Pražský závislostní korpus aneb Co tady před padesáti lety nebylo

Barbora Hladká, Praha

1. Úvod

Nabídku přispět do speciálního čísla Pokroků, které vychází u příležitosti 50. narozenin MFF UK, jsme rádi přijali hned ze dvou důvodů. První nám otvírá možnost plynule navázat na příspěvek z předloňských Pokroků (Panevová a kol., 2000); druhý dává příležitost představit něco, čím se MFF může pochlubit až nyní, co tady rozhodně před padesáti lety nebylo.

Pokud si máme počítače spojit s něčím pro ně neodlučitelným, tak by to měla být především jejich rychlost ve vypořádávání s náročnými operacemi, které jim předkládá člověk jakožto „zařízení“ přirozeně chytřejší, ale bohužel výrazně pomalejší. Většinová představa chápe čísla jako hlavní elementy zmíněných operací; pro naše potřeby opusťme svět náročných matematických výpočtů a přejděme od čísel k textům (i když jedním dechem dodáváme, že od textů se budeme muset „pokorně“ k číslům zase vrátit), s nimiž se každodenně setkáváme v knihách, novinách, časopisech, na reklamních letácích aj., ale již v elektronizované podobě.

S čísly se počítá — co se dělá s texty (kromě toho, že se čtou . . .)? Na různé aktivity spojené s texty se můžeme podívat ze dvou základních pohledů — pohled jazykovědný (teoretický) a pohled aplikační (praktický). Při pohledu jazykovědném se písemné projevy (nezapomeňme ale i na projevy mluvené) jazyka zkoumají a analyzují s cílem jazyku porozumět. I v případě jazykových aplikací, ve kterých právě počítače představují „moc vykonavatelskou“, nám jde o správné uchopení jazyka. Už jenom prosté vyhledávání v textech (jako příklad jedné z mnoha jazykových aplikací) nabízených prostředím Internetu o objemu několika miliónů slov kontrastuje s nereálnou představou projití takového množství jedním člověkem (dvěma, třemi, . . .). Nastupují tedy různé vyhledávací programy a výkonné počítače. Pokud nás zajímají výskyty jednotlivých slovních tvarů, znalost jazyka může být mizivá. Pokud ale pracujeme s jazykem tzv. flexívním, jakým čeština bezesporu je, potřebujeme mít znalosti z tvarosloví (morfologie) jazyka. Pro ilustraci nás mohou zajímat věty, ve kterých se mluví o *mateřském jazyce*. Bohaté tvarosloví češtiny nám „podsouvá“ kromě vět s výskytem spojení *mateřský jazyk* přirozeně i věty s výskyty spojení *mateřského jazyka*, *mateřskému jazyku*, *mateřském jazyku/jazyce*, *mateřským jazykem*, *mateřské jazyky*, *mateřských*

Mgr. Barbora Vidová-Hladká, Ph. D. (1971), Centrum komputační lingvistiky MFF UK, Malostranské nám. 25, 118 00 Praha 1, e-mail: hladka@ckl.mff.cuni.cz

jazyků, *mateřským jazykům*, *mateřských jazycích*, *mateřskými jazyky*. Samozřejmě, že můžeme ručně zadávat vyhledávacímu programu jednu formu spojení za druhou, ale proč by generování (např. skloňování podstatných jmen, časování sloves, ...) všech přípustných tvarů nemohl za nás provést sám vyhledávací program, tedy, zjednodušeně řečeno, počítač?

Složitější situace nastává, chceme-li např. získat přehled všech škol s rozšířenou výukou španělštiny. Převedeno do vyhledávací terminologie, zajímají nás aspoň ty věty, ve kterých jako podmět figuruje spojení *škola*, přísudek *vyučuje/nabízí/pořádá/...* a předmět *španělštinu/hodiny španělštiny*. Vyhledávání tohoto typu se neobejde bez zapracování znalostí syntaxe daného jazyka. A opět, podobně jako v předchozím příkladě, požadujeme tuto znalost od vyhledávacího programu.

