

# Generative Models on Text

## Large Language Models

Introduction to NLP and Linguistic Concepts

Marion Di Marco

April 29, 2025

# Outline

---

## What is NLP?

### Challenges in NLP

### Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics


### Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation

# Introduction: What is NLP?


 dinosaur with plates on back

All Images Videos Products Web News Books More Tools



Name T rex Blue Fossils And tall Meaning Facts With diamond shaped

## Stegosaurus

**Stegosaurus** is one of the most easily recognized dinosaurs of the Jurassic. The plates on its back and spikes on its tail make it one of the most unique creatures of the Mesozoic (Age of Dinosaurs). 11 Jan 2024

 National Park Service (.gov)  
<https://www.nps.gov/places/stegosaurus-ungulates>

[Stegosaurus ungulates \(U.S. National Park Service\)](#)



About featured snippets Feedback

People also ask

What dinosaur has armored plates on its back?

What is the name of the dinosaur with spikes on its back?

What are plated dinosaurs called?

What kind of dinosaur has spines on its back?

Feedback

→ search engines

# Introduction: What is NLP?

English ▼

↔

German ▼

Some species, like Stegosaurus, are believed to have used the large back plates to regulate their body temperature.

×

Man geht davon aus, dass einige Arten wie Stegosaurus die großen Rückenplatten zur Regulierung ihrer Körpertemperatur genutzt haben.

🔊

📄 🔊

→ translation

# Introduction: What is NLP?

**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.

→ language models

# Introduction: Natural Language Processing

---

- Human language has many layers and challenges
- Humans can understand and produce language effortlessly
  - understanding of linguistic structure
  - world knowledge
  - ability to generalize and interpret context
- What about computers?
  - good at detecting statistical patterns
- How well can computers capture language?
- NLP: Combines the fields of computer science, AI and linguistics

# Outline

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation

# Challenges in NLP: Word Senses

## SAME WORD · DIFFERENT MEANINGS



## SAME WORD · DIFFERENT MEANINGS



## SAME WORD · DIFFERENT MEANINGS



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



# Challenges in NLP: 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*

<i>Wach Stube</i>	→ guard room
<i>Wachs Tube</i>	→ tube of wax

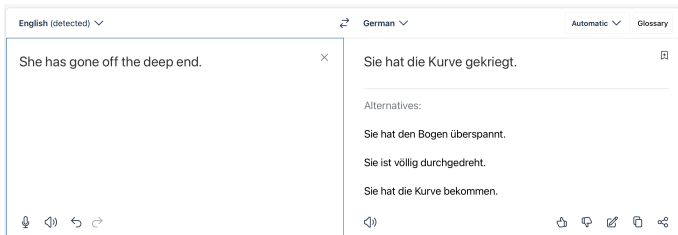
# Challenges: Non-compositional Phrases

---

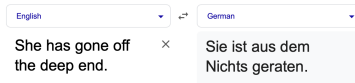
- Some words cannot be interpreted literally, but only in their context
- For example: idiomatic expressions
  - she's gone off the deep end*
  - er hat nicht mehr alle Tassen im Schrank*
- All words in the phrases have a distinct meaning that is not related to the meaning of the phrase (*crazy/verrückt*)
- Context: need to consider the entire phrase to derive the meaning
- Challenging for many NLP tasks

# Challenge: Translate Non-Compositional Phrases

DeepL:



Google Translate:



PONS:

She has gone off the deep end. | Sie ist aus dem Tiefpunkt verschwunden.

# Challenge: Translate Non-Compositional Phrases

## 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 <b>never</b> said she stole my money.	I never said she stole my money. Never.
I never <b>said</b> she stole my money.	I probably hinted, but did not say it.
I never said <b>she</b> stole my money.	It might be someone associated with her.
I never said she <b>stole</b> 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 <b>my</b> money.	She stole someone else's money.
I never said she stole my <b>money</b> .	She stole my credit card.

# Outline

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation

# Levels of Linguistics

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

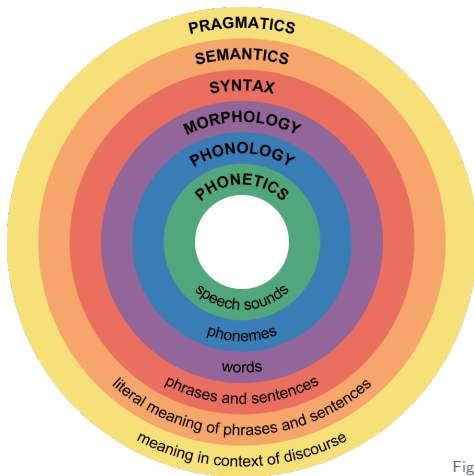


Figure from [https://commons.wikimedia.org/wiki/File:Major\\_levels\\_of\\_linguistic\\_structure.svg](https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg)

