

# 1. NLP

## Počítačové zpracování přirozeného jazyka

Jakub Machura

Masarykova univerzita  
Filozofická fakulta  
machura@phil.muni.cz

# Organizace předmětu

---

<https://is.muni.cz/auth/predmet/phil/podzim2020/PLIN034>

# NLP = Natural Language Processing

---

# NLP = Natural Language Processing

---

Co je NLP?

# NLP = Natural Language Processing

---

Co je NLP?

mluvené slovo × strojově čitelný text

# NLP = Natural Language Processing

---

Co je NLP?

mluvené slovo × strojově čitelný text

analýza × syntéza jazyka

# NLP

---

Kam zapadá syntax a syntaktická analýza?

# Dílčí úkoly analýzy jazyka

---



# Dílčí úkoly analýzy jazyka

---

## Tokenizace

# Dílčí úkoly analýzy jazyka

---

## Tokenizace

*„Chcete-li mi to dát, neváhejte!“*

# Tokenizace

---

*„Chcete-li mi to dát, neváhejte!“*

”

Chcete

-

li

mi

to

dát

,

neváhejte

!

“

# Tokenizace

---

*ohlas*

# Tokenizace

---

## *ohlas*

- imperativ slovesa *ohlásit*
- nom./akuz. substantiva *ohlas*
- 2. os. sg. fem. minulého času slovesa *ohnout*

# Větná segmentace

---

# Větná segmentace

---

- explicitně vyznačený začátek i konec věty

# Větná segmentace

---

- explicitně vyznačený začátek i konec věty

např. XML:     <s>     </s>



# Větná segmentace

---

- explicitně vyznačený začátek i konec věty

např. XML:      <s>      </s>

# Větná segmentace

---

Caesar byl zavražděn r. 43 př. Kr. Řím byl tehdy na pokraji převratu.

# Větná segmentace

---

Caesar byl zavražděn r. 43 př. Kr. Řím byl tehdy na pokraji převratu.

# Větná segmentace

---

Caesar byl zavražděn r. 43 př. Kr. **Ř**ím byl tehdy na pokraji převratu.

**Jak to vyřešit?**

# Větná segmentace

---

Caesar byl zavražděn r. 43 př. Kr. **Ř**ím byl tehdy na pokraji převratu.

Jak to vyřešit?

**Další problémy?**

# Morfologická analýza

---

# Morfologická analýza

---

lemma, lemmatizace

# Morfologická analýza

---

## lemma, lemmatizace

Na	na
vyzvání	vyzvání ( <del>vyzváněť</del> )
svého	svůj
předsedy	předseda
jsme	být
odešli	odejít ( <del>odeslat</del> )
.	.



# Morfologická analýza

---

lemma, lemmatizace

tag, tagging

# Morfologická analýza

---

lemma, lemmatizace

tag, tagging, tagger

desambiguace

# Tagging

---

- většinou založena na statistických modelech, někdy kombinováno s pravidly

# Tagging

---

- většinou založena na statistických modelech, někdy kombinováno s pravidly
- ruční anotace

# Tagging

---

- většinou založena na statistických modelech, někdy kombinováno s pravidly
- ruční anotace
- statistika četnosti značek

# Tagging

---

- většinou založena na statistických modelech, někdy kombinováno s pravidly
- ruční anotace
- statistika četnosti značek
- „natrénování“ taggeru

# Desambiguace

---

stochastická/statistická

# Desambiguace

---

stochastická/statistická

**založená na ling. pravidlech**



# Desambiguace

---

stochastická/statistická

založená na ling. pravidlech

**hybridní**

# Desambiguace

---

Syntaktická desambiguace

# Desambiguace

---

## Syntaktická desambiguace

*František hrál v altánu šachy se svým ruským přítelem.*

# Desambiguace

---

## Syntaktická desambiguace

*František hrál v altánu šachy se svým ruským přítelem.*

# Desambiguace

---

Sémantická desambiguace

# Desambiguace

---

Sémantická desambiguace

*využívat zařízení*

# Desambiguace

---

Sémantická desambiguace

*využívat zařízení*

*dělat chyby ve skloňování*

**Parsing = Syntaktická analýza**

---



# Parsing

---

Cíle:

