Aleš Horák

E-mail: hales@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- HPSG – Head-driven Phrase Structure Grammar
- Metagramatika systému synt

HPSG – Head-driven Phrase Structure Grammar

- HPSG, **Head-driven Phrase Structure Grammar** – Pollard & Sag, 1994
- navazuje na Gazdar, **Generalized Phrase Structure Grammar**, 1985
- **lexikalizovaná** teorie generativní gramatiky přirozeného jazyka
- *neterminály* CFG jsou nahrazeny **příznakovými strukturami**
- založená na **omezeních** (constraints)
- modeluje jazyk pomocí **deklarativních** omezení typovaných struktur
- **příznaky** jsou propojeny pomocí **strukturního sdílení**, tedy předáváním proměnných mezi podstrukturami dané struktury
- HPSG je **nederivační**, na rozdíl od jiných formalismů, kde jsou různé úrovně syntaktické struktury sekvenčně odvozovány pomocí transformačních operací

HPSG – Head-driven Phrase Structure Grammar – pokrač.

- gramatika je v HPSG modelována pomocí **uspořádaných příznakových struktur**, které korespondují s typy výrazů přirozeného jazyka a jejich částmi
- cílem teorie je detailní specifikace, které příznakové struktury jsou **přípustné**
- příznakové struktury definují **omezení**
  hodnoty příznaků mohou být jednoho ze čtyř typů
  - atomy
  - příznakové struktury
  - množiny příznakových struktur ($\{\ldots\}$)
  - nebo seznamy příznakových struktur ($<\ldots>$)

HPSG – lexikální hlava

- **slova** (lexikální položky) obsahují **hodně informací** – podle psycholingvistiky se podobá *zpracování v lidském mozku*
- **lexikální hlava** – základní prvek frázové struktury HPSG
  lexikální hlava = jedno slovo, jehož položka specifikuje informace, které určují základní gramatické **vlastnosti fráze**, kterou hlava zastupuje
  gramatické vlastnosti zahrnují:
  - morfologické informace (part-of-speech, POS)

    N zastupuje NP, VP zastupuje S, V zastupuje VP
  - relace závislosti (např. valenční rámec slovesa)
- lexikální hlava obsahuje také klíčové **sémantické informace**, které sdílí se zastupovanou frází

HPSG – struktury

HPSG struktury jsou **typované příznakové struktury**
zapisují se pomocí AVM – **příznaky** velkými písmeny, **typy** malými

$$\begin{bmatrix} \text{word} & \textit{\% slovo jde} \\ \text{PHON} & \langle \text{jde} \rangle \\ \text{SYNSEM} & \begin{bmatrix} \text{synsem} & \\ \text{CATEGORY} & \begin{bmatrix} \text{category} & \\ \text{HEAD} & \begin{bmatrix} \text{verb} & \\ \text{VFORM} & \text{finite} \\ \text{AUX} & - \\ \text{INV} & - \\ \text{PRD} & - \end{bmatrix} \\ \text{VALENCE} & \begin{bmatrix} \text{val} & \\ \text{SUBJECT} & \langle \text{NP[nom]::[1]} \rangle \\ \text{COMPLEMENTS} & \langle\rangle \\ \text{SPECIFIER} & \langle\rangle \end{bmatrix} \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{jít} & \\ \text{KDO[1]} & \begin{bmatrix} \text{index} & \\ \text{PERSON} & \text{3rd} \\ \text{NUMBER} & \text{singular} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

HPSG – syntaktické kategorie

symboly **syntaktických kategorií** – zkratky určitých příznakových popisů:

$$\text{NP::[1]} = \begin{bmatrix} \text{synsem} & \\ \text{LOCAL} & \begin{bmatrix} \text{local} & \\ \text{CATEGORY} & \begin{bmatrix} \text{category} & \\ \text{HEAD} & \text{noun} \\ \text{VALENCE} & \begin{bmatrix} \text{valence} & \\ \text{SUBJ} & \langle\rangle \\ \text{COMPS} & \langle\rangle \\ \text{SPR} & \langle\rangle \end{bmatrix} \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{nom-obj} & \\ \text{INDEX} & \text{[1]} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

$$\text{VP} = \begin{bmatrix} \text{synsem} & \\ \text{LOCAL} & \begin{bmatrix} \text{local} & \\ \text{CATEGORY} & \begin{bmatrix} \text{category} & \\ \text{HEAD} & \text{verb} \\ \text{VALENCE} & \begin{bmatrix} \text{valence} & \\ \text{SUBJ} & \langle \text{[synsem]} \rangle \\ \text{COMPS} & \langle\rangle \\ \text{SPR} & \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

