Od segmentů ke klauzím v češtině – analýza vybraných jevů*

Vladislav Kuboň and Markéta Lopatková

Ústav formální a aplikované lingvistiky
Univerzita Karlova v Praze, Česká republika
{vk,lopatkova}@ufal.mff.cuni.cz

Abstrakt *Článek se zabývá problémem skládání klauzí v českých souvětích z jednotlivých segmentů identifikovaných pomocí spojek a interpunkčních znamének. Množství segmentů je obvykle větší než počet klauzí, proto je při syntaktické analýze souvětí nutné rozpoznat a spojit jednotlivé segmenty do klauzí a určit vzájemné postavení těchto klauzí. Článek navrhuje a předkládá k diskusi určitá pravidla, která vycházejí z české gramatiky a z lexikálně syntaktických vlastností českých slov. Tato pravidla se opírají o analýzu jevů, důležitých pro určení vzájemného vztahu českých klauzí, a o jejich frekvenci. Pravidla jsou vytvořena převážně na základě dat, získaných pro tento úkol z Pražského závislostního korpusu.*

1 Úvodní poznámky – hranice a segmenty

Syntaktická analýza souvětí představuje jednu z cest k vylepšení výsledků nejrůznějších analyzátorů, ať již jsou založeny na tradičních metodách ručního psaní gramatik nebo na metodách statistických. V obou případech totiž informace o složení souvětí, o vzájemném vztahu jednotlivých klauzí i o složení jednotlivých klauzí mohou podstatným způsobem zjednodušit celý proces syntaktické analýzy. Bez ohledu na použitou metodu se totiž ukazuje, že čím delší a složitější souvětí, tím nižší úspěšnost analýzy. Podpůrné argumenty pro toto tvrzení je pro pravidlové gramatiky možné nalézt například v práci [1], pro metody statistické v práci [2].

V článku [3] jsme navrhli způsob rozdělení českých souvětí na tzv. **segmenty**, jednoznačně definované úseky, které vždy zcela určitě patří do jedné české klauze. Tento přístup byl v příspěvku [4] dále modifikován pro potřeby automatického zpracování a ručních anotací – na základě morfologické analýzy jsou určeny tzv. **hranice segmentů**, tedy souřadicí spojky a interpunkce. Rozdělení souvětí na segmenty pomocí takto definovaných hranic je možné díky poměrně striktním pravidlům, která existují v české gramatice pro interpunkci a pro používání souřadicích (a podřadicích) spojek; tyto výrazy jednoznačně oddělují jednotlivé segmenty. V případech, kdy samotná identifikace hranice není jednoznačná, je možné pomocí morfologického značkovače (taggeru) poměrně spolehlivě rozhodnout, zda se jedná o hranici či nikoli (např. výraz *jak* může být buď souřadicí spojkou, a tedy hranicí; nebo je podřadicí spojkou, zájmenným příslovcem či podstatným jménem, v takovém případě ho za hranici nepovažujeme).

Stejný článek také konstatuje, že spojování jednotlivých segmentů do klauzí a určování jejich vzájemného vztahu je problémem mnohem obtížnějším než nalezení všech segmentů v souvětí. Hlavním cílem tohoto článku je analyzovat tento problém a navrhnout řešení, které by se dalo v budoucnu implementovat.

