ITAT 2016 Proceedings, CEUR Workshop Proceedings Vol. 1649, pp. 26–33
http://ceur-ws.org/Vol-1649, Series ISSN 1613-0073, © 2016 M. Plátek, K. Oliva

Redukční analýza A-stromů s minimalistickými omezeními.*

Martin Plátek[1] a Karel Oliva[2]

[1] MFF UK Praha, Malostranské nám. 25, 118 00 Praha, Česká Republika
martin.platek@ufal.mff.cuni.cz
[2] UJČ ČAV Praha, Letenská, 118 00 Praha, Česká Republika
oliva@ujc.cas.cz

Abstrakt: Tento příspěvek navazuje na náš loňský příspěvek na ITATu. Zpracovává novým způsobem redukční analýzu na A-stromech, které jsou formalizací stromů, zpracovaných metodikou pro analytickou rovinu Pražského závislostního korpusu (PDT). Redukční analýza A-stromů sestává z minimálních korektních redukcí, které používají pouze elementární operace delete a shift.

Hlavním cílem je vyvinout formální prostředky, které by exaktně zachycovaly lingvisticky pozorované minimalistické vlastnosti jednotlivých parametrů stromové redukční analýzy stromů ve formátu PDT a dovolily následně realizovat podobná pozorování na různých přirozených, či umělých jazycích.

Pomocí pozorování lingvistického typu upřesňujeme strukturálně-složitostní vlastnosti A-stromů se závislostmi a koordinacemi. Zvýrazňujeme vlastnosti, kterými se závislosti a koordinace liší.

1 Úvod

V této práci zavádíme a studujeme exaktní pojem (úplné) redukční analýzy A-stromů (URAS). A-stromy modelují stromy analytické roviny Pražského závislostního korpusu (PDT). URAS obsahuje všechny korektní redukce, které lze zařadit do lingvisticky korektní (manuální) redukční analýzy na A-stromech. URAS používá operace delete a shift a jeho redukce jsou minimalizovány s ohledem na počet těchto operací. Postupně zavádíme různé další omezující parametry, které je možno minimalizovat a užívat pro jemnější aproximace lingvisticky intuitivní redukční analýzy. Zavádíme tříčlennou škálu stability pro omezené URAS. Za korektní omezené URAS považujeme ty, co jsou stabilní alespoň v tom nejslabším smyslu. Stabilita pomáhá hledat spodní odhady pro intuitivní redukční analýzu. Typ stability určuje větší či menší vzdálenost od neomezené URAS.

Zavedené pojmy používáme pro klasifikaci pozorování lingvistického typu. Pozorujeme množiny A-stromů, které odpovídají českým větám a jsou zpracovány metodikou analytické roviny PDT. Odkrýváme tak řadu strukturálních vlastností takovýchto A-stromů. Povšimněme si, že prezentovaná pozorování jsou smysluplná a netriviální na konečných i nekonečných jazycích (množinách). To je ve spojitosti s lingvistikou velmi užitečné. Prezentujeme strukturální pozorování a nekombinujeme je (zatím) s pozorováními statistického typu.

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

1.1 Neformální úvod do redukční analýzy.

V této sekci neformálně představujeme redukční analýzu A-stromů se závislostmi a s koordinacemi. Redukční analýzou českých vět a jejímu modelování se zabýváme již delší dobu. Jako základní variantu redukční analýzy předkládáme úplnou redukční analýzu A-stromů (URAS). Navazujeme na články z minulých let (viz [2, 1, 3]). Při zavádění variant redukčních analýz zvýrazňujeme jejich minimalistický charakter.

URAS je založena na postupném zjednodušování A-stromu po minimálních krocích. URAS definuje všechny možné posloupnosti větných redukcí – každá redukce spočívá ve *vypuštění* několika uzlů, nejméně však jednoho uzlu analyzovaného A-stromu. V A-stromě vypouštíme tak, abychom z A-stromu získali opět A-strom a každá cesta v novém A-stromě byla podposloupností cesty v původním A-stromě. Viz např. obrázky z příkladu 1. V některých redukcích může být kromě vypouštění použita operace *shift*, která přesune nějaký uzel na novou pozici v A-stromě.

V našich lingvistických pozorováních budeme rozlišovat vypouštění listů a vypouštění vnitřních uzlů. Kořeny se v URAS nevypouští. Intuitivně i v URAS u většiny závislostních jevů stačí používat vypouštění listů. Ukážeme, že redukce koordinací v PDT s vypouštěním listů nevystačí.

Metoda URAS je popsaná následujícími zásadami:

(i) URAS je složena z jednotlivých redukcí; redukce používají operace dvou typů : (1) vypuštění (delete) a (2) přesun (shift); To znamená, že tvary jednotlivých slov (i interpunkčních znamének), jejich morfologické charakteristiky i jejich syntaktické kategorie se nemění během jednotlivých redukcí.

(ii) Struktura, která je korektním A-stromem, musí být korektním A-stromem i po redukci.

(iii) Redukce nepatří do předem vytipované množiny zakázaných redukcí. Příkladem zakázané redukce je vynechání samotného zvratného ´se´.

(iv) Uvažujeme jen nezmenšitelné redukce, t.j. vynecháme-li z libovolné redukce jednu či více

operací, nastane porušení principu zachování gramatické správnosti (ii) nebo redukce se stane zakázanou a tím poruší princip (iii).

(v) URAS obsahuje všechny možné redukce splňující zásady (i) až (iv).

URAS tvoří základ, z kterého budeme odvozovat další varianty redukční analýzy, tak aby odpovídaly některým typům lingvistické (minimalisticé) intuice. Tento záměr budeme rozvíjet především ve formální části.

V následujících odstavcích nejprve uvedeme jeden příklad ilustrující URAS. Příklad se týká jen redukcí, které zjednodušují závislosti. Později budou následovat příklady, týkající se koordinací. Příklady nejprve poslouží pro úvod do problematiky, později (ve výsledkové části) jako separační příklady pro taxonomii redukčních analýz. A-stromy na obrázcích v našich příkladech jsou oproti stromům z PDT trochu zjednodušené. Za prvé: neobsahují identifikační uzel, který nenese žádnou syntaktickou informaci a neodpovídá žádnému slovu věty. Za druhé: značka 'Coord' je nahrazena značkou 'Cr' a za třetí vynecháváme morfologické značky.