# 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

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation

# 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*<sup>1</sup>
  - Finnish: *istahtaisinkohankaan (I wonder if I should sit down for a while after all)*<sup>2</sup>

<sup>1</sup><https://www.duden.de/sprachwissen/sprachratgeber/Die-langsten-Woerter-im-Dudenkorpus>

<sup>2</sup>[https://en.wikipedia.org/wiki/Finnish\\_language](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)	<i>(to) sense</i>
duygu	<i>sensation</i>
duygusal	<i>sensitive</i>
duygusallaş(-mak)	<i>(to) become sensitive</i>
duygusallaştırıl(-mak)	<i>(to) be made sensitive</i>
duygusallaştırılmış	<i>the one who has been made sensitive</i>
duygusallaştırılamamış	<i>the one who could not have been made sensitive</i>
duygusallaştırılamamışlardan	<i>from the ones who could not have been made sensitive</i>

- 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!

# Tokenization

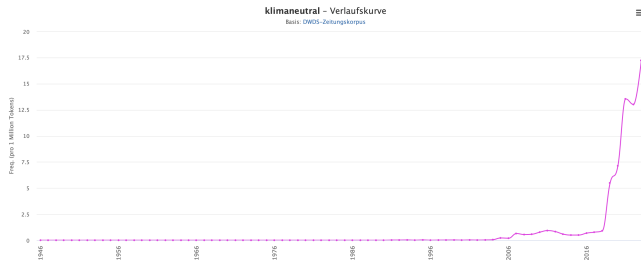
---

- For NLP tasks
  - consistent representation of the data as a sequence of tokens
  - keep the vocabulary as small as possible
- Do not blow up the vocabulary with different forms such as *house* and *house*, and *house!* and *“house”*
- Tokenization: breaking raw text into words assuming words as they appear on the surface level as tokens
- Languages with similar concepts of words than English: essentially splitting off punctuation
- Writing systems without spaces or languages with highly complex words: segmentation is more challenging

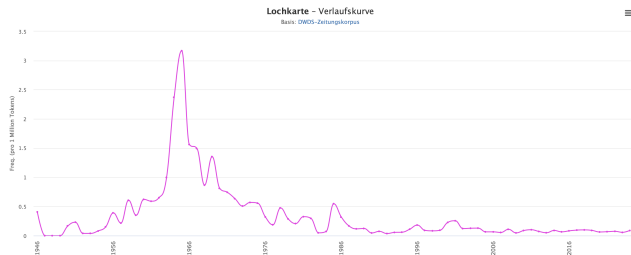
# What are the Words of a Language?

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

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



“climate neutral”



“punch card”

# Corpora and Word Distribution

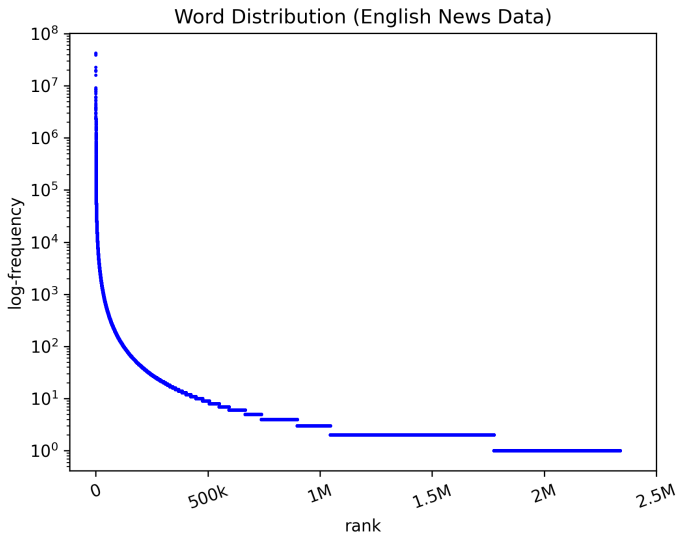
- The vocabulary of a language is fluid
- In practice: text corpus with a fixed set of words
- Continually update with new data → larger corpora
  
- English news data (33M sentences):

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	yesterday
3	yellow-beaked
3	worried
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<sub>Pref</sub> über<sub>Part</sub> setzen<sub>Verb</sub> bar<sub>Suffix\_ADJ</sub> keit<sub>Suffix\_N</sub>

