PLIN034 Algoritmický popis syntaxe

5. Gramatické formalismy

Jakub Machura

Masarykova univerzita
Ústav českého jazyka
machura@phil.muni.cz

podzim 2020

GRAMATICKÉ FORMALISMY

- velké množství přístupů
- nejrozšířenější gramatiky:
    - závislostní
    - kategoriální
    - stromové
    - lexikální funkční
    - gramatiky příznakových struktur

ZÁVISLOSTNÍ FORMALISMY

- vhodné pro popis jazyků s volným slovosledem
- vztah závislosti mezi řídícími a závislými větnými členy
- neexistují žádné neterminály, pouze lexikalizované uzly
- využití valence nebo subkategorizace
- typicky vztah mezi slovesem a jeho možnými doplněními:

  NOSIT

  = koho|co

  = komu & koho|co

ZÁVISLOSTNÍ FORMALISMY

- FGP/FGD (Functional Generative Description)
- UDG, Unification Dependency Grammar – Maxwell
- MTT, Meaning- Text Theory – Mel'čuk
- WG, Word Grammar – Hudson
- Lexicase – Starosta
- FG, Functional Grammar, Dik
- LG, Link Grammar – Temperley, Carnegie Mellon University (http://www.link.cs.cmu.edu/link/)
- DUG, Dependancy Unification Grammar – Halliday

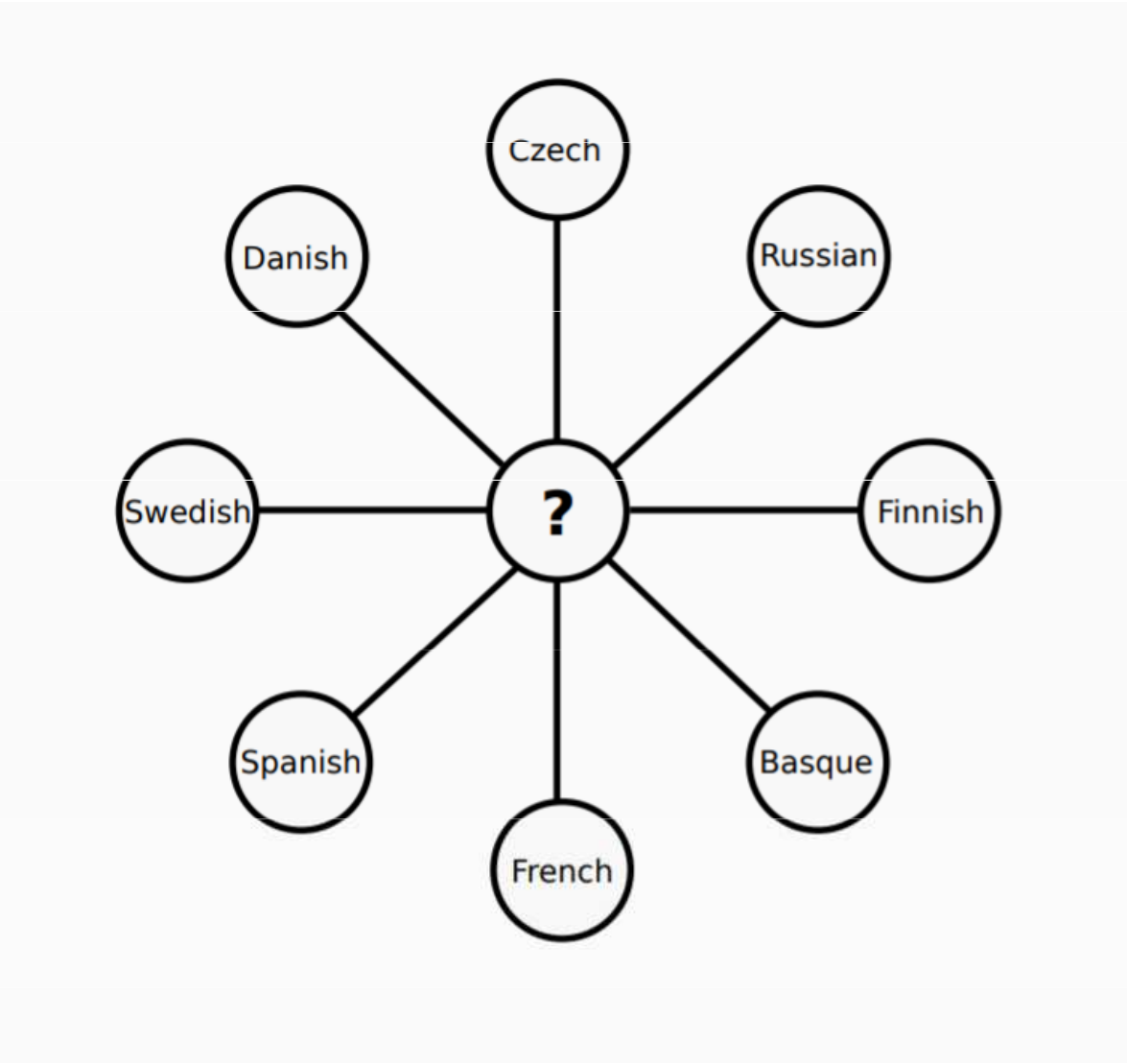
UNIVERSAL DEPENDENCIES

- universaldependencies.org
- „Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing more than 150 treebanks in 90 languages.“
- společná pravidla, jednotný manuál
- snaha vytvořit jednotný systém pro anotaci jakéhokoli (lidského) jazyka tak, aby bylo možné jazyky vzájemně srovnávat.

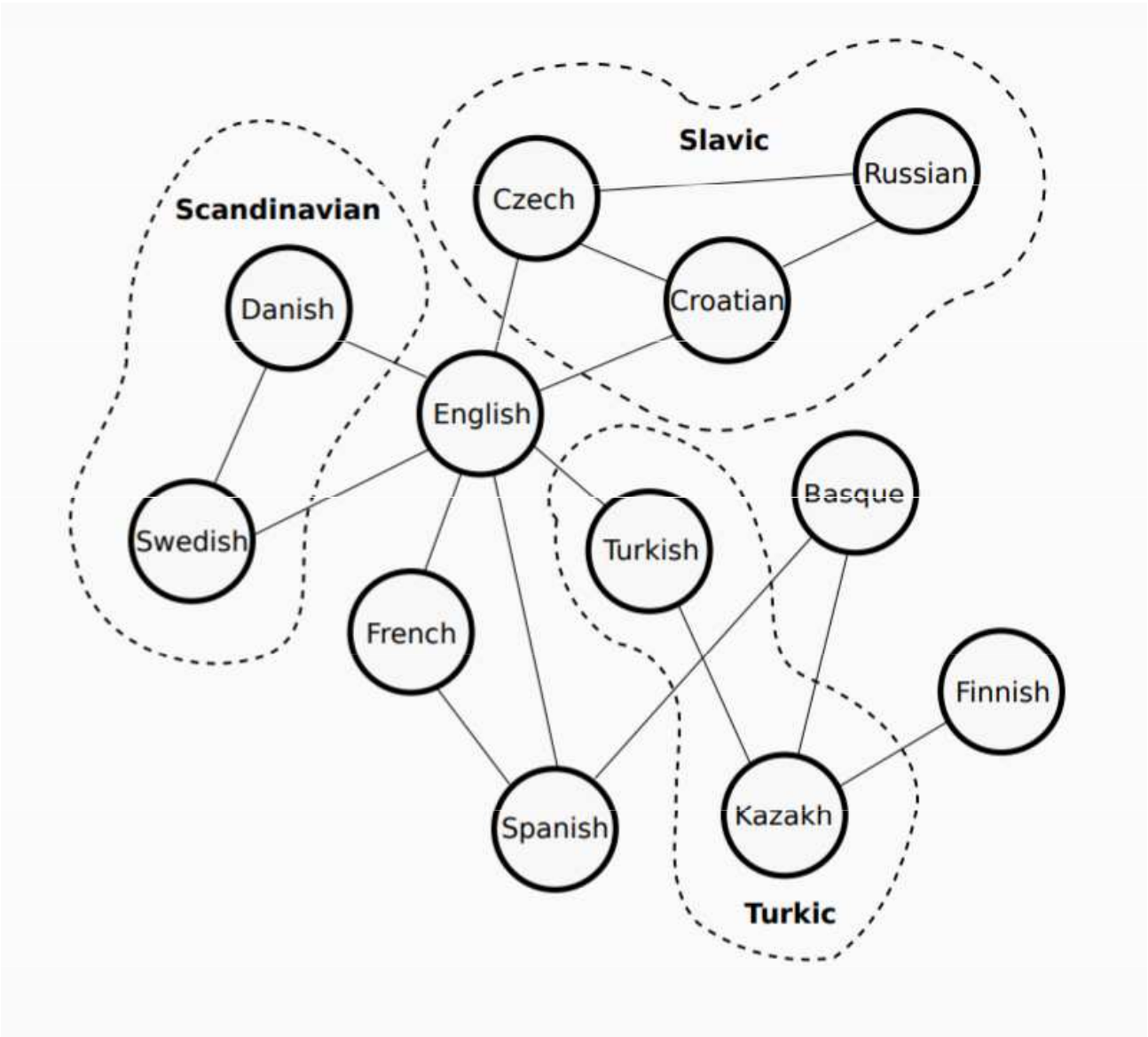
UNIVERSAL DEPENDENCIES

- snaha vytvořit jednotný systém pro anotaci jakéhokoli (lidského) jazyka tak, aby bylo možné jazyky vzájemně srovnávat
- společná pravidla, jednotný manuál
- nutné zjednodušení, částečná ztráta informace

