Dekomprese v popisu jazyka *aneb* hlubiny i mělčiny deklarativně

Alexandr Rosen
Ústav teoretické a komputační lingvistiky
Universita Karlova v Praze
alexandr.rosen@ff.cuni.cz

http://utkl.ff.cuni.cz/~rosen/public/THESIS

4. března 2002

„A constraint-based approach to dependency syntax applied to some issues of Czech word order“

„Deklarativní formalizace teorie závislostní syntaxe s aplikací na některé problémy českého slovosledu“

- Potíže při vynořování z hlubin – přechod mezi rovinami (z hloubkové na povrchovou)
- Složková syntax a CF gramatika: derivace = reprezentace
- Závislostní syntax: záleží na formalismu, navíc ve FGD je reprezentace abstraktnější (pomocná slova) $\longrightarrow$ roviny popisu
- Vztah mezi povrchovým řetězcem a jeho reprezentací – záležitost teoretického popisu
- Tradiční přístup: stratifikační a procedurální, jde to jinak?

Vývoj

„Lexikální informace ve strojovém překladu“

$\longrightarrow$

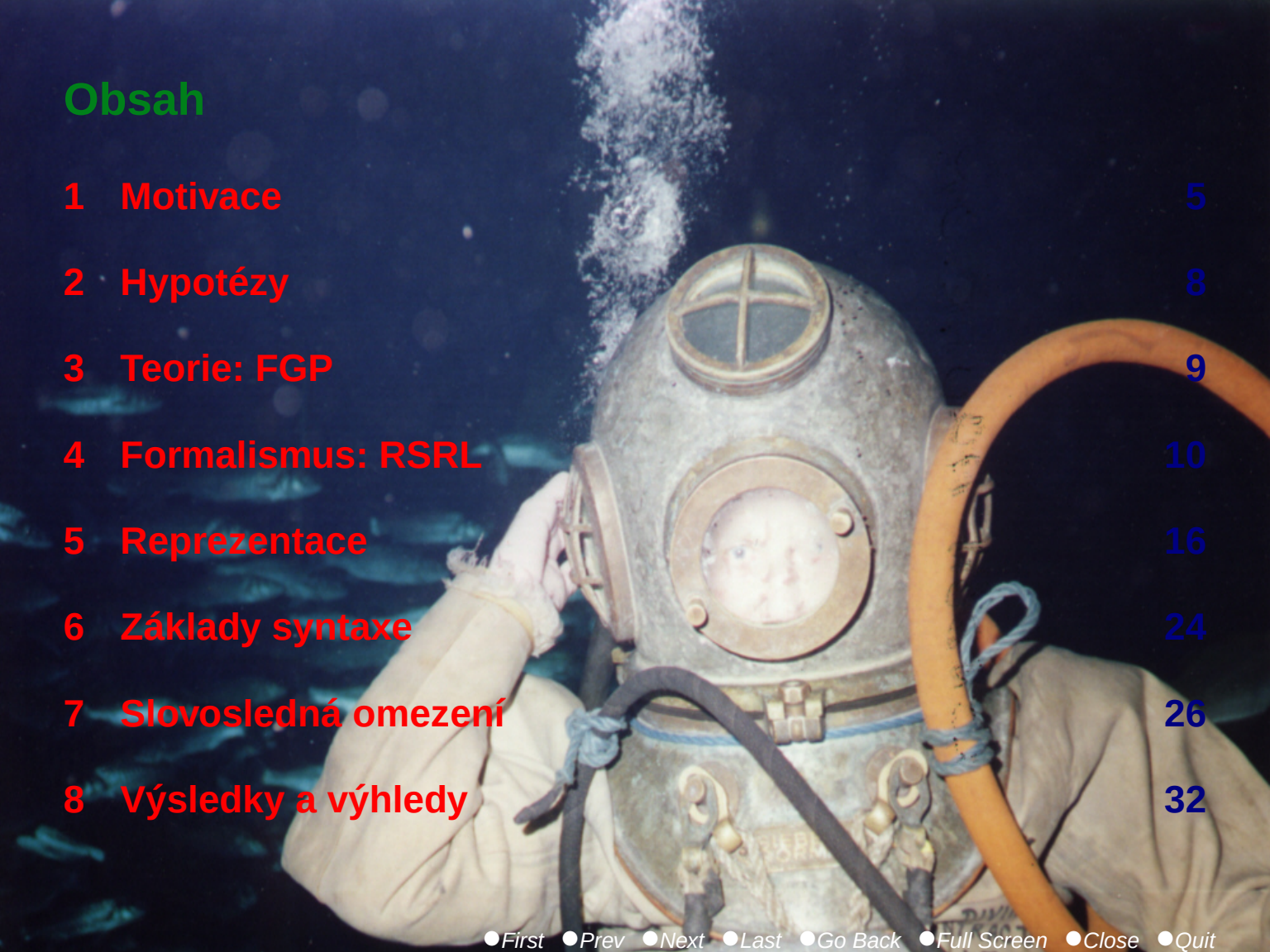
„Strojový překlad jako pokusný záhonek“

$\longrightarrow$

„Závislostní gramatika v deklarativní formalizaci, aplikovaná na kontrastivní popis dvou jazyků“

$\longrightarrow$

„Pokus o popis některých jevů českého slovosledu pomocí FGP v deklarativní formalizaci“



Obsah

1. Motivace

(jsou-li abstraktní entity pro poznání a popis jazyka nutné ...)

