Pokus o formální popis českého slovosledu

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

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

26. 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“

Obsah

1 Proč?

Předpoklad:

rozdíl mezi kompetencí (*langue*) a performancí (*parole*)

- popis (gramatika) kompetence:
  - které řetězce fonémů/grafémů patří a které nepatří do jazyka X
  - co tyto řetězce znamenají (jaký mají vztah k reprezentaci významu)
- popis (gramatika) performance:
  - jak se tyto znalosti využívají při „jazykových aktivitách“

Může korpus nahradit gramatiku?
– Jak z něj „vycucnout“ implicitně obsažená pravidla gramatiky a jak je zobecnit?

Automaticky = statisticky?

- Neoznačkovaný korpus: jaké řetězce se vyskytly, v jakém kontextu a situaci, jak často
- Označkovaný korpus: jaké kategorie, konstrukce, významy se vyskytly, v jakém kontextu a situaci, jak často

Longum iter est per preacepta, breve et efficax per exempla.
– Seneca

Es gibt nur die Beispiele
– Wittgenstein

Ale: statistické metody – jen simulace vědomého poznání

Gramatika:

- jak skládat delší výrazy z kratších
- jaký je vztah mezi povrchovým řetězcem a jeho reprezentací

Možnosti:

- Derivační (stratifikační) vs. nederivační (monostratální) přístup
- Složková vs. závislostní syntax
- Složková syntax a CF gramatika: derivace = reprezentace
- FGP: derivace ≠ reprezentace, reprezentace abstraktnější tradiční přístup: stratifikační a procedurální, jde to jinak?

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

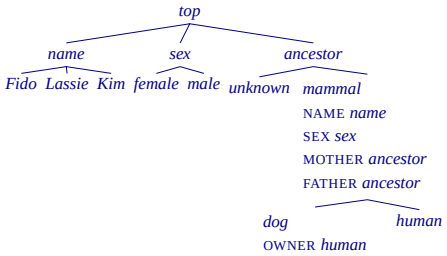
Gramatika popisuje jazyk nepřímo, pracuje s modelem jazyka. Skládá se ze „*signatury*“ a „*teorie*“.

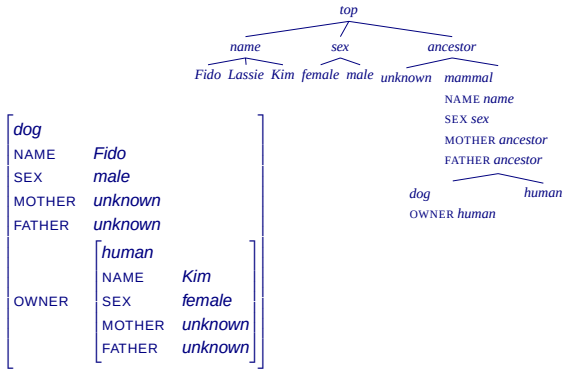
Signatura

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.

Navíc: seznamy, množiny.

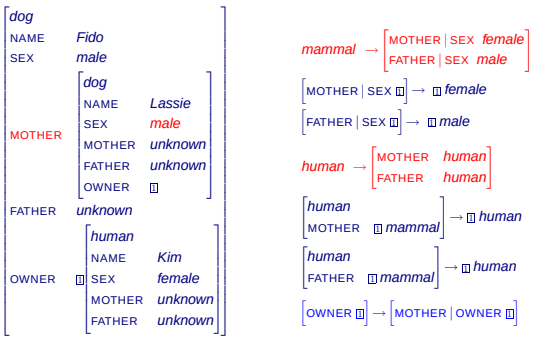




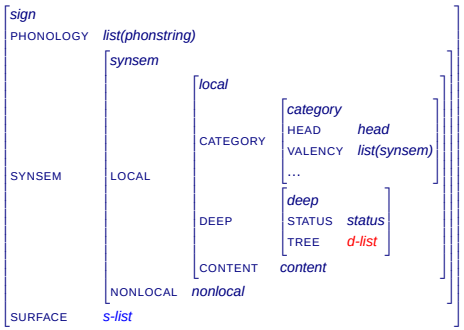
Teorie

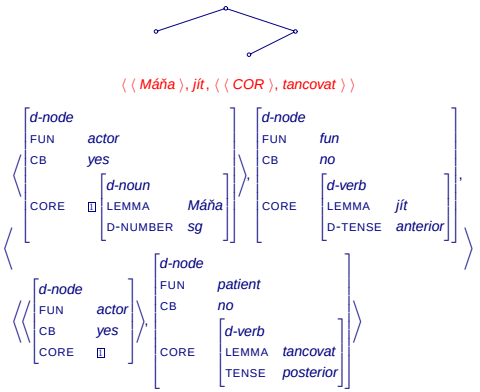
klade omezení na objekty modelu definované v signatuře.