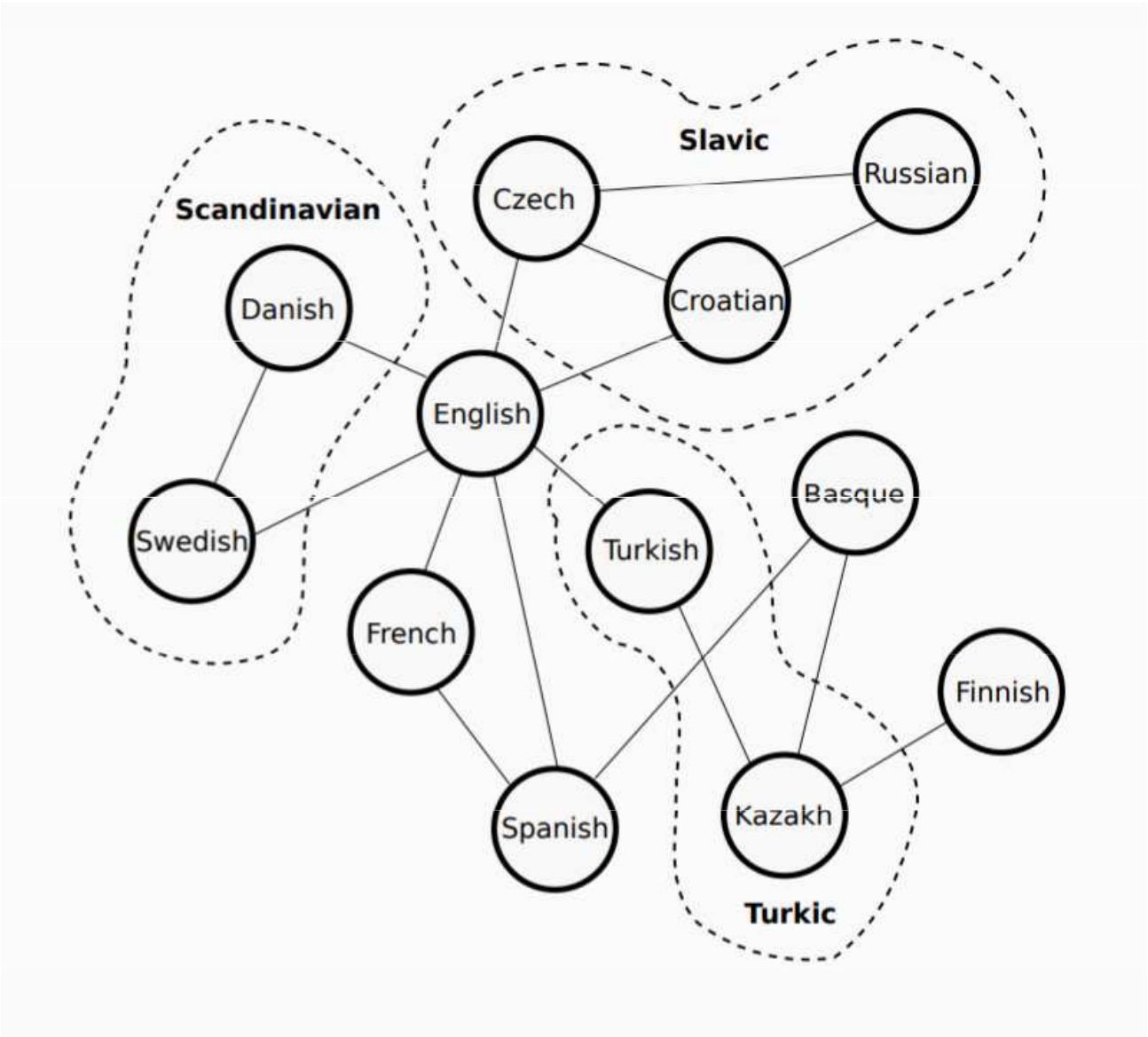
UNIVERSAL DEPENDENCIES



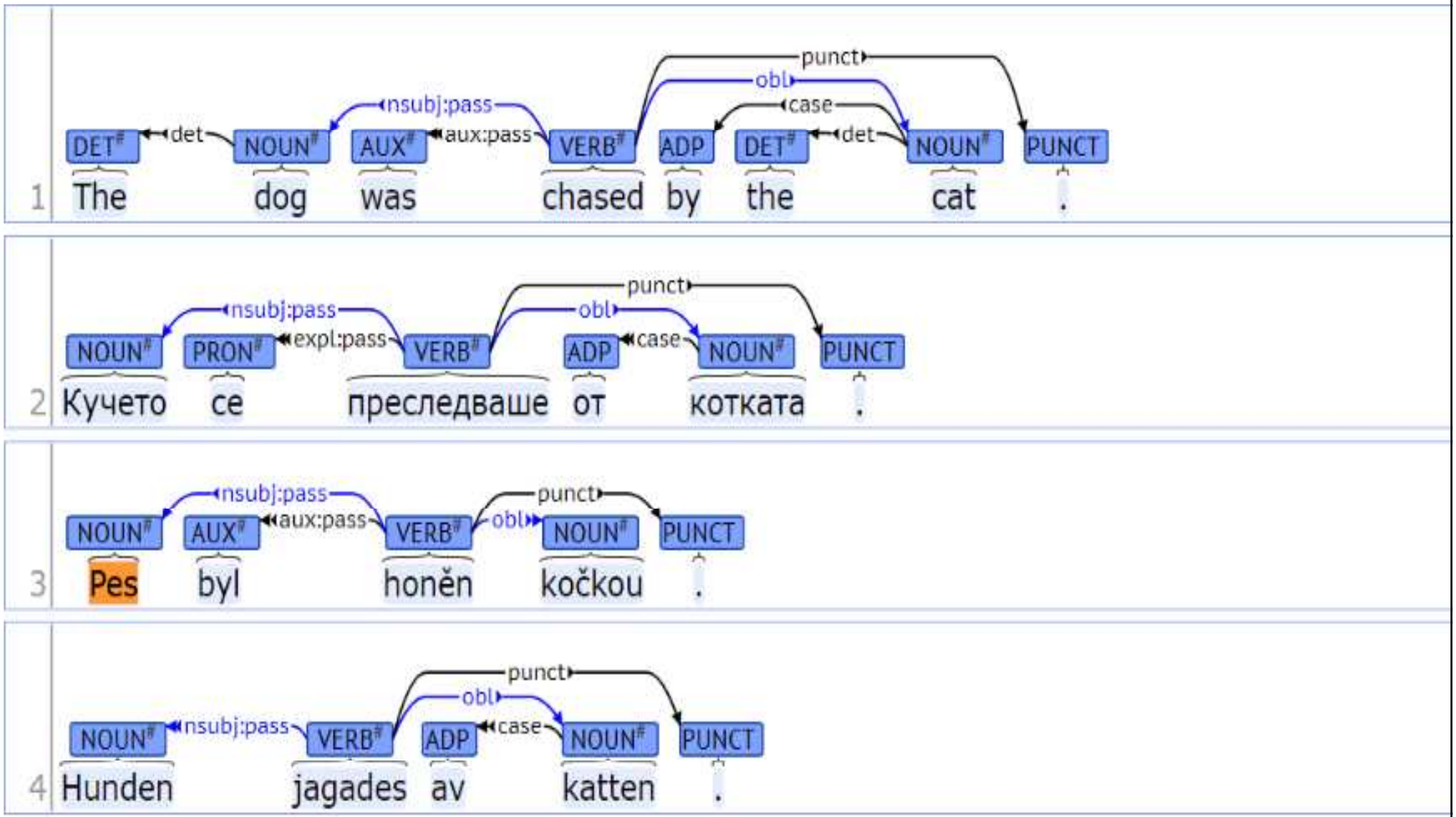
UNIVERSAL DEPENDENCIES



UNIVERSAL DEPENDENCIES



UNIVERSAL DEPENDENCIES



UNIVERSAL POS TAGS

- gramatiky pro jednotlivé jazyky založené na podobných principech
- detaily značkování ale často nejsou převoditelné 1:1
- sjednocení – minimalistická Google Universal Tagset

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Universal Features

- značky z Universal Tagset vymezují základní třídy
- lexikální a gramatické vztahy popisují Universal Features

UNIVERSAL DEPENDENCIES

1	Správkyně	Správkyně	NOUN	Case=Nom|Gender=Fem|Number=Sing|Polarity=Pos
2	dědictví	dědictví	NOUN	Case=Gen|Gender=Neut|Number=Sing|Polarity=Pos
3	Nováková	Nováková	PROPN	Case=Nom|Gender=Fem|NameType=Sur|Number=Sing|Polarity=Pos
4	označila	označit	VERB	Aspect=Perf|Gender=Fem,Neut|Number=Plur,Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act
5	pondělní	pondělní	ADJ	Case=Acc|Degree=Pos|Gender=Neut|Number=Sing|Polarity=Pos
6	rozhodnutí	rozhodnutí	NOUN	Case=Acc|Gender=Neut|Number=Sing|Polarity=Pos
7	za	za	ADP	AdpType=Prep|Case=Acc
8	potěšující	potěšující	ADJ	Aspect=Imp|Case=Acc|Gender=Neut|Number=Sing|Polarity=Pos|Tense=Pres|VerbForm=Part|Voice=Act
9	.	.	PUNCT	_

JAZYKOVÉ INSTRUKCE PRO UD

- každý jazyk má uvedené instrukce pro:
  - tokenizaci (hranici slov)
  - morfologické značky
  - syntax – základní a rozšířené závislosti