Jaký je vztah mezi povrchovým řetězcem fonémů/grafémů a jeho reprezentací?

- Stochastické metody – simulace vědomého poznání
- Teoretická lingvistika – různé hypotézy

Formalismus

- Standardní formalizace FGP je stratifikační s procedurálními prvky (generování tektogramatického zápisu, překladové složky)
- Problémy:
  - paralelní přístup k informacím z více rovin
  - interpretace neúplných výrazů
  - preference generování na úkor analýzy

Slovosled

- Slovosled odráží více faktorů z více rovin
  (slovosledné faktory: aktuální členění, ‚povrchová' pravidla gramatiky)
- Slovosledné faktorů spolu ‚soutěží'
  (jejich role se liší v různých jazycích – V. Mathesius)

Řešení?

- ‚Constraint-based' formalismy umožňují paralelní přístup k více rovinám
- FGP nabízí adekvátní teoretickou koncepci a hloubkovou reprezentaci, včetně aktuálního členění

2. Hypotézy

- Teorie není nerozlučně svázána s formalismem
- FGP lze kombinovat s deklarativním formalismem
- Tato kombinace umožňuje lépe popsat působení slovosledných faktorů

„deklarativní formalismus“:
Relational Speciate Re-entrant Logic – RSRL
Paul King (1989), Frank Richter (2000)

- **Cíl práce:**
  popsat podstatnou část slovosledných jevů v češtině pomocí:
  - FGP jako teoretického východiska
  - RSRL jako formálního jazyka

3. Teorie: FGP

- rozčlenění popisu jazyka na více rovin
- hranice mezi systémem jazyka a sémantickými + pragmatickými interpretacemi
- úloha výpovědní dynamičnosti a aktuálního členění v jazykovém významu
- reprezentace jazykového významu v podobě tektogramatického stromu: vztahy závislosti mezi autosémantickými slovy, hloubkový slovosled, kontextová zapojenost

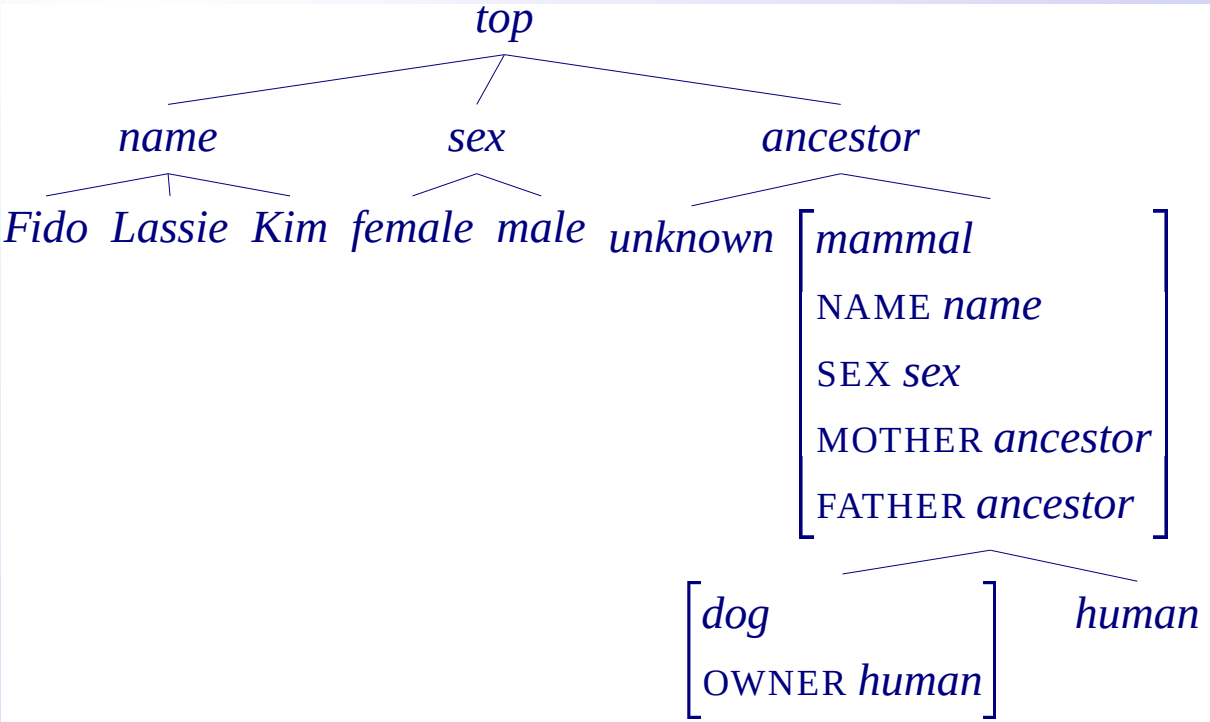
4. Formalismus: RSRL

Gramatika popisuje jazyk nepřímo, pracuje s modelem jazyka.
Skládá se ze „*signatury*“ a „*teorie*“.

