ROZHLEDY

Barbora Hladká – Zdeňka Urešová – Alla Bémová

Syntaktická proměna Českého akademického korpusu

The syntactic transformation of the Czech Academic Corpus

ABSTRACT: The idea of the Czech Academic Corpus (CAC) came to life in 1971 thanks to the Department of Mathematical Linguistics within the Czech Language Institute. By the mid 1980s, a total of 540,000 words were morphologically and syntactically annotated manually. After the Prague Dependency Treebank (PDT) – the largest annotated treebank of Czech written texts – was built, the conversion from CAC to PDT format began. The main goal was to make the CAC and the PDT compatible, and thus to enable the integration of the CAC into the PDT. The second version of the CAC is thus a complete conversion of the internal format and annotation schemes. The conversion of syntactic annotation began three years after the syntactic annotation of PDT was finished. Such a situation is exceptional because, to our knowledge, there is no other language for which such a significant amount of data is being annotated in two subsequent projects. This article summarizes the experience acquired during the conversion of the CAC syntactic annotation.

Key words: corpus, syntactic annotation, annotation guidelines, annotation checking

Klíčová slova: korpus, syntaktická anotace, pokyny k anotaci, oprava anotací

Předkládaným příspěvkem bezprostředně navazujeme na stať Proměna Českého akademického korpusu, která byla publikována ve Slově a slovesnosti v roce 2006 (Hladká – Králík, 2006). Nyní popisujeme zkušenosti se syntaktickou proměnou Českého akademického korpusu, při které byly původní syntaktické anotace ponechány stranou a texty byly nově anotovány dle koncepce Pražského závislostního korpusu. Proměna, neboli anotace, byla zahájena tři roky poté, co byla dokončena syntaktická anotace již zmíněného Pražského závislostního korpusu – největšího anotovaného korpusu psané češtiny. Tento tříletý časový odstup je do jisté míry kuriózní; neznáme jiný jazyk, pro který by po anotování velkého objemu dat (více než jeden milion slov) proběhla anotace dalších dat, sice objemu menšího, ale rovněž nezanedbatelného (statisíce slov).

Syntaktickou anotací Českého akademického korpusu jsme vstoupili podruhé do stejné řeky. Doufáme, že zkušenost, kterou si odnášíme, bude přínosná pro všechny jazykovědce.

1. Český akademický korpus

Český akademický korpus (dále ČAK) je morfologicky a syntakticky ručně anotovaný soubor vybraných textů, který obsahuje 600 tisíc slov ve 33 tisících větách. Jeho anotování, resp. anotování Korpusu věcného stylu, jak zněl původní název, bylo zahájeno

v roce 1975 v Ústavu pro jazyk český AV ČR pod vedením Marie Těšitelové. Cílem bylo získat materiál, na jehož základě budou zpracovány kvantitativní charakteristiky češtiny té doby (Těšitelová, 1983, 1984; Těšitelová – Uhlířová – Králík, 1984).

V době původní anotace ČAKu byly známy další tři anotované korpusy, a sice Brownův korpus textů americké angličtiny (Francis – Kucera, 1979), korpus LOB textů britské angličtiny (Atwell – Leech – Garside, 1984) a korpus švédských textů Talbanken (Teleman, 1974; Nilsson – Hall – Nivre, 2005). Oba korpusy angličtiny obsahují jeden milión morfologicky anotovaných slov a švédský Talbanken obsahuje 350 tis. slov, anotovaných morfologicky i syntakticky. ČAK převyšoval uvedené tři korpusy v množství lingvistické informace, kterou původní anotace jeho textů zachycovaly.

Původní anotace ČAKu byla dokončena v roce 1987. Krátce poté, na počátku devadesátých let, došlo v počítačovém zpracování přirozeného jazyka k nástupu empirických metod, nejčastěji pojmenovávaných *korpusové metody*. Pro češtinu byly první korpusové metody využity k řešení automatického morfologického značkování českých textů. Díky těmto experimentům přešel ČAK pod patronaci Ústavu formální a aplikované lingvistiky MFF UK.

Na přelomu tisíciletí se v kontextu projektu Pražského závislostního korpusu (PZK)[1] a ostatních „spřízněných" pražských anotačních projektů (Prague Arabic Dependency Treebank,[2] Prague English Dependency Treebank,[3] Prague Czech-English Dependency Treebank[4]) začalo na Český akademický korpus nahlížet v novém světle. Jak anotační schémata, tak i technologie použité pro jeho reprezentaci byly konfrontovány s metodami aplikovanými ve zmíněných projektech. Aby mohl být ČAK přičleněn k PZK, což bylo vzhledem k jeho původnímu objemu 540 000 slov vyhodnoceno jako přínosné pro automatické zpracování češtiny, bylo třeba zajistit vzájemnou kompatibilitu obou korpusů. Z tohoto důvodu jsme změnili nejen vnitřní formát ČAKu, ale do značné míry i jeho anotace.

Proměny byly realizovány postupně a jejich výsledky byly uceleně prezentovány formou samostatných CD-ROMů. První proměna ČAKu – konverze vnitřního formátu a morfologických anotací – vyústila publikováním tzv. první verze Českého akademického korpusu (ČAK 1.0; Vidová Hladká et al., 2007).[5] Následně prošly proměnou i syntaktické anotace a byl vytvořen ČAK 2.0 (Vidová Hladká et al., 2008a; Vidová Hladká et al., 2008b),[6] který obsahuje morfologicky a syntakticky anotované psané texty a pouze morfologicky anotované mluvené texty (tuto asymetrii vysvětlujeme v odd. 5).

Anotační manuál a anotační schémata proměny ČAKu jsou výsledkem důkladného studia české syntaxe, a to jak na základě mluvnic, tak i na základě empirických pozorování jazyka. Je třeba si uvědomit, že s ohledem na bohatost a složitost češtiny nelze

[1] Viz online na adrese <http://ufal.mff.cuni.cz/pdt2.0>.
[2] Viz online na adrese <http://ufal.mff.cuni.cz/padt>.
[3] Viz online na adrese <http://ufal.mff.cuni.cz/pedt>.
[4] Viz online na adrese <http://ufal.mff.cuni.cz/pcedt>.
[5] Viz online na adrese <http://ufal.mff.cuni.cz/rest/CAC/cac_10.html>.
[6] Viz online na adrese <http://ufal.mff.cuni.cz/rest/CAC/cac_20.html>.

anotační manuál a anotační schémata chápat jako jednou provždy dané dogma. Zcela jistě v případě dalšího anotačního projektu projdou revizí, jejímž výsledkem budou nové úpravy a doplňky anotačních pouček.

V průběhu anotace ovšem není v zájmu její konzistence možné anotační schéma měnit. Právě tento fakt jsme respektovali i při proměnách ČAKu. Zdůrazňujeme, že naším primárním cílem nebyla integrace ČAKu s Pražským závislostním korpusem, ale „aktualizace" anotací ČAKu vzhledem k současným studiím češtiny v kontextu jejího počítačového zpravování. Protože praktickým důsledkem těchto studií byl právě PZK, mohla být úloha „aktualizace" ČAKu chápána i jako integrace těchto dvou korpusů.

První proměně ČAKu byl věnován citovaný příspěvek z roku 2006 (Hladká – Králík, 2006); lingvistickým a praktickým aspektům druhé proměny se budeme věnovat v tomto článku.

