Toy Example




- Parse trees: illustrate the grammatical structure of a sentence
- Different grammar formalism to express the structure of a sentence (for example: dependency structures, constituency grammar )

# Syntax across Languages

- Linguistic concepts are realized differently across languages

- **Isolating Languages**
  - each morpheme is also a word and vice versa
  - for example, Chinese and Vietnamese
  - Mandarin Chinese: 一天, yì tiān “one day”, 三天, sān tiān “three day”.
    no inflection for number in English: *one day, three days*

- **Analytic languages**
  - low ratio of morphemes to words
  - syntactic information is mainly expressed by means of function words
  - syntactic functions (subject, object) are assigned via word order
  - for example English, Norwegian, Danish

# Syntax across Languages

- **Synthetic languages**
  - grammatical information represented by (inflectional) morphology
  - relatively free word order
  - for example Slavic languages, German, Finnish, Turkish
- **Agglutinative languages**
  - combine one or more morphemes into one word
  - each morpheme is individually identifiable as a meaningful unit



- **Fusional languages**
  - morpheme combinations do not remain distinct and fuse together
  - one morpheme to denote numerous grammatical or syntactic features

Illustration from `https://opentextbc.ca/psyclanguage/chapter/morphology-of-different-languages/`

# Morphological Typology

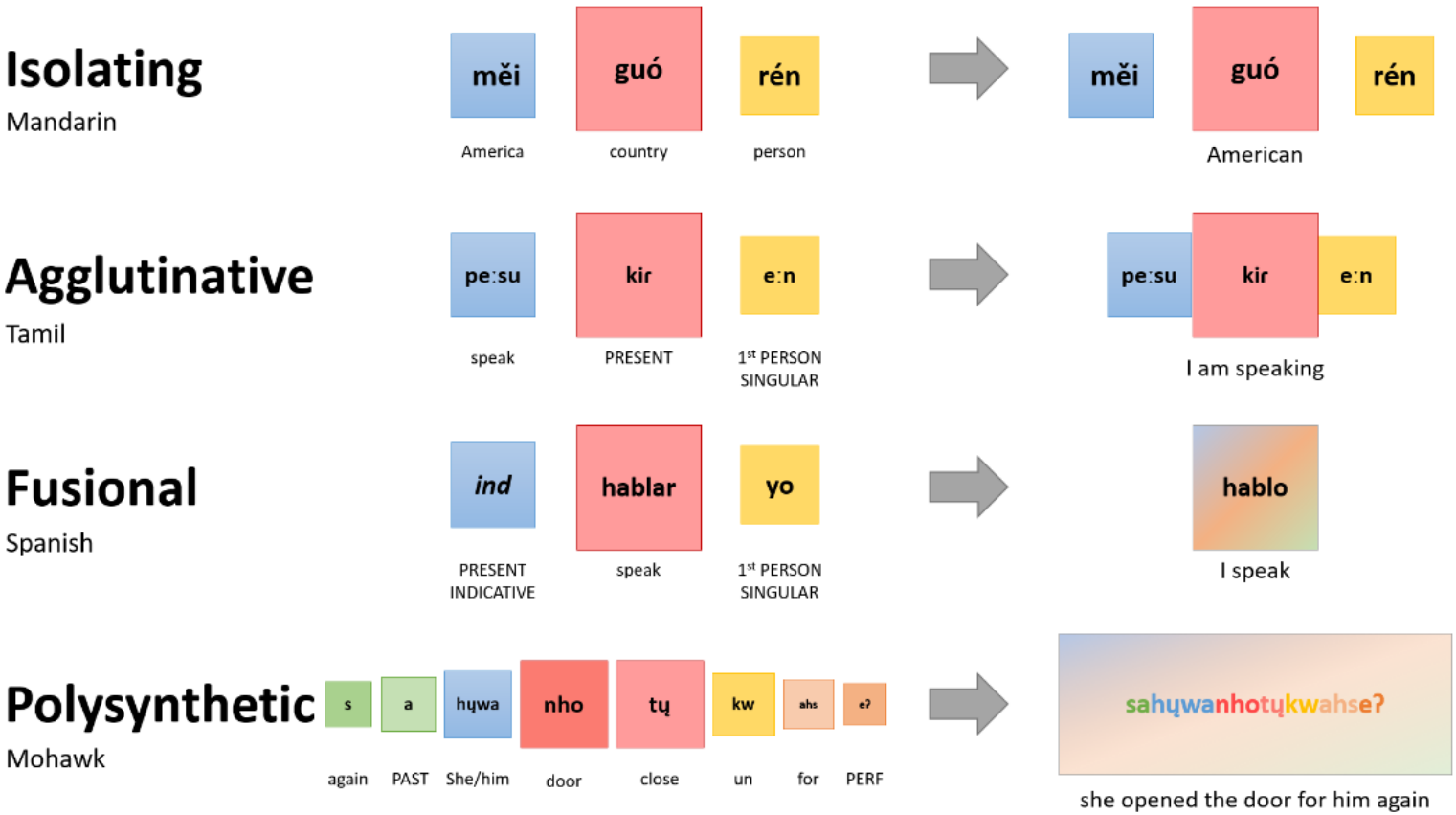



Illustration from `https://opentextbc.ca/psyclanguage/chapter/morphology-of-different-languages/`