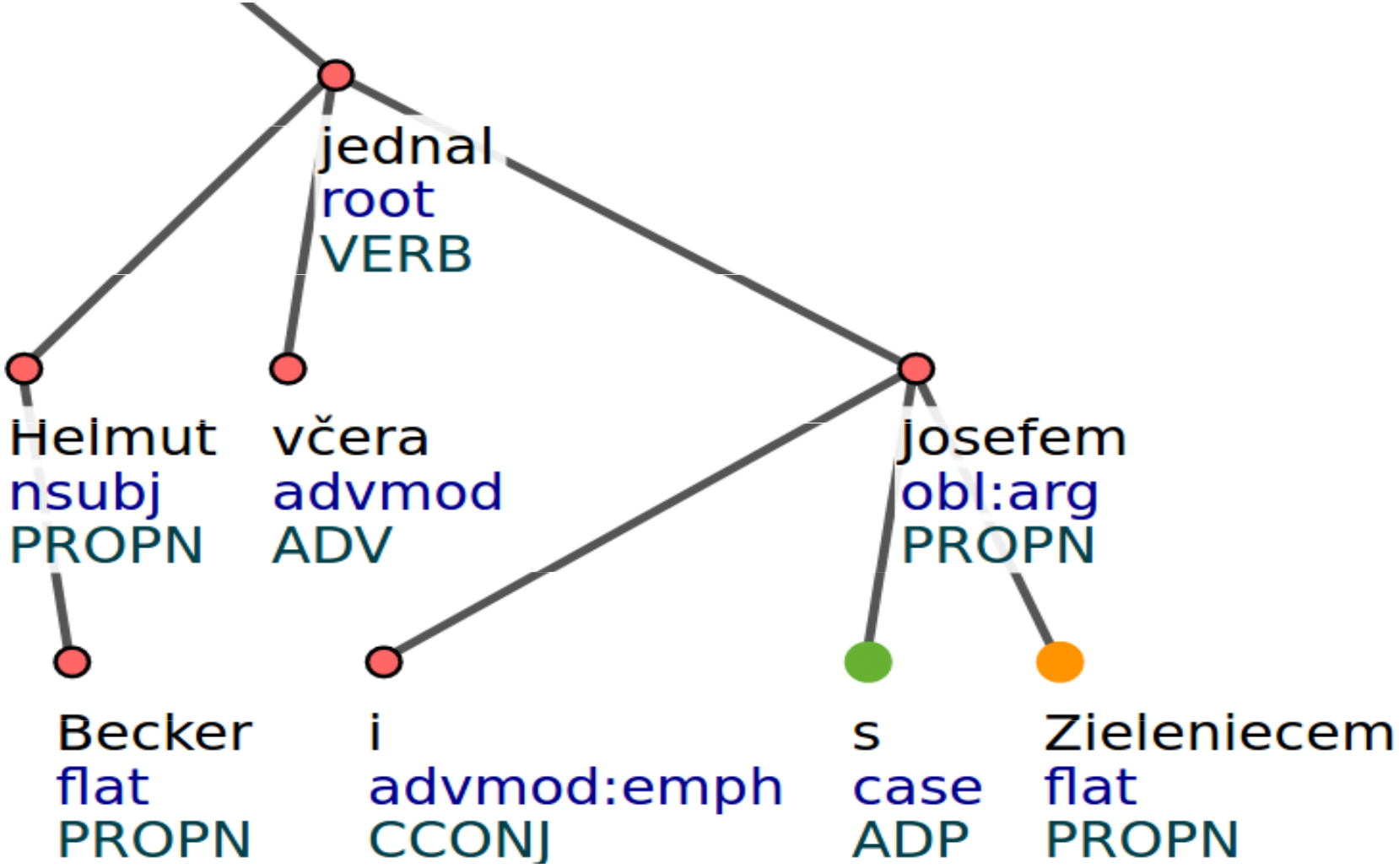
- pro češtinu: www.universaldependencies.org/cs/
- cíl instrukcí: sjednocení anotací napříč jazyky
- obsahuje i instrukce netypické pro jazyk – např. v češtině značkování některých zájmen jako determiner

  nebo expandování slov – kdybych = když + bych

SYNTAKTICKÁ ANOTACE V UD

- dává se přednost vztahům mezi plnovýznamovými slovy, funkční slova jsou upozaděna
- obsahová (plnovýznamová) slova jsou v závislostní struktuře primární, funkční slova jsou na nich závislá, na funkčních slovech nemohou být závislá žádná jiná slova
- syntaktické funkce patří do seznamu Universal Dependency Relations (http://universaldependencies.org/u/dep/index.html)
- koordinace je asymetrická (první člen koordinace reprezentuje celou koordinaci, další členy a spojka jsou závislé na něm)

SYNTAKTICKÁ ANOTACE V UD



SYNTAKTICKÁ ANOTACE V UD

They buy and sell books.

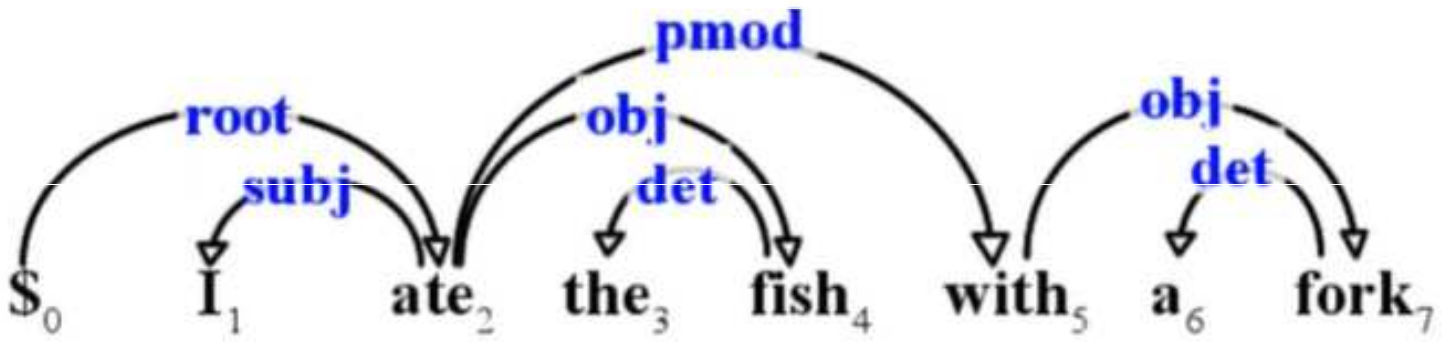
1	They	they	PRON	PRP	Case=Nom|Number=Plur	2	nsubj	2:nsubj|4:nsubj
2	buy	buy	VERB	VBP	Number=Plur|Person=3|Tense=Pres	0	root	0:root
3	and	and	CONJ	CC	_	4	cc	4:cc
4	sell	sell	VERB	VBP	Number=Plur|Person=3|Tense=Pres	2	conj	0:root|2:conj
5	books	book	NOUN	NNS	Number=Plur	2	obj	2:obj|4:obj
6	.	.	PUNCT	.	_	2	punct	2:punct

VYUŽITÍ UNIVERSAL DEPENDENCIES

- srovnání lingvistických fenoménů napříč jazyky
- testování syntaktické analýzy na různých jazycích
- vícejazyčná syntaktická analýza – paralelní dokumenty
- snadné porozumění rozdílům v anotaci

STROJOVÉ UČENÍ A ZÁVISLOSTNÍ FORMALISMUS

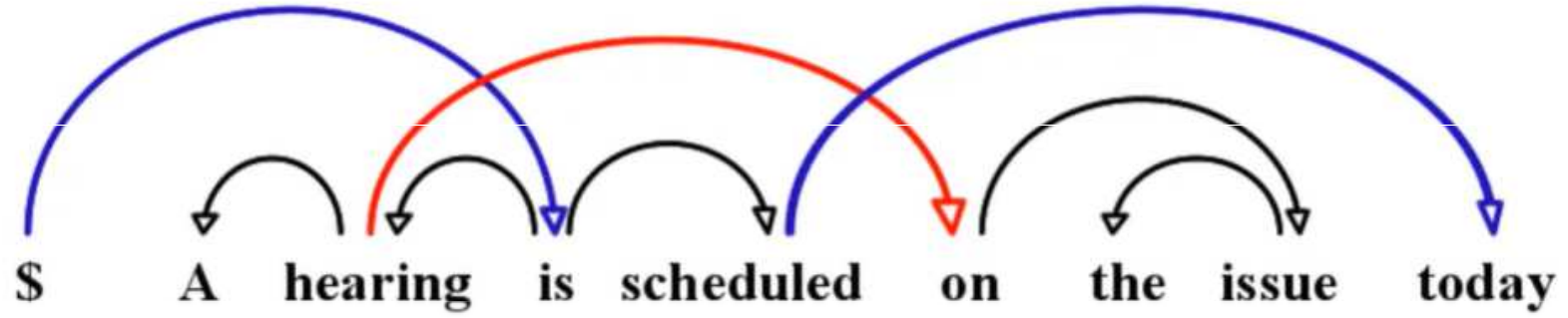
- jedna hrana pro každé slovo
- 2 až 3 informace pro učení



- head
- dependant
- type = edge label

STROJOVÉ UČENÍ A ZÁVISLOSTNÍ FORMALISMUS

- problém u neprojektivních konstrukcí



Example from "Dependency Parsing" by Kübler, Nivre, and McDonald, 2009

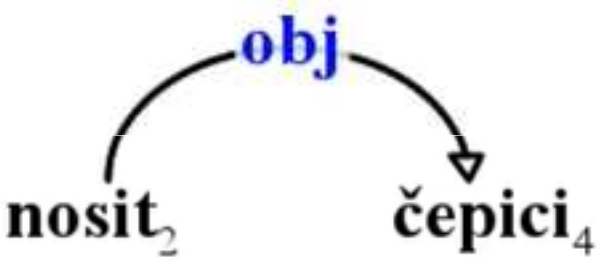
VYHODNOCENÍ

- velmi lehce rozpoznáme, jestli je vazba správná, porovnáme s treebankem
- ale jak moc je správná, resp. špatná?

http://universaldependencies.org/udw17/pdf/UDW11.pdf

VYHODNOCENÍ

- 4 metriky:



UAS – Unlabeled attachement score – words with correct head

LAS – Labeled attachement score – words with correct head and type

RA – Root Accuracy – analysis with correct root

CM – Complete Match rate – fully correct analyses

TAG, LTAG

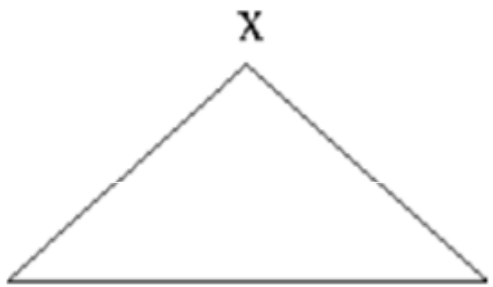
- formalismus založen na syntaxi frázové struktury
- Tree Adjoining Grammar – Joshi, Levy a Takahashi (*TAG formalism*, 1975)
- Lexicaloized TAG – Joshi a Schabes (1991)
- pracuje se přímo se stromy, a ne s řetězci slov

TAG

- stavební prvky analýzy nejsou slova a neterminály, ale částečně specifikované syntaxové stromy které podléhají několika přípustným stromovým operacím
- množina počátečních stromů – základní stavební prvky
- složitější věty odvozovány s použitím pomocných stromů

TAG

počáteční (*initial*) strom:



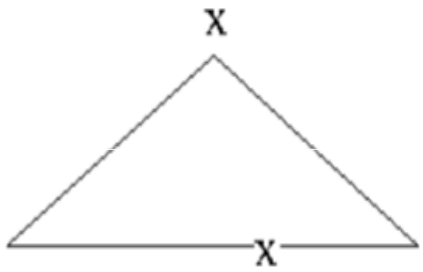
počáteční strom typu X = jeho kořen je označen termem X

- neobsahují rekurzi – popisují složkovou strukturu jednoduchých vět, jmenných skupin, předložkových skupin…

1. všechny nelistové uzly odpovídají neterminálům
2. všechny listové uzly odpovídají terminálům nebo neterminálům určeným k substituci