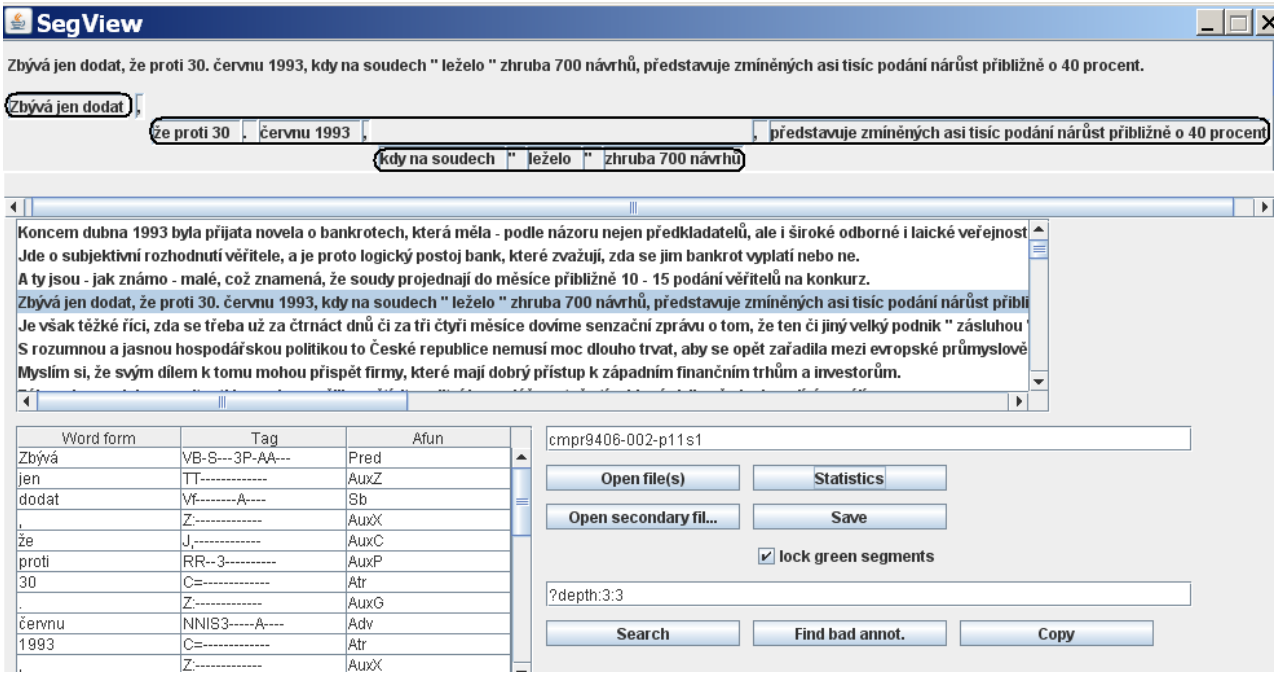
2 Segmentace a syntaktická analýza

Syntaktická analýza jazyků s volným slovosledem, mezi něž patří i čeština, se potýká s celou řadou problémů. Jedním ze základních problémů je správné určení **jednotlivých klauzí** v souvětí a jejich **vzájemných vztahů**. Tento problém můžeme ilustrovat například větou *Vyskytl se i případ, kdy nájemník neplatil nájem po určitou dobu, kdy byl nezaměstnaný, a po nalezení zaměstnání dluh uhradil.* Poslední segment *po nalezení zaměstnání dluh uhradil* můžeme analyzovat dvojím způsobem, buď jako klauzi koordinovanou s hlavní klauzí *Vyskytl se i případ [...] a po nalezení zaměstnání dluh uhradil.*, nebo jako klauzi tvořící s další klauzí koordinované přívlastkové věty *kdy nájemník neplatil nájem po určitou dobu [...] a po nalezení zaměstnání dluh uhradil.*[1]

Dalším významným problémem syntaktické analýzy (nejen) češtiny je obtížnost určení dosahu **vnořených větných konstrukcí** (vedlejších vět, vsuvek). Ty mají sice obvykle velmi snadno rozpoznatelný začátek (podřadicí spojka, vztažné zájmeno, zájmenné příslovce apod.), jejich konec se však rozpoznává již mnohem hůře. Vezměme si například následující souvětí *S úlevou nám sdělil, že ztracený Petr přišel, s chabou omluvou, až včera večer.* Poslední segment *až včera večer* je příslovečným určením času,

* Výsledky, o kterých referujeme v tomto článku, byly dosaženy za podpory grantu GAČR č. 405/08/0681.

[1] Poznamenejme, že z hlediska syntaktické analýzy jsou obě analýzy správné; preference druhé analýzy je dána až na úrovni porozumění větě v kontextu promluvy (pragmatiky).