Náročnější aplikace typu strojový překlad, tj. překlad z jednoho jazyka do druhého s pomocí počítače — tedy počítač jako asistent překladatele, by měla pro oba dva jazyky s sebou nést komplexní znalosti nejen tvarosloví a syntaktické stavby vět, ale i znalosti z oblasti významu (tj. sémantiky).

Ze své podstaty počítač sám nic neudělá; pokud po něm něco chceme, musíme jeho kroky jednoznačně „nastavit“ jeden po druhém. Pro ilustraci: požadujeme po vyhledávacím programu, aby zvládl tvarosloví češtiny, tedy nejen generování (ve smyslu příkladu s *mateřským jazykem* uvedeným výše), ale i analýzu (pro každou slovní formu zjistit možná morfologická čtení, tj. určit možné slovní druhy a možné hodnoty patřičných morfologických kategorií, jako je rod, číslo, pád, osoba, čas, ...). Je přirozené, že jedna slovní forma může mít více morfologických čtení bez ohledu na větný kontext; větný kontext poté zpravidla jednoznačně určí to jediné správné morfologické čtení. Na příkladu věty *Zajímá se o mateřský jazyk* vidíme dané zjednoznačnění velice názorně — jazyk je podstatné jméno, rodu mužského neživotného, čísla jednotného v pádě prvním nebo čtvrtém; čtvrtý pád je tím správným v kontextu věty. (Obtížnější, ale na štěstí i vzácnější jsou věty, ve kterých se víceznačnost řeší až kontextem širším, např. *Školy úřady upozornily, že situace je vážná.*) Již zjednoznačněná morfologická čtení jsou důležitá pro syntaktický větný rozbor, tj. pro určení větných členů. Víme např., že podmět v české větě nemůže být v šestém pádě.

V zásadě existují dva základní přístupy použité v automatických jazykových aplikacích — přístup založený na znalostech a přístup založený na strojovém učení (tzv. samoučící se metoda). První přístup vyžaduje důkladné znalosti např. z morfologie a syntaxe jazyka, které sám autor systému „zakóduje“ do pravidel. Ve druhém přístupu autor na základě svých znalostí připraví tzv. trénovací data (označkovaný korpus) a matematickým (nejčastěji statistickým) aparátem se vygeneruje model dat.

Zastavme se u strojového učení trochu déle. Když učení, tak musíme předem vědět, čemu se chceme naučit a z čeho budeme znalosti čerpat. V naší konkrétní situaci chceme počítač naučit např. přiřazovat jednoznačná morfologická čtení v daných větných kontextech a budeme ho tomu učit z ručně morfologicky označkovaného korpusu. Korpus textů na tomto místě představovat nebudeme; poznamenejme jen, že obohatíme-li korpus o jakékoli další přídavné informace (např. o jednoznačná morfologická čtení), dostáváme označkovaný korpus.

Zkusme si napsat zkušební test. Otázky na to, co jsme přímo nastudovali ze studijních materiálů, by nás zaskočit neměly (i když připouštíme, že i v takových případech může dojít k omylu v odpovědi). Na druhou stranu se můžeme setkat s otázkou, na kterou explicitní odpověď v učebnici rozhodně nenajdeme. Nabízejí se dvě krajní možnosti, jak se s neznalostí poprat — zapojit intuici a pokusit se správnou odpověď odvodit ze známých znalostí (načerpaných z jiných zdrojů), nebo jen tak něco „plácnout".

Metody strojového učení se potýkají se stejným problémem. Relativně často viděné situace v trénovacích datech — v našem případě v označkovaném korpusu — se dají naučit velmi dobře. Přirozeně se ale může stát, že se v nových textech, které jsou automaticky značkovány na základě naučených znalostí, objeví situace v učících datech neviděná nebo pouze „zahlédnutá" — v takových situacích procedury hádají (neočekávejme od automatických procedur zapojení intuice, žádnou totiž nemají).

Doufáme, že po tomto polidštěném pohledu na chybovost automatických procedur jsme vyvolali jistou dávku tolerance k chybovosti morfologických čtení např. u slov Českého národního korpusu (SYN2000[1]), 2000), který je přímo přístupný přes hlavní www stránku Ústavu českého národního korpusu, FF UK. Procedura použitá na přiřazení morfologických čtení českých slov (Hajič, 2002) pracuje s 92–93% přesností.

