Konverze syntaktických anotací
Českého akademického korpusu - Jaké to bylo?

Alla Bémová, Zdeňka Urešová

prosinec 2008

1. Podoba analyzovaného textu
2. Analýza mluvených projevů
3. Organizace anotování
4. Anotátoři
5. Anotační pravidla – manuál
6. Kontrola anotací

---

1. Podoba analyzovaného textu

Konverze syntaktických anotací Českého akademického korpusu (ČAK) do podoby kompatibilní s Pražským závislostním korpusem (PZK) musela řešit řadu problémů, protože výchozí data určená ke konverzní proceduře neodpovídala požadavkům a pravidlům pro anotaci analytické roviny PZK.

Podrobně o ČAK, jeho historii, anotacích, konverzi morfologických dat atd. referuje článek v SaS (Hladká, Králík, 2006[1]), zde jenom stručně zopakujeme některé podstatné skutečnosti. Cílem akademického korpusu, na němž se pracovalo v 70. - 80. letech v Ústavu pro jazyk český, bylo soustředit dostatečně reprezentativní textový materiál 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 vypuštěny

[1] Barbora Hladká, Jan Králík: *Proměny Českého akademického korpusu*. Slovo a slovesnost, 67:179-194, 2006.

(např. ciferné výrazy a interpunkce) nebo se jim nevěnovala pozornost (např. se nerozlišovala malá a velká písmena).

Tyto pro původní akademický korpus nepodstatné informace jsou však pro analytickou rovinu PZK nepostradatelné[2], a proto musely být do vstupního textu ČAK, určeného pro analýzu v rámci PZK, doplněny. Reprezentace větné struktury na analytické rovině má totiž podobu stromu (orientovaného acyklického grafu s jedním kořenem), v němž je nutné všechna slova věty zachovat. Nesmí se ztratit ani interpunkční symboly, které jsou ve výsledné analytické struktuře stejně jako všechna slova ohodnoceny příslušnou analytickou funkcí.

Jelikož se původní texty, které byly podkladem pro ČAK, nezachovaly, musely se dokumenty rekonstruovat na základě existujících anotací. To znamenalo všechny texty přečíst a provést nutné korektury: kontrolu velkých/malých písmen, označení nesrozumitelného textu, označení místa, kde chybí ciferný výraz, označení místa, kde chybí interpunkce.

Na rekonstrukci textů a jejich korekturách pracovali studenti filologických oborů Filozofické fakulty. Lidský faktor vždy přináší určitý druh nedůslednosti, a zpracování tak velkého množství materiálu se neobejde bez chyb. Z tohoto pohledu je třeba k datům určeným ke konverzi přistupovat.

[2] Texty pro anotování na všech třech úrovních PZK (morfologické, analytické a tektogramatické) byly převzaty z Českého národního korpusu a zachovávají jeho formát. Jsou rozděleny na slova (slovní tvary), věty a odstavce, explicitně je označena interpunkce a částečně zachována grafická informace z původního textu. Pracuje se i s čísly psanými číslicemi, rozlišují se velká a malá písmena s příslušnou gramatickou rolí.

1.1. Charakter chyb vstupního textu

Vstupní data ČAK obsahovala z hlediska PZK následující nedostatky:

A. Chybějící výrazy ve větě:

Například věta *Mezi řekami [ ] a Tigridem žily [ ] [ ] lety dva odlišné [ ].* obsahuje čtyři chybějící výrazy. Aby se analytická funkce nemusela přiřazovat prázdnému uzlu[3], má anotátor možnost předpokládané chybějící slovo (podle vlastního uvážení) doplnit do atributu `guessed_form`, pro tyto účely doplněného do vnitřní reprezentace ČAK. Místo konkrétního slovního tvaru může anotátor doplnit obecnější informaci, např. že jde o substantivum, adjektivum, nebo o infinitní tvar slovesa. Rekonstruovaná podoba výše uvedené věty by mohla vypadat takto: *Mezi řekami Eufratem a Tigridem žily před několika/mnoha/xy lety dva odlišné národy/kmeny*.

Doplňování chybějících slov je žádoucí také v případech, kdy nám pomůže odstranit případnou homonymii věty. Např. 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*).

