J. Yaghob (Ed.): ITAT 2015 pp. 43–50
Charles University in Prague, Prague, 2015

Redukční analýza a Pražský závislostní korpus*

Martin Plátek[1], Dana Pardubská[2], and Karel Oliva[3]

[1] MFF UK Praha, Malostranské nám. 25, 118 00 Praha, Česká Republika
martin.platek@ufal.mff.cuni.cz
[2] FMFI UK Bratislava, Mlynská dolina, 84248 Bratislava
pardubska@dcs.fmph.uniba.sk
[3] UJČ ČAV Praha, Letenská, 118 00 Praha, Česká Republika
oliva@ujc.cas.cz

Abstrakt: Cílem tohoto příspěvku je uvést, formálně zavést a exaktně pozorovat větnou redukční analýzu svázanou s redukční analýzu D-stromů. Tímto způsobem upřesníme strukturální vlastnosti D-stromů se závislostmi a koordinacemi z Pražského závislostního korpusu (PDT). Zvýrazňujeme vlastnosti, kterými se závislosti a koordinace liší. Snažíme se pracovat metodou, která je blízká metodám matematické lingvistiky, a to především těm, které formulují omezující podmínky pro syntaxi přirozených jazyků. Ukazujeme nové možnosti takových formulací.

1 Úvod

Postupně se věnujeme *větné redukční analýze* (RA) a její vazbě na *redukční analýzu D-stromů (*RADS*)*, abychom získali nové formální prostředky vhodné pro studium strukturálních vlastností D-stromů. Na základě těchto prostředků formulujeme pozorování o D-stromech v Pražském závislostním korpusu (PDT viz [1]). Tento článek vznikl ve spolupráci s Markétou Lopatkovou, která nám pomocí vybíraných příkladů zprostředkovala přístup do PDT a často s námi diskutovala, zvláště o problematice redukcí stromů z PDT s koordinacemi.

1.1 Neformální úvod do (manuální) redukční analýzy českých vět a redukční analýzy jejich D-stromů

V této sekci se pokusíme čtenáře neformálně uvést do problematiky manuální redukční analýzy vět a poukázat na souvislosti s redukční analýzou D-stromů, které těmto větám odpovídají. Redukční analýzou českých vět a jejímu modelování se zabýváme již delší dobu (viz např. [3, 5]), naopak explicitní zmínky o redukční analýze D-stromů se objevují ponejprv na loňském ITATU (viz [4, 2]). Při formalizaci obou typů redukčních analýz zvýrazňujeme jejich minimalistický charakter a využíváme ho při strukturální charakterizaci D-stromů.

RA je založena na postupném zjednodušování analyzované věty po malých krocích, viz [3, 5]. RA definuje možné posloupnosti větných redukcí – každá redukce RA spočívá ve *vypuštění* několika slov, nejméně však jednoho slova analyzované věty. V některých redukcích může být kromě vypouštění použita operace *shift*, která přesune nějaké slovo na novou pozici ve větě.

Metoda (manuální) redukční analýzy, studovaná v tomto příspěvku, dodržuje následující zásady:

(i) tvary jednotlivých slov (i interpunkčních znamének), jejich morfologické charakteristiky i jejich syntaktické kategorie se nemění během RA;

(ii) gramaticky správná věta (přesněji její čtení) musí zůstat správná i po redukci;

(iii) vynecháme-li z libovolné redukce jednu či více operací vypuštění nebo shift, nastane porušení principu zachování správnosti (ii);

(iv) předložkové vazby (např. 'o otce'), se vynechávají celé (jinak je možný posun významu, často i změny v pádech);

(v) věta, která obsahuje správnou větu (nebo její permutaci) jako svoji (případně nesouvislou) podposloupnost, musí být dále redukována;

(vi) redukce používají operaci shift jenom v případech vynucených principem zachování korektnosti, tedy v případech, kdy vynechání shiftu by vedlo k nekorektnímu větnému slovosledu;

(vii) syntaktická struktura věty po redukci zachovává strukturu věty před redukcí.

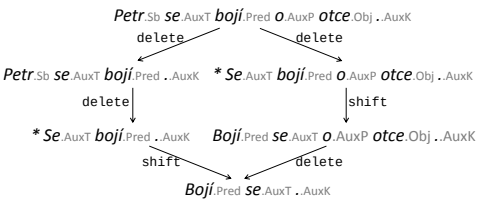
Novým prvkem mezi zásadami pro větnou redukční analýzu oproti [5] je položka (vii). Syntaktická struktura zde znamená větný rozbor odpovídající stromům z Pražského závislostního korpusu (D-strom). Tato zásada fakticky formuluje základní vztah mezi větnou redukční analýzou a redukční analýzou D-stromů. Výše uvedené zásady postupně upřesníme ve formální části příspěvku.

V následujících odstavcích uvedeme serii příkladů ilustrujících prvky redukční analýzy, které se týkají redukcí zjednodušující jak závislosti, tak především koordinace. Všimněme si, že redukce koordinací budou ve dvou aspektech složitější než redukce závislostí. Pozorování koordinačních jevů a formalizace těchto pozorování je hlavní novinkou a přínosem tohoto příspěvku.

*Příspěvek prezentuje výsledky dosažené v rámci projektu agentury GAČR číslo GA15-04960S.

D-stromy na našich obrázcích se liší od D-stromů z PDT jen ve dvou aspektech. Za prvé: neobsahují identifikační uzel, který nenese žádnou syntaktickou informaci a neodpovídá žádnému slovu věty. Za druhé: značka 'Coord' je nahrazena značkou 'Cr'.

Příklad 1.
(1) Petr.Sb se.AuxT bojí.Pred o.AuxP otce.Obj ..AuxK

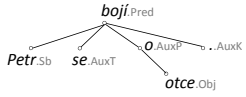


Obrázek 1: Schema RA pro větu (1).

Z obrázku 1 vidíme, že věta (1) může být v prvním kroku redukována dvěma způsoby:
(i) buď vypuštěním předložkové vazby 'o otce'; této větné redukci odpovídá redukce D-stromu T_1 z obrázku 2 na D-strom T_2 z obrázku 3,
(ii) nebo vypuštěním podmětu (subjektu) 'Petr', to však vede k větě se špatným slovosledem. Gramatické české věty nemohou začínat klitikou. To vede k použití přesunu klitiky 'se' na druhou pozici ve větě. Získáme tak korektní větu 'Bojí se o otce.' Této větné redukci odpovídá redukce D-stromu T_1 na D-strom T_4 z obrázku 5.

