

Zpracování přirozeného jazyka

Aleš Horák

E-mail: hales@fi.muni.cz
<http://nlp.fi.muni.cz/uui/>

Obsah:

- Komunikace
- Gramatiky a syntaktická analýza
- 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**

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**

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 ⇐ **jazyk** je pevně spojen s **myšlením**
komunikace se tvoří pomocí **řečových aktů** (*speech acts*) jako jeden z typů agentových akcí
cíl komunikace –

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 ⇐ **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?”
- **příkázat/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

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(Po ,
 $\neg Na_živu(Wumpus_1, S_3)$)

generování

“Wumpus je mrtvý.”

syntéza

[v u m p u s j e m r t v ý i]

Mluvčí

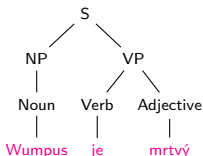
vnímání

analýza

zjednoznačnění

Posluchač

$\neg Na_živu(Wumpus_1, S_3)$

syntaktická
analýza:

“Wumpus
je mrtvý.”

sémantická interpretace: $\neg Na_živu(Wumpus, Ted)$
 $Unavený(Wumpus, Ted)$

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

zahrnutí

$Tell(KB,$
 $\neg Na_živu(Wumpus_1, S_3))$

Obsah

- 1 Komunikace
 - Přirozený jazyk – prostředek komunikace
 - Řečové akty
 - Komunikační fáze (při informování)
- 2 Gramatiky a syntaktická analýza
 - Gramatiky pro analýzu jazyka
 - Lexikon pro agenta ve Wumpusově jeskyni
 - Syntaktický strom
 - Test na shodu
 - Syntaktická analýza pomocí strojového učení
- 3 Analýza přirozeného jazyka
 - Význam syntaktické analýzy
 - Problémy při analýze přirozeného jazyka
 - Reálná syntaktická analýza přirozeného jazyka
- 4 PA026 – Projekt z umělé inteligence

Gramatiky a syntaktická analýza

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)

$$\underline{A}S\underline{B} \rightarrow \underline{AA}a\underline{BB}$$

umí $a^n b^n c^n$

- **rekurzivně vyčísitelné** (recursively enumerable) – bez omezení

ekvivalentní síle **Turingova stroje**

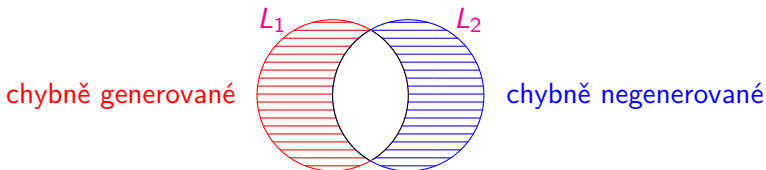
přirozený jazyk byl dlouho pokládán za bezkontextový → nyní prokázáno, že obsahuje **kontextové prvky**

Slido

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 ($|L_1 \cap L_2|/|L_1|$)
- kombinová **F-míra** – harmonický průměr $2 \cdot \frac{\text{přesnost} \cdot \text{pokrytí}}{\text{přesnost} + \text{pokrytí}}$

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

Gramatiky pro analýzu jazyka

využívané pro **syntaktickou analýzu**

- pro lokální varianty – **regulární** gramatiky (regulární výrazy, např. pro *extrakci informací*)
- pro vyjmenované větné struktury – **bezkontextové** gramatiky
- pro plný jazyk – (**mírně**) **kontextové** gramatiky
- praktické nástroje – většinou **rozšíření bezkontextových gramatik** (CFG):
 - Prolog – definite clause grammars, **DCG**
 - Java, Python – **ANTLR** (**AN**Other **T**ool for **L**anguage **R**ecognition)

```
grammar Expr;
prog:  (expr NEWLINE)* ;
expr:  expr ('*' | '/') expr
      | expr ('+' | '-') expr
      | INT
      | '(' expr ')'
      ;
NEWLINE: [\r\n]+ ;
INT: [0-9]+ ;
```

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'.

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

```
True
```

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

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: sl_druh(lemma, pád, číslo, rod) → slovo.

adj('chytrý', '1', 'j', 'mž') → 'chytrý'.

adj('chytrý', '2', 'j', 'mž') → 'chytrého'.

adj('chytrý', '1', 'mn', 'mž') → 'chytrí'.

