Průvodce PDT 2.0

Jan Hajič, Eva Hajičová, Jaroslava Hlaváčová,
Václav Klimeš, Jiří Mírovský, Petr Pajas, Jan Štěpánek,
Barbora Vidová Hladká, and Zdeněk Žabokrtský

20. června 2006

Obsah

Kapitola 1

Úvod

Tento průvodce představuje Pražský závislostní korpus, verzi 2.0 (PDT 2.0). Smyslem průvodce je seznámit zájemce v krátkosti s obsahem a základními myšlenkami PDT 2.0. Poskytuje přehled dat a nástrojů, včetně odkazů na podrobnější dokumentaci, tutoriály, formální specifikace a další reference. K dispozici je ve dvou formátech: HTML a PDF.

Webovou stránku PDT 2.0 najdete na <http://ufal.mff.cuni.cz/pdt2.0>. Můžete navštívit také stránku <http://ufal.mff.cuni.cz/pdt2.0update>, kde v budoucnu najdete případné opravy dat, nové verze nástrojů apod.

1.1 Co je PDT 2.0

Pražský závislostní korpus (PDT) je probíhající projekt pro ruční anotaci velkého množství českých textů bohatou lingvistickou informací, sahající od morfologie přes syntax až po sémantiku/pragmatiku a ještě dále.

PDT verze 2.0 je následník verze 1.0. PDT verze 1.0 obsahovala ruční anotaci morfologie a povrchové syntaxe (viz <http://ufal.mff.cuni.cz/pdt/> nebo webové stránky Linguistic Data Consortium (LDC), <http://www.ldc.upenn.edu>, katalogové číslo LDC2001T10). Verze 2.0 přidává hloubkovou syntax a sémantiku, aktuální členění, koreferenci a lexikální sémantiku založenou na valenčním slovníku. Verze 2.0 přináší navíc aktualizaci verze 1.0, a to v původním formátu dat pro použití těmi, kdo se starou verzí pracují.

PDT 2.0 obsahuje velké množství českých textů (2 milióny slov) s provázanými anotacemi na úrovni morfologie (2 milióny slov), povrchové syntaxe (1,5 mil. slov) a hloubkové syntaxe a sémantiky (0,8 mil. slov). Korpus využívá nejnovější anotační techniky (oddělené anotace s použitím XML, RelaxNG, viz sekce 3.4 a celá kapitola 3).

PDT 2.0 vychází z dlouhodobé pražské lingvistické tradice a je vhodný pro současné potřeby výzkumu v oblasti počítačové lingvistiky (viz také sekce 1.2). Obsahuje rovněž softwarové nástroje pro prohledávání korpusu, anotaci dat a jazykovou analýzu. K dispozici je i rozsáhlá dokumentace.

Tato verze PDT završuje desetileté období výzkumu a vývoje v Ústavu formální a aplikované lingvistiky (ÚFAL) a jeho Centra počítačové lingvistiky (viz sekce 1.3). V nedávné době byl projekt doplněn vydáním Pražského arabského závislostního korpusu, <http://www.ldc.upenn.edu>, katalogové číslo LDC2004T23, a paralelního Pražského česko-anglického závislostního korpusu, <http://www.ldc.upenn.edu>, katalogové číslo LDC2004T25. První z doplňujících projektů ukazuje, že české specifikace mohou být uzpůsobeny pro typologicky odlišný jazyk, druhý projekt staví na ruční anotaci korpusu Penn Treebank a je určen pro experimenty se strojovým překladem mezi dvěma jazyky, hlavně mezi češtinou a angličtinou.

PDT 2.0 slouží především těmto dvěma cílům:

- aplikovat teoretické výsledky Pražské lingvistické školy na velké množství skutečných jazykových „příkladů", a tím explicitně ověřit a zachovat teorii závislostně založeného *funkčně generativního popisu (FGD)* (viz také sekce 1.2),
- umožnit použití metod strojového učení pro vytvoření rozumně spolehlivých nástrojů automatické analýzy a generování jazykových dat.

Zatímco pro dosažení prvního cíle by možná stačilo vybrat pouze několik příkladů pro každý lingvistický jev, druhý cíl nepochybně vyžaduje zpracování velkého množství přirozeně se vyskytujících

posloupností vět. Statistiky, získané z takových dat, mohou být ovšem s výhodou použity zpětně pro lingvistický výzkum.

