

Concepts and Applications in NLP

Introduction and Linguistic Concepts

Marion Di Marco

October 22, 2024

Outline

What is NLP?

Challenges in NLP

Linguistic Concepts

Words and Morphology

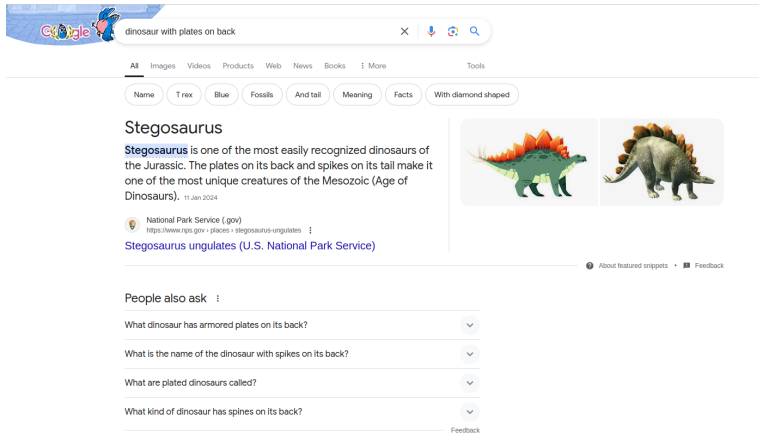
Parts of Speech

Sentences and Syntax

Semantics

Seminar: Outline and Organization

Introduction: What is NLP?



The screenshot shows a Google search interface. The search bar contains the text "dinosaur with plates on back". Below the search bar, there are tabs for "All", "Images", "Videos", "Products", "Web", "News", "Books", "More", and "Tools". A filter bar below the tabs includes buttons for "Name", "T rex", "Blue", "Fossils", "And tall", "Meaning", "Facts", and "With diamond shaped". The search results for "Stegosaurus" are displayed. The main result includes a title "Stegosaurus", a short description: "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).", a date "11 Jan 2024", and a source: "National Park Service (.gov) https://www.nps.gov/places/stegosaurus-ungulates". Below the main result, there is a section "People also ask" with four questions and dropdown arrows: "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?", and "What kind of dinosaur has spines on its back?". There are also links for "About featured snippets" and "Feedback".


Google 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 unguulates (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

- Natural Language Processing:
making human language accessible to computers

Tasks and applications

information retrieval
machine translation
question answering
paraphrasing
sentiment analysis
text generation
summarization
...

Analysis tasks

tokenization and word segmentation
part-of-speech tagging
named-entity recognition
syntactic parsing
semantic role labeling
coreference resolution
word sense disambiguation
...

- Combines the fields of computer science, AI and linguistics

NLP Tasks and Applications

- **Data**

- text corpora (general language, domain-specific, parallel, ...)
- further knowledge sources: lexicons, dictionaries, grammars, ...

Optional: linguistic analysis or annotation

- refine your raw corpus

- **Train**

- Learn statistical patterns relevant for the task

- **Apply and Evaluate**

- comparison against a baseline or benchmark
- human evaluation

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?
- How well can computers handle the challenges of language?

Outline

What is NLP?

Challenges in NLP

Linguistic Concepts

Words and Morphology

Parts of Speech

Sentences and Syntax

Semantics

Seminar: Outline and Organization

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 → how to translate?
- *a criminal lawyer*
 - is the lawyer criminal?
 - does the lawyer practice criminal law?
- *local phone compay, live fish transporter*
- Compound structure: *Mädchenhandelsschule*
 - [Mädchen [Handel Schule]]: trade school for girls
 - [Mädchen Handel] Schule]: school for girl trade
- *Wachstube*

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

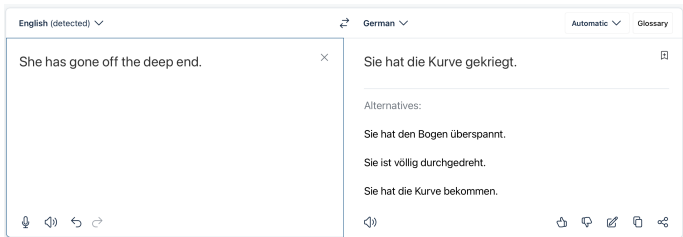
Challenges: Non-compositional Phrases

- 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

Challenges: Non-Compositional Phrases

- *to go off the deep end* seems to be difficult to translate

DeepL:



Google Translate:



PONS:

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

Challenges: 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 übergeschwappt.
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.

