Vztahy mezi segmenty – segmentační schémata českých vět

Markéta Lopatková[1] a Tomáš Holan[2]

[1] ÚFAL MFF UK, Praha, lopatkova@ufal.mff.cuni.cz
[2] KSVI MFF UK, Praha, Tomas.Holan@mff.cuni.cz

Abstrakt *Syntaktická analýza vět přirozeného jazyka, základní předpoklad mnoha aplikovaných úkolů, je složitá úloha, a to zejména pro jazyky s volným slovosledem. Přirozeným krokem, který snižuje složitost vstupních vět, může být vytvoření modulu, ve kterém se určí struktura souvětí ještě před úplnou syntaktickou analýzou. Navrhujeme využít pojem segmentů, snadno automaticky rozpoznatelných úseků vět. Určujeme 'segmentační schémata' popisující vzájemné vztahy mezi segmenty – zejména souřadná a apoziční spojení, podřadná spojení, případně vsuvky.*
V tomto článku představujeme vývojový rámec, který umožňuje vyvíjet a testovat pravidla pro automatické určování segmentačních schémat. Popisujeme dva základní experimenty – experiment se získáváním segmentačních schémat ze stromů Pražského závislostního korpusu a experiment se segmentačními pravidly aplikovanými na prostý text. Dále navrhujeme míry pro vyhodnocování úspěšnosti segmentačních pravidel.

1 Motivace

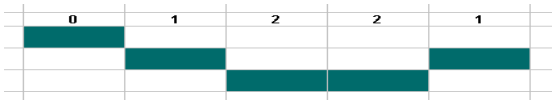
Syntaktická analýza vět přirozeného jazyka, základní předpoklad mnoha aplikovaných úkolů, je složitá úloha, a to zejména pro jazyky s volným slovosledem. Dlouhodobé úsilí přineslo řadu nástrojů pro parsing, které mají poměrně vysokou spolehlivost u krátkých a relativně jednoduchých vět, bylo však ukázáno, že jejich úspěšnost pro složitější souvětí výrazně klesá (viz např. [**?**]).

Přirozeným krokem, který snižuje složitost vstupních vět a zejména souvětí, může být vytvoření modulu mezi morfologickou a syntaktickou analýzou, ve kterém by se určila struktura souvětí. Navrhujeme využít pojem segmentů, tedy lingvisticky motivovaných, přitom však snadno automaticky rozpoznatelných úseků vět, tak jak byl tento pojem navržen v [**?**] a modifikován např. v [**?**]. Tam byla popsána i výchozí sada pravidel, na jejichž základě lze (nedeterministicky) určit segmentační schéma zachycující vztahy jednotlivých segmentů v souvětí. Dalším krokem před samotnou syntaktickou analýzou by byl odhad jednotlivých klauzí, ze kterých se dané souvětí skládá.

Ukažme si základní myšlenku na příkladu konkrétní věty z tisku (**??**). Větu nejprve rozdělíme na jednotlivé segmenty; zde za hranice segmentů považujeme interpunkční čárku, souřadicí spojku *a*, závorky a koncovou tečku (ve větě podtrženy). Potom zachytíme jejich možné vzájemné vztahy – rozlišujeme souřadná a apoziční spojení, podřadná spojení, případně vsuvky. Takto získáme **segmentační schéma**, které umožňuje vymezit základní syntaktickou strukturu souvětí ještě před úplnou syntaktickou analýzou.