[3] K terminologii *uzel* apod. viz manuál pro anotování analytické roviny PZK: Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, Alevtina Bémová, Jan Štěpánek, Petr Pajas, Jiří Kárník: Anotace na analytické rovině. Návod pro anotátory, In: *UFAL/CKL technical report*, ÚFAL/CKL MFF UK, Prague, TR-2004-23. 2004.

Svá specifika mají rekonstrukce úseků textu s cifernými výrazy. Po číselných výrazech vyjádřených ciframi zůstává v textu většinou jen stopa v podobě znaku *[]*. Problém při doplňování chybějících ciferných výrazů vzniká proto, že takové výrazy zachycujeme ve stromové struktuře věty na analytické rovině PZK 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 něm 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 zavěšení 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 prázdný uzel.

B. Přebývající výrazy ve větě

Některé věty naopak obsahovaly výrazy navíc. 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. Rekonstruovaná věta by pak vypadala následovně: *Rozměrné*

kamenné bloky těchto staveb…jsou dokladem zručnosti a značné vyspělosti lidí této doby…

C. Neoprávněné rozdělení věty

Z hlediska kompatibility textů ČAK a PZK bylo třeba zkontrolovat také formát a grafickou podobu vět. Ve vstupních textech docházelo k neoprávněnému rozdělování vět. Např. 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 vyjadřující přináležitost obou strukturních částí dané konstrukce.

D. Nadbytečná interpunkce

Podobně jako se v textech objevovaly nadbytečné výrazy, najdeme v nich i nadbytečnou interpunkci, např. čárky.
Např. ve větě *Milia lze otevřít jemným skalpelem a vytlačit, obsah hydradenomy, je nutno odstranit většinou chirurgicky* je čárka po slově *hydradenomy* nadbytečná, 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: jde o syntagma ze dvou slov *politicky výchovná*.

Nutno dodat, že ve vztahu k celkovému množství zpracovaného materiálu není množství nesrovnalosti různého druhu nijak podstatné, nicméně s možnými chybnými strukturami vstupních dat se musí počítat.

2. Analýza mluvených projevů

Vážný problém při anotaci dat pro analytickou rovinu představují mluvené projevy, jejichž přepisy tvoří v ČAK zhruba třetinu celkového objemu textů. Jejich struktura 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 pro anotaci analytické roviny nepočítají s anotací mluvených projevů, prostředky pro zachycení specifických jevů mluvené řeči chybí. Na analytické rovině neumíme zachytit např. bezdůvodné opakování stejného slova nebo koktání.

V následujícím příkladovém textu 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 patří:

- nedořečené věty (fragmenty), elipsy: „*A to je 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: „*ony jsou umístěny v jednom…*“, „*univerzitě, na, v Univerzitě svobodné,…*“
- nadbytečná deiktická slova: „*…ale ta, to centrum…*“
- intonační výplně: „*no…*“

- české otázky presumptivní (anglické 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 anakolut: *„…přeci nemohu nic vědět, a nic.*"

Jak již bylo uvedeno výše, pro adekvátní zachycení těchto typických rysů mluveného projevu jsou na analytické rovině PZK jen velmi omezené prostředky. Např. slova v roli intonačních výplní by mohla být ohodnocena analytickou funkcí *AuxO,* která se používá pro nadbytečný (odkazovací nebo emotivní) element. Většina prvků ústního projevu by však musela dostat analytickou funkci *ExD* (*Ex-Dependent*), která pouze upozorňuje na to, že se jedná o strukturu neúplnou, tj. o elipsu, v níž chybí řídící slovo. Nic dalšího o struktuře tato analytická funkce nevypovídá.

Z uvedených důvodů nejsou texty mluvených projevů v ČAK 2.0 syntakticky anotovány. Mluvené texty (nikoli však texty ČAK, ke kterým není možné využít audio, resp. originální nahrávky) se zpracovávají jiným způsobem v rámci projektů zabývajících se tzv. rekonstrukcí mluvené řeči (projekty PIRE[4] a Companions[5]). V systému „rekonstrukce mluvené řeči" se text nejprve převede na gramaticky správnou podobu a teprve pak se syntakticky anotuje, čímž se vyhneme problémům uvedeným výše v této kapitole.