Outline

What is NLP?

Challenges in NLP

Linguistic Concepts

Words and Morphology

Parts of Speech

Sentences and Syntax

Semantics

Seminar: Outline and Organization

Levels of Linguistics

Linguistics: the scientific study of language and its structure

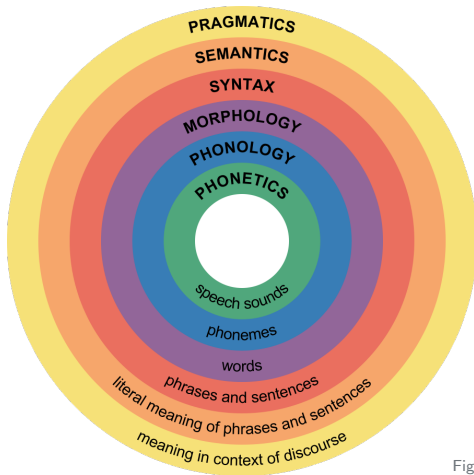


Figure from https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

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

→ Focus on written language

Outline

What is NLP?

Challenges in NLP

Linguistic Concepts

Words and Morphology

Parts of Speech

Sentences and Syntax

Semantics

Seminar: Outline and Organization

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-Woerter-im-Dudenkorpus>

²https://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>
duygusallaştırılmamış	<i>the one who could not have been made sensitive</i>
duygusallaştırılmamışlardan	<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!

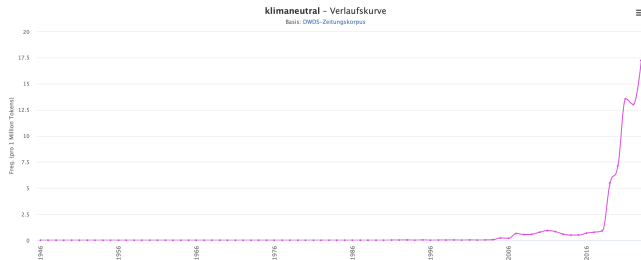
Tokenization

- 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

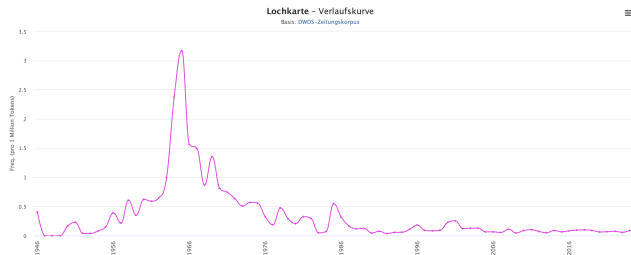
What are the Words of a Language?

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

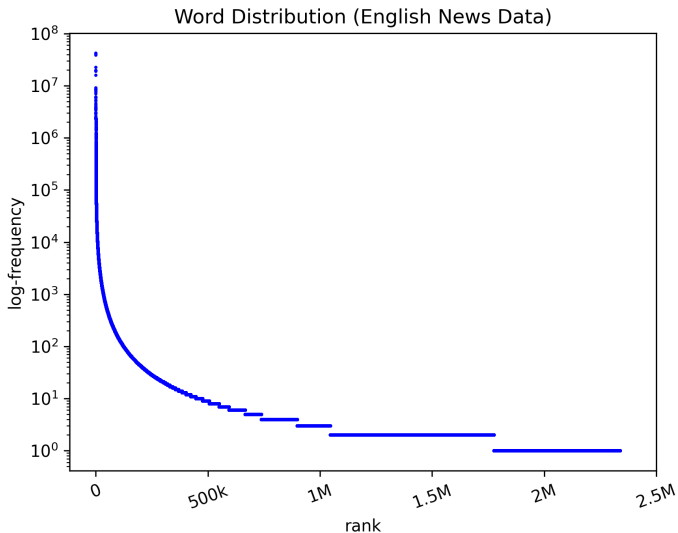
- English news data (33M sentences):

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



Some Terminology – Overview

- **Morphemes:** smallest meaningful constituents

unattainable_{Adj} → un|attain|able → un_{NegPrefix} attain_{Verb} able_{AdjSuffix}

- **Lexeme:** dictionary words

houses → house

is, was, are → (to) be

- **Word-form:** lexeme + grammatical features

read + “third person, singular, present tense” = reads

- **Paradigm:** set of word-forms belonging to a lexeme

(to) make: {make, makes, made, making}

- **Word families**

read, reader, unreadable, readability, ...