Budoucnost PDT není zatím přesně určena. Zvažováno je několik možných budoucích zaměření (samozřejmě, pokud finanční zdroje dovolí): přidání mluvených dat; přidání hlubší a širší anotace obzvláště pro koreferenci, informační strukturu a diskurz; anotace jiného (hodně odlišného) jazyka; ruční anotace češtiny/angličtiny na dalších paralelních textech s použitím stejné (tektogramatické) reprezentace; a přidání dalších vrstev anotace (reprezentace znalostí založená na obsahu výpovědi).

1.2 Historické pozadí projektu

Pražská škola funkční a strukturní lingvistiky se narozdíl od ostatních evropských škol lingvistického strukturalizmu vyznačuje (kromě jiného) svou otevřeností novým trendům a myšlenkám. Historie Pražské školy se formálně datuje od roku 1926, kdy tak vynikající lingvisté jako Vilém Mathesius, Roman Jakobson a Bohumil Trnka založili Pražský lingvistický kroužek. Výzkum razil cestu v několika směrech. Nejprve ve fonologii, která byla možná první mezinárodně vysoce uznávanou oblastí. Brzy se zde objevily také (s kladným mezinárodním ohlasem) originální příspěvky k jazykové typologii, tvoření slov, funkčnímu rozvrstvení jazyka, k obecným lingvistickým otázkám jako je rozlišení centra a periferie v jazykovém systému a v neposlední řadě také pokusy o systematický popis informační struktury věty (funkční větná perspektiva, aktuální členění).

Činnost Pražského lingvistického kroužku nebyla omezena geograficky. K zásadám Kroužku se otevřeně hlásila řada lingvistů ze zahraničí. Jedním z nich byl Lucien Tesnière, francouzský lingvista, kterého je možno oprávněně nazývat „otcem závislostní syntaxe“. Tesnièrův přístup nalezl vysoce kladné přijetí i mimo Kroužek, obzvláště v práci českého syntaktika Vladimíra Šmilauera, jehož *Novočeská skladba* je neopominutelným zdrojem informací pro všechny, kdo českou syntax studují.

Inspirace Pražské školy nalezla své pokračování také v novém lingvistickém paradigmatu explicitního popisu jazyka, jmenovitě ve funkčně generativním popisu (FGD), navrženém Petrem Sgallem v šedesátých letech dvacátého století a následně rozpracovaném jím samým a jeho spolupracovníky (rozsáhlé pojednání na toto téma nabízí kniha *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, 1986). Systém FGD se vyznačuje třemi typickými vlastnostmi:

- použitím závislostní syntaxe,
- zahrnutím hloubkové syntaktické roviny (tektogramatiky) do lingvistického popisu,
- specifikací formálního popisu informační struktury věty (aktuálního členění) a jeho začleněním do popisu jazyka.

1.3 Vývoj projektu

Projekt vlastně vznikl ve foyer malého hotelu v Dublinu v Irsku na konci března roku 1995, během 7. ročníku konference evropské pobočky ACL. Malá skupina nás se tam tehdy rozhodla usilovat o vytvoření podobného projektu, jakým byl tehdy nedávno vydaný anglický Penn Treebank, ale založeného na pražské závislostní tradici, s úplnou morfologickou analýzou a s vyhlídkou postupného rozšiřování anotace (více historických souvislostí viz sekce 1.2).

Prvním úkolem bylo finanční zajištění projektu. Měli jsme štěstí a získali jsme současně dva granty od Grantové agentury České republiky a jeden projekt Ministerstva školství, všechny začínající v roce 1996: jeden malý grant pro sepsání specifikace korpusu, jeden meziinstitucionální projekt na podporu Českého národního korpusu (našeho zdroje nezpracovaných textů) a nakonec projekt nazvaný „Laboratoř jazykových dat“ pro vlastní provádění anotace.

Teorie vyžadovala tříúrovňové pojetí anotace, s morfologickou, analytickou a tektogramatickou rovinou. Kromě morfologické roviny, jejíž návrh využíval již existující systém tagů pro češtinu, byly pokyny pro anotaci jen kusé a bylo jasné, že jejich dopracování bude muset probíhat současně s anotací tak, jak se budou nacházet nové jevy a problémy. Nicméně již od počátku jsme přijali několik „neporušitelných“ principů:

- morfologická anotace bude prováděna na jednotlivých slovech; nebudeme se pokoušet analyzovat např. složené slovesné tvary,