TAG

pomocný (*auxiliary*) strom:

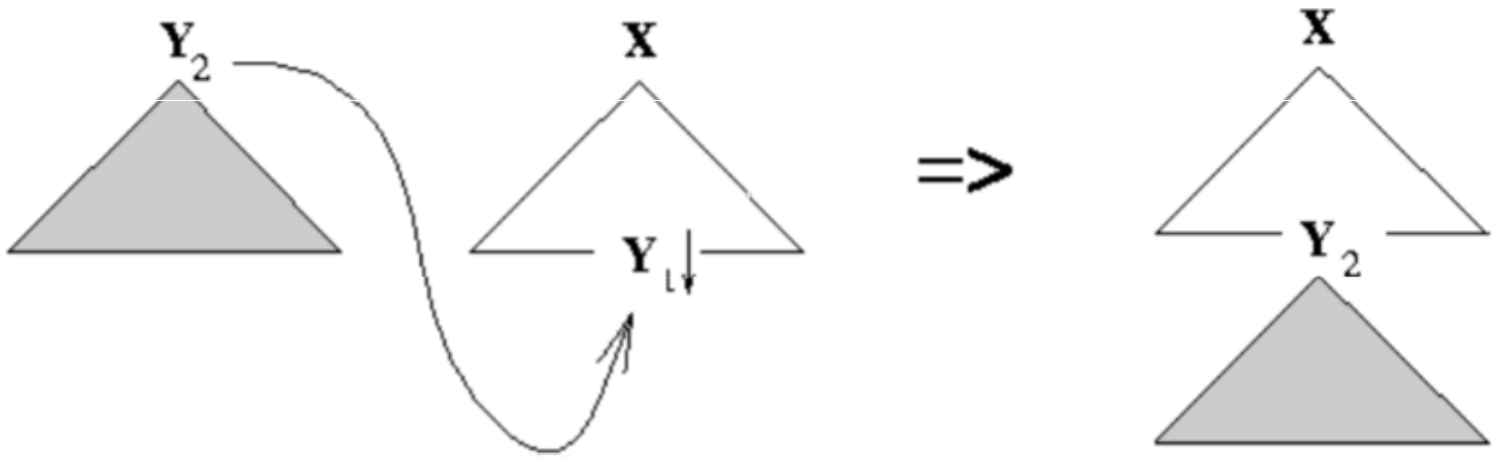


- reprezentují rekurzivní stromy, popisují větné struktury, které se připojují k základním strukturám (např. příslov. určení)

1. všechny nelistové uzly odpovídají neterminálům
2. všechny listové uzly odpovídají terminálům nebo neterminálům určeným k substituci kromě právě jednoho neterminálního uzlu (patový uzel, *foot node*)
3. patový uzel má stejné označení jako kořenový uzel (slouží k připojení stromu k jinému uzlu)

TAG

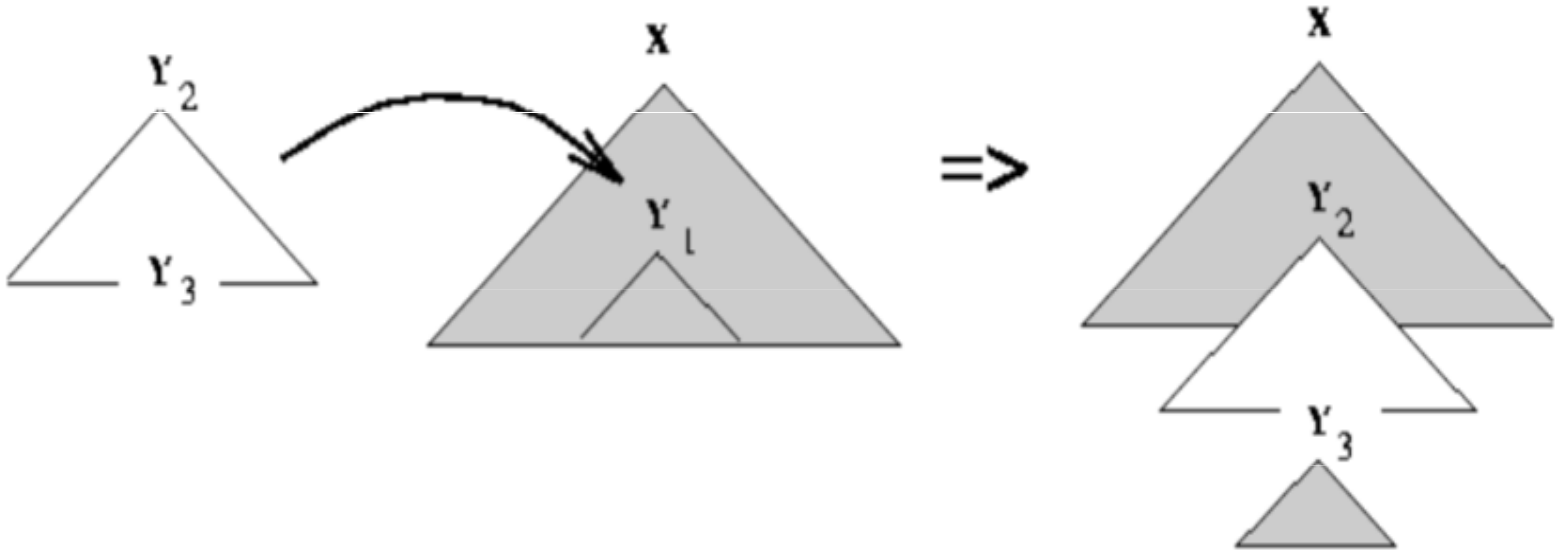
dvě operace – substituce a připojení (*adjunction*)
operace substituce – nahrazuje označený neterminál v listech nějakého stromu stromem, jehož kořen nese stejné označení



$Y_1\downarrow$ – označený pro substituci

TAG

operace připojení – vložení pomocného stromu, popisujícího rekurzi neterminálu X, se stromem, který obsahuje uzel označený rovněž X



DEFINICE TAG

- TAG G = (I, A, S) je:
  - množina I konečných počátečních stromů
  - množina A pomocných stromů
  - typ stromu S – neterminál označující větu

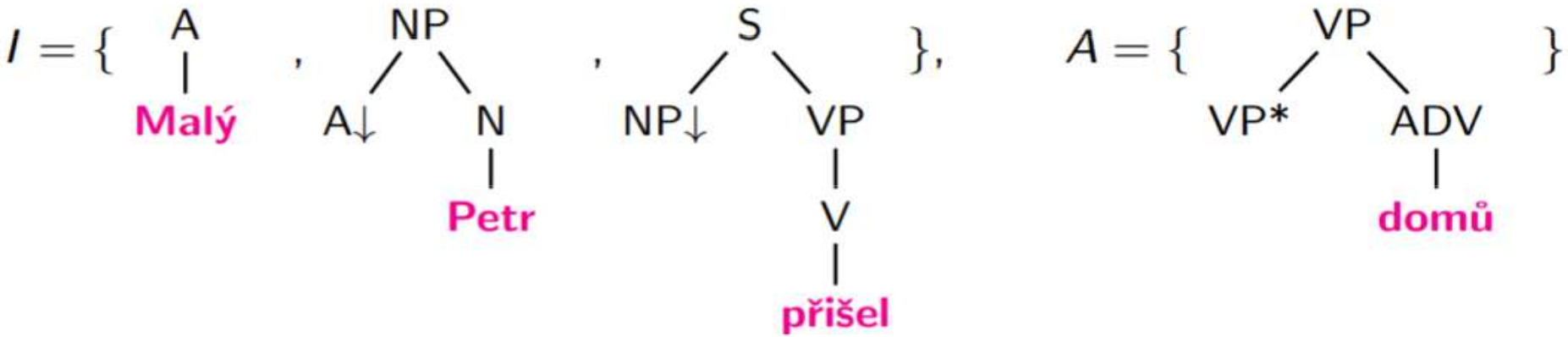
- množina stromů $\mathcal{T}(G)$ TA gramatiky G = množina všech stromů odvoditelných z počátečních stromů typu S z I, jejichž spodní okraj sestává čistě z terminálních uzlů (všechny substituční uzly byly doplněny)
- jazyk řetězců $\mathcal{L}(G)$ generovaných TA gramatikou G = množina všech terminálních řetězců na spodním okraji stromů v $\mathcal{T}(G)$.

LEXIKALIZACE TAG

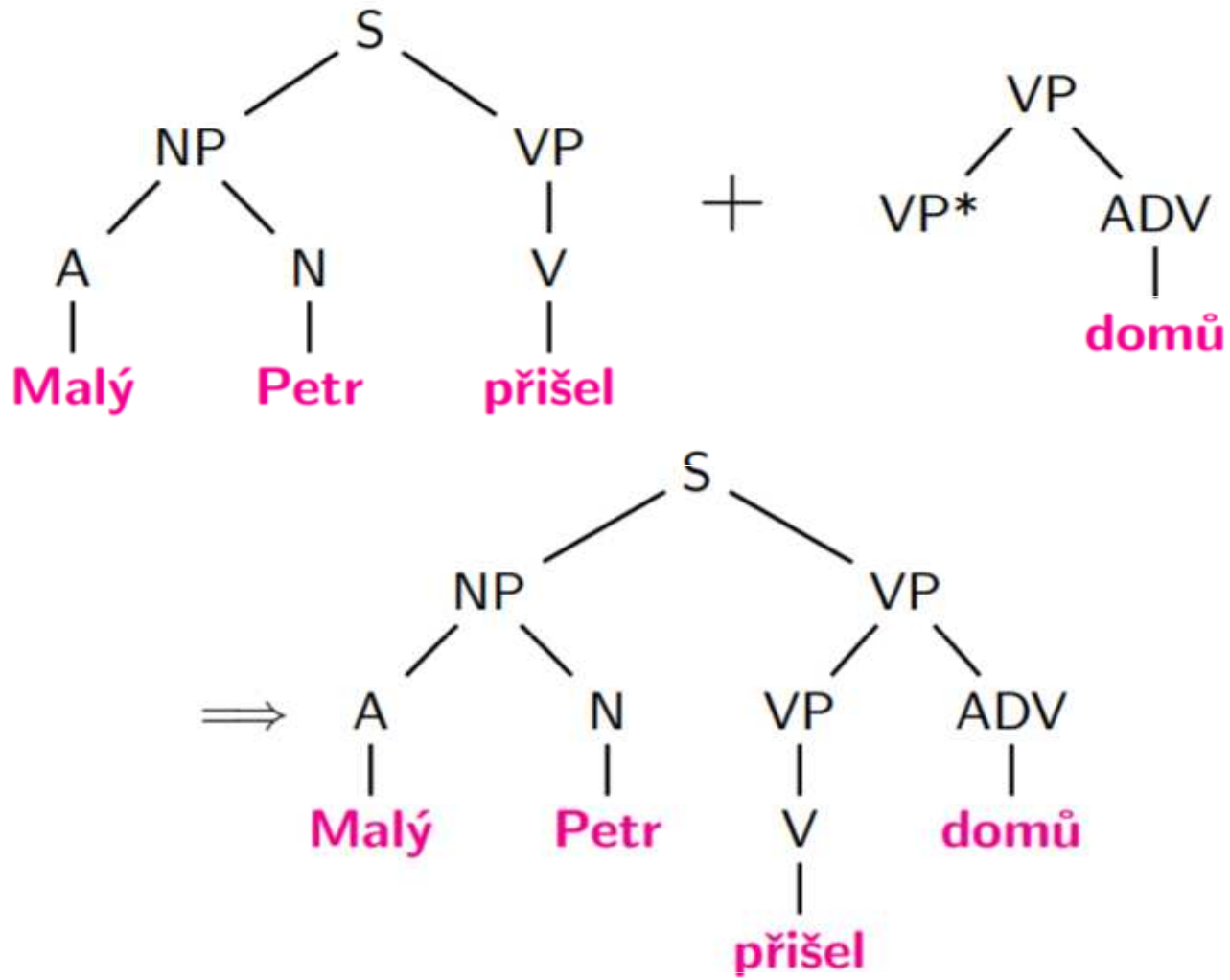
- LTAG je lexikalizovanou variantou formalismu TAG