Obr. 1. Editor SegView: segmentační schéma věty *Zbývá jen dodat, že proti 30. červnu 1993, kdy na soudech ,,leželo" zhruba 700 návrhů, představuje zmíněných asi tisíc podání nárůst přibližně o 40 procent* (klauze vyznačeny elipsami, maximální úroveň zanoření 2).

které následuje po vsuvce *s chabou omluvou* a syntakticky může tvořit klauzi jak s hlavní klauzí *S úlevou nám sdělil [...] až včera večer.*, tak s vedlejší klauzí *že ztracený Petr přišel [...] až včera večer.* Rozřešení této víceznačnosti je možné až na základě porozumění kontextu, z čistě syntaktického hlediska musíme uvažovat obě varianty. Tyto konstrukce představují velký problém pro všechny způsoby analýzy včetně analýzy založené na statistických metodách.

Nejzávažnějším jevem, který ztěžuje syntaktickou analýzu, jsou ovšem **koordinační**, případně **přístavkové (apoziční) konstrukce**. Z hlediska čistě syntaktické analýzy nejde ani tak o problém rozlišení těchto dvou jevů, ale hlavně o rozlišení **větně členské** a **větné koordinace**, protože toto rozlišení určuje typ vzájemného vztahu mezi jednotlivými klauzemi. Vezměme například úsek věty obsahující jmennou frázi *... Šavrda, kapitán, psycholog a kouč ...*, která může popisovat jak několik členů realizačního týmu (jako prostá členská koordinace), tak i složený přístavek rozvíjející vlastní jméno *Šavrda*; teprve v případě, že k celé konstrukci přidáme zprava výraz *v jedné osobě*, bude jasné, že se jedná o druhý případ; bude to ale jasné pouze na základě porozumění danému úseku, z čistě syntaktického hlediska opět zůstanou ve hře obě varianty. Problém rozlišení větné a větně členské koordinace podrobněji ilustrujeme v následující sekci.

Tyto a další konstrukce (podrobněji popsané například v práci [1]) přímo vyzývají k zapojení určitého mezikroku mezi morfologickou a syntaktickou analýzu. Tento mezikrok by umožnil nalézt určité dobře definované celky (segmenty), které by bylo možné pomocí speciálních pravidel pospojovat do klauzí. Vzhledem k tomu, že počáteční úseky klauzí obvykle dobře definují roli klauze v souvětí (různé příznaky podřízenosti, zejména podřadicí spojky, vztažná zájmena apod., se obvykle nacházejí na začátku prvního segmentu klauze), by tato pravidla mohla pomoci stanovit i vzájemné postavení jednotlivých klauzí v souvětí, a tím výrazně ulehčit další kroky syntaktické analýzy (vložené klauze je možné zpracovávat nezávisle). Neboť ať se již rozhodneme pro jakoukoli metodu analýzy, její úspěšnost bude pro kratší jednoduché věty vyšší [2].

2.1 Dostupná data a nástroje

Aby bylo možné chování segmentů zkoumat, bylo nutné vytvořit odpovídající data. Dostupné syntakticky anotované korpusy (pro češtinu zejména Pražský závislostní korpus (PDT), viz [5]) se totiž většinou soustřeďují na vztahy mezi jednotlivými slovy, chybí v nich explicitní znázornění vztahů mezi většími větnými celky. Proto jsme se rozhodli data z PDT upravit do podoby vhodnější pro naše experimenty.

počet segmentů	počet vět	počet klauzí								
		1	2	3	4	5	6	7	8	9
1	942	942								
2	804	396	408							
3	583	165	236	182						
4	400	100	124	107	69					
5	275	48	81	7	4	29				
6	171	26	30	45	35	25	10			
7	85	10	22	24	14	7	6	2		
8	61	12	7	13	13	9	5	1	1	
9	40	7	8	7	6	3	4	1	2	2
10	26	3	1	6	2	3	3	3	4	1
11	24	1	3	3	3	2	3	4	3	2
vce	33	11	5	2	3	2	4	4	1	1

Tab. 1. Počty klauzí a segmentů v datech.

Použili jsme jednak automatickou metodu, podrobně popsanou v článku [6], jednak ruční anotaci [7]. Z dat získaných ruční anotací vycházíme i v tomto příspěvku – jde o soubor 3 444 vět z PDT, u kterých byla určena struktura souvětí. Při zpracování dat se vycházelo z pojmu segmentu, anotátoři určovali jednak vztahy mezi jednotlivými segmenty (zda a v jakém vztahu jsou části vět vyjádřené jednotlivými segmenty, tedy zda vyjadřují části věty souřadně spojené či je mezi nimi vztah řídící-závislá klauze, případně zda jde o vsuvku), jednak vyznačovali jednotlivé větné klauze.

