Univerzita Karlova v Praze
Filozofická fakulta
Ústav teoretické a komputační lingvistiky

Alexandr Rosen

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

Teze disertační práce

filologie – matematická lingvistika

vedoucí práce a předseda oborové rady
– prof. PhDr. Petr Sgall, DrSc.

2001

1 Úvod

Tato práce si klade dva hlavní cíle. Prvním z nich je integrovat do jednoho lingvistického rámce dvě různorodé složky: *funkční generativní popis* (FGP), lingvistickou teorii založenou na závislostní syntaxi a vycházející z Pražské školy, a *Relational Speciate Re-entrant Logic* (RSRL), formální jazyk vhodný pro budování deklarativních gramatik. Druhým cílem je ukázat, že kombinaci těchto dvou složek lze použít pro popis povrchového slovosledu, jehož podoba je dána jako výsledek působení různých činitelů, které mají původ v různých rovinách popisu jazyka.

Jinými slovy, cílem je popsat podstatnou část slovosledných jevů v češtině pomocí FGP jako teoretického východiska a RSRL jako formalismu, který pracuje s typovanými strukturami a množinou deklarativně formulovaných omezení. Od takového přístupu lze očekávat, že umožní adekvátním způsobem vyjádřit působení povrchových slovosledných pravidel i *hloubkového slovosledu*, který odráží aktuální členění výpovědi.

Adekvátní popis vztahu mezi hloubkovou syntaktickou strukturou a její povrchovou realizací je důležitým úkolem teoretické lingvistiky. Jako východisko byly přijaty premisy FGP: rozčlenění popisu jazyka na více rovin, hranice mezi systémem jazyka a jeho sémantickými a pragmatickými interpretacemi, úloha výpovědní dynamičnosti a aktuálního členění v *jazykovém významu*. Tento pojem se vztahuje k rovině hloubkové syntaxe, kde je struktura věty vyjádřena jako závislostní strom. Uzly tohoto stromu jsou ohodnoceny údaji, které reprezentují lingvistický význam plnovýznamových slov [7].

V rámci obvyklé formalizace FGP je jazyk popsán gramatikou generující hloubkovou strukturu a překladovými složkami, které spolu s přesunovacími pravidly zajišťují komunikaci mezi formálně oddělenými rovinami popisu.

Tato stratifikační specifikace s jistými procedurálními prvky má své slabiny: preferuje jeden z možných způsobů svého využití (generování, tvoření výpovědí, na úkor analýzy, jejich porozumění), neumožňuje současný přístup k informacím na všech rovinách popisu ani interpretaci neúplných výrazů. Proto je v této práci nahrazena specifikací deklarativní, která umožňuje paralelní reprezentaci a popis více rovin pro věty, větné členy i jednotlivá slova. Jde o formální jazyk, s nímž pracují lingvistické teorie, které si říkají *constraint-based*, např. HPSG – viz [3] a [4]. Tento formalismus, vybavený řádně definovanou syntaxí a sémantikou, využívá jako svého hlavního popisného aparátu systém *typů*, uspořádaných do dědičné hierarchie a doplněných atributy s možností sdílení hodnot. Gramatika formalizovaná tímto způsobem klade omezení na typované sestavy rysů, které slouží jako modely událostí a objektů jazykové skutečnosti.

Druhý úkol se týká empirických dat: český slovosled je stále zdrojem řady nevyřešených otázek. Podle Viléma Mathesia [1] určují slovosled nejen v češtině, ale také v angličtině (a pravděpodobně v lidských jazycích obecně) činitelé různé povahy. Rozdíly mezi jazyky s takzvaným volným a pevným slovosledem jsou způsobeny relativní vahou těchto činitelů. Takové pojetí je slučitelné s deklarativním formalismem, a má také velmi blízko k pojetí vztahu povrchového a hloubkového slovosledu ve FGP, podle níž je povrchový slovosled určován spolupůsobením hloubkových a povrchových činitelů.

Stručně řečeno, jde o obhajobu tvrzení, že deklarativní ‚constraint-based' formalizací FGP lze slovosledné (a prosodické) jevy (alespoň) v češtině vyřešit snáze než ve stratifikačním formalismu s procedurálními prvky.

Vzhledem k tomu, že jde o spojení dvou odlišných lingvistických tradic, obsahuje práce rozsáhlé úvodní partie, které mají posloužit jako základ pro následující specifikaci gramatiky. Specifikace gramatiky pak začíná stanovením pravidel, jimiž se řídí skládání větších celků z menších. Následuje formalizace vztahu mezi hloubkovým a povrchovým slovosledem a interakce hloubkového slovosledu s povrchovými slovoslednými pravidly. Tento úkol lze řešit na abstraktní rovině, neboť FGP poskytuje v otázkách formalizace aktuálního členění a podobných tématech dostatečnou oporu. Obtížněji se hledá empirická opora a odpovídající formální analýza pro povrchové jevy.

