

# Zpracování přirozeného jazyka

Aleš Horák

E-mail: [hales@fi.muni.cz](mailto:hales@fi.muni.cz)  
<http://nlp.fi.muni.cz/uui/>

Obsah:

- ▶ Komunikace
- ▶ Gramatiky
- ▶ Analýza přirozeného jazyka
- ▶ PA026 – Projekt z umělé inteligence

## Přirozený jazyk – prostředek komunikace

**komunikace** = cílená výměna informace pomocí produkce a vnímání (sdílených) **pokynů**

- zvířata – až stovky pokynů (šimpanz, delfín, ...)
- člověk – potenciálně neomezené množství, díky **přirozenému jazyku**

2 náhledy na **přirozený jazyk**:

- ▶ **klasický (před 1953)** – jazyk se skládá z vět, které jsou buď pravdivé nebo nepravdivé (srovnej s logikou)
- ▶ **moderní (po 1953)** – užití jazyka je jedna z možných **akcí**  
Wittgenstein (1953) **Philosophical Investigations**  
Searle (1969) **Speech Acts**

Turingův test založen na jazyku  $\Leftarrow$  **jazyk** je pevně spojen s **myšlením**  
**komunikace** se tvoří pomocí **řečových aktů** (*speech acts*) jako jeden z typů agentových akcí

**cíl komunikace** – **změnit** akce ostatních agentů

# Řečové akty

## KOMUNIKAČNÍ SITUACE

Mluvčí (*speaker*) → Promluva (*utterance*) → Posluchač (*hearer*)

řečové akty směřují k naplnění cílů mluvčího:

- informovat (inform) “Před tebou je jáma.”
- ptát se (query) “Vidíš zlato?”
- přikázat/žádat (command/request) “Zvedni to.”
- slíbit/svěřit se s plánem (promise, commit to plan) “Rozdělím se s tebou o zlato.”
- potvrdit (acknowledge) “OK”

plánování řečových aktů vyžaduje znalosti:

- komunikační situace
- sémantiky a syntaxe (sdílených konvencí)
- informace o Posluchači – cíle, znalosti, rozumnost

## Komunikační fáze (při informování)

průběh promluvy je možné rozložit na fáze:

- **záměr** (intention)  $M$  chce informovat  $P_o$ , že  $P_r$
- **generování** (generation)  $M$  vybírá slova  $W$  pro vyjádření  $P_r$
- **syntéza** (synthesis)  $M$  říká slova  $W$
  
- **vnímání** (perception)  $P_o$  vnímá  $W'$
- **analýza** (analysis)  $P_o$  odvozuje možné významy  $P_{r_1}, \dots, P_{r_n}$
- **zjednoznačnění** (disambiguation)  $P_o$  vybírá zamýšlený význam  $P_{r_i}$
- **zahrnutí** (incorporation)  $P_o$  zahrne  $P_{r_i}$  do své báze znalostí

Může přitom vzniknout chyba?

- neupřímnost ( $P_o$  nevěří  $P_r$ )
- víceznačnost promluvy ( $P_o$  zvolí špatné  $P_{r_i}$ )
- různé pochopení aktuální situace (zamýšlený význam mezi  $P_{r_i}$  není)

# Komunikační fáze – příklad

záměr

Vědět( $P_0$ ,  
 $\neg Na\_živu(Wumpus_1, S_3)$ )

generování

“Wumpus je mrtvý.”

syntéza

MLUVČÍ

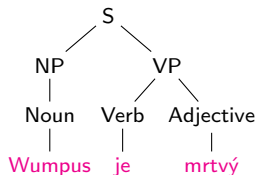
[w u m p u s j e m r t v ý:]

vnímání

analýza

zjednoznačnění POSLUCHAČ

$\neg Na\_živu(Wumpus_1, S_3)$

syntaktická  
analýza:

“Wumpus  
je mrtvý.”

zahrnutí

Tell(KB,  
 $\neg Na\_živu(Wumpus_1, S_3)$ )

šémantická  
interpretace:  $\neg Na\_živu(Wumpus, Ted)$   
 $Unavený(Wumpus, Ted)$

pragmatická  
interpretace:  $\neg Na\_živu(Wumpus_1, S_3)$   
 $Unavený(Wumpus_1, S_3)$