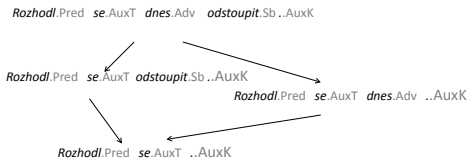
Všimněme si, že korektní A-strom zcela určuje jednu korektní českou větu i s jejím korektním značkováním.

Příklad 1. *Zde ilustrujeme URAS k větě (1). Nevypouštíme zvratnou částici se, neboť vypuštění pouhé zvratné částice považujeme za zakázanou redukci. Zde vůbec nepoužíváme shift.*
(1) Rozhodl.Pred se.AuxT dnes.Adv odstoupit.Sb ..AuxK

Obrázek 1 reprezentuje schema větné redukční analýzy (tzv. UPRA). Isomorfní (velmi podobné) schema mají URAS A-stromů $T1_1$ a $T1_2$ z obrázku 2. Obrázek 1 zde zastupuje i tato schemata.

V jednotlivých redukcích URAS A-stromu $T1_1$ se vypouštějí jen listy, tedy redukcemi nevznikají nové hrany. U $T1_2$, při redukci položky 'odstoupit.Sb', se vypouští vnitřní uzel A-stromu, tedy vzniká nová hrana a to v tomto případě signalizuje změnu významu. To není žádoucí.

Vznikne tak strom $T1_3$, viz obr. 3. Doplňme, že $T1_1$ a $T1_2$ lze redukovat na $T1_4$ a $T1_3$ lze také redukovat na $T1_5$. Obrázky těchto redukcí jsme vynechali. K tomuto příkladu patří ještě obrázek 4, zobrazující redukci $T1_4$ na $T1_5$.

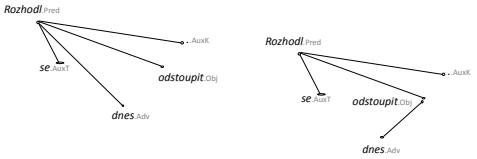


Obrázek 1: UPRA věty (1).

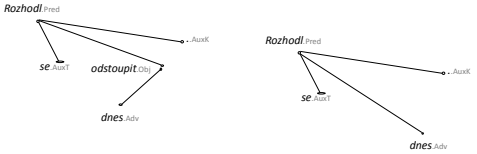
2 Formalizace redukční analýzy.

Zde zavedeme obecné formální pojmy, které mohou sloužit k formulaci redukčních vlastností jak závislostních stromů přirozených jazyků, tak i podobných struktur u programovacích a dotazovacích jazyků. V podsekcích, prezentujících pozorování lingvistického typu, se budeme věnovat formulaci redukčních vlastností stromů analytické roviny PDT. Značka $\subset$ znamená v celém příspěvku vlastní podmnožinu.

Formalizace redukční analýzy analytických stromů se neobejde bez formalizace lexikální analýzy.



Obrázek 2: A-stromy $T1_1$ a $T1_2$ nad větou (1).



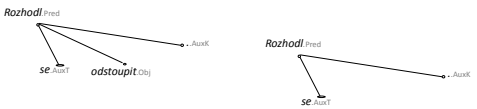
Obrázek 3: Redukce $T1_2$ na $T1_3$.

2.1 Formalizace lexikální analýzy

Při formalizaci lexikální analýzy rozlišujeme tři konečné množiny slov a značek. Σ_p označuje tzv. vlastní slovník [1], který obsahuje jednotlivé slovní formy a interpunkční znaménka daného jazyka. Σ_c označuje tzv. kategoriální seznam, tedy množinu syntakticko-morfologických značek. Hlavní slovník $\Gamma \subseteq \Sigma_p \times \Sigma_c$ reprezentuje zjednoznačněnou lexikální analýzu daného jazyka.

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 2. *V našich pozorováních analytické roviny PDT pracujeme s hlavním slovníkem označeným jako Γ_{PDT}, Σ_{pPDT} označuje vlastní slovník a Σ_{cPDT} označuje kategoriální seznam značek, užívaných v PDT.*



Obrázek 4: Redukce $T1_4$ na $T1_5$.

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

Výše definované pojmy ilustrujeme na příklade, který vychází z příkladu 1.
{ Rozhodl, se, dnes , odstoupit, . } $\subset \Sigma_{pPDT}$,
{ Pred, AuxT, Adv, Sb, AuxK} $\subset \Sigma_{cPDT}$,
{ [Rozhodl,Pred], [se,AuxT],
[dnes,Adv], [odstoupit,Sb], [.,AuxK]} $\subset \Gamma_{PDT}$.

Jednotlivým položkám hlavního slovníku z tohoto příkladu přiřazujeme jména (b_1 atd.), která budeme v dalších příkladech užívat jako zkratky.
b_1*= [Rozhodl,Pred],* b_2*=[se,AuxT],* b_3*= [dnes,Adv],* b_4*=[odstoupit,Sb],* b_5*=[.,AuxK].*

V abecedě kategorií v tomto příkladě jsou využity jen jednoduché závislostní kategorie (ne všechny). Kategorie mohou být složené z více značek. Kategorie pro koordinace budou obsahovat značky 'Cr', nebo 'Co'.

Věty v našich příkladech končí sentinelem (ukončením věty), který se během redukční analýzy ani nevypouští, ani nepřesunuje. Je to [.,AuxK].

2.2 R-seznamy a A-stromy.

V následující části budeme reprezentovat věty pomocí tzv. R-seznamů a jejich syntaktické struktury pomocí A-stromů. R-seznamy a A-stromy jsou datové typy, vhodné pro používání operací delete a shift. Na R-seznamech a A-stromech zavádíme uniformním způsobem redukce, založené právě na operacích delete a shift. Redukční seznamy (R-seznamy) zjemňují pojem řetězu a A-stromy nesou více informace než R-seznamy. A-strom a R-seznam se skládají z uzlů, které v PDT reprezenují výskyty lexikálních jednotek (slov, interpunkčních znamének a jejich značek) v příslušné větě.

V A-stromu jsou pomocí stromové struktury reprezentovány syntaktické vztahy, pomocí R-seznamu, jež je součástí každého A-stromu, je reprezentováno pořadí slov.