Potom pokračují redukce podobným způsobem v obou větvích, až dospějeme k neredukovatelné správné větě 'Bojí se.' . Této fázi odpovídají redukce D-stromů T_2 a T_4 na D-strom T_3 z obrázku 4.

Předchozí příklad ilustruje přirozenou souvislost mezi větnou redukční analýzou věty (1) a redukční analýzou D-stromu [illegible] obrázku 2.

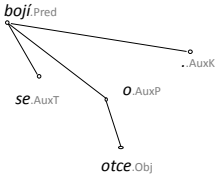


Obrázek 2: Závislostní strom T_1.

Obrázek 3: T_2, vzniklé redukcí z T_1.

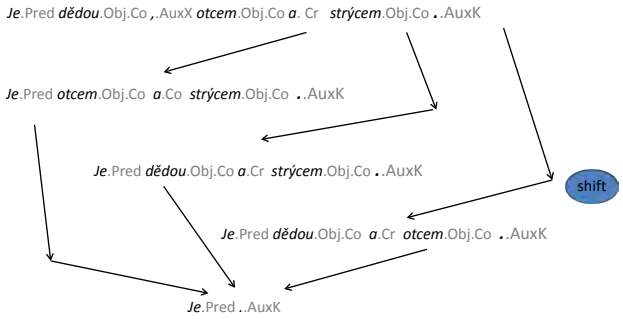
Příklad 2. *Na obrázku 6 vidíme schema redukční analýzy věty (2). Věta (2) obsahuje trojnásobnou koordinaci předmětů. Povšimněme si, že dalšímu zjemnění schematu zabraňují kategorie (značky), použité podle vzoru PDT. Značka 'Cr' znamená koordinující symbol (slovo), 'Co' značí koordinované slovo, či symbol. Schematu na obrázku odovídají redukce D-stromů, které reprezentují obrázky 7*

Obrázek 4: T_3 vzniklé redukcí se shiftem z T_2 nebo redukcí bez shiftu z T_4.



Obrázek 5: T_4, vzniklé redukcí z T_1.

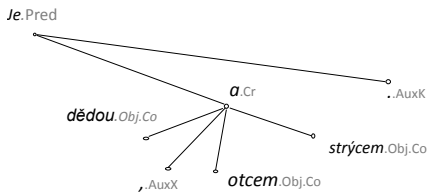
až 11. Všechny tři redukce D-stromu Tc_1 odstraňují (při zjednodušování trojnásobné koordinace na dvojnásobnou) dva nesouvisející uzly (podstromy). Třetí redukce navíc používá shift. Tyto redukce se liší od předchozího příkladu, kde všechny redukce odtrhly jediný úplný souvislý podstrom. Zbylé redukce dvojnásobných koordinací se realizují odtržením souvislého úplného podstromu, určeného jejich vrcholem, podobně jako u redukcí v předchozím příkladě, týkající se závislostí.



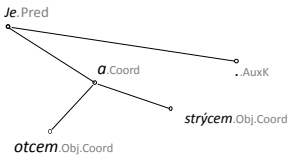
Obrázek 6: RA věty (2) s vícenásobnou koordinací.

Příklad 3. *Na obrázku 12 vidíme schema redukční analýzy věty (3). Toto schema znázorňuje jedinou redukci, která odstraňuje koordinovaná příslovečná určení, která jsou závislá na koordinovaných predikátech. Odpovídající redukci D-stromu ilustrují obrázky 13 a 14.*

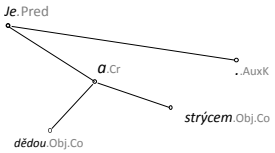
Příklad 4. *Na obrázku 15 vidíme schema redukční analýzy věty (4). Věta (4) je věta s vloženou koordinací. D-stromy zachycující odpovídající redukční analýzu D-stromů jsou na obrázcích 16 až 18. Vložená koordinace se v D-stromě Tcz_3 zjednodušuje tak, že se vyjme jedna hrana s řídícím uzlem se značkou 'Cr:Co' (ve složitějších případech i to co na ní visí). To odpovídá dvěma redukcím ve větné redukční analýze z obrázku 15. Tento typ redukce je nový oproti předchozím případům a je vynucen principy zachování korektnosti a minimality ve větné redukční analýze.*



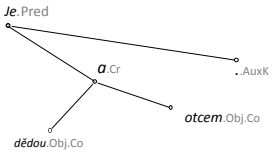
Obrázek 7: D-strom Tc_1.



Obrázek 8: Tca_2, vzniklé redukcí z Tc_1.

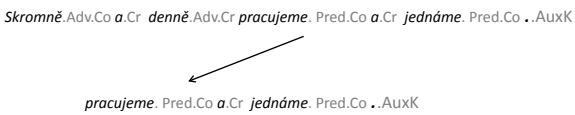


Obrázek 9: Tcb_2, vzniklé redukcí z Tc_1.

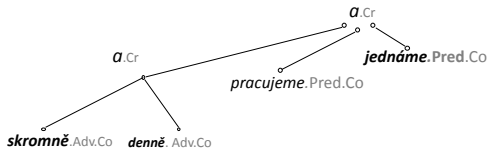


Obrázek 10: Tcc_2, vzniklé redukcí z Tc_1.

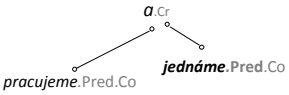
Obrázek 11: Tc_3, vzniklé redukcí z Tca_2, Tcb_2 a Tcc_2.



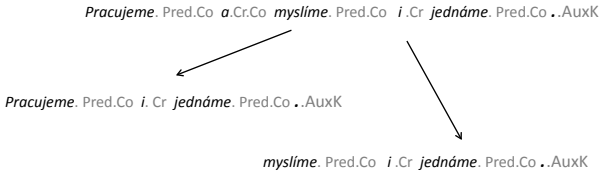
Obrázek 12: RA závislé koordinace na řídící koordinaci.