# Gramatiky

zvířata používají místo vět izolované symboly  $\Rightarrow$  omezená sada komunikovatelných situací  $\rightarrow$  žádná generativní kapacita

**gramatika** specifikuje skladební strukturu složených pokynů – definuje **formální jazyk** pokynů

**formální jazyk** = množina **řetězců** (vět) **terminálních symbolů** (slov)

2 náhledy na vztah věty a gramatiky:

- $S$  je správný řetězec/věta z jazyka  $\Leftrightarrow S$  je **analyzovatelný** danou gramatikou
- příslušná gramatika **generuje**  $S$   $\Leftrightarrow S$  je správný řetězec/věta z jazyka

gramatika je zadána jako množina **přepisovacích pravidel**

$$S \rightarrow NP \ VP$$

$$Pronoun \rightarrow \textit{já} \mid \textit{ty} \mid \textit{on} \mid \dots$$

v tomto příkladu:  $S$                       **větný symbol** – kořenový symbol gramatiky  
 $NP, VP$                       **neterminály**  
 $\textit{já}, \textit{ty}, \dots$                       **terminály**

# Typy gramatik

- ▶ regulární (regular)      **neterminál** → **terminál**[neterminál]

$$S \rightarrow aS$$

$$S \rightarrow b$$

ekvivalentní síle **konečných automatů**, neumí  $a^n b^n$

- ▶ bezkontextové (context-free)      **neterminál** → **cokoliv**

$$S \rightarrow aSb$$

ekvivalentní síle **zásobníkových automatů**, umí  $a^n b^n$ , neumí  $a^n b^n c^n$

- ▶ kontextové (context-sensitive) – víc termů na levé straně (*kontext* neterminálu)

$$\underline{ASB} \rightarrow \underline{AAaBB}$$

umí  $a^n b^n c^n$

- ▶ rekurzivně vyčíslitelné (recursively enumerable) – bez omezení  
ekvivalentní síle **Turingova stroje**

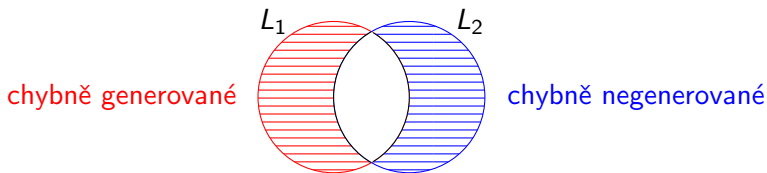
**přirozený jazyk** byl dlouho pokládán za bezkontextový → nyní prokázáno,

že obsahuje **kontextový jazyk**

## Přesnost a pokrytí gramatiky

u složitějších jazyků (např. přirozených)

→ jazyk  $L_1$  (generovaný gramatikou) se liší od zamýšleného jazyka  $L_2$



kvalita gramatiky:

- **pokrytí** – procento vět jazyka  $L_2$  generovatelných gramatikou ( $|L_1 \cap L_2|/|L_2|$ )
- **přesnost** – procento generovaných vět, které jsou správné věty jazyka  $L_2$  ( $|L_1 \cap L_2|/|L_1|$ )

**tvorba gramatiky** ... postupný proces zvyšování pokrytí a přesnosti gramatiky přirozených jazyků – velmi rozsáhlé a přesto většinou nepopisují plně ani angličtinu ☹



## DC gramatiky – gramatiky uspořádaných klauzulí

### Gramatiky uspořádaných klauzulí:

- ▶ *Definite-Clause Grammars*, DCG
- ▶ významná aplikace Prologu – *syntaktická analýza*
- ▶ DCG jsou rozšířením bezkontextových gramatik (CFG)
- ▶ jejich implementace využívá *rozdílových seznamů*

# DC gramatika – příklad 1

gramatika vět typu “The young boy sings a song.”

*% 1. část -- pravidla*

sentence --> noun\_phrase, verb\_phrase.

noun\_phrase --> determiner, noun\_phrase2.

noun\_phrase --> noun\_phrase2.

noun\_phrase2 --> adjective, noun\_phrase2.

noun\_phrase2 --> noun.

verb\_phrase --> verb.

verb\_phrase --> verb, noun\_phrase.

*% 2. část -- lexikon*

determiner --> [the].                      noun --> [boy].

determiner --> [a].                         noun --> [song].

verb --> [sings].                         adjective --> [young].