- pro anotaci bude přímo použit systém tagů existujícího morfologického slovníku pro češtinu, vyvinutého na ÚFALu,
- jednotkou anotace povrchové syntaxe (analytické roviny) bude rovněž slovo, se vztahem 1:1 vůči jednotkám morfologické roviny; součástí anotace nebudou „stopy", náhrady elips ani nic podobného,
- závislostní anotace bude použita nejen pro rovinu hloubkové syntaxe (tektogramatickou rovinu), ale rovněž pro rovinu analytickou,
- tektogramatická rovina bude obsahovat všechno, co teorie nabízí, tedy aktuální členění, koreferenci a další podrobnou anotaci; v souladu s teorií a cíli hloubkové reprezentace bude umožněno „vkládání" a „mazání" uzlů (ve vztahu k nižším rovinám),
- funkce členů závislých na slovese (případně i na podstatném či přídavném jméně) bude určována na základě valence.

Formát pro anotovaná data byl vytvořen jakožto rozšíření SGML formátu používaného v Českém národním korpusu, pojmenovaného *CSTS*. Dalším krokem bylo určení organizace anotace. Začali jsme současnou anotací dvou nižších rovin (morfologie a analytické syntaxe). Anotace tektogramatické roviny musela být odložena až do dokončení dvou nižších rovin. Současně byly vytvářeny nástroje pro anotaci. Jedním z prvních byl `Graph`, grafický editor stromů, používající náš vlastní formát dat (nazývaný *FS*), který není založený na SGML, ale je značně obecný a prostorově úsporný.

Anotace morfologické a analytické roviny byla prováděna především pracovníky s lingvistickým vzděláním. Jelikož nebyly k dispozici úplné anotační pokyny, konaly se každý týden schůzky týmu anotátorů, kde byly probírány vzniklé problémy a s okamžitou platností přijímána rozhodnutí o způsobu další anotace. Později byl z řad anotátorů vybrán jeden koordinátor a další dva anotátoři museli být vyčleněni pro řešení technických otázek celého procesu.

Morfologická anotace každého textu byla prováděna dvěma anotátory, tedy dvakrát. Výsledky pak byly porovnávány a slévány do konečné anotace. Aby byla zajištěna co nejvyšší konzistence, celé slévání prováděl jeden anotátor. Anotátoři vybírali z možných lemmat a tagů, nabízených českým morfologickým slovníkem bez jakéhokoliv předzpracování či preference tagů. Na morfologické rovině tak byly ručně anotovány téměř dva milióny slov.

Anotace analytické roviny byla provedena jen jednou, ale s použitím velkého počtu automatických testů konzistence, včetně testů překračujících hranice rovin. Zpočátku jsme nepoužívali žádné automatické předzpracování textů. Později byly závislostní funkce předběžně přiřazovány ručně psanými skripty. V roce 1998 byla pro letní JHU Language Engineering Workshop v Baltimoru sestavena testovací verze korpusu, nazvaná PDT 0.5 (obsahovala přibližně 380 tisíc anotovaných slovních jednotek). Na workshopu byl vytvořen první český parser (data byla zkonvertována pro mírně upravený Collinsův parser lexikalizované angličtiny). Od roku 1999 byla data určená pro anotaci nejprve předzpracována tímto parserem a anotátoři prováděli pouze opravy jeho výstupu, což přineslo přibližně 30% zrychlení anotace. Na analytické rovině tak bylo ručně anotováno přes 1,5 mil. slovních jednotek, čímž se dosáhlo velikosti Penn Treebanku.

Spojení morfologické a analytické roviny byl složitý proces a trval déle než rok. Zahrnoval i rozsáhlé kontroly konzistence dat, závěrečné úpravy anotačních návodů (a jejich překlad do angličtiny), jakož i konečnou přípravu CD-ROM k publikaci v roce 2001 pod názvem Pražský závislostní korpus, verze 1.0. Během tohoto období byl také vytvořen `TrEd`, nový nástroj pro editaci korpusu, nezávislý na platformě.

Anotace tektogramatické roviny (již s použitím `TrEdu`) začala v roce 2000, současně se založením Centra komputační lingvistiky, v době, kdy původní finanční zdroje byly vyčerpány. Zpočátku se zdálo příliš náročné plně pokrýt celá plánovaná data (část dat PDT 1.0, cca 50 tis. vět). Anotace byla rozdělena do čtyř oblastí:

- závislostní struktura ve formě závislostního stromu, včetně sémantického označkování a anotace valence,
- aktuální členění,
- koreference (gramatická a část textové),
- gramatické atributy uzlů ve stromě (neobsažené v předchozích bodech).

