Systém pro syntaktické značkování velkých textových korpusů[1]

Tomáš Jelínek
Ústav teoretické a komputační lingvistiky
Filozofické fakulty Karlovy univerzity

Abstract
Syntactic annotation of corpora is a useful corpus exploitation tool, presently limited to small corpora. The purpose of our project, entitled *Syntactic Annotation of Czech Corpora,* is to provide a large syntactically annotated corpus with customizable representation. In this paper, I present the methods used for automatic syntactic annotation: a stochastic parser followed by a rule-based automatic correction module. The usefulness of such an annotation is demonstrated on frequency tables of syntactic functions of Czech nouns with respect to their case and preposition, which were extracted from the SYN2005 corpus, with additional syntactic annotation.

0. Úvod

Pro jazykový výzkum založený na rozsáhlých korpusech je mnohdy neefektivní využívat pouze „čistý text“, tvary slov bez jakékoli interpretace. Práci často velmi usnadní další informace, které byly do původního textu vneseny, aniž by uživateli bránily využívat text neinterpretovaný. Od publikace prvního velkého korpusu současné češtiny SYN2000 jsou všechny velké korpusy psané češtiny v rámci projektu *Český národní korpus* (ČNK) opatřeny slovnědruhovým a morfologickým značkováním a lemmatizací. Tento nástroj, jehož kvalita se postupně zlepšuje, využívá stále více uživatelů (přibližně 30 % dotazů do korpusů řady SYN v roce 2009 se zakládalo na lemmatech nebo morfologických značkách). Systém tohoto morfologického značkování je podrobně popsán v jiném příspěvku v této publikaci (*Systém jazykového značkování korpusů současné psané češtiny*).

V tomto příspěvku chci představit návrh povrchového syntaktického značkování (označení syntaktických funkcí větných členů a závislostní struktury věty), jež se dosud v korpusech ČNK nepoužívá, ale které náš tým připravuje v projektu *Syntaktická anotace českých korpusů* (viz pozn. 1). Chci také ukázat příklad možného využití takového značkování, konkrétně statistiku syntaktických funkcí u jednotlivých pádů substantiv (s předložkami a bez nich).

1. Syntaktické značkování textu

Syntaktické značkování závisí na rozdíl od morfologického značkování mnohem více na tom, na jakém teoretickém východisku je založeno, přesto se na většině základních charakteristik slov a jejich zapojení do větné struktury lze shodnout, rozdíly jsou spíše ve vlastnostech, na něž je kladen větší či menší důraz, či ve způsobu zobrazení. V projektu *Syntaktická anotace českých korpusů* usilujeme o to, aby si uživatel sám mohl nastavit, jaké zobrazení větných struktur a jaké funkce chce používat. Tímto zobrazením se podrobněji zabývá příspěvek *Syntakticky anotovaný korpus českých textů* v této publikaci. V této kapitole představím automatický systém, který bude použit v procesu značkování textů.

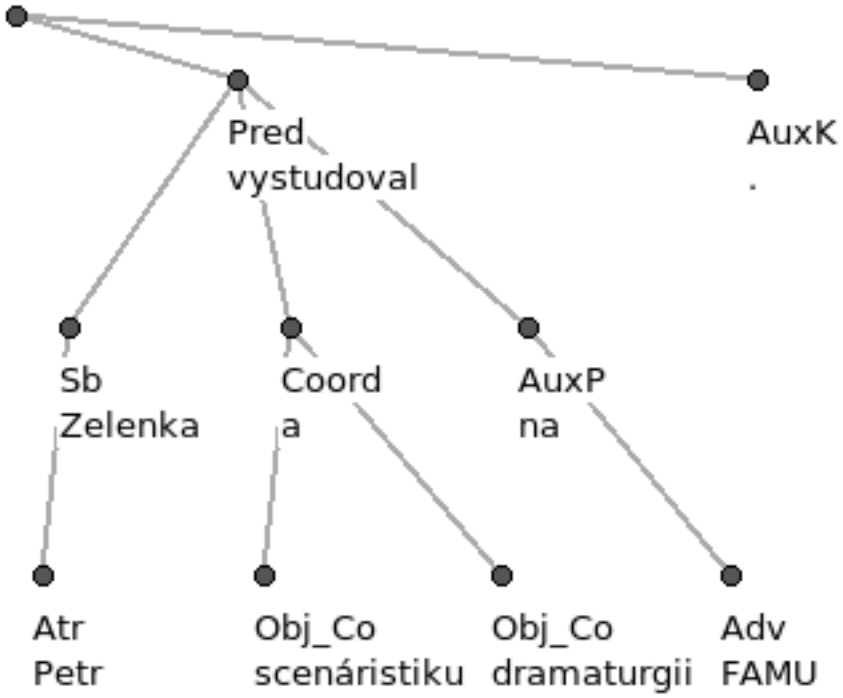
1.1. Pracovní formát pro automatické syntaktické značkování: PDT

Pro vlastní proces syntaktického značkování jsme zvolili formát analytické roviny PDT – Prague Dependency Treebank, Pražský závislostní korpus (Hajič et al. 2006), protože tento korpus představuje

1 Tento příspěvek byl realizován s podporou grantu GAČR P406/10/0434.

v současné době nejlépe anotovaný závislostní korpus s dostatečným rozsahem (cca 1,5 miliónu slovních tvarů). Některé automatické metody syntaktického značkování, o nichž budu mluvit dále, totiž potřebují jako východisko kvalitní manuálně označkovaný korpus. Formát PDT lze snadno převést do formátu, který budeme používat v projektu *Syntaktická anotace českých korpusů*. Analytická rovina PDT používá závislostní struktury a víceméně „tradiční" syntaktické funkce (Šmilauer 1966): podmět (Obj), přísudek (Pred), jmenný přísudek (Pnom), předmět (Obj), příslovečné určení (Adv), přívlastek (Atr), doplněk (Atv) spolu s dalšími funkcemi pro neplnovýznamová slova předložky (AuxP), podřadicí spojky (AuxC), koordinační spojky (Coord), pomocné sloveso (AuxV) aj. Následující obrázek ukazuje grafické znázornění závislostní struktury používané v PDT.

Grafické znázornění závislostního stromu ve formátu PDT



2. Systém automatického syntaktického značkování

Korpusy ČNK, jež chceme označkovat syntakticky, jsou příliš rozsáhlé (celkem přes miliardu slov) na to, aby byly označkovány manuálně, a tak je nutné se spolehnout na automatické metody. Protože je však správné určení závislostních struktur ve větě mnohem složitější úkol než lemmatizace nebo morfologické značkování, jsou výsledky automatického syntaktického značkování výrazně méně spolehlivé. V současné době obsahují cca 15 % chyb a ani při aplikaci sofistikovanějších metod značkování se v dohledné době nedosáhne méně než desetiprocentní chybovosti. Značkování tak nelze použít jako jediný nekorigovaný základ pro jazykový výzkum, ale jen jako pomůcku k vyhledání příkladů a protipříkladů pro určitý jev; hodí se pro předběžnou práci spíše než pro definitivní závěry ohledně zkoumaného jevu. V části 5 chci ukázat, že jsou oblasti, v nichž lze syntaktické značkování i se současným vysokým procentem chyb s úspěchem použít. Lze si také představit řadu výzkumných úkolů, jež by bez víceméně správně rozpoznané syntaktické struktury nebyly vůbec uskutečnitelné.