[4] http://www.clsp.jhu.edu/research/pire
[5] http://www.companions-project.org

3. Organizace anotování

Anotátoři dostali k anotování texty ČAK po automatické proceduře[6], která provedla syntaktickou analýzu včetně klasifikace větných členů. Nejprve proběhla první testovací fáze anotování, která měla jednak odhalit, zda jsou takto připravená data vhodná k anotování, tj. jak kvalitní výstup procedury je, a jednak zjistit, zda máme vhodné anotátory.

Dva anotátoři zpracovávali paralelně stejné soubory, které pak speciální program porovnával a diskrepance mezi anotátory opravoval třetí nezávislý anotátor. Přínos tohoto procesu není jednoznačný. Jisté je, že porovnání krom jiného upozornilo na nedostatky či nepřesnosti manuálu a na homonymní jevy.

Tento způsob anotování byl ovšem časově velmi náročný, takže se od něho po zacvičení anotátorů muselo upustit. Anotace dále probíhaly tak, že anotátoři zpracovávali každý jiná data. Všechny anotované soubory pak kontroloval jiný anotátor.

4. Anotátoři

Práci anotátora považujeme za náročnou z několika hledisek :

1. Vyžaduje určitou úroveň odborných znalostí, a to jak z pohledu morfologie tak i syntaxe. Předpokládá jistou zkušenost s větným rozborem, tedy schopnost analyzovat větu tak, aby vyloučením jevů vedlejších a vymezením podstatných

[6] Automatická procedura je označována jako parser a pro ČAK byl použit parser Ryana McDonalda: Ryan McDonald, Fernando Pereira, Kiril Ribarov, Jan Hajič. **Non-Projective Dependency Parsing using Spanning Tree Algorithms**. In *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP)*. Vancouver, BC, Canada, Oct. 6-8: Association of Computational Linguistics, 2005. s. 523-530.

vznikl jasný obraz její vnitřní struktury. K tomu je vhodné, aby anotátor měl náležitou jazykovou intuici. Takové požadavky by nejlépe splňovali zejména studenti jazykovědných oborů vysokých škol.

2. Vyžaduje zvládnutí pravidel anotačního manuálu. Předcházející odborná výchova anotátorů se může poněkud lišit od konvencí a pravidel přijatých pro anotování na analytické rovině PZK. Ostatně i jednotlivé 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, přizpůsobit je pokynům manuálu. Naráží pak na následující překážky : (i) jevy homonymní, (ii) možnost různé interpretace stejného pravidla, (iii) jevy, které z různých důvodů nejsou v manuálu adekvátně zachyceny.

3. Vyžaduje vysokou míru soustředění. Samotný proces anotování je náročný jak na čas, tak i na pozornost. Anotátor sleduje jednak správnost stromové struktury a jednak správnost přiřazení analytických funkcí jednotlivým uzlům. Porozumění textu z neznámého oboru vyžaduje mnohdy několikeré čtení analyzované věty. Časově náročná je anotace syntaktické struktury dlouhých vět (vyskytl se strom, který měl přes 200 uzlů!). Neméně časově náročné je i anotování vět s elipsami. Nejde zde tolik o samotné porozumění obsahu věty, jako spíše o budování její stromové struktury. Věty s elipsami způsobují často neprojektivitu a vzdálenost mezi členem řídícím a členem závislým je někdy více než 10 uzlů.

Zároveň je třeba přiznat, že ačkoliv je práce anotátora náročná a mnohdy vyčerpávající, má více rutinní než tvůrčí charakter. Nemělo by být proto překvapením, že vznikl problém se získáváním anotátorů. Lze předpokládat, že svou roli zde sehrál i lidský faktor. Pocit frustrace z toho, že vynaložené úsilí a

péče neodpovídají konečnému výsledku (např. když se při namátkové kontrole zjistí, že v anotovaných souborech zůstaly chyby) některé anotátory odradil. Mnozí potenciální anotátoři, kteří měli požadované odborné znalosti a kteří zvládli překonat úskalí manuálu, nakonec práci vzdali ve stádiu, kdy se mělo po testovací fázi přistoupit ke skutečnému anotování.

Tento organizační problém přinesl vážné důsledky: vyvolal v harmonogramu prací časovou prodlevu, která bohužel vedla k určité uspěchanosti v závěrečné fázi projektu.