- „porozumět“ gramatice př. jaz.
- odhalit povrchovou strukturu  
(větný rozbor)

# Parsing

---

Výsledky:

- orientované grafy (tzv. stromy)

závislostní × složkový

# Parsing

---

## Překážky:

- pro čj bohatá morfologie a rel. volný slovosled

# Parsing

---

## Překážky:

- pro čj bohatá morfologie a rel. volný slovosled
- **velké množství teoretických východisek**

# Parsing

---

## Překážky:

- pro čj bohatá morfologie a rel. volný slovosled
- velké množství teoretických východisek
- **subjektivita syntaxe**

# Parsing

---

## Překážky:

- pro čj bohatá morfologie a rel. volný slovosled
- velké množství teoretických východisek
- **subjektivita syntaxe**

*Faxu škodí **především** přetížené telefonní linky.*

# Parsing

---

Víceznačnost:

# Parsing

---

Víceznačnost:

1. Předložkové fráze (PP)



# Parsing

---

Víceznačnost:

1. Předložkové fráze (PP)

*Charles talked about cooking with Britney Spears.*

# 1. Předložkové fráze (PP)

---

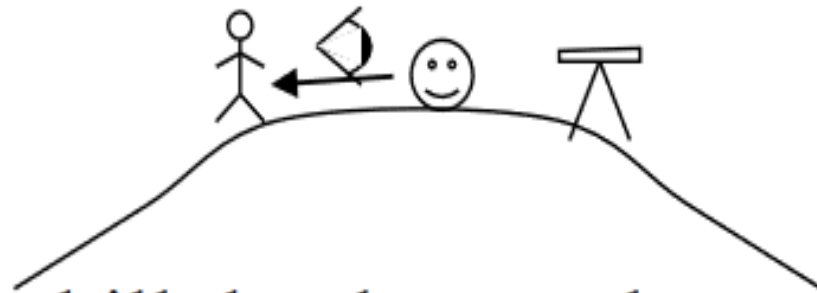
*I saw the man on the hill with the telescope.*



# 1. Předložkové fráze (PP)

---

*I saw the man on the hill with the telescope.*

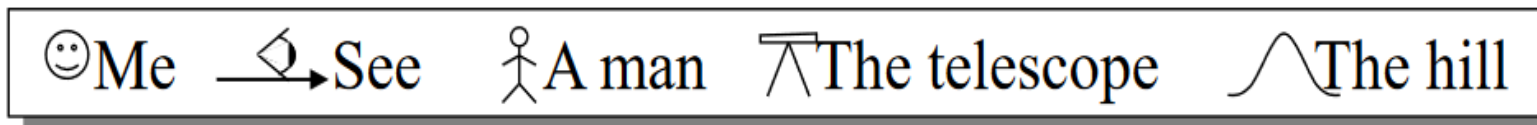


“I was on the hill that has a telescope  
when I saw a man.”

# 1. Předložkové fráze (PP)

---

*I saw the man on the hill with the telescope.*



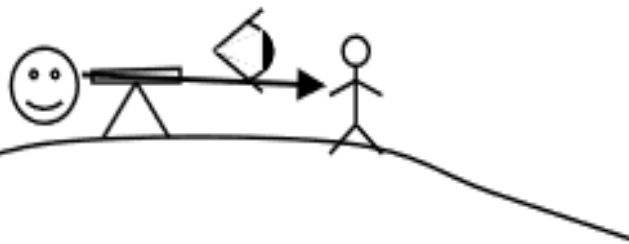
“I saw a man who was on the hill that has a telescope on it.”