2.1 Možné postupy automatického značkování

Stejně jako pro morfologické značkování (viz kapitola *Systém jazykového značkování korpusů současné psané češtiny* v této publikaci) lze pro automatické syntaktické značkování použít postupy stochastické (statistické) a postupy založené na lingvistických pravidlech. Na rozdíl od morfologického značkování je však vytvoření čistě lingvisticky založeného značkovacího programu značně obtížné: existující pravidlový modul používaný pro morfologickou disambiguaci je založen primárně negativně, odstraňuje takové interpretace slovních tvarů, jež dohromady tvoří negramatické konstrukce. Ovšem negativní postup v případě závislostní struktury by vyžadoval nejprve vytvořit všechny potenciální struktury ve větě, kterých by i u relativně jednoduché věty mohlo být tolik, že by jejich zpracování přesáhlo možnosti současných výpočetních systémů, a i po zpracování mnoha syntaktickými pravidly by mohly pro každou jen trochu delší větu zbývat stovky různých interpretací. Pro úspěšnou pozitivní konstrukci závislostních struktur by zase muselo být sepsáno obrovské množství pravidel pro různé dílčí struktury (nebo by musel být zpracován mimořádně komplexní slovník), což by vyžadovalo mnohaletou práci několika lingvistů, systém by i tak byl pravděpodobně málo flexibilní. Proto jsme se rozhodli nejprve použít nejlepší současný stochastický závislostní parser (syntaktický analyzátor) a na základě rozboru jeho výsledků vytvořit systém lingvistických pravidel pro snížení celkového počtu chyb.

3. Statistický závislostní parser a analýza jeho chyb

Statistický závislostní parser pro analýzu češtiny (ve formátu PDT), který má v současné době nejlepší výsledky, je MST (maximum spanning tree) Parser Ryana McDonalda (McDonald et al. 2005) se specifickým nastavením pro češtinu (Novák et al. 2007). Podrobný popis principů tohoto stochastického programu by zabral příliš prostoru a byl by pro většinu čtenářů příliš technický, spokojíme se tu vysvětlením, že tento parser funguje na podobném principu jako stochastické taggery, o nichž je řeč v příspěvku *Systém jazykového značkování korpusů současné psané češtiny*. Parser vychází z ručně označkovaných „trénovacích“ dat (korpus PDT). Je nastaven, aby v těchto datech sledoval určité typy proměnných a vztahů (například slovní druh větného členu, pád, lemma, syntaktickou funkci, pořadí ve větě atd.). Výskyt různých kombinací těchto proměnných v „trénovacích“ datech si zaznamenává a získává tak rozsáhlou databázi pravděpodobností výskytu různých struktur. Potom je parser spuštěn na nový text označkovaný pouze morfologickými značkami a lemmaty, kde z možných struktur volí tu strukturu, pro niž je celkový součin pravděpodobností jednotlivých dílčích kombinací nejvyšší.

3.1 Chyby parseru a jejich příčiny

Chyby, kterých se parser dopouští, mají dva hlavní důvody: nejvíce chyb je nejspíš způsobeno tím, že parser při svém posuzování trénovacích dat nemůže nikdy obsáhnout celou větu, ale přiřazuje pravděpodobnosti jen dílčím vztahům mezi větnými členy a jejich funkcemi. Při posuzování neznámého textu pak může jako nejpravděpodobnější vyjít struktura, která je negramatická (kupříkladu dva nekoordinované subjekty v jedné jednoduché větě); druhým důvodem je relativně malý rozsah trénovacích dat: některé struktury a mnoho lemmat se v trénovacích datech neobjeví a parser pak „neví“, jak je v neznámém textu značkovat. V některých chybách lze snadno rozpoznat správné struktury z trénovacích dat. V následující větě z korpusu SYN2010 označkovaného MST Parserem je například chybně určena funkce (subjekt) u slova *ostrovy* v předložkové frázi (slovo *ostrovy* nemůže mít funkci *Sb*, mj. z toho důvodu, že předložková fráze nevyjadřuje neurčitou kvantifikaci a sloveso není v neutru singuláru):

*Na **ostrovy/Sb** se **přeplavili** z Jutska, Angelnu a z Dolního Saska.*

Tato chyba mohla být způsobena podobností s následující strukturou z trénovacích dat (PDT), která je sice správně označkována, svým výskytem však parser „mate“ (v soupisu možných dílčích struktur se může objevit kombinace subjektu v předložkové frázi se slovesem v množném čísle):

*Na **třicet/Sb** ázerbájdžánských vojáků a dva **Arménci/Sb byli zabiti** v sobotu během bojů na severovýchodě Náhorního Karabachu...*

Zde se předložková fráze *na třicet* s funkcí *Sb* vyjadřující neurčitou kvantifikaci objevuje ve větě se slovesem v množném čísle. Sloveso se totiž shoduje s druhým z koordinovaných subjektů, nemusí tedy být v neutru singuláru.

Část chyb má příčinu už v chybném morfologickém značkování. Byl-li substantivu chybně určen pád, pravděpodobně bude chybně určena i jeho syntaktická funkce. Tak je tomu například v následujícím příkladu, kde je chybně určená syntaktická funkce u slova *čas* způsobena chybně určeným pádem (akuzativ místo nominativu):

*Ale na přemýšlení jí nezbýval čas/**Obj/akuzativ**.*

3.2 Automatická identifikace chyb parseru

Manuální rozbor několika tisíc vět z korpusu SYN2005 syntakticky označkovaného zmíněným parserem ukázal, že některé typy chyb se systematicky opakují a je možné je vyhledávat automaticky. Vytvořil jsem program, který některé typické chyby automaticky vyhledával v celém korpusu, abych mohl určit, které chyby jsou frekventované a je nutné je prioritně opravit a které typy chyb lze prozatím zanedbat. Vyhledávací algoritmus byl poměrně hrubý, některé struktury označené jako chybné byly v rámci formalismu PDT ve skutečnosti v pořádku, ale základní přehled o chybách parseru jsem tak získal. Automaticky jsem byl schopen identifikovat cca 4 % tokenů v korpusu jako chybně označkovaných nebo s chybnou závislostí, což odpovídá přibližně 25 % předpokládaných chyb. V následující tabulce jsou uvedeny poměry výskytů nejčastějších chyb mezi identifikovanými chybami.

Automaticky vyhledané chyby v korpusu SYN2005 označkovaném MST Parserem

Typ chyby	
Sloveso označené jako hlavní sloveso ve větě závislé na jiném větném členu	25,3 %
Nekompatibilní syntaktické funkce závislé na jedné koord. spojce (Adv + Atr, Obj + Atr...)	13,2 %
Dva nekoordinované subjekty závislé na jednom slovese	11,5 %
Neodůvodněné označení zájmena „se“ jako částice u slovesa reflex. tantum	10,7 %
Substantivum v nominativu označené jako předmět (Obj)	5,8 %
Chybná syntaktická funkce v předložkové frázi závislé na slovese	5,6 %
Substantivum v jiném pádě než nom. označené jako subjekt (Sb; mimo gen. záporový aj.)	4,2 %
Neshodný přívlastek (Atr) závislý na slovese (většinou v předložkových frázích)	3,1 %
Syntaktická substantiva v nekompatibilních pádech závislá na jedné koord. spojce	1,8 %
Substantivum v akuzativu označené jako předmět (Obj) závislé na netranzitivním slovese	1,7 %
Substantivum v akuzativu označené jako předmět (Obj) závislé na reflexivním slovese, vyjma „učit“	1,4 %
Substantivum v akuzativu označené jako předmět závislé na modálním či fázovém slovese s infinitivním předmětem	1,0 %

Podrobný rozbor frekventovaných chyb ukázal, že některé typy chyb lze s velkou mírou jistoty opravovat na základě lingvistických pravidel. Vždy je však mnohem obtížnější určit, jak správně opravit identifikovanou chybnou strukturu, než chybu pouze nalézt. Lingvistické opravy korpusu označkovaného stochastickým parserem představuje další část této kapitoly.