Nakonec se práce na anotování ujali anotátoři (většinou anotátorky) ze Slovenska. V rámci projektu Slovenského národního korpusu pracují na podobném projektu pro slovenštinu. Vzhledem k příbuznosti a blízkosti českého a slovenského jazyka jsou základní principy budování označkovaného závislostního korpusu velmi podobné. Proto se využití jejich odborné zkušenosti nabízelo jako vhodné řešení.

Slovenští anotátoři tedy splňují požadavek odborné připravenosti a do značné míry i požadavek znalosti českého manuálu. Zůstává však otázkou, do jaké míry je pro ně, jakožto nerodilé mluvčí, překážkou samotná čeština. Oproti předchozí slovenské generaci, která se setkávala s češtinou denně, může dnešní mladá generace osamostatněného Slovenska vnímat češtinu už jako jazyk cizí.

Některé chyby při budování českých stromových struktur nasvědčují tomu, že správné porozumění textu dělalo slovenským anotátorům občas určité problémy. Avšak celkový stav není znepokojivý. Ostatně je obtížné odlišit, zda chybně vybudovaná stromová struktura je zapříčiněna nedostatečnou znalostí jazyka, nebo je to jen důsledek nepozorného čtení textu.

Odlišnou jazykovou intuicí by bylo možné vysvětlit 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 spisovné češtiny. Avšak zároveň je třeba přihlédnout k tomu, že nevlastní předložky jsou živá kategorie, která je v pohybu, a v jejich vymezení neexistují přesné hranice, takže i rodilí mluvčí je identifikují různě.

Odlišnost mezi češtinou a slovenštinou existuje ve vymezení některých gramatických kategorií. Slovenské anotátorky např. pracují s užším pojetím v chápání doplňku, než je tomu v české tradici.

Stejné vysvětlení by mohlo mít jejich počáteční váhání se zavěšením netypické české spojky *–li (-li* má ve slovenštině jiný ekvivalent – samostatnou spojku *ak*), nebo chyby ve větných strukturách se spojkou *neboť* (slovenské anotátorky s ní zacházejí jako se spojkou podřadicí, ale v češtině se považuje za spojku souřadicí).

5. Anotační pravidla – manuál[7]

Pravidla pro anotaci na analytické rovině PZK zachovávají, pokud je to možné, tradiční pojmy českých mluvnic. Vychází se zejména ze syntaxe V. Šmilauera „Novočeská skladba“.

Šmilauerův přístup se promítá především do obdobného pojetí základních syntaktických funkcí. Seznam syntakticko-analytických funkcí, které se v PZK používají, je však mnohem širší, protože vzhledem k nepočítačovému Šmilauerovu pojetí české syntaxe jsme museli tradiční česká pravidla na mnoha

[7] Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, Alevtina Bémová, Jan Štěpánek, Petr Pajas, Jiří Kárník: Anotace na analytické rovině. Návod pro anotátory, In: *UFAL/CKL technical report*, ÚFAL/CKL MFF UK, Prague, TR-2004-23. 2004.

místech rozšířit, popřípadě přeformulovat. Přesto korpus obsahuje jevy, které nejsou ani tradičními gramatikami (zaměřenými na lidského uživatele popsatelné), 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 ovšem 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 také jinak vymezují vzájemné hranice jednotlivých větných členů, zvláště objektu, adverbiálního určení a doplňku.

Jevy z reálných textů, s nimiž se v syntaktických příručkách nepočítá :

- 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 mnohem jemnější dělení najdeme až na tektogramatické rovině PZK .

- Pro reprezentaci koordinačních a apozičních vztahů mezi jednotlivými větnými členy, případně mezi větnými strukturami, jsou k dispozici funkce pro koordinaci (*Coord*) a apozici (*Apos*), pro označení parentetických částí vět se používá přípony *_Pa*.

- Z „klasického“ repertoáru větných členů se vymyká analytická funkce *ExD* (*Ex-Dependent*), která se používá pro označení výrazu v eliptické konstrukci, připisuje se tedy tehdy, chybí-li k nějakému závislému členu jeho řídící uzel.