## Podobnosti mezi DCG a CFG

### Formální podobnosti mezi DCG a CFG:

CFG:

- ▶ pravidla tvaru  $x \rightarrow y$ , kde  $x \in N$  je neterminál a  $y \in (N \cup T)^*$  je konečná posloupnost terminálů a neterminálů

DCG:

- ▶ pravidla tvaru  $\langle \mathbf{hlava} \rangle \dashrightarrow \langle \mathbf{tělo} \rangle$ , kde  $\langle \mathbf{hlava} \rangle$  je opět neterminál a  $\langle \mathbf{tělo} \rangle$  je opět konečná posloupnost terminálů a neterminálů
- ▶ pravidlo  $\langle \mathbf{hlava} \rangle \dashrightarrow \langle \mathbf{tělo} \rangle$  znamená, že jedním z možných tvarů  $\langle \mathbf{hlavy} \rangle$  je  $\mathbf{tělo}$ , neboli:  
 $\langle \mathbf{hlavu} \rangle$  je možné vyjádřit jako  $\langle \mathbf{tělo} \rangle$

## Rozdíly a rozšíření DCG oproti CFG

### DCG:

1. **Neterminál** může být téměř libovolný term, kromě *seznamu*, *proměnné* a *čísla*.
2. **Terminál** může být libovolný term, s tím, že terminály a posloupnosti terminálů uzavíráme do hranatých závorek – jako **seznamy**.
3. Pravá strana pravidla může obsahovat **dodatečné podmínky** v podobě prologovských podcílů. Tyto podmínky uzavíráme do složených závorek.
4. Levá strana pravidla může dokonce vypadat i tak, že neterminál je následován posloupností terminálů.
5. Tělo pravidla smí obsahovat řez.

## Analýza v Prologu pomocí append

- ▶ věta = seznam slov **[the,young,boy,sings,a,song]**
- ▶ **pravidlová část** – neterminál chápeme jako unární predikát, jehož argumentem je ta větná složka, kterou daný neterminál popisuje

```
sentence(S) :- append(NP,VP,S),
                noun_phrase(NP), verb_phrase(VP).
```

...

- ▶ **slovníková část, lexikon** – reprezentujeme pomocí faktů:

```
determiner([the]).          noun([boy]).
determiner([a]).           ...
```

## Efektivněji – rozdílové seznamy

přepis gramatiky do Prologu pomocí **rozdílových seznamů**:

```
sentence(S,S0) :- noun_phrase(S,S1), verb_phrase(S1,S0).
```

```
noun_phrase(S,S0) :- determiner(S,S1), noun_phrase2(S1,S0).
```

```
noun_phrase(S,S0) :- noun_phrase2(S,S0).
```

```
noun_phrase2(S,S0) :- adjective(S,S1), noun_phrase2(S1,S0).
```

```
noun_phrase2(S,S0) :- noun(S,S0).
```

```
verb_phrase(S,S0) :- verb(S,S0).
```

```
verb_phrase(S,S0) :- verb(S,S1), noun_phrase(S1,S0).
```

```
determiner([the|S],S).
```

```
noun([boy|S],S).
```

```
determiner([a|S],S).
```

```
noun([song|S],S).
```

```
verb([sings|S],S).
```

```
adjective([young|S],S).
```

```
?- sentence([the,young,boy,sings,a,song], []).
```

Yes

## Lexikon pro agenta ve Wumpusově jeskyni

Gramatika přímo na slovech je příliš rozsáhlá. Řešením je rozdělení slov do **kategorií**:

podst. jméno:	<i>Noun</i>	→	zápach   vánek   třpyt   nic   wumpuse   jáma   zlato   ...
sloveso:	<i>Verb</i>	→	jsem   je   vidím   cítím   působí   zapáchá   jdu   ...
příd. jméno:	<i>Adjective</i>	→	levý   pravý   východní   jižní   ...
příslovce:	<i>Adverb</i>	→	tady   tam   blízko   vpředu   vpravo   vlevo   východně   jižně   vzadu   ...
vl. jméno:	<i>Name</i>	→	Petr   Honza   Brno   FI MU   ...
zájmeno:	<i>Pronoun</i>	→	já   ty   mě   toho   ten   ta ...
předložka:	<i>Preposition</i>	→	do   v   na   u   ...
spojka:	<i>Conjunction</i>	→	a   nebo   ale   ...
číslice:	<i>Digit</i>	→	0   1   2   3   4   5   6   7   8   9

kategorie můžeme dělit na **otevřené** (vyvíjející se) a **uzavřené** (stálé)