Obrázek 13: Tcz_2



Obrázek 14: Tcz_{22}, vzniklé redukcí z Tcz_2.



Obrázek 15: AR věty s vloženou koordinací.

2 Formalizace

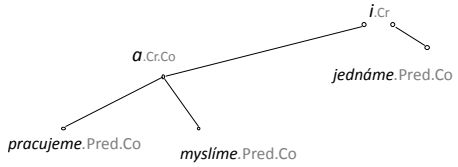
Formalizace RA přirozených jazyků začíná formalizováním lexikální analýzy těchto jazyků. Lexikální analýza kromě jiného umožňuje rozlišovat možnosti uplatnení jednotlivých typů redukcí.

2.1 Lexikální analýza

Při formalizaci lexikální analýzy pracujeme se třemi abecedami (slovníky)- konečnými množinami slov. Σ_p, tzv. slovník [1], se využívá na modelování jednotlivých slovních forem. Σ_c označuje abecedu kategorií, například syntaktických značek v PDT. Kombinací dostáváme hlavní slovník $\Gamma \subseteq \Sigma_p \times \Sigma_c$, který umožňuje odstraňovat lexikomorfologické nejednoznačnosti jednotlivých slovních forem. Lexiko-morfologicky zjednoznačněná věta tedy vstupuje do RA jako retězec nad slovníkem Γ.

Projekce z Γ^* do Σ_p^* resp. do Σ_c^* přirozeně definujeme pomocí homomorfismů: *slovníkovým homomorfismem* $h_p : \Gamma \to \Sigma_p$ a *kategoriálním homomorfismem* $h_c : \Gamma \to \Sigma_c$: $h_p([a,b]) = a$ a $h_c([a,b]) = b$ pro všechny $[a,b] \in \Gamma$.

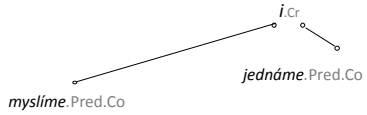
Příklad 5. *Definované pojmy ilustrujeme na příklade, který vychází z příkladu 1*
Slovník: $\Sigma_p^1 =$ *{ Petr, se, bojí , o, otce, . }*
Abeceda kategorií: $\Sigma_c^1 =$ *{ Sb, AuxT, Pred, AuxP, Obj, AuxK}*
Hlavní slovník: $\Gamma^1 = \{b_1$= *[Petr,Sb],* b_2=*[se,AuxT],* b_3= *[bojí,Pred],* b_4=*[o,AuxP],* b_5= *[otce,Obj],* b_6=*[.,AuxK]*$\}$



Obrázek 16: Tcz_3

[1] Index p při označení abecedy se vztahuje na anglickou verzi, kde se používá slovo proper

Obrázek 17: Tcz_{31}, vzniklé redukcí z Tcz_3.



Obrázek 18: Tcz_{32}, vzniklé redukcí z Tcz_3.

V abecedě kategorií v tomto příkladě jsou jen závislostní kategorie (ne všechny). Koordinační kategorie vznikají kombinacemi se značkami 'Cr', 'Co'.

2.2 Formální RA

V této sekci zavádíme postupně formální redukční analýzu vět (řetězů) RA a formální redukční analýzu pro D-stromy.

Nejprve zavedeme na jazyce L tzv. DS-redukci $\succ_L$. Nechť u,v jsou řetězce. Říkáme, že u je větší než v vzhledem k jazyku L a označujeme $u >_L v$ pokud:

- $u,v \in L$ a $|u| > |v|$;
- v je permutace nějaké podposloupnosti u.

Říkáme, že v je DS-redukce u vzhledem k jazyku L a označujeme $u \succ_L v$ pokud:

- $u >_L v$ a neexistuje žádné $z \in L$ takové, že $u >_L z >_L v$, t.j., platí princip minimality redukcí.

Reflexívní a tranzitívní uzávěr relace $\succ_L$ označujeme $\succ_L^*$. Částečné uspořádání $\succ_L$ přirozeně definuje

- $L^0_\succ = \{v \in L \mid \neg\exists u \in L : v \succ_L u\}$ - množinu ireducibilních vět jazyka L
- $L^{n+1}_\succ = \{v \in L \mid \exists u \in L^n_{\succ_L} : u \succ_L v\} \cup L^n_\succ, n \in N$ - množina těch vět z jazyka, které je možné zredukovat na ireducibilní větu z jazyka posloupností DS-redukcí délky nanejvýš $n+1$.

Množinu $\succ_L$=$\{u \succ_L v \mid u,v \in L\}$ nazveme množinou DS-redukcí jazyka L. Analogicky pro větu w jazyka L nazveme $\succ_L(w)$ =$\{u \succ_L v \mid w \succ_L^* u\}$ DS-redukční množinou věty w.

Fakt: $\succ_L$ aj $\succ_L(w)$ jsou jednoznačne určené L, resp. w a L.

Přistupme k formalizaci (minimalistické) redukční analýzy. Říkáme, že relace $\rhd_L \subseteq \succ_L$ je *DS-(redukční) analýza jazyka* L pokud $L = L^0_\succ \cup \{v \mid \exists u,z : \; v \rhd_L u \rhd_L^* z \in L^0_\succ\}$. Analogicky definujeme DS-analýzu $\rhd_L(w)$ pro $w \in L$; $\rhd_L(w)$ =$\{u \rhd_L v \mid w \rhd_L^* u\}$.

Uvědomme si, že zatím co jazyk L je jednoznačne určený pomocí $\rhd_L$ a $L^0_\succ$, věta $w \in L$ může mít více DS-analýz. Různé DS-analýzy věty w v lingvistice odpovídají různému čtení (porozumění) této věty.

Relace $\rhd_L$ určuje velikost zkrácení, které je možné dosáhnout jedním krokem redukce. Říkáme, že $\rhd_L$ a L jsou *k-omezené* pokud délka slov z $L^0_\succ$ je nejvýše k a $|u|-|v| \leq k$ pro všechny $u \rhd_L v \in \rhd_L$.

Bylo by zvláštní, kdyby v DS-redukci přirozeného jazyka byly ireducibilní věty dlouhé, pričemž všechny redukce z $\rhd_L$ by zkracovaly věty jen málo. Zajímáme se proto hlavně o takové DS-analýzy, v kterých $\forall w \in L^0_\succ$ existují u, v, $u \rhd_L v$ takové, že $|u|-|v| \geq |w|$. Takovým DS-analýzám říkáme *proporcionální*.