R-seznam. Nechť I je konečná množina přirozených čísel, Γ konečná abeceda a $V \subseteq (I \times \Gamma)$, kde V reprezentuje totální zobrazení množiny I do Γ. Nechť *ord* je úplné uspořádání množiny V. Říkáme, že *ord* je redukčním seznamem (R-seznamem) na Γ. Zapisujeme ho jako seznam prvků z V. Prvky R-seznamu označujeme jako uzly. Množinu R-seznamů, která vznikla všemi možnými uspořádáními množiny V, označujeme jako $ord(V)$.
Nechť $u \in V$, pak $u = [i,a]$, kde $i \in I$, $a \in \Gamma$. Říkáme, že i je *indexem uzlu* u. Slouží k jednoznačné identifikaci uzlu. Říkáme, že a je *symbolem uzlu* u.

A-strom. A-strom nad Γ je trojice $s = (V,E,ord)$, kde (V,E) je orientovaný strom, jehož (maximální) cesty začínají v listech a končí v kořeni, V je konečná množina jeho uzlů, $E \subset V \times V$ konečná množina jeho hran a $ord \in ord(V)$. Říkáme, že *ord* je R-seznamem A-stromu s. Píšeme $R(s) = ord$.

Projekce. Je-li $ord = ([i_1,a_1],\cdots,[i_n,a_n])$, tak $w = a_1 \cdots a_n$ je řetěz (resp. věta), který označujeme $Str(s) = w$ nebo $Str(ord) = w$, a říkáme, že w je řetězem (projekcí) A-stromu s nebo řetězem (projekcí) R-seznamu *ord*.

Normalizace. Říkáme, že A-strom $s = (V,E,ord)$ (R-seznam *ord*) je *normalizovaný*, pokud *ord* má tvar $ord = ([1,a_1],[2,a_2],\cdots,[n,a_n])$. *Normalizace* A-stromu $s = (V,E,ord)$ je takový normalizovaný A-strom $s_1 = (V_1,E_1,ord_1)$, pro který (V,E) a (V_1,E_1) jsou izomorfní a $Str(s) = Str(s_1)$. Všimněme si, že normalizace A-stromu je jednoznačně daná.

Ekvivalence. Dva A-stromy (R-seznamy) jsou ekvivalentní, pokud mají stejnou normalizaci. Ekvivalentní A-stromy často nebudeme rozlišovat.

Operace shift a delete zavedeme tak, že převedou A-strom na A-strom.

Delete. Operace $dl(i)$ vyřadí z množiny V a z R-seznamu *ord* uzel tvaru $[i,a_i]$ a získá tím množinu V_1 a R-seznam ord_1. Z A-stromu $s = (V,E,ord)$ operace $dl(i)$ udělá A-strom $s_1 = (V_1,E_1,ord_1)$ tím, že vyřadí uzel tvaru $[i,a_i]$ jak z množiny V, tak z R-seznamu *ord*. Dále vyřadí z E všechny dvojice hran tvaru $([j,a_j],[i,a_i])$ a $([i,a_i],[k,a_k])$ (pokud existují). Každou takovou dvojici hran nahradí v E_1 jedinou hranou tvaru $([j,a_j],[k,a_k])$. Viz příklad 3.

Shift. Operace $sh(i,j)$ přesune v R-seznamu *ord* uzel s indexem i před uzel s indexem j. Vytvoří tak nový R-seznam ord_2. Provedeme-li operaci $sh(i,j)$ na A-strom $s = (V,E,ord)$, získáme tím A-strom $s_2 = (V,E,ord_2)$. Operace shift mění v A-stromě pouze R-seznam, tedy slovosled. Viz příklad 4.

Poznámka. Připomeňme si, že operace mají být voleny tak, že posledním uzlem trvale zůstává sentinel.

2.3 URAS (Úplná redukční analýza A-stromu).

Zavádíme URAS s možností regulace pomocí množiny (významově) zakázaných redukcí. Příkladem zakázané redukce A-stromů z PDT, je vynechání předložky z předložkové vazby, či vynechání samotné zvratné částice.

Značení. Nechť Γ je konečná abeceda. $T(\Gamma)$ značí množinu všech A-stromů na Γ. Nechť $\mathsf{T} \subseteq T(\Gamma)$. Říkáme, že T tvoří T-jazyk na Γ. Množinu R-seznamů $R(\mathsf{T}) = \{R(t) \mid t \in \mathsf{T}\}$ nazýváme R-jazykem T-jazyka T. Analogicky, jazyk $Str(\mathsf{T}) = \{Str(t) \mid t \in \mathsf{T}\}$ nazýváme Str-jazykem T. Nechť $Z \subset \{(s,t) | s,t \in \mathsf{T}\}$ je daná množina zakázaných redukcí na T. Označíme $Str(Z) = \{(Str(s),Str(t)) | (s,t) \in Z\}$ a $R(Z) = \{(R(s),R(t)) | (s,t) \in Z\}$.

Redukce. Nyní zavedeme k T-jazyku T a dané množině zakázaných redukcí Z redukce typu $\vdash^Z_T$. Nechť s,t jsou A-stromy. Říkáme, že s je přímo redukovatelné na t podle T a Z a píšeme $s \vdash^Z_T t$ pokud:

- $s,t \in \mathsf{T}$ a $|Str(s)| > |Str(t)|$ a (s,t) není ze Z;
- t je získáno z s provedením množiny operací vypuštění (deletů) Dl a následně postupným provedením shiftů z uspořádané množiny Sh. Dl je povinně neprázdná, Sh může být prázdná.

- Libovolný uzel je přesouván pomocí *Sh* maximálně jednou.
- **Operační nezmenšitelnost redukce**. Pokud bychom vynechali při aplikaci na s jednu nebo více operací z *Dl* nebo z *Sh*, získali bychom A-strom z takový, že $z \notin \mathsf{T}$, nebo $(s,z) \in Z$.
- Jako $DL(s,t)$ označujeme množinu uzlů A-stromu s, vypuštěnou během redukce $s \vdash_T^Z t$ a říkáme, že je DL-množinou redukce $s \vdash_T^Z t$. O *Sh* říkáme, že je SH-sekvencí redukce $s \vdash_T^Z t$.

Doplňující pojmy. Reflexívní a tranzitívní uzávěr relace $\vdash_T^Z$ označujeme $\vdash_T^{Z\,*}$. Částečné uspořádání $\vdash_T^Z$ přirozeně definuje

- $T^0_{\vdash_T^Z} = \{v \in \mathsf{T} \mid \neg\exists u \in \mathsf{T} : v \vdash_T^Z u\}$ - množina neredukovatelných A-stromů T-jazyka T.
- $T^{n+1}_{\vdash_T^Z} = \{v \in \mathsf{T} \mid \exists u \in T^n_{\vdash_T^Z} : u \vdash_T^Z v\} \cup T^n_{\vdash_T^Z}$, $n \in N$ - množina A-stromů z T, které je možné zredukovat na neredukovatelný A-strom z Tposloupností URAS-redukcí délky nanejvýš $n+1$.