Signatura

definuje hierarchii typů – množin objektů modelu.

- Každému objektu modelu je přiřazen právě jeden maximálně specifický typ.
- Typy mají předepsané vlastnosti – atributy. Předepsané hodnoty atributů jsou opět typy.
- Objekty modelu musí mít všechny předepsané atributy, jejichž hodnoty musí být maximálně specifické typy.
- Typy stojící v hierarchii níže dědí všechny specifikace svých nadtypů.


top
name
sex
ancestor
Fido
Lassie
Kim
female
male
unknown
mammal
NAME name
SEX sex
MOTHER ancestor
FATHER ancestor
dog
OWNER human
human


dog
NAME Fido
SEX male
MOTHER unknown
FATHER unknown
OWNER
human
NAME Kim
SEX female
MOTHER unknown
FATHER unknown

Teorie

klade omezení na objekty modelu definované v signatuře.

- Teorie se skládá z formulí, každá formule musí být splněna všemi objekty modelu.
- Formule se skládají z typů a atributů (definovaných v signatuře) a logických symbolů.
- Logické symboly jsou obvyklé spojovací výrazy, proměnné, negace a kvantifikátory s dosahem přes komponenty daného objektu.
- V teorii lze definovat a používat relace.

$$\begin{bmatrix} \textit{dog} & \\ \text{NAME} & \textit{Fido} \\ \text{SEX} & \textit{male} \\ \text{MOTHER} & \begin{bmatrix} \textit{dog} & \\ \text{NAME} & \textit{Lassie} \\ \text{SEX} & \textit{male} \\ \text{MOTHER} & \textit{unknown} \\ \text{FATHER} & \textit{unknown} \\ \text{OWNER} & \boxed{1} \end{bmatrix} \\ \text{FATHER} & \textit{unknown} \\ \text{OWNER} & \boxed{1} \begin{bmatrix} \textit{human} & \\ \text{NAME} & \textit{Kim} \\ \text{SEX} & \textit{female} \\ \text{MOTHER} & \textit{unknown} \\ \text{FATHER} & \textit{unknown} \end{bmatrix} \end{bmatrix}$$

$$\textit{mammal} \rightarrow \begin{bmatrix} \text{MOTHER} \mid \text{SEX} & \textit{female} \\ \text{FATHER} \mid \text{SEX} & \textit{male} \end{bmatrix}$$

$$\left[\text{MOTHER } \textit{mammal}\right] \rightarrow \left[\text{MOTHER} \mid \text{SEX } \textit{female}\right]$$

$$\left[\text{FATHER } \textit{mammal}\right] \rightarrow \left[\text{FATHER} \mid \text{SEX } \textit{male}\right]$$

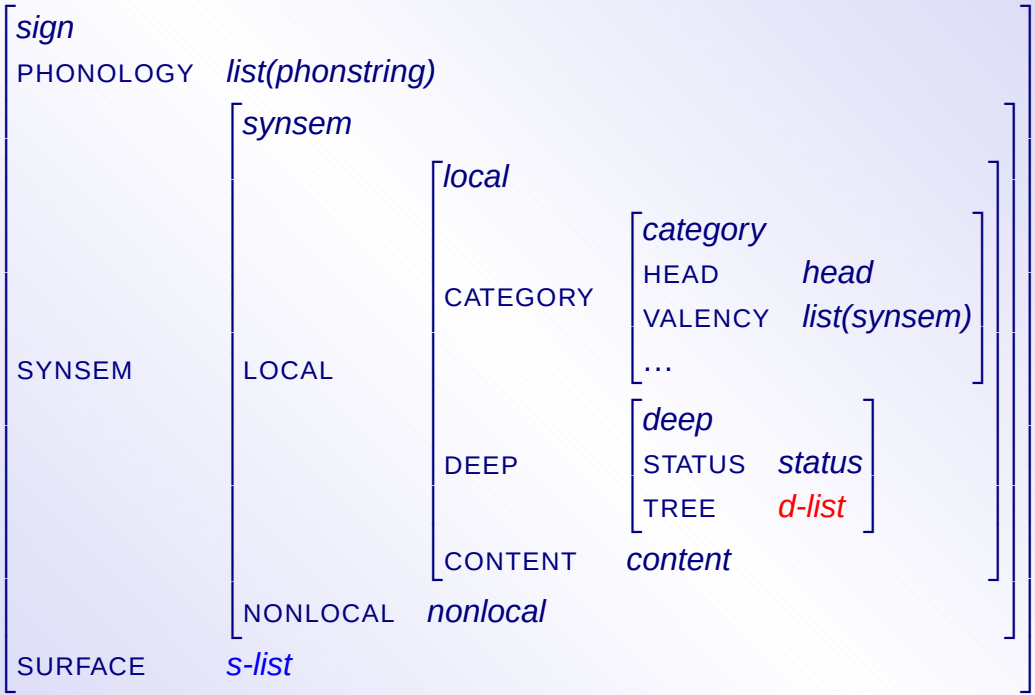
$$\textit{human} \rightarrow \begin{bmatrix} \text{MOTHER} & \textit{human} \\ \text{FATHER} & \textit{human} \end{bmatrix}$$