4. Opravy výsledků stochastického parseru založené na lingvistických pravidlech

Při rozboru chyb syntaktického parseru jsem zjistil, že některé z nich se často opakují a je přitom možné najít algoritmus, který může chyby víceméně spolehlivě opravit. Parser o větě „neuvažuje" jako lingvista, nemá k dispozici údaje o hranicích vět v souvětí, o valencích sloves atd., pouze relativní pravděpodobnosti jednotlivých struktur. Některé údaje (jako valenci) by bylo možné vložit přímo do parseru, ale jeho „rozhodování" by se tak stalo ještě komplexnějším a ne nutně úspěšnějším.

Bylo proto efektivnější vytvořit samostatný programový modul, který na základě lingvistických poznatků chyby identifikuje a známé typy chyb pomocí pečlivě sestaveného algoritmu opravuje. Tento modul je dosud v přípravné fázi, z plánovaných několika desítek pravidel bylo implementováno pouze deset, některé z nich je ještě nutné zpřesnit nebo doplnit. Pro pochopení charakteru a potřebného rozsahu práce na opravném modulu uvedu tři příklady typických chyb MST Parseru a jejich řešení v modulu pro automatickou opravu založenou na lingvistických pravidlech: dva nekoordinované subjekty závislé na jednom slovese; akuzativní předmět nebo předmět v předložkové frázi závislý na modálním či fázovém slovese, které má také infinitivní předmět; chybné určení syntaktické funkce substantiva v předložkové frázi závislé na slovese. Na závěr této části pak ukážu, jak byla dosud implementovaná pravidla úspěšná při opravách korpusových textů.

4.1 Dva subjekty závislé na jednom slovese

Jednou z častých chyb parseru je struktura, v níž jsou na jednom slovese závislé dva nekoordinované subjekty. Při současném (dosud nejúspěšnějším) nastavení parseru totiž program nemůže ověřit, zda již slovesu jeden subjekt nepřiřadil. Oprava této snadno identifikovatelné chyby je však relativně složitá, podrobnou analýzou věty je třeba určit, které z několika možných řešení je nejlepší. Poměrně často je chyba způsobena chybným morfologickým značkováním: i to může modul opravit. V současné verzi opravného modulu je sedm možností, jak nalezenou chybu odstranit (ne vždy úspěšně a ne každou takovou chybu je modul v současnosti schopen opravit, i když ji identifikuje). Jednotlivé podtypy chyby a jejich dílčí řešení představuji v pořadí od nejčastěji používaného.

4.1.1 Dva subjekty závislé na tranzitivním slovese, oba uvnitř klauze

Nejčastější podtyp chyby má příčinu v chybném morfologickém značkování, jedno substantivum má chybně určený pád: nominativ místo akuzativu. Parser na základě chybných dat určil chybně funkci (subjekt):

> ***Prosby/NNFP1/Sb*** *přednášejí i ostatní* ***členové/Sb*** *domácnosti.*

V takové konfiguraci ověří opravný modul podmínky pro opravu:
a) alespoň jedno ze substantiv je pádově homonymní (akuzativ – nominativ)
b) na slovese není závislý jiný předmět ve čtvrtém pádě ani reflexivum „se"
c) substantivum není vyjádřením míry nebo času.

Za těchto podmínek změní modul značku i funkci homonymního substantiva:

Prosby/NNFP4/Obj *přednášejí i ostatní členové domácnosti.*

Kdyby podmínku a) splňovala obě substantiva a obě se přitom mohla shodovat se slovesem v rodě i čísle, opraví modul **druhé** substantivum (vyšší pravděpodobnost úspěchu), ve větě

Zločiny/Sb *otců pronásledují* ***děti/NNFP1/Sb****.*

bude tedy opravena značka i funkce u substantiva *děti*.

Zločiny/Sb *otců pronásledují* ***děti/NNFP4/Obj****.*

Kdyby naopak struktura nesplňovala podmínku c), případně ani c), ani b), lze i tak opravit značku a funkci u syntaktického substantiva vyjadřujícího čas či míru, ovšem ne na předmět, nýbrž na příslovečné určení. Větu

Debaty, ***které/Sb*** *se už* ***týdny/NNIP1/Sb*** *vedou kolem projektu nové Národní knihovny, ...*

opraví modul takto

Debaty, ***které/Sb*** *se už* ***týdny/NNIP4/Adv*** *vedou kolem projektu nové Národní knihovny, ...*

4.1.2 Dva subjekty závislé na slovese *být*, oba uvnitř klauze

Jsou-li na jednom slovese *být/bývat* závislé dva subjekty (v nominativu) a na slovese přitom není závislý jmenný přísudek, lze předpokládat, že jeden ze subjektů by správně měl být označkován jako jmenný přísudek.

Je-li jedním ze subjektů zájmeno *to*, bude jako jmenný přísudek označen druhý subjekt bez ohledu na shodu se slovesem (první je původní věta, druhá opravená):

Podle židovského letopočtu je ***to/Sb rok/Sb*** *2448.*
Podle židovského letopočtu je ***to/Sb rok/Pnom*** *2448.*

Jinak bude opravena funkce u toho syntaktického substantiva, které se neshoduje se slovesem v rodě a čísle:

Rákoska/Sb *bylo krátké tlusté* ***pravítko/Sb*** *z červeného dřeva...*
Rákoska/Pnom *bylo krátké tlusté* ***pravítko/Sb*** *z červeného dřeva...*

Shodují-li se obě syntaktická substantiva, bude opravena funkce u druhého substantiva v pořadí (opět nejde o opravu jistě správnou, ale pravděpodobnější)

Služba/Sb *lidem je* ***povinnost/Sb****.*
Služba/Sb *lidem je* ***povinnost/Pnom****.*

4.1.3 Dva subjekty těsně vedle sebe, pojmenování osoby

Tvoří-li dva nekoordinované subjekty stojící ve větě těsně vedle sebe dvojici, jež se shoduje v rodě, čísle i pádě, a oba subjekty jsou substantiva, a to buď vlastní jména, nebo substantiva ze seznamu generických označení osob (jako *pan, paní, soudruh, doktorka, prezident* aj.), musí být v rámci formalismu PDT první z nich označen jako shodný přívlastek závislý na druhém jménu. V následující větě tvoří zvýrazněná jména dvojici, z níž (ve zvoleném zápisu) první je přívlastek, druhé subjekt. Chybnou větu je tedy třeba opravit:

*Při úctě k jiným náboženstvím věříme, že je **Ježíš/Sb Kristus/Sb** spasitelem všech lidí.*

Systém změní nejen označení syntaktické funkce, ale i závislost (první substantivum už není závislé na slovese, ale na druhém substantivu).