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: sl_druh(lemma, pád, číslo, rod) → slovo.

adj('chytrý', '1', 'j', 'mž') → 'chytrý'.

adj('chytrý', '2', 'j', 'mž') → 'chytrého'.

adj('chytrý', '1', 'mn', 'mž') → 'chytrí'.

- 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
<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
<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>
<i>RelClause</i>	→	<i>',' který' VP</i>	%	,který + zapáchá

Syntaktický strom

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

Východní

jáma

tady

působí

vánek

Syntaktický strom

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

Adjective

|
Východní

Noun

|
jáma

Adverb

|
tady

Verb

|
působí

Noun

|
vánek

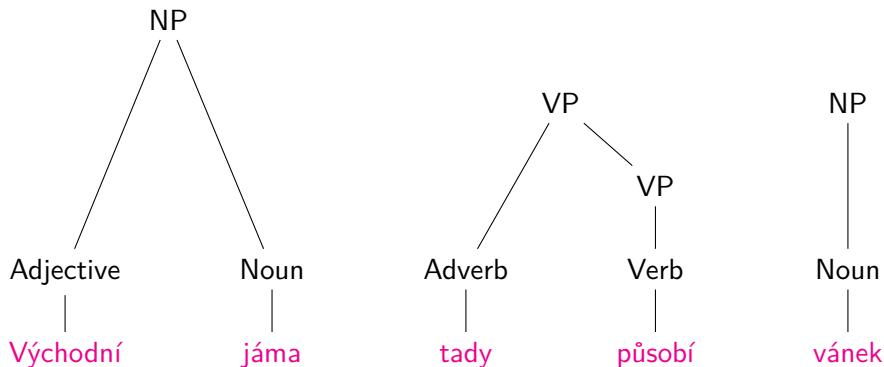
Syntaktický strom

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



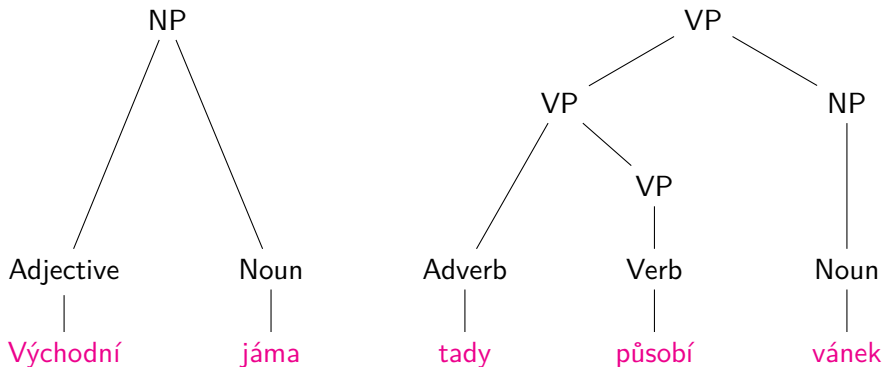
Syntaktický strom

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



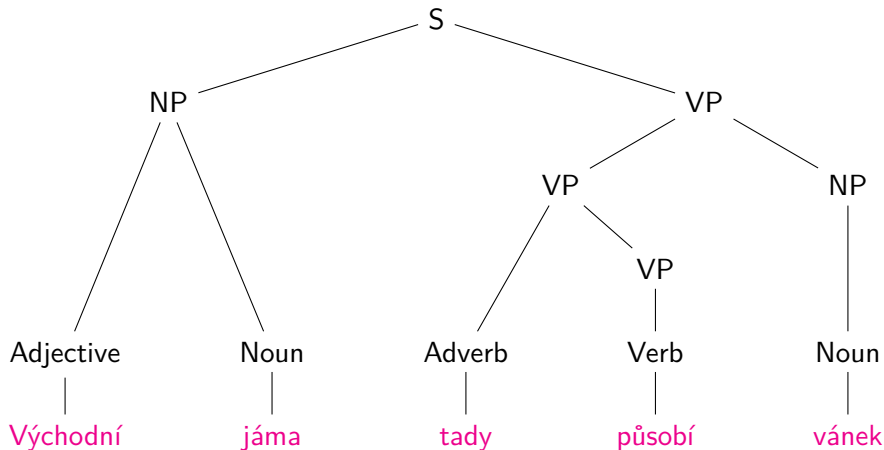
Syntaktický strom

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



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).`

`sentence(s(N,V)) → noun_phrase(N), verb_phrase(V).`

`noun_phrase(np(D,N)) → determiner(D), noun_phrase2(N).`

`noun_phrase(np(N)) → noun_phrase2(N).`

`noun_phrase2(np2(A,N)) → adjective(A), noun_phrase2(N).`

`noun_phrase2(np2(N)) → noun(N).`

`verb_phrase(vp(V)) → verb(V).`

`verb_phrase(vp(V,N)) → verb(V), noun_phrase(N).`

`determiner(det(the)) → 'the'.`

`determiner(det(a)) → 'a'.`

`adjective(adj(young)) → 'young'.`

`noun(noun(boy)) → 'boy'.`

`noun(noun(song)) → 'song'.`

`verb(verb(sings)) → 'sings'.`

Konstrukce derivačního stromu

Neterminály opatříme argumentem:

`sentence(sentence(NP,VP)) → noun_phrase(NP), verb_phrase(VP).`