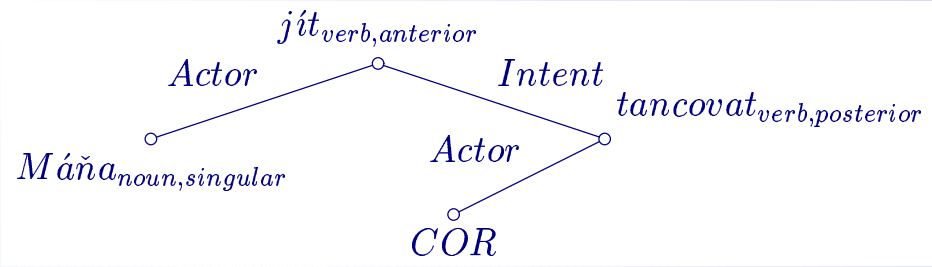
$$\begin{bmatrix} \textit{human} \\ \text{MOTHER } \textit{mammal} \end{bmatrix} \rightarrow \left[\text{MOTHER } \textit{human}\right]$$

$$\begin{bmatrix} \textit{human} \\ \text{FATHER } \textit{mammal} \end{bmatrix} \rightarrow \left[\text{FATHER } \textit{human}\right]$$

$$\textit{dog} \rightarrow \begin{bmatrix} \text{OWNER } \boxed{1} \\ \text{MOTHER} \mid \text{OWNER } \boxed{1} \end{bmatrix}$$

5. Reprezentace





⟨ ⟨ *Máňa* ⟩, *jít*, ⟨ ⟨ *COR* ⟩, *tancovat* ⟩ ⟩

$$\left\langle \left\langle \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{actor} \\ \text{CB} & \textit{yes} \\ \text{CORE} & \boxed{1}\begin{bmatrix} \textit{d-noun} & \\ \text{LEMMA} & \textit{Máňa} \\ \text{D-NUMBER} & \textit{sg} \end{bmatrix} \end{bmatrix} \right\rangle, \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{root} \\ \text{CB} & \textit{no} \\ \text{CORE} & \begin{bmatrix} \textit{d-verb} & \\ \text{LEMMA} & \textit{jít} \\ \text{D-TENSE} & \textit{anterior} \end{bmatrix} \end{bmatrix} \right\rangle,$$

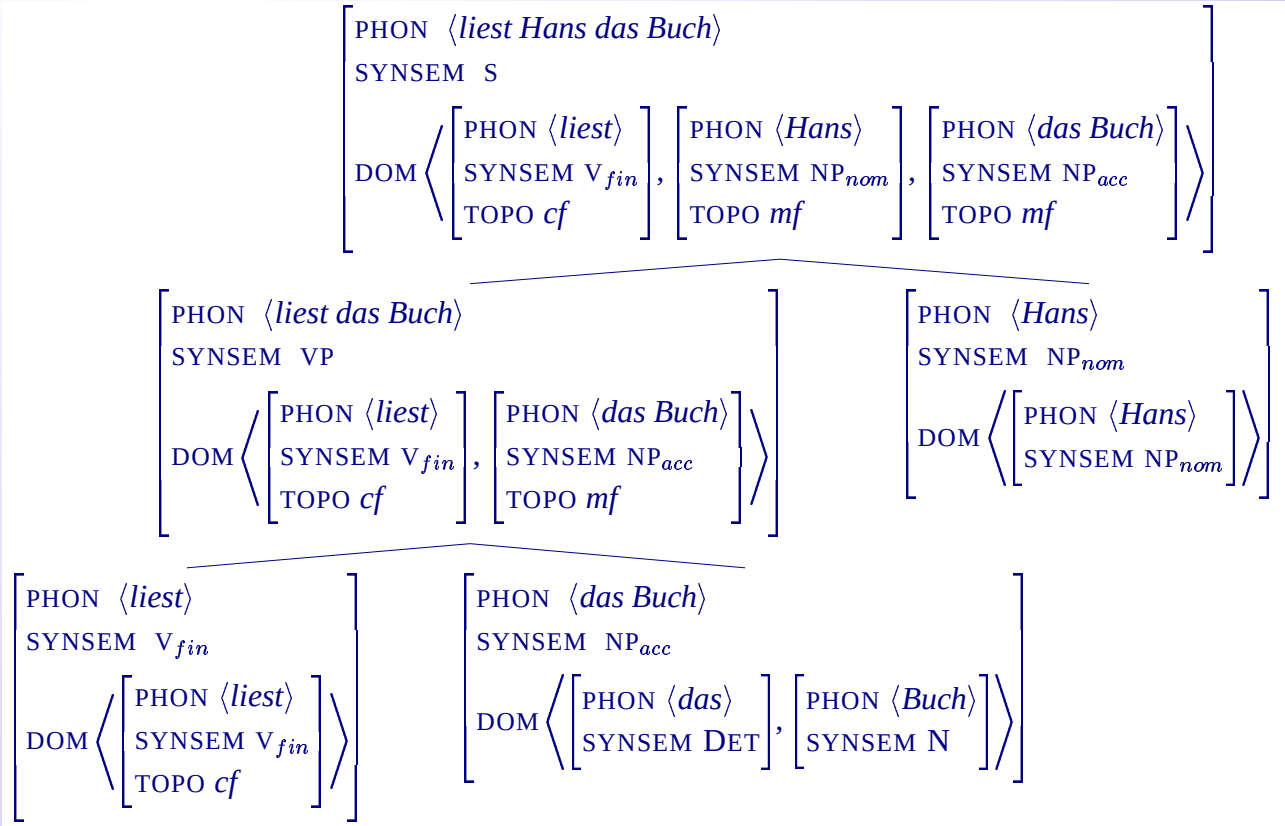
$$\left\langle \left\langle \left\langle \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{actor} \\ \text{CB} & \textit{yes} \\ \text{CORE} & \boxed{1} \end{bmatrix} \right\rangle, \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{intent} \\ \text{CB} & \textit{no} \\ \text{CORE} & \begin{bmatrix} \textit{d-verb} & \\ \text{LEMMA} & \textit{tancovat} \\ \text{D-TENSE} & \textit{posterior} \end{bmatrix} \end{bmatrix} \right\rangle \right\rangle$$