Pro strukturní víceznačnost (v případech, kdy *věcně* daná konstrukce ovšem vyjadřuje totéž, tj. má *stejný význam* bez ohledu na formální syntaktické vyjádření) bylo třeba vzhledem k potřebě exaktního počítačového popisu zavést tzv. kombinované funkce. Označení *AtrAdv* nebo *AdvAtr* vyjadřuje (při shodném významu) (pseudo-)nejistotu mezi závislostí adverbiální a adnominální, označení *ObjAtr* nebo *AtrObj* mezi závislostí objektovou a adnominální. Označení *AtrAtr* znamená, že za řídící slovo atributu je možné beze změny významu vybrat kterékoli z bezprostředně předcházejících syntaktických substantiv.

- Speciální analytické funkce dostávají i „slova“, 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 ve stromové struktuře jsou také čárka (*AuxX*[8]), grafické symboly (*AuxG*), koncová interpunkce (*AuxK*) a technický kořen stromu (*AuxS*). Všechna táto „slova“ mohou mít ve stromové struktuře odpovídající syntaktickou funkci.

V důsledku toho oproti tradičnímu pojetí dochází nutně k dalším četným odchylkám i při budování stromové struktury:

- Na symbolu kořenu stromu (*AuxS*) je zavěšen predikát, který řídí celou větnou strukturu. Skládá-li se z více slov, funkci *Pred* dostává jen významové sloveso. Ostatní členy predikátu jsou zavěšeny na něm, a to buď jako *AuxV* nebo *Pnom*.

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

U predikátu vyjádřeného zvratným slovesem se zavěšuje částice *se* (*AuxR* nebo *AuxT*) na toto sloveso. U predikátu složeného (*může pracovat, hodlá překládat*) se používá pro infinitiv funkce *Obj*, která označuje klasický objekt.

- Nejzásadnější změnou oproti tradičním gramatikám je jiný vztah mezi subjektem a predikátem věty. Protože při syntaktickém anotování korpusu vycházíme ze závislostní syntaxe, která charakterizuje větnou stavbu na základě valence, především valence slovesa, je sloveso v PZK chápáno jako centrum věty. Vztah mezi slovesem (predikátem) a podmětem (subjektem) vidíme jako jeden z druhů vztahu závislosti. Opouštíme tudíž tradiční pojetí základní větné dvojice. Subjekt se chápe jako závislý na predikátu a zavěšuje se na něj stejně jako ostatní rozvíjející členy predikátu.

- Pro zavěšení koordinačních a apozičních vztahů byla přijata následující konvence :
- nositelem koordinační funkce jsou spojky souřadicí, spojovací výrazy a také interpunkční znaky, např. čárka. Dostávají funkci *Coord* a jsou řídícím uzlem celého koordinačního řetězu. Koordinační členy se zavěšují na uzel s hodnotou *Coord* jako kdyby to byly členy závislé;

- apozice v PZK se ve stromové struktuře zachycuje stejně jako koordinace. Za spojovník slov ve vztahu apozice se pokládají jak slovní výrazy, tak i grafické prostředky, jako čárka, pomlčka, dvojtečka i závorka. Apoziční členy se zavěšují na spojovací výraz apozice jako členy závislé.

Pravidla manuálu byla formulována ve snaze popsat všechny jazykové jevy, které se v analyzovaných textech vyskytly. Avšak vzhledem k povaze přirozeného jazyka není možné předvídat v pravidlech všechno. V textech se mohou vyskytnout větné struktury, které nejsou v pravidlech popsány.

V takových případech je, jak již bylo uvedeno výše, anotace ponechána na jazykovém citu anotátorů a na jejich individuálním rozhodnutí.

Příkladem nedostatečně explicitně formulovaného pravidla jsou pokyny k anotování doplňku. Podle Šmilauera se rozlišuje doplněk určující (nevazebný) a doplněk doplňující (vazebný). Naše pojetí doplňku je užší, za doplněk považujeme pouze Šmilauerův doplněk určující. Označujeme ho podle zavěšení buď atributem *Atv* (pokud jsou přítomny oba členy, k nímž se tento doplněk vztahuje, zavěšujeme ho 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) považujeme za objekt a označujeme atributem *Obj*, zavěšujeme ho pak jako ostatní objekty pouze na sloveso.