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Semantics

- **Semantics:** study of linguistic meaning
- Lexical semantics:
  - analysis of word meanings and compositionality
  - relations between words
- Formal semantics
  - relies on logic and mathematics
  - provides precise frameworks of the relation between language and meaning

$$\frac{\text{All humans are mortal}\quad \text{Socrates is a human}}{\text{Socrates is mortal}}$$

- **Pragmatics**: investigates how people use language in communication

"It's cold in here, isn't it?" (looks towards the open window)

# Semantic Roles

- XYZ corporation bought the stock.
- They sold the stock to XYZ corporation.
- The stock was bought by XYZ corporation.
- The purchase of the stock by XYZ corporation...
- The stock purchase by XYZ corporation...

- Purchase event: described by the verbs *bought*, *sold*
- Participants: *XYZ Corp* and *some stock*
- Semantic roles $\neq$ syntactic subject/object
- Semantic role labeling: the task of assigning roles to spans in sentence

Example from Jurafsky and Martin

# Outline

Introduction: What are LLMs?

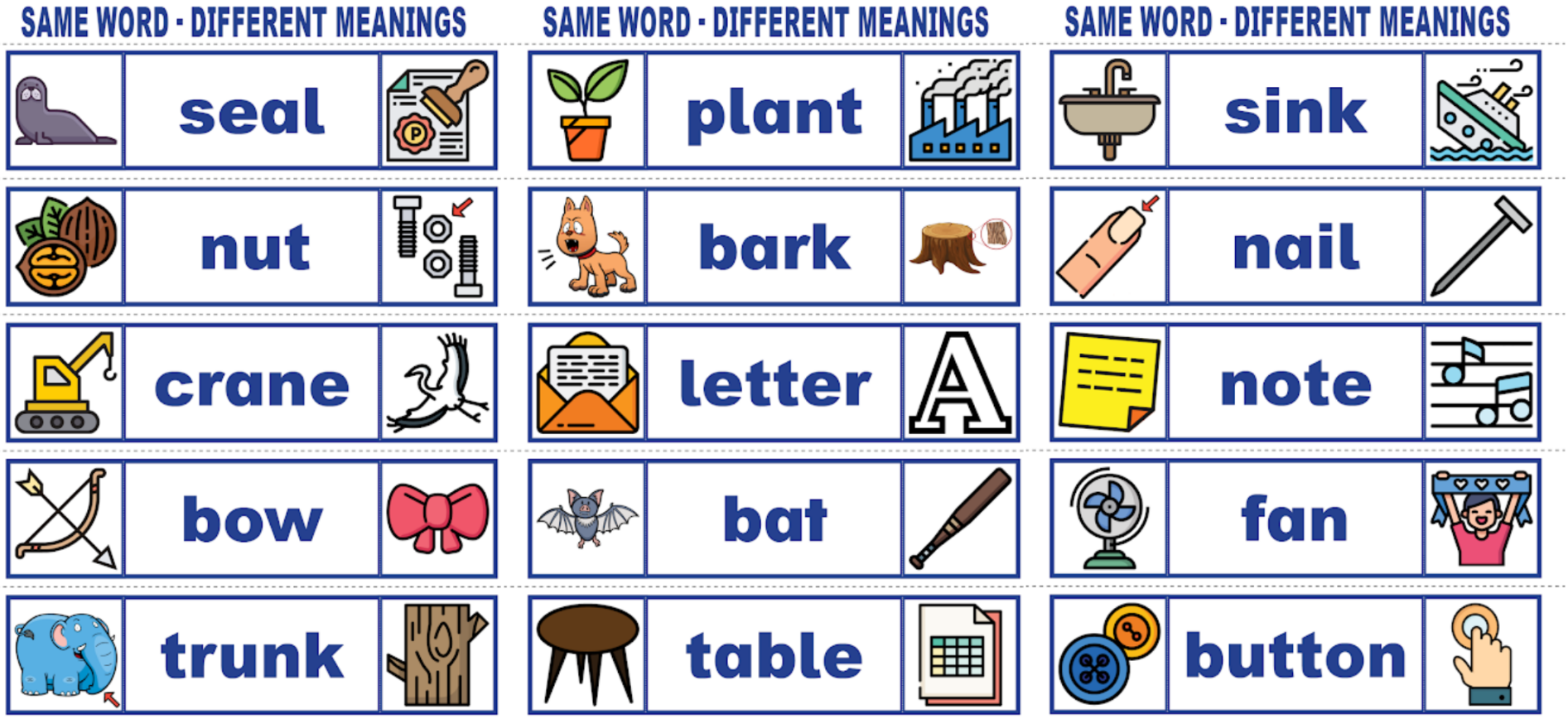
Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Word Senses




Illustrations from https://www.englishforkidz.com/2020/01/vocabulary-cards-word-different-meanings.html

# Challenges: Structure and Attachment

- *The boy saw the man with a telescope.*
  - who has the telescope?
  - note: in some languages, there is no ambiguity

- *a criminal lawyer*
  - is the lawyer criminal?
  - does the lawyer practice criminal law?