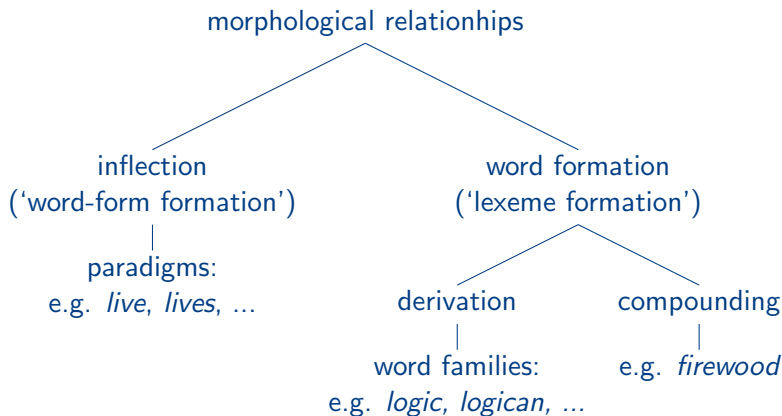
- *untranslatability* →

un<sub>Pref</sub> translate<sub>Verb</sub> able<sub>Suffix\_ADJ</sub> ity<sub>Suffix\_N</sub>

# 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



# Verbal Inflection

- **English**

*I am swim-m-ing*

- we know the meaning of (to) *swim*
- *-ing*: this event is taking place at the time of the utterance
- why the extra *m*?

- **Turkish**

*Ben yüz-üyor-um*

*I.Nom swim-Prog-1P.Sg*

- *yüz* means 'swim'
- *-üyor* corresponds to English *-ing*
- *-um* indicates the person

⇒ Inflected Turkish verb contains more information

# 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

# Morphological Complexity – Czech Nominal 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](https://en.wikipedia.org/wiki/Czech_declension)

# Outline

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech**

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation

# 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
- Part-of-Speech tagging: labeling the POS tags of words in a text
- Well-established strategy:
  - annotate a large amount of text with POS-tags
  - train a tagger on the annotated data
- No trivial task:
  - words that appear the same can occur in different functions, for example *to house* (VERB) ↔ *the house* (NOUN)
  - classify previously unseen words



# POS Tagging – Example

<b>word</b>	<b>POS</b>
When	WRB
the	DT
space	NN
shuttle	NN
was	VBD
approved	VRB
in	IN
1972	CD
,	,
NASA	NNP
officials	NN
predicted	VBD
that	IN
they	PP
would	MD
launch	VB
one	CD
every	DT
week	NN
or	CC
two	CD
.	SENT

# 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/>

# Function Words and Content Words

## Content words

⇒ open-class words

- Words with lexical content
  - Nouns → refer to entities
  - Verbs → actions
  - Adjectives → attributes of entities
  - Adverbs → attributes of actions
- Continually evolving non-finite set of words

## Function words

⇒ closed-class words

- Words with little to no lexical meaning
- Provide sentence structure: express grammatical relations (prepositions, pronouns, articles, auxiliary verbs, ...)
- Small set of words, make up a large part of the overall word count

# Outline

---

What is NLP?

Challenges in NLP

**Linguistic Concepts**

Words and Morphology

Parts of Speech

**Sentences and Syntax**

Semantics

Large Language Models ... some Notes on

Training Data

Subword Segmentation

Evaluation

# 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

# Phrases

- Phrase: meaningful unit of words grouped together
- **Noun Phrases:** words grouped around a noun (= head of the phrase)
  - a zebra
  - a cute little cat
  - the dog that bit the postman
  - a 100-year old turtle with dark green spots
- **Prepositional phrases**
  - in the supermarket
  - on Wednesday
  - on a plane to London
- **Verbal phrases**
  - read a book
  - sleeps

# Grammars: Toy Example

Alice reads a book.

Alice sleeps.

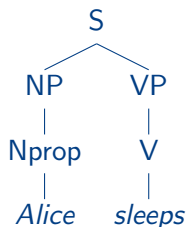
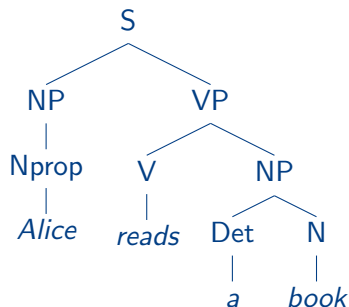
S	→	NP VP	non-terminal symbols
VP	→	V NP	
VP	→	V	
NP	→	Det N	
NP	→	Nprop	