(1) *S tím byly trochu problémy , protože starosta v řeči rád zdůrazňoval své vzdělání (však studoval až v Klatovech a v Roudnici), a Víťa tedy občas nutně trochu tápal .*

První segment tvoří hlavní větu souvětí (neobsahuje žádný podřadicí výraz a je zde určité sloveso *byly*). Umístíme ho proto na nultou, základní úroveň v segmentačním schématu. Druhý segment je uvozen podřadicí spojkou *protože* a obsahuje určité sloveso *zdůrazňoval*, můžeme ho určit jako segment podřízený prvnímu segmentu a umístit na první nižší úroveň ve schématu. Následuje otevírací závorka, kterou interpretujeme jako začátek vsuvky, třetí segment tedy umístíme opět o úroveň níž. Čtvrtý segment je oddělen souřadnou spojkou *a*, umístíme ho proto na stejnou úroveň jako třetí segment (třetí segment obsahuje určité sloveso *studoval*, čtvrtý určité sloveso neobsahuje, půjde tedy zřejmě o koordinaci větně členskou). Uzavírací závorka uzavírá vsuvku tvořenou 3. a 4. segmentem. Vzhledem k tomu, že poslední segment obsahuje slovo *tedy*, chápeme trojici *, a tedy* jako příznak souřadnosti, pátý segment analyzujeme jako koordinaci buď k prvnímu, nebo k druhému segmentu. Segmentační schéma můžeme vyjádřit graficky – na obrázku **??** je jedno z možných schémat pro větu (**??**).



Obrázek 1. Segmentační schema 01221

Pokud bychom uměli před úplnou syntaktickou analýzou dostatečně bezpečně určit vztahy mezi jednotlivými segmenty, případně určit možnou strukturu klauzí, syntaktická analýza tohoto souvětí by se zjednodušila a mohla by vykazovat lepší výsledky.

Navíc se ukazuje, že při řadě důležitých aplikačních úloh – jako je např. vyhledávání informací, určování struktury dokumentů a jejich hlavních a vedlejších témat – není potřeba plná syntaktická analýza. Bylo by proto zajímavé zkoumat, nakolik se můžeme omezit pouze na zpracování segmentů na "vyšších" úrovních (a zanedbat hlouběji zanořené struktury).

O přínosech obdobných metod pro analýzu typologicky odlišných jazyků svědčí např. [**?**] či [**?**].

Hlavním cílem tohoto článku je představit vývojový rámec, který umožňuje dále testovat pravidla pro automatické určování segmentačních schémat. Po vymezení pojmu segmentu a segmentačního schématu (oddíl **??**) popisujeme dva základní experimenty – experiment se získáváním segmentačních schémat ze stromů Pražského závislostního korpusu (oddíl **??**) a experiment se segmentačními pravidly aplikovanými na prostý text (oddíl **??**). Následuje oddíl **??**, ve kterém představujeme vhodné míry pro vyhodnocování úspěšnosti segmentačních pravidel. Porovnáváme schémata získaná aplikací těchto dvou sad pravidel s ručně anotovaným vzorkem složitějších vět (i když jsme si vědomi pouze orientačního charakteru těchto výsledků).

2 Vymezení segmentů a jejich příznaky, segmentační schéma

Pod pojmem věta / souvětí zde rozumíme část textu, která je větou v typografickém smyslu (začíná velkým písmenem a končí koncovou interpunkcí, nejčastěji tečkou, otazníkem či vykřičníkem).[3] Jde tedy o posloupnost lexikálních jednotek (v anglicky psané literatuře označované jako 'tokens') $w_1 w_2 \ldots w_n$, kde každá položka w_i reprezentuje buď jednu slovní formu daného přirozeného jazyka, nebo interpunkční znaménko (tečku, čárku, otazník, závorku, pomlčku, dvojtečku, středník, ...). Předpokládáme, že ke každé lexikální jednotce je k dispozici její morfologická analýza.

Na základě morfologické informace a slovní formy rozčleníme větu na jednotlivé úseky a těmto úsekům přiřadíme syntaktické příznaky.[4]

Hranice segmentu

Hranice jsou takové tokeny (slovní formy nebo interpunkce) či jejich posloupnosti, které rozdělují větu na jednotlivé dobře rozeznatelné části označované jako segmenty.