`sentence(s(N,V)) → noun_phrase(N), verb_phrase(V).`

`noun_phrase(np(D,N)) → determiner(D), noun_phrase2(N).`

`noun_phrase(np(N)) → noun_phrase2(N).`

`noun_phrase2(np2(A,N)) → adjective(A), noun_phrase2(N).`

`noun_phrase2(np2(N)) → noun(N).`

`verb_phrase(vp(V)) → verb(V).`

`verb_phrase(vp(V,N)) → verb(V), noun_phrase(N).`

`determiner(det(the)) → 'the'.`

`determiner(det(a)) → 'a'.`

`adjective(adj(young)) → 'young'.`

`noun(noun(boy)) → 'boy'.`

`noun(noun(song)) → 'song'.`

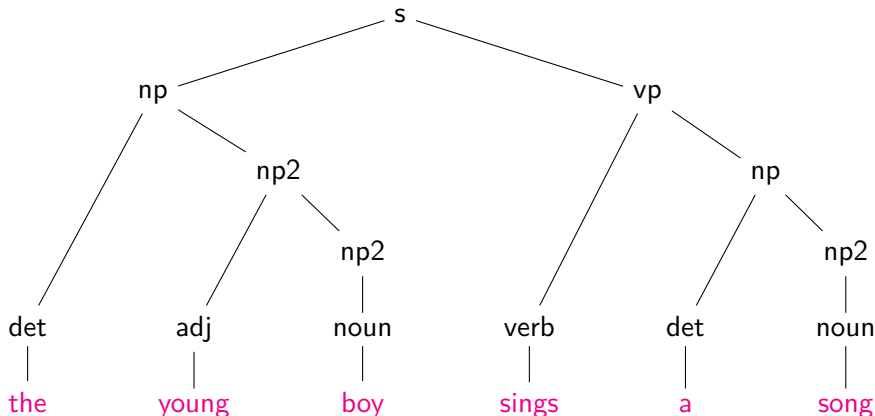
`verb(verb(sings)) → 'sings'.`

`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')))))`

Derivační strom analýzy v 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']).  
True
```

```
sentence(_, ['a', 'boy', 'sing']).  
True
```

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']).  
True
```

```
sentence(,['a', 'boy', 'sing']).  
True
```

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

```
sentence(sentence(NP,VP)) → noun_phrase(NP, Num), verb_phrase(VP, Num).
```

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).`

`determiner(det(the), _) → 'the'.`
`determiner(det(a), sg) → 'a'.`
`verb(verb(sings), sg) → 'sings'.`
`verb(verb(sing), pl) → 'sing'.`
`adjective(adj(young)) → 'young'.`

`noun(noun(boy), sg) → 'boy'.`
`noun(noun(song), sg) → 'song'.`
`noun(noun(boys), pl) → 'boys'.`
`noun(noun(songs), pl) → 'songs'.`

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), _) → 'the'.</code>	<code>noun(noun(boy), sg) → 'boy'.</code>
<code>determiner(det(a), sg) → 'a'.</code>	<code>noun(noun(song), sg) → 'song'.</code>
<code>verb(verb(sings), sg) → 'sings'.</code>	<code>noun(noun(boys), pl) → 'boys'.</code>
<code>verb(verb(sing), pl) → 'sing'.</code>	<code>noun(noun(songs), pl) → 'songs'.</code>
<code>adjective(adj(young)) → 'young'.</code>	