Kromě zachycení vztahu mezi povrchovým slovosledem a hloubkovou reprezentací obsahuje práce i pokusy o více či méně důsledné řešení některých dalších jevů patřících výhradně do povrchové roviny. Kromě pravidelných případů souvislých realizací závislostních podstromů jde také o nespojité komparativní konstrukce, některé případy vzdálených závislostí, rozdělené předložkové fráze a postavení pomocných (synsémantických) slov. Podrobněji se práce věnuje obtížnému tématu větných příklonek.

2 Organizace

Práce se skládá z osmi kapitol a čtyřdílné přílohy. První tři kapitoly podávají úvod a základní orientaci v problematice: po identifikaci problémů, které práce hodlá řešit a proč (kapitola 1) následuje pojednání o volbě lingvistické teorie a formálního metajazyka a porovnání několika předchozích závislostních přístupů k řešení problémů slovosledu (kapitola 2). V kapitole 3 je pak podrobněji představena teorie a formalismus, které tvoří základ navrhovaného řešení.

Tři následující kapitoly tvoří jádro celé práce. V kapitole 4 jsou uvedeny teoretické a empirické otázky, které budou předmětem řešení v rámci formálního popisu. Kromě toho je zde také několik poznámek týkajících se některých pojmů: vztahu lingvistického významu a sémantiky, rozdílu mezi řídícím členem a pojmem *head* a možnosti řešení koordinace a apozice. Nejobsáhlejší poznámka se ovšem týká slovosledu. Na základě Mathesiových slovosledných principů je navržena klasifikace slovosledných jevů, která se později promítne v navrženém formálním popisu. V kapitole 5 je specifikována podoba dvou důležitých součástí formálního popisu: reprezentace hloubkové (tektogramatické) a povrchové (morfématické) roviny. Zde je také ve stručnosti popsána úloha a organizace slovníku. V následující kapitole je tak možno přistoupit k formálním definicím. Jde o definice (a) obecných principů, které určují podobu reprezentace výrazů jako funkce částí, z nichž jsou výrazy složeny, (b) principů vymezujících podobu reprezentace hloubkové a povrchové roviny a vztahu mezi nimi, a konečně (c) řady povrchových omezení slovosledu specifických pro češtinu.

V kapitole 7 se systém aplikuje na postavení českých větných příklonek s mnoha příklady. Poslední kapitola obsahuje souhrn výsledků a perspektivy další práce na otevřených tématech.

V příloze jsou shrnuty všechny formální části popisu, spolu s definicemi relací a soupisu tektogramatických funkcí.

3 Teorie a její formalizace

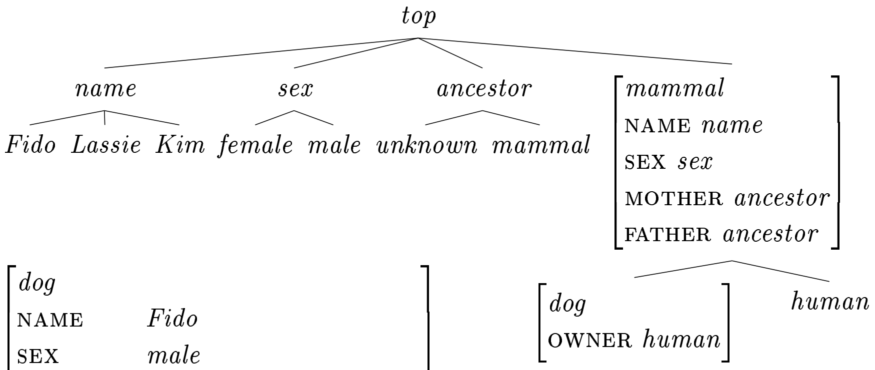
Vyjdeme-li z požadavku deskriptivní adekvátnosti, lze uvést několik argumentů ve prospěch závislostního pojetí hloubkové syntaxe jako roviny, která je (a) dostatečně abstraktní a tedy zbavená povrchových jevů, jako je např. shoda nebo forma a postavení pomocných slov, ale zároveň je (b) pevně ukotvená v systému jazyka. Hloubkovou reprezentaci tohoto druhu nabízí právě FGP [7], [6]. Kromě specifikace závislostních vztahů mezi významovými slovy věty podává hloubková reprezentace i informace o aktuálním členění. Každý uzel závislostního stromu je označen jako *kontextově zapojený* nebo *kontextově nezapojený*. Všechny uzly jsou pak uspořádány podle hierarchie výpovědní dynamičnosti, přičemž nejdynamičtější z nich (obvykle ty, které přinášejí novou informaci) jsou nejvíce vpravo.

Rovina závislotní syntaxe je jen jednou z rovin FGP. Vztah mezi ní a posloupností znaků nebo zvuků se obvykle popisuje procedurálně pomocí překladových složek a přesunovacích pravidel, které fungují postupně v pořadí od vyšších k nižším rovinám popisu. Takto uspořádaný popis jazyka je obtížné použít pro analýzu povrchového řetězce, paralelní využívání informací z více rovin a interpretaci neúplných výrazů. Z těchto důvodů je v této práci vztah mezi rovinou hloubkové syntaxe a povrchovým řetězcem popsán deklarativně. Větné členy jsou modelovány jako typované sestavy rysů, určované souborem omezení – gramatikou a slovníkem. Každý větný člen je tak reprezentován a popisován jako jediný objekt o více dimenzích, které odpovídají rovinám popisu.