V následujících experimentech považujeme za **elementární hranice** následující symboly:

interpunkce: čárka, dvojtečka, středník, otazník, vykřičník, pomlčka (všech délek), otevírací a uzavírací závorka (všech druhů), svislítko, uvozovky (všech typů); tedy symboly *, : ; ? ! - ( ) [ ] | {} ‘ ’ “ ”, ‘ ,, “*

interpunkce na konci věty

souřadné spojky: morfologická značka začínající dvojicí J^ (viz [**?**])

Pokud ve větě následuje několik elementárních hranic bezprostředně za sebou, označujeme maximální neprázdnou posloupnost těchto hranic jako **(složenou) hranici**.

Segmentem S potom rozumíme maximální neprázdnou posloupnost po sobě jdoucích slovních forem $w_1 w_2 \ldots w_s$, která neobsahuje žádnou hranici.

Pro určování jednotlivých segmentů využíváme následující **doplňující pravidla:**

- Každá věta začíná a končí hranicí (pokud by nebyla hranice na začátku či na konci věty, doplňuje se prázdná hranice).
- Pracujeme se zjednoznačněnou koncovou tečkou (tj. uvnitř věty tečka není hranicí).

Poznamenejme zde, že hranice, určené na základě morfologické analýzy, nejsou nutně jednoznačné. Interpunkční znaménka jsou v užívaném morfologickém slovníku jednoznačná, to ale neplatí pro souřadné spojky (např. slovní forma *ale* je jednak souřadicí spojka, jednak tvar patřící ke třem substantivním lemmatům *ala*). Obecně proto připouštíme nejednoznačnou segmentaci věty. Pro češtinu ovšem existují velmi spolehlivé automatické nástroje, tzv. taggery, které vybírají pro každý token právě jednu morfologickou značku (nejvyšší publikovaná přesnost (accuracy) na prvních dvou pozicích morfologické značky je 99.36%, viz [**?**]). Proto budeme nadále pracovat s daty již předem morfologicky zpracovanými (správnou morfologickou analýzu přebíráme z PDT), a tedy s jednoznačně určenými hranicemi segmentů.

Příznaky segmentu

Morfologická analýza textu obsahuje ovšem řadu dalších (více či méně spolehlivých) informací, které lze využít pro určování vztahů mezi jednotlivými segmenty. Tyto informace zachycujeme pomocí **příznaků**, které přidělujeme jednotlivým segmentům – zatím pracujeme pouze s příznakem podřízenosti. Předpokládáme, že dalšími důležitými příznaky bude například příznak souřadnosti a příznak určitého slovesa.

[3] Rozčleněním textu na věty se zde nezabýváme, přebíráme je z Pražského závislostního korpusu.

[4] Zde se odchylujeme od návrhu v [**?**], kde se uvažovaly tzv. separátory, které sloužily jednak jako hranice, jednak jako příznaky podřízenosti.

Příznak podřízenosti (PP). Příznak podřízenosti se přiděluje takovému segmentu, který obsahuje (alespoň jednu) slovní formu, jejíž morfologická značka má první dvě pozice z následujícího výčtu (pro zájmena a číslovky), nebo je jedním z uvedených zájmenných příslovcí:

podřadicí spojka: `J,`
tázací/vztažné zájmeno: `P4`, `PE`, `PJ`, `PK`, `PQ`, `PY`
číslovka: `C?`, `Cu`, `Cz`
zájmenné příslovce: *jak, kam, kde, kdy, proč, kudy*

Segmentační schéma

Vztahy mezi segmenty popisujeme pomocí segmentačních schémat, která zachycují **úrovně vnoření jednotlivých segmentů**. Základní myšlenka úrovní segmentů je jednoduchá:

- Segmenty tvořící (všechny) hlavní klauze souvětí mají úroveň 0;
- Segmenty, tvořící klauze závislé na klauzích se segmenty na úrovni k, mají úroveň $k + 1$ (úroveň vnoření je vyšší);
- Segmenty, tvořící koordinované, příp. aponované výrazy, mají stejnou úroveň;
- Segmenty, tvořící vsuvky (např. obsahy závorek), mají úroveň vnoření zvýšenu o 1 oproti segmentům, které je obklopují.