2. Syntaktická proměna ČAKu v kostce

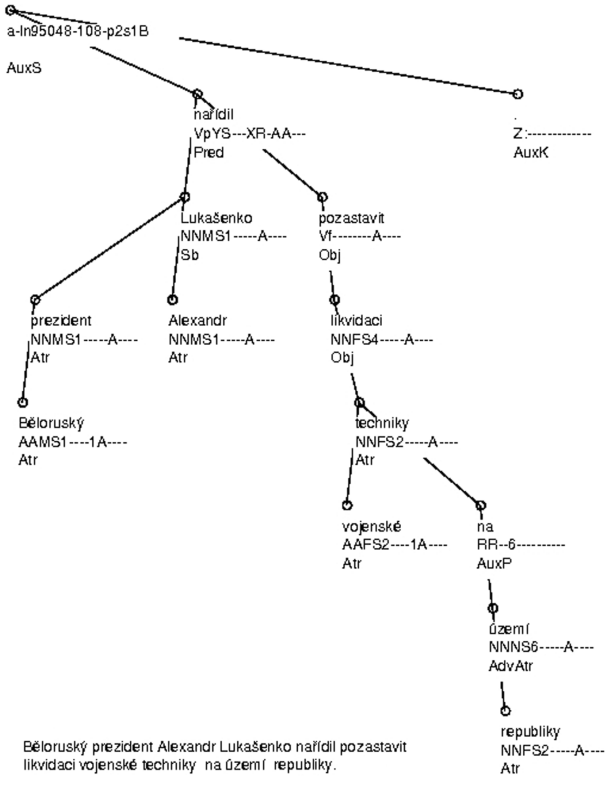
Pilotní studie (Ribarov – Bémová – Hladká, 2006), která dospěla k poznání, že syntaktická anotace použitá v době vzniku ČAKu je natolik nekompatibilní s koncepcí PZK a zároveň i z hlediska dnes zkoumaných jevů neúplná (zejména v oblasti úlohy interpunkce, funkcí vedlejších vět, ignorování některých pomocných slov apod.), podpořila následující strategii syntaktické proměny: odhlíží se od původních syntaktických anotací, texty jsou zpracovány automatickou syntaktickou analýzou (McDonald et al., 2005) a automaticky přiřazené anotace – již odpovídající koncepci PZK – jsou následně zkontrolovány a ručně opraveny anotátory. Tato strategie je časově efektivní zejména u kratších vět, jejichž automatická syntaktická analýza je téměř bezchybná. Poznamenejme, že na rozdíl od anotace syntaktické byla manuální anotace morfologie převzata po konverzi téměř v plném rozsahu (Hladká – Králík, 2006).

Zpočátku byly automaticky přiřazené anotace textů kontrolovány a opravovány dvěma anotátory. V okamžiku, kdy byli anotátoři dostatečně zacvičeni, kontroloval každý text pouze jeden anotátor. Celkově však byla kontrola dvoustupňová, protože texty a jejich anotace přebíral po anotátorovi arbitr, tj. nejzkušenější anotátor, a anotaci opakovaně kontroloval a opravoval. Spolu s ručními opravami probíhaly i kontroly a opravy automatické, které byly navrženy pro kontrolu anotací PZK.

Pro úplnost dodáváme, že syntaktická proměna ČAKu, dále *transformace*, příp. *konverze*, se týkala jeho verze 1.0.

3. Koncepce Pražského závislostního korpusu

Koncepce anotování Pražského závislostního korpusu, kterou pro všechny pražské anotační projekty zkracujeme na k o n c e p c i P Z K, pracuje se třemi anotačními rovinami: morfologickou, syntakticko-analytickou a hloubkově syntaktickou, tzv. tektogramatickou (Hajič et al., 2006). Anotacím na jednotlivých rovinách odpovídají reprezentační struktury dvojího typu – pro morfologickou rovinu je to lineární seznam a pro syntaktickou a tektogramatickou rovinu je to závislostní strom. Anotační roviny jsou



Obr. 1: Ukázka morfologické a syntaktické anotace v PZK

doplněny nultou, neanotační slovní nebo též lexikální rovinou, která obsahuje text v jeho původní podobě.

Pro téma našeho článku je relevantní rovina analytická. V závislostním stromu analytické roviny, tzv. a-stromu (orientovaný acyklický graf s jedním kořenem), jsou zachována všechna slova věty a interpunkční symboly. Ve výsledném stromu jsou všechny uzly ohodnoceny příslušnou analytickou funkcí (viz přílohu Analytické funkce v PZK; pro ilustraci viz obr. 1).

4. Vynechávky v textech ČAKu

Původním cílem projektu Českého akademického korpusu bylo shromáždění dostatečně reprezentativního textového materiálu pro všestrannou kvantitativní charakteristiku češtiny té doby. Z hlediska takto formulovaného záměru se některé prvky textů jevily jako nepodstatné, a proto byly z analyzovaného materiálu buď zcela vypuštěny (např. ciferné výrazy a interpunkce), nebo zanedbány (např. nerozlišovala se malá a velká písmena). Anotace textů PZK probíhala v rámci experimentálního ověřování definice formální reprezentace analýzy českých vět, a proto i různé technické záležitosti textů hrály důležitou roli. Vzhledem k tomu klasifikujeme tento druh odlišností textů ČAKu od PZK jako nedostatky, které dále v přehledu shrnujeme.

A. Chybějící výrazy ve větě

Věta *Mezi řekami [] a Tigridem žily [] [] lety dva odlišné [].* obsahuje v ČAKu čtyři chybějící slova. Zástupný symbol „[]“ za chybějící slova, interpunkci a ciferné výrazy byl korektory doplněn do textů ČAKu při jeho první proměně (Hladká – Králík, 2006). Zástupné symboly tvoří v a-stromech samostatné uzly s nenaplněnými atributy, ale jsou součástí syntaktické struktury a měly by dostat analytickou funkci. Aby se analytické funkce nepřiřazovaly prázdným symbolům, nahrazovali je anotátoři konkrétní informací dle vlastního uvážení. Někdy bylo ovšem obtížné doplnit konkrétní slovní tvar, proto anotátoři pracovali i s obecnější informací, např. že jde o substantivum, adjektivum, nebo o infinitní tvar slovesa. Technická realizace náhrady zástupného symbolu probíhala tak, že anotátor doplňoval informaci (tj. všechny možnosti, které ho napadly) do speciálního atributu, jenž byl pro tyto účely doplněn do vnitřní reprezentace ČAKu. Rekonstruovaná podoba výše uvedené věty by mohla vypadat takto (anotátor poznamenal do speciálního atributu možnosti zvýrazněné podtržením): *Mezi řekami Eufratem a Tigridem žily před několika / mnoha lety dva odlišné národy / kmeny.*

Doplňování chybějících slov je žádoucí také v případech, kdy pomůže odstranit případnou homonymii věty. Například ve větě *Palácové soubory, vytvářené na vyvýšené terase [] širokým schodištěm, jsou souměrné s rozsáhlými sloupovými síněmi…* lze chybějící slovní tvar doplnit přinejmenším dvěma způsoby. Pokud je to *terasa se širokým schodištěm*, doplněným chybějícím slovem je předložka *se*, která se na analytické rovině ohodnotí analytickou funkcí pro předložku (AuxP). Pokud je to *terasa přístupná širokým schodištěm*, je doplněným chybějícím slovem adjektivum *přístupná* s přiřazenou analytickou funkcí pro přívlastek (Atr).

Problém při doplňování chybějících ciferných výrazů vznikal proto, že takové výrazy jsou v a-stromech zachyceny dvojím způsobem. Pro výrazy obsahující základní číslovky 2, 3 a 4 (nebo výrazy, které těmito číslovkami končí) je reprezentantem číselného výrazu počitatelný objekt a číslovka je v závislostním stromě zavěšena na tomto objektu jako jeho přívlastek. Pro výrazy obsahující číslovky od 5 výše je tomu naopak – reprezentantem celého číselného výrazu je sama číslovka a počitatelný objekt je jejím přívlastkem. V případě jako *[] stoly – [] stolů* (1. a 2. pád plurálu), tj. pokud je počitatelné substantivum přítomno, lze na základě pádu substantiva rozpoznat, zda jde o první nebo druhou variantu zachycení číselného výrazu. Avšak v případech jako *Vláda Peršanů trvala [] století*., kde je pádový tvar substantiva *století* homonymní (*dvě století/pět století*), taková možnost rozlišení není a na přesné doplnění číselných výrazů se muselo rezignovat. Totéž platí i pro konstrukce, v nichž za číselným výrazem následuje označení míry (*cm*, *m*, *km*, *kg*, *t*, *l*, atd.) nebo další zástupný symbol.

B. Přebývající výrazy ve větě

Některé věty v ČAKu obsahují přebývající výrazy. Například v následující větě přebývá sloveso *vypnout*: *Rozměrné kamenné bloky těchto [] hladce opracované kamennými nástroji jsou dokladem zručnosti vypnout značné vyspělosti lidí této doby…* Místo slovesa *vypnout* by mohla být doplněna spojka *a* nebo čárka: *Rozměrné kamenné bloky těchto staveb … jsou dokladem zručnosti a/, značné vyspělosti lidí této doby…*

C. Neoprávněně rozdělené věty

V textech ČAKu docházelo někdy k neoprávněnému rozdělování vět. Například v původním textu *…jejich analýzu důsledně dovršil X.Y., který napsal. Abstrahujeme-li od užitné hodnoty zbožních těles, zbude jim jen jediná vlastnost, že jsou totiž produkty práce…* byla na místě tečky rozdělující věty pravděpodobně dvojtečka, která odpovídá napojení obou strukturních částí dané konstrukce.