→ počáteční i pomocné stromy obsahují v listech jednu nebo více tzv. lexikálních kotev – uzly, které jsou přiřazeny (ukotveny) k určitým slovům lexikonu

lexikalizované stromy (*substituční uzly* – ↓, *patové uzly* – *):



LEXIKALIZACE TAG



TAG, LTAG

- díky použití operace připojení mají TAG a LTAG větší generativní sílu než bezkontextové gramatiky

  (CFG $\subset$ MCSL)

  generování mírně kontextových jazyků
  (mildly context-sensitive languages)

- vlastnost konstantního růstu – pokud uspořádáme řetězce jazyka vzestupně podle délky, potom rozdíl dvou po sobě jdoucích délek nemůže být libovolný (každá délka je lineární kombinací konečného počtu pevných délek).
- analyzovatelnost v polynomiálním čase $O(n^6)$ vzhledem k délce vstupu

TAG, LTAG

- i jiné formalismy umí MSCL:
  - LIG, Linear Indexed Grammars – Gazdar, 1985
  - HG, Head Grammar – Pollard, 1984
  - CCG, kombinatorické kategoriální gramatiky

KATEGORIÁLNÍ GRAMATIKY

- categorial grammar, CG
- skupina teorií syntaxe a sémantiky PJ s velkým důrazem na lexikon
- neobsahuje pravidla pro kombinování slov → lexikální kategorie slov tvoří funkce, které určují, jak se dané kategorie kombinují s jinými výrazy a je výsledkem aplikace podvýrazů na sebe

pěkný := NP/N ... funkce, která má argument N a vrací NP

KATEGORIÁLNÍ GRAMATIKY

- opírají se o princip kompozicionality:
  Význam složeného výrazu je jednoznačně určen významy částí tohoto výrazu a způsobem, jakým jsou tyto části složeny dohromady.
- základy kategoriální g. položili ve 30. letech 20. stol. polští logikové Leśniewski a Ajdukiewicz
- hlavní uplatnění nalezly tyto gramatiky v lingvistice zejm. při popisu jazyků s pevným slovosledem (např. angl.).
- první použití CG pro popis PJ: Jehošua Bar-Hillel (1953)

KATEGORIÁLNÍ GRAMATIKY

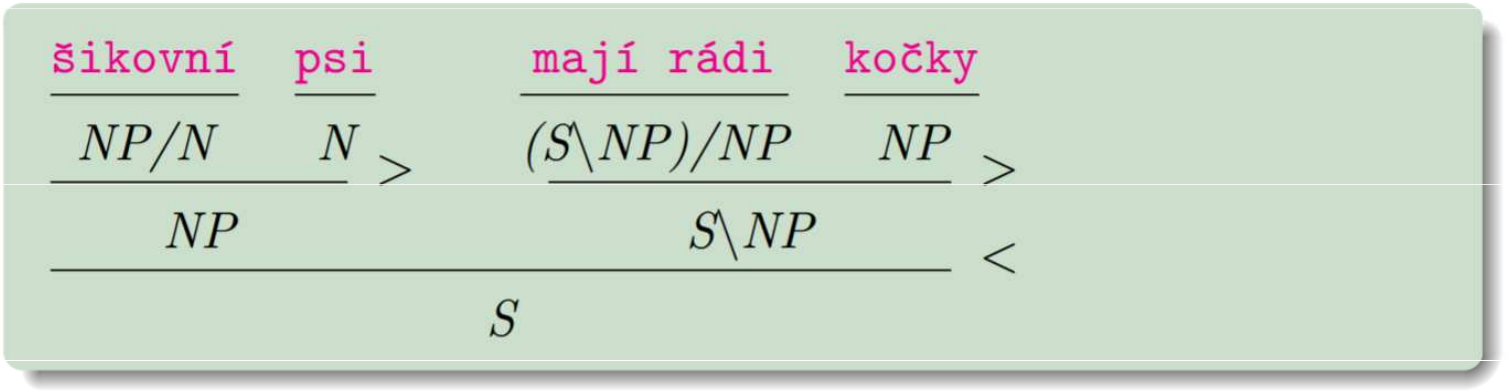
- CG jsou tvořeny
  a) množinou tzv. základních syntaktických kategorií (neterminálních symbolů)
  např. N (podstatné jméno), N\S (intranzitivní sloveso)

- každé kategorii je přiřazena množina slov jazyka (terminálních symbolů) patřících do této kategorie

  b) elementárních operací vytvářejících z těchto kategorií odvozené syntaktické struktury a majících podobu jednostranného „krácení abstraktních zlomků“

KATEGORIÁLNÍ GRAMATIKY

- „krácení zlomků“



- intranzitivní sloveso lze zachytit jako „zlomek“ N/S (nalevo vyžaduje N, podmět): N N/S ⟶ S

KATEGORIÁLNÍ GRAMATIKY

- CG jsou ekvivalentní CFG
- výhoda: rozšířitelné o sémantickou komponentu
- existují ale rozšíření CG, která vedou k systémům s vyšší vyjadřovací silou, než mají standardní CG
- klíčový problém: nespojité větné části (neprojektivita)
- řešení pomocí rozšíření CG – přídavné kombinatorické operátory založené na typech
  - např. CCG (kombinatorické CG) přidává pravidla odpovídající jednoduchým operacím nad kategoriemi

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)

- teorie organizovaná na lexikalistické hypotéze
- Kaplan a Bresnan, 1982
- dva typy syntaktických struktur
  - vnější, složková, c-struktura:
    slovosled a syntaktické složky
  - vnitřní, funkční, f-struktura:
    syntaktické funkce (podmět, předmět apod.)

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)

- jazyky se výrazněji odlišují v organizace fráze, v pořadí a způsobech realizace gram. funkcí
- abstraktnější, funkcionální organizace jazyků se odlišuje daleko méně (např. se běžně objevují funkce podmět, předmět atd.)

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)

- složková struktura – zachycuje frázovou dominanci a prioritu a je reprezentována jako strom frázové struktury (CFG strom)
- funkční struktura – zachycuje syntaktickou strukturu typu predikát-argumenty a je reprezentován ***maticí*** dvojic *atribut-hodnota*
- f-struktura obsahuje:
  - příznaky: čas, rod, číslo…
  - funkce: PRED, SUBJ, OBJ, jejichž hodnoty mohou být jiné f-struktury

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)

Zjednodušená f-struktura věty *Lucka snědla všechny buchty*

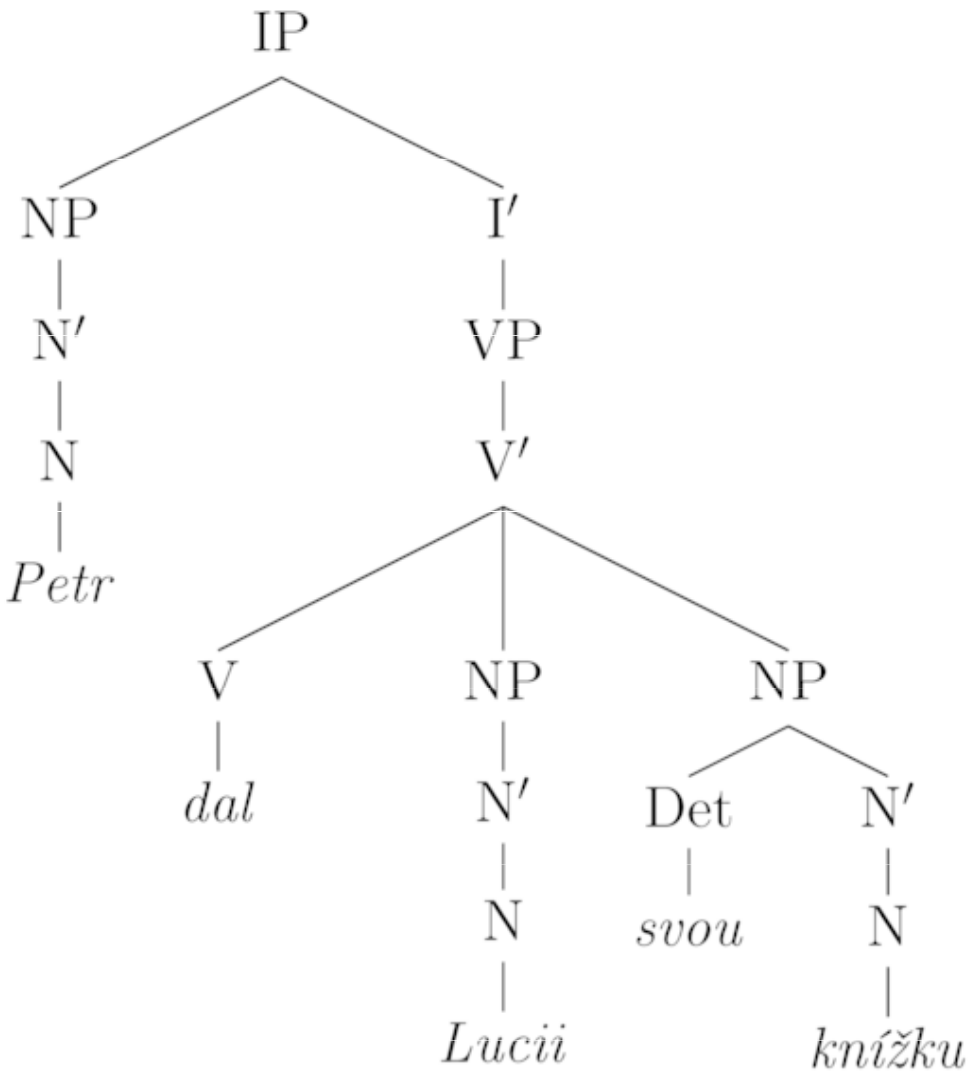
$$\begin{bmatrix} \text{PRED} & \text{`}\mathbf{sníst}\langle \text{SUBJ,OBJ} \rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`}\mathbf{Lucka}\text{'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{SPEC} & \begin{bmatrix} \text{PRED} & \text{`}\mathbf{všechen}\text{'} \end{bmatrix} \\ \text{PRED} & \text{`}\mathbf{buchta}\text{'} \end{bmatrix} \end{bmatrix}$$