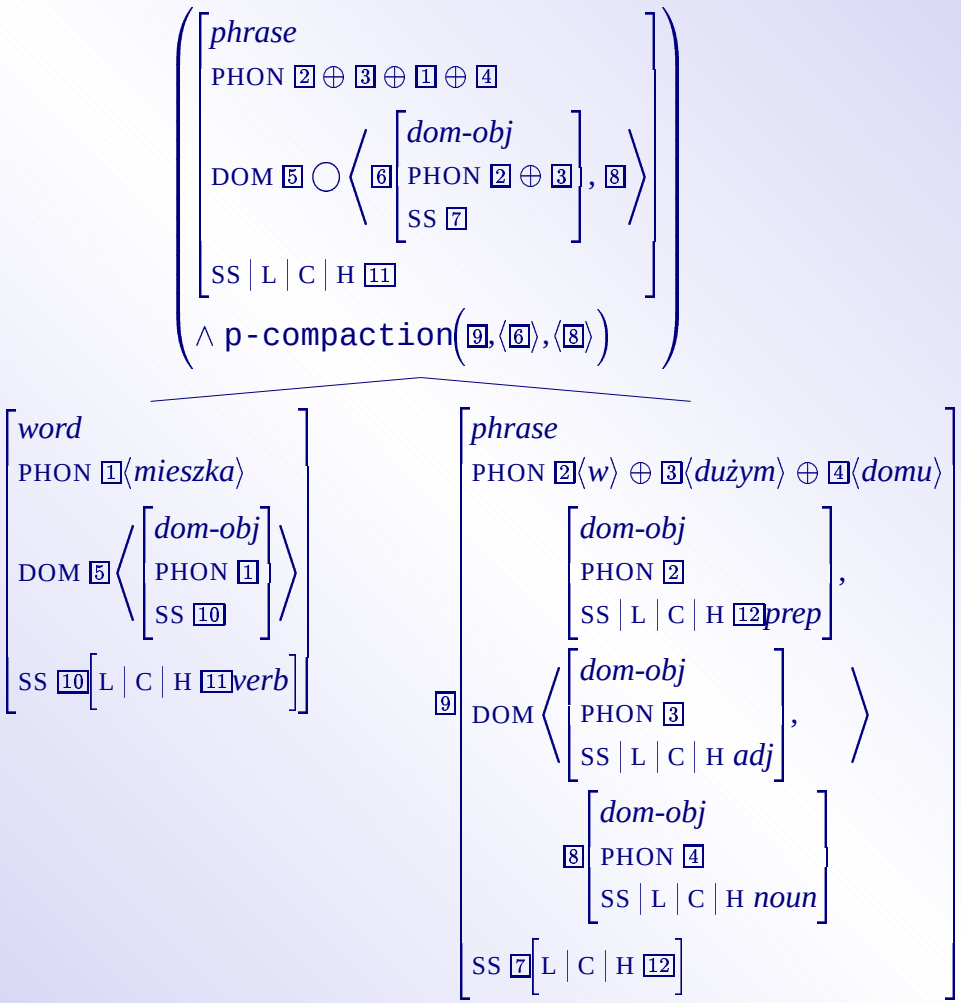
Povrchový řetězec

Inspirace č. 1: Mike Reape, Andreas Kathol

- Horizontální pořadí terminálů v derivačním stromě se nerovná povrchovému slovosledu
- Slovosledné domény, objekty domén
- Relace *shuffle* a *compaction*
- Topologická pole

$$\textit{sign} \rightarrow \left[\begin{array}{l}\textsc{phonology}\ \boxed{1} \oplus \ldots \oplus \boxed{n} \\ \textsc{domain}\ \left\langle \left[\begin{array}{l}\textit{domain-object} \\ \textsc{phonology}\ \boxed{1}\end{array}\right], \ldots, \left[\begin{array}{l}\textit{domain-object} \\ \textsc{phonology}\ \boxed{n}\end{array}\right] \right\rangle\end{array}\right]$$