D. Nadbytečná interpunkce

Kromě nadbytečných výrazů najdeme v textech ČAKu i nadbytečnou interpunkci (tzn. že názor korektora a anotátora na umístění interpunkce se lišil). Ve větě *Milia lze otevřít jemným skalpelem a vytlačit, obsah hydradenomy, je nutno odstranit většinou chirurgicky* je čárka za slovem *hydradenomy* nadbytečná a činí strukturu gramaticky nesprávnou. Podobně v konstrukci *…účast pracujících na řízení ideové a politicky, výchovné práce a hospodářské politiky* je čárka mezi slovy *politicky* a *výchovné* doplněna nesprávně; místo ní by mohla být buď pomlčka: *politicky-výchovná práce*, nebo nic, pak jde o dvouslovné syntagma *politicky výchovná*.

Zejména z kapacitních důvodů nebyly odlišnosti A–D syntaktickými anotátory ani nikým jiným dosud opravovány. V poměru k celkovému množství zpracovaného materiálu není množství těchto odlišností a nesrovnalostí naštěstí nijak podstatné.

5. Mluvené texty ČAKu

Vážný problém z hlediska syntaktické anotace představují mluvené projevy, jejichž přepisy tvoří v ČAKu zhruba třetinu celkového objemu textů. Struktura mluvených projevů má naprosto specifický charakter a podstatně se liší od textů psaných. Ústní projev obsahuje velké množství neúplných vět, které jen částečně tvoří smysluplnou strukturu, a také velké množství tzv. „intonačních vycpávek“, jež se do struktury věty vůbec nezačleňují. Pravidla Pražského závislostního korpusu pro anotaci analytické roviny nepočítají s anotací mluvených projevů, a tak zde prostředky pro zachycení specifických jevů mluvené řeči chybějí. Například nejsou pravidla pro zachycení bezdůvodného opakování stejného slova nebo pro zachycení koktání. V následujícím příkladovém textu z ČAKu najdeme hned několik typických znaků ústního projevu: *A to jsou trošku, jedna je, jedna má světlou budovu a druhá má tmavou budovu, ony jsou umístěny v jednom, v jednom areále, ale ta, to centrum, patřilo té, bylo to v bloku Univerzity vlámské, a já jsem se ptala na univerzitě, na, v Univerzitě svobodné, že, no a to přeci oni vědí, to nanejvýš, to prostě jediné, když je Univerzita vlámská, tak o tom oni přece nemohou nic vědět, a nic.* Mezi typické znaky ústního projevu v tomto textu patří:

– nedořečené věty (fragmenty): *A to jsou trošku…*
– opakované začátky (tzv. restarty): *…jedna je, jedna má…*
– opakovaná slova uprostřed vět: *jsou umístěny v jednom, jednom areále…*
– nadbytečná a nesprávně užitá gramatická slova: *univerzitě, na, v Univerzitě svobodné,…*
– nadbytečná deiktická slova: *…ale ta, to centrum…*
– diskurzní částice: *no…*

– české otázky presumptivní (dovětky typu *že ano, že ne…*): *…na Univerzitě svobodné, že…*
– nadbytečné konektory: *…když je to Univerzita vlámská, tak to o tom…*
– porušení koherence výpovědi, „roztrhané“ syntaktické schéma propozice: *…ale ta, to centrum, bylo to v bloku…*
– syntaktické chyby typu anakolutu: *…přeci nemohu nic vědět, a nic.*

Jak již bylo uvedeno výše, pro adekvátní zachycení těchto rysů mluveného projevu existují v pravidlech anotace analytické roviny PZK jen velmi omezené prostředky. Například slova v roli intonačních výplní by mohla být ohodnocena analytickou funkcí AuxO, která je používána v psaných textech pro označení nadbytečných (odkazovacích nebo emotivních) elementů. Většině prvků ústního projevu by však musela být přiřazena analytická funkce ExD (*Ex-Dependent*), která pouze upozorňuje na to, že se jedná o strukturu neúplnou, tj. o elipsu, v níž chybí řídící slovo. Tato analytická funkce nic dalšího o struktuře nevypovídá.

S ohledem na důvody vyjmenované výše bylo rozhodnuto ponechat mluvené projevy v ČAK 2.0 bez syntaktických anotací. Podrobněji viz Hladká – Urešová (2009).

6. Anotátoři

Sestavení a zacvičení anotátorského týmu je klíčovou otázkou jakéhokoli anotačního projektu. Připomeňme čtenářům několik požadavků dokládajících náročnost práce anotátora:

(1) Anotování vyžaduje určitou úroveň odborných jazykových znalostí z oblasti morfologie i syntaxe. Předpokládá jistou zkušenost s větným rozborem, tedy schopnost analyzovat větu tak, aby vznikla správná větná struktura. Tyto požadavky by mohli splňovat zejména studenti jazykovědných oborů vysokých škol.

(2) Anotování vyžaduje zvládnutí pravidel anotačního manuálu. Předcházející odborná výchova anotátorů se může poněkud lišit od konvencí a pravidel přijatých pro dané anotování. Ostatně i jednotlivé tradiční mluvnice se liší v názorech na některé syntaktické jevy, jak o tom svědčí např. rozdíly ve vymezení hranice mezi objektem a adverbiálním určením. Anotátor tudíž musí korigovat svoje dosavadní znalosti a návyky a přizpůsobit je pokynům manuálu. I přes detailní znalost pravidel anotace se anotátor setkává s jevy, jako je homonymie či možnost různé interpretace stejného pravidla, a musí se s nimi umět vyrovnat. Musí umět najít správné řešení také pro jevy, které z různých důvodů nejsou v manuálu zachyceny adekvátně.

(3) Anotování vyžaduje vysokou míru soustředění. Samotný proces anotování je náročný jak na čas, tak i na pozornost. Anotátor sleduje správnost stromové struktury a správnost přiřazení analytických funkcí jednotlivým uzlům. Porozumění textu z neznámého oboru vyžaduje mnohdy opakované čtení analyzované věty. Anotace syntaktické struktury dlouhých vět je navíc i časově náročná. Neméně časově náročné je anotování vět s elipsami. Nejde zde tolik o samotné porozumění obsahu věty, jako spíše o budování její stromové reprezentace. Věty s elipsami způsobují často neprojektivitu a vzdálenost mezi členem řídícím a členem závislým je někdy více než deset uzlů.

Na základě zkušeností chceme připomenout, že práce anotátora je náročná a mnohdy vyčerpávající, má spíše rutinní než tvůrčí charakter. V každém případě je to však práce přínosná, přinášející nové pohledy a poznatky.

Pro anotování ČAKu jsme kontaktovali studenty Filozofické fakulty UK. Naše snaha byla bohužel neúspěšná; několik studentů sice začalo anotovat, ale spolupráce s nimi nepřesáhla zkušební období. Poté se anotování ČAKu ujali mladí anotátoři ze Slovenska, kteří v rámci projektu Slovenského národního korpusu (SNK)[7] pracují na podobném projektu pro slovenštinu. Vzhledem k příbuznosti českého a slovenského jazyka jsou základní principy budování anotovaného závislostního korpusu velmi podobné. Proto se využití jejich odborné zkušenosti nabízelo jako vhodné řešení. Slovenští anotátoři splňovali požadavek odborné připravenosti a do značné míry i požadavek znalosti českého manuálu. Zůstává však otázkou, do jaké míry byla pro ně, jakožto nerodilé mluvčí, překážkou při anotování syntaxe českého jazyka samotná čeština. Oproti předchozí slovenské generaci, která se s češtinou setkávala denně, může dnešní mladá generace osamostatněného Slovenska vnímat češtinu už jako jazyk cizí. Celkový výsledek práce slovenských anotátorů však nebyl znepokojivý (podrobněji viz odd. 7).