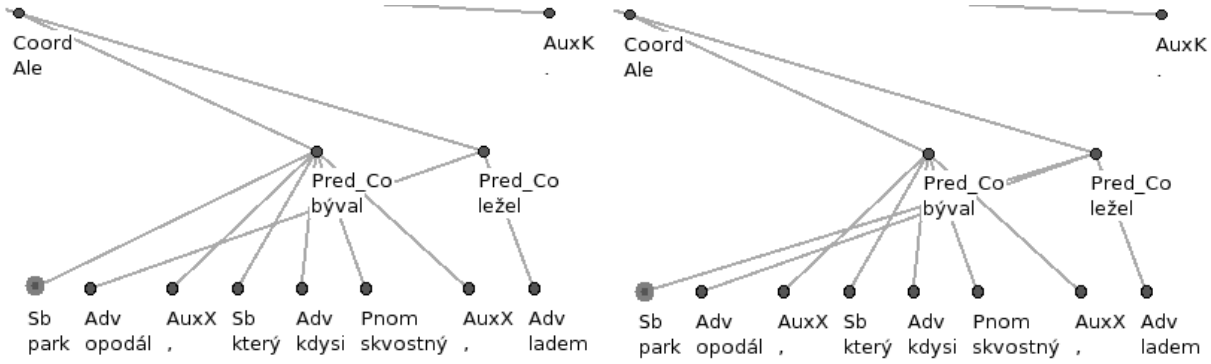
*Při úctě k jiným náboženstvím věříme, že je **Ježíš/Atr Kristus/Sb** spasitelem všech lidí.*

4.1.4 Dva subjekty závislé na jednom slovese, jeden přes jednu hranici klauze

Je-li větný člen závislý na slově, jež patří do sousední klauze, je závislostní struktura pravděpodobně chybná (ve formátu PDT kromě spojek a sloves, které v souvětích reprezentují celé vedlejší věty). Závislost může přecházet přes hranice klauzí jen tehdy, když mezi větnými členy stojí vnořená vedlejší věta, jež by sama měla být ukončena další hranicí klauze. Pokud tedy vazba jednoho ze subjektů v identifikované chybě přechází přes právě jednu hranici klauzí a tento subjekt nemá ve „své“ části věty žádné sloveso, je třeba hledat jiné, vhodnější sloveso, na kterém by měl být subjekt závislý. Takové sloveso nesmí být samo ve vnořené vedlejší větě (tedy musí následovat po čárce, po níž nestojí hypotaktická spojka ani vztažné zájmeno či příslovce), nesmí mít jiný subjekt a musí se se subjektem, pro nějž hledáme řešení, shodovat v čísle i v rodě. Nalezne-li opravný modul takové sloveso, může na něj subjekt „převěsit“ a chybu se dvěma subjekty tak opravit, jako v následujícím příkladu, kde sice nebyla opravena celá chybná struktura, ale chyba se dvěma subjekty byla odstraněna (původní struktura vlevo, opravená vpravo):

*Ale **park/Sb** opodál, **který/Sb** kdysi býval skvostný, ležel ladem.*

Slovo *park*, původně závislé na slovese *býval*, je po opravě závislé na slovese *ležel*:



4.1.5 Dva subjekty závislé na slovese *být*, jeden subjekt přes 2 hranice klauzí

Je-li na jednom slovese *být/bývat* závislý jeden subjekt uvnitř klauze a druhý přes dvě hranice klauzí, přičemž vzdálený subjekt není ve vedlejší větě (uvozené hypotaktickou spojkou či vztažným zájmenem nebo příslovcem), pravděpodobně jsou správně určeny závislosti, ale chybně určena funkce. Opravný modul změní syntaktickou funkci u subjektu, který se neshoduje se slovesem, popř. u druhého subjektu v pořadí:

*Neboť **štěstí/Sb**, které není trvalé, není pravé **štěstí/Sb**.*
*Neboť **štěstí/Sb**, které není trvalé, není pravé **štěstí/Pnom**.*

4.1.6 Dva subjekty závislé na jednom slovese, jeden subjekt přes hranici klauzí

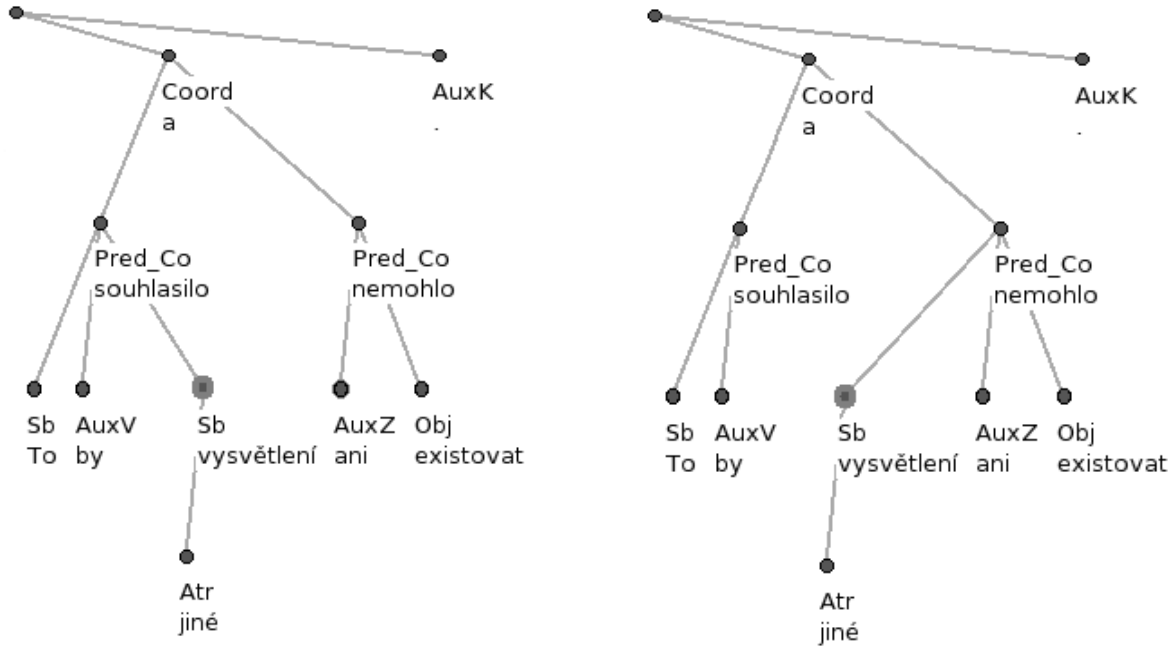
Nejméně často se vyskytují struktury, kde jsou dva subjekty závislé na témž slovese, přičemž závislost

jednoho subjektu jde přes hranici klauze, přestože v rámci jeho vlastní klauze sloveso je. Pokud je to možné (vzhledem ke shodě aj.), převěsí se subjekt na nejbližší sloveso.

Chyba, kdy jde závislost přes jednu hranici klauze, se občas objevuje ve výsledcích parseru. Vytvořili jsme pro ni samostatné pravidlo, které se někdy provede dříve než dílčí pravidlo pro dva subjekty se závislostí přes jednu hranici klauze, možná i proto je frekvence jeho použití relativně nízká.

V příkladu je znázorněna oprava chybné struktury. Chybu parseru možná způsobila přítomnost slova „ani“, které někdy funguje jako hranice klauzí, jindy ne:

To/Sb *by souhlasilo a jiné* ***vysvětlení/Sb*** *ani nemohlo existovat.*



Slovo *vysvětlení*, původně závislé na slovese *souhlasilo*, je po opravě závislé na slovese *existovat*.