Použitý formalismus [5] byl původně vyvinut pro HPSG, ale jeho vlastnosti nejsou v rozporu s premisami FGP, což je dáno mimo jiné i tím, že RSRL lze použít v mnoha jiných kontextech mimo své původní určení. Vůči RSRL lze vznést námitku, že je jako prostředek popisu jazyka příliš silný, což dokládají jeho výpočetní vlastnostmi. Na to lze odpovědět tak, že není úkolem formalismu omezovat vyjadřovací možnosti teorie, ba že tomu je právě naopak: jedině teorie (a gramatika) může omezit možnosti, které připadají v úvahu. Implementaci je pak třeba řešit jako oddělený problém.

Gramatika ve formátu RSRL se skládá ze dvou částí: *signatury* a *teorie*. V *signatuře* se definují typy objektů, které se mohou objevit v modelu a které jsou dále vymezeny *teorií*. Tyto objekty mají podobu typovaných sestav rysů, které jsou interpretovány jako reprezentace množin jazykových událostí a jejich vlastností v reálném světě. Pro jednoduchost bude uveden nelingvistický příklad.

(1)



(2)
$$\begin{bmatrix} \textit{dog} & \\ \textsc{name} & \textit{Fido} \\ \textsc{sex} & \textit{male} \\ \textsc{mother} & \begin{bmatrix} \textit{dog} & \\ \textsc{name} & \textit{Lassie} \\ \textsc{sex} & \textit{male} \\ \textsc{mother} & \textit{unknown} \\ \textsc{father} & \textit{unknown} \\ \textsc{owner} & \boxed{1} \end{bmatrix} \\ \textsc{father} & \textit{unknown} \\ \textsc{owner} & \boxed{1} \begin{bmatrix} \textit{human} & \\ \textsc{name} & \textit{Kim} \\ \textsc{sex} & \textit{female} \\ \textsc{mother} & \textit{unknown} \\ \textsc{father} & \textit{unknown} \end{bmatrix} \end{bmatrix}$$

(3)
$$[\textsc{mother} \mid \textsc{sex}\ \boxed{1}] \rightarrow \boxed{1}\ \textit{female}$$

(4)
$$\begin{bmatrix} \textit{human} & \\ \textsc{mother} & \boxed{1}\ \textit{mammal} \end{bmatrix} \rightarrow \boxed{1}\ \textit{human}$$

(5)
$$[\textsc{owner}\ \boxed{1}] \rightarrow [\textsc{mother} \mid \textsc{owner}\ \boxed{1}]$$

Signatura (1) definuje hierarchii typů v modelu jednoduchého světa psů a jejich pánů. Některé typy mají atributy, jejichž předepsané hodnoty jsou opět typy. Ke všem svým atributům dědí daný typ všechny atributy a předepsané hodnoty od typu, který je mu v hierarchii nadřízen. Sestavy rysů v modelu mohou obsahovat jen ty typy, které jsou v hierarchii nejníže, tedy maximálně specifické typy, a to se všemi předepsanými atributy, jejichž hodnoty musí být opět maximálně specifické typy.

V (2) je sestava rysů, která vyhovuje signatuře (1). Je modelem psa jménem Fido, jehož majitel (Kim) je stejný jako majitel Fidovy matky Lassie (číslo v rámečku $\boxed{1}$ koindexuje identické části struktury). Tato sestava rysů vyhovuje i celé gramatice, protože její *teorie* je prázdná a žádná další omezení tedy nejsou kladena. Signatura však připouští i některé objekty, které by se patrně neměly v modelu objevit, jako např. matka, která je mužského pohlaví (2). Tomu lze odpomoci implikací uvedenou v (3), která představuje první formuli *teorie*, jeden z jejích *popisů*. Každý popis musí být splněn všemi objekty v modelu. (3) tedy znamená: pro každý objekt s atributem MOTHER (t.j., objekt typu *mammal*) a hodnotou tohoto atributu takovou, že jedním z jejích atributů je SEX (opět musí jít jen o objekt typu *mammal*), hodnota tohoto atributu SEX musí být *female*. Podobný popis lze použít také pro vyloučení psa jako matky (a ovšem i otce) člověka (4). Popis (5) pak stanoví méně samozřejmý fakt, že každý majitel psa je také majitelem matky tohoto psa.

RSRL obsahuje také obvyklé logické spojovací výrazy, proměnné, negaci, a možnost použít explicitní kvantifikaci s dosahem přes komponenty daného objektu.

4 Slovosledné principy

Vilém Mathesius přišel s koncepcí slovosledných principů, které působí ve vzájemné interakci. Tyto principy jsou univerzální, ale v různých jazycích mají různou váhu. Pro každý jazyk lze specifikovat částečné uspořádání těchto principů. Toto uspořádání rozhoduje o vítězi v případě, že existuje více principů, jejichž uplatnění by vedlo k odlišným slovosledům.