V manuálu (str. 60, kapitola 3.3.4.6 Objekt po slovesech sponových a polosponových) je uveden seznam sponových a polosponových sloves (celkem 9), 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).*

Nabízí se otázka, zda jde o konečný seznam sloves nebo zda to jsou jen příklady. Během anotace textů se objevila analogická slovesa, která manuál neuvádí a bylo by vhodné o ně manuál doplnit. Z následujících příkladů je patrné, že by se mělo doplnit např. sloveso *považovat* a *prohlásit*: *Pokládat něco za diskriminaci* je podle manuálu objekt a *považovat něco za diskriminaci* už podle manuálu objekt není. *Označit návrh za špatný* je podle manuálu objekt a *prohlásit návrh za špatný* už není (tj. měl by být anotován jako doplněk).
Jak je vidět z uvedených příkladů, popis doplňku v manuálu je opravdu nedostatečný a dostává anotátora do úzkých.

Z hlediska dostatečné popisnosti a úplnosti najdeme v manuálu další prohřešky. Například nikde není výslovně řečeno, jak se zavěšuje rematizator u substantiva s nepravou předložkou (*pouze na úkor studia*). nebo kam se zavěšuje *se / si* u složeného predikátu typu *začíná se rozednívat, musí si uvědomit, začíná si zvykat,* nebo také jakým způsobem se zachycuje přimá řeč po slovesech typu *lamentoval, přihlasil se, pohrozil.*

Nutno přiznat, že tištěná podoba manuálu je nedostatečně přehledná. Některá závažná sdělení, která by měla mít formu pravidla, jsou zmíněna jen mimochodem a anotátor je nucen číst mezi řádky. Proto je vhodnější pracovat s elektronickou podobou manuálu, protože se v ní snadněji vyhledají všechna místa, kde se o daném jevu pojednává.

Závěrem této části je třeba dodat, že manuál pro reprezentaci větných struktur na analytické rovině vznikal během anotování PZK a dostal definitivní podobu až v době, kdy práce na anotaci analytické roviny byly v podstatě dokončeny. Neprošel tedy stádiem připomínek, kdy by bylo možno některá pravidla lépe formulovat, korigovat nevhodné umístění některých pokynů nebo chybějící pravidla do manuálu dodat.

6. Ruční kontrola anotací

Kontrola anotací je důležitou, časově náročnou součástí každého anotačního procesu. Má za úkol opravit případné chyby (velkou část z nich detekovaných automatickými kontrolními procedurami) ve zpracovaných datech, ale také zajistit jednotnost přístupu při řešení jevů homonymních, diskutabilních a sporných, tzn. zajistit co nejvyšší konzistencí zpracovaných dat.

Kontrolní opravy se týkají jak samotné podoby stromové struktury, tak i přiřazení analytických funkcí.

6.1 Chyby

Při kontrolách se odstraňují především chyby vzniklé nepozorností, přehlédnutím nebo opomenutím. Domníváme se, že při zpracování tak rozsáhlého souboru dat se těmto chybám nelze úplně vyhnout.

Opravují se také 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* zavěsily anotátorky 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., kdežto 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 větě. 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.*

Při opravách je dále třeba věnovat pozornost neshodám, které jsou způsobeny nedůsledností pokynů manuálu, a zajistit konzistenci v zachycení dat.

Jak už bylo řečeno, popsat pomocí pravidel všechny jazykové jevy, které se v korpusu vyskytují, není možné. Dokonce i v případech, kdy jsou pravidla formulována explicitně, zůstává prostor pro intuici a individuální přístup anotátora.

Prostor pro anotátorovu intuici skýtá např. vytváření stromové struktury, která obsahuje více koordinačních vztahů. V manuálu (kapitola 3.5.1.3 Vícenásobná koordinace, str. 143) se dočteme, 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 člení 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 vlastně dvojí řešení. Buď se všechny koordinační členy zavěsí na poslední *a* ve větě, nebo se čárka před slovem *zákonité* zavěsí jako koordinační znak na slovo *vztahy* a jednotlivá tři koordinační spojení jsou dceřiné uzly koordinační čárky.

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.