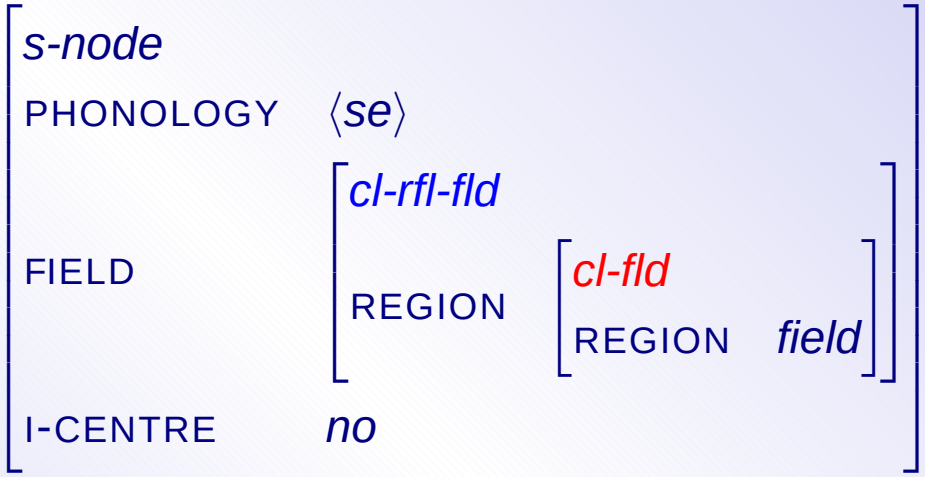

PHON ⟨liest Hans das Buch⟩
SYNSEM S
DOM ⟨ PHON ⟨liest⟩ SYNSEM V_fin TOPO cf , PHON ⟨Hans⟩ SYNSEM NP_nom TOPO mf , PHON ⟨das Buch⟩ SYNSEM NP_acc TOPO mf ⟩
PHON ⟨liest das Buch⟩
SYNSEM VP
DOM ⟨ PHON ⟨liest⟩ SYNSEM V_fin TOPO cf , PHON ⟨das Buch⟩ SYNSEM NP_acc TOPO mf ⟩
PHON ⟨Hans⟩
SYNSEM NP_nom
DOM ⟨ PHON ⟨Hans⟩ SYNSEM NP_nom ⟩
PHON ⟨liest⟩
SYNSEM V_fin
DOM ⟨ PHON ⟨liest⟩ SYNSEM V_fin TOPO cf ⟩
PHON ⟨das Buch⟩
SYNSEM NP_acc
DOM ⟨ PHON ⟨das⟩ SYNSEM DET , PHON ⟨Buch⟩ SYNSEM N ⟩

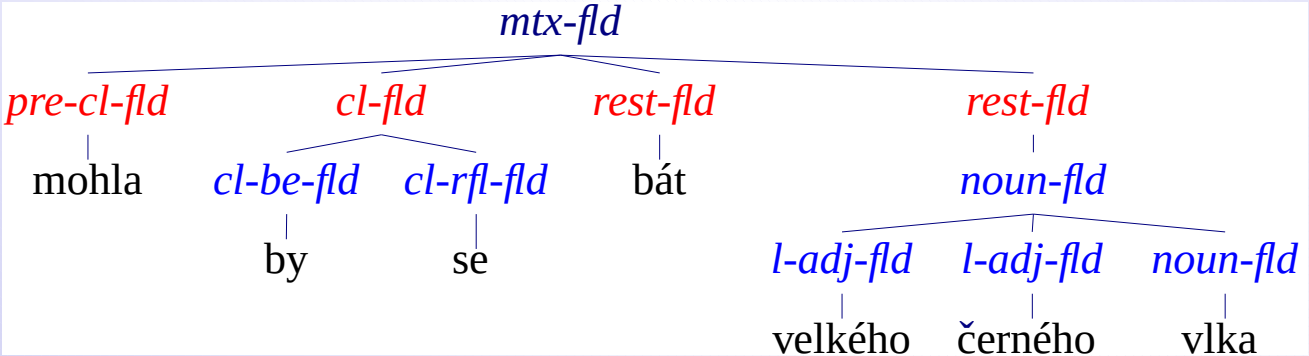

phrase
PHON 2 ⊕ 3 ⊕ 1 ⊕ 4
DOM 5 ◯ ⟨ 6 [dom-obj PHON 2 ⊕ 3 SS 7] , 8 ⟩
SS | L | C | H 11
∧ p-compaction(9, ⟨6⟩, ⟨8⟩)
word
PHON 1⟨mieszka⟩
DOM 5 ⟨ [dom-obj PHON 1 SS 10] ⟩
SS 10 [L | C | H 11 verb]
phrase
PHON 2⟨w⟩ ⊕ 3⟨dużym⟩ ⊕ 4⟨domu⟩
9 DOM ⟨ [dom-obj PHON 2 SS | L | C | H 12 prep] , [dom-obj PHON 3 SS | L | C | H adj] , 8 [dom-obj PHON 4 SS | L | C | H noun] ⟩
SS 7 [L | C | H 12]