Všimněme si, že redukční analýza české věty z příkladu 1 vyhovuje podmínkám kladeným na proporcionální 2-omezenou DS-analýzu, zatímco redukční analýza české věty z příkladu 2 je proporcionální 3-omezenou DS-analýzou.

DS-analýzu budeme považovať za relevantní model skladby přirozených i umělých jazyků, pokud to bude DS-analýza konečných, anebo nekonečných semi-lineárnych jazyků, které jsou proporcionální a k-ohraničené pro nejaké neveliké k.

2.3 D-struktury a D-stromy

V následující části zavedeme tzv. D-struktury a D-stromy, ktoré jsou grafovou reprezentací struktury vět a jejich odvození.[2] D-struktura reprezenuje syntaktické jednotky (slova a jejich kategorie použité v príslušné větě) jako vrcholy grafu a vzájemné syntaktické vztahy mezi nimi hranami; pořadí slov je určené totálním uspořádáním vrcholů.

D-struktura na Γ je trojice $D = (V,E,ord(V))$, kde (V,E) je orientovaný acyklický graf, V konečná množina jeho vrcholů a $E \subset V \times V$ konečná množina jeho hran. *Vrchol* $u \in V$ je dvojice $u = [i,a]$, kde $a \in \Gamma$ je symbol (slovo) spolu s přirazenými kategoriemi, i (*index/ identifikační číslo*) je přirozené číslo sloužící pro jednoznačnou identifikaci vrcholu u a $ord(V)$ je totální uspořádání na V, obvykle popsané uspořádaným seznamem prvků z V.

Hrany D-struktury interpretujeme jako syntaktické vztahy mezi odpovídajícími lexikálními jednotkami, uspořádání $ord(V)$ reprezentuje pořadí slov v modelované větě. Je-li $ord(V) = \{[i_1,a_1],\cdots,[i_n,a_n]\}$, tak $w = a_1 \cdots a_n$ je řetězec (resp. věta), který označujeme $St(D) = w$, a říkáme, že je projekcí D-struktury D.

Říkáme, že D-struktura $D = (V,E,ord(V))$ je *normalizovaná*, pokud $ord(V) = ([1,a_1],[2,a_2],\cdots,[n,a_n])$ pro nejaké $a_1,\cdots,a_n$. *Normalizace* D-struktury $D = (V,E,ord(V))$ je taková normalizovaná D-struktura $D_1 = (V_1,E_1,ord(V_1))$, pro kterou (V,E) a (V_1,E_1) jsou izomorfní a $St(D) = St(D_1)$. Všimněme si, že normalizace D-struktury je jednoznačně daná.

Dve D-struktury jsou ekvivalentní pokud mají stejnou normalizaci. Ekvivalentní D-struktury obvykle nebudeme rozlišovat. Uvidíme, že nenormalizované D-struktury (stromy) získáme z normalizovaných pomocí operací, které zavedeme.

[2] prefix Dje převzatý z anglických pojmů Delete a Dependency.

Vzhledem k charakteru zkoumané problematiky budeme většinou pracovat se stromovými D-strukturami. Říkáme, že D-struktura $D = (V,E,ord(V))$ nad Γ je D-strom nad Γ pokud (V,E) je kořenový strom (t.j., všechny maximální cesty (V,E) začínají v listech a končí v jediném kořeni).

Budeme pracovat s redukcemi D-stromů - relace $\sqsupset$ a $\vdash$ definované na D-stromech souvisí s realizací různých typů redukcí. Nechť $D = (V,E,ord(V))$, $D_1 = (V_1,E_1,ord(V_1))$ jsou D-stromy.

$D \sqsupset D_1$ pokud
(1) (V_1,E_1) je podstrom (V,E)
(2) V_1 obsahuje kořen D
(3) $ord(V_1)$ je permutace podposloupnosti $ord(V)$.

$D \vdash D_1$, pokud podmínku (1) nahradíme dvěma podmínkami
(1a) $V \subset V_1$
(1b) $\forall v_1, v_2 \in V_1$ platí, že pokud existuje cesta z v_1 do v_2 ve stromě (V,E) tak existuje také cesta z v_1 do v_2 i ve stromě (V_1,E_1).

Příklad 6. *Následuje popis* D*-stromů T_1 a T_2, které reprezentují obr. 2 a obr. 3:*

$T_1 = (V_1,E_1,ord(V_1))$*, pričemž*

$V_1 = \{[1,b_1],[2,b_2],[3,b_3],[4,b_4],[5,b_5],[6,b_6]\}$

$E_1 = \{([1,b_1],[3,b_3]),\ ([2,b_2],[3,b_3]),\ ([4,b_4],[3,b_3]),\ ([5,b_5],[4,b_4]),\ ([6,b_6],[3,b_3])\}$,

$ord(V_1) = ([1,b_1],[2,b_2],[3,b_3],[4,b_4],[5,b_5],[6,b_6])$

$T_2 = (V_2,E_2,ord(V_2))$*, pričemž*

$V_2 = \{[1,b_1],[2,b_2],[3,b_3],[6,b_6]\}$

$E_2 = \{([1,b_1],[3,b_3]),\ ([2,b_2],[3,b_3]),\ ([6,b_6],[3,b_3])\}$

$ord(V_2) = ([1,b_1],[2,b_2],[3,b_3],[6,b_6])$

Je snadno vidět, že $T_1 \sqsupset T_2$.

Takřka všechny neformální redukce z kapitoly jedna vedou k realizaci relace $\sqsupset$. Neplatí to jen pro redukce na obr. 17 a 18. Tyto redukce splňují obecnější relaci $\vdash$. Tyto dvě relace reprezentují dvě varianty zachování zbylé D-struktury, vzniklé zmenšením při uplatnění redukcí redukční analýzy na D-stromech.