FGP považuje jeden z těchto principů, totiž princip aktuálního členění, za primární. Aktuální členění dané výpovědi přitom lze vyjádřit různými způsoby: slovosledně, prosodicky nebo určitou syntaktickou konstrukcí. Aktuální členění dané výpovědi je spolu s dalšími složkami lingvistického významu součástí hloubkové reprezentace. Skládá-li se gramatika: (a) z množiny omezení kladených na možné tektogramatické reprezentace, (b) z množiny omezení kladených na možné povrchové řetězce, a (c) z množiny omezení kladených na vztahy mezi tektogramatickými reprezentacemi a povrchovými řetězci, pak (c) představuje právě onu klíčovou složku, která zprostředkovává kontakt mezi hloubkovou reprezentací aktuálního členění a jeho povrchovou realizací za podmínek určovaných ostatními činiteli. Aktuální členění se přitom projeví na povrchu, kdykoli je to možné, tedy pokud není neutralizováno jiným principem. Vzhledem k tomu, že použitý formalismus umožňuje pracovat se všemi údaji paralelně, slovoslední činitelé mohou působit paralelně a ve vzájemné interakci. Tyto úvahy vedou k následujícím předběžným slovosledným omezením (PS: povrchový slovosled, HS: hloubkový slovosled):

a. Relativní PS každé dvojice slov odpovídá HS odpovídajících sémantémů, pokud neplatí některý z bodů uvedených ve speciálních slovosledných podmínkách (viz níže).

b. Pomocné slovo je v postavení těsně vedle příslušného významového v pořadí daným syntaktickým omezením, pokud neplatí některý z bodů uvedených ve speciálních slovosledných podmínkách (viz níže).

c. Pro relativní PS každé dvojice pomocných slov v jedné slovosledné doméně platí speciální slovosledné podmínky.

Speciální slovosledné podmínky:

a. Slovo na první pozici v PS je intonační centrum výpovědi a je součástí vlastního fokusu.
b. Syntaktické omezení vyžaduje jiný slovosled.
c. Přízvučný rytmus vyžaduje jiný slovosled.
d. Slovo je na první pozici v PS, slovosledná doména je větší než doména jeho hloubkového podstromu a slovo je součástí vlastního nebo kontrastivního topiku.

Pokud se dvě nebo více podmínek utkají ve sporu o různý slovosled, rozhodne pořadí principů specifické pro daný jazyk. U všech speciálních slovosledných podmínek musí být splněny specifické podmínky týkající se lokálnosti.

5 Architektura

Problém, jak kompozicionálně popsat vztah mezi povrchovým řetězcem a tektogramatickou reprezentací je řešen pomocí ploché derivační struktury, v níž je pomocné slovo syntaktickou sestrou řídící dcery (*head daughter*) a závislých uzlů (s výjimkou případů, kdy pomocné slovo může být dcerou jiného pomocného slova, jako v případě analytické morfologie sloves). Výraz odpovídající tektogramatickému uzlu nebo podstromu má podobu sestavy rysů typu *sign*. Jeho struktura je naznačena v obr. 1.

Tato struktura je velice podobná typu *sign* v HPSG. Je tomu tak proto, aby bylo možné snadno převzít řešení použitá v rámci této teorie, která jsou v souladu s koncepcí FGP. Nic podstatného však na této podobnosti nezávisí.

Typ *deep* reprezentuje tektogramatickou rovinu a typ *surface* rovinu morfématickou (řetězec objektů představujících morfémy, seřazených podle povrchového slovosledu). Typ *status* má dva podtypy: *embedded* a *unembedded*. Druhý z nich obsahuje atributy, které jsou relevantní pro celou větu. Dále struktura obsahuje části, které vyjadřují, zda a jak se příslušný výraz může spojovat s jinými výrazy: *category* a – volitelně – jeho sémantickou interpretaci (*content*).

Typ *sign* má také dva podtypy: *lexical* a *non-lexical* (odpovídající typům HPSG *word* a *phrase*). Typ *non-lexical* má dva další atributy, které zaznamenávají bezprostřední syntaktické složky výrazu a vyjadřují tak strukturu derivačního stromu: atribut HEAD-DAUGHTER s předepsanou hodnotou typu

$$\begin{bmatrix} \textit{sign} & \\ \text{PHONOLOGY} & \textit{list(phonstring)} \\ \text{SYNSEM} & \begin{bmatrix} \textit{synsem} & \\ \text{LOCAL} & \begin{bmatrix} \textit{local} & \\ \text{CATEGORY} & \begin{bmatrix} \textit{category} & \\ \text{HEAD} & \textit{head} \\ \text{VALENCY} & \textit{list(synsem)} \\ \ldots & \end{bmatrix} \\ \text{DEEP} & \begin{bmatrix} \textit{deep} & \\ \text{STATUS} & \textit{status} \\ \text{TREE} & \textit{d-list} \end{bmatrix} \\ \text{CONTENT} & \textit{content} \end{bmatrix} \\ \text{NONLOCAL} & \textit{nonlocal} \end{bmatrix} \\ \text{SURFACE} & \textit{s-list} \end{bmatrix}$$