Segmentační schéma můžeme vyjádřit graficky (viz obrázek **??**) nebo jako vektor čísel (pro větu (**??**) dostaneme dva vektory odpovídající dvěma segmentačním schématům (01220) a (01221)).

3 Experimenty s automatickým určováním segmentačních schémat

3.1 Jak získat segmenty ze syntaktických stromů?

Zde popíšeme možný postup, jak z analytické roviny Pražského závislostního korpusu[5] (PDT, [**?**]) určovat segmentační schémata pro jednotlivé věty. Analytická rovina PDT zachycuje povrchovou syntaxi, tedy v zásadě tytéž informace, které by měly umožnit segmentaci vět a vymezení možných úrovní jednotlivých segmentů.

Věta na analytické rovině je zachycena jako závislostní strom – uzly odpovídají slovním tvarům a interpunkci, hrany primárně závislostním vztahům (hrany považujeme za orientované od závislého k řídícímu uzlu). Mimo závislostní vztahy jsou zde zachyceny též vztahy koordinace a apozice – uzel odpovídající koordinační spojce (příp. jinému spojovacímu výrazu nebo interpunkci) je rodičem jednotlivých koordinovaných výrazů, podobně spojovací výraz uvádějící apozici je rodičem aponovaných výrazů. Typy vztahů reprezentované hranami jsou dále specifikovány tzv. analytickými funkcemi. V závislostní reprezentaci věty tedy nenajdeme uzly, které by odpovídaly konceptu segmentů, přesto by měla obsahovat informace, na jejichž základě lze segmentační schéma věty vymezit.

Pro popis pravidel potřebujeme zavést pojem cesty mezi segmenty a dále skupiny segmentů. Řekneme, že pro větu W **vede hrana ze segmentu** S_i **do segmentu** S_j $(S_i, S_j \subset W)$, pokud pro nějaké slovo $u \in S_i$ existuje slovo $v \in S_j$ takové, že v závislostním stromě T věty W existuje cesta z u do v. Dále řekneme, že pro větu W **vede cesta ze segmentu** S_i **do segmentu** S_j $(S_i, S_j \subset W)$, pokud existuje posloupnost segmentů $S_i = S_{p_1} \,..\, S_{p_m} = S_j,\ S_{p_k} \subset W (k = 1 \,..m)$ taková, že pro každé $k = 1 \,..\, m-1$ vede hrana ze segmentu S_{p_k} do segmentu $S_{p_{k+1}}$. **Skupina segmentů** G je taková množina segmentů věty W, ve které pro každou dvojici segmentů $S_i, S_j \in G$ platí, že z S_i vede cesta do S_j (a tedy zároveň z S_j vede cesta do S_i).

Při získávání segmentačního schématu pro jednotlivé věty z PDT postupujeme následujícím způsobem.

Určení segmentů: Prvním krokem pro získání segmentačního schématu je určení hranic a segmentů.

Skupiny segmentů: Určíme jednotlivé skupiny segmentů.

Úroveň nula: Všechny segmenty ze skupin segmentů, ze kterých vede cesta do kořene závislostního stromu T buď přímá (tj. hrana) nebo pouze přes uzly reprezentující (elementární) hranice, dostanou úroveň nula.

Koordinace a apozice: Sousedí-li se segmentem S_i, pro který známe úroveň, segment S_j s neznámou úrovní a odpovídá-li hranice mezi nimi koordinačnímu či apozičnímu výrazu (např. uzel reprezentující elementární hranici má analytickou funkci Coord či Apos), dostane segment S_j stejnou úroveň jako segment S_i.

Hlouběji vnořené úrovně: Všechny segmenty ze skupin segmentů, které nemají stanovenu úroveň a ze kterých vede cesta buď přímá (tj. hrana) nebo pouze přes uzly reprezentující elementární hranice do segmentu s úrovní k, dostanou úroveň $k + 1$.

Koordinace a apozice: Opět zkontrolujeme sousedy všech segmentů se známou úrovní (viz výš).

Tento postup opakujeme, dokud nejsou určeny úrovně vnoření všech segmentů.