## Morfologická analýza

- ▶ v češtině u lexikonu nestačí prostý výčet tvarů – je nutná **morfologická analýza** (morfologie=tvarosloví)
- ▶ skloňovaná a časovaná slova se rozkládají na **segmenty**

**pří-lež-it-ost-n-ými:**

*pří* – prefix; *lež* – kořen; *it, ost, n* – suffixy; *ými* – koncovka

- ▶ **základní tvar** slova (*lemma*), podle koncovky se určují **gramatické kategorie**  
*% slovník základních gramatických kategorií -- pád, číslo, rod*  
*% adj(+Slovo, +Lemma, +Pád, +Císlo, +Rod)*  
 adj(chytrý, chytrý, 1, sg, mz).      adj(chytrého, chytrý, 2, sg, mz).  
 adj(chytří, chytrý, 1, pl, mz).
- ▶ reálná morfologická analýza ČJ – program MAJKA na FI MU  
<http://nlp.fi.muni.cz/projekty/wwwajka/>

```
ajka>nejneuvěřitelněji
<s> nej-ne=uvěřiteln==ěji= (1022)
<l>uvěřitelně
<c>k6xMeNd3
```

```
ajka>hnát
<s> ==hná=t= (618)
<l>hnát
<c>k5eAmFaI
<s> =hnát=== (1030)
<l>hnát
<c>k1gInSc1,k1gInSc4
```

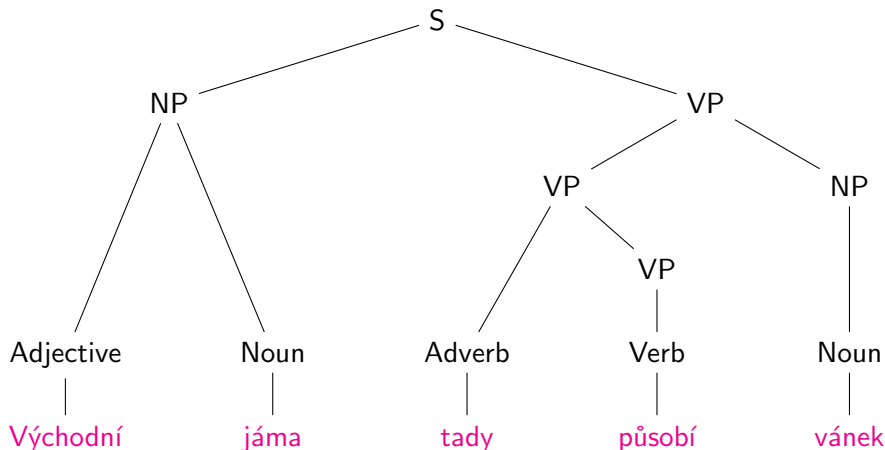


# Gramatická pravidla pro agenta ve Wumpusově jeskyni

<i>S</i>	→	<i>NP VP</i>	%	já + cítím vánek
		<i>S Conjunction S</i>	%	já cítím vánek + a + já jdu
			%	na východ
<i>NP</i>	→	<i>Pronoun</i>	%	já
		<i>Noun</i>	%	jáma
		<i>Adjective Noun</i>	%	levá jáma
		<i>Pronoun NP</i>	%	toho + wumpuse
		<i>Noun Digit ',' Digit</i>	%	pole + 3,4
		<i>NP PP</i>	%	jáma + na východě
		<i>NP RelClause</i>	%	toho wumpuse + ,který
			%	zapáchá
<i>VP</i>	→	<i>Verb</i>	%	zapáchá
		<i>VP NP</i>	%	cítím + vánek
		<i>VP Adjective</i>	%	je + třpytivý
		<i>VP PP</i>	%	jdu + na východ
		<i>VP Adverb   Adverb VP</i>	%	jdu + dopředu
<i>PP</i>	→	<i>Preposition NP</i>	%	na + východ
<i>RelClause</i>	→	<i>',' který' VP</i>	%	,který + zapáchá

## Syntaktický strom

**syntaktický strom** vzniká během **syntaktické analýzy** a dává **záznam** o jejím průběhu:



## Konstrukce derivačního stromu

Neterminály opatříme argumentem:

`sentence(sentence(NP,VP)) --> noun_phrase(NP), verb_phrase(VP).`

Převod do podoby klauzulí:

`sentence(sentence(NP,VP),S,S0) :- noun_phrase(NP,S,S1), verb_phrase(VP,S1,S0).`

# DC gramatika s konstrukcí stromu analýzy

$\text{sentence}(s(N,V)) \rightarrow \text{noun\_phrase}(N), \text{verb\_phrase}(V).$   
 $\text{noun\_phrase}(np(D,N)) \rightarrow \text{determiner}(D), \text{noun\_phrase2}(N).$   
 $\text{noun\_phrase}(np(N)) \rightarrow \text{noun\_phrase2}(N).$   
 $\text{noun\_phrase2}(np2(A,N)) \rightarrow \text{adjective}(A), \text{noun\_phrase2}(N).$   
 $\text{noun\_phrase2}(np2(N)) \rightarrow \text{noun}(N).$   
 $\text{verb\_phrase}(vp(V)) \rightarrow \text{verb}(V).$   
 $\text{verb\_phrase}(vp(V,N)) \rightarrow \text{verb}(V), \text{noun\_phrase}(N).$

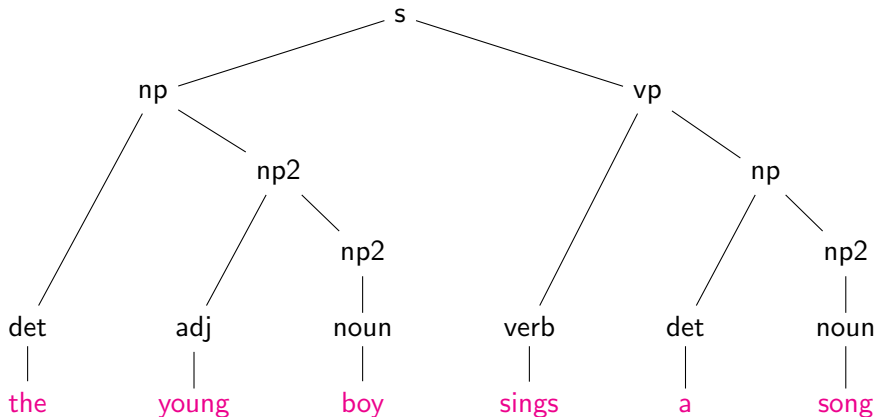
$\text{determiner}(\text{det}(\text{the})) \rightarrow [\text{the}].$   
 $\text{determiner}(\text{det}(\text{a})) \rightarrow [\text{a}].$   
 $\text{adjective}(\text{adj}(\text{young})) \rightarrow [\text{young}].$   
 $\text{noun}(\text{noun}(\text{boy})) \rightarrow [\text{boy}].$   
 $\text{noun}(\text{noun}(\text{song})) \rightarrow [\text{song}].$   
 $\text{verb}(\text{verb}(\text{sings})) \rightarrow [\text{sings}].$

?–  $\text{sentence}(\text{Tree}, [\text{the}, \text{young}, \text{boy}, \text{sings}, \text{a}, \text{song}], [])$ .  
 $\text{Tree} = s(np(\text{det}(\text{the}), np2(\text{adj}(\text{young}), np2(\text{noun}(\text{boy})))),$   
 $\text{vp}(\text{verb}(\text{sings}), np(\text{det}(\text{a}), np2(\text{noun}(\text{song}))))$

# Derivační strom analýzy v DC gramatikách

?– sentence(Tree, [the, young, boy, sings, a, song], []).

Tree=s(np(det(the), np2(adj(young), np2(noun(boy)))),  
vp(verb(sings), np(det(a), np2(noun(song)))))



## Test na shodu

Pokud však rozšíříme slovník:

`noun(noun(boys)) --> [boys].`

`verb(verb(sing)) --> [sing].`

Narazíme na problém se shodou v čísle:

?- `sentence(.,[a, young, boys, sings],[]).`

Yes

?- `sentence(.,[a, boy, sing],[]).`

Yes

Proto rozšíříme neterminály o další argument **Num**, ve kterém můžeme testovat shodu:

`sentence(sentence(NP,VP)) --> noun_phrase(NP,Num), verb_phrase(VP,Num).`

## DC gramatika s testy na shodu

`sentence(sentence(N,V)) --> noun_phrase(N,Num), verb_phrase(V,Num).`  
`noun_phrase(np(D,N),Num) --> determiner(D,Num), noun_phrase2(N,Num).`  
`noun_phrase(np(N),Num) --> noun_phrase2(N,Num).`  
`noun_phrase2(np2(A,N),Num) --> adjective(A), noun_phrase2(N,Num).`  
`noun_phrase2(np2(N),Num) --> noun(N,Num).`  
`verb_phrase(vp(V),Num) --> verb(V,Num).`  
`verb_phrase(vp(V,N),Num) --> verb(V,Num), noun_phrase(N,Num1).`

<code>determiner(det(the), _) --&gt; [the].</code>	<code>noun(noun(boy),sg) --&gt; [boy].</code>
<code>determiner(det(a), sg) --&gt; [a].</code>	<code>noun(noun(song),sg) --&gt; [song].</code>
<code>verb(verb(sings), sg) --&gt; [sings].</code>	<code>noun(noun(boys),pl) --&gt; [boys].</code>
<code>verb(verb(sing), pl) --&gt; [sing].</code>	<code>noun(noun(songs),pl) --&gt; [songs].</code>
<code>adjective(adj(young)) --&gt; [young].</code>	

?- `sentence(_, [a, young, boys, sings], []).`

No

?- `sentence(_, [the, boys, sings, a, song], []).`

No

?- `sentence(_, [the, boys, sing, a, song], []).`

Yes

## Podmínky v těle pravidel

DC gramatiky mohou mít pomocné **podmínky** v těle pravidel – libovolný

### Prologovský kód

CFG pro vyhodnocení aritmetického výrazu:

$$\begin{array}{l}
 E \rightarrow T + E \quad | \quad T - E \quad | \quad T \\
 T \rightarrow F * T \quad | \quad F / T \quad | \quad F \\
 F \rightarrow (E) \quad | \quad f
 \end{array}$$

zapíšeme **včetně výpočtu** hodnoty výrazu:

$\text{expr}(X) \text{ --> term}(Y), [+], \text{expr}(Z), \{X \text{ is } Y+Z\}.$

$\text{expr}(X) \text{ --> term}(Y), [-], \text{expr}(Z), \{X \text{ is } Y-Z\}.$

$\text{expr}(X) \text{ --> term}(X).$

$\text{term}(X) \text{ --> factor}(Y), [*], \text{term}(Z), \{X \text{ is } Y*Z\}.$

$\text{term}(X) \text{ --> factor}(Y), [/], \text{term}(Z), \{X \text{ is } Y/Z\}.$

$\text{term}(X) \text{ --> factor}(X).$

$\text{factor}(X) \text{ --> ['('], expr}(X), [')'].$

$\text{factor}(X) \text{ --> [X], \{integer}(X)\}.$

?-  $\text{expr}(X, [3, +, 4, /, 2, -, '(, 2, *, 6, /, 3, +, 2, ')'], []).$

$X = -1$

%  $3 + 4/2 - (2*6/3 + 2) = -1$



## Generativní síla DCG

Generativní (rozpoznávací) síla DCG je větší než CFG

např. jazyk  $a^n b^n c^n$ :

$abc \rightarrow a(N), b(N), c(N).$

$a(0) \rightarrow [].$

$a(s(N)) \rightarrow [a], a(N).$

$b(0) \rightarrow [].$

$b(s(N)) \rightarrow [b], b(N).$

$c(0) \rightarrow [].$

$c(s(N)) \rightarrow [c], c(N).$

$?- abc(X, []).$

$X = [] ;$

$X = [a, b, c] ;$

$X = [a, a, b, b, c, c] ;$

$X = [a, a, a, b, b, b, c, c, c] ;$

...