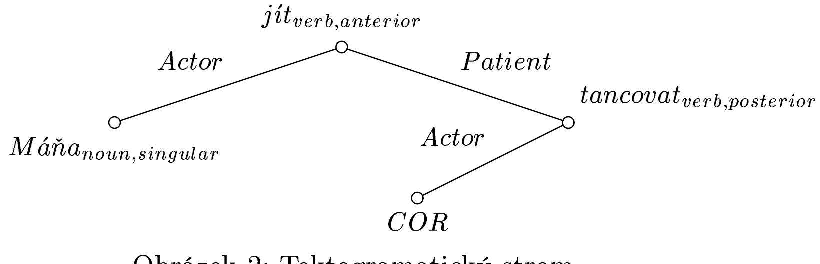
Obrázek 1: Schéma typu *sign*

sign a atribut NONHEAD-DAUGHTERS s předepsanou hodnotou seznam objektů typu *sign*. Řetězec fonémů (v našem případě grafémů) daného výrazu je znázorněn jako hodnota atributu PHONOLOGY.

Tektogramatická reprezentace je reprezentována jako rekurzivní struktura, seznam (*d-list*), který se skládá z jednoho objektu typu *d-node*, který reprezentuje řídící uzel, a dalších seznamů typu *d-list*, které reprezentují závislé podstromy.

(6) Máňa šla tancovat

Tektogramatický strom věty (6) je znázorněn na obr. 2. Jeho linerizovaný zápis je schématicky zachycen v (7). Lomené závorky uzavírají seznam.



Obrázek 2: Tektogramatický strom

(7) $\langle\ \langle[\textit{Máňa}]\rangle,\ [\textit{jít}],\ \langle\ \langle[\textit{COR}]\rangle,\ [\textit{tancovat}]\ \rangle\ \rangle$

Reprezentace tektogramatické roviny pomocí vnořených seznamů je podmíněná projektivitou tektogramatických stromů. Příklad (8) ukazuje ohodnocení uzlů, jejich „vnitřek“. Příklad zobrazuje jen základní údaje: tektogramatickou funkci, kontextovou zapojenost, tektogramatický slovní druh, lemma a jeden z gramatémů relevantních pro daný slovní druh.

(8)

$$\left\langle \begin{array}{l} \left\langle \begin{bmatrix} \textit{d-node} & \\ \textsc{fun} & \textit{actor} \\ \textsc{cb} & \textit{yes} \\ \textsc{core} & \boxed{1} \begin{bmatrix} \textit{d-noun} & \\ \textsc{lemma} & \textit{Máňa} \\ \textsc{d-number} & \textit{sg} \end{bmatrix} \end{bmatrix} \right\rangle, \begin{bmatrix} \textit{d-node} & \\ \textsc{fun} & \textit{fun} \\ \textsc{cb} & \textit{no} \\ \textsc{core} & \begin{bmatrix} \textit{d-verb} & \\ \textsc{lemma} & \textit{jít} \\ \textsc{d-tense} & \textit{anterior} \end{bmatrix} \end{bmatrix}, \\ \left\langle \left\langle \begin{bmatrix} \textit{d-node} & \\ \textsc{fun} & \textit{actor} \\ \textsc{cb} & \textit{yes} \\ \textsc{core} & \boxed{1} \end{bmatrix} \right\rangle, \begin{bmatrix} \textit{d-node} & \\ \textsc{fun} & \textit{patient} \\ \textsc{cb} & \textit{no} \\ \textsc{core} & \begin{bmatrix} \textit{d-verb} & \\ \textsc{lemma} & \textit{tancovat} \\ \textsc{tense} & \textit{posterior} \end{bmatrix} \end{bmatrix} \right\rangle \end{array} \right\rangle$$

Povrchový řetězec slov je formálně vyjádřen jako prostý seznam (*s-list*). Jeho struktura a struktura jeho prvků je inspirována seznamy slovosledných domén [2], v nichž se pořadí a těsnost vzájemného postavení (adjacence) prvků seznamu (objektů domény) stanoví pomocí vnořitelných *topologických polí* a sdílení struktur (vyjádřeného koindexací). V seznamu typu *s-list* (10), který představuje větu (9), jsou tyto údaje zakódovány v objektech reprezentujících jednotlivá slova (objekty typu *s-node*) jako hodnoty atributů F(IELD) a R(EGION). Atribut I-C vyjadřuje údaj o *informačním centru* věty.

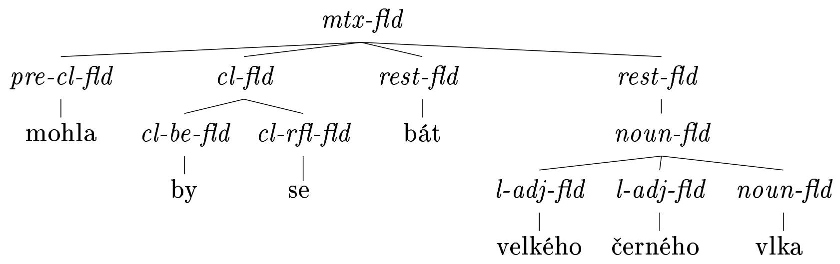
(9) Mohla by se bát velkého černého vlka