- 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.



5 Reprezentace



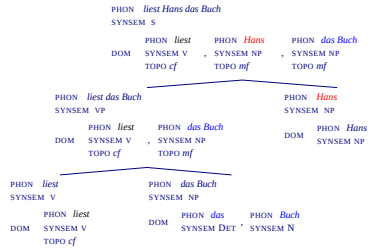


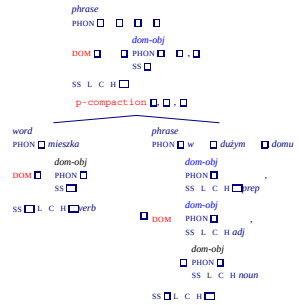
Povrchový řetězec

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

$$\textit{sign} \rightarrow \begin{bmatrix} \text{PHONOLOGY} & \boxed{1} \oplus \ldots \oplus \boxed{n} \\ \text{DOMAIN} & \left\langle \begin{bmatrix} \textit{domain-object} \\ \text{PHONOLOGY } \boxed{1} \end{bmatrix}, \ldots, \begin{bmatrix} \textit{domain-object} \\ \text{PHONOLOGY } \boxed{n} \end{bmatrix} \right\rangle \end{bmatrix}$$



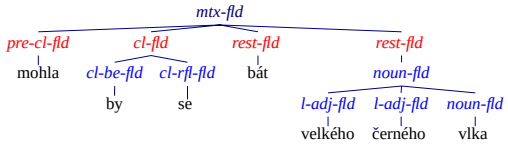


Inspirace č. 2: Gerald Penn

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

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

$$\begin{bmatrix} \textit{s-node} & \\ \text{PHONOLOGY} & \langle \textit{se} \rangle \\ \text{FIELD} & \begin{bmatrix} \textit{cl-rfl-fld} \\ \text{REGION} \quad \begin{bmatrix} \textit{cl-fld} \\ \text{REGION} \quad \textit{field} \end{bmatrix} \end{bmatrix} \\ \text{I-CENTRE} & \textit{no} \end{bmatrix}$$



$$\left\langle \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{mohla} \rangle \\ \text{F} \begin{bmatrix} \textit{pre-cl-fld} \\ \text{R } \boxed{1} \textit{ mtx-fld} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{by} \rangle \\ \text{F} \begin{bmatrix} \textit{cl-be-fld} \\ \text{R } \boxed{2} \begin{bmatrix} \textit{cl-fld} \\ \text{R } \boxed{1} \end{bmatrix} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{se} \rangle \\ \text{F} \begin{bmatrix} \textit{cl-rfl-fld} \\ \text{R } \boxed{2} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{bát} \rangle \\ \text{F} \begin{bmatrix} \textit{rest-fld} \\ \text{R } \boxed{1} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \right.$$

$$\left. \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{velkého} \rangle \\ \text{F} \begin{bmatrix} \textit{l-adj-fld} \\ \text{R } \boxed{3} \end{bmatrix} \\ \text{I-C } \textit{yes} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{černého} \rangle \\ \text{F} \begin{bmatrix} \textit{l-adj-fld} \\ \text{R } \boxed{3} \end{bmatrix} \\ \text{I-C } \textit{yes} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{vlka} \rangle \\ \text{F} \begin{bmatrix} \textit{noun-fld} \\ \text{R } \boxed{3} \begin{bmatrix} \textit{noun-fld} \\ \text{R } \begin{bmatrix} \textit{rest-fld} \\ \text{R } \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{I-C } \textit{yes} \end{bmatrix} \right\rangle$$

6 Syntaktická „kostra“

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.

$$\textit{non-lexical} \rightarrow \left(\begin{bmatrix} \text{SYNSEM} \mid \text{LOCAL} \mid \text{DEEP} \mid \text{TREE } \boxed{0} \\ \text{HEAD-DAUGHTER} \mid \text{SYNSEM} \mid \text{LOCAL} \mid \text{DEEP} \mid \text{TREE } \boxed{1} \\ \text{NONHEAD-DAUGHTERS } \boxed{2} \end{bmatrix} \right.$$
$$\wedge\ \textsc{collect\_dlists}(\boxed{2}, \boxed{3})$$
$$\wedge\ \textsc{append}(\boxed{1}, \boxed{3}, \boxed{4})$$
$$\left. \wedge\ \textsc{permute}(\boxed{4}, \boxed{5}) \right)$$