4.1.7 Úspěšnost dílčích algoritmů v rámci opravy dvou subjektů

Pro přehled uvádíme tabulku s údaji o frekvenci použití a o úspěšnosti dílčích pravidel při zpracování korpusu. Úspěšnost jsme hodnotili na vzorcích (několik desítek oprav pro každý typ), počet oprav vychází ze statistik pořízených na celém korpusu SYN2010 označkovaném nejprve MST Parserem, potom korigovaném pomocí opravného pravidlového modulu. Počet oprav byl přepočítán na 1 milión tokenů.

Jednotlivé zásahy dílčích opravných pravidel byly hodnoceny jako *pozitivní*, jestliže chybnou strukturu skutečně opravily; jako *neutrální*, byla-li chybná struktura sice identifikována a změněna, ale žádoucí oprava byla jiná, takže struktura zůstala chybná; jako *negativní*, byla-li správně označkovaná struktura identifikována jako chybná a změněna, čímž byla vytvořena nová chyba.

Úspěšnost dílčích algoritmů v rámci pravidla pro opravu dvou subjektů na jednom slovese

podtypy chyby 2 Sb závislé na jednom slovese	oprava	počet oprav	+	0	–
tranz. sloveso, 2 Sb v jedné klauzi	N1→ N4; Sb → Obj	393	53 %	43 %	4 %
sloveso *být*, 2 Sb v jedné klauzi	2. Sb → Pnom	273	85 %	15 %	0 %
sloveso *být*, 1. Sb „to“, 2. Sb substantivum	2. Sb → Pnom	220	95 %	5 %	0 %

podtypy chyby 2 Sb závislé na jednom slovese	oprava	počet oprav	+	0	–
2 Sb vedle sebe, vlastní jména aj.	1. Sb → Atr	48	92 %	8 %	0 %
1. Sb přes 1 hranici klauzí, 2. Sb v klauzi	změna řídícího uzlu	43	77 %	23 %	0 %
1. Sb přes 2 hranice klauzí, 2. Sb v klauzi	1 ze Sb → Pnom	10	96 %	4 %	0 %
1. Sb přes hranici klauzí	změna řídícího uzlu	8	100 %	0 %	0 %
CELKEM		995	84 %	15 %	1 %

počet oprav = počet zásahů v korpusu přepočítaný na 1 milión tokenů
+ = podíl správných oprav, které původní chybu ve struktuře zcela správně odstraní
0 = podíl oprav, které rozpoznají chybnou strukturu a pozmění ji, změna ale nevede k vytvoření správné struktury (chybně zvolená oprava)
– = podíl oprav, které chybně identifikují správnou strukturu jako chybnou a změní ji, takže výsledná struktura je chybná

4.2 Předmět závislý na modálním či fázovém slovese

Poněkud méně častou, ale nezanedbatelnou chybou parseru jsou případy, kdy je syntaktické substantivum v akuzativu (bez předložky) závislé na modálním či fázovém slovese, které má infinitivní předmět, jako v této příkladové větě, kde zájmeno *to* závisí na modálním slovese *mohu* místo na plnovýznamovém slovese *potvrdit*:

> ***Mohu to/Obj*** *jenom potvrdit.*

Chybná je také podobná struktura, kde na modálním či fázovém slovese závisí místo na slovese plnovýznamovém předmět v předložkové frázi (plnovýznamové sloveso je valenční):

> *Pravděpodobně* ***o něm/Obj*** *budete* ***chtít*** *všem povědět.*

V druhém případě je oprava chyby jednoduchá: je-li na modálním či fázovém slovese závislá předložková fráze s funkcí předmětu a zároveň sloveso v infinitivu s odpovídající valencí (předložka a pád), změní opravný modul závislost předložkové fráze z modálního slovesa na valenční sloveso v infinitivu.

V případě s předmětem v bezpředložkovém akuzativu je však náprava chyby složitější. Stejně jako u výše analyzované chyby se dvěma subjekty je častou příčinou této chyby nesprávné morfologické značkování pádově homonymních slov, zvláště homonymie nominativ/akuzativ. Nemá-li modální či fázové sloveso podmět, je-li ve třetí osobě a shoduje-li se potenciální tvar nominativu u zkoumaného předmětu se slovesem v rodě a čísle, je pravděpodobnější, že je třeba opravit morfologickou značku a funkci předmětu, spíše než měnit závislosti ve větě. Má-li plnovýznamové sloveso v infinitivu již jiný předmět v akuzativu, je toto řešení víceméně jediné možné, jako v následující větě:

> *Mohou tyto* ***zkušenosti/NNFP4/Obj*** *člověka naplnit?*
> *Mohou tyto* ***zkušenosti/NNFP1/Sb*** *člověka naplnit?*

4.2.1 Úspěšnost dílčích algoritmů v rámci pravidla pro opravu předmětu závislého na modálním či fázovém slovese

podtypy chyby 2 Sb závislé na jednom slovese	oprava	počet oprav	+	0	–
akuzativní předmět ve větě s tranz. slovesem	změna závislosti	93	100 %	0 %	0 %
předložková fráze ve větě s valenčním slovesem	změna závislosti	30	100 %	0 %	0 %
homonymie ak./nom., změna pádu i funkce	N4 → N1, Obj → Sb	23	76 %	23 %	0 %
CELKEM		146	91 %	9 %	0 %

počet oprav = počet zásahů v korpusu přepočítaný na 1 milión tokenů
+ = podíl správných oprav, které původní chybu ve struktuře zcela správně odstraní
0 = podíl oprav, které rozpoznají chybnou strukturu a pozmění ji, změna ale nevede k vytvoření správné struktury (chybně zvolená oprava)
– = podíl oprav, které chybně identifikují správnou strukturu jako chybnou a změní ji, takže výsledná struktura je chybná

4.3 Chybné určení syntaktické funkce substantiva v předložkové frázi závislé na slovese

Třetím příkladem řešení opakované chyby MST Parseru je chybné určení syntaktické funkce substantiva v předložkové frázi závislé na slovese. Hranice mezi příslovečným určením a předmětem není zvláště v případě předložkových frází zcela přesná, přesto lze v naprosté většině případů rozhodnout, jakou funkci má předložková fráze mít. Mezi předměty se řadí případy, kdy volba slovesa určuje také předložku a pád jeho předmětu, mezi příslovečná určení patří případy časového, místního aj. určení, v nichž je předložka a pád určen významem okolnostního určení.

U často se vyskytujících valenčních sloves i u frekventovaných příslovečných výrazů jsou chyby výjimečné (substantiva v lokálu v předložkové frázi s předložkou *o* závislé na slovese *mluvit* jsou v 99 % označené jako předmět; spojení *o půlnoci* závislé na slovese je v 97 % příslovečné určení), protože se dostatečně často vyskytovaly v trénovacích datech, zatímco u méně častých valenčních sloves a příslovečných určení je procento chyb mnohem vyšší (substantiva v lokálu v předložkové frázi s předložkou *po* závislé na slovese *dychtit* jsou v 17 % označena chybně jako příslovečná určení; *o Letnicích* je ve 36 % označeno chybně jako předmět).

Pro opravu těchto chyb jsem z korpusů SYN2005 a SYN2010 (předběžně syntakticky označkovaných) získal rozsáhlé seznamy valenčních sloves (rozdělené podle předložky a pádu) a seznamy ustálených adverbiálních spojení (také rozdělené podle předložky a pádu, např. *o víkendu*, *o Vánocích*; *za chvíli*, *za týden*). Tyto seznamy slouží jako základ pro opravná pravidla. Oprava může jít oběma směry: při splnění následujících podmínek se mění označení funkce předložkové fráze z *Obj* na *Adv* nebo z *Adv* na *Obj*.