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

False

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), _) → 'the'.</code>	<code>noun(noun(boy), sg) → 'boy'.</code>
<code>determiner(det(a), sg) → 'a'.</code>	<code>noun(noun(song), sg) → 'song'.</code>
<code>verb(verb(sings), sg) → 'sings'.</code>	<code>noun(noun(boys), pl) → 'boys'.</code>
<code>verb(verb(sing), pl) → 'sing'.</code>	<code>noun(noun(songs), pl) → 'songs'.</code>
<code>adjective(adj(young)) → 'young'.</code>	

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

False

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

False

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), _) → 'the'.</code>	<code>noun(noun(boy), sg) → 'boy'.</code>
<code>determiner(det(a), sg) → 'a'.</code>	<code>noun(noun(song), sg) → 'song'.</code>
<code>verb(verb(sings), sg) → 'sings'.</code>	<code>noun(noun(boys), pl) → 'boys'.</code>
<code>verb(verb(sing), pl) → 'sing'.</code>	<code>noun(noun(songs), pl) → 'songs'.</code>
<code>adjective(adj(young)) → 'young'.</code>	

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

False

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

False

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

True

Generativní síla gramatik

Generativní (rozpoznávací) síla analyzačních gramatik je často větší než CFG např. jazyk $a^n b^n c^n$:

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

$a(0) \rightarrow \square. \quad \# \in$
 $a(s(N)) \rightarrow 'a', a(N).$

$b(0) \rightarrow \square.$
 $b(s(N)) \rightarrow 'b', b(N).$

$c(0) \rightarrow \square.$
 $c(s(N)) \rightarrow 'c', c(N).$

Generativní síla gramatik

Generativní (rozpoznávací) síla analyzačních gramatik je často větší než CFG
např. jazyk $a^n b^n c^n$:

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

$a(0) \rightarrow [].$ $\# \in$
 $a(s(N)) \rightarrow 'a', a(N).$

$b(0) \rightarrow [].$
 $b(s(N)) \rightarrow 'b', b(N).$

$c(0) \rightarrow [].$
 $c(s(N)) \rightarrow 'c', c(N).$

$abc(X, []).$

Generativní síla gramatik

Generativní (rozpoznávací) síla analyzačních gramatik je často větší než CFG např. jazyk $a^n b^n c^n$:

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

$a(0) \rightarrow [].$ $\# \in$
 $a(s(N)) \rightarrow 'a', a(N).$

$b(0) \rightarrow [].$
 $b(s(N)) \rightarrow 'b', b(N).$

$c(0) \rightarrow [].$
 $c(s(N)) \rightarrow 'c', c(N).$

$abc(X, []).$
 $X = []$

Generativní síla gramatik

Generativní (rozpoznávací) síla analyzačních gramatik je často větší než CFG např. jazyk $a^n b^n c^n$:

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

$a(0) \rightarrow []$. $\# \in$
 $a(s(N)) \rightarrow 'a', a(N)$.

$b(0) \rightarrow []$.
 $b(s(N)) \rightarrow 'b', b(N)$.

$c(0) \rightarrow []$.
 $c(s(N)) \rightarrow 'c', c(N)$.

$abc(X, [])$.
 $X = []$
 $X = ['a', 'b', 'c']$

Generativní síla gramatik

Generativní (rozpoznávací) síla analyzačních gramatik je často větší než CFG např. jazyk $a^n b^n c^n$:

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

$a(0) \rightarrow [].$ $\# \in$
 $a(s(N)) \rightarrow 'a', a(N).$

$b(0) \rightarrow [].$
 $b(s(N)) \rightarrow 'b', b(N).$

$c(0) \rightarrow [].$
 $c(s(N)) \rightarrow 'c', c(N).$

$abc(X, []).$

$X = []$

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

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

Generativní síla gramatik

Generativní (rozpoznávací) síla analyzačních gramatik je často větší než CFG např. jazyk $a^n b^n c^n$:

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

$a(0) \rightarrow [].$ $\# \in$
 $a(s(N)) \rightarrow 'a', a(N).$

$b(0) \rightarrow [].$
 $b(s(N)) \rightarrow 'b', b(N).$

$c(0) \rightarrow [].$
 $c(s(N)) \rightarrow 'c', c(N).$

$abc(X, []).$

$X = []$

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

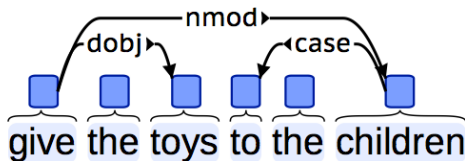
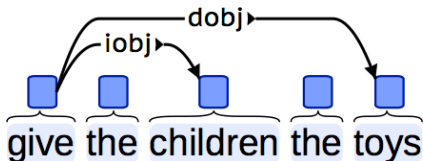
$X = ['a', 'a', 'b', 'b', 'c', 'c']$

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

...