Podrobnější informace o aplikaci a výsledcích metod strojového učení na automatické zpracování češtiny jsou k dispozici v (Panevová a kol, 2000).

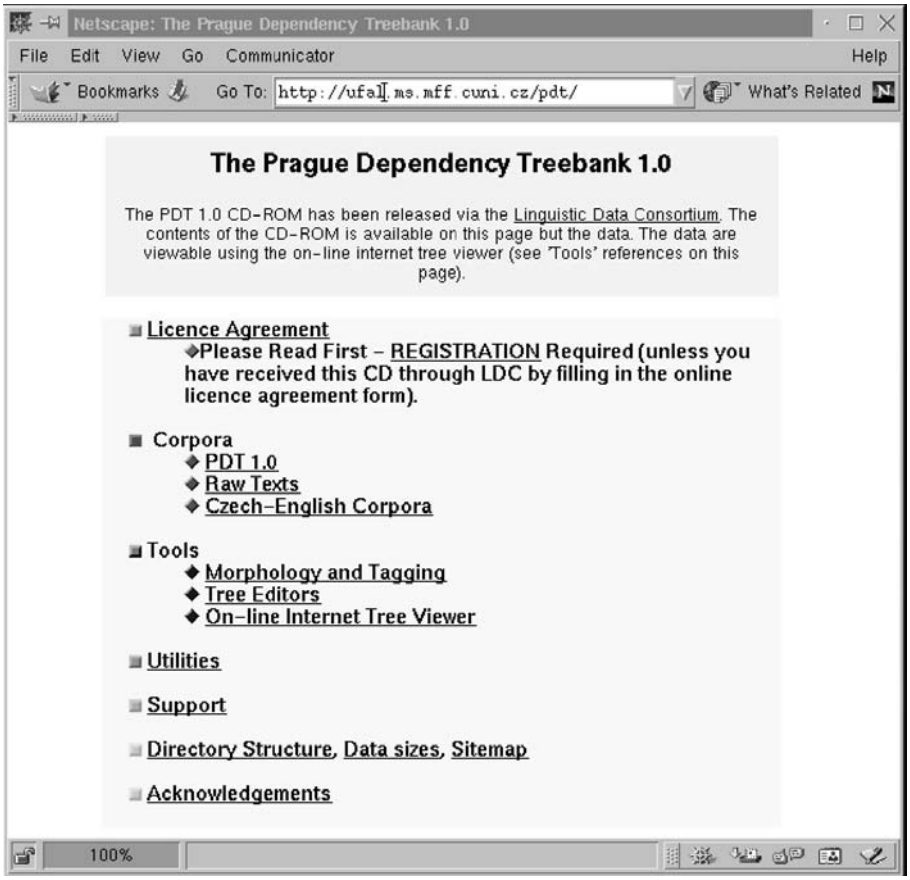
2. Pražský závislostní korpus, verze 1.0 (CD ROM)

V kontextu anotovaných korpusů nyní nastává vhodná chvíle uvést Pražský závislostní korpus (konkrétně jeho první publikovanou (CD-ROM) verzi, PZK 1.0), který představuje druhý největší, ručně označkovaný korpus na světě (hned za anglickým Penn Treebank (1992), kterým jsme se nechali inspirovat). Dříve než představíme to nejcennější, co CD nabízí, tedy PZK, zastavíme se u popisu samotného CD.

2.1. Co „magický" kotouč přináší

Do všech jeho „komnat" je možné nahlédnout přes hlavní uvítací stránku (viz obr. 1). Z pohledu interiéru CD pokrývá složku datovou (korpusovou) a složku nástrojů. Dominantou datové složky je rozhodně již zmíněná první verze Pražského závislostního korpusu. Jí sekundují paralelní česko-anglický korpus textů z výběru Reader's Digest (450 článků, 53 tis. paralelních vět) a texty (ve své původní podobě o objemu 39 mil. slov) z deníku Lidové noviny (ročníky 1991–95). Jako společný vnitřní formát nabízených korpusů jsme zvolili značkovací kódování SGML.