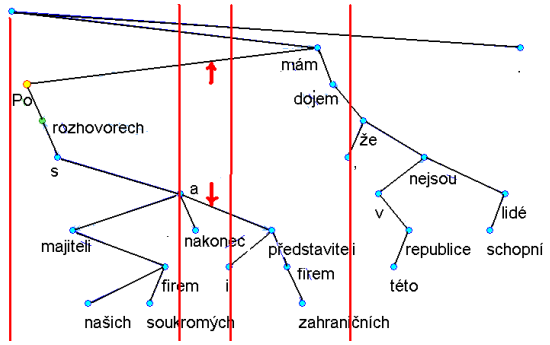
Tato výchozí sada pravidel pro získání segmentů z PDT určí pro každou vstupní větu reprezentovanou analytickým stromem právě jedno segmentační schéma (ne nutně správné).

[5] http://ufal.mff.cuni.cz/pdt2.0/

Ukažme si postup na konkrétní větě (**??**), jejíž analytický strom je na obrázku 2.

(2) _ *Po rozhovorech s majiteli našich soukromých firem a nakonec i představiteli firem zahraničních mám dojem , že v této republice nejsou schopní lidé .*

Věta (**??**) se skládá ze čtyř segmentů (ve větě hranice podtrženy, na obrázku odděleny svislými čarami). První a třetí segment tvoří skupinu (hrana z uzlu *po* do uzlu *mám* a zároveň cesta z uzlu *představiteli* do uzlu *s*, viz šipky). Tyto dva segmentu mají úroveň 0, neboť z uzlu *mám* vede hrana do kořene stromu. Druhý segment získá též úroveň 0, neboť jeho hranice s prvním segmentem odpovídá koordinačnímu výrazu. Čtvrtý segment získá úroveň vnoření 1, neboť z něj vede hrana do třetího segmentu s již stanovenou úrovní 0. Nalezené segmentační schéma věty tedy je (0001) (v tomto případě jde o správné segmentační schéma).



Obrázek 2. Analytický strom věty (**??**) s vyznačenými segmenty.

3.2 Výchozí sada segmentačních pravidel pro text

Výchozí sada (heuristických) segmentačních pravidel byla publikována v [**?**]. Tato pravidla jsme zpřesnili a implementovali, abychom mohli porovnat jimi získaná segmentační schémata s výsledky segmentace na základě stromů PDT a s ruční anotací. Předpokládáme, že zpracovávaný text je již rozdělen na jednotlivé věty.

Při určování segmentačního schématu z prostého textu postupujeme vždy od začátku věty a chceme pro každý segment určit jeho úroveň vnoření.

Dále uvedená pravidla určují, jaká úroveň vnoření má být přiřazena prvnímu segmentu. Dále určují, jak se může měnit úroveň vnoření při překročení té které hranice.

Protože pravidla nedávají vždy jednoznačnou odpověď (například čárka může ukončovat jednu nebo více úrovní vnoření), namísto jedné úrovně vnoření každému segmentu přiřazujeme **interval** úrovní, ve kterých se může nacházet. Výsledné intervaly určují množinu segmentačních schémat.

Jednotlivé segmenty mohou být odděleny složenou hranicí, tedy posloupností několika elementárních hranic. V takovém případě jsou pravidla uplatňována postupně na jednotlivé elementární hranice. Proto pravidla neurčují, jaká bude úroveň vnoření následujícího segmentu, ale to, jak se změní globální průběžný ukazatel úrovně při zpracování elementární hranice. Segmentu je nakonec přiřazena taková úroveň, resp. takový interval úrovní, jaký byl nastaven po zpracování poslední elementární hranice před začátkem tohoto segmentu.

Pokud se podmínky dále uvedených pravidel odkazují na následující segment, odkazují se všechny na první segment ležící za složenou hranicí.

Začátek věty: Pokud první segment nemá přiřazen PP, bude umístěn na základní úrovni 0. Má-li PP, bude umístěn na úrovni 1.