7 Slovosled

- 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 slovoslednými 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

Oblast	Pole	Pořadí	Obsazení
matrix-fld	*pre-cl-fld*	1	1
	cl-fld	2	$\leq$1
	rest-fld	3	neurčeno
	fin-fld	4	$\leq$1

⟨ [*s-node*
PHONOLOGY ⟨hoří⟩
FIELD [*pre-cl-fld*
REGION *matrix-fld*]
I-CENTRE *yes*] ⟩

⟨ [*s-node*
PHONOLOGY ⟨*děti*⟩
FIELD [*noun-fld*
R [*pre-cl-fld*
R 1 *matrix-fld*]]
I-CENTRE *no*], [*s-node*
PHONOLOGY ⟨*spí*⟩
FIELD [*rest-fld*
R 1]
I-CENTRE *yes*] ⟩

field
- *matrix-fld*
- *embedded-fld* REGION *field*
  - *pre-cl-fld*
  - *cl-fld*
  - *rest-fld*
  - *fin-fld*
  - *dep-fld*
    - *noun-fld*
    - *pp-fld*
    - *adj-fld*
    - *adv-fld*
    - *inf-fld*
    - *aux-fld*
    - *emb-cls-fld*
      - *sconj-cls-fld*
      - *wh-cls-fld*

Oblast	Pole	Pořadí	Obsazení
pre-cl-fld	*dep-fld*	1	1
rest-fld	*dep-fld*	any	any

Oblast	Pole	Pořadí	Obsazení
noun-fld	*l-adj-fld*	1	any
	noun-fld	2	$\geq$1
	emb-cls-fld	3	$\leq$1
	h-adj-fld	$\geq$4	any
	pp-fld	$\geq$4	any
pp-fld	*prep-fld*	1	1
	noun-fld	2	1
l-adj-fld	*adv-fld*	1	any
	adj-fld	2	1
h-adj-fld	*adv-fld*	1	any
	adj-fld	2	1
	pp-fld	$\geq$3	any
	emb-cls-fld	$\geq$3	any
	noun-fld	$\geq$3	any
adv-fld	*adv-fld*	1	$\leq$1
	adv-fld	2	1

Oblast	Pole	Pořadí	Obsazení
cl-fld	*cl-lis-fld*	1	$\leq$2
	cl-be-fld	2	$\leq$1
	cl-rfl-fld	3$\vee$4	$\leq$1
	cl-ethdat-fld	3$\vee$4	$\leq$1
	cl-freedat-fld	5	neurčeno
	cl-dat-fld	6	neurčeno
	cl-acc-fld	7	neurčeno
	cl-gen-fld	8	neurčeno
	cl-ins-fld	9	neurčeno
	cl-nom-fld	neurčeno	$\leq$1
	cl-uz-fld	neurčeno	$\leq$1
	cl-pry-fld	neurčeno	$\leq$1
	cl-vsak-fld	neurčeno	$\leq$1

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 Příklady

Příklad 1: šplhání příklonek

1. A clitic can climb to a higher 2P unless it is governed by a finite verb (1), a deverbative (gerund) (2), an adjectival participle (3), or an adverbial participle (4).

(1) a. Šéf *ho* nařídil zbavit všech výsad.
b. Šéf nařídil, aby *ho* zbavili všech výsad.
c. *Šéf *ho* nařídil, aby zbavili všech výsad.

(2) a. Dědeček nemá rád dětské ušklíbání *se* nad polévkou.
b. *Dědeček se nemá rád dětské ušklíbání nad polévkou.

(3) a. Uvítali bychom více takových kajících *se* hříšníků.
b. *Uvítali bychom se více takových kajících hříšníků.

(4) a. Ředitel vzhlédl od dopisu, tváře *se* ustaraně.
b. *Ředitel se vzhlédl od dopisu, tváře ustaraně.

2. A clitic may only climb through a domain governed by an infinitive.
3. A more deeply embedded clitic cannot climb over a less deeply embedded clitic (5).