HPSG – lexikální položky

velké množství akcí je v **lexikonu**:

$$\text{JÍT} \begin{bmatrix} \text{CATEGORY} & \begin{bmatrix} \text{HEAD} & \text{verb} \\ \text{VALENCE} & \begin{bmatrix} \text{SUBJ} & \langle \text{NP::[1]} \rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix} \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{jít} & \\ \text{KDO} & \text{[1]} \end{bmatrix} \end{bmatrix}$$

$$\text{DÁT} \begin{bmatrix} \text{CATEGORY} & \begin{bmatrix} \text{HEAD} & \text{verb} \\ \text{VALENCE} & \begin{bmatrix} \text{SUBJ} & \langle \text{NP::[1]} \rangle \\ \text{COMPS} & \langle \text{NP::[2],NP::[3]} \rangle \end{bmatrix} \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{dát} & \\ \text{KDO} & \text{[1]} \\ \text{CO} & \text{[2]} \\ \text{KOMU} & \text{[3]} \end{bmatrix} \end{bmatrix}$$

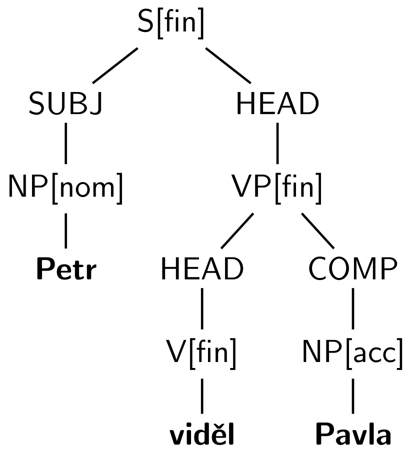
HPSG – fráze

reprezentace **frází** – v HPSG obdoba reprezentace **slov**
navíc příznak **DAUGHTERS** – struktura členů fráze

$$\begin{bmatrix} \text{phrase} & \\ \text{SYNSEM} & \text{S[fin]} \\ \text{DTRS} & \begin{bmatrix} \text{SUBJ-DTR} & \begin{bmatrix} \text{phrase} & \\ \text{PHON} & \langle \text{Petr} \rangle \\ \text{SYNSEM} & \text{NP[nom]} \end{bmatrix} \\ \text{HEAD-DTR} & \begin{bmatrix} \text{phrase} & \\ \text{PHON} & \langle \text{viděl Pavla} \rangle \\ \text{SYNSEM} & \text{VP[fin]} \\ \text{DTRS} & \begin{bmatrix} \text{HEAD-DTR} & \begin{bmatrix} \text{word} & \\ \text{PHON} & \langle \text{viděl} \rangle \\ \text{SYNSEM} & \text{V[fin]} \end{bmatrix} \\ \text{COMP-DTR} & \begin{bmatrix} \text{phrase} & \\ \text{PHON} & \langle \text{Pavla} \rangle \\ \text{SYNSEM} & \text{NP[acc]} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

HPSG – fráze – pokrač.

pro snazší **čtení popisů frází** používáme **stromový zápis**:



ve skutečnosti se ovšem jedná o **příznakovou strukturu**, ne strom!