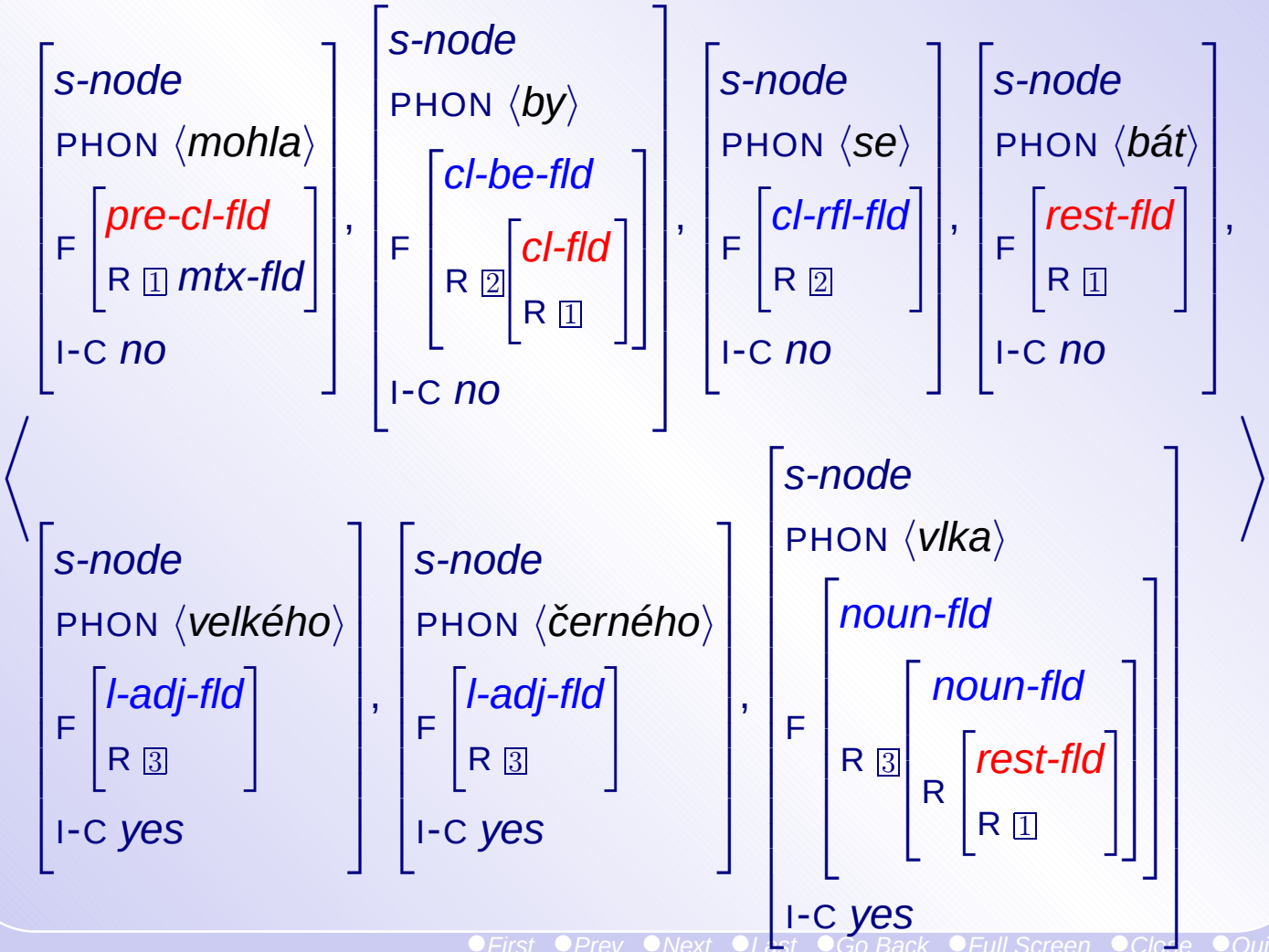
Inspirace č. 2: Gerald Penn

- Objekty domén odpovídají jednotlivým slovům
- Příslušnost ke slovosledné oblasti je vyjádřena sdílením hodnot
- To umožňuje hierarchickou strukturu topologických polí a nezávislost na syntaktických objektech

$$\begin{bmatrix} \textit{s-node} & \\ \text{PHONOLOGY} & \textit{list(phonstring)} \\ \text{FIELD} & \begin{bmatrix} \textit{field} & \\ \text{REGION} & \textit{field} \end{bmatrix} \\ \text{I-CENTRE} & \textit{boolean} \end{bmatrix}$$