(5) a. Pavel *se* snažil *mu* pomoci *ho* najít.
b. Pavel *se mu* snažil *ho* pomoci najít.
c. Pavel *se mu ho* snažil pomoci najít.
d.*?Pavel *se ho* snažil *mu* pomoci najít.

4. Two phonologically identical clitics with different governors either do not co-occur in a single clitic cluster or haplologize (6).

(6) a. Kamila mi slíbila to vrátit MNĚ.
b. *Kamila mi mi to slíbila vrátit.
c. Kamila mi to slíbila vrátit.
d. Kamila mi slíbila mi to vrátit.

5. Two reflexives *si* and *se* can haplologize yielding *si*, if the reflexive *si* originates in a more embedded domain (7).

(7) a. Jan *si* bál vzít kravatu. [KO]
b. Snažím *si* to představit.
c. *Snažím *se* to představit.
d. Styděla *si* sednout do první řady.
e. *Styděla *se* sednout do první řady.
f. *Troufla *si* usadit v první řadě.
g. *Troufla *se* usadit v první řadě.

6. The order of pronominal clitics with the same morphological case co-occurring in a single cluster due to clitic climbing corresponds to the level of embedding of their governors.

Jeden z příkladů, který se zatím nepodařilo formálně popsat:
pořadí nepřízvučných zájmen v dativu

(8) a. [Podařilo se *mi*] [*mu* amputovat pravou zadní nohu].
b. [Podařilo se *mu*] [*mi* udělat velkou radost].

(9) a. Včera se *ti jí* to konečně povedlo vysvětlit.
b. Včera se *jí ti* to konečně povedlo vysvětlit.

(10) a. Vysvětlit se *mi jí* to konečně povedlo až včera.
b. Vysvětlit se *jí mi* to konečně povedlo až včera.

Příklad z korpusu:

(11) Naštěstí se mu jim podařilo chlapce v bezvědomí vyrvat a utéci s ním za plot. (LN 96)

Příklad 2: modifikátory substantiv v první pozici

(12) a. Jaká *se vám* vybaví představa?
b. Vážně nevím, kterou *si* vybral nevěstu.

(13) a. PĚKNÁ *se* nám vybaví představa.
b. BOHATOU *si* vybral nevěstu.
c. TAKOVÝ *se mi* líbí básník.

Ale proč ne taky tohle:

(14) a. *TENHLE *mi* člověk slíbil peníze.
b. *VYSOKÝ *mi* slíbil peníze člověk.

Příklad 3: rozdělené předložkové fráze

(15) O jakou se jedná soutěž?

(16) a. O jak dotovanou soutěž se jedná?
b. O jak dotovanou se jedná soutěž?
c. *O jak se jedná dotovanou soutěž?
d. ?O velmi dobře dotovanou se jedná soutěž.

(17) O jakou jsi myslel, že se jedná soutěž?

(18) PP COMPACTION:

$$\left(\begin{array}{l}\left[\begin{array}{l}\textsc{synsem}\,|\,\textsc{local}\,|\,\textsc{category}\,|\,\textsc{head}\ \textit{noun}\\ \textsc{nonhead-daughters}\ \boxed{1}\\ \textsc{surface}\ \boxed{2}\langle[\textsc{field}\ \textit{prep-fld}]\rangle \bigcirc \boxed{3} \bigcirc \boxed{1}\end{array}\right]\\ \wedge\ \textsc{member}([\textsc{surface}\ \boxed{2}], \boxed{4})\\ \wedge\ \textsc{member}([\textsc{surface}\ \boxed{3}], \boxed{4})\end{array}\right)$$

$$\rightarrow \exists \boxed{5}\left(\begin{array}{l}\left(\begin{array}{l}\textsc{append}(\boxed{2}, \boxed{3}, \boxed{5})\\ \wedge\ \textsc{region}(\boxed{5}, \textit{pre-cl-fld})\\ \wedge\ \textsc{region\_setup}(\boxed{4}, \textit{noun-fld})\end{array}\right)\\ \vee\left(\begin{array}{l}\textsc{shuffle}(\boxed{5}, \boxed{4}, \boxed{3})\\ \wedge\ \textsc{region\_setup}(\boxed{5}, \textit{noun-fld})\end{array}\right)\end{array}\right)$$

9 Výsledky a výhledy

- formalizace FGP pomocí RSRL
- popis vztahu hloubkového a povrchového slovosledu
- popis interakce hloubkového slovosledu s povrchovými pravidly
- kompatibilita s Mathesiovými slovosledními 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 ...