Anotátoři používali speciální editor SegView;[2] ukázka anotace pomocí tohoto editoru je na obrázku 1. SegView umožňuje kromě vlastních anotací též vyhledávání zajímavých příkladů – kromě triviálních dotazů na formu, lemma či tag SegView dovoluje vyhledávání zajímavých struktur, jako je například souvětí o určitém počtu klauzí, souvětí s určitou hloubkou zanoření či souvětí s maximálním ‚skokem' mezi jednotlivými segmenty. Základní statistiky dokládající frekvenci jevů, důležitých pro určení vzájemného vztahu českých klauzí, představíme v následující sekci.

3 Analýza vybraných jevů

3.1 Kvantitativní analýza

První typ analýzy dat, který nám umožňuje nástroj SegView, je analýza kvantitativní. Ta nám pomohla zjistit některé vlastnosti českých textů, které jsou důležité pro vytvoření algoritmu pro rozklad souvětí na klauze a zjištění jejich vzájemných vztahů. Její výsledky jsou zachyceny v tabulce 1. Celkem jsme ve 3 444 analyzovaných větách identifikovali 10 746 segmentů, které tvoří 6 571 klauzí.

[2] Zde bychom chtěli poděkovat autorovi editoru Petru Homolovi, který též zajistil technickou podporu při anotacích a při vyhodnocování dat.

Čísla obsažená v tabulce 1 na první pohled dokumentují nijak překvapivou skutečnost, že jednoduché věty a souvětí složená z jedné až dvou klauzí s maximálně dvěma segmenty představují podstatnou část dat. Jedná se celkem o 1 746 vět, tedy o nepatrně více než polovinu celkového množství. Tyto věty jsou z hlediska zpracování poměrně triviální, protože i u vět složených ze dvou klauzí odpadá problém nalezení konce obou klauzí, jejich vzájemný vztah (souřadnost nebo podřízenost, případně vsuvka) se obvykle také dá velmi snadno určit na základě charakteru spojení obou klauzí.

Na druhém konci škály nalezneme několik zajímavých extrémních případů. Jedním z nich je bezesporu i věta s maximálním počtem segmentů (27) z celé množiny dat. Tato věta zároveň sestává pouze z jediné klauze, nejedná se tedy o souvětí. Je to věta lnd94101-082-p1s13 *Tenis Atlanta - 2. kolo: Chang - Mattar 6 : 3, 7 : 5, Martin - Dunn 6 : 3, 6 : 2, Agassi - Reneberg 4 : 6, 6 : 2, 6 : 4, Washington - Connors 6 : 4, 3 : 6, 6 : 1.* Tato věta (podobně jako další extrémní příklady nesouladu mezi počtem segmentů a klauzí) představuje velmi specifický případ, se kterým ovšem musíme v naší analýze také počítat, neboť všechny věty obsažené v PDT jsou reálnými větami a v reálném textu se samozřejmě nejrůznější speciální případy vyskytují, a to dokonce ve nezanedbatelném množství.

Podobným typem věty je i věta mf920901-025-p3s4, která má čtyři klauze a dvacet segmentů: *Oslovili jsme lidi vesměs známé, zajímavé a talentované (mj. Jireš, Špáta, Vihanová, Vorel, Němec, Císařovský, Pavlásková, Svěrák, Chaun, Kačírek, Koutecký) s tím, že každý měl zároveň navrhnout ,,svůj objekt", hrdinu portrétu, který by rád osobně natočil.*

Výhodou je, že tyto extrémní případy jsou snadno i v běžném textu identifikovatelné pomocí relativně jednoduchých neligvistických pravidel – sportovní výsledky, dlouhé ‚nákupní' seznamy, nejrůznější ta-

bulky apod. se dají automaticky rozpoznat, vyznačují se např. vysokým počtem velmi krátkých segmentů a malým počtem určitých sloves. Je tedy zřejmě správné předřadit celému procesu lingvisticky motivované analýzy právě identifikaci několika častých typů ‚podezřelých vět'. Lingvisticky motivovaná analýza se potom může soustředit na jádro celého problému, tedy zejména na věty nacházající se v tabulce 1 zhruba uprostřed.