Nechť T je nejaká množina D-stromů na Γ. Říkáme, že T tvoří T-jazyk na Γ a píšeme $\mathsf{T} \subseteq T(\Gamma)$. Analogicky, množinu $St(\mathsf{T}) = \{St(t) \mid t \in \mathsf{T}\}$ nazýváme *projekcí* T, množina $h_p(St(\mathsf{T})) = \{h_p(St(t)) \mid t \in \mathsf{T}\}$ je vlastní jazyk pro T, a $h_c(St(\mathsf{T})) = \{h_c(St(t)) \mid t \in \mathsf{T}\}$ je kategoriální jazyk pro T.

Zavedeme tři operace pro práci s D-stromy. Umožní nám realizovat typ redukcí čistě závislostních i redukce různých typů koordinací.

Najjednodušší operací je tzv. *shift*, což je takový posun některého vrcholu D-stromu $D = (V,E,ord(V))$ na nové místo v $ord(V)$, který zachová stromovou strukturu D, tedy zachová všechny uzly z V a všechny hrany z E.

Druhou operaci nazveme *UNC*, z anglického *upper-node-cut*. Je typická pro redukce závislostí a při jejím zavádění si pomůžeme jednodušší operací *LNC*, z anglického *lower-node-cut*. Operace UNC i LNC jsou určené uzlem u D-stromu různým od kořene. Tento uzel jednoznačně určuje rozklad D-stromu D na dva podstromy:

1) $T_L(u,D)$ označuje výsledek LNC aplikovaného na D v uzlu u ; je to podstrom stromu D, který tvoří uzly ležící na nějaké cestě z listu do u (včetně u). Pořadí uzlů v $T_L(u,D)$ je určené pořadím v D.

2) $T_U(u,D)$ označuje výsledek UNC aplikovaného na D v uzlu u; je to maximální podstrom D obsahující kořen D a všechny uzly mimo $T_L(u,D)$. Pořadí uzlů je určené poradím v D. UNC tedy transformuje D na D-strom $T_U(u,D)$.

Poslední operací je *UEC*, z anglického *upper-edge-cut.* Použití této operace jsme videli při redukci (odstraňování) vložených koordinací z obr. 17 a 18. Nechť (u,v) a (v,v_1) jsou takové hrany D-stromu D, že existuje právě jeden uzel $u_1 \neq u$ a hrana (u_1,v) vedoucí do v. Operace UEC aplikovaná na D podle hrany (u,v) vytvoří D-strom $T_E((u,v),D)$. $T_E((u,v),D)$ získáme následujícím způsobem: nejprve aplikací UNC-operace vytvoříme $T_U(u,D)$ a následně z něj odstraníme uzel v spolu s hranami (u,v) a (v,v_1). Potom spojíme vrcholy u_1, v_1 novou hranou (u_1,v_1) a získáme tak D-strom, který označujeme $T_E((u,v),D)$.

Nyní zavádíme formální redukce a redukční analýzu na D-stromech tak, abychom pokryli jak závislostní, tak koordinační jevy z PDT.

Nechť $\mathsf{T} \subseteq T(\Gamma)$, $t_1,t_2 \in \mathsf{T}$. Symbolem $\vdash_T$ budeme označovat zúžení operace $\vdash$ na T[3]

Říkáme, že t_1 je *NES-redukované na* $t_2 \in \mathsf{T}$ a označujeme $t_1 \hookrightarrow_{NES} t_2$, pokud redukci $t_1 \vdash_T t_2$ umíme popsat pomocí množiny O_N UNC-operací a/nebo množiny O_E UEC-operací, případně následovanými množinou shiftů O_S. Navíc, $O_N \cup O_E$ je neprázdná, každý uzel je operací z O_s přesouvaný nejvýše jednou, O_s může být prázdná.

Pokud v predchozí definici nepovolíme UEC-operace, budeme říkat, že t_1 je *NS-redukované* na $t_2 \in \mathsf{T}$ a označovat $t_1 \hookrightarrow_{NS} t_2$.

Pokud při redukci nepovolíme ani shifty, budeme hovořit o *N-redukci* a označovat $t_1 \hookrightarrow_N t_2$.

Redukce typu NES, NE a N mohou být, v principu, aplikované na libovolné D-stromy. Nás však zajímají redukce D-stromů daného T-jazyka, proto vyžadujeme, aby i po aplikování zmíněných redukcí byl vzniklý strom platným D-stromem zkoumaného jazyka. Při definování pojmu redukce proto přidávame parametr T.

Nechť $X \in \{NES,NS,N\}$, $\mathsf{T} \subseteq T(\Gamma)$. Říkáme, že t_1 je (X,T)-redukované na t_2 a píšeme $t_1 \gg_{(T,X)} t_2$ pokud:

- $t_1,t_2 \in \mathsf{T}$
- $t_1 \hookrightarrow_X t_2$ a neexistuje $z \in \mathsf{T}$ tak, aby $t_1 \hookrightarrow_X z \hookrightarrow_X t_2$, t.j., platí princíp minimality redukcí.

[3] Při $\vdash_T$ tedy vyžadujeme, aby t_1 i t_2 byli z T.

Tranzitívní, reflexívní uzávěr $\gg_{(T,X)}$ označujeme $\gg^*_{(T,X)}$. Tranzitívní, anti-reflexívní uzávěr $\gg_{(T,X)}$ označujeme $\gg^+_{(T,X)}$. V situaci, kdy je Tzřejmé z kontextu, hovoříme jen o NES-, NE, resp. N-redukci.

Uvědomme si, že T a X jednoznačně určují množinu $\gg_{(T,X)} = \{u \gg_{(T,X)} v \mid u,v \in T\}$, ktorou považujeme za redukční analýzu T-jazyka T. Říkáme, že $\gg_{(T,X)}$ je X-redukcí T. Všimněme si rozdílu oproti DS-analýze retězcových jazyků, která nebývá jednoznačně určená svým jazykem.

Nechť $X \in \{NES, NS, N\}$.

$(T,X)^0_{\gg} = \{t \in T \mid \neg\exists s \in L : t \gg_{(T,X)} s\}$,

$(T,X)^{n+1}_{\gg} = \{v \in T \mid \exists u \in (T,X)^n_{\gg} : v \gg_{(T,X)} u\} \cup T^n_{\gg}$.

Nechť $t \in T$. Píšeme $\gg_{(T,X)}(t) = \{u \gg_{(T,X)} v \mid t \gg^*_{(T,X)} u\}$. Říkáme, že $\gg_{(T,X)}(t)$ je X-analýza (redukční) D-stromu t.