- *local phone compay*, *live fish transporter*

- *Wachstube*

| | |
|---|---|
| *Wach\|Stube* | → guard room |
| *Wachs\|Tube* | → tube of wax |

# Challenges: Non-compositional Phrases

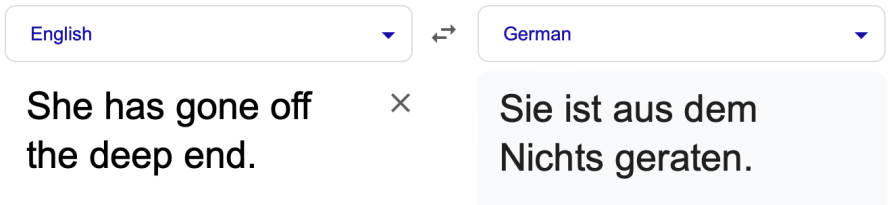
- Some words cannot be interpreted literally, but only in their context
- For example: idiomatic expressions

  *she's gone off the deep end*
  *er hat nicht mehr alle Tassen im Schrank*

- All words in the phrases have a distinct meaning that is not related to the meaning of the phrase (*crazy/verrückt*)
- Context: need to consider the entire phrase to derive the meaning
- Challenging for many NLP tasks

# Challenge: Translate Non-Compositional Phrases

## Google Translate:




## GPT-4o:

**SYSTEM**

Translate into German.

**USER**

She has gone off the deep end.

**ASSISTANT**

Sie hat den Verstand verloren.

**SYSTEM**

Translate into German. Please give 3 possibilities.

**USER**

She has gone off the deep end.

**ASSISTANT**

1. Sie ist übergeschnappt.
2. Sie dreht durch.
3. Sie hat den Verstand verloren.

# Challenges in NLP: Understanding and Interpretation

| | |
|---|---|
| **I** never said she stole my money. | It was probably someone else but definitely not me who said that. |
| I **never** said she stole my money. | I never said she stole my money. Never. |
| I never **said** she stole my money. | I probably hinted, but did not *say* it. |
| I never said **she** stole my money. | It might be someone associated with her. |
| I never said she **stole** my money. | She borrowed some money from me a long time ago and I was just wondering why she has yet to return it. |
| I never said she stole **my** money. | She stole someone else's money. |
| I never said she stole my **money**. | She stole my credit card. |

# Confusing Headlines

Violinist Linked to JAL Crash Blossoms

Local High School Dropouts Cut in Half

Hospitals Are Sued by 7 Foot Doctors

**Noun** [ edit ]

**crash blossom** (*plural* **crash blossoms**)

1. (*linguistics*) A sentence, often a news headline, that is subject to incorrect interpretation due to syntactic and/or lexical ambiguity. [quotations ▼]

Examples from `https://web.stanford.edu/ jurafsky/NLPCourseraSlides.html` (Introduction.pdf)

`https://en.wiktionary.org/wiki/crash_blossom`

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Pretraining LLMs

- Text contains enormous amounts of knowledge!

- Language models are trained on huge amounts of data, often on multilingual training data
- **No explicit linguistic information**
- Multilingual LLMs: no explicit language information

- LLMs are not trained on words, but **subwords**
  - efficiency: restricted vocabulary size
  - handle unknown words

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# What is an LLM?

- A model that can predict next words from previous words
- Given a context of words: LLM assigns probability distribution over the possible next words