Je-li jako příslovečné určení označeno substantivum v předložkové frázi závislé na slovese, které má odpovídající valenci, a nepatří-li mezi typická příslovečná určení s daným pádem a předložkou, bude syntaktická funkce změněna:

*Přesto pokračovali v **tažení/Adv**.*
*Přesto pokračovali v **tažení/Obj**.*

Předložková fráze v další větě však patří mezi typická příslovečná určení s předložkou *v* a s lokálem, takže tato věta opravena nebude:

*V **roce/Adv** 1952 pokračovala likvidace živností opět pomalým tempem...*

Oprava směřující opačným směrem (změna z *Obj* na *Adv*) je založena na podobném principu: nemá-li řídící sloveso předložkové fráze odpovídající valenci a/nebo patří-li předložková fráze mezi

typicky adverbiální, bude syntaktická funkce změněna:

*Ve **svátek/Obj** Jana Husa chci připomenout jeho smrt jako fakt života.*
*Ve **svátek/Adv** Jana Husa chci připomenout jeho smrt jako fakt života.*

Výsledky testování pravidla ukazují, že je ještě nutné doplnit do opravného modulu změnu morfologické značky místo změny funkce v případě chybně určeného pádu: u pádově homonymního substantiva s pádově homonymní předložkou (např. akuzativ/lokál) závislého na slovese, které má valenci s danou předložkou, ale s jiným pádem, je třeba změnit pád v morfologické značce:

*Většina Bachovy tvorby musela na **vydání/NNNS6/Obj** tiskem čekat víc než století.*

Sloveso čekat sice nemá valenci s předložkou *na* a lokálem, ale má valenci s *na* a akuzativem, je tedy třeba opravit pád substantiva a předložky, ne syntaktickou funkci:

*Většina Bachovy tvorby musela na **vydání/NNNS4/Obj** tiskem čekat víc než století.*

4.3.1 Úspěšnost dílčích algoritmů v rámci pravidla pro opravu chyb v syntaktické funkci předložkové fráze

podtypy chyby v synt. funkci předl. fráze	oprava	počet oprav	+	0	–
Adv. závislé na valenčním slovese	Adv → Obj	1822	96 %	0 %	4 %
Obj. závislý na nevalenčním slovese	Obj → Adv	878	77 %	15 %	8 %
CELKEM		2700	90 %	5 %	5 %

počet oprav = počet zásahů v korpusu přepočítaný na 1 milión tokenů
\+ = podíl správných oprav, které původní chybu ve struktuře zcela správně odstraní
0 = podíl oprav, které rozpoznají chybnou strukturu a pozmění ji, změna ale nevede k vytvoření správné struktury (chybně zvolená oprava)
– = podíl oprav, které chybně identifikují správnou strukturu jako chybnou a změní ji, takže výsledná struktura je chybná

4.4 Ostatní implementovaná opravná pravidla a celková úspěšnost

Uvedené tři příklady opravných pravidel stačí k objasnění principů, na nichž pravidla fungují. Pro úplnost uvádím přehled všech dosud implementovaných pravidel se stručnou charakteristikou a tabulku úspěšnosti celého systému. Dosud se podařilo vytvořit deset pravidel, pro dalších asi dvacet frekventovaných typů chyb existují návrhy opravných algoritmů, které bude třeba otestovat a zavést do opravného systému. Implementovaná pravidla uvádím v pořadí podle počtu provedených oprav v korpusu SYN2010.

4.4.1 Oprava chybné funkce u reflexivního *se*

Pravidlo opravuje chybně určenou syntaktickou funkci *AuxT* u reflexivního zájmena *se*. Funkce *AuxT* označuje u reflexiva součást slovesa, které je reflexivum tantum (*usmát se*). Nepatří-li sloveso mezi reflexiva tantum, musí být funkce opravena. Pravidlo rozlišuje mezi užitím zájmena *se* v reflexivním pasivu a jeho užitím jako reflexivního předmětu (*mýt se*) a přiřazuje zájmenu správnou funkci.

4.4.2 Chybná syntaktická funkce v předložkové frázi

Opravné pravidlo bylo podrobně popsáno v části 4.3. Ověřuje a koriguje syntaktické funkce mezi *Obj* a

Adv v předložkových frázích závislých na slovese.

4.4.3 Dva subjekty závislé na jednom slovese

V části 4.1 byly podrobně popsány dílčí algoritmy tohoto pravidla, které hledá vhodné opravy pro struktury se dvěma nekoordinovanými subjekty závislými na témže slovese.

4.4.4 Větný člen závislý na vzdáleném slovese

Jak již bylo řečeno v předchozím výkladu, stochastický parser v současném nastavení nedokáže pracovat s hranicemi klauzí. Někdy tedy vytváří chybné struktury, v nichž závislost překračuje hranice klauzí, přestože skutečný řídící uzel je v téže klauzi jako uzel závislý. Pravidlo vyhledává a opravuje struktury, v nichž je větný člen závislý na slovese, od něhož je oddělen nejméně jednou hranicí klauzí, přestože je mu nablízku jiné sloveso v určitém tvaru, které je ve stejné klauzi. Změní závislost větného členu ze vzdáleného slovesa na závislost na nejbližším slovese, případně i změní syntaktickou funkci.

4.4.5 Akuzativní předmět závislý na netranzitivním slovese

Pravidlo vyhledává nepředložkové akuzativní předměty závislé na netranzitivních slovesech. Často to jsou časová určení v akuzativu, u těch mění syntaktickou funkci na *Adv*, jindy jde o důsledek chybného značkování a je třeba změnit nejen syntaktickou funkci, ale i morfologickou značku (například změnit *Obj* na *Sb* a akuzativ na nominativ).

4.4.6 Chyby u víceslovných předložkových výrazů

Opravné pravidlo řeší různé problémy se značkováním víceslovných předložkových výrazů. Pro výrazy neodůvodněně označené jako víceslovné předložky hledá správnou syntaktickou funkci; výrazy, které by měly být označené jako víceslovné předložky, ale nebyly, označuje správně.

4.4.7 Předmět závislý na modálním či fázovém slovese

Pravidlo popsané v části 4.2 vyhledává předměty závislé na modálním či fázovém slovese, které má také infinitivní předmět (plnovýznamové sloveso v infinitivu). Mění syntaktickou funkci předmětu, popř. morfologickou značku nebo závislost.

4.4.8 Vedlejší věta označená jako hlavní

Pravidlo vyhledává vedlejší věty (uvozené podřadicí spojkou nebo vztažným zájmenem), které byly parserem označeny jako věty hlavní. Je-li to možné, opraví závislosti tak, aby jako hlavní byla označena věta, kterou nelze považovat za vedlejší, a pro vedlejší větu nalezne správný řídící uzel a správnou syntaktickou funkci.

4.4.9 Chybná závislost v předložkové frázi se zájmenem

V kolokacích typu *předložka – zájmeno – substantivum* se parser dopouští chyb v určení závislosti, protože nedokáže rozlišit mezi zájmeny fungujícími jako syntaktická substantiva (*sebe*, *jemuž*), zájmeny fungujícími obvykle jako syntaktická adjektiva (*svému*, *nějaký*) a zájmeny, která mohou fungovat jako syntaktická substantiva i adjektiva podle kontextu (*všem*, *toho*), mimo jiné proto, že některé morfologické značky mezi těmito typy také nerozlišují (*nic* a *žádný* patří podle morfologických značek ke stejnému druhu zájmen).