(10)

$$\left\langle \begin{array}{l} \begin{bmatrix} \textit{s-node} \\ \textsc{phon}\ \langle \textit{mohla} \rangle \\ \textsc{f} \begin{bmatrix} \textit{pre-cl-fld} \\ \textsc{r}\ \boxed{1}\ \textit{mtx-fld} \end{bmatrix} \\ \textsc{i-c}\ \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \textsc{phon}\ \langle \textit{by} \rangle \\ \textsc{f} \begin{bmatrix} \textit{cl-be-fld} \\ \textsc{r}\ \boxed{2} \begin{bmatrix} \textit{cl-fld} \\ \textsc{r}\ \boxed{1} \end{bmatrix} \end{bmatrix} \\ \textsc{i-c}\ \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \textsc{phon}\ \langle \textit{se} \rangle \\ \textsc{f} \begin{bmatrix} \textit{cl-rfl-fld} \\ \textsc{r}\ \boxed{2} \end{bmatrix} \\ \textsc{i-c}\ \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \textsc{phon}\ \langle \textit{bát} \rangle \\ \textsc{f} \begin{bmatrix} \textit{rest-fld} \\ \textsc{r}\ \boxed{1} \end{bmatrix} \\ \textsc{i-c}\ \textit{no} \end{bmatrix}, \\ \begin{bmatrix} \textit{s-node} \\ \textsc{phon}\ \langle \textit{velkého} \rangle \\ \textsc{f} \begin{bmatrix} \textit{l-adj-fld} \\ \textsc{r}\ \boxed{3} \end{bmatrix} \\ \textsc{i-c}\ \textit{yes} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \textsc{phon}\ \langle \textit{černého} \rangle \\ \textsc{f} \begin{bmatrix} \textit{l-adj-fld} \\ \textsc{r}\ \boxed{3} \end{bmatrix} \\ \textsc{i-c}\ \textit{yes} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \textsc{phon}\ \langle \textit{vlka} \rangle \\ \textsc{f} \begin{bmatrix} \textit{noun-fld} \\ \textsc{r}\ \boxed{3} \begin{bmatrix} \textit{noun-fld} \\ \textsc{r} \begin{bmatrix} \textit{rest-fld} \\ \textsc{r}\ \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \textsc{i-c}\ \textit{yes} \end{bmatrix} \end{array} \right\rangle$$

Strukturu polí v (10) je lépe vidět ve stromovém znázornění (11).

(11)



V příkladu jsou použita tato topologická pole a oblasti: pre-klitické (počáteční) pole, klitické pole (skládající se zde z pomocného slovesa a reflexiva) a dvě „zbytková pole" (*rest-fld*), jedno pro sloveso a druhé pro jmennou skupinu. Celá věta je jedno pole (oblast): matrix. Pro každou oblast, která se skládá z více polí je definováno pořadí těchto polí. Některá pole se musí vyskytovat vždy vedle sebe (to se týká např. polí ve jmenné skupině a klitických polí, příslušná pole určují oblast, ve které se musí vyskytovat vedle sebe, koindexací). Vzhledem k tomu, že všechna slova daného výrazu jsou uvedena jako prvky jediného seznamu, je možné klást omezení na slovosled a adjacenci monotónním způsobem.

6 Omezení týkající se derivační struktury

Po specifikaci teoretických a formálních základů a definici nezbytných částí signatury lze formulovat omezení týkající se objektů modelu. Skladbu nelexikálních objektů typu *sign* definují čtyři principy: Deep List Composition Principle, Surface List Composition Principle, Valency Principle a Phonology Principle. Následující příklady se týkají prvního z nich. Jeho formální vyjádření (13) lze parafrázovat takto:

(12) 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.

(13)

$$\textit{non-lexical} \;\rightarrow\; \left(\begin{array}{l} \begin{bmatrix} \textsc{synsem} \mid \textsc{local} \mid \textsc{deep} \mid \textsc{tree}\ \boxed{5} \\ \textsc{head-daughter} \mid \textsc{synsem} \mid \textsc{local} \mid \textsc{deep} \mid \textsc{tree}\ \boxed{1} \\ \textsc{nonhead-daughters}\ \boxed{2} \end{bmatrix} \\ \wedge\ \texttt{collect\_dlists}(\boxed{2}, \boxed{3}) \\ \wedge\ \texttt{append}(\boxed{1}, \boxed{3}, \boxed{4}) \\ \wedge\ \texttt{permute}(\boxed{4}, \boxed{5}) \end{array} \right)$$

Vkládání seznamů typu *d-list* je formalizováno pomocí tří relací. Relace `collect_dlists/2` identifikuje *d-list* každé neřídící dcery a přidá ho do seznamu $\boxed{3}$. S tímto „seznamem seznamů typu *d-list*" se sloučí *d-list* řídící dcery ($\boxed{1}$). Výsledkem je $\boxed{4}$, formálně *d-list*. Permutace tohoto seznamu tvoří *d-list* syntaktické matky ($\boxed{5}$) a musí splňovat všechna ostatní omezení kladená na seznamy typu *d-list*. Věta (14) její částečná reprezentace v (15) ukazuje výsledek uplatnění tohoto principu.[1]