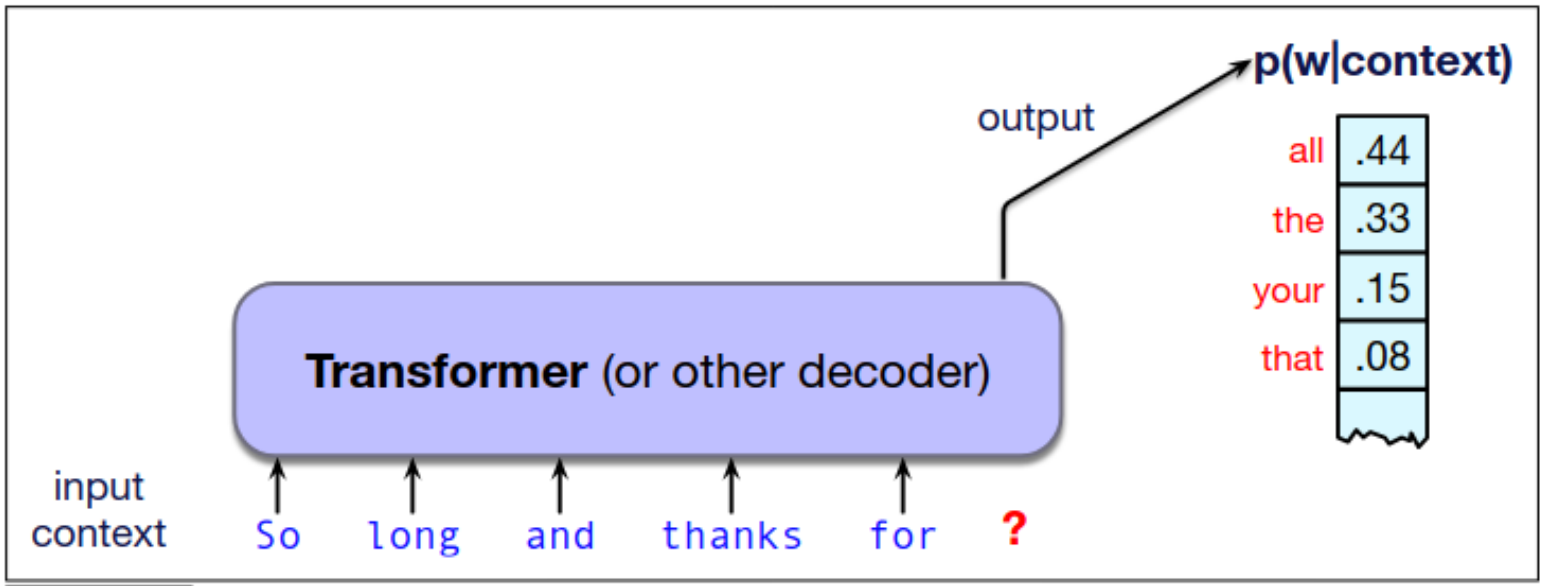



Figure from Jurafsky and Martin (2026)

# Generating with LLMs

- *Generate* text by sampling from the distribution

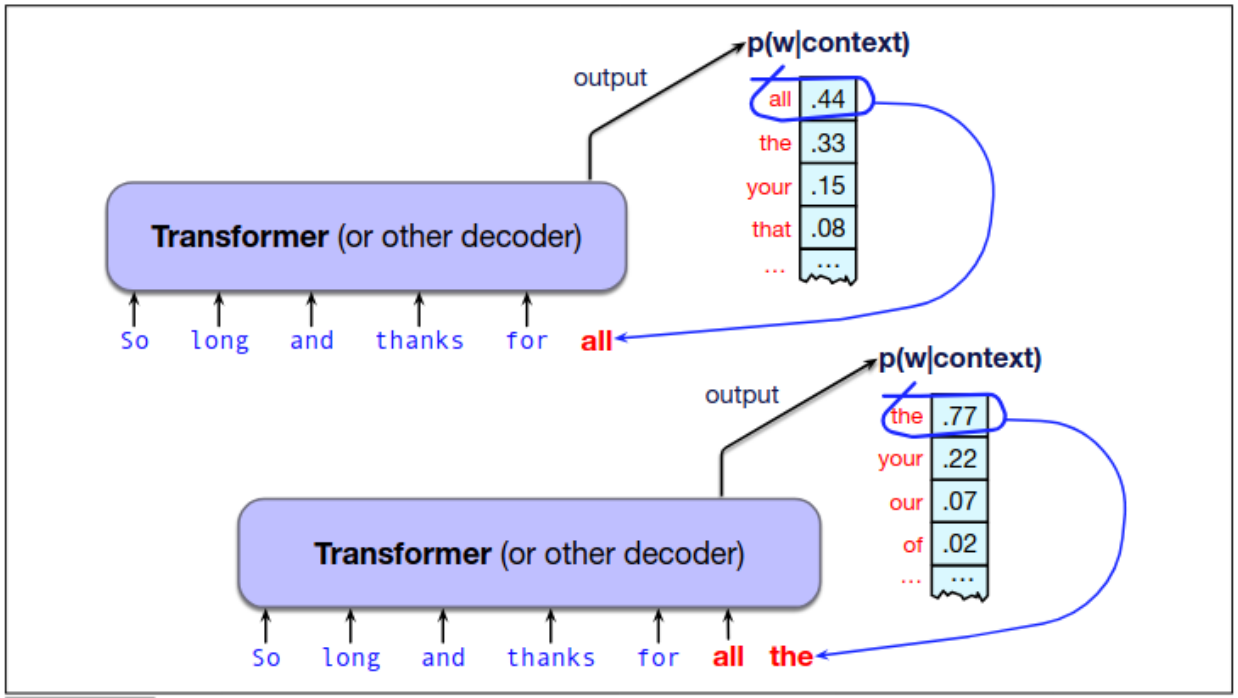



**Figure 7.2** Turning a predictive model that gives a probability distribution over next words into a generative model by repeatedly sampling from the distribution. The result is a left-to-right (also called autoregressive) language model. As each token is generated, it gets added onto the context as a prefix for generating the next token.