4.4.10 Chybný subjekt v předložkové frázi

Tento typ chyby jsem uvedl jako příklad v části 3.1. Předložkové fráze mohou mít funkci subjektu, ale jen za velmi specifických podmínek (kvantifikace, správná předložka, shoda přísudku). Nejsou-li tyto podmínky splněny, ale předložková fráze má přesto funkci subjektu, pravidlo funkci opraví.

4.4.11 Celková úspěšnost implementovaných opravných pravidel

implementovaná pravidla v opravném modulu	oprava	počet oprav	+	0	–
oprava chybné funkce u reflexivního „se“	AuxT → Obj / AuxR	5174	64 %	26 %	10 %
chybná synt. funkce v předložkové frázi	Obj → Adv; Adv → Obj	2700	90 %	5 %	5 %
dva subjekty závislé na jednom slovese	např. Sb → Obj, N1→ N4	995	84 %	15 %	1 %
větný člen závislý na vzdáleném slovese	změna závislosti	794	91 %	9 %	0 %
akuzativní předmět závislý na netranz. slovese	např. Obj → Sb, N4 → N1	290	82 %	18 %	0 %
chyby u víceslovných předložkových výrazů	změna funkce / závislosti	203	98 %	2 %	0 %
předmět závislý na modálním či fáz. slovese	změna závislosti aj.	146	91 %	9 %	0 %
vedlejší věta označená jako hlavní	změna závislosti	111	89 %	7 %	4 %
chybná závislost v PP se zájmenem	změna závislosti	77	79 %	16 %	5 %
chybný subjekt v předložkové frázi	Sb → Obj/Adv	17	100 %	0 %	0 %
CELKEM		10507	85 %	12 %	3 %

počet oprav = počet zásahů v korpusu přepočítaný na 1 milión tokenů
+ = podíl správných oprav, které původní chybu ve struktuře zcela správně odstraní
0 = podíl oprav, které rozpoznají chybnou strukturu a pozmění ji, změna ale nevede k vytvoření správné struktury (chybně zvolená oprava)
– = podíl oprav, které chybně identifikují správnou strukturu jako chybnou a změní ji, takže výsledná struktura je chybná

5. Syntaktické funkce substantiv podle jejich pádu a předložky

V poslední části tohoto příspěvku předvedu příklad možného využití syntaktického značkování korpusu. Již v počátcích české korpusové lingvistiky zjišťovala Marie Těšitelová[2] z korpusu, jaký je v češtině vztah mezi syntaktickými funkcemi a pády. S dobře definovanými syntaktickými funkcemi pomůže taková statistika lépe rozumět jak pádovému systému češtiny, tak větné skladbě aj. Práce kolektivu M. Těšitelové však byla nutně omezena možnostmi výpočetních systému osmdesátých let. Statistika obsahuje například jeden zásadní nedostatek: nerozlišuje mezi předložkovými a nepředložkovými pády; můžeme tak například zjistit, že ve zkoumaném korpusu bylo 11,77 % příslovečných určení vyjádřeno akuzativem, ale nevíme, kdy bylo vyjádřeno bezpředložkovým akuzativem (*Nechci, abys zase hnil **celé léto** u počítače*) a kdy akuzativem s předložkou (*Mnozí z nich **za celý život** nevytáhli paty z těch několika údolí a kopců*).

Nová statistika na řádově větším korpusu a s rozlišením pádů na předložkové a nepředložkové může původní údaje zpřesnit. Ze syntakticky anotovaného korpusu lze samozřejmě získat i mnohem podrobnější statistiky, na ty však v tomto příspěvku není místo.

Ze syntakticky anotovaného korpusu lze takovou statistiku získat snadno. Protože však syntaktické značkování dosud není zcela spolehlivé, bylo nutné manuálně ověřit jednotlivé kombinace funkcí a pádů na vzorcích a upravit data podle těchto vzorků. Tři níže uvedené tabulky tak obsahují údaje o kombinacích pádu a syntaktických funkcí z korpusu SYN2005 označkovaného MST parserem a opraveného pravidlovým syntaktickým modulem; tyto údaje byly dále upraveny po manuální kontrole vzorků.

2 Těšitelová M. a kol.: *Kvantitativní charakteristiky současné češtiny.* Praha, 1985.

5.1 Definice použitých syntaktických funkcí

V tabulkách uvádím jen šest základních (frekventovaných) funkcí substantiv, které se objevují v korpusu, ostatní jsou zanedbány. Pro správné porozumění údajům v tabulkách je nutné upřesnit význam používaných syntaktických funkcí a zvláště hranice mezi přechodnými jevy, jak jsou použity v provizorně syntakticky označkovaném korpusu. Definice jsou zjednodušené, formální, kladu důraz jen na rozdíly oproti „tradičním, školským“ definicím.

Podmět (*Sb*) je tradiční „člen predikační dvojice“, většinou v nominativu, zřídka v bezpředložkovém genitivu (*Je toho názoru, že není vážných **problémů** mezi stranami*). Podmět v předložkové frázi (*Kolem 30 **centimetrů** sněhu pokrývá svahy ve skiareálu Zadov*) vyjadřující neurčitou kvantifikaci se vyskytuje zcela výjimečně, ve zkoumaných datech představuje u substantiv méně než 0,001 % výskytů.

Syntaktickou funkcí pro jmennou část verbonominálního přísudku (*Pnom*) se v anotovaném korpusu označuje výhradně bezpředložková substantiva se sponou *být/bývat*, většinou v nominativu či instrumentálu, zřídka v genitivu (*Jeho otec byl toho **názoru**, že...*). Předložkové fráze rozvíjející sponu jsou značeny jako příslovečná určení.

Přívlastek (*Atr*) u substantiv zahrnuje ve formátu PDT jednak neshodné přívlastky, jednak součásti adordinačních spojení typu *pan Jan Novák,* kde se za řídící člen považuje poslední substantivum, všechny předchozí členy jsou označeny jako shodné přívlastky.

Předmět (*Obj*) je větný člen, jehož tvar a pád je určen slovesem (adjektivem), zasažený dějem slovesa. Mezi předměty se řadí také „původce děje“ u pasivních sloves a některé jiné vazební členy tradičně řazené mezi příslovečná určení (určení původu a výsledku).

Příslovečné určení (*Adv*) je nevazebný větný člen, jeho pád (a předložka) je dán významem, ne vazbou slovesa. Abstraktní, přenesené výrazy jsou často spíše předměty než příslovečnými určeními.

Elipsa (*ExD*) zahrnuje především substantiva v nevětných výpovědích, dále substantiva ve strukturách s elidovaným členem (aktuální elipsa). Rozpoznání struktur s aktuálně elidovaným členem je pro parser obtížné a není ani snadné je opravit na základě pravidel; mimo nevětné výpovědi bylo nutné tuto funkci často opravovat. Po manuální korekci tak většina případů pochází ze struktur neobsahujících sloveso (názvy, výčty, odpovědi v dialogu, vokativy aj.).