V následující sekci budeme navíc ještě vázat použití jednotlivých typů operací na (ne)přítomnost koordinačních značek v určujících hranách a uzlech těchto operací. O takových typech omezení uplatnění operací jsme zatím nemluvili.

2.4 Principy, vlastnosti a pozorování

Zde zavedeme principy, které nám umožní formulovat požadavky na redukční analýzu na D-stromech a formulovat pozorování o jejich plnění na stromech z PDT. Při těchto pozorováních uplatníme možnost porovnávat NES-analýzy, NS-analýzy a N-analýzy D-stromů a využijeme tato porovnání pro charakterizaci (klasifikaci) těchto D-stromů.

Princip S-kompatibility. Nech $X \in \{NES, NS, N\}$. Pokud platí, že $t_1 \gg_{(T,X)} t_2$ a zároveň platí, že $Str(t_1) \succ_{Str(T)} Str(t_2)$, tak říkáme, že redukce $t_1 \gg_{(T,X)} t_2$ je S-kompatibilní. Neformálně řečeno, pokud redukci D-stromů odpovídá řetězová redukce na řetězech získaných projekcí ze stromů, která je vztažena k jazyku řetězů $Str(T)$, daných množinou stromů T.

Podobně říkáme, že $\gg_{(T,X)}(t)$ je S-kompatibilní, pokud všechny jeho X-redukce jsou S-kompatibilní a pokud za předpokladu $u \in (T,X)^0_{\gg}$ a $t \gg^*_{(T,X)} u$ platí, že $Str(u) \in Str(T)^0_{\succ}$.

Říkáme, že X-analýza $\gg_{(T,X)}$ je S-kompatibilní pokud všechny její D-stromy mají S-kompatibilní X-analýzu.

Fakt. Vidíme, že $\gg_{(T,X)}(t)$ je S-kompatibilní pokud $Str(\gg_{(T,X)}(t)) = \{Str(u) > Str(v) \mid u \gg_{(T,X)} v \in \gg_{(T,X)}(t)\}$ tvoří DS-analýzu věty $Str(t)$ vzhledem k jazyku $Str(T)$.

Princip S-kompatibility je tak požadavkem, který zaručuje přirozený vztah mezi větnou DS-analýzou a X-analýzami na D-stromech.

Fakt. Uvažujeme NS-analýzu A D-stromu t. Platí, že uzel u, který je ve stromě t na cestě ke kořenu blíže než uzel v, nemůže být v žádne větvi NS-analýzy A vypuštěn dříve než v.

Tento fakt přímo vyplývá z definice UNC-operace.

Předchozí fakt zpřesňuje intuitivně vnímané vlastnosti (ne)závislostí v (čistě) závislostních stromech.

Následující dva principy jsou blízké algebraickému principu konfluence.

Princip Tl-kompatibility. Požadujeme, aby všechny větve v NES-analýze A stromu t byly stejně dlouhé a v každé větvi byl použit stejný počet UNC-operací a UEC-operací.

Následujicí princip je přísnější. Odlišuje čistě závislostní D-stromy od D-stromů s koordinacemi.

Princip Ta-kompatibility (Formulace závislostniho principu). Tento princip uvažuje pouze D-stromy t, které nemají koordinační znaky, a jejichž NES-analýzy jsou i NS-analýzami a zároveň splňují princip Tl-compatibility. Dále zde požadujeme, aby množina UNC-operací užitých v dané NS-analýze A byla určena libovolnou větví z A (t.j. v každé větvi byla ta množina stejná) a aby všechny větve z A končily stejnou neredukovatelnou větou (algebraickou terminologií A tvoří svaz).

Další dva principy formulují volnější předpoklady, jak by měla redukční analýza reprezentovat tvar analyzovaného D-stromu, ve kterém jsou i koordinační značky.

Princip Tb-kompatibility. Pokud máme NES-analýzu A D-stromu t a dva různé uzly u, v D-stromu t, které jde redukovat jako určující uzly dvou UNC-operací a přitom nevede cesta mezi u a v, tak požadujeme, aby během A mohla být dříve provedena kterákoliv z těchto UNC-operací (tj. aby existovaly dvě větve z A, kde v první větvi je provedena dříve redukce s u a v té druhé větvi je dříve provedena redukce s v.)

Princip Tc-kompatibility. Nechť máme NES-analýzu A D-stromu t, dvě hrany e_1, e_2 stromu t, které neleží (oběma uzly) na jedné cestě v t a e_1, e_2 jde obě redukovat jako určující hrany UEC-operací. Požadujeme, aby během A mohla být dříve provedena kterákoliv z těchto UNC-operací (tj. existují dvě větve z A , kde v první je provedena dříve redukce s e_1 a v té druhé je redukována dříve e_2. Poznamenejme, že v jedné větvi nemusí být nutně provedeny obě tyto redukce.

Říkáme, že X-analýza $\gg_{(T,X)}$ je k-omezená, pokud počet vypuštěných uzlů v jednotlivých X-redukcích z $\gg_{(T,X)}$ nepřesahuje k a $(T,X)^0_{\gg}$ neobsahuje D-strom s více uzly než k.

Analogicky lze zavést k-omezenou X-analýzu jednotlivého stromu.

Říkáme, že X-analýza $\gg_{(T,X)}(t)$ D-stromu t je proporcionální, pokud $Str(\gg_{(T,X)}(t))$ je proporcionální.

Máme také možnost měřit složitost X-redukcí pomocí počtu operací užitých v jednotlivých X-redukcích.