- Conditioning on priming context and subsequently generated outputs
- Causal (autoregressive) LMs

Figure from Jurafsky and Martin (2026)

# Training Stages

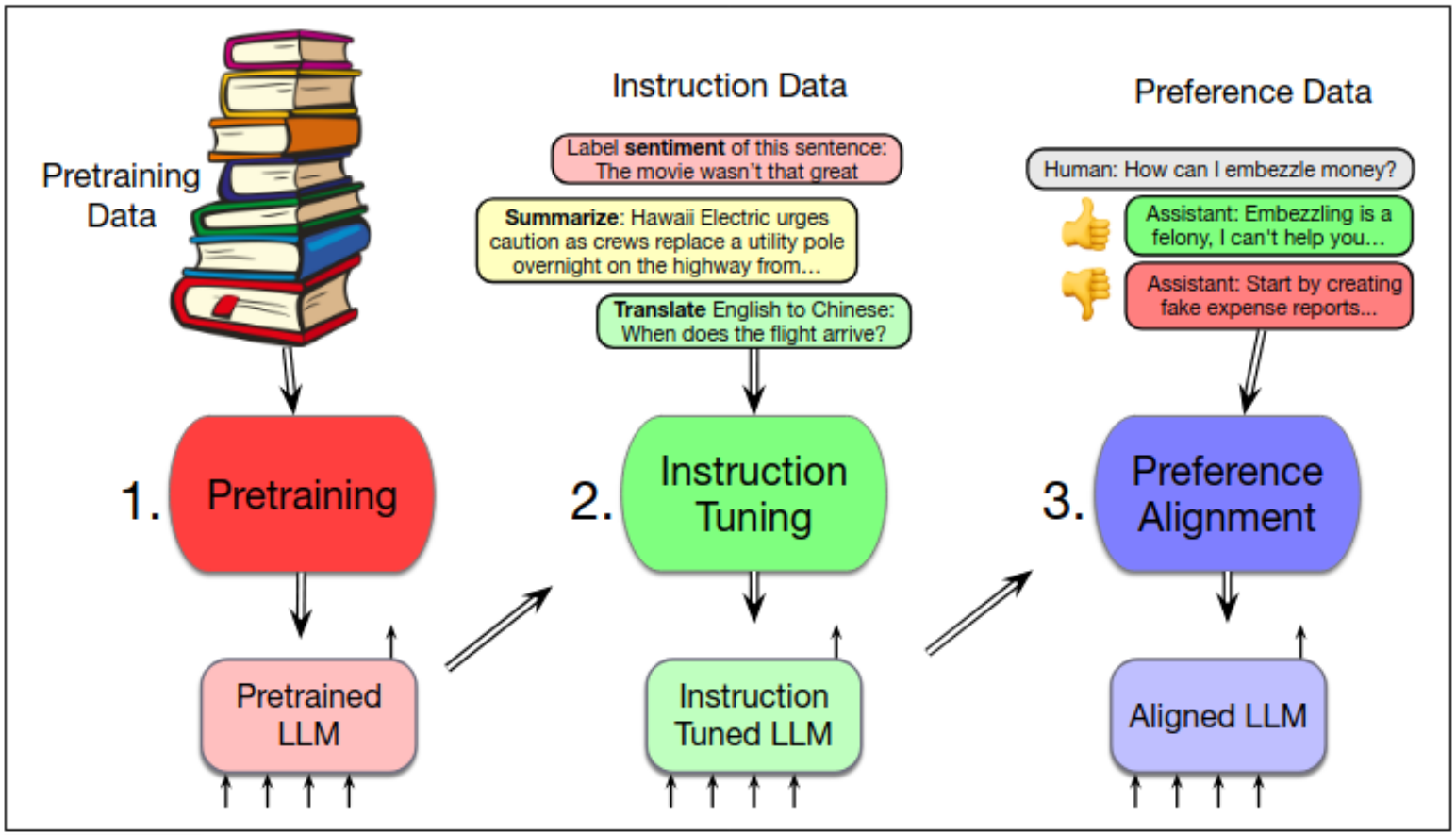



**Figure 7.12** Three stages of training large language models: pretraining, instruction tuning, and preference alignment.

Figure from Jurafsky and Martin (2026)

# Training Stages

- **Pretraining**
  - train a model on next-word prediction
  - self-supervised training: at each time step $t$, predict word $t+1$
- **Instruction tuning, supervised finetuning**
  - train the model to follow instructions
  - question answering, translation, summarization, code writing, ...
  - trained on instructions + responses to instructions
- **Preference alignment**
  - trained the model to make it maximally helpful and less harmful
  - trained on preference data: context and two potential continuations labeled as "accepted" or "rejected"
  - reinforcement learning to prefer the accepted continuation