Čárka: Pokud následující segment nemá PP, bude dolní mez intervalu u ukazatele úrovně zachována, horní mez bude nastavena na 0 (odpovídá konci libovolně mnoha vnořených klauzí). Má-li následující segment PP, bude ukazatel úrovně vnoření o 1 zvětšen (odpovídá začátku vnořené klauze či její části).

Otevírací závorka (libovolného druhu): Pokud následující segment nemá PP, bude úroveň vnoření o 1 zvýšena (začátek vsuvky). Má-li následující segment PP, bude úroveň zvýšena o 2 (vsuvka s hlouběji vnořenou strukturou).

Uzavírací závorka (libovolného druhu): Pokud jí předchází dosud neuzavřená otevírací závorka stejného druhu, bude úroveň vnoření (příp. interval možných úrovní) nastavena na úroveň platnou před zpracováním otevírací závorky, jinak se úroveň nezmění (toto omezení odpovídá třeba případům výčtů *a)... b)...*).

Koordinační spojka: Úroveň zůstává beze změny.

Dvojtečka: Pokud následující segment nemá PP, zůstane horní mez intervalu úrovně nezměněná (odpovídá koordinaci či apozici), dolní mez bude zvýšena o 1 (začátek vnořené klauze či její části). Má-li následující segment PP, bude horní mez zvýšena o 1 a dolní mez o 2 (hlouběji vnořená (část) klauze).

Otazník, vykřičník: Dolní mez u ukazatele úrovně bude snížena o 1, horní mez bude nastavena na 0 (konec libovolně mnoha vnořených klauzí).

Středník: Dolní mez u ukazatele úrovně bude zachována, horní mez bude nastavena na 0.

Svislítko, pomlčka, uvozovky: Úroveň zůstává beze změny.

Pokud v takto získaném schématu nemá žádný segment úroveň nulu, ‚posuneme' celé schéma o úroveň výše (posun -1).

Tato pravidla určí pro každou morfologicky analyzovanou vstupní větu množinu segmentačních schémat.

4 Vyhodnocování a rozbor výsledků

4.1 Testovací data a možné míry úspěšnosti

V předchozích oddílech jsme popsali základní experimenty s automatickým určováním segmentačních schémat z prostého textu a ze stromů PDT. Abychom mohli tato základní pravidla dále vyvíjet a zlepšovat, potřebovali jsme vytvořit testovací sadu vět se správně určenými segmentačními schématy.

Vybrali jsme proto z vývojovych dat PDT 2.0 (tzv. dtest data, 5 228 vět) věty, které obsahují alespoň pět segmentů (707 vět). Z těchto vět jsme pro každou desátou větu ručně určili segmentační schéma. Tím jsme získali 71 poměrně strukturně složitých vět s označenými segmentačními schématy.

Zdůrazněme zde, že tento výběr vět (průměrně 6,49 segmentu na větu) do značné míry zhoršuje naměřené výsledky oproti náhodnému výběru, neboť vylučuje věty z hlediska segmentace jednoduché (průměrný počet segmentů na větu v celých dtest datech činí 2,72).

Poznamenejme dále, že mnohé z testovacích vět jsou homonymní, mají více možných syntaktických stromů, v PDT je však zachycena vždy pouze jedna z možných struktur. Při určování segmentačního schématu jsme se drželi struktury, kterou zachytili anotátoři PDT. Každé větě tedy bylo přiřazeno jediné schéma (např. větě (**??**) bylo přiřazeno pouze segmentační schéma (01221)).

Chceme-li vyhodnotit úspěšnost navržených pravidel, nabízí se několik možných přístupů. Jako nejjednodušší vhodná míra se jeví míra počítající shodu úrovně vnoření jednotlivých segmentů, dále tuto základní míru značíme ρ.