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)

Zjednodušená f-struktura věty *Včera v parku Lucka snědla všechny buchty*

$$\begin{bmatrix}
\text{PRED} & \text{`}\mathbf{sníst}\langle\text{SUBJ,OBJ}\rangle\text{'} \\
\text{TENSE} & \text{PAST} \\
\text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{`}\mathbf{Lucka}\text{'}\end{bmatrix} \\
\text{OBJ} & \begin{bmatrix}\text{SPEC} & \begin{bmatrix}\text{PRED} & \text{`}\mathbf{všechen}\text{'}\end{bmatrix} \\ \text{PRED} & \text{`}\mathbf{buchta}\text{'}\end{bmatrix} \\
\text{ADJ} & \left\{\begin{bmatrix}\text{PRED} & \text{`}\mathbf{včera}\text{'}\end{bmatrix}, \begin{bmatrix}\text{PRED} & \text{`}\mathbf{v}\langle\text{OBJ}\rangle\text{'} \\ \text{OBJ} & \begin{bmatrix}\text{PRED} & \text{`}\mathbf{park}\text{'}\end{bmatrix}\end{bmatrix}\right\}
\end{bmatrix}$$

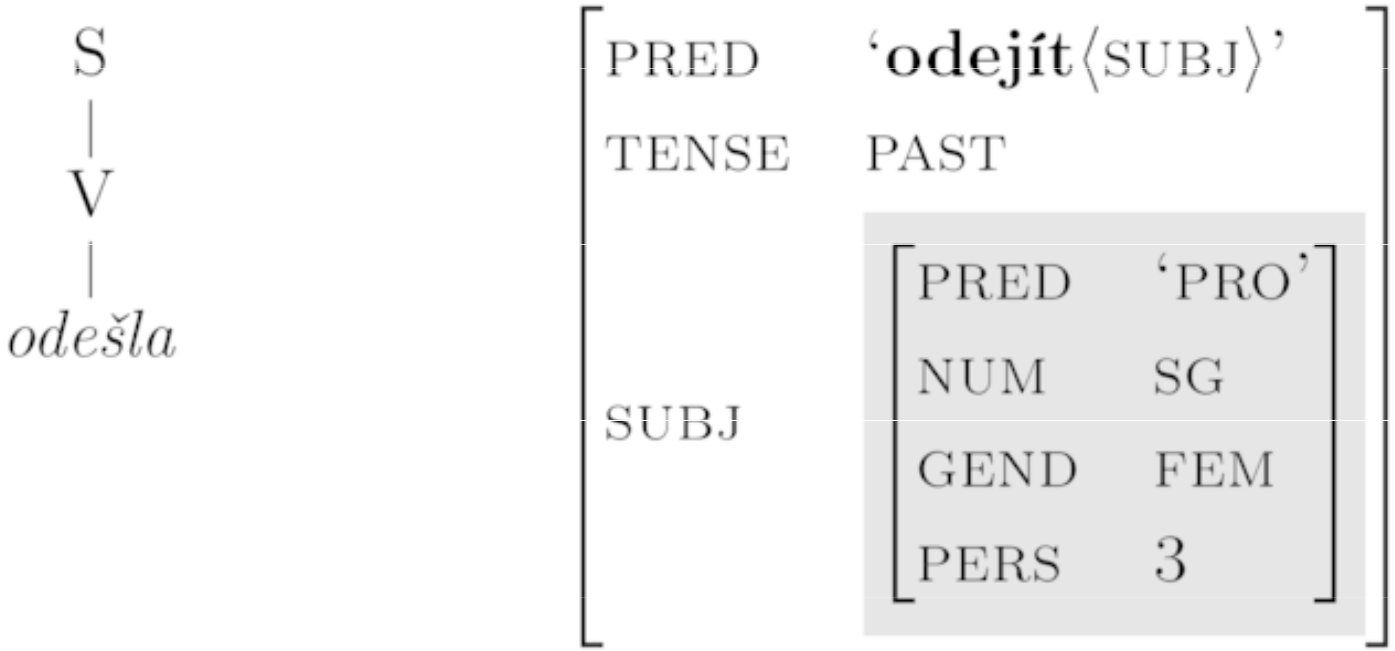
LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)



$$\begin{bmatrix} \text{PRED} & \text{`}\mathbf{dát}\langle \text{SUBJ}, \text{OBJ}, \text{OBJ}_{\text{theme}} \rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & [\ \text{PRED `}\mathbf{Petr}\text{'}\ ] \\ \text{OBJ} & [\ \text{PRED `}\mathbf{Lucie}\text{'}\ ] \\ \text{OBJ}_{\text{theme}} & \begin{bmatrix} \text{SPEC} & [\ \text{PRED `}\mathbf{svůj}\text{'}\ ] \\ \text{PRED} & \text{`}\mathbf{knížka}\text{'} \end{bmatrix} \end{bmatrix}$$

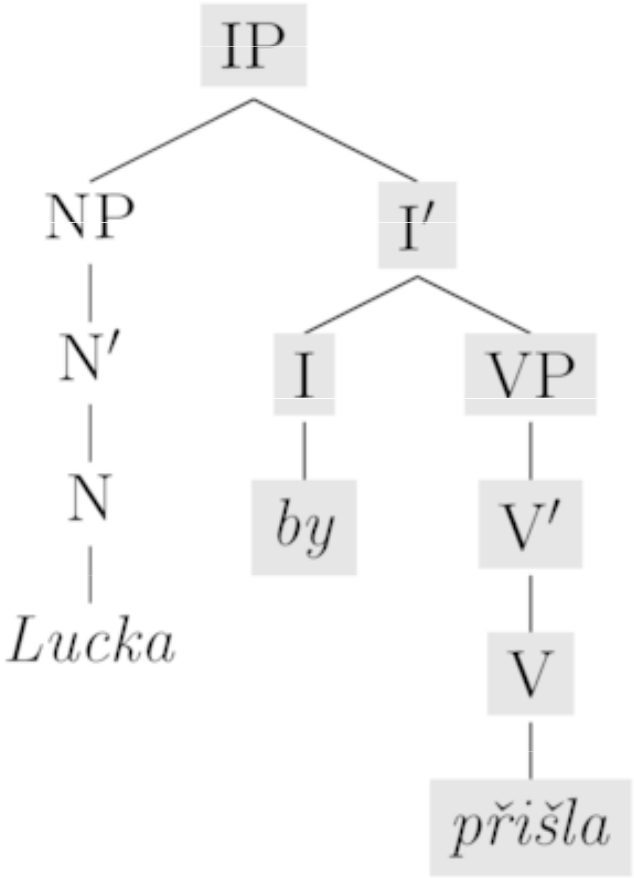
Vztah mezi c-strukturou a f-strukturou

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)



F-struktura nemusí odpovídat uzlu c-struktury

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)



$$\begin{bmatrix} \text{PRED} & \text{‘}\mathbf{přijít}\langle\text{SUBJ}\rangle\text{’} \\ \text{TENSE} & \text{PRES} \\ \text{MOOD} & \text{COND} \\ \text{SUBJ} & [\ \text{PRED}\ \text{‘}\mathbf{Lucka}\text{’}\ ] \end{bmatrix}$$

Více uzlů c-struktury může odpovídat jedné f-struktuře

LEXIKÁLNÍ FUNKČNÍ GRAMATIKY (LFG)

příklady:

$$\begin{array}{lccc} S \rightarrow & NP & & VP \\ & (\uparrow \text{SUBJ}) = \downarrow & & \uparrow = \downarrow \end{array}$$

$$\begin{array}{lccc} VP \rightarrow & V & & (NP) \\ & \uparrow = \downarrow & & (\uparrow \text{OBJ}) = \downarrow \end{array}$$