# Training Data

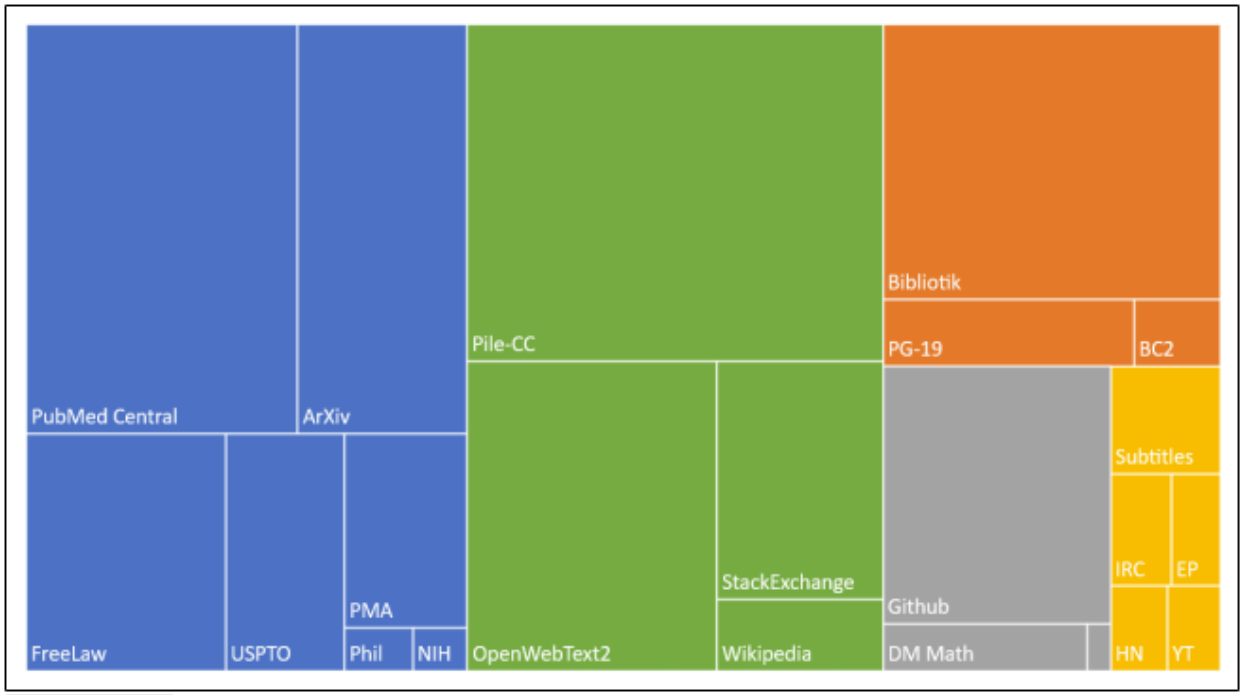



**Figure 10.5** The Pile corpus, showing the size of different components, color coded as academic (articles from PubMed and ArXiv, patents from the USPTA; internet (webtext including a subset of the common crawl as well as Wikipedia), prose (a large corpus of books), dialogue (including movie subtitles and chat data), and misc.. Figure from Gao et al. (2020).

- LLMs are mainly trained on web data
  - quality
  - safety

Figure from Jurafsky and Martin (2026)

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Subword Segmentation

- LLM pretraining data is segmented into subwords,
  for example using BPE (Byte Pair Encoding) Sennrich et al. (2016)
- Frequency-based compression algorithms:
  - start with small vocabulary
  - iteratively merge the most common tuples
    until the desired vocabulary size is reached

  → keep frequent words intact, segment less frequent ones
- Example: `t h e c a t s a t o n t h e m a t`

  assuming "`t h`" is the most frequent tuple given an EN corpus:

  `th e c a t s a t o n th e m a t`
- Segmented words: playing → `play ##ing`

# Subword Segmentation: Examples (GPT)

```
Many words map to one token, but some don't: indivisible.

The Nile crocodile (Crocodylus niloticus) is a large crocodilian native
 to freshwater habitats in Africa. It is widely distributed in sub
-Saharan Africa.

Das Nilkrokodil ist das größte Krokodil Afrikas und erreicht normaler
weise Längen von 3 bis 4 m.
```

## Segmentation examples – related words:

```
decipher
deciphers
deciphered
deciphering
decipherable
decipherment
indecipherable
```

```
entschlüsseln
entschlüsselbar
unentschlüsselbar
Entschlüsselung
verschlüsseln
unverschlüsselt
```

# Example Segmentation: German vs. Czech

| Tokens | Characters |
| --- | --- |
| 144 | 504 |

Es besteht aus der Sonne, acht sie umkreisenden Planeten (von innen nach
außen: Merkur, Venus, Erde, Mars, Jupiter, Saturn, Uranus und Neptun),
deren natürlichen Satelliten, den Zwergplaneten, anderen Kleinkörpern (K
ometen, Asteroiden und Meteoroiden) und aus unzähligen Gas- und Staubteil
chen, die durch die Anziehungskraft der Sonne an diese gebunden sind.
Die Internationale Astronomische Union definiert den Pluto seit 2006 als
Zwergplanet und nicht mehr als den äußersten Planeten des Sonnensystems.

| Tokens | Characters |
| --- | --- |
| 228 | 526 |

Sluneční soustava je planetární systém hvězdy známé pod názvem Slunce, ve
kterém se nachází planeta Země.
Systém tvoří především 8 planet, 5 trpasličích planet, přes 150 měsíců
planet (především u Jupiteru, Saturnu, Uranu a Neptunu) a další menší tě
lesa jako planetky, komety, meteoroidy apod., které jsou soustředěny před
evším v Hlavním pásu uvnitř soustavy a Kuiperově pásu na jejím okraji. Te
oreticky sluneční soustavu ještě obklopuje Oortův oblak.
Sluneční soustava je součástí Galaxie tradičně též nazývané Mléčná dráha.