URAS. Pro A-strom $s \in \mathsf{T}$ a zakázanou množinu Z nazveme $\mathrm{URAS}(s,T,Z) = \{u \vdash_T^Z v \mid s \vdash_T^{Z\,*} u\}$ (úplnou) redukční analýzou s podle T a Z.

Větev. Nechť $B = (s_1, s_2, \cdots, s_n)$ je posloupnost A-stromů taková, že $s_1 \vdash_T^Z s_2$, $s_2 \vdash_T^Z s_3$, $\cdots$, $s_{n-1} \vdash_T^Z s_n$ a $s_n \in T^0_{\vdash_T^Z}$.
Říkáme, že B je větví $\mathrm{URAS}(s,T,Z)$ a n je její délka.

DL-sekvence a DL-charakteristika. Nechť Dl_i je Dl-množinou redukce $s_i \vdash_T s_{i+1}$ pro $1 \le i < n$ a Dl_n je množinou uzlů A-stromu s_n.
Píšeme $Dl(B) = (Dl_1, DL_2, \cdots, Dl_{n-1})$ a říkáme, že $Dl(B)$ je DL-sekvencí větve B.
Množina $Ch(B) = (\{Dl_1, DL_2, \cdots, Dl_{n-1}\})$ je DL-charakteristikou větve B.
DL-charakteristika a DL-sekvence se liší tím, že u DL-charakteristiky nezáleží na pořadí redukčních množin, ale u DL-sekvence ano.
Vidíme, že pro $1 \le i < j < n$ jsou Dl_i a Dl_j disjunktní.

2.4 Algebraické vlastnosti závislostí a koordinací u analytických stromů PDT.

Touto podsekcí začíná výsledková část příspěvku. Předkládáme výsledky dvou typů. Nejčastěji prezentujeme lingvistická pozorování, formulovaná pomocí zavedeného aparátu. Získali jsme je (neúplným) procházením materiálu z PDT. K pozorováním jsme nenašli žádné výjimky a nevěříme, že se nějaké najdou. Pozorování by měla být podnětem ke (korpusově lingvistické) diskusi.

Druhým typem výsledků jsou tvrzení a důsledky matematického charakteru. Vycházejí z rozboru prezentovaných (lingvistických) příkladů a z vlastností zavedeného aparátu.

T_P v následujícím textu označuje množinu korektních A-stromů s koordinacemi a závislostmi, zpracovaných metodikou analytické roviny PDT. Rozhodnout o tom, zda daný A-strom patří do T_P, by měli umět lidé (lingvisté, anotátoři), ovládající češtinu a metodiku PDT.

ZP označuje množinu zakázaných redukcí pro analytickou rovinu PDT.

Příklad 3. *Tento příklad navazuje na příklady 1 a 2. Obsahuje formalizaci A-stromů $T1_2$ a $T1_3$ a tím i popis redukce $T1_2 \vdash_{T_P}^{ZP} T1_3$:*
$T1_2 = (V_2, E_2, ord_2)$*, pričemž*
$V_2 = \{[1,b_1],[2,b_2],[3,b_3],[4,b_4],[5,b_5]\}$
$E_2 = \{([2,b_2],[1,b_1]),\ ([3,b_3],[4,b_4]),\ ([4,b_4],[1,b_1]),\ ([5,b_5],[1,b_1])\}$,
$ord_2 = ([1,b_1],[2,b_2],[3,b_3],[4,b_4],[5,b_5])$
$T1_3 = (V_3, E_3, ord_3)$*, pričemž*
$V_3 = \{[1,b_1],[2,b_2],[3,b_3],[5,b_5]\}$
$E_3 = \{([2,b_2],[1,b_1]),\ ([3,b_3],[1,b_1]),\ ([5,b_5],[1,b_1])\}$
$ord_3 = ([1,b_1],[2,b_2],[3,b_3],[5,b_5])$
Vidíme, že $T1_2$ je normalizovaný a že $T1_3$ normalizovaný není, protože vznikl z $T1_2$ vypuštěním uzlu $[4,b_4]$.

Následují strukturální pozorování A-stromů z T_P. Pozorování odrážejí syntaktické vlastnosti českých vět a anotátorskou metodiku pro analytickou rovinu PDT. Naše příklady tato pozorování ilustrují.

Pozorování 1. *Nechť s je A-strom z T_P. Všechny větve $URAS(s,T_P,ZP)$ mají stejnou délku.*

Pozorování 2. *Nechť s je A-strom z T_P, který neobsahuje koordinace (tj. značky ´Cr´ a ´Co´). Všechny větve $URAS(s,T_P,ZP)$ mají nejen stejnou délku, ale i stejnou DL-charakteristiku. Navíc $URAS(s,T_P,ZP)$ obsahuje jediný neredukovatelný A-strom. Tedy $URAS(s,T_P,ZP)$ lze považovat za (algebraickou strukturu zvanou) svaz.*

Pozorování 3. *Nechť s je A-strom z T_P, který neobsahuje koordinace a r_1, r_2 jsou dvě různé redukce z $URAS(s,T_P,ZP)$. Platí, že r_1 a r_2 mají disjunktní DL-množiny.*

Pozorování 4. *Nechť s je A-strom z T_P, který obsahuje koordinaci alespoň tří členů. Existují dvě větve $URAS(s,T_P,ZP)$ s různou DL-charakteristikou.*

Pozorování 5. *Nechť s je A-strom z T_P, který obsahuje koordinaci alespoň tří členů. Existují dvě redukce z $URAS(s,T_P,ZP)$, které nemají disjunktní DL-množiny. Průnik těchto DL-množin obsahuje uzel se spojkou nebo čárkou se značkou ¨AuxX¨.*

Předchozí dvě pozorování jsou ilustrovány příkladem 4.

Tvrzení 1. *Existuje $t \in T_P$, jehož URAS obsahuje více než jeden neredukovalelný A-strom.*

Předchozí tvrzení lze dokázat pomocí A-stromu k větě 'Přišel, viděl, zvítězil.'.