Morphology

- Morphology: studies the internal structure and composition of words
- **Inflectional morphology:**
addition of a morpheme to express grammatical categories
 - does not change the core lexical meaning of the words
number: *house* → *houses*
tense: *machen* → *machte*
- **Derivational morphology:**
forming a new word from existing words
 - changes the lexical interpretation of the word
addition of particle: *ab* + *machen* → *abmachen* ('off make': *remove*)
adjectivization: *fold*_{verb} + *-able* → *foldable*_{adj}

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
Apfel + Kuchen → Apfelkuchen (apple cake)
- More morphological variation: larger vocabulary of surface forms

Example: French Verbal Inflection

Inflection paradigm for the French verb *voir* (to see)

INDICATIF

Présent

je	vois
tu	vois
il/elle/on	voit
nous	voyons
vous	voyez
ils/elles	voient

Imparfait

je	voyais
tu	voyais
il/elle/on	voyait
nous	voyions
vous	voyiez
ils/elles	voyaient

Passé simple

je	vis
tu	vis
il/elle/on	vit
nous	vîmes
vous	vîtes
ils/elles	virent

Futur simple

je	verrai
tu	verras
il/elle/on	verra
nous	verrons
vous	verrez
ils/elles	verront

SUBJONCTIF

Présent

que	je	voie
que	tu	voies
qu'	il/elle/on	voie
que	nous	voyions
que	vous	voyiez
qu'	ils/elles	voient

Imparfait

que	je	visse
que	tu	visses
qu'	il/elle/on	vît
que	nous	visussions
que	vous	visseriez
qu'	ils/elles	visissent

CONDITIONNEL

Présent

je	verrais
tu	verrais
il/elle/on	verrait
nous	verrions
vous	verriez
ils/elles	verraient

FORMES IMPERSONNELLES

Infinitif

voir

Participe présent

voyant

Participe passé

vu(e)

- In addition: composed tenses
- In contrast: (to) see, sees, saw, seen, seeing

Morphological Complexity

- Large vocabulary → data sparsity
 - some forms only occur infrequently or even not at all
- Generally challenging for NLP applications

- Interpretation of a seen form:
 - what does the particular realization of a word mean?
- Generation of an appropriate form:
 - what should a form look like in the given context?

- Just add more training data?
 - more data certainly helps ...
 - ... but still puts morphologically rich languages at a disadvantage
- Ideally: generalization

Handling Morphological Complexity

- Simplify complex form to their lemma or stem
 - stemming or lemmatization
 - core lexical meaning remains
 - language/context specific surface realization is removed
- Different strategies for different applications
 - inflectional features (tense, plural, case, ...) might be kept
 - machine translation
 - (monolingual) language modeling
 - context vectors
- Compound splitting
 - (statistical) machine translation
- Generation: generate forms given stem and features

Outline

What is NLP?

Challenges in NLP

Linguistic Concepts

Words and Morphology

Parts of Speech

Sentences and Syntax

Semantics

Seminar: Outline and Organization

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

<u>word</u>	<u>POS</u>
When	WRB
the	DT
space	NN
shuttle	NN
was	VBD
approved	VBN
in	IN
1972	CD
,	,
NASA	NP
officials	NNS
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

- Words with lexical content
 - Nouns → refer to entities
 - Verbs → actions
 - Adjectives → attributes of entities
 - Adverbs → attributes of actions
- Open-class words

Function words

- Words with little to no lexical meaning
- Provide the structure of a sentence: express grammatical relations between content words
- For example prepositions, pronouns, articles, auxiliary verbs, ...
- Closed-class words

What does that mean for NLP applications?