# Morphological Complexity – Czech Inflection

- Inflection paradigm for the Czech adjective *mladý* (*young*)

| | | Masculine animate | Masculine inanimate | Feminine | Neuter |
|---|---|---|---|---|---|
| Sg. | Nominative | mladý | | mladá | mladé |
| | Genitive | mladého | | mladé | mladého |
| | Dative | mladému | | mladé | mladému |
| | Accusative | mladého | mladý | mladou | mladé |
| | Vocative | mladý! | | mladá! | mladé! |
| | Locative | mladém | | mladé | mladém |
| | Instrumental | mladým | | mladou | mladým |
| Pl. | Nominative | mladí | mladé | | mladá |
| | Genitive | mladých | | | |
| | Dative | mladým | | | |
| | Accusative | mladé | | | mladá |
| | Vocative | mladí! | mladé! | | mladá! |
| | Locative | mladých | | | |
| | Instrumental | mladými | | | |

Figure from https://en.wikipedia.org/wiki/Czech_declension

# Subword Segmentation in Multilingual LMs

- Large multilingual LMs are often "English-centric"
  → tokenization often based on English vocabulary
- Segmentation affects the performance for languages other than English

  Armengol-Estapé (2022)
- Cost of using LLMs for under-represented languages
  - LLM usage is charged per token
  - many short tokens → expensive and loss of context

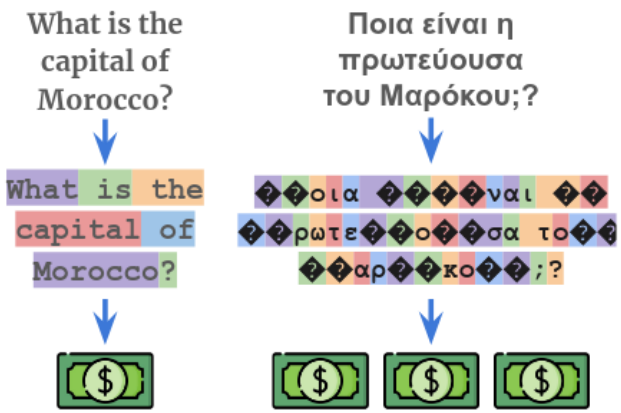



Figure from *Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models.* Ahia et al. (2023)