2.5 UPRA (úplná větná redukční analýza.)

Abychom mohli dát do souvislosti URAS se starším pojmem, větnou redukční analýzou, zavádíme úplnou větnou redukční analýzu (UPRA), viz [3]. Do UPRA vstupuje věta ve formě R-seznamu. UPRA zavádíme zcela analogicky jako URAS.

Redukce. Mějme jazyk L a R-seznam u takový, že $Str(u) \in L$. Říkáme, že u je R-seznamem k jazyku L a píšeme $u \in R(L)$. Nechť $U \subset \{(u,v)|u,v \in R(L)\}$ je daná množina zakázaných redukcí.

Zavedeme k $R(L)$ a dané U redukce $\succ_L^U$. Nechť $u,v \in R(L)$. Říkáme, že u je redukovatelné na v podle L a U a označujeme $u \succ_L^U v$, pokud:

- $|Str(u)| > |Str(v)|$ a (u,v) není z U;
- R-seznam v je získán z u provedením množiny operací vypuštění (deletů) Dl a následně postupným provedením shiftů z uspořádané množiny Sh. Dl je povinně neprázdná, Sh může být prázdná.
- Libovolný uzel je přesouván pomocí Sh maximálně jednou.
- **Operační nezmenšitelnost redukce**. Pokud bychom vynechali při aplikaci na u jednu nebo více operací z Dl nebo z Sh, získali bychom R-seznam z takový, že $Str(z) \notin L$, nebo $(u,z) \in U$.
- Jako $Dl(u,v)$ označujeme množinu uzlů R-seznamu u, vypuštěnou provedením množiny deletů Dl a říkáme, že $Dl(u,v)$ je DL-množinou redukce $u \succ_L^U v$. O Sh říkáme, že je SH-sekvencí redukce $u \succ_L^U v$.

UPRA. Nechť $w \in R(L)$ a $U \subset \{(u,v)|u,v \in R(L)\}$ je daná množina zakázaných redukcí. UPRA(w,L,U) $=\{u \succ_L^U v \mid w \succ_L^U {}^* u\}$ nazveme úplnou redukční analýzou w k jazyku L a množině nekorektních redukcí U.

Zbývající potřebné pojmy pro UPRA lze zavést zcela analogicky jako pro URAS.

2.6 Nesouvislosti a stabilita redukcí.

Zavádíme dvě míry nesouvislosti redukcí, které se vzájemně doplňují. S ohledem na tyto a další míry zavádíme několik typů stability pro URAS, které nám dovolí klasifikovat omezená URAS jako stabilní, nebo nestabilní. Stabilita URAS pro jednotlivé A-stromy je formálním kriteriem pro lingvistickou adekvátnost redukční analýzy, s ohledem na daná omezení. Budeme hledat maximální omezení taková, která zachovávají alespoň nejslabší typ stability. Následuje několik formálních definic.

Graf redukce. Mějme redukci $s \vdash_T^Z t$, kde $s = (V,E,or)$, a její DL-množinu $DL(s,t)$. Píšeme $G(s,t) = (DL(s,t), \{(a,b) \in E | a,b \in DL(s,t)\})$ a říkáme, že $G(s,t)$ je DL-grafem redukce $s \vdash_T^Z t$.

Počet komponent redukce. Nechť i je počet komponent DL-grafu $G(s,t)$. Budeme psát, že $pk(s,t) = i$ a říkat, že i je počet komponent redukce $s \vdash_T^Z t$.

URAS s omezeným počtem komponent. Nechť i je přirozené číslo. Označíme jako URAS$(s,T,Z;pk \leq i)$ podmnožinu URAS(s,T,Z), která obsahuje všechny redukce z URAS(s,T,Z), které nemají více komponent než i.

Vidíme, že neredukovatelné stromy v URAS$(s,T,Z;pk \leq i)$ mohou být pro některá i jiné (větší), než ty z URAS(s,T,Z).

Říkáme, že URAS$(s,T,Z;pk \leq i)$ je pro dané i T-stabilní, pokud URAS$(s,T,Z;pk \leq i)$ = URAS(s,T,Z).

Říkáme, že URAS$(s,T,Z;pk \leq i)$ je pro dané i CH-stabilní, pokud množina charakteristik URAS$(s,T,Z;pk \leq i)$ a URAS(s,T,Z) je stejná.

URAS$(s,T,Z;pk \leq i)$ je pro dané i Mn-stabilní, pokud každý neredukovatelný strom z URAS$(s,T,Z,pk \leq i)$ je i neredukovatelným stromem URAS(s,T,Z).

Požadavky na stabilitu jsou seřazeny od nejsilnější k nejslabší. Nahlédneme, že stejně můžeme užívat zavedené typy stability pro další typy redukčních omezení.

Počet komponent je jednou přirozenou mírou nesouvislosti redukce A-stromu. Budeme používat ještě jednu míru nesouvislosti redukce, která měří velikost mezer mezi komponentami. Následují další formální definice.

Velikost mezer v redukci. Jako $Sv(s,t)$ budeme označovat nejmenší souvislý (bez ohledu na orientaci) podgraf A-stromu s, který obsahuje DL-graf $G(s,t)$. Nechť j je počet uzlů, které obsahuje $Sv(s,t)$ navíc oproti $G(s,t)$. Píšeme $ns(s,t) = j$ a říkáme, že redukce $s \vdash_T^Z t$ má velikost mezer j.

URAS s omezením na velikost mezer. Nechť i je přirozené číslo. Označíme jako URAS$(s,T,Z : ns \leq i)$ podmnožinu URAS(s,T,Z), která obsahuje všechny redukce z URAS(s,T,Z), které nemají velikost mezer větší než i.

Omezení můžeme i skládat. Např. URAS$(s,T,Z;pk \leq i,ns \leq j)$ = URAS$(s,T,Z;pk \leq i) \cap$ URAS$(s,T,Z;ns \leq j)$.

Množiny stromů stabililní s ohledem na omezení. Budeme používat následující typy značení pro množiny A-stromů splňující daná omezení.

Např. TRAS$(T,Z;pk \leq 1,ns \leq 0;$ T-st $) = \{t \in T |$ URAS$(t,T,Z;pk \leq 1,ns \leq 0)$ je T-stabilní $\}$.