3.2 Lingvistická analýza

Kromě kvantitativní analýzy jsme se pokusili lingvisticky zanalyzovat konkrétní věty, abychom získali určitou představu o tom, co je určující pro spojování segmentů do klauzí a kam tedy nasměrovat další výzkum. Při práci jsme vyšli i z článků popisujících podobný výzkum pro jiné jazyky. Například v článku [8] autoři navrhují algoritmus spojování segmentů do klauzí ve slovinštině. Zaměřují se zejména na koordinace uvnitř klauzí a koordinace mezi klauzemi. Jejich přístup však bohužel nelze přímo přejmout, protožed používají hodně netriviálních heuristik, které jsou poněkud neprůhledné a většinou nemají přímou linvistickou oporu. Autoři také relativně opomíjejí ostatní typy vztahů, které mohou pomoci pospojovat jednotlivé segmenty do klauzí (valence apod.). Podívejme se tedy, jak by bylo možné tento problém řešit v češtině.

Uveďme si nejprve základní fakta, která o českých souvětích můžeme vyčíst z dostupných dat. Především platí, že **počet klauzí** v českých souvětích se obecně odvíjí od **počtu určitých sloves** v souvětí. Není vždy totožný, např. nejrůznější nadpisy, výčty, vsuvky, texty v závorkách apod. nemusejí žádné sloveso obsahovat, přesto je budeme za klauze považovat. Podobně složitá je i otázka přechodníků. Ve většině případů budou mít samostatný segment, oddělený čárkou od zbytku souvětí, ovšem není tomu tak vždy. Například ve Šmilauerově Novočeské skladbě [9] nalezneme několik příkladů (zejména ze starší literatury), kdy jeden segment pobsahuje jak přechodník, tak i hlavní (určité) sloveso – *Nasytiv se chlebem usnul* (Jirásek) či *Chlapec směje se dobře mu odpověděl* (Němcová). V těchto případech tedy počet klauzí bude menší než počet určitých tvarů sloves (do kterých budeme počítat i přechodníky). Ve větách jako mf920924-004-p2s14A *Letos se ale neujalo nic.* nebo ln95040-062-p2s9 *Stejně jako další z legend, kterými hoteliér láká hosty do lokálu.* naopak nalezneme více klauzí než sloves. Zdá se tedy, že bude velmi těžké stanovit nějaký obecně platný vzorec pro vztah počtu klauzí a určitých sloves, spíš bude nutné pracovat s konkrétními typy konstrukcí, které budeme ve větách identifikovat (například vložené výrazy neobsahující slovesa ohraničené z obou stran závorkami jsou v našem pojetí klauzemi a přitom jsou velmi snadno identifikovatelné).

Fakt, že počet klauzí zhruba odpovídá počtu určitých sloves, může velmi pomoci při zjišťování, zda určitá koordinace je větná nebo členská. Vezměme například větu cmpr9406-002-p4s1 z PDT *Koncem dubna 1993 byla přijata novela o bankrotech, která měla – podle názoru nejen předkladatelů, ale i široké odborné i laické veřejnosti – vyvolat dominový efekt krachu podniků, které si vzájemně neplatí.* Tato věta v sobě obsahuje několik zajímavých jevů. Obsahuje tři určitá slovesa, ale sedm segmentů (kombinace, *ale i* je považována za jednu hranici mezi segmenty). Pro snazší orientaci si je očíslujeme:

1. *Koncem dubna 1993 byla přijata novela o bankrotech*
2. *která měla*
3. *podle názoru nejen předkladatelů*
4. *široké odborné*
5. *laické veřejnosti*
6. *vyvolat dominový efekt krachu podniků*
7. *které si vzájemně neplatí*

Už počet segmentů jasně napovídá, že jich několik zřejmě tvoří jednu klauzi. Protože v souvětí je několik souřadicích spojek, je vysoce pravděpodobné, že bude obsahovat členskou koordinaci. Dvojice spojovníků (–) také pomáhá určit rozsah této koordinace: vzhledem k tomu, že mezi nimi není jediné určité sloveso, ale zato několik souřadicích spojek, dá se celkem spolehlivě určit, že celý úsek mezi spojovníky patří ke stejné klauzi.