---

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

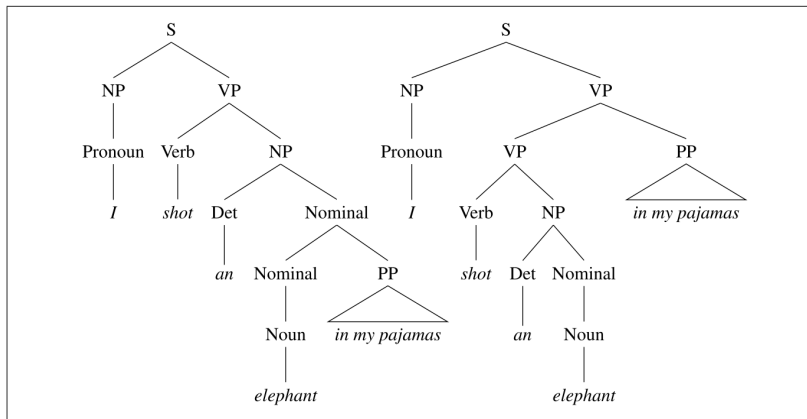


# Grammars: Toy Example



- 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 )

# Structural Ambiguities



*I shot an elephant in my pajamas.*

→ who is wearing the pajamas?

# Useful Resources: Universal Dependency Treebank

- UDP: developing cross-linguistically consistent treebank annotation for many languages
- Tree structures for English, Bulgarian, Czech and Swedish

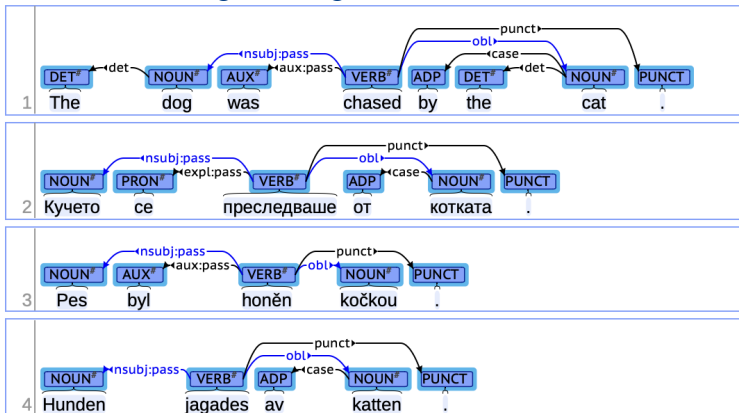


Figure from <https://universaldependencies.org/introduction.html>

# 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 (e.g., prepositions, modifiers)
  - syntactic functions (subject, object) are assigned via word order
  - for example English, Norwegian, Danish

# Syntax across Languages

- **Synthetic languages**

- grammatical information is synthesized into one word by means of (inflectional) morphology (e.g. grammatical case instead of prepositions)
- 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

**Adamlar tanıştım**  
indirect object   instrumental case suffix   verb stem   past tense suffix   indicator of subject  
I met with the man

**Adamın kitabı**  
possessor   genitive suffix   possessed noun   possessive ending  
Man's book

- **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

## Isolating

Mandarin



## Agglutinative

Tamil



## Fusional

Spanish



## Polysynthetic

Mohawk



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

# Outline

---

What is NLP?

Challenges in NLP

**Linguistic Concepts**

Words and Morphology

Parts of Speech

Sentences and Syntax

**Semantics**

Large Language Models ... some Notes on

Training Data

Subword Segmentation

Evaluation

# 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

All humans are mortal

Socrates is a human

---

Socrates is mortal

- **Pragmatics:** investigates how people use language in communication
  - "It's cold in here, isn't it?" (looks towards the open window)



# Lexical Semantics: Semantic Role Labeling

- 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

# Lexical Semantics: Word Similarity and Relatedness

- **Word similarity:**

(near) synonyms  $\leftrightarrow$  similar words

*cat* and *dog* are not synonyms, but similar words

- Example from the SimLex-999 dataset

Hill et al. (2015)

vanish	disappear	9.8
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

- **Word relatedness:** *coffee* and *cup* are not similar, but are related  
( $\rightarrow$  co-participating in the event of drinking coffee out of a cup)

# Lexical Semantics: Word Sense Disambiguation

bank <sup>1</sup>	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	“he cashed a check at the bank”, “that bank holds the mortgage on my home”
bank <sup>2</sup>	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	“they pulled the canoe up on the bank”, “he sat on the bank of the river and watched the currents”

Example from Jurafsky and Martin

# Outline

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation

# Large Language Models

---

- Idea: text contains enormous amounts of knowledge
- Pretraining 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 ...