7. Anotační pravidla – manuál

Manuál pro syntaktickou anotaci analytické roviny (Hajič et al., 2004) vznikal souběžně s anotováním PZK. Anotační pravidla zachovávají, pokud je to možné, tradiční pojmy českých mluvnic. Vychází se zejména z Novočeské skladby V. Šmilauera (1947). Šmilauerův přístup se promítá především do obdobného pojetí základních syntaktických funkcí. Seznam syntakticko-analytických funkcí, které se v PZK používají (viz Přílohu Analytické funkce v PZK), je však mnohem širší, protože vzhledem k „nepočítačovému“ Šmilauerovu pojetí české syntaxe musela být tradiční česká pravidla na mnoha místech rozšířena, popřípadě přeformulována. Přesto PZK obsahuje jevy, které nejsou ani tradičními gramatikami (zaměřenými na lidského uživatele), ani manuálem pro analytickou anotaci (zaměřeným na exaktní počítačové zpracování) popsány. V těchto výjimečných případech byla rozhodnutí ponechána na jazykovém citu anotátorů, kteří pak rozhodovali jednotlivě.

Základní analytické funkce odpovídají klasickým větným členům, jak jsou známy i ze školního větného rozboru: subjekt (Sb), predikát (Pred), včetně predikátu nominálního (Pnom), objekt (Obj), adverbiální určení (Adv), atribut (Atr), doplněk (Atv, AtvV). Oproti Šmilauerovu pojetí se však jinak vymezují vzájemné hranice jednotlivých větných členů, zvláště objektu, adverbiálního určení a doplňku. Dále souhrnně uvádíme přehled jevů z reálných anotovaných textů, které jsou v tradičních českých syntaktických příručkách buď popsány způsobem odlišným od pravidel pro analytickou anotaci PZK, nebo se s nimi vůbec nepočítá:

[7] Viz online na adrese <http://korpus.juls.savba.sk/>.

– Funkce přívlastku (Atr) je v PZK širší. Používá se nejen pro klasický přívlastek, ale i pro složky adres a jmen, pro složky cizojazyčného textu a pro složky číselných výrazů.
– Funkce příslovečného určení (Adv) zůstává naproti tomu na analytické rovině bez dalšího třídění, její jemnější dělení najdeme až na tektogramatické rovině PZK.
– Pro reprezentaci koordinačních a apozičních vztahů mezi jednotlivými větnými členy, případně mezi větnými celky, jsou k dispozici funkce pro koordinaci (Coord) a apozici (Apos); pro označení parentetických částí vět se používá přípony _Pa.
– Z tradičního repertoáru větných členů se vymyká analytická funkce ExD (Ex-Dependent), která se používá pro označení výrazu v eliptické konstrukci. Přiřazuje se tehdy, chybí-li nějakému závislému členu řídící uzel.
– Pro strukturní víceznačnost (v případech, kdy daná konstrukce vyjadřuje věcně totéž, tj. má stejný význam bez ohledu na formální syntaktické vyjádření) bylo třeba vzhledem k potřebě exaktního počítačového popisu zavést tzv. kombinované funkce. Analytická funkce AtrAdv nebo AdvAtr vyjadřuje (při shodném významu) (pseudo-)nejistotu mezi závislostí adverbiální a adnominální, zatímco analytická funkce ObjAtr nebo AtrObj vyjadřuje (pseudo-)nejistotu mezi závislostí objektovou a adnominální. Analytická funkce AtrAtr znamená, že za řídící slovo atributu je možné beze změny významu vybrat kteréhokoli z rodičů syntaktických substantiv.
– Speciální analytické funkce dostávají i „slova“ (v analytické stromové reprezentaci jsou to uzly), která samostatnými větnými členy většinou nejsou: pomocné sloveso *být* (AuxV), zvratné *se* u reflexiv tantum (AuxT), zvratné *se* u reflexivního pasiva (AuxR), předložka (AuxP), spojka podřadicí (AuxC), odkazovací (emotivní) element (AuxO), zdůrazňovací slovo (AuxZ), částice vztahující se k celé výpovědi (AuxY). Samostatnými uzly v našem pojetí jsou také čárka (AuxX)[8], grafické symboly (AuxG), koncová interpunkce (AuxK) a technický kořen stromu (AuxS).

V důsledku výše jmenovaných odlišností od tradičního pojetí dochází k četným odchylkám také při vytváření stromové reprezentace povrchově syntaktické struktury textu:

– Na symbolu kořene stromu (AuxS) je zavěšen predikát, který řídí celou větnou strukturu. Skládá-li se z více slov, funkci Pred dostává významové sloveso. Ostatní členy predikátu jsou zavěšeny na něm, a to buď s přiřazenou analytickou funkcí AuxV, nebo Pnom. U predikátu vyjádřeného zvratným slovesem se zavěšuje částice *se* (AuxR nebo AuxT) na toto sloveso. U predikátu složeného (*může pracovat*, *hodlá překládat*) se infinitivu přiřazuje funkce Obj, která označuje objekt.
– Nejzásadnější odlišností oproti tradičním gramatikám je jiné zachycení vztahu mezi subjektem a predikátem věty. Subjekt se považuje za rozvíjející větný člen stejně jako ostatní členy predikátu a zavěšuje se na svůj řídící uzel, tj. na predikát. Opouští se zde tudíž tradiční pojetí základní větné dvojice.

[8] Čárka je ohodnocena funkcí AuxX, pokud nevyjadřuje jinou analytickou funkci: čárka může být např. nositelem koordinace, pak je ohodnocena funkcí Coord, nositelem apozice, pak je ohodnocena funkcí Apos.

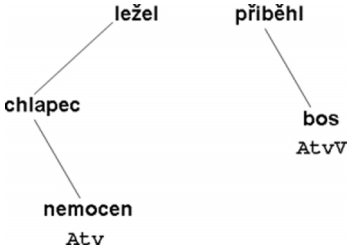
– Pro zavěšení koordinačních a apozičních vztahů byla přijata následující konvence:
  - Nositelem koordinační funkce jsou spojky souřadicí, spojovací výrazy a interpunkční znaky (např. čárka), které dostávají funkci Coord a jsou řídicím uzlem celého koordinačního řetězu. Koordinační členy se zavěšují na uzel s hodnotou Coord tak, jako kdyby to byly členy závislé.
  - Apozice se v PZK zachycuje stejně jako koordinace. Za spojovací výrazy apozice se pokládají jak slovní výrazy, tak i grafické prostředky, jako je čárka, pomlčka, dvojtečka i závorka. Apoziční členy se na tyto spojovací výrazy zavěšují jako členy závislé.

Anotační manuál pro PZK byl formulován souběžně se syntakticko-analytickou anotací PZK a ucelenější podobu dostal až v době, kdy anotace byly v podstatě ukončeny. Bezprostředně po sepsání manuálu nezbýval prostor na stádium připomínek, které by umožnilo některá pravidla lépe formulovat, nevhodné umístění některých pokynů korigovat nebo chybějící pravidla do manuálu doplnit. Potřebnou revizí tedy anotační manuál prošel až při anotaci ČAKu.

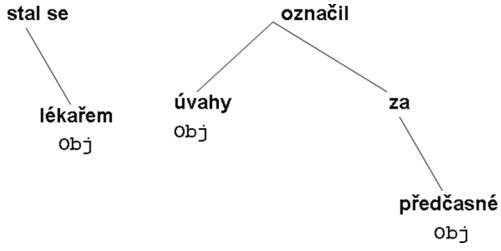
Vytvoření pravidel manuálu je vždy vedeno snahou popsat všechny jazykové jevy, které se v analyzovaných textech mohou vyskytnout. Avšak vzhledem k povaze přirozeného jazyka není možné předvídat v pravidlech všechno. Zároveň každý další průchod anotovanými daty s sebou přináší hodnocení předchozích rozhodnutí, buď jejich schválení, nebo jejich opravu, což se promítá i do anotačních pravidel. Tímto vysvětlením směřujeme k tomu, že nedostatky v manuálu není možné chápat jako prohřešky, ale jako přirozené důsledky formalizace různorodostí přirozeného jazyka. Pro ilustraci uvádíme příklady několika nedostatků manuálu, které byly při anotaci ČAKu zrevidovány:

(a) Pokyny k anotování doplňku: Podle Šmilauera se rozlišuje doplněk určující (nevazebný), např. *chlapec ležel nemocen*, *přiběhl bos*, a doplněk doplňující (vazebný), např. *stal se lékařem*, *označil úvahy za předčasné*. Pojetí doplňku v manuálu PZK je užší. Za doplněk je považován pouze Šmilauerův doplněk určující. Podle zavěšení je označován buď atributem Atv (pokud jsou přítomny oba členy, k nímž se tento doplněk vztahuje, je doplněk zavěšen z technických důvodů na jméno), nebo atributem AtvV (pokud je jmenný řídící člen elidován, zavěsí se doplněk na svůj druhý řídící člen, jímž je sloveso). Šmilauerův doplněk doplňující (vazební) je v manuálu považován za objekt, je označován atributem Obj a je zavěšen stejně jako ostatní objekty pouze na sloveso.