[1]) SYN2000 je reprezentativní vyvážený korpus současné psané češtiny, obsahující 100 mil. slov.



Obr. 1. Úvodní obrazovka CD ROM „PDT 1.0“.

Složka nástrojů v sobě nese nástroje vytvořené pro potřeby anotování (tzv. stromové editory[2])), pro potřeby prohlížení anotovaných dat (on-line internetovský prohlížeč[3])) a pro potřeby morfologického zpracování textů (morfologická analýza a automatické procedury morfologického značkování založené na přístupu strojového učení). Přirozeně, že český jazyk není jediným jazykem, kterým se světová komputační lingvistika zaobírá; přece jenom těch, kteří se věnují angličtině, je drtivá většina. Tento fakt ale nic neubírá na intenzivním zájmu otestovat metodologie originálně „ušité“ na angličtinu na typologicky odlišný jazyk (co když přece jenom je primární základ všech jazyků společný... ?). Pro metody strojového učení aplikované na angličtinu na sebe odpovědnost učících dat bere již zmíněný PennTreebank, jehož vnitřní formát se od námi použitého formátu liší. Proto vznikla automatická procedura pro převod textových dat z PZK do formátu PennTreebanku tak, aby otestování vhodnosti či

[2]) Ukázku interaktivní obrazovky ilustrují obrázky 2 a 3.

[3]) Anotovaná data nejsou ke stažení přes Internet, ale prostřednictvím on-line prohlížeče je možné v nich „listovat“.

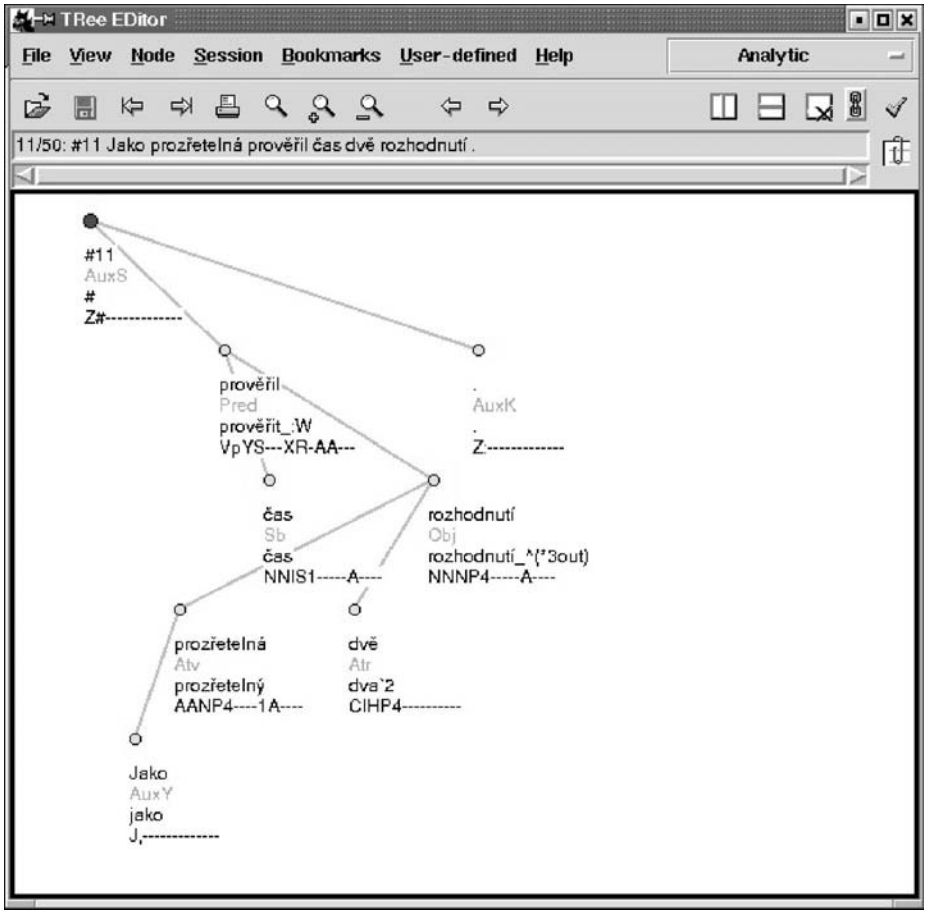
nevhodnosti dané metody pro češtinu bylo možné realizovat bez únavné přípravy toho správného formátu dat. Podobně užitečných procedur je ve složce nástrojů k dispozici celá řada.