s-node
PHONOLOGY ⟨se⟩
FIELD
cl-rfl-fld
REGION
cl-fld
REGION field
I-CENTRE no


mtx-fld
pre-cl-fld
cl-fld
rest-fld
rest-fld
mohla
cl-be-fld
cl-rfl-fld
bát
noun-fld
by
se
l-adj-fld
l-adj-fld
noun-fld
velkého
černého
vlka


s-node
PHON ⟨mohla⟩
F pre-cl-fld
R 1 mtx-fld
I-C no
s-node
PHON ⟨by⟩
F cl-be-fld
R 2 cl-fld
R 1
I-C no
s-node
PHON ⟨se⟩
F cl-rfl-fld
R 2
I-C no
s-node
PHON ⟨bát⟩
F rest-fld
R 1
I-C no
s-node
PHON ⟨velkého⟩
F l-adj-fld
R 3
I-C yes
s-node
PHON ⟨černého⟩
F l-adj-fld
R 3
I-C yes
s-node
PHON ⟨vlka⟩
F noun-fld
R 3 noun-fld
R rest-fld
R 1
I-C yes

6. Základy syntaxe

Vlastně pravidla derivační struktury, tedy skladby nelexikálních objektů typu *sign*.

- Deep List Composition Principle
- Surface List Composition Principle
- Valency Principle
- Head Principle
- Phonology Principle

Derivační struktura je plochá (všechny závislé uzly, včetně funkčních slov, jsou sestry řídícího uzlu). Výjimku funkční slova modifikovaná jiným funkčním slovem.

Deep List Composition Principle

V každém objektu typu *non-lexical* je *d-list* syntaktické matky roven seznamu *d-list* řídící dcery, do něhož jsou vloženy seznamy *d-list* ostatních dcer.

non-lexical $\rightarrow$

$$\left(\begin{array}{l} \left[\begin{array}{l} \text{SYNSEM} \mid \text{LOCAL} \mid \text{DEEP} \mid \text{TREE } \boxed{5} \\ \text{HEAD-DAUGHTER} \mid \text{SYNSEM} \mid \text{LOCAL} \mid \text{DEEP} \mid \text{TREE } \boxed{1} \\ \text{NONHEAD-DAUGHTERS } \boxed{2} \end{array} \right] \\ \wedge\ \text{COLLECT\_DLISTS}(\boxed{2}, \boxed{3}) \\ \wedge\ \text{APPEND}(\boxed{1}, \boxed{3}, \boxed{4}) \\ \wedge\ \text{PERMUTE}(\boxed{4}, \boxed{5}) \end{array} \right)$$

7. Slovosledná omezení

- Omezení hloubkového slovosledu
- Omezení povrchového slovosledu
- Omezení vztahu mezi hloubkovým a povrchovým slovosledem

Omezení hloubkového slovosledu

- V každém nezanořeném stromě musí být aspoň jeden NB uzel.
- V každém stromě musí řídící uzel v horizontálním pořadí předcházet všechny NB uzly.
- V každém stromě musí NB uzly v horizontálním pořadí následovat po CB uzlech podle systémového uspořádání.

Omezení vztahu mezi hloubkovým a povrchovým slovosledem

Uplatní se pro každou dvojici tektogramatických uzlů, není-li povrchová pozice žádného z nich určena povrchovými slovosledními pravidly (dáno topologickým polem).
Existují 3 možnosti:

- Relativní pořadí uzlů je stejné (uplatní se princip aktuálního členění).
- Je-li první z uzlů NB a na prvním místě v seznamu *d-list*, může být v *pre-cl-fld* vyšší klauze (narušení principu členské sounáležitosti).
- Je-li druhý z uzlů na posledním místě v seznamu *d-list*, může být v seznamu *s-list* na jiném než posledním místě a označen jako součást intonačního centra (uplatní se princip důraznosti).

Obecná omezení povrchového slovosledu

Matrix Compaction Principle – všechny struktury *s-node* ve větě jsou částí jednoho pole *matrix-fld*

Planarity Principle – spojitost polí (‚projektivita‘)

Topological Order, Field Existence, Field Uniqueness Principles – definice skladby topologických oblastí:

- pořadí, existence ($\geq$1) a jedinečnost ($=$1) polí v oblasti
- konkrétní hodnoty definovány relacemi specifikovanými pro přehlednost v podobě tabulek

viz table.pdf
viz table1.pdf

Specifická omezení povrchového slovosledu

- komparativní konstrukce
  (*menší vesnice než Lhota*)
- některé případy vzdálených závislostí
  (*Koho jsi říkal, že Marie myslela, že Pavel pozve?*
- rozdělené předložkové fráze
  (*O jakou se jedná soutěž?*)
- postavení příklonek, haplologie
  (*Miloš se jim nakonec rozhodl omluvit.*)

8. Výsledky a výhledy

- formalizace FGP pomocí RSRL
- popis vztahu hloubkového a povrchového slovosledu
- popis interakce hloubkového slovosledu s povrchovými pravidly
- kompatibilita s Mathesiovými slovosledными principy
- formalizace několika zákonitostí povrchového slovosledu s interakcí syntaktických, diskursních a prozodických faktorů

Co dál:

- implementace
- více jevů
- více jazyků
- modifikace?

A to je konec ...