V manuálu je pro ilustraci uveden seznam sponových a polosponových sloves, po nichž je třeba doplněk doplňující (vyjádřený formami: instrumentál, *za* + akuzativ, *jako* + nominativ) považovat za objekt (*stát se*, *jmenovat*, *pokládat*, *označit*, *připadat*, *zdát se*, *shledat*, *zůstávat*, *jevit se*). Během anotace textů ČAKu se objevila analogická, v seznamu neuvedená, slovesa, u nichž anotátor váhá, zda má považovat jejich rozvití za objekt nebo za doplněk. Například slovesa *považovat* a *prohlásit*: *Pokládat něco za diskriminaci* (Obj) a *považovat něco za diskriminaci* (?). Analogický příklad sloves *označit a prohlásit*: *Označit návrh za špatný* (Obj) a *prohlásit návrh za špatný* (?). Aby se vyloučila nejistota anotátora, je manuál v tomto směru rozšířen, případně opraven.



Obr. 2: Dva způsoby anotace nevazebného doplňku v PZK



Obr. 3: Dva způsoby anotace vazebného doplňku v PZK

To znamená, že je jednoznačně řečeno, že každý Šmilauerův doplněk doplňující bude anotován jako objekt.

(b) V původním manuálu nebylo výslovně řečeno, co je řídícím uzlem pro rematizátor u substantiva s nepravou předložkou (*pouze na úkor studia*) nebo pro částici *se*/*si* u složeného predikátu (*začíná se rozednívat*, *musí si uvědomit*, *začíná si zvykat*).

8. Ruční kontrola anotací a vzniklé problémy

Kontrola anotací je důležitou, časově náročnou součástí každého anotačního procesu. Aby bylo dosaženo co nejvyšší konzistence zpracovaných dat, má za úkol jednak opravit případné chyby (velkou část z nich detekovaných automatickými kontrolními procedurami; Štěpánek, 2006), jednak zajistit jednotnost přístupu při řešení jevů homonymních, diskutabilních a sporných. Kontrolní opravy se týkají jak podoby stromové struktury – reprezentace stromu, tak i přiřazení analytických funkcí.

Anotace Českého akademického korpusu byla specifická tím, že na ní pracovali slovenští anotátoři. Jakkoli si je čeština se slovenštinou blízká, oba jazyky se přece jen liší a není překvapením, že odlišnosti najdeme také v syntaktickém pojetí obou jazyků. Odborné jazykové znalosti a jazyková intuice slovenských anotátorek tedy odpovídaly do jisté míry odlišnému lingvistickému přístupu danému slovenskou syntaktickou tradicí (např. v pracích J. Kačaly, 1998, J. Oravce a E. Bajzíkové, 1982, J. Nižníkové a M. Sokolové, 1998).

Odlišnou jazykovou intuicí by bylo možné vysvětlit například značný počet neoprávněně přidaných nevlastních předložek (např. *v kontextu s*…, *z oboru*…), které nejsou uvedeny ani v seznamu nepravých předložek anotačního manuálu, ani ve Slovníku spisovného jazyka českého (Havránek, 1989). Zároveň je třeba přihlédnout k tomu, že problematika nevlastních předložek je i pro rodilé mluvčí velmi složitá. Také odlišnosti ve vymezení některých gramatických kategorií v obou jazycích ovlivnily anotaci různých jevů. Slovenští anotátoři pracovali například s poněkud jiným pojetím doplňku. Stejné vysvětlení by mohlo mít také jejich počáteční váhání nad zavěšením netypické české spojky *-li* (*-li* má ve slovenštině jiný ekvivalent – samostatnou spojku *ak*). Podobný důvod mají pravděpodobně i chyby ve větných strukturách se spojkou *neboť* (slovenští anotátoři ji chápali jako spojku podřadicí, zatímco česká tradice ji považuje

za spojku souřadicí). Občasné váhání při určování složeného přísudku lze vysvětlit tím, že klasická slovenská syntax pracuje s větším počtem neplnovýznamových pomocných sloves, které jsou součástí složeného přísudku.

8.1. *Chybné porozumění a nedostatečné znalosti*

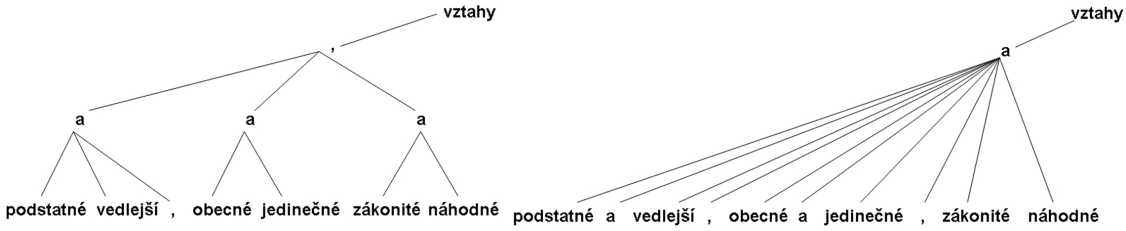
Opravují se především chyby, které vznikly nedostatečným porozuměním textu. Například ve větě *Prezident České republiky propůjčil mistru sportu, nadporučíku Fr. Venclovskému vyznamenání Za statečnost, za prokázanou osobní odvahu a příkladnou bojovnost při plavbě kanálem La Manche* zavěsili anotátoři členy *Za statečnost*, *za odvahu a bojovnost* jako koordinované rozvíjející přívlastky ke slovu *vyznamenání.* Do názvu zmíněného vyznamenání však patří jen spojení *Za statečnost,* a tudíž přívlastkem (Atr) ke slovu *vyznamenání* je pouze tento rozvíjející člen, zatímco spojení *za odvahu a bojovnost* jsou adverbiální určení (Adv) k predikátu *propůjčil*, tzn. *propůjčil (komu co) za odvahu a bojovnost.*

Někdy dochází k chybám také proto, že anotátor nemá dostatečné znalosti potřebné k jednoznačnému porozumění textu. Například ve větě *Pětadvacetiletý knihkupec Václav Klement a strojník Václav Laurin z nedalekého Turnova se roku [] dohodli na společné výrobě jízdních kol.* je pro správné zavěšení větného členu *z Turnova* třeba vědět, zda z Turnova pocházeli oba jmenovaní. Pokud Klement a Laurin oba z Turnova byli, zavěšuje se spojení *z Turnova* jako uzel s funkcí Atr na koordinační spojku *a*, ale pokud z Turnova byl jen Václav Laurin, bude spojení *z Turnova* zavěšeno sice opět s funkcí Atr, ale tentokrát jen na člen *Laurin.*

8.2. *Koordinace*

Kontrola anotace koriguje (odstraňuje) nejednotnost při budování stromových struktur. Tato nejednotnost nastává v důsledku individuální nebo doslovné aplikace pravidel manuálu. Příkladem může být stromová struktura, která obsahuje více koordinačních vztahů. V manuálu (Hajič et al., 2004) je uvedeno, že má-li koordinace více členů (a tedy i více koordinačních znaků), dostává funkci Coord koordinační znak, který je v dané konstrukci nejvíce vpravo. Ostatní koordinační znaky se zavěšují jako jeho dceřiné uzly. Toto pravidlo platí, pokud jsou všechny členy koordinačního spojení rovnoprávné.

Může ale nastat případ, kdy autor textu záměrně rozděluje koordinační členy do určitých seskupení. V tomto případě vzniká otázka, zda má anotátor striktně dodržovat pravidlo manuálu a všechny členy koordinace zavěsit na poslední koordinační znak nebo zda má postupovat podle svého jazykového cítění a při zavěšení sledovat autorovo členění koordinace. Například ve větě *…je nutno odhalit vzájemné souvislosti, jež pomohou rozlišit podstatné a vedlejší, obecné a jedinečné, zákonité a náhodné vztahy činitelů zkoumaného procesu …* se nabízí dvojí řešení. Buď se čárka před slovem *zákonité* zavěsí jako koordinační znak na slovo *vztahy* a jednotlivá tři koordinační spojení vytvoří dceřiné uzly koordinační čárky, nebo se všechny koordinační členy zavěsí na poslední *a* ve větě – viz obrázek níže.