Zkoumáme-li ovšem výsledky experimentů, zjišťujeme, že v řadě případů chyba při určování úrovně pro jeden segment má za následek chybně určené úrovně u dalších segmentů, přestože vztahy mezi jednotlivými segmenty jsou určeny správně. Například věta (**??**) má správné segmentační schéma (2233110), pomocí pravidel z PDT pak bylo navrženo schéma (1122000); přestože byly správně určeny skoro všechny vztahy mezi segmenty, úrovně se shodují u jediného segmentu (hranice opět podtrženy).

(3) *„ Když to odečtete od výplaty spolu se ztrátou při výměně slovenských korun za české a za pojištění , které se musí platit tam i u nás , nezbude manželovi z výplaty ani polovina ," zlobí se paní Krajčová .*

Tento nedostatek základní míry lze odstranit tím, že povolíme posun celého výsledného schématu tak, aby se shodovaly úrovně co největšího počtu segmentů (pro větu (**??**) je to posun hodnoceného schématu o +1, který nám dá shodu na šesti segmentech). Tuto míru značíme σ.

Protože nás primárně zajímají právě vztahy mezi segmenty, je vhodné uvažovat i míru (značíme δ), která měří správnost rozdílu úrovní dvou sousedních segmentů (např. schémata (101) a (211) mají stejné vztahy mezi prvním a druhým segmentem (rozdíl -1), ale různé vztahy mezi druhým a třetím segmentem).

Přestože hlavním cílem tohoto příspěvku není implementace konkrétní sady pravidel, ale vytvoření vhodného prostředí pro vývoj a testování těchto pravidel, podívejme se na úspěšnost dvou základních sad pravidel popsaných výše (jakkoliv mají tyto výsledky pouze informativní povahu danou nejen neúplností zatím navržených pravidel, ale i malým rozsahem testovacích dat).

4.2 Úspěšnost pravidel pro segmentační schémata ze syntaktických stromů

Základní sada pravidel pro získání segmentů z PDT určí pro každou vstupní větu právě jedno segmentační schéma. Při vyhodnocování úspěšnosti proto stačí *přesnost*, anglicky *accuracy*.[6] Úspěšnost těchto pravidel je shrnuta v tabulce **??**.

Tabulka 1. Úspěšnost základní sady pravidel pro získání segmentů ze stromů PDT (při povoleném posunu ±2)

accuracy	základní míra		míra s posunem	
# segmentů	# správně	ρ	# správně	σ
461	264	0,57	335	0,73

Vyhodnocujeme-li rozdíl úrovní dvou sousedních segmentů, dosahují pravidla úspěšnost $\delta = 0,70$ (správně bylo určeno 274 ze 390 vztahů mezi segmenty).

Zmiňme zde tři hlavní problémy, které snižují úspěšnost výchozích pravidel pro automatické určování segmentačních schémat ze stromů PDT.

1. Větný člen, který tvoří samostatný segment, má přiřazenu úroveň vnoření o 1 vyšší než segment s jeho řídícím členem. Např. větám *Petr, který nikdy nelže, tentokrát zalhal.* či *Včera, kdy*

[6] Míry pokrytí (*recall*) a přesnost (*precision*) se rovnají.

tak pršelo, přišli. se správným segmentačním schématem (010) bude přiřazeno nesprávné segmentační schéma (120).

2. Zatím neřešena je speciální (ale poměrně častá) česká konstrukce, kdy za sebou následují dva podřadící výrazy (podtržené), jako např. *Nevěděl, že když jsem se probral k vědomí, zavolal jsem policii.*
3. Dalším rozšířeným jevem, na který je potřeba se soustředit, jsou koordinační (příp. apoziční) spojení více než dvou členů.

4.3 Úspěšnost pravidel pro segmentační schémata z textu

Vyhodnocování úspěšnosti při určování segmentačního schématu z prostého textu spočívá v počítání, zda pro jednotlivé segmenty výsledný interval úrovní obsahuje správnou úroveň vnoření toho kterého segmentu (zatím tedy měříme pouze tzv. *recall*), viz tabulku **??**. (Vzhledem k nejednoznačně určeným vztahům mezi segmenty zde neuvádíme míru δ.)