Příklad 7. D-*strom reprezentující obrázek 4:*

$T_3 = (\{[2,b_2],[3,b_3],[6,b_6]\},$
$\{([2,b_2],[3,b_3]),([6,b_6],[3,b_3])\},([3,b_3],[2,b_2],[6,b_6]))$

D-*strom representující obrázek 5:*

$T_4 = (\{[2,b_2],[3,b_3],[4,b_4],[5,b_5],[6,b_6]\},$
$\{([2,b_2],[3,b_3]),([4,b_4],[3,b_3]),([5,b_5],[4,b_4]),$
$([6,b_6],[3,b_3])\},$
$([3,b_3],[2,b_2],[4,b_4],[5,b_5],[6,b_6]))$

Příklad 8. *Vidíme, že D-strom T_1 má jen značky odpovídající závislostem (nemá značky Cr, Co pro koordinace). Let $R_2 = \{T_1, T_2, T_3, T_4\}$, kde* D*-stromy T_1, T_2, T_3, T_4 byly popsány v předchozích příkladech.*

Vidíme, že

$\gg_{(R_2,NES)} = \{T_1 \gg_{(R_2,NES)} T_2,\ T_2 \gg_{(R_2,NES)} T_3,$
$T_1 \gg_{(R_2,NES)} T_4,\ T_4 \gg_{(R_2,NES)} T_3\},$

a dále že $\gg_{(R_2,NES)}$ je rovno nejen $\gg_{(R_2,NES)}(T_1)$ ale, i $\gg_{(R_2,NS)}(T_1)$.

Platí, že $(R_2, NES)^0_{\gg} = \{T_3\}$.

$\gg_{(R_2,NES)}(T_1)$ je tedy NS-analýzou věty T_1, ale není její N-analýzou, jelikož NS-redukce $T_2 \gg_{(R_2,NS)} T_3$ a $T_1 \gg_{(R_2,NS)} T_4$ používají shift.

Vidíme také, že $\gg_{(R_2,NS)}(T_1)$ je S-kompatibilní, a že její redukce používají jedinou UNC-operaci a maximálně jeden shift.

$\gg_{(R_2,NS)}(T_1)$ je také Ta-kompatibilní, Tb-kompatibilní (a triviálně Tc-kompatibilní a Tl-kompatibilní), 2-omezená, a proporcionální.

Vymezení čistě závislostních D-stromů. Podobné vlastnosti jako má NS-analýza D-stromu T_1 požadujeme po všech čistě závislostních D-stromech (obsahují jen hrany (uzly) se závislostními kategoriemi (značkami)).
Čistě závislostní D-stromy mají NS-analýzu, jejíž redukce obsahují jedinou operaci UNC a nejvýše tři shifty. Každá NS-analýza čistě závislostního D-stromu má být S-kompatibilní, Ta-kompatibilní, Tb-kompatibilní (triviálně i Tc-kompatibilní a Tl-kompatibilní) a proporciální vzhledem k množině všech korektních NS-redukcí korektních čistě závislostních stromů. Toto formální vymezení závislostních stromů odpovídá rozšířenému intuitivnímu vnímání závislostí a je logickým vzorem i pro vymezení D-stromů s koordinacemi.

Pozorování a poznámka. V PDT jsme nezpozorovali žádnou odchylku proti předchozímu vymezení u D-stromů s čistě závislostními značkami. Pokud však budeme uvažovat jen N-analýzu D-stromu T_1, tak ta není ani S-kompatibilní, ani Ta-kompatibilní. Pozorování příkladů tohoto typu nás vedla k rozšíření původně užívané N-analýzy na vhodnější NS-analýzu, kterou lze uplatňovat zřejmě na celou třídu čistě závislostních D-stromů při zachování výše požadovaných principů.

Příklad 9. *V tomto příkladě budeme pozorovat D-strom Tc_1 z obrázku 9, jeho NES-analýzu A_1 na obrázcích 10 až 13 a jeho DS-analýzu z obrázku 6. Tc_1 neobsahuje uzel s dvojicí značek Cr, Co, ani hranu, která má oba uzly se značkou Co.*

Vidíme, že A_1 D-stromu Tc_1 je NS-analýzou (nepoužívá UEC-operace).

A_1 je S-kompatibilní, Tl-kompatibilní a Tb-kompatibilní (triviálně i Tc-kompatibilní) a proporciální.

A_1 je NS-analýzou věty (D-stromu) s trojnásobnou (nezapuštěnou) koordinací.

A_1 není Ta-kompatibilní, protože množiny UNC-operací v jednotlivých větvích nejsou stejné.

A_1 obsahuje redukce, které používají dvě UNC-operace. Tím se liší od závislostních redukcí, které používají jen jednu UNC-operaci.

Všimněme si, že určující uzly dvou UNC-operací v jedné redukci visí na stejném uzlu (se značkou Cr) a odstraněné podstromy tvoří souvislý úsek v uspořádání uzlů.

Povšimněme si ještě, že budeme-li uvažovat N-analýzu A_2 D-stromu Tc_1, tak přijdeme o poslední větev se shiftem. A_2 je také S-kompatibilní, Tl-kompatibilní, Tb-kompatibilní a proporciální. A_2 má tedy také pěkné vlastnosti.

Vymezení závislostně-koordinačních D-stromů bez vložených koordinací. Podobné vlastnosti jako má NS-analýza D-stromu Tc_1 požadujeme po všech D-stromech bez vložených koordinací. Má to být NS-analýza, která je S-kompatibilní, Tl-kompatibilní a Tb-kompatibilní (triviálně i Tc-kompatibilní). Může používat dvě UNC-operace v jedné redukci, které odstraňují dva vedlejší podstromy visící na jednom uzlu.

Pozorování. V PDT jsme zatím nezpozorovali žádnou odchylku proti předchozímu vymezení. Pokud však budeme uvažovat jen NS-analýzu D-stromu Tc_1, která bude pracovat s jedinou UNC-operací v redukci, tak ta není S-kompatibilní.

Poznamenejme, že malou technickou změnou v metodě zobrazování vícenásobných koordinací v PDT bychom dosáhli toho, že by pro zachování S-kompability u redukcí tohoto jevu by nebylo třeba použít více než jednu UNC-operaci.

Příklad 10.

V tomto příkladě budeme pozorovat D-strom Tcz_3 z obrázku 16, jeho NES-analýzu A_3 na obrázcích 16 až 18 a jeho DS-analýzu z obrázku 8.

Tcz_3 obsahuje uzel s dvojicí značek Cr, Co i hranu, která má oba uzly se značkou Co.

A_3 je S-kompatibilní, Tl-kompatibilní a Tb-kompatibilní i Tc-kompatibilní.

A_3 je NES-analýzou věty (D-stromu) s vloženou koordinací, kde UEC-operace jsou uplatněny na hrany u kterých mají oba uzly značku Co, tedy hrany vložené koordinace. Řídící uzel těchto hran mívá ještě značku Cr.