Pokud tedy spojíme segmenty 3, 4 a 5 do jednoho celku, zůstane nám v souvětí pět segmentů na tři určitá slovesa. Zajímavé je, že nyní v každém segmentu nalezneme sloveso – tři určité slovesné tvary a jeden infinitiv v segmentu 6. Tento infinitiv může těžko stát sám o sobě, a vzhledem k tomu, že sloveso *měla* v segmentu 2 se jako modální sloveso pojí s infinitivem, je možné segmenty 2 a 6 spojit do jediné klauze.

Spojíme-li segmenty 2 a 6 do jednoho celku a segmenty 3, 4 a 5 do celku druhého, zůstanou nám v souvětí čtyři segmenty – kandidáti na klauze. Z hlediska syntaktické analýzy je už potom celkem lhostejné, zda budeme s vloženou skupinou 3, 4 a 5 zacházet jako se vsuvkou a analyzovat ji zvlášť, či zda ji budeme považovat za nedílnou součást klauze 2, 3, 4, 5 a 6. V tomto případě nám tedy velký rozdíl počtu segmentů a určitých sloves pomáhá určit moment, kdy se spojováním můžeme přestat – jakmile se počet složených segmentů přiblíží k počtu určitých sloves.

Zkusme tento postup uplatnit ještě na jednom příkladu. Vezměme větu cmpr9407-005-p10s1 z PDT:

Abyste mohla tento nárok s úspěchem ve stanovené lhůtě uplatnit, bylo by třeba, abyste byla nejenom československou, a později českou občankou, ale měla i trvalý pobyt na území ČR. a rozdělme ji na segmenty.

1. *Abyste mohla tento nárok s úspěchem ve stanovené lhůtě uplatnit*
2. *bylo by třeba*
3. *abyste byla nejenom československou*
4. *později českou občankou*
5. *měla*
6. *trvalý pobyt na území ČR*

V tomto souvětí nalezneme šest segmentů a čtyři určitá slovesa. V případě segmentů 3 a 4 na základě morfologické informace (shodné pády adjektiva a jmenné skupiny) relativně snadno identifikujeme členskou koordinaci *československou a později českou občankou* a tyto segmenty spojíme. Při hledání dalších kandidátů na spojení objevíme souřadicí spojku *i* spojující segmenty 5 a 6. Tato spojka není standardní slučovací spojkou, má zde spíše význam stupňovací. To je ovšem informace, kterou nemáme při analýze souvětí k dispozici. Proto si spíše všimneme toho, že tato spojka spojuje slovesný tvar na levé straně s nominální skupinou na straně pravé, kde je segment obsahující slovesný tvar uvozen spojkou *ale*; můžeme proto s vysokou pravděpodobností určit, že jde skutečně o stupňovací souřadné spojení vyjádřené dvojicí *ale i* (v distantním postavení). Kdyby v segmentu 6 byl určitý slovesný tvar, jednalo by se naopak jednoznačně o koordinaci klauzí. Z uvedeného příkladu tedy vyplývá, že u řady souřadicích spojek je nutné vzít v úvahu nejen jejich morfologickou značku, ale i lexikální hodnotu.

Toto tvrzení podporuje i další příklad, věta cmpr9406-002-p18s1A z PDT: *Je však těžké říci, zda se třeba už za čtrnáct dnů či za tři čtyři měsíce dovíme senzační zprávu o tom, že ten či jiný velký podnik „zásluhou“ příslušné banky zbankrotoval.*. Jednotlivé segmenty vypadají takto:

1. *Je*
2. *těžké říci*
3. *zda se třeba už za čtrnáct dnů*
4. *za tři čtyři měsíce dovíme senzační zprávu o tom*
5. *že ten*
6. *jiný velký podnik*
7. *zásluhou*
8. *příslušné banky zbankrotoval*