Obr. 4: Dvojí možnost anotace koordinace

Podobné odchylky od pravidla o koordinaci mohou nastat také v případě větné koordinace. Existují souvětí, v nichž je subjekt nebo predikát společný jen pro část účastníků koordinace. V těchto souvětích je třeba opět porušit pravidlo a zavěsit jednotlivé koordinační celky tak, aby stromová struktura odpovídala sémantickému členění koordinačních částí souvětí. Dokladem tohoto jevu je například následující souvětí: *První stupeň jakosti dostalo [] výrobků, do druhého stupně bylo zařazeno [], do třetího [] výrobků*. Toto souvětí se skládá ze tří klauzí, přičemž jen dvě poslední klauze jsou vázány společným predikátem. Poslední koordinační znak souvětí proto nemůže být na vrcholu stromové struktury, jak by předepisovalo doslovné chápání pravidla manuálu.

Domníváme se, že uvedené příklady jsou jasným dokladem toho, že koordinační členy lze seskupovat různě, tj. například ((A a B) nebo C), nebo (A a (B nebo C)), a že pravidlo o řídícím uzlu koordinace je třeba zpřesnit v tom smyslu, že se vztahuje vždy jen na jednu úroveň koordinace poté, co celá koordinace je (někdy až na více úrovní, tj. hierarchicky) rozdělena na sémanticky související části.

8.3. *Homonymní konstrukce a konstrukce se spornou interpretací*

Při kontrolách anotace je také prostor pro korekce homonymních, různě interpretovatelných a sporných konstrukcí. V těchto případech nelze mluvit vysloveně o chybách, protože jde spíše o různé pohledy na daný jazykový jev. V klasických mluvnicích jsou k takovým jevům uváděny komentáře. Při anotaci PZK se však musí vybrat jen jedno řešení, které, pokud je to možné, je ve shodě s pravidly manuálu.

8.3.1. *Hranice mezi objektem a adverbiálním určením (Obj/Adv)*

Zdrojem nejednoznačností při výběru analytické funkce je různý pohled na rozlišení objektu (Obj) a adverbiálního určení (Adv). Různé mluvnice vedou hranici mezi objektem a adverbiálním určením rozdílně a neexistují kritéria, podle nichž by se dalo zvolené řešení prohlásit za definitivně správné. Například původ a výsledek děje jsou Šmilauerem chápány jako zvláštní typy příslovečných určení, kdežto my tato určení řadíme k objektům, např. *získala informace od studentů* (Obj), *vyrostl v muže* (Obj).

V PZK se při rozhodování, zda vybrat analytickou funkci Obj nebo Adv, opíráme o práce J. Panevové (1980, 1998). V nejistých případech mohli anotátoři používat jako

pomůcku valenční slovník PDT-Vallex (srov. Hajič et al., 2003; Hajič – Urešová, 2003; Urešová, 2009), který byl na základě analyzovaných textů vypracován pro tektogramatickou rovinu PZK. Tento slovník ovšem obsahuje jen data z PZK, takže pokud se v korpusu ČAK vyskytla nová slovesa, popřípadě nové významy sloves, anotátor příslušné valenční rámce v PDT-Vallexu nenalezl.

8.3.2. *Rozlišení konstrukcí stavových a dějových*

Při kontrolách anotace se ukázalo, že značné problémy dělalo anotátorům rozlišení stavu a děje, a to zvláště tehdy, když existující kritéria nevyznívala jednoznačně a kontext pro rozlišení nebyl dost prokazatelný. Zdánlivě jednoduchá struktura věty *Hrad byl vystavěn* má dvojí možnost zachycení ve stromové struktuře. Buď se důraz klade na děj (v kontextu *hrad byl vystavěn za Karla IV* jde o trpný tvar slovesa *vystavět*), nebo se důraz klade na stav (v kontextu *hrad byl tehdy už vystavěn* jde o slovesně jmenný predikát). V prvním případě funkci Pred dostává trpný tvar významového slovesa a tvar pomocného slovesa *být* je zavěšen na něm s funkcí AuxV. V druhém případě funkci Pred dostává tvar slovesa *být* a participium *vystavěn* je zavěšeno na něm s funkcí Pnom. Tento sémantický rozdíl lze zpravidla poznat na základě kontextu. Ovšem ne vždy je rozhodování mezi stavovou a dějovou interpretací struktury zřejmé a větu lze chápat obojím způsobem. Například větu *Všechna naše setkání byla naplněna vřelým přátelstvím.* lze chápat buď jako stav, tzn. jako variantu věty *Setkání byla přátelská* (*byla naplněna* lze považovat za predikát nominální), nebo jako děj, tzn. jako větu, jež lze parafrázovat jako *Vřelá přátelství* (Subj) *naplnila naše setkání* (Obj) (*byla naplněna* je pasivum slovesa *naplnit*). U těchto konstrukcí je z uvedených důvodů proto třeba počítat s nejednoznačností (nekonzistencí) v anotovaných datech. Podle našeho názoru teprve následný pohled na anotovaná data a důkladný statistický a věcný rozbor pomůže vyjasnit, nakolik je tento rozdíl opravdu sémanticky relevantní a nakolik by bylo možno k dosažení konzistence v některých nejasných případech použít konvencí.

8.3.3. *Identifikace* se *a* si *(AuxT/AuxR)*

Částice *se* může být ve shodě s manuálem ohodnocena analytickými funkcemi AuxT, AuxR nebo Obj. Částice *si* může mít funkci AuxT, Obj, Adv nebo AuxO. Analytickou funkci AuxT dostávají částice *se*, *si* tehdy, když sloveso bez nich neexistuje (*bát se*, *pospíšit si*). Analytickou funkci AuxR dostává částice *se*, jde-li o reflexivní pasivum (*tancuje se*, *píše se*, *diskutuje se*). Objektem je částice *se* např. u sloves *bránit se* (Obj), *mýt se* (Obj). Částici *si* považujeme za objekt např. u sloves *přečíst si* (Obj), *stanovit si* (Obj). Jako AuxO označujeme takové *si*, které je nadbytečné a může být vypuštěno, aniž dojde ke ztrátě smyslu nebo gramatičnosti (*jít si* (AuxO) *na výlet*), i když stylisticky může jistou informaci nést. Částice *se* a *si* se zavěšují na sloveso, k němuž patří, ať už mají jakoukoli z uvedených funkcí.

Ačkoliv kritéria přiřazování analytických funkcí částici *se* a částici *si* jsou formulována velmi zřetelně, při praktickém anotování konkrétního textu vyšlo najevo, že se

aplikují velmi obtížně. Je to zejména proto, že u velkého množství sloves vystupují tyto částice ve více funkcích a anotátor musí umět vybrat jedinou vhodnou. Například částice *se* u slovesa *soustředit* může mít následující výskyty s různým ohodnocením: *Sbírky mincí se soustředily do prvního patra* (AuxR), *Soustředil se na práci* (AuxT), *Vojska se soustředila před bránami města* (AuxT/AuxR). U posledního příkladu existuje dvojí interpretace částice *se* (a to i ve velmi podobných kontextech), což vede k nekonzistenci anotace. Podobně může dojít k různé interpretaci částice *se* u slovesa *rozšiřovat* ve větách *Podstatně se rozšiřuje pěstování karotky a petržele* (nejspíše funkce AuxR) a *Infekce se rozšiřuje z Německa* (nejspíše funkce AuxT).

8.3.4. *Hranice mezi subjektem a predikátem nominálním (Sb/Pnom)*

Při kontrole anotovaných dat se dále ukázalo, že je-li predikát vyjádřen tvary slovesa *být* (sponou), není vždy snadné rozlišit, který člen věty je subjektem (Sb) a který je predikátem nominálním (Pnom), pokud jsou oba tyto členy v nominativu. Například ve větě *Tiskárna* (Sb) *je moderní zařízení* (Pnom) anotátor pravděpodobně nezaváhá, kterému členu přiřadit kterou funkci, ale např. ve větě *Jeho plochy jsou rovnostranné trojúhelníky* stojí před otázkou, zda subjektem je člen *plochy* nebo člen *trojúhelníky*.