Analogicky TRAS$(T,Z;pk \leq 1;$ CH-st $) = \{t \in T |$ URAS$(t,T,Z;pk \leq 1)$ je CH-stabilní $\}$. Podobně budeme popisovat množiny A-stromů z T parametrizované dalšími omezeními a různými typy stability ze škály T-stabilní, CH-stabilní, Mn-stabilní.

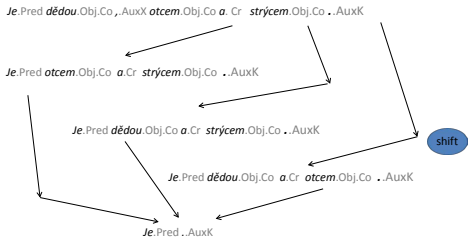
2.7 Rozlišení závislostí a koordinací pomocí (ne)souvislosti.

Předchozí pojmy a následující příklady využijeme k formulaci nových pozorování o PDT.

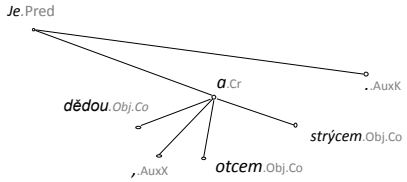
Příklad 4. *Tento příklad ilustruje redukce vícenásobných koordinací a použití grafově nesouvislé redukce v URAS.*

(3) Je.Pred dědou.Obj.Co ,.AuxX otcem.Obj.Co a..Cr strýcem.Obj.Co..AuxK

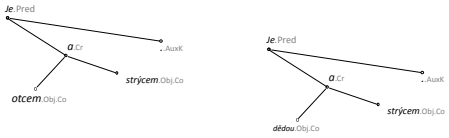
Na obrázku 5 vidíme schema UPRA věty (3) podle stromu $T3_1$, jazyka T_p a prázdné zakázané množiny. Schema stejného tvaru má i schema URAS A-stromu $T3_1$. Věta (3) obsahuje trojnásobnou koordinaci předmětů. Povšimněme si, že dalšímu zjemnění schematu zabraňují kategorie (značky), použité podle vzoru PDT. Značka 'Cr' znamená koordinující symbol (slovo), 'Co' značí koordinované slovo, či symbol. Schematu na obrázku odpovídají redukce A-stromů, které jsou reprezentovány obrázky 4 až 8. Všechny tři redukce A-stromu $T3_1$ vypouštějí (při zjednodušování trojnásobné koordinace na dvojnásobnou) dva nesouvisející listy (podstromy). Třetí redukce navíc používá shift. Zbývající redukce dvojnásobných koordinací se realizují postupným vypouštěním listů, které tvoří souvislý úplný podstrom.



Obrázek 5: UPRA věty (3) podle $T3_1$.



Obrázek 6: A-strom $T3_1$.



Obrázek 7: $T3_2$ a $T3_3$ vzniklé redukcemi z $T3_1$.

Snadno ověříme z definic následující tvrzení.

Tvrzení 2. *Vidíme, že*
$URAS(T1_1, T_P, ZP; pk \leq 1)$ *je T-stabilní,*
$URAS(T3_1, T_P, ZP; pk \leq 2)$ *je T-stabilní a*
$URAS(T3_1, T_P, ZP; pk \leq 1)$ *není Mn-stabilní.*

Obrázek 8: Vlevo $T3_4$, vzniklý redukcí z $T3_1$ a vpravo $T3_5$ vzniklý redukcemi z $T3_2$, $T3_3$ a $T3_4$.

Z předchozích tvrzení vyplývá následující důsledek.

Důsledek 1. *Vidíme, že*
$TRAS(T_P, ZP, pk \leq 1; T\text{-}st) \subset TRAS(T_P, ZP; pk \leq 2; T\text{-}st)$

Následují výsledky našeho pozorování T_P, které se týkají nesouvislostí.

Pozorování 6. *Nechť $s \in T_P$. $URAS(s, T_P, ZP; pk \leq 2)$ je T-stabilní.*

Pozorování 7. *Nechť $s \in T_P$ je A-strom bez koordinací. $URAS(s, T_P, ZP; pk \leq 1)$ je T-stabilní.*

Pozorování 8. *Nechť $s \in T_P$ je A-strom s alespoň trojnásbnou koordinací. $URAS(s, T_P, ZP; pk \leq 1)$ není Mn-stabilní.*

Poznámky k předchozímu pozorování. Podobně jako u $T3_1$, každá alespoň trojnásobná koordinace z PDT vyžaduje alespoň jednu redukci se dvěma komponentami. Pokud povolíme redukce s maximálně jednou komponentou, bude každý neredukovatelný strom z $URAS(s, T_P, ZP; pk \leq 1)$ minimálně o jednu nevykonanou redukci větší, než příslušný neredukovatelný strom z $URAS(s, T_P, ZP)$.

Pozorování o velikosti mezer jsou analogická pozorováním o počtu komponent. Důležité pozorování je, že koordinace dovolují redukcím jen velikost mezer rovnou jedné a stromy bez koordinací dovolují redukcím jen jedinou komponentu.

Pozorování 9. *Vypozorovali jsme, že*
$TRAS(T_P, ZP, ns \leq 0; T\text{-}st) = TRAS(T_P, ZP; pk \leq 1; T\text{-}st)$,
$TRAS(T_P, ZP; ns \leq 1; T\text{-}st) = TRAS(T_P, ZP, pk \leq 2; T\text{-}st)$
$= T_P$.

Pozorování 10. *Nechť $s \in T_P$ je A-strom bez koordinací. $URAS(s, T_P, ZP; ns \leq 0)$ je T-stabilní. Vidíme, že i $URAS(s, T_P, ZP; pk \leq 1, ns \leq 0)$ je T-stabilní.*

Pozorování 11. *Nechť $s \in T_P$ je A-strom s alespoň trojnásobnou koordinací. Platí, že $URAS(s, T_P, ZP; ns \leq 0)$ není Mn-stabilní.*

2.8 URAS s omezeními míry (ne)listovosti.

Snažíme se minimalizovat při redukcích změny hran (změny významu), takže se snažíme redukovat stromy bez

koordinací tak, že vypouštíme v jistém pořadí jen listy. Pojmy zaváděné v tomto odstavci zavádíme za dvojím účelem. Prvním účelem je dát prostředky pro formální aproximaci intuitivní redukční analýzy stromů bez koordinací. Druhým účelem je exaktně zachytit fakt, že redukce vložených koordinací nutně používají vypuštění vnitřního uzlu a charakterizovat složitost tohoto faktu. Při redukci vložených koordinací se význam redukovaného stromu nijak nemění.