Tři určitá slovesa nalezneme v segmentech 1, 4 a 8. Spojka *či* spojuje segmenty 5 a 6 do jednoho celku, protože mezi podřadicí spojkou *že* a spojkou *či* není žádné sloveso, které by se dalo koordinovat se slovesem napravo od této spojky, tudíž se musí jednat o koordinaci členskou. Jediné sloveso napravo od segmentu 5 také napovídá, že uvozovky mezi segmenty 6, 7 a 8 mají pouze kosmetickou úlohu a všechny segmenty 5, 6, 7 a 8 lze spojit do jednoho celku (navíc v případě, že by uvozovky signalizovaly přímou řeč, a tedy jinou klauzi, by se uvozovky kombinovaly s interpunkcí). Bude se tedy jednat o jedinou klauzi (zleva je oddělena od zbytku souvětí podřadicí spojkou a napravo od ní už je pouze tečka za souvětím). Tímto se dostaneme k počtu 5 zbývajících segmentů. Jasným kandidátem na spojení je potom segment 2, který nemůže stát osamoceně a musí být spojen se segmentem 1. Slovo *však* zde tedy má roli příslovce, nikoli souřadicí spojky. Segmenty 3 a 4 můžeme spojit na základě faktu, že sloveso *dovědět se* je reflexivum tantum a proto reflexivní částice *se* v segmentu 3 patří do stejné klauze. Po tomto spojení již počet segmentů odpovídá počtu určitých sloves a spojování je dokončeno.

3.3 Strukturní analýza

Dostupná data umožňují též zajímavá pozorování, která se týkají struktury segmentů a úrovně jejich zanoření. Ukazuje se například, že prototypicky může být **zanoření segmentu nejvýš o jednu úroveň hlubší oproti předchozímu segmentu**. Toto pravidlo bylo v datech porušeno pouze u 12 vět, z nichž v 9 případech šlo o jevy příbuzné jevu popsanému v [10], kdy se v jednom segmentu nacházejí dva ‚příznaky podřízenosti', např. podřadicí spojky (jako je dvojice spojek *že když* ve větě ln95041-042-p7s6 *Zjistili jsme, že když žijeme v Čechách, měli bychom hrát muziku pro českého posluchače.*). Další 3 případy se týkaly závislých klauzí v přímé řeči (např. věta ln95040-032-p2s14 *Zdeněk Müller, trenér Kladna: „Jestli mám někoho pochválit, pak útočníka Tona a Chlada v brance.“*). Protože oba tyto případy lze s vysokou správností určit na základě morfologické analýzy, dává nám toto pozorování důležitou informaci o přípustném tvaru segmentačního schématu věty.

Podobně lze velmi bezpečně určit **úroveň prvního segmentu** – prototypicky je prvni segment, který neobsahuje příznak podřízenosti, na základní úrovni. První segment s úrovní 1 se v analyzovanách datech vyskytl v 207 větách; v těchto případech rozbor ukázal, že první segment měl následující charakteristiky:

příznak podřízenosti	105
fragment v závorkách	15
přímá řeč (s párovými uvozovkami)	33
přímá řeč (jen koncové uvozovky)	26
polopřímá řeč	11

První segment na úrovni 2 se v analyzovaných datech vyskytl pouze ve 4 případech, vždy šlo o závislou

klauzi (s příznakem podřízenosti) v přímé řeči. Na nižších úrovních se první segment nevyskytl vůbec, přestože takový případ nelze zcela vyloučit (první segment na úrovni 3 ve větě *„A že když se bavím s osmnáctiletými kluky, připadám si jako instituce, přiznávám se,“ smál se trenér.*).

4 Návrh postupu

Z výše uvedených příkladů vyplývá několik pozorování, která je nutno zohlednit při vytváření algoritmu spojování.

Ukazuje se, že lingvistické analýze je vhodné předřadit **identifikaci vět s ne zcela standardní větnou strukturou** (seznamy, adresy, výčty apod.).

Při vlastní lingvistické analýze se jako nejdůležitější jeví fakt, že naprosto klíčovou roli hraje schopnost určit, zda se pro konkrétní slučovací spojku v daném kontextu jedná o **koordinaci větnou nebo členskou**. K tomuto rozhodnutí je nutné vzít v úvahu zejména přítomnost či naopak nepřítomnost určitých sloves v koordinovaných segmentech, shodu mezi koordinovanými segmenty apod.

Velmi důležitá je také **lexikální hodnota** samotné **spojky**, už v procesu anotace vyšlo najevo, že u některých spojek (*však*, *proto*, *či* apod.) nestačí brát v úvahu morfologickou značku a je nutné s každým tímto slovem zacházet individuálně.