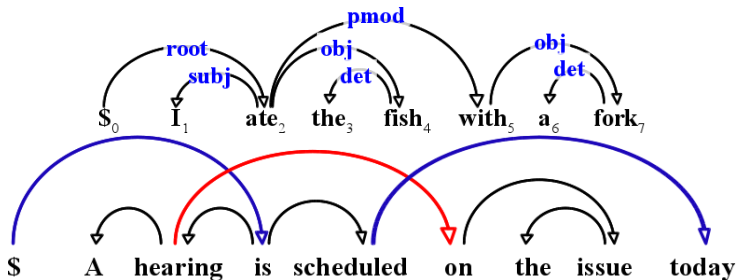
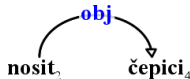
Syntaktická analýza pomocí strojového učení

- využití **anotovaných stromových korpusů** (*treebanks*)
- lidé **anotují** textový korpus – doplní **syntaktické stromy**
- **strojové učení** hledá **pravidla/váhy** parametrů
- **univerzální** napříč jazyky (do jisté míry)
- anotování je **drahé**
- **modifikace** pro různé účely je obtížnější
- často **není dost dat**



Závislostní analýza

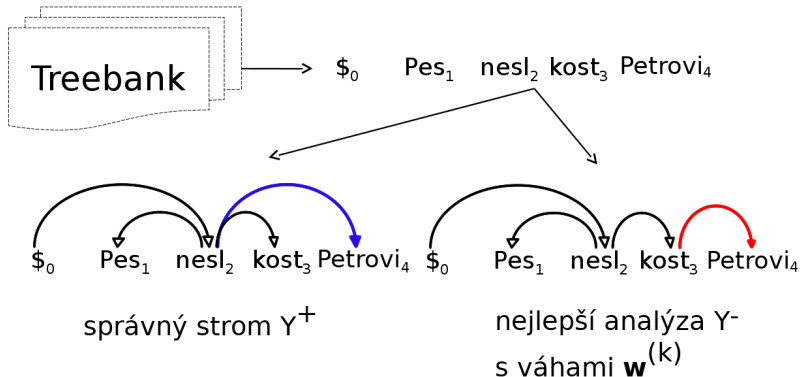
- jedna hrana pro každé slovo
 - **hlava** – řídicí slovo
 - **závislé/rozvíjející** slovo – modifikátor
 - **typ** – popisek hrany
- obtížné pro **neprojektivní stromy**



Example from “Dependency Parsing” by Kübler, Nivre, and McDonald, 2009

Online učení skóre hrany

učení **matice vah rysů w**



$$w^{(k+1)} = w^{(k)} + f(X, Y^+) - f(X, Y^-)$$

Obsah

- 1 Komunikace
 - Přirozený jazyk – prostředek komunikace
 - Řečové akty
 - Komunikační fáze (při informování)
- 2 Gramatiky a syntaktická analýza
 - Gramatiky pro analýzu jazyka
 - Lexikon pro agenta ve Wumpusově jeskyni
 - Syntaktický strom
 - Test na shodu
 - Syntaktická analýza pomocí strojového učení
- 3 Analýza přirozeného jazyka
 - Význam syntaktické analýzy
 - Problémy při analýze přirozeného jazyka
 - Reálná syntaktická analýza přirozeného jazyka
- 4 PA026 – Projekt z umělé inteligence

Význam syntaktické analýzy

- analýza **syntaxe** je **podkladem** 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
- metafory

Víceznačnost

- *ambiguity*
- **víceznačnost** může být **lexikální**, **syntaktická**, **sémantická** a **referenční**

Víceznačnost

- *ambiguity*
- **víceznačnost** může být **lexikální**, **syntaktická**, **sémantická** a **referenční**
- lexikální – “**stát**,” “**žena**,” “**hnát**”

Víceznačnost

- *ambiguity*
- **víceznačnost** může být lexikální, syntaktická, sémantická a referenční
- lexikální – “stát,” “žena,” “hnát”
- syntaktická – “Jím špagety s masem.”