V mnoha případech pomáhá anotátorovi následující školní pomůcka pro rozlišení subjektu a nominální části predikátu: ten větný člen, který lze z nominativu proměnit v instrumentál, je součástí predikátu, a tudíž je ohodnocen analytickou funkcí Pnom. Tato pomůcka se však dobře aplikuje jen tehdy, jde-li o text, kterému anotátor dobře rozumí. Zejména ve specializovaných odborných textech je využití takové transformace problematické, neboť anotátorovi k posouzení takové proměny chybí porozumění obsahu. Vzhledem k tomu mu připadá reálné převést do instrumentálu jak první, tak druhý nominativ. Dokladem těchto na posouzení obtížných příkladů jsou např. následující věty: *Takové nesouměrné plochy* (Pnom/Sb?) *jsou nepravidelné mnohoúhelníky, obecné trojúhelníky, různoběžník, kosodélník*. *Krystaly kamence jsou obyčejně pravidelné osmistěny* (Pnom/Sb?). Ačkoliv klasické mluvnice připouštějí v takových případech dvojí výklad (tzn. homonymní strukturu), v korpusu je třeba se rozhodnout jen pro jedno řešení vyplývající z porozumění obsahu. Je zřejmé, že anotátoři, kteří nemohou být odborníky ve všech myslitelných oborech, se mohou v řešení lišit, a je tedy třeba počítat s jistou nekonzistencí při přidělování této funkce.

8.3.5. *Nepravé (nevlastní) předložky*

Během kontrol anotace se potvrdilo, že živé, produktivní, a proto proměnlivé jazykové jevy, jako je např. kategorie nevlastních předložek, jsou při anotaci velmi problematické. Identifikace nevlastních předložek a jejich odlišení od užití nepředložkového způsobovaly anotátorům problémy a nutno podotknout, že anotování tohoto jevu patří k nejméně konzistentním.

Při kontrolách anotace se projevila dvojí mylná tendence ve vymezení nepravých předložek. Buď se skutečná nepravá předložka nepovažuje za předložku (např. anotace

spojení *v oblasti* je nesprávně: *v* (AuvP) *oblasti* (Adv) *vědy* (Atr)), nebo se předložkové sousloví neoprávněně považuje za jednu (nepravou) předložku (např. *v asociaci s…, v jednotě s…*). Anotátoři měli k dispozici speciální přílohu manuálu, v níž je výčtem uveden seznam sousloví, která se za nevlastní předložku považují. Ovšem jako většina podobných seznamů je i tento seznam neúplný a anotátoři museli svůj názor ověřovat i v jiných normativních příručkách českého jazyka.

Poměrně vysoká nejednoznačnost anotace nepravých předložek odráží skutečnost, že významy předložkových sousloví, která se za předložku už považují, a předložkových sousloví, která mezi předložky ještě nepatří, jsou si natolik blízké, že k záměně může dojít velmi snadno. Zmíněný problém ilustrují následující příklady spojení (nejprve je uvedeno spojení považované manuálem za nepravou předložku, za ním v závorce následuje sousloví, které v seznamu předložek není): *bez zřetele k okolnostem* (*bez zřetele na vliv prostředí*), *úměrně ke kvalifikaci* (*úměrně schopnostem*), *v oblasti vědy* (*z oblasti pohledávek*), *ve spolupráci s klubem* (*v součinnosti se spolkem*), *společně s problémy* (*současně s vývojem*).

I kdybychom přílohu nevlastních předložek v manuálu znova aktualizovali, domníváme se, že bychom se omylům v anotaci nepravých předložek opět nevyhnuli. Oddělit předložkové sousloví od nepředložkového užití (např. ve spojeních *z hlediska praxe*, *v oblasti vědy* jde o nevlastní předložku, zatímco *z hlediska historického*, *v oblasti povodní* nikoliv) je totiž velmi obtížné, a to nejen proto, že samotná kategorie nevlastních předložek není v české lingvistice definitivně vymezena. Při anotaci tohoto lingvistického jevu totiž navíc často selhává i intuice rodilého mluvčího.

9. Závěr

Syntaktická proměna Českého akademického korpusu má za sebou dvě třetiny cesty, které reprezentuje dokončená anotace psaných textů. Anotace mluvených textů zůstává otevřená. Stejně tak zůstává otevřena otázka atributů prázdných uzlů a otázka oprav již odhalených chyb.

Myšlenka projektu tzv. rekonstrukce mluvené řeči (Mikulová, 2008), kterou iniciovaly výsledky systémů automatického rozpoznávání řeči, se jeví jako vhodné řešení pro doplnění syntaktických anotací mluvených textů ČAKu. Rekonstrukce probíhá tak, že text se nejprve převede na gramaticky správnou větu a teprve pak se syntakticky anotuje podle již existujících pravidel pro psané texty, čímž se eliminují problémy uvedené v odd. 5. Do systému čtyř rovin popsaných v odd. 2 je pro rekonstruované texty doplněna mezi roviny lexikální a morfologickou další rovina, rovněž charakteru lexikálního. Rekonstrukce mluvených textů se pak uskutečňuje přechodem z lexikální roviny do roviny nově doplněné. Díky odkazům mezi jednotkami rovin existuje přehled o tom, jak se z původního řetězce zformovala gramaticky správná věta, a žádná pro uživatele podstatná informace se tak neztrácí, neboť syntaktickou strukturu věty lze zobrazit (a to doslova, s doprovodnými grafickými vizuálními nástroji)[9] jak nad zre-

[9] Viz online na adrese: <http://ufal.mff.cuni.cz/pdt2.0/doc/tools/tred/index.html>.

konstruovanou větou, tak i nad větou původní. Vrátíme-li se k větě uvedené v odd. 5, její rekonstruovaná podoba by mohla vypadat následovně (tučně jsou zvýrazněna doplněná nebo odstraněná nebo změněná slova, příp. písmena): *A **(to)** jsou trošku **rozdílné**,**(jedna je,)** jedna má světlou budovu a druhá má tmavou budovu.**(, ony)** Jsou umístěny **(v jednom,)** v jednom areále, ale **(ta,)** to centrum **(, patřilo té,)** bylo **(to)** v bloku Univerzity vlámské**(,)** a já jsem se ptala na **(univerzitě, na, v)** Univerzitě svobodné. **(, že, no a to přeci oni nevědí, to nanejvýš, to prostě jedině,) K**dyž je to Univerzita vlámská, tak o tom oni přece nemohou nic vědět **(, a nic).***

Pražský závislostní korpus obsahuje 68 tisíc vět syntakticky anotovaných a psané texty ČAKu jich obsahují 24 tisíc. Přičleněním ČAKu k PZK se celkový objem syntakticky anotovaných českých vět významně navýší, čímž je k dispozici více materiálu k vyhledávání různých jazykových jevů a více trénovacích dat pro metody strojového učení. Nicméně otázkou zůstává, máme-li k PZK přičlenit psané i mluvené texty, nebo máme-li mluvené texty ponechat stranou jako materiál pro další studium rozdílností psaných a mluvených projevů.

Syntaktická anotace ČAKu přispěla k revizi původního analytického manuálu, což je přínosné i z toho pohledu, že se nyní dá chápat jako počítačová učebnice současné české syntaxe.

Kromě syntaktické anotace mluvených textů ČAKu nás v blízké době projekt ruční syntaktické anotace dalších českých textů nečeká. Avšak pokud si dobře vzpomínáme, něco podobného jsme si mysleli také v okamžiku dokončení syntaktických anotací PZK.

Poděkování

Konverze Českého akademického korpusu probíhala za podpory těchto projektů: projekt Grantové agentury Akademie věd České republiky č. 1ET101120413, projekty Ministerstva školství, mládeže a tělovýchovy České republiky č. MSM-0021620838, ME 838 a projekt Grantové agentury Univerzity Karlovy v Praze č. GAUK 52408/2008. Srdečně děkujeme Jarmile Panevové a Evě Hajičové za jejich poznámky a komentáře, které nás vedly k vylepšování příspěvku.