HPSG – dobře utvořené příznakové struktury

dobře utvořené příznakové struktury musí splňovat **omezení daná gramatikou**

příznaková struktura je **dobře utvořená** ⇔:

- každý uzel splňuje **omezení geometrie příznaku**
- každá uzel vstupního slova splňuje **omezení** některé **lexikální položky**
- každý frázový uzel splňuje **frázová omezení** – *omezení přímé dominance* (immediate dominance), *omezení hlavových příznaků* (head feature), *valenční omezení*, ...

omezení geometrie příznaku specifikují:

- s jakými **typy** se pracuje
- jaká je použitá **typová hierarchie** – který typ je podtypem jiného typu
- pro každý typ – jaké příznaky přísluší tomuto typu
- pro každý typ a každý příznak – jakých typů mohou být hodnoty tohoto příznaku

HPSG – deklarace typu

pro popis omezení geometrie příznaku se používají **typové deklarace**:

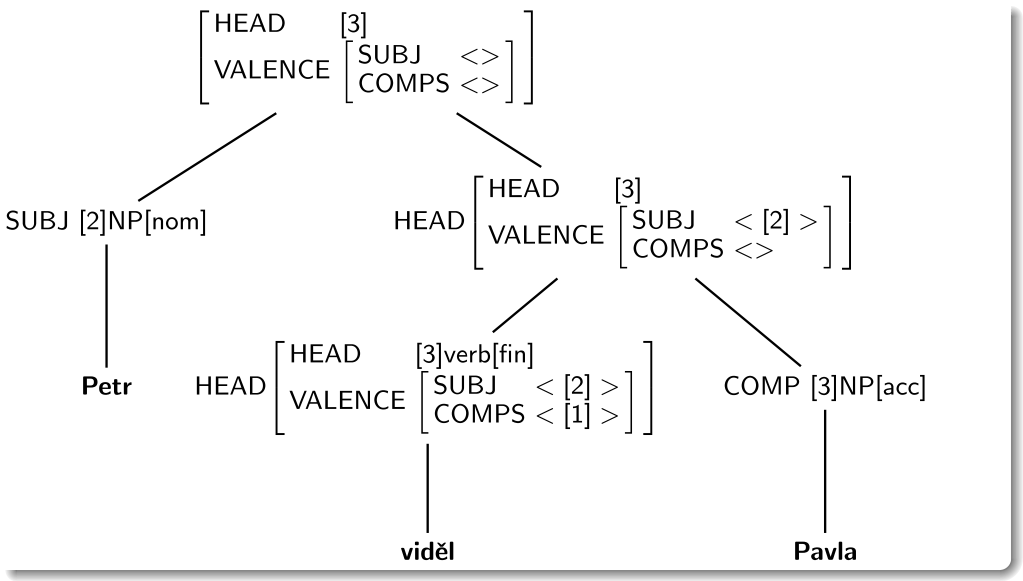
```
category: [HEAD: head, VALENCE: valence]

head     # příznaková struktura složená z příznakových struktur
   noun: [CASE: case]
   verb: [VFORM: vform, AUX: boolean, INV: boolean]
   prep: [PFORM: pform]
   ...

vform     # jednoduchý příznak, forma slovesa – možné hodnoty:
   fin     # určitý tvar slovesa
   inf     # neurčitý tvar slovesa – infinitive
   ...

case     # jednoduchý příznak, gramatický pád
   nom     # 1. pád, nominativ
   acc     # 4. pád, akuzativ
   ...
```


HPSG – dobře utvořená slova a fráze

- každé vstupní **slovo** musí splňovat nějakou **lexikální položku**
- **fráze** musí splňovat **frázová omezení**:
  - **omezení přímé dominance** – schéma head-subject, schéma head-complement, ...
  - **omezení hlavových příznaků** – pro každou frázi, která má hlavu, musí být hlavové příznaky fráze shodné s hlavovými příznaky potomka, který je hlavou
  - **valenční omezení** – pro každý z valenčních příznaků (SUBJECT, COMPLEMENTS, ...) – hodnota příznaku na hlavové frázi musí odpovídat hodnotě na potomku, který je hlavou, mínus ty příznaky, které jsou splněny některým z nehlavových potomků

HPSG – dobře utvořené příznakové struktury

omezení ve větě ‘`Petr viděl Pavla.`’:



Metagramatika systému synt

3 formy (meta)gramatiky:

- metagramatika (G1)
  - pravidla s kombinatorickými konstrukty + globální omezení pořadí
  - akce (= gramatické testy + kontextové akce)
  - česká lingvistická tradice – závislostní struktury, kontrola shody, pravidla pro pořadí slov, . . .
- generovaná gramatika (G2)
  - bezkontextová pravidla
  - akce
- expandovaná gramatika (G3)
  - jen bezkontextová pravidla

Metagramatika – kombinatorické konstrukty

kombinatorické konstrukty se používají pro generování variant pořadí daným terminálů a neterminálů

hlavní kombinatorické konstrukty:

- `order()` generuje všechny možné permutace zadaných komponent
- `first()` argument musí být na prvním místě
- `rhs()` doplní všechny pravé strany svého argumentu

```
/* budu se ptát */
clause ===> order(VBU,R,VRI)

/* který ... */
relclause ===> first(relprongr) rhs(clause)
```


Metagramatika – typy pravidel

- `->` normální CF pravidlo
- `-->` vložit `intersegment` mezi každé dva prvky
- `==>` + kontrola správného pořadí příklonek
- `===>` intersegmenty na začátku a konci RHS, spojky, . . .

```
ss -> conj clause
/* budu muset číst */
futmod --> VBU VOI VI
/* byl bych býval */
cpredcondgr ==> VBL VBK VBLL
/* musím se ptát */
clause ===> VO R VRI
```

`clause` pravidla se zadávají pomocí **pravidlových vzorů**

Metagramatika – globální omezení pořadí

globální omezení pořadí zakazuje některé kombinace pořadí preterminálů

`%enclitic` – které preterminály jsou brány jako **příklonky**
`%order` – zajišťuje dodržení precedence zadaných preterminálů

```
/* jsem, bych, se */
%enclitic = (VB12, VBK, R)

/* byl — četl, ptal, musel */
%order VBL = {VL, VRL, VOL}
```


Metagramatika – generativní konstrukty

skupina výrazů `%list_*` – produkují nová pravidla pro seznamy (s oddělovači/bez oddělovačů, s různými testy na shody, …)

```
/* (nesmim) zapomenout udelat - to forget to do */
%list_nocoord vi_list
vi_list -> VI

%list_coord_case_number_gender np
/* krasny pes a mala kocka - beautiful dog and small cat */
np -> left_modif np
```

koncovky `*_case`, `*_number_gender` and `*_case_number_gender` určují typ shody

Metagramatika – pravidlové vzory

pravidla pro slovesné skupiny – cca 40% všech pravidel metagramatiky
pravidlové vzory `%group` – definují časté skupiny konstrukcí v pravidlech

```
%group verbP={
  V:    verb_rule_schema($@,"(#1)")
        groupflag($1,"head"),
  VR R: verb_rule_schema($@,"(#1 #2)")
        groupflag($1,"head"),
}

%template clause ====> order(RHS)

/* ctu/ptam se - I am reading/I am asking */
clause %> group(verbP) vi_list
  verb_rule_schema($@,"#2")
  depends(getgroupflag($1,"head"), $2)
```


Metagramatika – pravidlové vzory – pokrač.

- předchozí příklad – skupina `verbP` = dvě skupiny preterminálů (`V` a `VR R`) s příslušnými akcemi
- v `clause` expanduje postupně na obě pravé strany
- `(get)groupflag` – odkaz na prvek uvnitř `%group`
- **vzor celého pravidla** – speciální pravidlová šipka `%>`
  `%template` definuje vzor každého pravidla s `%>`

Metagramatika – úrovně pravidel

- používá se pro **ohodnocení** výstupních stromů pro jejich **třídění**
- doplněk trénování na velkých **stromových korpusech** (které nejsou ☹)
- zadané **lingvistou** – specialistou na vývoj gramatiky
- **základní úroveň** – 0, **vyšší úrovně** – méně frekventované fenomeny
- pravidla vyšších úrovní mohou být v průběhu analýzy **zapnuté**/**vypnuté**

```
3:np -> adj_group
  propagate_case_number_gender($1)
```


Gramatika G2 – kontextové akce

- gramatické **testy na shody** – pád, rod, číslo
- **testy na zanoření vedlejších vět** – `test_comma`
- akce pro specifikaci **závislostních hran**
- akce **typové kontroly** logických konstrukcí

```
np -> adj_group np
  rule_schema($@, "lwtx(awtx(#1) and awtx(#2))")
  rule_schema($@, "lwtx([[awt(#1),#2],x])")
```