Shrnout výsledek pětiletého projektu do dvou odstavců bez pocitu uzardění či dokonce viny z opomenutí poděkovat všem těm, kteří mají zásluhu na tom, že se můžeme chlubit (o přílišném zjednodušení ani nemluvíme), lze jen a jen díky poslední stopě kotouče, do které jsou „vylisována“ nejen jména všech těch, kteří se na budování PZK podíleli, ale samozřejmě i názvy všech organizací, které svými finančními prostředky umožnily se do takové práce pustit a dotáhnout ji do úspěšného konce.

2.2. Charakteristika PZK 1.0

PZK 1.0 (Hajič a kol., 1998, PZK, 2001) je ručně označkovaný (anotovaný) korpus českých textů (převážně publicistického stylu) o celkovém objemu cca 1,8 mil. slov. Anotování korpusu probíhalo ve dvou stupních: stupeň odpovídající tvarosloví a stupeň odpovídající větné stavbě (skladbě). Pro potřeby tvaroslovného anotování byla navržena množina tzv. morfologických značek jako reprezentantů morfologických čtení. Spolu se značkami se pro slovo určuje i základní forma, tzv. lema — identifikace lexikální jednotky (v tzv. slovníkovém tvaru — nominativ singuláru, infinitiv). Při syntaktickém anotování korpusu se vychází ze závislostní syntaxe, která charakterizuje větnou stavbu na základě valence (kombinatorických možností), především valence slovesa, protože sloveso je chápáno jako vlastní centrum stavby věty. Pro potřeby anotování byly vyvinuty softwarové uživatelsky přátelské nástroje.

Na obrázku 2 je velice pěkně vidět jak oba dva stupně anotování, tak i hlavní okno anotačního programu. V jeho hlavní části je zobrazen tzv. analytický závislostní strom pro konkrétní větu z článku v jistých novinách (*Jako prozřetelná prověřil čas dvě rozhodnutí*). Tato věta je v obvyklé podobě v okně anotačního programu zobrazena v horním řádku. Každému slovu věty (včetně např. tečky na jejím konci) odpovídá uzel analytického stromu (uzly jsou znázorněny kolečky), z technických důvodů je ve stromu navíc uzel, který slouží k identifikaci věty v rámci korpusu. Úsečky (hrany) mezi uzly znázorňují syntaktické závislosti, známé z větného rozboru, ale jsou tu obdobně vyznačeny i vztahy další, týkající se pomocných slov a interpunkce. Přibližme si údaje uvedené u jednotlivých uzlů. Na prvním řádku pod každým uzlem je zopakováno příslušné slovo přesně tak, jak se vyskytuje ve větě. Na druhém řádku je tzv. analytická funkce daného slova, která vyjadřuje druh závislosti na slově řídícím (např. značka *Pred* odpovídá přísudku, *Sb* podmětu, *Obj* předmětu, *Atr* přívlastku, *Atv* doplňku). Na posledních dvou řádcích je uvedeno lema, tj. identifikace základního tvaru slova, a morfologická značka pro daný tvar slova vyskytující se ve větě. Morfologická značka je patnáctipoziční řetězec, ve kterém každé pozici odpovídá právě jedna morfologická kategorie (první pozice — slovní druh, druhá pozice — slovní poddruh, třetí pozice — rod, . . . , osmá pozice — osoba, . . .). Protože morfologická značka zachycuje pouze hodnoty morfologických kategorií relevantních pro příslušný slovní druh, jsou např. u slova *rozhodnutí* obsazeny hodnoty relevantní pro podstatné jméno (NNNP4-----A----).