# 1. Předložkové fráze (PP)

---

*I saw the man on the hill with the telescope.*

☺ Me → See    👤 A man    🔭 The telescope    ~ The hill

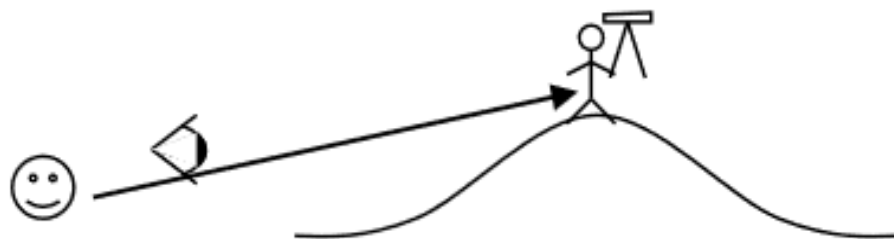


“I was on the hill when I used the telescope to see a man.”

# 1. Předložkové fráze (PP)

*I saw the man on the hill with the telescope.*

☺ Me → See    👤 A man    🔭 The telescope    ~ The hill

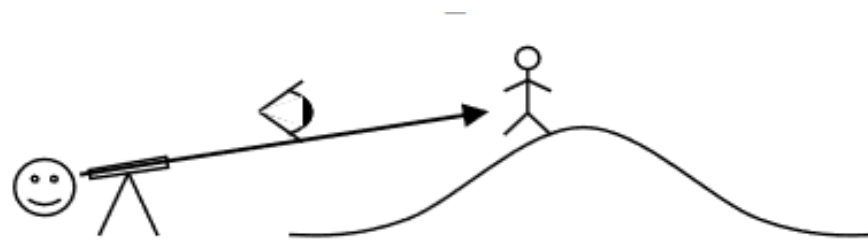
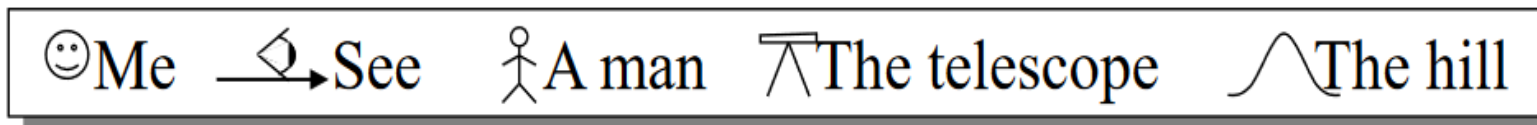


“I saw a man who was on a hill and who had a telescope.”

# 1. Předložkové fráze (PP)

---

*I saw the man on the hill with the telescope.*



“Using a telescope, I saw a man who was on a hill.”

## 2. Elipsa (gap)

---

*Marie má ráda fyziku, ale nesnáší chemii.*

*[Mary likes Physics but hates Chemistry.]*



## 2. Elipsa (gap)

---

*Marie má ráda fyziku, ale nesnáší chemii.*

*[Mary likes Physics but hates Chemistry.]*

### 3. Koordinační konstrukce

---

*Small boys and girls are playing.*

*Dřevěná vrata a okna natřel nabílo.*

## 4. Slovnědruhová homonymie

---

*She ran up a large bill.*

*She ran up a large hill.*

## 4. Slovnědruhová homonymie

---

*She ran up a large bill.* [částici]

*She ran up a large hill.* [předložka]

## 4. Slovnědruhov<sup>á</sup> homonymie

---

*She ran up a large bill.* [částici]

*She ran up a large hill.* [předložka]

*Umyl se úplně celý.*

*Umyl se žínkou nádobí.*

## 4. Slovnědruhová homonymie

---

*She ran up a large bill.* [částici]

*She ran up a large hill.* [předložka]

*Umyl se úplně celý.* [zvratné zájmeno]

*Umyl se žínkou nádobí.* [předložka]

## 4. Slovnědruhová homonymie

---

*Frightening kids can cause troubles.*

[gerundium vs. adjektivum]

## 4. Slovnědruhová homonymie

---

*Frightening kids can cause troubles.*

[gerundium vs. adjektivum]

Zdraví nemocnému nevěří.