`rule_schema` – schéma pro tvorbu logické konstrukce ze subkonstrukcí
projdou jenom kombinace, které **typově vyhovují** danému schématu

Expandovaná gramatika G3

- překlad testů na shody do CF pravidel
- v češtině – 7 gramatických pádů, dvě čísla a 4 rody → 56 možných variant pro plnou shodu mezi dvěma prvky

počty pravidel

metagramatika G1	253
gramatika G2	3091
expandovaná gramatika G3	11530

Systém synt – příklad logické analýzy

vyhodnocení `rule_schema` pro np '`pečené kuře`'

```
4, 6, -npnl -> . left_modif  np .: k1gNnSc145
agree_case_number_gender_and_propagate OK
rule_schema: 2 nterms, 'lwtx(awtx(#1) and awtx(#2))'
And constrs, Abstr and Exi vars are just gathered
1 (1x1) constructions:
```

$\lambda w_2 \lambda t_3 \lambda x_4([\textbf{pečený}_{w_2 t_3}, x_4] \ \wedge\ [\textbf{kuře}_{w_2 t_3}, x_4]) \ldots (o\iota)_{\tau\omega}$

```
And constrs: none added
Exi vars: none added
```

Systém synt – příklad logické analýzy – pokrač.

vyhodnocení `verb_rule_schema` pro celou `clause`

verb_rule_schema: 3 groups
no acceptable subject found: supplying an inexplicit one
inexplicit subject: k3xPgMnSc1,k3xPgInSc1: $On \ldots \iota$
Clause valency list: jíst <v>#1:(1)hA-#2:(2)hPTc1,   ...
Verb valency list: jíst <v>#2:hH-#1:hPTc4ti
Matched valency list: jíst <v>#2:(1)hH-#1:(2)hPTc4ti
time span: $\lambda t_{12}\mathbf{dnes}_{tt_{12}} \ldots (o\tau)$
frequency: $\mathbf{Onc} \ldots ((o(o\tau))\pi)_\omega$
verbal object: $x_{15} \ldots (o(o\pi)(o\pi))$
present tense clause:
$\lambda w_{17}\lambda t_{18}(\exists i_{10})(\exists x_{15})(\exists i_{16})([\mathbf{Does}_{w_{17}t_{18}}, On, [\mathbf{Imp}_{w_{17}}, x_{15}]] \ \wedge \ [\mathbf{večeře}_{w_{17}t_{18}}, i_{10}] \ \wedge$
$[\mathbf{pečený}_{w_{17}t_{18}}, i_{16}] \ \wedge \ [\mathbf{kuře}_{w_{17}t_{18}}, i_{16}] \ \wedge \ x_{15} =$
$[\mathbf{jíst}, i_{16}]_{w_{17}} \ \wedge \ [[\mathbf{k}_{w_{17}t_{18}}, i_{10}]_{w_{17}}, x_{15}]) \ldots \pi$
clause:
$\lambda w_{19}\lambda t_{20}[\mathbf{P}_{t_{20}}, [\mathbf{Onc}_{w_{19}}, \lambda w_{17}\lambda t_{18}(\exists i_{10})(\exists x_{15})(\exists i_{16})([\mathbf{Does}_{w_{17}t_{18}}, On, [\mathbf{Imp}_{w_{17}}, x_{15}]] \wedge$
$[\mathbf{večeře}_{w_{17}t_{18}}, i_{10}] \ \wedge \ [\mathbf{pečený}_{w_{17}t_{18}}, i_{16}] \ \wedge \ [\mathbf{kuře}_{w_{17}t_{18}}, i_{16}] \ \wedge \ x_{15} =$
$[\mathbf{jíst}, i_{16}]_{w_{17}} \ \wedge \ [[\mathbf{k}_{w_{17}t_{18}}, i_{10}]_{w_{17}}, x_{15}])], \lambda t_{12}\mathbf{dnes}_{tt_{12}}] \ldots \pi$