Syntaktická analýza

(přirozeného jazyka)

Syntaxe

- popis dobře utvořených posloupností
  v PJ věty, větné fráze
  Velcí psi štěkali. vs. ~~Velký psi štěkaly~~.
- nástroj popisu: gramatika (pravidla)

S → NP VP	ADJ → 'velcí'
NP → ADJ N	N → 'psi'
VP → V	V → 'štěkali'

- výstup SyA: typicky stromy
  složkové, závislostní
- návaznosti morfologie → syntaxe → sémantika (prolínání)

Gramatiky

- hierarchie
  regulární (KA, neumí $a^n b^n$), CF (ZA, neumí $a^n b^n c^n$), kontextové, typ 0
- PJ není bezkontextový
  důk. prof. Novotného
  Norsko, Čína a USA mají postupně hlavní města Oslo, Peking a Washington a nacházejí se v Evropě, Asii a Americe.
- přesto se používají CFG
  s případnými rozšířeními

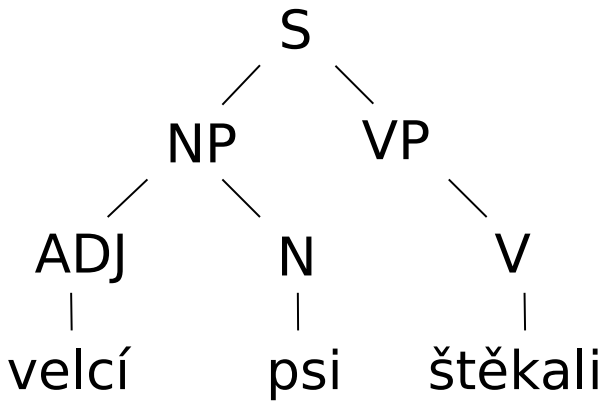
Terminologie

- fráze (skupina)
  jednotka větší než slovo, ale menší než věta, např. jmenná fráze
  'velcí psi', 'ta naše česká povaha'
- lexikální kategorie
  neterminál přepisující se přímo na terminál N → 'psi'
- frázová kategorie
  neterminál nevyjadřující lexikální kategorii NP → ADJ N
- větný člen
  lexikální nebo frázová kategorie

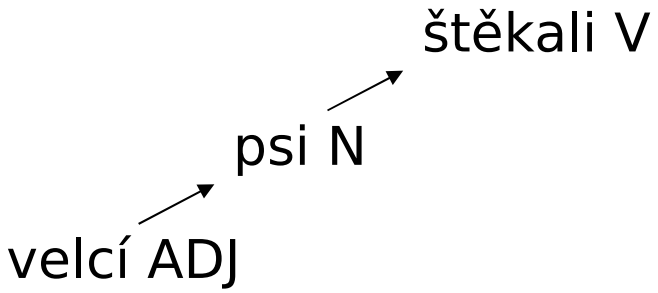
Stromy

zachycují větnou strukturu

- složkové stromy
  zachycují povrchovou strukturu



- závislostní stromy
  dvojice řídicí prvek – závislý prvek



Krok k sémantice

nalezení **sémantických rolí** větných členů

- tzv. hloubková struktura věty
- agens (kdo je původce činnosti), patiens (komu se něco děje), donor (kdo dává), ...
- vychází se z gramatických rolí větných členů (podmět, předmět, ...), k jejich určení pomáhá gramatický pád
- není jednoduché ani určení gramatických rolí, natož sémantických

Zachycení shody

potřebujeme testovat shodu

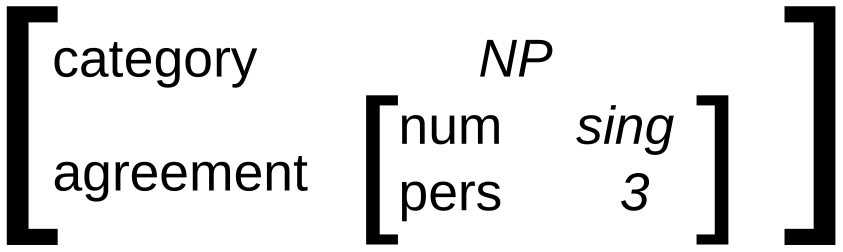
číslo, pád, rod, podmět – přísudek,...