5.2 Tabulky zastoupení syntaktických funkcí a pádů substantiv v rámci korpusu SYN2005

Uvádím tři tabulky, zobrazující tatáž data jako podíly z různých základů. V první tabulce je základem pád (samostatně pád předložkový a bezpředložkový). Údaje (v procentech) v první tabulce tedy ukazují, jak často je určitý pád využit k realizaci různých syntaktických funkcí. V druhé tabulce je základem syntaktická funkce, tabulka ukazuje, jak často je syntaktická funkce vyjádřena substantivem v určitém pádu (popř. s předložkou). Ve třetí tabulce je základem součet všech výskytů substantiv. Procenta jsou zaokrouhlená na jedno desetinné místo, přesnější údaje by vzhledem k chybovosti neměly smysl. Tabulky nepotřebují podrobný komentář, záleží na čtenáři, které údaje ho zajímají a jaké závěry z nich vyvodí.

Zastoupení syntaktických funkcí v rámci pádů (bez předložky a s předložkou)

	Sb	Pnom	Atr	Obj	Adv	ExD	
Nom.	77,2	6,7	11,9	0,2	0,0	17,9	100
Gen.	0,2	0,0	95,8	3,5	0,4	0,1	100
prep Gen.	0,0	0,0	19,5	3,3	77,1	0,1	100

Dat.	0,0	0,0	3,1	93,8	2,4	0,7	100
prep Dat.	0,0	0,0	17,6	26,3	55,8	0,3	100
Ak.	0,0	0,0	0,4	95,2	3,7	0,7	100
prep Ak.	0,0	0,0	29,6	25,6	44,1	0,7	100
Vok.	0,0	0,0	10,9	0,0	0,0	89,1	100
prep Lok.	0,0	0,0	27,5	5,1	67,2	0,2	100
Instr.	0,0	22,7	10,5	25,6	40,2	1,0	100
prep Instr.	0,0	0,0	46,1	10,3	43,4	0,2	100
všechny pády	18,6	2,9	31,3	20,2	21,1	5,9	100

Zastoupení pádů (bez předložek nebo s předložkami) v rámci syntaktických funkcí

	Sb	Pnom	Atr	Obj	Adv	ExD	všechny funkce
Nom.	99,8	66	11,1	0,3	0,0	88,7	29,2
Gen.	0,2	0,2	60,2	3,4	0,4	0,5	19,7
prep Gen.	0,0	0,0	4,8	1,3	28,1	0,1	7,7
Dat.	0,0	0,0	0,2	8,0	0,2	0,2	1,7
prep Dat.	0,0	0,0	1,1	2,5	5,1	0,1	1,9
Ak.	0,0	0,0	0,2	66,9	2,5	1,7	14,2
prep Ak.	0,0	0,0	5,5	7,4	12,1	0,7	5,8
Vok.	0,0	0,0	0,2	0,0	0,0	6,7	0,4
prep Lok.	0,0	0,0	9,8	2,8	35,4	0,4	11,1
Instr.	0,0	33,8	1,5	5,6	8,4	0,8	4,4
prep Instr.	0,0	0,0	5,6	2	7,9	0,1	3,8
	100	100	100	100	100	100	100

Zastoupení kombinace syntaktických funkcí a pádů v rámci všech substantiv

	Sb	Pnom	Atr	Obj	Adv	ExD	celkem
Nom.	18,5	1,9	3,5	0,1	0,0	5,2	29,2
Gen.	0,0	0,0	18,9	0,7	0,1	0,0	19,7
prep Gen.	0,0	0,0	1,5	0,3	5,9	0,0	7,7
Dat.	0,0	0,0	0,1	1,6	0,0	0,0	1,7
prep Dat.	0,0	0,0	0,3	0,5	1,1	0,0	1,9
Ak.	0,0	0,0	0,1	13,5	0,5	0,1	14,2
prep Ak.	0,0	0,0	1,7	1,5	2,6	0,0	5,8
Vok.	0,0	0,0	0,0	0,0	0,0	0,4	0,4
prep Lok.	0,0	0,0	3,1	0,6	7,5	0,0	11,1
Instr.	0,0	1,0	0,5	1,1	1,8	0,0	4,4
prep Instr.	0,0	0,0	1,8	0,4	1,7	0,0	3,8
celkem	18,6	2,9	31,3	20,2	21,1	5,9	100

5.3 Další možné směry výzkumu vztahu pádů a syntaktických funkcí

Korpusy řady SYN jsou složeny z různých typů textů, jedním z dalších směrů výzkumu by tak mělo být srovnání frekvencí pádů a syntaktických funkcí podle žánrů. Dále by bylo vhodné podrobněji rozdělit „předložkové pády“ podle konkrétních předložek a přizpůsobit tomu statistiky syntaktických funkcí (víme, že předložka *o* s lokálem většinou vyjadřuje předmět, kdežto předložka *v* s lokálem většinou vyjadřuje příslovečné určení, ale nevíme, v jakém poměru). Ze syntakticky anotovaného korpusu bude možné získat podrobné soupisy valenčních sloves a adverbiálních konstrukcí pro jednotlivé kombinace předložky a pádu aj.

6. Závěr

Syntaktické značkování korpusů může zajímavě rozšířit existující nástroje pro práci s korpusem. Cílem projektu *Syntaktická anotace českých korpusů* je umožnit, aby uživatelé využívali syntakticky označkovaný korpus a mohli si sami zvolit zobrazení a funkce, které potřebují. V tomto příspěvku jsem představil proces automatického syntaktického značkování, které bude použito pro zpracování rozsáhlých textových korpusů synchronní češtiny. Značkování bude provedeno stochastickým MST parserem, frekventované typy chyb parseru budou korigovány automatickým opravným modulem založeným na lingvistických pravidlech. Vývoj tohoto opravného modulu dosud nebyl ukončen, ale již teď znatelně snižuje počet chyb v syntaktickém značkování. V závěru článku jsem jako příklad využití syntaktického korpusu představil statistiky syntaktických funkcí substantiv ve vztahu k jejich pádu a použití předložky.

Literatura

Český národní korpus – SYN2005, 2005. *SYN2010*, 2010. Ústav Českého národního korpusu FF UK, Praha. Dostupný z WWW: <http://ucnk.ff.cuni.cz>.

Hajič J. et al, 2006, *Prague Dependency Treebank 2.0.* CD-ROM, Philadelphia: Linguistic Data Consortium.

Hajič, J. 2004, *Complex Corpus Annotation: The Prague Dependency Treebank.* Jazykovedný ústav Ľ. Štúra, SAV, Bratislava.

Hajič J., J. Panevová, E. Buráňová, Z. Urešová, A. Bémová, 1999, *Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory*. Dostupné z WWW: <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/html/index.html>

Hnátková M., P. Jäger, T. Jelínek, V. Petkevič, A. Rosen, H. Skoumalová, 2011, Syntakticky anotovaný korpus českých textů. In *Korpusová lingvistika 2011. III Gramatika a značkování korpusů*, eds V. Petkevič, A. Rosen, Nakladatelství Lidové noviny, Praha.

Jelínek T., V. Petkevič, 2011, Systém jazykového značkování korpusů současné psané češtiny. In *Korpusová lingvistika 2011. III Gramatika a značkování korpusů*, eds V. Petkevič, A. Rosen, Nakladatelství Lidové noviny, Praha.

McDonald R., F. Pereira, K. Ribarov, J. Hajič, 2005, Non-projective dependency parsing using

spanning tree algorithms. In *HLT'05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, 523–530.

Novák V., Z. Žabokrtský, 2007, Feature Engineering in Maximum Spanning Tree Dependency Parser. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*. Berlin, Heidelberg: Springer-Verlag, LNCS 4629, 92–98.

Šmilauer V., 1966, *Novočeská skladba*. SPN, Praha.