Uvažujeme-li NS-analýzu A_4 D-stromu Tcz_3, tak vidíme, že A_4 není S-kompatibilní, jelikož nemá na rozdíl od odpovídající DS-analýzy z obrázku 8 žádné redukce.

Vymezení závislostně-koordinačních D-stromů. Podobné vlastnosti jako má NES-analýza D-stromu Tcz_3 požadujeme po všech D-stromech s koordinacemi a závislostmi. Má to být NES-analýza, která je S-kompatibilní, Tl-kompatibilní a Tb-kompatibilní (triviálně i Tc-kompatibilní). Může používat UEC-operace s určující hranou jejíž oba uzly nesou značku Co (jiné UEC-operace nejsou povoleny).

Pozorování. V PDT jsme zatím nezpozorovali žádnou odchylku proti předchozímu vymezení.

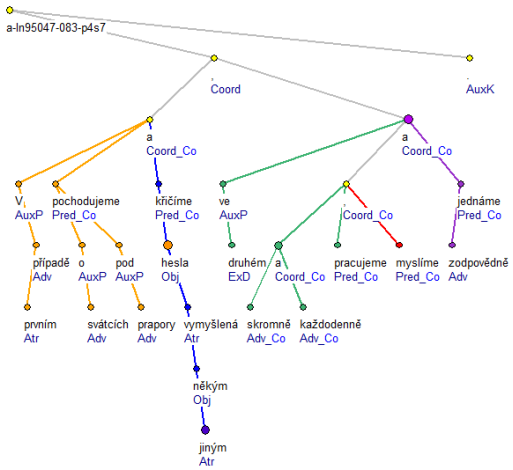
Pozorování obr. 19. Na obrázku 19 je jeden z autentických stromů z PDT. Podle vzoru tohoto stromu vznikly naše obrázky 12 až 18 pro tři různé typy redukcí koordinací.

Připomeňme, že symbol Coord z obrázku 19 je symbol Cr na našich obrázcích, symbol Coord_Co je v našich obrázcích nahrazen symbolem Cr.Co. Symbol Coord_Co je značkou, která má označovat řídící uzel (otce) vložené koordinace. V obrázku 19 je tento symbol jednou užit nesprávně, a to pro frázi 'skromě a Coord_Co každodenně'. Tato fráze zde není vloženou koordinací, ale koordinovanou závislostí podobně jako na obrázku 13.

NES-analýzou získáme z obrázku 19 několik dále neredukovalných vět s koordinacemi, které mají bez identifikačního uzlu a uzlu pro tečku jen tři uzly. NES-analýza bude S-kompatibilní, Tl-kompatibilní a Tb-kompatibilní i Tc-kompatibilní.

Použijeme-li na stejný D-strom jen NS-analýzu, nedostaneme se u redukovaných a dále neredukovatelných D-stromů pod sedm uzlů. Toto poslední pozorování připomíná pozorování z [2], kde se implicitně uvažují redukce, používající maximálně jednu UNC-operaci a žádnou UEC-operaci.

NS-analýza D-stromu z obrázku 19 nemůže být S-kompatibilní.



Obrázek 19: Autentický D-strom z PDT.

2.5 Shrnutí

V tomto příspěvku jsme exaktně zavedli pojmy větné redukční analýzy a tři typy redukční analýzy D-stromů. Formulovali jsme požadavky na kompabilitu větné redukční analýzy a redukční analýzy D-stromů. Našli jsme operace a typy redukcí, které dovolí provádět redukční analýzu D-stromů se závislostmi a koordinacemi stejně jemně a se stejnými k-omezeními jako větnou redukční analýzu. To je hlavní přínos tohoto příspěvku. Při formulaci typů redukčních analýz pro D-stromy jsme vycházeli z pozorování D-stromů z Pražského závislostního korpusu (PDT) a to D-stromů, které kromě modelování závislostí, modelují také složené koordinace. Tři (přesněji čtyři) typy redukčních analýz D-stromů nám dávají přirozenou taxonomii závislostních a koordinačních jevů zachycených D-stromy z PDT.

Domníváme se, že zavedený aparát dovolí hlouběji porozumět neprojektivitě a jejím mírám a volnosti slovosledu. To bude jedno z témat, kterým se budeme zabývat v blízké budoucnosti.

Dále se domníváme, že uvedená metoda by měla pomoci při odhalování nekonzistencí (či chyb) v PDT, podobně jako to bylo v případě D-stromu z obrázku 19.

V blízké budoucnosti bychom také rádi zahrnuli do metody redukční analýzy zbylé syntaktické jevy, které jsou v PDT rozlišeny. Máme na mysli hlavně koordinace s elipsami.

Na závěr děkujeme Markétě Lopatkové za poskytování informací o PDT i za komentáře k poskytnutému materiálu a ochotu o něm diskutovat.

Reference

[1] Hajič, J. Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M.: Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia, 2006.

[2] Lopatková, M., Mírovský, J., Kubon, V.: Gramatické závislosti vs. koordinace z pohledu redukční analýzy. In: Proceedings of the Main Track of the 14th Conference on Information Technologies – Applications and Theory (ITAT 2014), with selected papers from Znalosti 2014 collocated with Znalosti 2014, Demanovska Dolina – Jasna, Slovakia, September 25–29, 2014., pages 61–67, 2014.

[3] Lopatková, M., Plátek, M., Kuboň, V.: Modeling syntax of free word-order languages: dependency analysis by reduction. In: Matoušek, V. et al., editor, Proceedings of TSD 2005, volume 3658 of *LNCS*, pages 140–147. Springer, 2005.

[4] Plátek, M.: Analysis by reduction of d-trees. In: Proceedings of the main track of the 14th Conference on Information Technologies – Applications and Theory (ITAT 2014), with selected papers from Znalosti 2014 collocated with Znalosti 2014, Demanovska Dolina – Jasna, Slovakia, September 25–29, 2014., pages 68–71, 2014.

[5] Plátek, M., Pardubská, D., Lopatková, M.: On minimalism of analysis by reduction by restarting automata. In: Formal Grammar – 19th International Conference, FG 2014, Tübingen, Germany, August 16–17, 2014. Proceedings, pages 155–170, 2014.