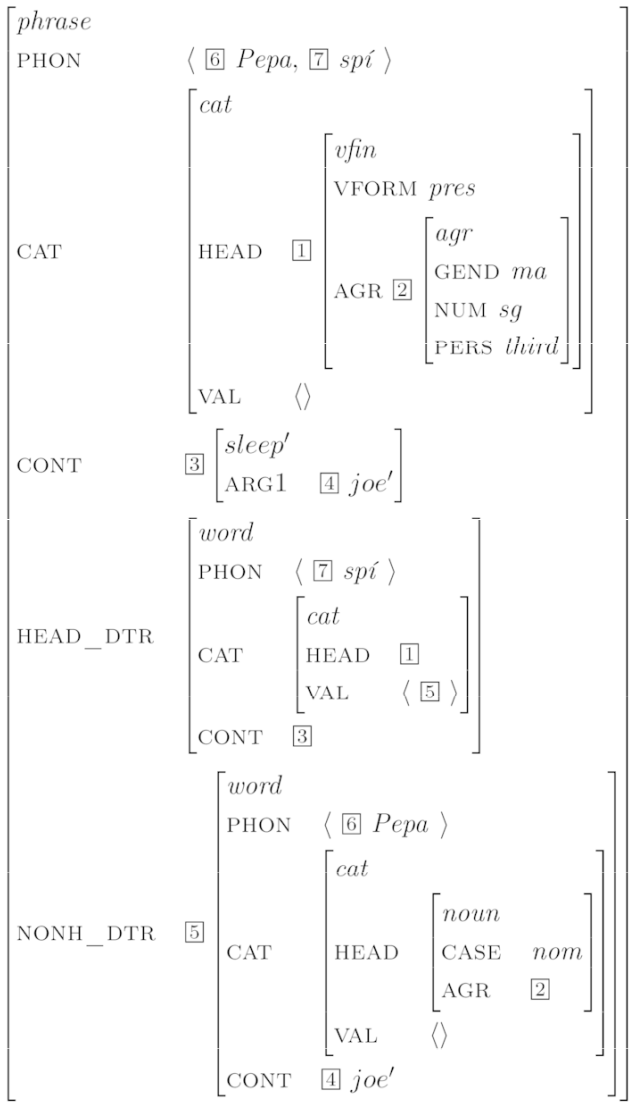
$$\begin{array}{lccc} NP \rightarrow & (DET) & & N \\ & \uparrow = \downarrow & & \uparrow = \downarrow \end{array}$$

výrazy $(\uparrow \text{SUBJ}) = \downarrow$, $\uparrow = \downarrow$ a $(\uparrow \text{OBJ}) = \downarrow$ jsou *funkční schémata*

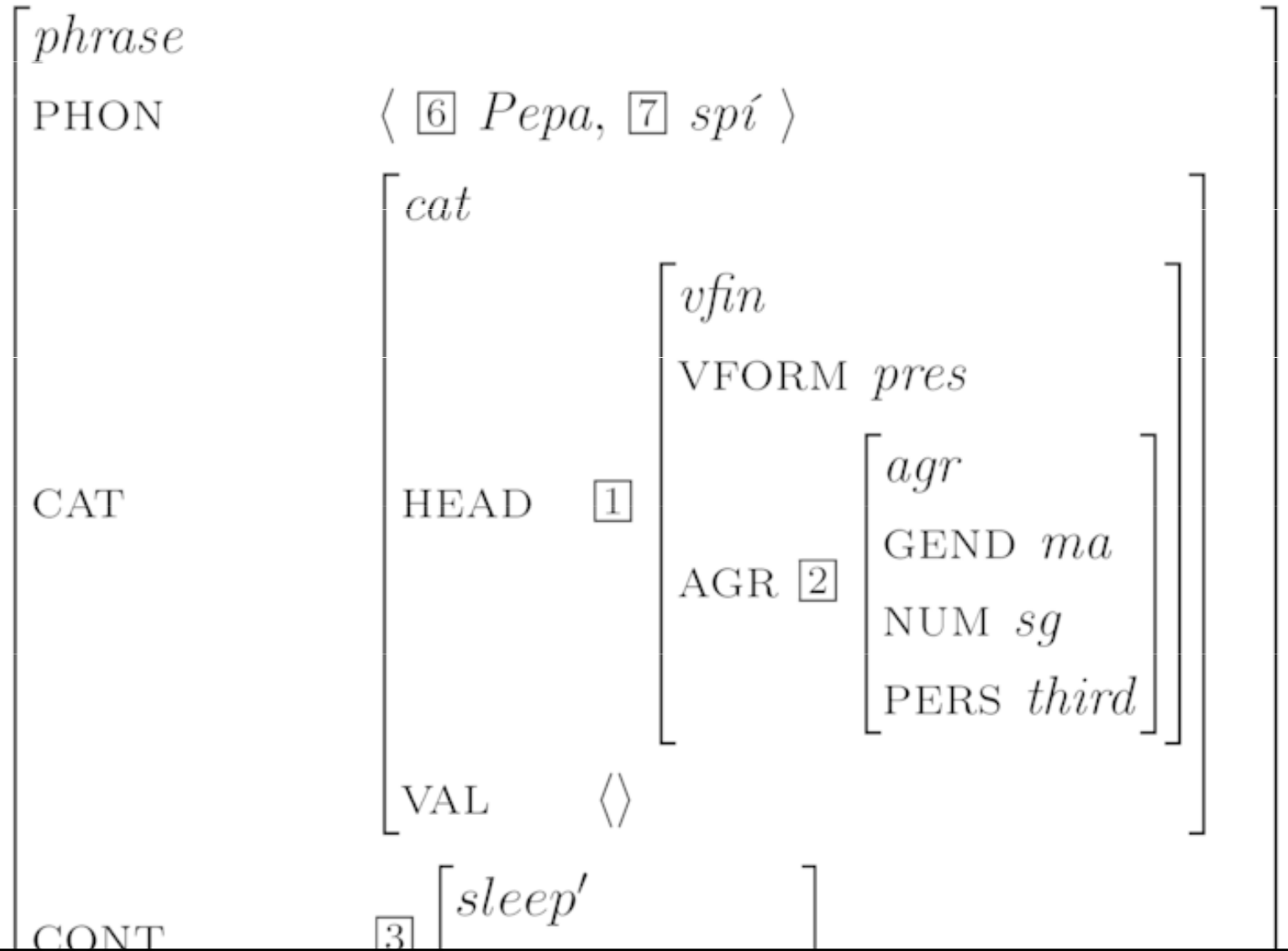
HPSG (Head-Driven Phrase Structure Grammar)

- Pollard & Sag, 1987, 1994
- navazuje na Generalized Phrase Structure Grammar (1985)
- lexikalismus
- gramatika jako celek obsahující pravidla i slovník
- analýza libovolného výrazu (věta, syntagma i slovo) na více rovinách popisu jazyka současně (včetně fonologie, morfologie, sémantiky i pragmatiky), a to jako jazykový znak
- gramatika je deklarativní, ne derivační: popisuje stav, ne operace vedoucí k výsledku

HPSG (HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR)



HPSG (Head-Driven Phrase Structure Grammar)



$$
\begin{bmatrix}
\text{CAT} & \begin{bmatrix}
cat & \\
\text{HEAD} & \boxed{1} \begin{bmatrix} vfin & \\ \text{VFORM} & pres \\ \text{AGR} & \boxed{2} \begin{bmatrix} agr & \\ \text{GEND} & ma \\ \text{NUM} & sg \\ \text{PERS} & third \end{bmatrix} \end{bmatrix} \\
\text{VAL} & \langle\rangle
\end{bmatrix} \\
\text{CONT} & \boxed{3} \begin{bmatrix} sleep' & \\ \text{ARG1} & \boxed{4}\ joe' \end{bmatrix} \\
\text{HEAD\_DTR} & \begin{bmatrix}
word & \\
\text{PHON} & \langle\ \boxed{7}\ spí\ \rangle \\
\text{CAT} & \begin{bmatrix} cat & \\ \text{HEAD} & \boxed{1} \\ \text{VAL} & \langle\ \boxed{5}\ \rangle \end{bmatrix} \\
\text{CONT} & \boxed{3}
\end{bmatrix} \\
 & \begin{bmatrix} word & \\ \text{PHON} & \langle\ \boxed{6}\ Pepa\ \rangle \end{bmatrix}
\end{bmatrix}
$$

$$
\begin{bmatrix}
\text{CONT} & \boxed{3}\begin{bmatrix} sleep' & \\ \text{ARG1} & \boxed{4}\ joe' \end{bmatrix} \\
\text{HEAD\_DTR} & \begin{bmatrix} word & \\ \text{PHON} & \langle\ \boxed{7}\ spí\ \rangle \\ \text{CAT} & \begin{bmatrix} cat & \\ \text{HEAD} & \boxed{1} \\ \text{VAL} & \langle\ \boxed{5}\ \rangle \end{bmatrix} \\ \text{CONT} & \boxed{3} \end{bmatrix} \\
\text{NONH\_DTR} & \boxed{5}\begin{bmatrix} word & \\ \text{PHON} & \langle\ \boxed{6}\ Pepa\ \rangle \\ \text{CAT} & \begin{bmatrix} cat & \\ \text{HEAD} & \begin{bmatrix} noun & \\ \text{CASE} & nom \\ \text{AGR} & \boxed{2} \end{bmatrix} \\ \text{VAL} & \langle\rangle \end{bmatrix} \\ \text{CONT} & \boxed{4}\ joe' \end{bmatrix}
\end{bmatrix}
$$

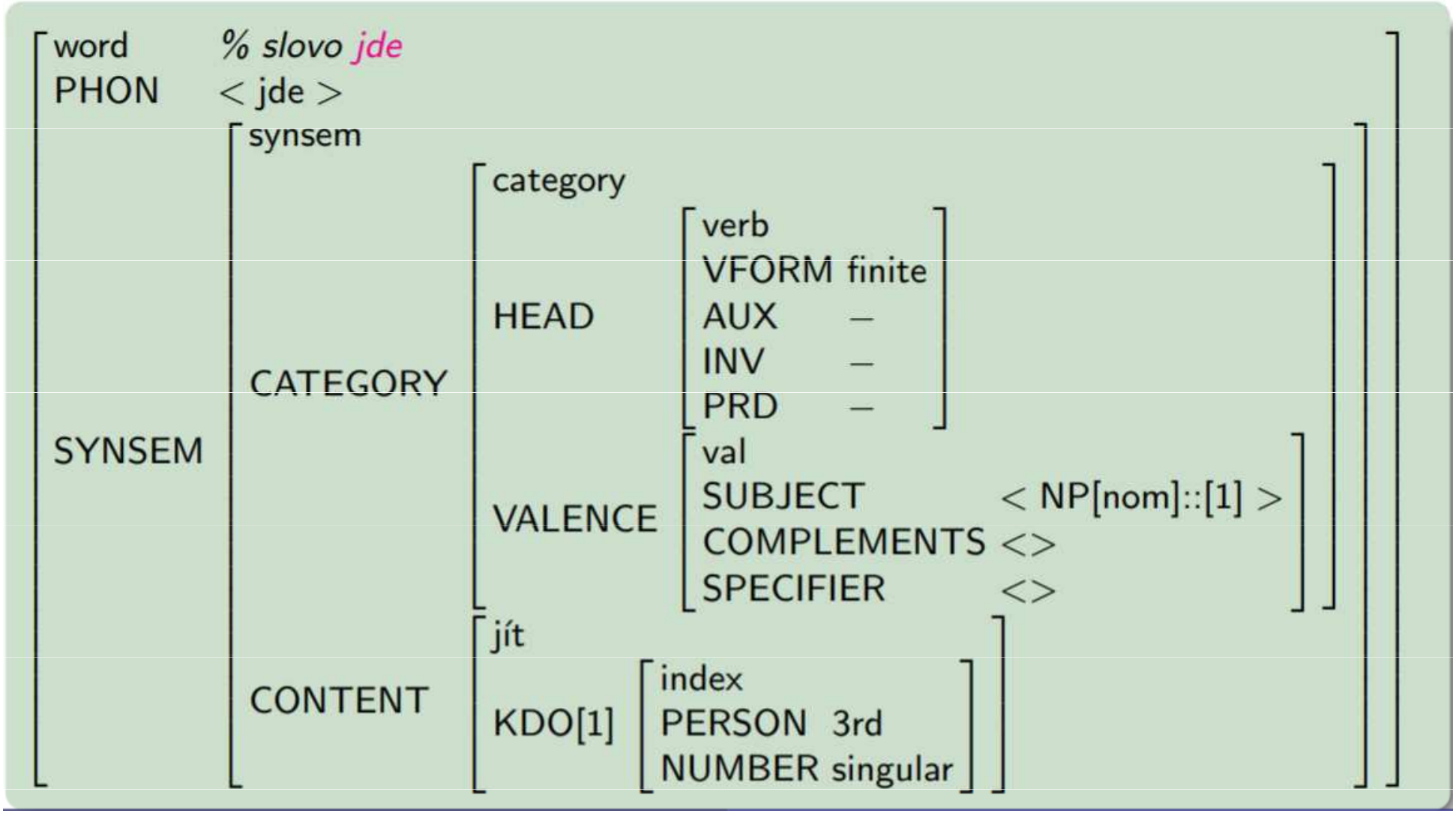
HPSG (Head-Driven Phrase Structure Grammar)