Další důležitá skupina pravidel popisuje **spojování segmentů** obsahujících určité slovesné tvary se segmenty obsahujícími určitá slova doplňující tyto slovesné tvary. Jako příklad lze uvést odloučené reflexivní částice, které můžeme spojovat s reflexivy tantum; dále jde zejména o slovesa s valenčními doplněními se specifickou formou, např. infinitivem (segmenty obsahující určité sloveso, v jehož valenčním rámci se vyskytuje infinitivní slovesné doplnění, a segment obsahující infinitiv tvoří s vysokou pravděpodobností klauzi). Slučovací pravidla budou mít různé priority, nejvyšší prioritu získají pravidla popisující členské koordinace.

Podstatnou roli hrají též **strukturní omezení**, která je nutno uplatňovat na tvar segmentačního schématu, tedy na možnou strukturu segmentů a na úroveň jejich zanoření; od struktury segmentů se pak odvíjí možné struktury klauzí.

Pokud již nepůjde uplatnit slučovací pravidla, uplatníme **speciální heuristiky**, vytvořené na základě specifických jevů identifikovaných v sekci 3. Ty budou zejména připojovat segmenty bez určitých sloves k segmentům tato slovesa obsahujícím, řešit případy klauzí uzavřených v závorkách apod.

5 Závěr a výhledy do budoucna

Jak již bylo uvedeno, tématem tohoto článku je popis probíhající analýzy dat, získaných transformací PDT do podoby zohledňující vzájemné postavení klauzí v českých souvětích. Tento soubor dat je dostatečně rozsáhlý na to, aby bylo možné vytvořit spolehlivá pravidla pro spojování segmentů do klauzí. V současné době máme též k dispozici nástroj umožňující prohledávání dat. Zautomatizování celého procesu formulace pravidel pro spojování segmentů do klauzí je prvořadým tématem pro další výzkum. Z dosud provedené analýzy vyplývá, že pro tento úkol je nejpodstatnější rozlišení mezi členskou a větnou koordinací u souřadicích spojek; převážná část souvětí s více segmenty v dosud analyzovaných datech obsahuje jednu nebo více členských koordinací. Důležité je také vytipování speciálních případů a jejich vyřešení ještě před startem hlavního algoritmu.

Literatura

1. V. Kuboň: *A robust parser for Czech.* PhD Thesis, Charles University in Prague, Prague, 2001.
2. D. Zeman: *Parsing with a statistical dependency model.* PhD Thesis, Charles University in Prague, Prague, 2004.
3. V. Kuboň, M. Lopatková, M. Plátek, P. Pognan: *A linguistically-based segmentation of complex sentences.* In: Wilson, D.C., Sutcliffe, G.C., (eds), Proceedings of FLAIRS 2007 Conference, Menlo Park, AAAI Press, 2007, 368–374.
4. M. Lopatková, T. Holan: *Vztahy mezi segmenty – segmentační schémata českých vět.* In: Vojtáš, P., (ed.), Proceedings of ITAT 2008, Košice, University of P.J. Šafárik, 2007, 15–22.
5. J. Hajič, E. Hajičová, J. Panevová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová: *Prague Dependency Treebank 2.0.* Linguistic Data Consortium, Philadelphia, PA, USA, 2006.
6. O. Krůza, V. Kuboň: *Obtaining hidden relations from a syntactically annotated corpus – from word relationships to clause relationships.* In: Proceedings of Flairs 2009, 2009.
7. M. Lopatková, N. Klyueva, P. Homola: *Annotation of sentence structure; capturing the relationship among clauses in czech sentences.* In: Proceedings of the Third Linguistic Annotation Workshop, Suntec, Singapore, Association for Computational Linguistics, August 2009, 74–81.
8. D. Marinčič, T. Šef, M. Gams: *Parsing with clause and intraclausal coordination detection.* Computing and Informatics (in press).
9. V. Šmilauer: *Novočeská skladba.* SPN, Praha, 1966.
10. Š. Lešnerová-Zikánová, K. Oliva: *Česká vztažná souvětí s nestandardní strukturou.* Slovo a slovesnost, **64**,(4), 2004, 241–252.