6.2 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 výslovně o chybách, protože jde spíše o různé pohledy na daný jazykový jev. V klasických mluvnicích jsou k takovým jevům uváděny komentáře. V anotačním procesu PZK se však musí vybrat jen jedno řešení, které, pokud je to možné, je ve shodě s pravidly a konvencemi manuálu.

6.2.1 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.
V PZK se při rozhodování, zda vybrat analytickou funkci *Obj* nebo *Adv* opíráme o práce J. Panevové (1980, 1998)[9]. V nejistých případech jsme se v případě ČAKu mohli opřít o výběr funkce podle valenčního slovníku PDT-Vallex, který byl na základě analyzovaných textů vypracován pro tektogramatickou rovinu PZK.
I přesto, že anotátor má k dispozici uvedené zdroje, je obtížné dosáhnout jednoznačnosti zpracování dat, protože ne u všech významů sloves, které se v korpusu vyskytly, je k dispozici valenční rámec.

6.2.2 Rozlišení konstrukcí stavových a dějových

Při kontrolách anotace vyšlo najevo, 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* (manuál, str. 34) 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*.

[9] PANEVOVÁ, J. (1980): *Formy a funkce ve stavbě české věty*. Praha, Academia a PANEVOVÁ, J. (1998a): Ještě k teorii valence. SaS, 59, s. 1-14.

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 takto zřejmé, protože větu lze chápat obojím způsobem. Například obsah věty *Všechna naše setkání byla naplněna vřelým přátelstvím* lze chápat buď stavově jako obrazně řečenou variantu věty *Setkání byla přátelská* (nejde zde o žádné naplňování a *byla naplněna* lze považovat za predikát nominální), nebo dějově jako větu, jež lze parafrázovat jako *Vřelá setkání(Obj) naplnila naše přátelství(Sb)* (jde o naplňování a *byla naplněna* je pasivum slovesa *naplnit,* i když tuto interpretaci lze považovat za zvláštní, nikoli však nemožnou**).**

U těchto konstrukcí je proto třeba počítat s nejednoznačností (nekonzistencí) v anotovaných datech.

6.2.3 Indentifikace částice *se a si* (*AuxT/AuxR*)

Částice *se* může být 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 reflexívní 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* i částice *si* zavěšují na sloveso, ke kterému patří, ať už mají jakoukoli z uvedených funkcí.

Ačkoliv kritéria přiřazování analytických funkcí částici *se* a částici *si* jsou formulována velmi zřetelně, při anotování konkrétního textu 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*). Poslední případ má možnou dvojí interpretaci (a to i ve velmi podobných kontextech), což vede k nekonzistenci.

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*).

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

Při kontrole anotovaných dat vyšlo najevo, že je-li predikát vyjádřen tvary slovesa *být* (sponou), není vždy lehké 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) asi anotátor 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 připouští v takových případech dvojí výklad (tzn. homonymní strukturu), v korpusu je třeba se rozhodnout a opět tedy počítat s nekonzistencí při přidělování této funkce.

6.2.5 K nepravým (nevlastním) předložkám

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 vyplynula na povrch dvojí mylná tendence v anotaci nepravých předložek: buď se nepravá předložka nepovažuje za nepravou předložku (např. anotace spojení *v oblasti* je nesprávně následující: *v* (*AuvP*) *oblasti* (*Adv*) *vědy* (*Atr*)), nebo se předložkové sousloví neoprávněně za předložku považuje (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. Je zřejmé, že tento seznam je neúplný a

anotátoři musí svůj názor ověřovat i v jiných normativních příručkách českého jazyka, např. ve Slovníku spisovné češtiny[10].

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), na úkor kvality (na újmu zdraví), 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).*

Domníváme se, že by bylo vhodné přílohu nevlastních předložek v manuálu aktualizovat a následně texty dodatečně zkontrolovat. Přesto bychom se však ani po dodatečné kontrole zřejmě nevyhnuli omylům v anotaci nepravých předložek. 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 vzhledem k tomu, že samotná kategorie nevlastních předložek není v české lingvistice definitivně vymezena, obtížné. Při anotaci tohoto lingvistického jevu navíc často selhává i intuice rodilého mluvčího.

[10] SSJČ. Slovník spisovného jazyka českého. Praha. Academia. 1989.