- neterminály „s parametry“

  např. v DCG lze přímo NP(g,n,c) → ADJ(g,n,c) N(g,n,c)

- sestavy rysů (feature structures)
  - množina dvojic atribut-hodnota, hodnoty mohou být komplexní
  - použití v řadě gramatických formalismů (GPSG, HPSG, LFG)
  - unifikace složitějších struktur

Př.: sestava rysů



- dva atributy (category, agreement)
  hodnota druhého je komplexní (sestava rysů)
- označuje se AVM (attribute value matrix)
- používá se i seznamová notace

Slovosled, nespojité fráze

v češtině musíme řešit

- volný slovosled

  Petr byl včera doma. Doma byl Petr včera. Včera byl doma Petr.

  (např. synt řeší kombinatorickými konstrukty)

- nespojité fráze
  - hlavně slovesné, ale nejen

    by se zde mělo utkání konat

    hrdliččin zval ku lásce hlas
  - možné řešení: mezerové gramatiky

Vlastní SyA

- vytvoření gramatiky ve zvoleném formalismu
  DCG, unifikační, závislostní, GPSG, HPSG, LFG, CG, TAG, metagramatika (synt),...
- výběr/vytvoření algoritmu pro analýzu
  chart (tabulková) – shora dolů, zdola nahoru, head-driven; CKY, zobecněný LR
  - gramatika + vstupní věta → patří/nepatří + reprezentace (strom)
  - úpravy gramatiky (normální formy)
  - rozpoznávání/generování
- pro češtinu
  klara, synt, DIS, Zuzana

synt: metagramatika

3 úrovně

- metagramatika G1: 253 pravidel
  kombinatorické konstrukty (generování pořadí), akce (gramatické testy, kontextové akce), závislostní struktury
- generovaná G2: 3091 pravidel
  bezkontextová + akce (shoda, zanoření vedlejších vět,...)
- expandovaná G3: 11530 pravidel
  pouze bezkontextová

DCG

Definite clause grammars

- jedny z nejstarších gramatik navržených pro PJ
- 2 rozšíření CFG: neterminály - složené termy (lze parametry), možnost volání procedur
- možnost zpracování nespojitých frází (gapping grammars)
- umí $a^n b^n c^n$

```
abc --> a(N), b(N), c(N).
a(0) --> [].
a(s(N)) --> [a], a(N).
b(0) --> [].
b(s(N)) --> [b], b(N).
c(0) --> [].
c(s(N)) --> [c], c(N).
```

Parciální analýza

partial/shallow parsing, chunking

Důvody:

- chyby něco nezpracuje MA, transformace zdrojů (oddělovače vět ap.), přímo ve zdrojových textech (gramatické chyby)
- příliš mnoho úplných analýz synt: 57 102 672 pro
  Obehnat Šalounův pomník mistra Jana Husa na pražském Staroměstském náměstí živým plotem z hustých keřů s trny navrhuje občanské sdružení Společnost Jana Jesenia.
- pro řadu aplikací stačí
- analyzátor DIS

DIS

- pravidla pro NG, PNG, VG
  získaná/ověřená v korpusech, řeší nespojité slovesné fráze
- DCG + vlastní implementace mezerových gramatik
- aplikace částečná desambiguace, kategorizace textů
- nadstavba VaDIS hledá některé gramatické role + povrchové valence → další desambiguace

DIS: příklad pravidla

```
% příklad: by se mohla zbavit

slovesná_skupina(Gap1) -->
  kond(S0,Lm0,k5,E0,P0,N0,tP,mC,A0),  % by/k5e?p?n_tPmCa?
  se(Se,sebe,k3,xX,nS,C1),            % sebe/k3xXnSc?
  gap([],Gap0),                       % gap
  sloveso(S2,Lm2,k5,E2,P2,N2,tM,mP,A2),  % modal/k5e?p?n_tMmPa?
   { check_n(N2,N0,P0),               % { check_n }
     modal(Lm2),
     check_p(P0,P2)
   },
  gap(Gap0,Gap1),                     % gap
  sloveso(S3,Lm3,k5,E3,mF,A3),        % k5e?mFa?
   { not_by(Lm3) }.
```

Valence

- další krok k sémantice
- povrchové zpracovává VaDIS

  opouštět
  = koho||co
  = koho||co & pro co
  = koho||co & kvůli čemu

  opouštět <v>hPTc4,hPTc4-hTc4r{pro},hPTc4-hTc3r{kvůli}

- hloubkové

  VerbaLex, experiment Zuzana