# Význam syntaktické analýzy

- ▶ analýza **syntaxe** je **nutná** pro analýzu **významu**
- ▶ většina teorií analýzy významu využívá **princip kompozicionality**:

*Význam složeného výrazu je funkcí významu jednotlivých podvýrazů*

- ▶ proces **sémantické analýzy**:
  - buď vychází z **výsledků** syntaktické analýzy
  - nebo **probíhá současně** se syntaktickou analýzou; pak může zasahovat i do tvorby syntaktického stromu

# Problémy při analýze přirozeného jazyka

- ▶ víceznačnost
- ▶ anaforické výrazy
- ▶ indexické výrazy
- ▶ nejasnost
- ▶ nekompozicionalita
- ▶ struktura promluvy
- ▶ metonymie
- ▶ metafora

# Víceznačnost

- ▶ *ambiguity*
- ▶ **víceznačnost** může být **lexikální**, **syntaktická**, **sémantická** a **referenční**
- ▶ lexikální – “**stát**,” “**žena**,” “**hnát**”
- ▶ syntaktická – “**Jím špagety s masem.**”  
“**Jím špagety se salátem.**”  
“**Jím špagety s použitím vidličky.**”  
“**Jím špagety se sebezapřením.**”  
“**Jím špagety s přítelem.**”
- ▶ sémantická – “**Jeřáb** je vysoký.” “Viděli jsme veliké **oko.**”
- ▶ referenční – “**Oni** přišli pozdě.” “Můžeš mi půjčit **knihu?**”  
“Ředitel vyhodil dělníka, protože (**on**) byl agresivní.”

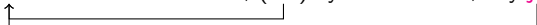
## Anaforické a indexické výrazy

### anaforické výrazy:

▶ *anaphora*

▶ používají **zájmena** pro odkazování na objekty zmíněné **dříve**

“Poté co se Honza s Marií rozhodli se vzít, (**oni**) vyhledali kněze, aby **je** oddal.”



“Marie uviděla ve výloze prstýnek a požádala Honzu, aby **jí ho** koupil.”



### indexické výrazy:

▶ *indexicals*

▶ odkazují se na údaje v **jiných částech** promluvy nebo **mimo** promluvu

“**Já** jsem **tady**.”

“Proč **jsi to** udělal?”

# Metafora a metonymie

## metafora:

- ▶ *metaphor*
- ▶ použití slov v **přeneseném významu** (na základě podobnosti), často systematicky

“Zkoušel jsem ten proces **zabít**, ale nešlo to.”

“Bouře se **vzteká**.”

## metonymie:

- ▶ *metonymy*
- ▶ používání **jména** jedné **věci** pro (často zkrácené) označení **věci jiné**

“Čtu **Shakespeara**.”

“**Chrysler** oznámil rekordní zisk.”

“Ten **pstruh na másle** u stolu 3 chce další pivo.”

# Nekompozicionalita

- ▶ *noncompositionality*
- ▶ příklady porušení pravidla kompozicionality u ustálených termínů nebo přednost jiného možného významu při určitých spojeních

“aligátoří boty,” “basketbalové boty,” “dětské boty”

“pata sloupu”

“červená kniha,” “červené pero”

“bílý trpaslík”

“dřevěný pes,” “umělá tráva”

“velká molekula”

## Reálná syntaktická analýza přirozeného jazyka

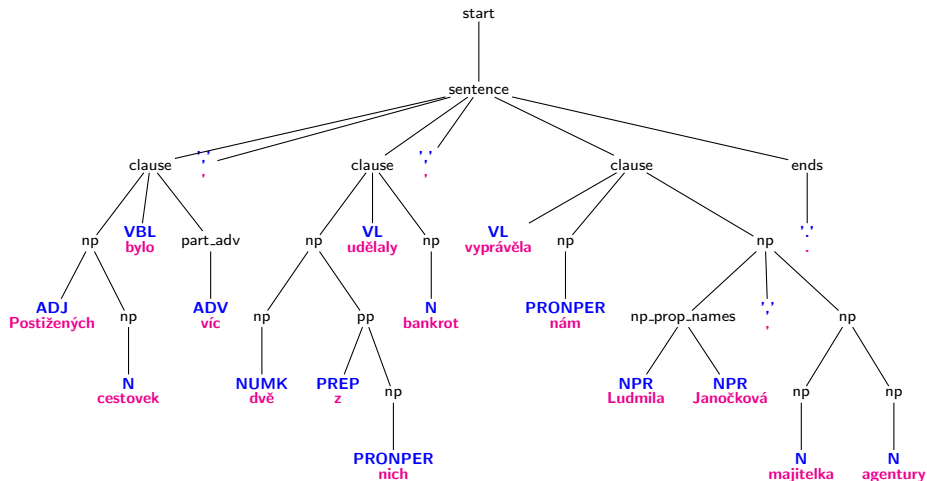
- ▶ velice **rozsáhlé gramatiky** (desítky až stovky tisíc pravidel)
- ▶ **silná víceznačnost** – někdy až obrovské množství (>milióny) možných syntaktických stromů