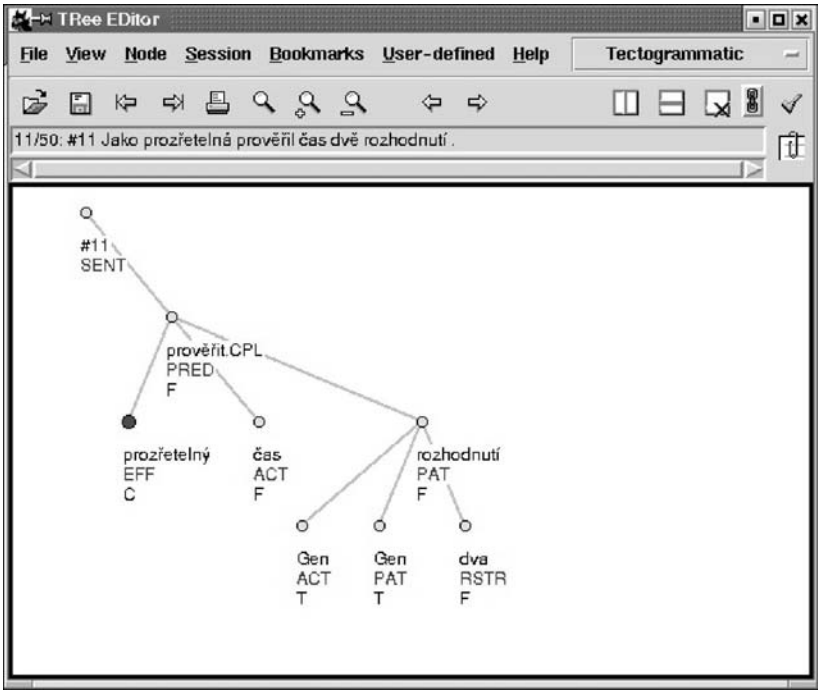


Obr. 2. Věta *Jako prozřetelná prověřil čas dvě rozhodnutí,* anotovaná na syntakticko-analytické rovině.

Nejdůležitějším kritériem, které musí být v případě vzniku takovéto banky textů (datového materiálu) respektováno, je kritérium konzistence. Ta může být zajištěna (alespoň do určité míry) buď existencí manuálu s (pokud možno) jednoznačnými pokyny pro anotátory, nebo velice omezeným počtem anotátorů. Při anotování PZK jsme uplatnili obě dvě strategie. Anotování na morfologické a syntakticko-analytické rovině zajišťovaly paralelně dva různé týmy anotátorů následujícím způsobem. Textový soubor určený k morfologické anotaci byl označkován dvěma anotátory (bez manuálu). Diskrepance, které se přirozeně vyskytly, byly vyřešeny jediným anotátorem. Další kontroly morfologického značkování vzhledem k morfologické analýze prováděli pouze dva anotátoři. Naopak „stromečkování“ (anotování na syntakticko-analytické rovině) se řídilo manuálem, který vznikal a postupně se dolaďoval během anotačního procesu. V závěrečné fázi byla provedena vzájemná kontrola morfologického a syntakticko-analytického značkování stejnými dvěma anotátory jako v případě morfologického značkování a jediným „stromečkovým“ anotátorem.

3. Na cestě od verze 1.0 k verzi 2.0

Pokud chceme počítač brát jako opravdu produktivního pomocníka překladatele, musíme do systému automatického překladu (nebo jiných náročných aplikací, zejména pro komunikaci s automatickým systémem) zapracovat to, co vlastně dané věty říkají, tedy význam. K tomu směřuje přechod od analytické stavby k významové, tzv. tektogramatické stavbě věty. V terminologii verzí PZK to znamená vydat se od verze první směrem k verzi druhé; prakticky je tato cesta možná díky vzniku Centra komputační lingvistiky[4]).



Obr. 3. Věta *Jako prozřetelná prověřil čas dvě rozhodnutí,* anotovaná na tektogramatické rovině.