Nechť o je nějaké uspořádání množiny Dl, kde Dl je DL-množinou nějaké redukce A-stromu s. Pak říkáme, že o realizuje Dl na s. Píšeme $o \in ord(Dl,s)$.

IN-stupněm operace $dl(i)$ na A-stromě s nazveme počet hran z E vcházejících do uzlu $[i,a_i]$. Všimněme si, že delete uzlu $[i,a_i]$ má IN-stupeň 0 právě tehdy, pokud $[i,a_i]$ je listem A-stromu s.

Uvažujme různé realizace množiny Dl, kde Dl je DL-množina na s. V různých realizacích Dl na s může mít $dl(i) \in Dl$ různou hodnotu svého IN-stupně, neboť $dl(i)$ může být prováděna na různých A-stromech.

Omezíme se jen na neklesající realizace DL-množin v redukcích, neboť realizace vypouštějící jen listy musí být neklesající. Budeme využívat faktu, že ke každé redukci existuje neklesající realizace.

Značení. Říkáme, že $o \in ord(Dl,s)$ je neklesající a píšeme $o \in Nord(Dl,s)$, pokud $o = (dl(i_1),dl(i_2),\cdots,dl(i_n))$ a $IN(dl(i_1)) \leq IN(dl(i_2)),\cdots,IN(dl(i_{n-1})) \leq IN(dl(i_n))$. Píšeme $IN(o) = (IN(dl(i_1)),IN(dl(i_2)),\cdots,IN(dl(i_n)))$.

Nechť $o \in Nord(Dl,s)$, první prvek z o je $dl(i)$, poslední prvek z o je $dl(j)$. Budeme psát $MinIN(o) = IN(dl(i))$ a $MaxIN(o) = IN(dl(j))$.

URAS se spodní mírou (ne)listovosti. Označíme jako $\mathrm{URAS}(s,T,Z;MinIn \leq i)$ podmnožinu $\mathrm{URAS}(s,T,Z)$, která obsahuje všechny redukce z $\mathrm{URAS}(s,T,Z)$, které mají neklesající realizaci o s $MinIN(o) \leq i$.

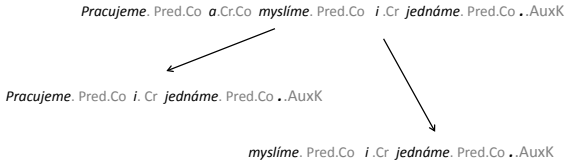
URAS s horní mírou (ne)listovosti. Označíme jako $\mathrm{URAS}(s,T,Z;MaxIN \leq i)$ podmnožinu $\mathrm{URAS}(s,T,Z)$, která obsahuje všechny redukce z $\mathrm{URAS}(s,T,Z)$, které mají neklesající realizaci o s $MaxIN(o) \leq i$.

2.9 Závislosti, vložená koordinace a (ne)listovost.

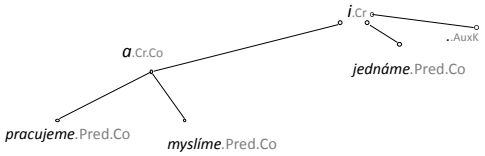
Příklad 5. *Tento příklad ilustruje redukce vložených koordinací.*

(5) Pracujeme.Pred.Co a.Cr.Co myslíme.Pred.Co i..Cr jednáme.Pred.Co..AuxK

Na obrázku 9 vidíme schema UPRA věty (5) podle $T5_1$, T_P a ZP. Věta (5) je věta s vloženou koordinací. A-stromy odpovídající redukcím jsou na obrázcích 10 až 12. Vložená koordinace se v A-stromě $T5_1$ zjednodušuje tak, že se vyjme jedna hrana s řídícím uzlem se značkou 'Cr.Co'. To odpovídá dvěma redukcím v UPRA z obrázku . Vidíme, že tyto redukce vypouštějí jeden list a jeden vnitřní uzel do kterého vchází jediná hrana.



Obrázek 9: UPRA věty s vloženou koordinací.



Obrázek 10: $T5_1$

Tvrzení 3. *Pro pro čistě závislostní strom $T1_1$ z příkladu 1 platí, že $URAS(T1_1,T_P,ZP;MaxIN \leq 0)$ je T-stabilní.*

Tvrzení 4. *Pro čistě závislostní strom $T1_2$ z příkladu 1 platí, že $URAS(T1_2,T_P,MaxIN \leq 0)$ není T-stabilní, ale je Mn-stabilní. Navíc $URAS(T1_2,T_P,MinIn \leq 1,MaxIN \leq 1)$ je T-stabilní.*

Důsledek 2. *Vidíme, že*

- $TRAS(T_P,ZP;MaxIN \leq 0;$ *T-st* $)$
  $\subset TRAS(T_P,ZP;MaxIN \leq 1;$ *T-st).*
- $TRAS(T_P,ZP;MinIN \leq 0;$ *T-st* $)$
  $\subset TRAS(T_P,ZP;MinIN \leq 1;$ *T-st).*

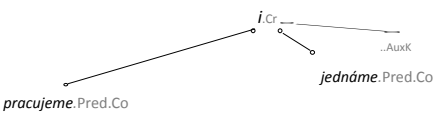
Pozorování 12. $TRAS(T_P,ZP;MinIN \leq 1;$ *T-st* $) \subset T_P$.

Tvrzení 5. *Pro $T5_1$ z příkladu 5 platí, že*
$URAS(T5_1,T_P,ZP;MinIN \leq 0)$ je T-stabilní,
$URAS(T5_1,T_P,ZP;MaxIN \leq 0)$ není Mn-stabilní a
$URAS(T5_1,T_P,ZP;MinIN \leq 0,MaxIN \leq 1)$
je T-stabilní.
$T5_1$ nese koordinaci vloženou do koordinace. Vidíme, že platí $T5_1 \in TRAS(T_P,ZP;MinIN \leq 0,MaxIn \leq 1;$ T-st $)$

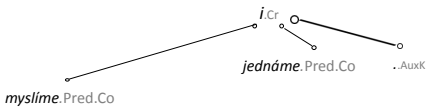
Pozorování 13. *Nechť $t \in T_P$ nese koordinaci vloženou do koordinace. Platí, že $URAS(t,T_P,ZP;MaxIN \leq 0)$ není Mn-stabilní.*

Důsledek 3. *Vidíme, že*

- $TRAS(T_P,ZP;MaxIN \leq 0;$ *T-st* $)$
  $\subset TRAS(T_P,ZP;MinIN \leq 0;MaxIn \leq 1;$ *Mn-st)*
- $TRAS(T_P,ZP;MinIN \leq 0;MaxIn \leq 1;$ *Mn-st*$)$
  $\subset TRAS(T_P,ZP;MinIN \leq 1;MaxIn \leq 1;$ *Mn-st* $)$



Obrázek 11: $T5_2$, vzniklé redukcí z $T5_1$.