Zdraví si musíme chránit.

Zdraví vás z Krušných hor.



# Základní termíny

---

**slovo**

# Základní termíny

---

slovo

**autosémantika, synsémantika**

# Základní termíny

---

slovo

autosémantika, synsémantika

**fráze, idiomy**

# Základní termíny

---

**lexikální symbol, lexikální kategorie** (lexical category)

# Základní termíny

---

## **lexikální symbol, lexikální kategorie** (lexical category)

- tzv. preterminál, speciální neterminál gramatiky, který se přímo přepisuje na terminálový řetězec znaků, tj. pravidla tvaru  $X \rightarrow w$

# Základní termíny

## lexikální symbol, lexikální kategorie (lexical category)

- tzv. preterminál, speciální neterminál gramatiky, který se přímo přepisuje na terminálový řetězec znaků, tj. pravidla tvaru  $X \rightarrow w$

N	→	pes		člověk		dům ...
V	→	nese		chodit		psal ...
ADJ	→	...				
PREP	→	...				
ADV	→	...				

# Základní termíny

---

## **frázová kategorie** (phrasal category)

- neterminální symbol gramatiky, který nevyjadřuje lexikální kategorii

# Základní termíny

---

## frázová kategorie (phrasal category)

- neterminální symbol gramatiky, který nevyjadřuje lexikální kategorii

ADJP	→	ADJP	ADJ
NP	→	ADJP	N
VP	→	V	NP
S	→	NP	VP



# Základní termíny

---

## **složka (konstituent, fráze)**

- lexikální nebo frázová kategorie

*Osvobození hrdinnou Sovětskou armádou jsme oslavili v letošním roce obzvláště důstojně.*

NP: Sovětskou armádou

NP: hrdinnou Sovětskou armádou

NP: Osvobození hrdinnou Sovětskou armádou

VP: jsme oslavili

PP: v letošním roce

AdvP: obzvláště důstojně

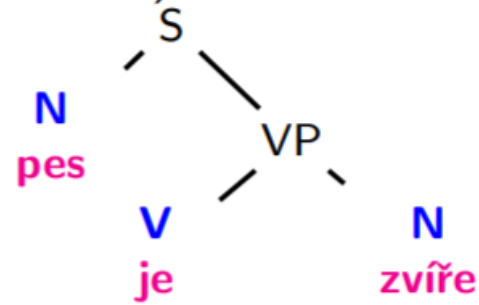
S: Osvobození hrdinnou Sovětskou armádou  
jsme oslavili v letošním roce obzvláště  
důstojně

# Základní termíny

## větná struktura

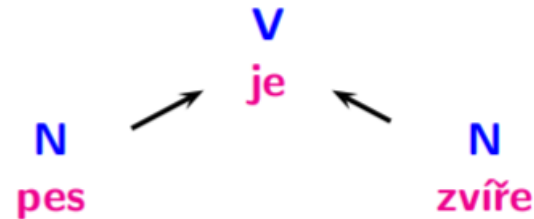
- **povrchová struktura** (*surface structure*)

derivační/složkový strom jako  
výsledek bezkontextové (CF)  
analýzy



- **závislostní struktura** (*dependency structure*)

zobrazuje závislosti mezi  
větnými členy



- **hloubková struktura** (*deep structure*) – sémantická interpretace fráze. Popisuje **role větných členů** (agens, patiens, donor, cause, ...)

# Základní termíny

---

**klauze (clause)**

# Literatura

---

Nový encyklopedický slovník češtiny online:

<https://www.czechency.org/>

hesla: Počítačové zpracování přirozeného  
jazyka, Tokenizace, Větná segmentace,  
Morfologická analýza, Lemmatizace,  
Desambiguace, Tagger, Parsing, Složka,  
Klauze