Na obrázku 3 je zachycen tektogramatický závislostní strom pro větu z obrázku 2 morfologicky a syntakticko-analyticky analyzovanou. Tektogramatická anotace vychází ještě důsledněji ze závislostní syntaxe. Tektogramatický strom opět zachycuje syntaktické závislosti, ty jsou ale ve srovnání s analytickou anotací jemněji významově členěné. Rozdílů je však víc - za povšimnutí stojí, že ve srovnání s předchozím obrázkem zde některé uzly ubyly, jiné zase přibyly. Při tektogramatickém značkování se nezachycují tzv. slova synsémantická, tj. slova samostatně nenesoucí význam (např. předložky), přidávají se však uzly odpovídající tzv. elidovaným (vypuštěným) členům, tj. informaci, která je zřejmá každému mluvčímu češtiny, který dané větě rozumí, ale která není vyjádřena v povrchové podobě věty.

[4]) Projekt Ministerstva školství, mládeže a tělovýchovy (identifikační číslo LN00A063)

Protože *při každém rozhodnutí rozhoduje někdo o něčem*, jsou do stromu přidány uzly závisející na slově *rozhodnutí*, a to uzel se značkou ACT (odpovídající tomu, kdo rozhoduje) a uzel se značkou PAT (o čem rozhoduje) — oba mají na prvním řádku lema Gen (tj. general, všeobecný), vyjadřující, že konkrétní osoba a věc, o které by rozhodovala, nemusí být známy. Závislostní vztahy v tektogramatickém stromu zachycují významovou závislost, dochází proto někdy k „přesunutí" závislostní hrany. V případě věty na obrázku 3 jde o uzel s lematem *prozřetelný* a značkou EFF (odpovídá tzv. výsledku děje); ten v tektogramatickém stromu závisí přímo na slovesu, protože vlastně vyjadřuje, k jakému závěru čas při svém prověřování dospěl. Uspořádání stromu zleva doprava spolu s údaji na posledním řádku zachycují tzv. aktuální členění věty, tedy členění na základ (informace prezentovaná větou jako známá, daná informace) a jádro výpovědi (nová informace). Určitě nezůstalo bez povšimnutí, že na obrázku 2 je jedna závislostní hrana „výrazně šikmá", a to hrana spojující slova *prozřetelná* a *rozhodnutí* (spojuje slova, která jsou od sebe ve větě vzdálená, a tuto hranu protíná aspoň jedna svislice spuštěná z nějakého uzlu ležícího nad hranou — zde jde o slova *prověřil* a *čas*). Tato realizace závislostního vztahu v analytickém stromě je příčinou toho, že v tektogramatickém stromu je uzel s lematem *prozřetelný* chápán jako tzv. kontrastivní základ výpovědi (značka C). Uzly se značkou F (fokus, ohnisko) zde představují autosémantická slova (slova nesoucí samostatný význam) přinášející novou informaci, uzly se značkou T (téma, základ) představují informaci danou. (Ještě poznámka pro zvídavé: tmavé (ve skutečnosti červené) kolečko na obrázcích 2 a 3 říká, s kterým uzlem uživatel anotačního programu právě pracuje.)

4. PZK 1.0 je o jazyce

Doufáme, že z předchozích řádků vyplývá alespoň rámcová představa o daném produktu. Už možná není tolik zřejmé, v čem spatřujeme jeho ojedinělost a význam.

PZK 1.0 je o jazyce — dokonce o jazyce mateřském, v jeho psané podobě. Tedy o jednom fenoménu, který nás spojuje a kterému se nikdo z nás nemůže vyhnout. Jde pouze o to, jaký a jak hluboký vztah s přirozeným jazykem navážeme. PZK vznikl v prostředí vědeckém — začněme tedy u vztahu profesního, který by už sám o sobě měl být hluboký. Bez existence textových korpusů vznikají lingvistické teorie často na příkladech osamocených vět, které se více méně rodí v hlavách jazykovědců. Přítomnost realistického datového materiálu — v našem případě textů, které se skutečně objevily na stránkách knih, časopisů, novin, ... a ve kterých je možné listovat „stiskem klávesy" — je výzvou pro otestování těchto teorií. Vzájemné soužití teoretické a počítačové lingvistiky tvoří kruh, jehož uzavřenost rozhodně není na závadu. Ba naopak, vzájemně se podporující výsledky[5]) rozhodně vedou k lepšímu porozumění jazyku. Ono lepší porozumění se snadněji ilustruje právě na aplikacích počítačové lingvistiky s výsledky povahy hmatatelnější — gramaticky/stylisticky opravený text,

[5]) Vůbec nemáme na mysli většinu komerčních produktů, v jejichž případě se bohužel o nějakém solidním teoretickém základu nedá mluvit.