(14) Pepa dneska pase sousedovu kozu

(15)
$$\begin{bmatrix} \textit{non-lexical} \\ \textsc{phon}\ \langle \textit{Pepa, dneska, pase, sousedovu, kozu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \langle\ \boxed{2}\langle[\textit{Pepa}]\rangle,\ \boxed{3}\langle[\textit{dnes}]\rangle,\ \boxed{1}[\textit{pást}],\ \boxed{4}\langle\ \boxed{5}[\textit{koza}],\ \boxed{6}\langle[\textit{soused}]\rangle\ \rangle\ \rangle \\ \textsc{hd}\ \begin{bmatrix} \textit{lexical} \\ \textsc{phon}\ \langle \textit{pase} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \langle \boxed{1} \rangle \end{bmatrix} \\ \textsc{nhd}\ \left\langle \begin{bmatrix} \textit{lexical} \\ \textsc{phon}\ \langle \textit{Pepa} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{2} \end{bmatrix}, \begin{bmatrix} \textit{lexical} \\ \textsc{phon}\ \langle \textit{dneska} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{3} \end{bmatrix}, \begin{bmatrix} \textit{non-lexical} \\ \textsc{phon}\ \langle \textit{sousedovu, kozu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{4} \\ \textsc{hd}\ \begin{bmatrix} \textit{lexical} \\ \textsc{phon}\ \langle \textit{kozu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \langle \boxed{5} \rangle \end{bmatrix} \\ \textsc{nhd}\ \left\langle \begin{bmatrix} \textit{lexical} \\ \textsc{phon}\ \langle \textit{sousedovu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{6} \end{bmatrix} \right\rangle \end{bmatrix} \right\rangle \end{bmatrix}$$

7 Tři druhy slovosledných omezení

Slovosledná omezení lze rozdělit do tří skupin: na ta, která se týkají jen jedné z rovin – tektogramatické nebo povrchové – a na ta, která se týkají vztahu mezi těmito dvěma rovinami. Omezení spolupůsobí stejně jako slovoslední činitelé. Není-li relativní pořadí dvou slov v povrchovém slovosledu stanoveno povrchovými omezeními, řídí se hloubkovým slovosledem, který je dán systémovým uspořádáním a specifikací kontextově zapojených a nezapojených uzlů.

Na seznamy typu *d-list* se uplatňují čtyři omezení, která dbají na to, aby v každém stromě byl alespoň jeden kontextově nezapojený uzel, aby v každém podstromě stál řídící uzel na správné pozici, aby kontextově nezapojené uzly následovaly po zapojených a byly uspořádány v souladu s aktuálním členěním.

Vztah mezi hloubkovým a povrchovým slovosledem je definován jediným principem, který se uplatní na každou dvojici uzlů v lokálním tektogramatic-

[1]Sestava rysů obsahuje některé zkrácené zápisy: (a) SS|L|D|T je zkratka za SYNSEM | LOCAL | DEEP | TREE, HD za HEAD-DAUGHTER, a NHD za NONHEAD-DAUGHTERS, (b) jsou uvedeny pouze atributy, které nějak souvisejí s principem DLCP, (c) fonologické podřetězce nejsou koindexovány, a (d) objekty typu *d-node* jsou uvedeny jen jako lemata.

kém stromě za předpokladu, že povrchová pozice ani jednoho z nich není určena povrchovými slovoslednými pravidly. Tuto informaci lze odvodit z toho, jaké topologické pole je danému uzlu přiděleno. Konsekvent tohoto principu obsahuje tři možnosti: (a) relativní pořadí obou uzlů je na obou rovinách stejné (uplatní se Mathesiův princip aktuálního členění), nebo (b) je-li první z nich kontextově nezapojený a na počáteční pozici v seznamu *d-list*, může mu být přiděleno preklitické topologické pole a může se objevit v doméně vyšší klauze (čímž dojde k narušení principu členské sounáležitosti), nebo (c) je-li druhý z nich v seznamu *d-list* na poslední pozici, může být umístěn v seznamu *s-list* na jiném než posledním místě a označen jako součást intonačního centra (uplatní se princip důraznosti).

Ostatní slovosledné zákonitosti jsou věcí povrchových omezení. Seznamy typu *s-list* podléhají pěti obecným a dalším specifickým omezením pro konstrukce určitého druhu. Obecná omezení na *s-list* jsou upravené verze obecných omezení na seznamy domén z [2]. Jde o principy Matrix Compaction, Planarity, Topological Order, Field Existence a Field Uniqueness. Tři poslední jsou zobecněné tak, aby umožňovaly odděleně definovat skladbu topologických oblastí: pořadí polí a kolikrát se pole určitého typu může nebo musí v oblasti objevit. V tabulce 1 je definována nejvyšší oblast *matrix-fld*.