- Content words:
 - continually evolving non-finite set of words
 - many existing words, with new words being introduced
 - depending on the language: further inflectional variants → morphology
- Need for large text corpora to span many topics and domains for sufficient coverage
- Function words:
 - comparatively small set of words
 - make up a large part of the overall word count
 - their interpretation is often context-dependent (for example, *that* as a determiner or relative pronoun)
 - depending in the language: different realization of linguistic concepts
→ morphology, sentence structure

Outline

What is NLP?

Challenges in NLP

Linguistic Concepts

Words and Morphology

Parts of Speech

Sentences and Syntax

Semantics

Seminar: Outline and Organization

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
 - clauses: group of words containing a subject and predicate and functioning as part of a complex sentence
 - 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

Phrases

- Clause: a sequence of words that have a subject and a verb
- Clauses fall between phrases and sentences

when it rained the kids went inside

- the sentence consists of two clauses
- the second clause can also occur as a sentence

- Alice remembered everything Bob said

Alice remembered Bob's explanation

for example: noun clauses can be replaced with a noun phrase

Two types of clauses:

- main clause or independent clause: can stand alone
- subordinate clause: relies on a main clause (i.e. is dependent on it)

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

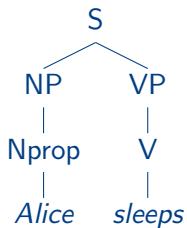
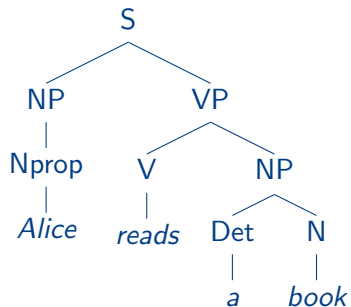
V → sleeps | reads terminal symbols

Det → a

N → book

Nprop → Alice

Grammars: Toy Example



Grammars and Parse Trees

- Different grammar formalism to express the structure of a sentence (for example, phrase structure grammar , dependency structures, ...)
- Parse trees: illustrate the grammatical structure of a sentence
- Phrase-structure grammar
 - Models the constituents in a sentence and how they are composed of other constituents and words
 - The inner nodes are non-terminal symbols (grammatical/lexical categories), and the leaves are terminal symbols (words)
- Dependency structures: display relationship (dependencies) between words
 - one word is the head of the sentence, dependent on a notional ROOT (mostly the verb of the main clause)
 - all other words are dependent on another word
 - all nodes are terminals

Dependency Grammar: Example

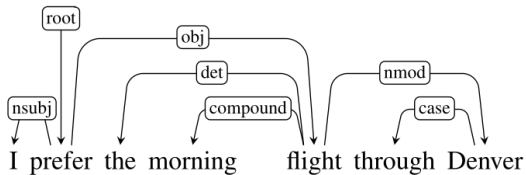


Figure from Jurafsky and Martin

Dependency vs. Constituency Trees: Example

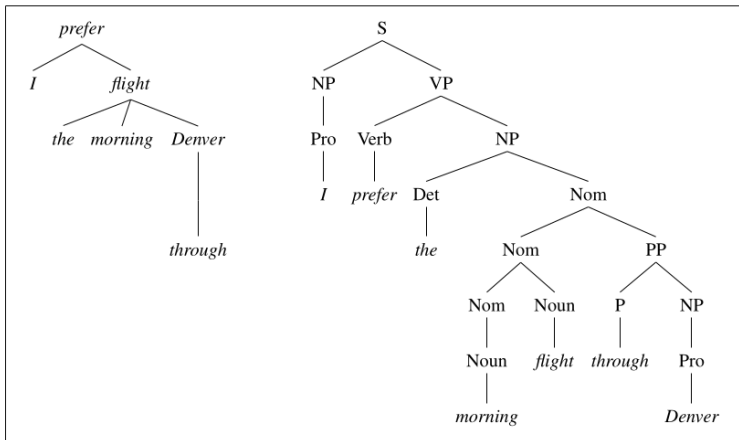
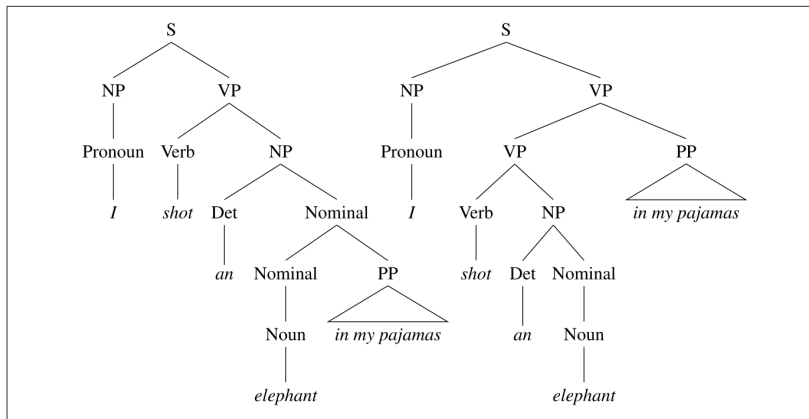