Obrázek 12: $T5_3$, vzniklé redukcí z $T5_1$.

- $TRAS(T_P, ZP; MinIN \leq 1, MaxIn \leq 1;$ *Mn-st* $) \subseteq T_P$.

Poznámka. Pro každé $t \in T_P$, které jsme pozorovali, bylo URAS$(t, T_P, ZP; MinIN \leq 1)$ Mn-stabilní. Neumíme odhadnout, zda existuje A-strom $t \in T_P$ takový, že URAS$(t, T_P, ZP; MinIN \leq 1)$ není Mn-stabilní, tedy zda TRAS$(T_P, ZP, MinIN \leq 1, MaxInPc \leq 1;$ Mn-st $) = T_P$.

2.10 Konzistence URAS a UPRA nad PDT

L_P značí množinu korektních českých vět (jen) s koordinacemi a závislostmi, která je korektně značkovaná metodikou analytické roviny PDT. Připomeňme, že T_P označuje množinu všech korektních A-stromů s koordinacemi a závislostmi, zpracovaných metodikou analytické roviny PDT. ZP označuje množinu zakázaných redukcí na T_P.

Následuje pozorování o konzistenci mezi URAS na T_P a UPRA na L_P.

Podle našich pozorování a naší notace platí, že $L_P = Str(T_P)$, $R(L_P) = R(T_P)$ a $UP = R(ZP)$.

Pozorování 14. *Nechť* $s \vdash_{T_P}^{ZP} t$, *pak* $R(s) \succ_{L_P}^{UP} R(t)$. *Nechť* $s,t \in T_P$ *a* $R(s) \succ_{L_P}^{UP} R(t)$, *pak* $s \vdash_{T_P}^{ZP} t$.

Předchozí pozorování formuluje vlastnost konzistence mezi UPRA a PRAS. Říká, že A-stromy z PDT jsou konstruovány v souladu s větnou redukční analýzou. Toto pozorování je naším základním pozorováním analytické roviny PDT. Přirozeně všechny zde prezentované příklady na URAS a UPRA splňují podmínky konzistence mezi URAS a UPRA.

2.11 Další omezení a výhledy do budoucna.

Následující omezení mají, na rozdíl od předchozích podobnou platnost pro URAS i pro UPRA.

URAS s omezením na počet deletů. Nechť i je přirozené číslo. Označíme jako URAS$(s, T, Z : dl \leq i)$ podmnožinu URAS(s, T, Z), která obsahuje všechny redukce z URAS(s, T, Z), které nemají počet deletů větší než i.

URAS s omezením na vzdálenost vypouštěných uzlů. Nechť k je přirozené číslo. Označíme jako URAS$(s, T, Z : ds \leq k)$ podmnožinu URAS(s, T, Z), která obsahuje všechny redukce z URAS(s, T, Z), které nemají vzdálenost mezi vypouštěnými uzly (podle uspořádání v R-seznamu) větší než k.

Příklad 6. *Uvažujme formální jazyk* $L_1 = \{a^n b^n | n > 0\}$. *Každému slovu (větě) tohoto jazyka přiřadíme A-strom* t_n *následujícím způsobem:*

a) kořenem t_n *bude nejlevější a,*

b) z každého a, které není kořenem vede hrana do jeho levého souseda,

c) z i-tého b vede hrana do i-tého a. Jiné hrany t_n *neobsahuje.*

Budiž $T_1 = \{t_n | n > 0\}$. *Vidíme, že* $TRAS(T_1, \emptyset; dl \leq 2, ds \leq 2, pk \leq 1, MaxIn \leq 0;$ *T-st*$) = T_1$.

Předchozí rovnost dává strukturálně-složitostní charakteristiku T-jazyka T_1. *Zmenšením kteréhokoliv parametru buď rovnost ztrácíme, nebo zmenšení parametru nemá smysl.*

Následující tvrzení není těžké nahlédnout.

Tvrzení 6. *Ke každému* $k \in \mathbb{N}$ *existuje regulární jazyk L, takový, že pro libovolný T-jazyk T takový, že* $Str(T) = L$ *platí, že* $TRAS(T, \emptyset; ds \leq k,$ *Mn-st* $) \neq T$.

Podobné tvrzení platí pro bezkontextové jazyky, které nejsou regulární.

Poznamenejme, že v následujícím zřejmém tvrzení mají označení UPRA$(u, L, \emptyset; ds \leq k)$ a Mn-stabilita analogický význam jako pro URAS.

Tvrzení 7. *Ke každému bezkontextovému jazyku L existuje* $k \in \mathbb{N}$ *takové, že pro libovolné* $u \in R(L)$ *platí, že* $UPRA(u, L, \emptyset; ds \leq k)$ *je Mn-stabilní.*

Předchozí příklad a tvrzení uvádíme, abychom poukázali na souvislosti našich lingvistických pozorování T_P a formální teorií (nekonečných) jazyků. Vidíme, že z pohledu formální redukční analýzy, nejsou A-stromy z T_P příliš složité. Přesto jsme (na základě lingvistického folklóru) očekávali jednodušší a uniformější výsledky.

V budoucnu plánujeme zavést míry neprojektivity a řetězové nesouvislosti založené na redukční analýze a konfrontovat tyto míry s PDT. Očekáváme, že se ukáže souvislost těchto měr s časovou složitostí redukční analýzy.

Reference

[1] Markéta Lopatková, Jirí Mírovský, and Vladislav Kubon. 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.

[2] Martin Plátek, Dana Pardubská, and Markéta Lopatková. 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.

[3] Martin Plátek, Dana Pardubská, and Karel Oliva. Redukční analýza a pražský závislostní korpus. In *Proceedings ITAT 2015: Information Technologies - Applications and Theory, Slovensky Raj, Slovakia, September 17-21, 2015.*, pages 43–50, 2015.