Tabulka 2. Úspěšnost základní sady pravidel pro získání segmentů z prostého textu (povolené posouvání ±2)

recall	základní míra		míra s posunem	
# segmentů	# správně	ρ	# správně	σ
461	297	0,64	349	0,76

Průměrný počet segmentačních schémat na jednu větu pro naše testovací data je 2,17 (průměrná míra víceznačnosti na celém dtestu je 1,32).

Zmiňme zde opět alespoň několik jevů, které výchozí sada pravidel nepostihuje adekvátně a které musí být předmětem bližší specifikace (ta však vyžaduje další podrobné lingvistické zkoumání).

1. Nejsou stanovena pravidla pro přímou a polopřímou řeč, např. ve všech nalezených segmentačních schématech věty (**??**) nebudou úrovně prvních čtyř segmentů dostatečně hluboko (nalezené úrovně (1122) místo správných (2233)).
2. Zatím není řešena koordinace několika klauzí s opakujícím se příznakem podřízenosti (např. věta *Jak účelně větrat, jak nepřetápět, jak spotřebu měnit a podle toho účtovat.* bude mít přiřazeno chybné schéma (0122) místo správného schématu (0000).)

Shrnutí

Segmentačního schéma zachycuje vztahy mezi jednotlivými segmenty věty, tedy lingvisticky motivovanými, přitom však snadno automaticky rozpoznatelnými úseky vět. Určuje tedy základní strukturu souvětí ještě před úplnou syntaktickou analýzou.

Příspěvek se soustřeďuje na popis rámce pro vývoj pravidel pro automatický odhad segmentačních schémat vět. Tento rámec umožňuje přesnou formulaci a zjemňování segmentačních pravidel. Dále jsme představili vhodné míry pro vyhodnocování úspěšnosti segmentační analýzy.

V této fázi vývoje byly implementovány dvě sady pravidel – pravidla pracující se syntaktickou strukturou Pražského závislostního korpusu a pravidla pro zpracování prostého textu. Porovnali jsme výsledky těchto pravidel s ručně anotovaným vzorkem vět z PDT, i když je zřejmé, že tyto výsledky mají pouze informativní charakter.

Provedené experimenty umožnily bližší specifikaci segmentačního schématu a pravidel pro ruční anotaci. Pro další výzkum segmentační analýzy a zpřesňování pravidel je potřeba vytvořit řádově větší testovací vzorek vět. Ukazuje se, že bez manuální anotace velké sady vět se v další fázi výzkumu neobejdeme.

Poděkování

Práce na tomto projektu je podporována Grantovou agenturou ČR, GAČR 405/08/0681 a částečně též programem Informatická společnost č. 1ET100300517.

Reference

1. Holan, T.: O složitosti Vesmíru. In Obdržálek, D., Štanclová, J., Plátek, M. (eds.) Malý informatický seminář MIS 2007. MatFyzPress, 2007. pp. 44-47
2. Kuboň, V.: Problems of Robust Parsing of Czech. Ph.D. Thesis, MFF UK, Prague, 2001
3. Kuboň, V., Lopatková, M., Plátek, M., Pognan, P.: A Linguistically-Based Segmentation of Complex Sentences. In Wilson, D.C., Sutcliffe, G.C.J. (eds.) Proceedings of FLAIRS Conference. AAAI Press, 2007. pp. 368-374
4. Jones B.E.M.: Exploiting the Role of Punctuation in Parsing Natural Text. In Proceedings of the COLING'94. Kyoto, 1994. pp. 421-425
5. Ohno, T., Matsubara S., Kashioka, H., Maruyama, T., Inagaki, Y.: Dependency Parsing of Japanese Spoken Monologue Based on Clause Boundaries In Proceedings of COLING and ACL. ACL, 2006. pp. 169-176
6. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). UK, Nakladatelství Karolinum, Praha, 2004
7. Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P.: The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In Proceedings of Balto-Slavonic NLP Workshop. ACL, Prague, 2007. pp. 67–74
8. Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M.: Prague Dependency Treebank 2.0. LDC, Philadelphia. 2006