# Large Language Models

---

- LLMs: remarkable performance on many NLP tasks due to knowledge obtained in pretraining
- Especially for tasks where text is produced
  - summarization
  - machine translation
  - question answering
  - chatbots

# Outline

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on  
Training Data

- Subword Segmentation

- Evaluation

# LLMs: General Background

---

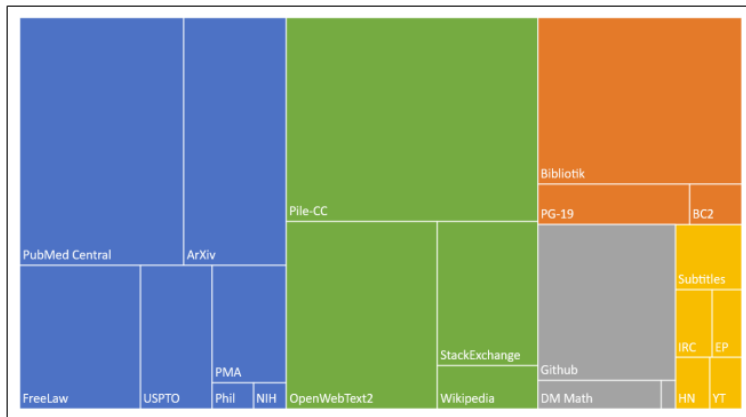
- Language models are trained on huge amounts of data, often on multilingual training data
- **No explicit linguistic information**
- Multilingual LLMs: no explicit language information
- LLMs are not trained on words, but **subwords**
  - efficiency: restricted vocabulary size
  - handle unknown words



# Training Data

- Automatically-crawled web data
- Common crawl: <https://commoncrawl.org>
  - for example the Colossal Clean Crawled Corpus (C4) Raffel et al. (2020)
  - 156 billion tokens of English
  - filtered in various ways (deduplicated, removing non-natural language like code, sentences with offensive words from a blocklist)
- Wikipedia
- Book corpora
- The Pile: 825 GB English corpus Gao et al. (2020)
- Dolma: 3 trillion tokens; web text, academic papers, code, books, encyclopedic materials, and social media Soldaini et al. (2024)

# 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. \(2020\)](#).

# Outline

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation

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

# Subword Segmentation: Example

## Segmentation from 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 normalerweise Längen von 3 bis 4 m.

# 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 (Kometen, Asteroiden und Meteoroiden) und aus unzähligen Gas- und Staubteilchen, 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 tělesa jako planety, komety, meteoroidy apod., které jsou soustředěny především v Hlavním pásu uvnitř soustavy a Kuiperově pásu na jejím okraji. Teoreticky sluneční soustavu ještě obklopuje Oortův oblak. Sluneční soustava je součástí Galaxie tradičně též nazývané Mléčná dráha.

# BPE vs. Morphological Segmentation

- Frequency-based segmentation strategies often not optimal translation and language modeling
- Linguistically guided segmentation can improve performance
  - for example Hoffmann et al. (2021), Hou et al. (2023)
  - lower perplexity and faster convergence
  - model size: smaller models trained on morphological segmentation comparable to larger models trained on BPE
- Many variants of language-specific (monolingual) LMs modeling the languages' properties

# Outline

---

What is NLP?

Challenges in NLP

Linguistic Concepts

- Words and Morphology

- Parts of Speech

- Sentences and Syntax

- Semantics

Large Language Models ... some Notes on

- Training Data

- Subword Segmentation

- Evaluation



# Benchmarks

- Standard evaluation tasks to assess performance of LLMs
- GLUE: Wang et al. (2018)  
General Language Understanding Evaluation benchmark  
contains diverse natural language understanding tasks
  - MNLI (Multi-Genre Natural Language Inference): predict whether sentence B is an entailment, contradiction, or neutral with respect to sentence A
  - ...
- SQuAD (Stanford Question Answering Dataset): Rajpurkar et al. (2016)  
for a question and a passage containing the answer,  
predict the span of the answer
- XNLI: cross-lingual natural language inference Conneau et al. (2018)  
for 15 languages (including low-resource languages)

# The End

---

- Questions?

- Next week: n-gram models

Reading: <https://web.stanford.edu/~jurafsky/slp3/3.pdf>  
(without section 3.7)

# 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
- Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze (2021). *Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words.* ACL 2021.
- Rico Sennrich, Barry Haddow, and Alexandra Birch (2016). *Neural machine translation of rare words with subword units.* ACL 2016.

# References

- Jue Hou, Anisia Katinskaia, Anh-Duc Vu, and Roman Yangarber (2023). *Effects of sub-word segmentation on performance of transformer language models*. EMNLP 2023.
- 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>