Figure 18.1 Dependency and constituent analyses for *I prefer the morning flight through Denver*.

Figure from Jurafsky and Martin

Structural Ambiguities



I shot an elephant in my pajamas.

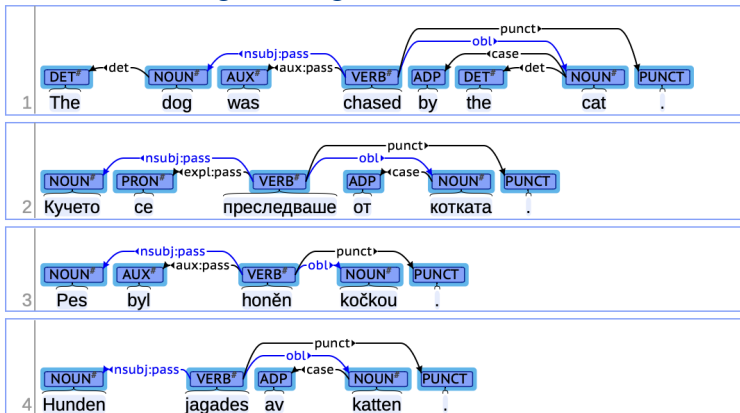
→ who is wearing the pajamas?

Syntax across Languages

- Linguistic concepts and processes are realized differently
- **Analytic languages**
 - 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
- **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
- Often no clear distinction: languages can have features of both groups

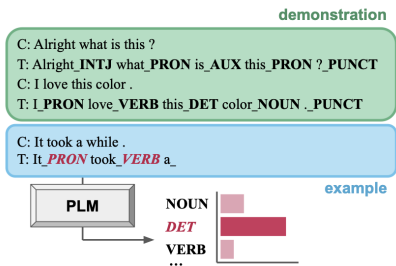
Universal Dependency Treebank

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



Linguistic Structure in Large Language Models

- Language models perform very well at many language tasks
- To what extent can these abilities be attributed to generalizable linguistic understanding vs. surface-level lexical patterns?
- Can we obtain linguistic structure from LMs?



Prompting Language Models for Linguistic Structure.
Blevins et al. (2023)

Figure 1: Sequence tagging via structured prompting. Each predicted label is appended to the context along with the next word to iteratively tag the full sentence.

Outline

What is NLP?

Challenges in NLP

Linguistic Concepts

Words and Morphology

Parts of Speech

Sentences and Syntax

Semantics

Seminar: Outline and Organization

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: Semantic Role Labeling

- Semantic roles help generalize different surface realizations

(20.3) *John broke the window.*

AGENT THEME

(20.4) *John broke the window with a rock.*

AGENT THEME INSTRUMENT

(20.5) *The rock broke the window.*

INSTRUMENT THEME

(20.6) *The window broke.*

THEME

(20.7) *The window was broken by John.*

THEME AGENT

- Selectional restrictions:
preferences that predicates express about their arguments for example
the theme of *eat* is generally something edible.

Lexical Semantics: Word Similarity and Relatedness

- **Word similarity:**

(near) synonyms \leftrightarrow similar words

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

- Similarity is useful for many semantic tasks for example question answering, paraphrasing and summarization

- 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 (i.e. they share no features), but are related (\rightarrow co-participating in the event of drinking coffee out of a cup)

Lexical Semantics: Word Sense Disambiguation

bank ¹	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 ²	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

Seminar: Outline and Organization

Seminar Outline

- Lectures
- Homework assignments

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 Atamanab, Matteo Negrib, Marco Turchib, 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