úseky textů výhradně s tematikou, která nás zajímá, slovníky s příklady použití slov v reálných kontextech, ...

Je-li k dispozici datový materiál o objemu srovnatelném s objemem PZK 1.0, je radostí testovat starší i nové jazykové teorie, stejně tak i nechat počítač „samostatně“ se učit zákonitostem jazyka.

Jiný vztah, který lze s přirozeným jazykem navázat, bychom označili jako školský. Nabízíme učitelům i žákům novou pomůcku do hodin českého jazyka. Mluvnický rozbor vět představuje tu náplň hodin češtiny, na kterou si jistě každý vzpomene jako na něco ne příliš záživného. PZK 1.0 obsahuje na 100 000 vět mluvnicky rozebraných — dostatek vět na to, aby se procvičování větné skladby stalo zábavnějším i díky příjemnému, variabilnímu (co do vizualizace skladby věty) editoru (dostupného na CD, viz výše) nejen pro prohlížení anotovaných dat, ale i pro vlastní určování základních skladebních dvojic „šoupáním“ myši na obrazovce.

5. ... aneb co tady před padesáti lety nebylo

Iniciátory myšlenky Pražského závislostního korpusu se staly významné osobnosti české komputační lingvistiky, jejichž příchod (lépe příchody, pro bližší vysvětlení viz (Sgall, 2002)) na půdu Matematicko-fyzikální fakulty nenastal rozhodně před padesáti lety. Nastal o něco později, a to v době, kdy se samotná korpusová lingvistika rodila; chce se nám říci světová korpusová lingvistika, protože vzhledem k jejímu hlavnímu poslání — budování korpusů — nebyl na takové ambice v československých podmínkách let šedesátých až osmdesátých prostor. Ovšem výzkumu formálního popisu a algoritmického zpracování češtiny nemohlo nic zabránit. A tak tedy v době, kdy již zmínění iniciátoři, obklopeni početným týmem spolupracovníků a stále se vylepšujícím výpočetním zázemím, mohli svůj teoretický výzkum zasadit do konkrétní aplikace, začal vznikat ojedinělý PZK.

Literatura

[1] Český národní korpus (SYN2000). `http://ucnk.ff.cuni.cz`

[2] Hajič, J.: *Disambiguation of Rich Inflection — Computational Morphology of Czech.* Charles University Press — Karolinum, 2002.

[3] Hajič, J., Hajičová, E., Panevová, J. Sgall, P.: *Syntax v Českém národním korpusu.* Slovo a slovesnost *59* (1998), 168–177.

[4] Hajič, J., Hladká, B.: *Morfologické značkování korpusu českých textů stochastickou metodou.* Slovo a slovesnost *58* (1997), 288–304.

[5] Kocek, J., Kopřivová M., Kučera, K. (editoři): *Český národní korpus — úvod a příručka uživatele.* FF ÚČNK 2000.

[6] Panevová, J. a kol.: *Počítačová lingvistika ve vztahu k informatice.* Pokroky mat. fyz. astronom. *45*, č. 3 (2000).

[7] Pražský závislostní korpus (PZK) 2001. `http://ufal.mff.cuni.cz/pdt`

[8] Penn Treebank.1992. `http://www.cis.upenn.edu/ treebank/`

[9] Sgall, P.: *Z fildy mezi matfyzáky, pro jistotu dvakrát. (Dětské krůčky pražské komputační lingvistiky.)* Jubilejní almanach k 50. výročí založení MFF UK (2002), 143–144.