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

Tabulka 1: Pole v nejvyšší oblasti

Specifická omezení seznamů typu *s-list* se téměř vždy opírají o specifikace topologických polí a v některých případech také o morfosyntaktické vlastnosti dotčených větných členů. Řeší několik vybraných případů nespojitých realizací tektogramatických podstromů: komparativní konstrukce (*menší vesnice než Lhota*), elementární případy vzdálených závislostí, rozdělené předložkové fráze (*O jakou se jedná soutěž?*) a také postavení větných příklonek na druhé pozici ve větě, včetně pokusu o řešení „šplhání“ příklonek a jejich haplologie. Příklonkám je věnována zvláštní kapitola s mnoha příklady.

Navržená řešení jsou všechna založena na specifikaci topologických polí v objektech typu *d-node*. To nemusí být optimální přístup, zvláště pro případy vzdálených závislostí a „šplhání“ příklonek a haplologii, kde může být výhodnější řešení založené na syntaktické struktuře a nikoli na povrchovém seznamu.

8 Závěr

Zbývá ještě spousta nedořešených úkolů. Mnohé nakousnuté problémy by zasluhovaly empirický výzkum s oporou korpusu a hlubší analýzu. Navrženým řešením by nepochybně prospělo řádné ověření, nejlépe v implementovaném systému. Přesto si dovolím uvést několik pozitivních výsledků:

a. Byly uvedeny argumenty ve prospěch spojení FGP jako nosné lingvistické teorie a RSRL jako adekvátního formálního jazyka. Tento rámec byl použit pro popis řady jevů českého slovosledu. Byl navržen popis vztahu mezi hloubkovým a povrchovým slovosledem s řešením interakce hloubkového slovosledu s povrchovými slovoslednými zákonitostmi.
b. Byla doložena kompatibilita Mathesiových slovosledných principů s použitým deklarativním formalismem. Vedoucí úloha aktuálního členění a hierarchie výpovědní dynamičnosti při stanovení povrchového slovosledu a prosodických charakteristik výpovědi byla vyjádřena v podobě principů, které působí v interakci s ostatními slovoslednými principy.
c. Kompozicionální základ vztahu mezi povrchovým řetězcem a tektogramatickou reprezentací byl popsán jako plochá derivační struktura. Tektogramatická reprezentace byla formalizována jako seznam s rekurzivně vnořenými seznamy, morfématický řetězec jako seznam s údaji o adjacenci a pořadí zakódovanými do jednotlivých prvků seznamu.
d. Empirická data a teoretické premisy byly formalizovány způsobem, který umožňuje vzájemnou interakci faktorů ovlivňujících povrchový slovosled. Hloubkový slovosled přitom ovlivňuje povrchový slovosled, kdykoli jiná omezení nejsou kladena. K tomu bylo použito několika obecných a většího počtu specifických omezení, která kladou podmínky na topologická pole, tektogramatickou reprezentaci a v některých případech také na morfosyntaktické vlastnosti dotčených větných členů.
e. Bylo shrnuto a částečně formalizováno několik zákonitostí, jimiž se řídí větné příklonky v češtině, které potvrzují interakci syntaktických, diskursních a prosodických faktorů, přičemž syntaktické faktory hrají hlavní roli.

Ne všechny cíle se podařilo splnit. Implementace popisu fragmentu češtiny představuje další logický krok, spolu s nezbytným přeformulováním některých omezení do podoby, která je výpočetně jednodušší, ale přitom intuitivní a teoereticky adekvátní. Takový krok je žádoucí jednak pro ověření popisu, ale také pro odhad možností dalšího vývoje.

Ne všechny zkoumané jevy se podařilo uspokojivě popsat. To se týká zejména jevů uvedených v kapitole o příklonkách. Výsledkem může být modifikace některých více či méně podstatných formálních nebo teoretických aspektů. Avšak před takovým krokem se počítá s rozsáhlejší aplikací na jiné jevy, nejen slovosledné a ne jen české.

Literatura

[1] Vilém Mathesius. On Information-Bearing Structure of the Sentence. In Susumu Kuno, editor, *Harvard Studies in Syntax and Semantics*, strany 467–480. 1975.

[2] Gerald Penn. Linearization and WH-extraction in HPSG: Evidence from Serbo-Croation. In Robert D. Borsley a Adam Przepiórkowski, editoři, *Slavic in HPSG*, Studies in constraint-based lexicalism, strany 149–182. CSLI Publications, Stanford, 1999a.

[3] Carl J. Pollard a Ivan A. Sag. *Information-Based Syntax and Semantics.* Číslo 13 v řadě CSLI Lecture Notes. CSLI, Stanford, California, USA, 1987.

[4] Carl J. Pollard a Ivan A. Sag. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, 1994.

[5] Frank Richter. A mathematical formalism for linguistic theories with an application in head-driven phrase structure grammar. Dissertation, submitted to the Neuphilologische Fakultät of the Universität Tübingen, Version of April 28th, 2000, 2000.

[6] Petr Sgall. Underlying structure of sentences and its relations to semantics. In Tilmann Reuther, editor, *Festschrift für Viktor Jul'evič Rozenberg*, strany 273–282. Wien, 1992. Wiener Slawistischer Almanach, Sonderband 33.

[7] Petr Sgall, Eva Hajičová a Jarmila Panevová. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects.* Reidel and Academia, Dordrecht and Praha, 1986. Editor: Jacob Mey.