Víceznačnost

- *ambiguity*
- **víceznačnost** může být **lexikální**, **syntaktická**, **sémantická** a **referenční**
- lexikální – “**stát**,” “**žena**,” “**hnát**”
- syntaktická – “**Jím špagety s masem.**”
 “**Jím špagety se salátem.**”

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.”

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.”

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.”

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**.”

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**

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.”

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.”



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.”

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.”

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

Anaforické a indexické výrazy

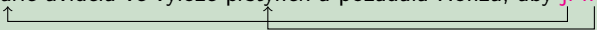
anaforické výrazy:

- *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**.”

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

Metafora a metonymie

metafora:

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

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

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**.”

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**

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**.”

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.”

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

Nekompozicionalita

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

“aligátóří boty,” “basketbalové boty,” “dětské boty”

Nekompozicionalita

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

“aligátóří boty,” “basketbalové boty,” “dětské boty”

“pata sloupu”

Nekompozicionalita

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

“aligátóří boty,” “basketbalové boty,” “dětské boty”

“pata sloupu”

“červená kniha,” “červené pero”

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”

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”

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

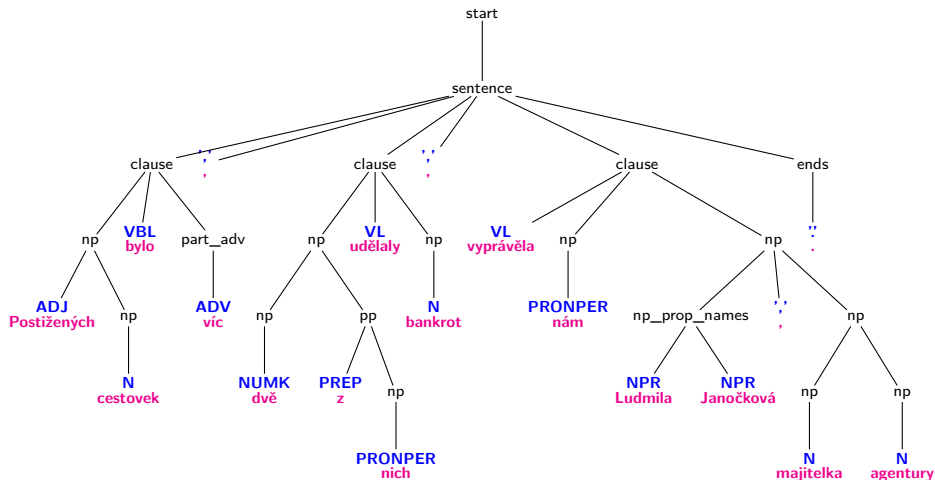
Reálná syntaktická analýza přirozeného jazyka

- 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 **sdržení** 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.

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

NLP – Natural Language Processing

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

Významné úkoly v NLP (předmět IA161)

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



<https://beta.openai.com/playground>

Obsah

- 1 Komunikace
 - Přirozený jazyk – prostředek komunikace
 - Řečové akty
 - Komunikační fáze (při informování)
- 2 Gramatiky a syntaktická analýza
 - Gramatiky pro analýzu jazyka
 - Lexikon pro agenta ve Wumpusově jeskyni
 - Syntaktický strom
 - Test na shodu
 - Syntaktická analýza pomocí strojového učení
- 3 Analýza přirozeného jazyka
 - Význam syntaktické analýzy
 - Problémy při analýze přirozeného jazyka
 - Reálná syntaktická analýza přirozeného jazyka
- 4 PA026 – Projekt z umělé inteligence

PA026 – Projekt z umělé inteligence

- navazuje na předmět *PB016 Úvod do umělé inteligence*
- volba programovacího jazyka není omezena
- samostatná volba tématu v rozsahu ≥ 1 semestru
- předmět probíhá jako prezentace a konzultace
- zajímavé výsledky (<http://nlp.fi.muni.cz/uiprojekt/>)
 - projekt [elnet](#) – > 5 let spolupráce na grantových projektech simulace elektrorozvodných sítí
 - projekt [plagiaty_z_webu](#) – vyhledávání shod s dokumenty na celém webu
 - projekt [robot_johnny_5](#) – sestavení a “oživení” robota – mobilního počítače
 - robot [Karel Pepper](#) – <https://nlp.fi.muni.cz/projects/pepper>