- neterminály jsou nahrazeny příznakovými strukturami
- založeno na omezeních (constrains)
- modeluje jazyk pomocí deklarativních omezení typovaných struktur
- příznaky jsou propojeny pomocí strukturního sdílení, předávání proměnných mezi podstrukturami dané struktury

HPSG (Head-Driven Phrase Structure Grammar)

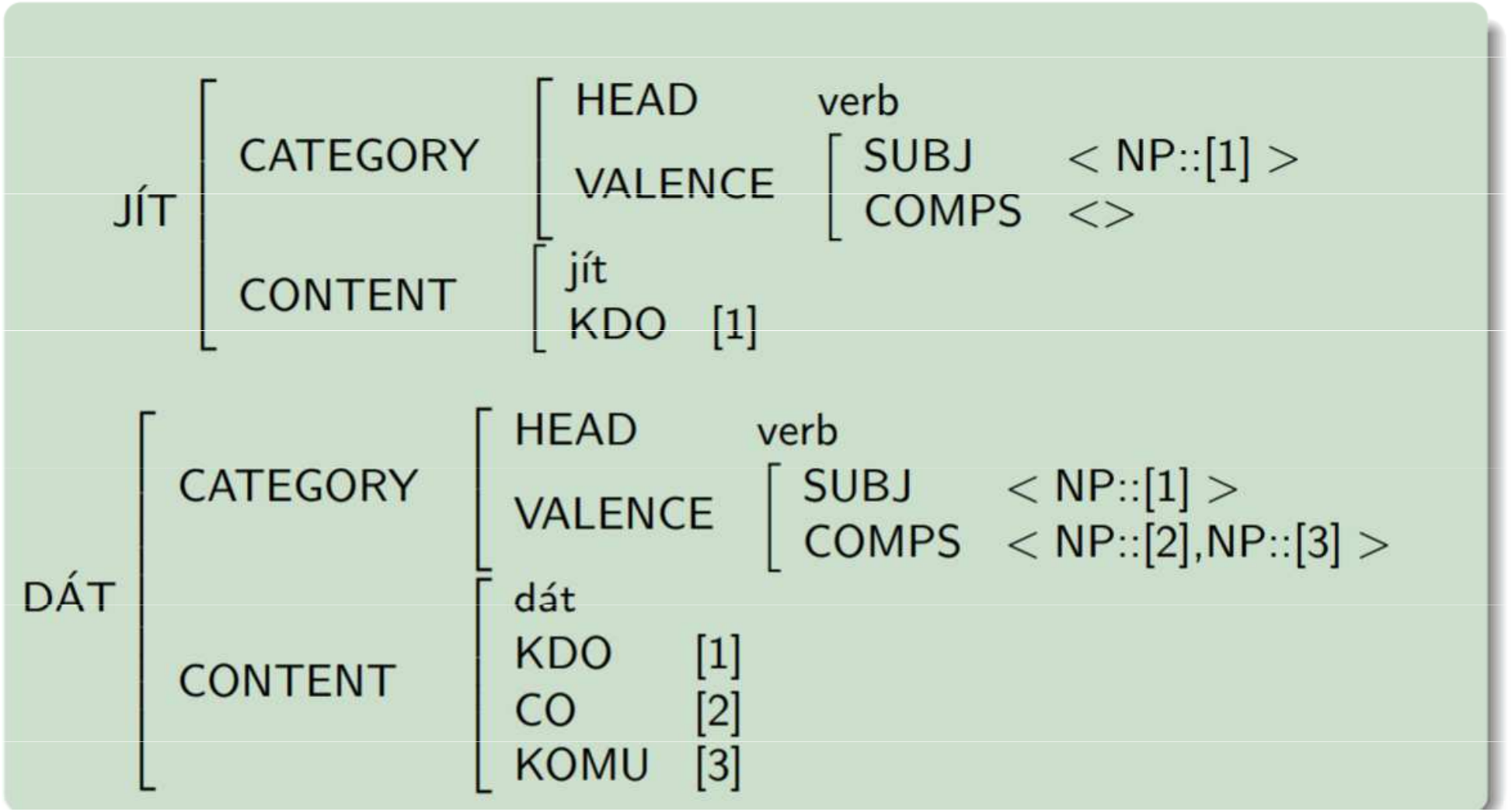
- lexikální hlava – základní prvek frázové struktury HPSG
- hlava určuje základní gramatické vlastnosti fráze
  - N zastupuje NP
  - VP zastupuje S
  - V zastupuje VP
  - relace závislostí (např. valenční rámec slovesa)

HPSG (HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR)



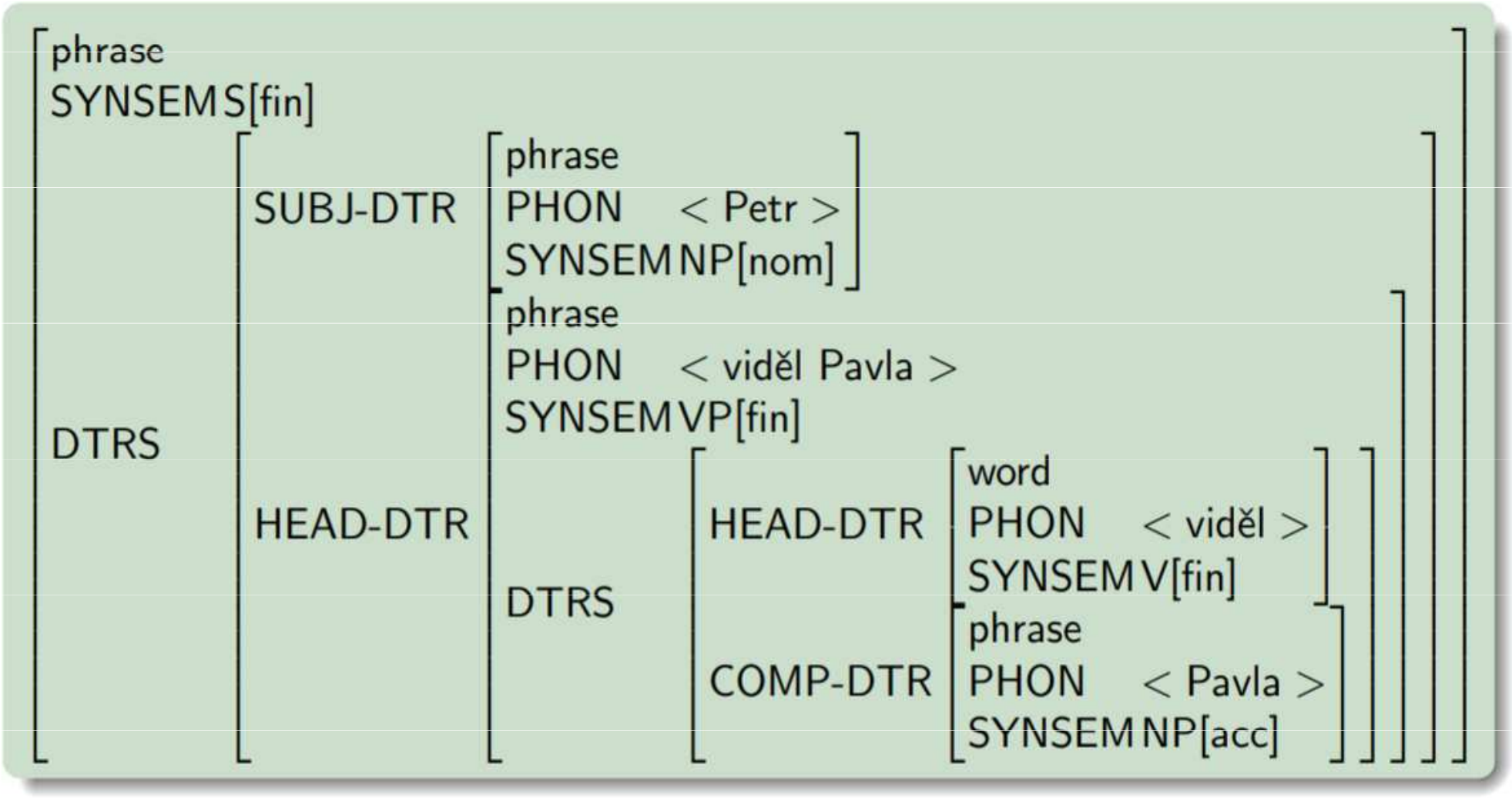
HPSG (HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR)

- velké množství akcí je v lexikonu



HPSG (HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR)

- reprezentace frází, příznak DAUGHTERS (struktura členů fráze)



Literatura

Nový encyklopedický slovník češtiny online:

https://www.czechency.org/

hesla: Formální gramatika, Lexikalistická hypotéza, Nelexikalistická hypotéza, LFG HPSG

https://universaldependencies.org/