*Obehnat Šalounův pomník mistra Jana Husa na pražském Staroměstském náměstí živým plotem z hustých keřů s trny navrhuje občanské sdružení Společnost Jana Jesenia.*

- ▶ existují efektivní algoritmy pro takové gramatiky  
např. **tabulkový analyzátor** (*chart parser*), běží v  $O(n^3)$ , tisíce slov/sekundu



# Příklad stromu analýzy v systému synt



<http://nlp.fi.muni.cz/projekty/wwsynt/>

## Příklad logické analýzy v systému synt

Když je pořádná zima s množstvím sněhu, ani velký nával návštěvníků přírodě příliš nevadí.

$$\begin{aligned}
 & \lambda w_1 \lambda t_2 \left[ \text{kdýž\_ani}_{w_1 t_2}, \right. \\
 & \quad \lambda w_3 \lambda t_4 (\exists i_5) \left( \left[ \text{pořádný}_{w_3 t_4}, i_5 \right] \wedge \left[ \text{zima}_{w_3 t_4}, i_5 \right] \wedge \right. \\
 & \quad \quad \left. \left[ \left[ \text{s}_{w_3 t_4}, \left[ \text{Of, množství, sních} \right]_{w_3 t_4}, i_5 \right] \right) \right), \\
 & \quad \lambda w_6 \lambda t_7 \left[ \text{Not}, \left[ \text{True}_{w_6 t_7}, \lambda w_8 \lambda t_9 (\exists x_{10}) (\exists i_{11}) (\exists i_{12}) \left( \right. \right. \right. \\
 & \quad \quad \left. \left[ \text{Does}_{w_8 t_9}, i_{12}, \left[ \text{Imp}_{w_8}, x_{10} \right] \right] \wedge \left[ \text{příroda}_{w_8 t_9}, i_{11} \right] \wedge \right. \\
 & \quad \quad x_{10} \subset \left[ \text{vadit}, i_{11} \right]_{w_8} \wedge \left[ \text{příliš}, x_{10} \right] \wedge \\
 & \quad \quad \left. \left. \left. \left[ \left[ \text{velký}, \left[ \text{Of, nával, návštěvník} \right]_{w_8 t_9}, i_{12} \right] \right] \right] \right] \right] \dots \mathcal{O}_{T\omega}
 \end{aligned}$$

# NLP – Natural Language Processing

část **umělé inteligence** zaměřená na **zpracování textu a řeči**

**Významné úkoly v NLP** (předměty **IB030** a **IA161**)

- ▶ **analýza** textu v přirozeném jazyce – morfologická, syntaktická, sémantická
- ▶ **generování** textu v přirozeném jazyce
- ▶ syntéza a rozpoznávání **řeči**
- ▶ strojový překlad (**Machine translation**)
- ▶ odpovídání na otázky (**Question answering**)
- ▶ komunikace člověk-stroj (**Man-machine communication**, **Chatbots**)
- ▶ získávání informací (**Information retrieval**)
- ▶ korektura textu (**Spell-checking**, **Grammar checking**)
- ▶ extrakce informací (**Information extraction**, **Text mining**)
- ▶ výtah z textu (**Text summarization**)
- ▶ určení typu dokumentu (**Text classification/clustering**)
- ▶ určení stylu dokumentu/autora (**Stylometry**, **Authorship attribution**)

## PA026 – Projekt z umělé inteligence

- ▶ navazuje na předmět *PB016 Úvod do umělé inteligence*
- ▶ volba programovacího jazyka ovšem není nijak omezena
- ▶ samostatná volba tématu v rozsahu  $\geq 1$  semestru
- ▶ předmět probíhá jako konzultace
- ▶ zajímavé výsledky (<http://nlp.fi.muni.cz/uiprojekt/>)
  - projekt [elnet](#) – > 5 let spolupráce na grantových projektech simulace elektrorozvodných sítí
  - projekt [plagiaty\\_z\\_webu](#) – reálné a funkční vyhledávání shod s dokumenty na celém webu
  - projekt [robot\\_johnny\\_5](#) – sestavení a “oživení” robota – mobilního počítače