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

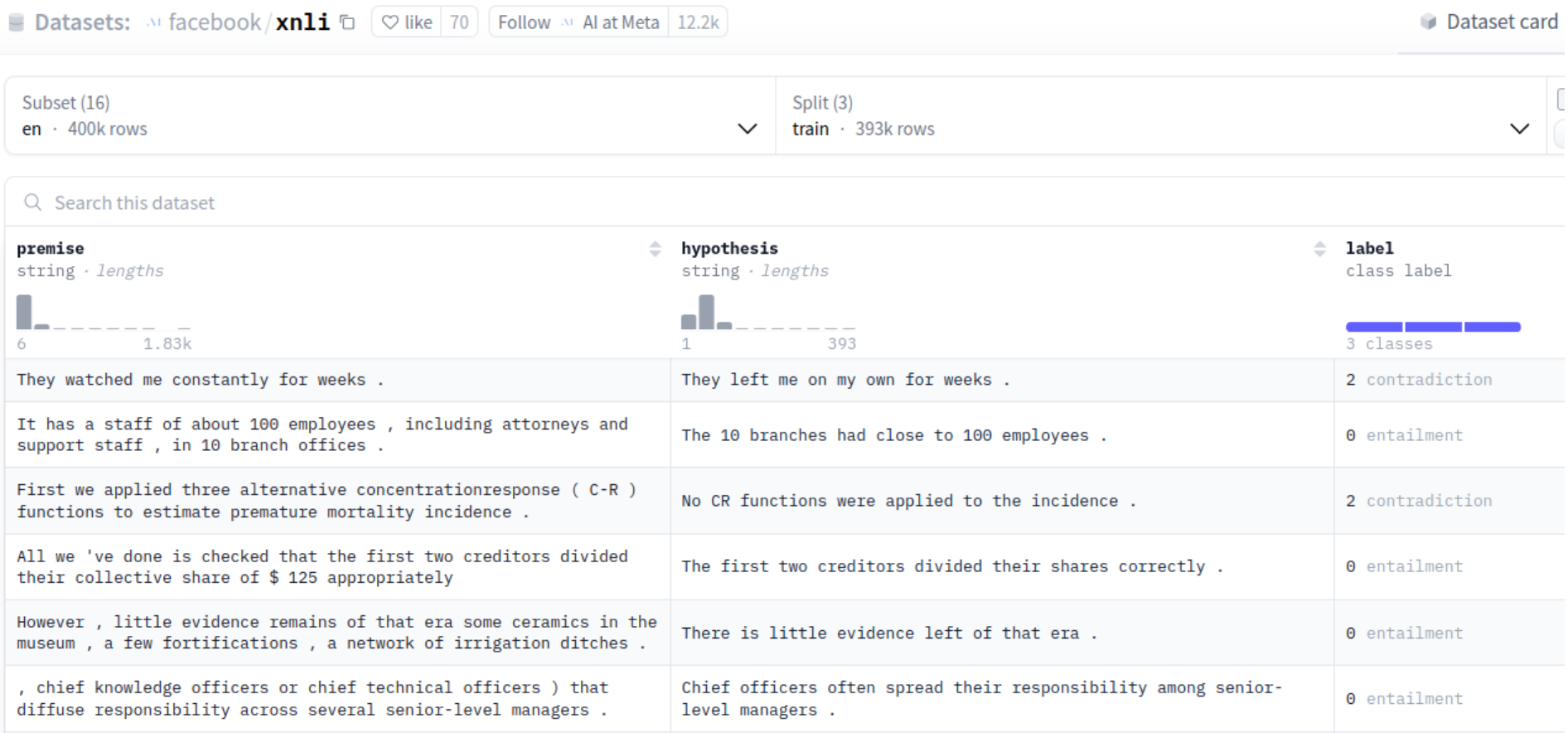
# Benchmarks and Evaluation

Standard evaluation tasks to assess performance of LLMs

**Some examples:**

- GLUE: Wang et al. (2018)
  General Language Understanding Evaluation benchmark
  contains diverse natural language understanding tasks
- XNLI: cross-lingual natural language inference Conneau et al. (2018)
  for 15 languages (including low-resource languages)
- MMLU: Measuring Massive Multitask Language Understanding
  Hendrycks et al. (2021)
- ...

# Dataset: XNLI




https://huggingface.co/datasets/facebook/xnli/viewer/en

# Dataset: MMLU

| | | | |
|---|---|---|---|
| The terrestrial planet cores contain mostly metal because | astronomy | [ "the entire planets are made mostly of metal.", "metals condensed first in the solar nebula and the… | 2 C |
| Why are the inner planets made of denser materials than the outer planets? | astronomy | [ "In the beginning when the protoplanetary disk was spinning faster centrifugal forces flung the lighter materials toward the outer parts of the solar nebula.", "In the inner part of the nebula only metals and rocks were able to condense because of the high temperatures whereas hydrogen compounds although more abundant were only able to condense in the cooler outer regions.", "Denser materials were heavier and sank to the center of the nebula.", "When the solar nebula formed a disk materials naturally segregated into bands and in our particular solar system the denser materials settled nearer the Sun while lighter materials are found in the outer part." ] | 1 B |
| What do meteorites reveal about the solar system? | astronomy | [ "They reveal that the early solar system consisted mostly of hydrogen and helium gas.",… | 2 C |
| Venus shows evidence of which of the following surface processes? | astronomy | [ "Impacts", "Erosion", "Volcanism", "A B and C" ] | 3 D |
| Planetary rings are | astronomy | [ "known to exist for all of the jovian planets.", "composed of a large number of individual particles… | 3 D |

https://huggingface.co/datasets/cais/mmlu/viewer/astronomy?row=30

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# The End – Summary

- Text data contains a lot of knowledge
- Pretraining on enormous amounts of data
- Language
  - humans can understand and produce language effortlessly
  - language is full of challenges → difficult to formalize
- Overview of linguistic concepts: words, part-of-speech, ...
- Subword segmentation for pretraining

**Questions?**

# Next Lecture

- Next week: n-gram models

  `https://web.stanford.edu/~jurafsky/slp3/3.pdf`

  (without section 3.7)

# Outline

Introduction: What are LLMs?

Language and Linguistic Concepts
- Words and Morphology
- Parts of Speech
- Sentences and Syntax
- Semantics
- Challenges and Ambiguities

Large Language Models: some Background
- Training
- Subword Segmentation
- Evaluation

Summary and Outlook

Credits and References

# Credits

Some content is based on *Speech and Language Processing*:
Daniel Jurafsky & James H. Martin (2926)

Large Language Models (Chapter 7)

`https://web.stanford.edu/~jurafsky/slp3/7.pdf`
`https://web.stanford.edu/~jurafsky/slp3/slides/llm25aug.pdf`

# References


- Dan Jurafsky and James H. Martin (2024)
  *Speech and Language Processing (3rd ed. draft)*
  `https://web.stanford.edu/ jurafsky/slp3/`
- Hill, F., R. Reichart, and A. Korhonen. (2015)
  *Simlex-999: Evaluating semantic models with (genuine) similarity estimation.*
  Computational Linguistics, 41(4):665–695.
- Duygu Ataman, Matteo Negri, Marco Turchi, Marcello Federico. (2017)
  *Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English*.
  The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 331-342. doi: 10.1515/pralin-2017-0031
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhard (2021). *Measuring Massive Multitask Language Understanding*. Proceedings of the International Conference on Learning Representations (ICLR).
- Rico Sennrich, Barry Haddow, and Alexandra Birch (2016). *Neural machine translation of rare words with subword units.* ACL 2016.

# References


- Alex Wang, Amanpreet Singh, et al. (2018): *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang (2016): *SQuAD: 100,000+ Questions for Machine Comprehension of Text* Proceedings of EMNLP 2016. `https://aclanthology.org/D16-1264/`
- Alexis Conneau, Guillaume Lample et al. (2018): *XNLI: Evaluating Cross-lingual Sentence Representations* `https://arxiv.org/abs/1809.05053`
- Raffel et al. (2019): *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. `https://arxiv.org/pdf/1910.10683v4`
- Gao et al. (2020): *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. `https://arxiv.org/abs/2101.00027`