Příloha: Analytické funkce v PZK

analytická funkce	popis
Pred	Predikát, resp. uzel, který nezávisí na jiném uzlu; věší se na kořen stromu.
Pnom	Predikát nominální, resp. jmenná část přísudku se sponou *být*.
AuxC	Spojka (podřadicí).
AuxK	Koncová interpunkce věty.
Sb	Subjekt (podmět).
AuxV	Pomocné sloveso *být*.
AuxO	Nadbytečný (odkazovací, emotivní) element.
ExD	Náhradní funkce pro technické hrany vedoucí místo od elidovaného členu k „pseudořídícímu“ slovu nebo pro hlavní člen věty bez predikátu.
Obj	Objekt (předmět).
Coord	Koordinační uzel (souřadné spojení).
AuxZ	Zdůrazňovací slovo.
AtrAtr	Řídicím slovem atributu může být díky strukturní víceznačnosti kterékoli z bezprostředně předcházejících (syntaktických) substantiv.

Adv	Adverbiale (příslovečné určení bez dalšího rozlišení).
Apos	Apozice (hlavní uzel).
AuxX	Čárka (ne však nositel koordinace).
AtrAdv	Strukturní víceznačnost mezi závislostí adverbiální (příslovečnou) a adnominální (zavěšení na substantivum) bez sémantických důsledků.
Atv	Doplněk (jen tzv. určující), technicky zavěšen na neslovesném členu.
AuxT	Zvratné *se*, neoddělitelné *se* – reflexivum tantum.
AuxG	Jiné grafické symboly, které neukončují větu.
AdvAtr	Jako AtrAdv, ale s opačnou preferencí.
AuxR	Zvratné *se*, které není Obj ani AuxT (tvoří pasivum reflexivní).
AuxY	Příslovce a částice, které nelze zařadit jinam.
AtrObj	Strukturní víceznačnost mezi závislostí objektovou a adnominální (zavěšení na substantivum) bez sémantických důsledků.
Atr	Atribut (přívlastek).
AuxS	Kořen stromu (#).
ObjAtr	Jako AtrObj, ale s opačnou preferencí.

LITERATURA

ATWELL, E. – LEECH, G. – GARSIDE, R. (1984): Analysis of the LOB Corpus: progress and prospects. In: J. Aarts – W. Meijs (eds.), *Corpus Linguistics.* Amsterdam: Rodopi, s. 41–52.

FRANCIS, W. N. – KUCERA, H. (1979): *Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers*. Providence: Department of Linguistics, Brown University.

HAJIČ, J. – PANEVOVÁ, J. – BURÁŇOVÁ, E. – URešOVÁ, Z. – BÉMOVÁ, A. – ŠTĚPÁNEK, J. – PAJAS, P. – KÁRNÍK, J. (2004): *Anotace na analytické rovině: Návod pro anotátory*. Technická zpráva, TR-2004-23. Praha: ÚFAL/CKL MFF UK.

HAJIČ, J. – PANEVOVÁ, J. – HAJIČOVÁ, E. – SGALL, P. – PAJAS, P. – ŠTĚPÁNEK, J. – HAVELKA, J. – MIKULOVÁ, M. – ŽABOKRTSKÝ, Z. – ŠEVČÍKOVÁ-RAZÍMOVÁ, M. (2006): *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

HAJIČ, J. – PANEVOVÁ, J. – URešOVÁ, Z. – BÉMOVÁ, A. – KOLÁŘOVÁ, V. (2003): PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*. Vaxjo: Vaxjo University Press, s. 57–68.

HAJIČ, J. – URešOVÁ, Z. (2003): Linguistic annotation: from links to cross-layer lexicons. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*. Vaxjo: Vaxjo University Press, s. 69–80.

HAVRÁNEK, B. (ed.) (1989): *Slovník spisovného jazyka českého, 1–8*. Vyd. 2. Praha: Academia.

HLADKÁ, B. – KRÁLÍK, J. (2006): Proměna Českého akademického korpusu. *Slovo a slovesnost*, 67, s. 179–194.

HLADKÁ, B. – URešOVÁ, Z. (2009): Syntactic annotation of transcriptions in the Czech Academic Corpus: then and now. In: M. Mahlberg – V. González-Díaz – C. Smith (eds.), *Proceedings of the Corpus Linguistics Conference (CL2009), University of Liverpool, UK, 20–23 July 2009*. University of Liverpool. Available online at: <http://ucrel.lancs.ac.uk/publications/cl2009/>.

KAČALA, J. (1998): *Syntaktický systém jazyka*. Pezinok: Formát.

MCDONALD, R. – PEREIRA, F. – RIBAROV, K. – HAJIČ, J. (2005): Non-projective dependency parsing using spanning tree algorithms. In: *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, s. 523–530.

MIKULOVÁ, M. (2008): *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny: Manuál pro anotátory*. Technická zpráva, TR-2008-38. Praha: ÚFAL MFF UK.

NILSSON, J. – HALL, J. – NIVRE, J. (2005): MAMBA meets TIGER: Reconstructing a treebank from antiquity. In: *Proceedings of NODALIDA 2005 Special Session on Treebanks for Spoken and Discourse* (= Copenhagen Studies in Language, 32). Joensuu, s. 119–132.

NIŽNÍKOVÁ, J. – SOKOLOVÁ, M., a kol. (1998): *Valenčný slovník slovenských slovies*. Prešov: Prešovská univerzita, Filozofická fakulta.

ORAVEC, J. – BAJZÍKOVÁ, E. (1982): *Súčasný slovenský jazyk: Syntax*. Bratislava: Slovenské pedagogické nakľadatelstvo.

PANEVOVÁ, J. (1980): *Formy a funkce ve stavbě české věty.* Praha: Academia.

PANEVOVÁ, J. (1998): Ještě k teorii valence. *Slovo a slovesnost*, 59, s. 1–14.

RIBAROV, K. – BĚMOVÁ, A. – HLADKÁ, B. (2006): When a statistically oriented parser was more efficient than a linguist: A case of treebank conversion. *The Prague Bulletin of Mathematical Linguistics*, 86, s. 21–28.

ŠMILAUER, V. (1947): *Novočeská skladba*. Praha: Státní pedagogické nakladatelství.

ŠTĚPÁNEK, J. (2006): *Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat)*. Nepublikovaná disertační práce. Praha: Matematicko-fyzikální fakulta Univerzity Karlovy.

TELEMAN, U. (1974): *Manual för grammatisk beskrivning av talad och skriven svenska*. Lund: Studentlitteratur.

TĚŠITELOVÁ, M. (ed.) (1983): *Frekvenční slovník češtiny věcného stylu*. Praha: Ústav pro jazyk český ČSAV.

TĚŠITELOVÁ, M. (1984): Kvantitativní analýza češtiny s pomocí moderní výpočetní techniky. *Naše řeč*, 67, s. 47–50.

TĚŠITELOVÁ, M. – UHLÍŘOVÁ, L. – KRÁLÍK, J. (1984): K automatickému zpracování textu při kvantitativní analýze přirozeného (českého) jazyka. *Slovo a slovesnost*, 45, s. 145–150.

UREŠOVÁ, Z. (2009): Building the PDT-Vallex valency lexicon. In: M. Mahlberg – V. González-Díaz – C. Smith (eds.), *Proceedings of the Corpus Linguistics Conference (CL2009), University of Liverpool, UK, 20–23 July 2009*. University of Liverpool. Available online at: <http://ucrel.lancs.ac.uk/publications/cl2009/>.

VIDOVÁ HLADKÁ, B. – HAJIČ, J. – HANA, J. – HLAVÁČOVÁ, J. – MÍROVSKÝ, J. – VOTRUBEC, J. (2007): *Český akademický korpus 1.0 – Průvodce*. Praha: Karolinum.

VIDOVÁ HLADKÁ, B. – HAJIČ, J. – HANA, J. – HLAVÁČOVÁ, J. – MÍROVSKÝ, J. – RAAB, J. (2008a): *Czech Academic Corpus 2.0 – CD-ROM*. Cat. num. LDC2008T22. Philadelphia: Linguistic Data Consortium.

VIDOVÁ HLADKÁ, B. – HAJIČ, J. – HANA, J. – HLAVÁČOVÁ, J. – MÍROVSKÝ, J. – RAAB, J. (2008b): Czech Academic Corpus 2.0 – Guide. *The Prague Bulletin of Mathematical Linguistics*, 89, s. 41–96.

Ústav formální a aplikované lingvistiky MFF UK
Malostranské náměstí 25, 118 00 Praha 1
<hladka@ufal.mff.cuni.cz>
<uresova@ufal.mff.cuni.cz>