Většina úsilí byla zaměřena na první oblast, neboť ostatní oblasti měly být anotovány jen na malé ukázkové části dat. Pomocí ručně psaných pravidel byly stromy analytické roviny předanotovány do té míry, pokud se vztah mezi analytickým a tektogramatickým stromem zdál být jasný. Byl vytvořen základ valenčního slovníku (zatím na papíře), aby byla zajištěna konzistence alespoň u nejfrekventovanějších sloves. Později byla vypracována XML verze valenčního slovníku, *PDT-VALLEX*, která byla rovněž propojena s editorem `TrEd`, aby mohli uživatelé pracovat se slovníkem přímo během editace; to také umožnilo přiřazovat správný valenční rámec k výskytům slov v korpusu. Mezitím pokročila práce na anotačních pravidlech a na testovací anotaci koreference a aktuálního členění a nakonec bylo rozhodnuto provést tyto anotace na celých datech. Ještě později, v roce 2004, byla i čtvrtá anotační oblast (přiřazení dalších gramatických informací, zahrnujících dalších 16 atributů u každého tektogramatického uzlu) poloautomaticky rozšířena na celá tektogramaticky anotovaná data, tedy 50 tisíc vět.

Narozdíl od anotování analytické roviny, v případě roviny tektogramatické byl anotační tým rozdělen na malé skupiny, které měly na starost jednotlivé oblasti anotace. To přinášelo i jisté obtíže - informace se někdy nedostaly ke všem, pro koho byly důležité. Po celou dobu pracovalo na projektu až 30 lidí současně. Vše bylo anotováno jen jednou, kromě úvodních testů mezianotátorské shody. Na data byly aplikovány podobné testy konzistence jako pro analytickou rovinu, s použitím složitých mezirovinových testů.

Po dokončení anotačního procesu v roce 2004 začala závěrečná fáze, která trvala rovněž déle než rok. Pro distribuci dat byl vytvořen úplně nový XML formát. Valenční lexikon PDT-VALLEX byl celý ručně zkontrolován a upraven pro slovesa a některé kategorie podstatných jmen (v obou případech jedním člověkem, aby byla zajištěna co největší konzistence). Bylo vytvořeno velké množství mezirovinových testů pro vyhledávání anotačních nekonzistencí, všechny nalezené případy byly ručně opraveny. Byl zvolen redaktor manuálu pro tektogramatické značkování, jehož úkolem bylo přepsat jednotlivé sekce pokynů (celkem přes 800 stran) jasnou formou s jednotnou terminologií tak, aby byl manuál v souladu s konečnou anotací dat. Manuál byl rovněž přeložen do angličtiny. V roce 2006 bylo CD-ROM dokončeno a posláno k publikaci do LDC.

1.4 O češtině

Čeština - jazyk textů zpracovaných v Pražském závislostním korpusu - patří do západní skupiny slovanských jazyků. Česky se mluví především v České republice, kde je čeština jediným úředním jazykem. Čeští rodilí mluvčí žijí rovněž v dalších evropských zemích, zvláště na Slovensku, a desítky tisíc českých mluvčích žijí v USA, Kanadě a Austrálii. Celkem má čeština přes 10 miliónů mluvčích.

Čeština je, podobně jako další slovanské jazyky, vysoce flexivní. Má sedm pádů a čtyři rody (jen pro skloňování podstatných jmen existuje 16 hlavních vzorů) a má volný slovosled (z čistě syntaktického pohledu): slova ve větě mohou být obvykle řazena několika způsoby. Slovosled však ovlivňuje význam věty.

Psaná čeština používá latinskou abecedu rozšířenou o několik písmen s diakritikou. Česká abeceda (celkem 82 znaky) je obsažena ve standardu Unicode; běžně používána jsou i kódování ISO 8859-2 (Latin 2), standardní 8-bitové kódování pro jazyky střední Evropy, a CP1250, jeho protějšek z MS Windows.

Více informací o češtině najdete na <`http://www.czech-language.cz`>.

1.5 Adresářová struktura

Tato sekce obsahuje stručný popis adresářové struktury distribuce PDT 2.0, a to až do druhé úrovně zanoření.

- `data/` – viz kapitola 3
  - `binary/` – kompletní anotovaná data (pouze na distribučním CD-ROM; viz sekce 3.6) ve formátu Perl Storable Format (viz sekce 3.4.2)
  - `filelists/` – několik předgenerovaných seznamů datových souborů (pouze na distribučním CD-ROM), viz sekce 3.6
  - `full/` – kompletní anotovaná data (pouze na distribučním CD-ROM; viz sekce 3.6) ve formátu PML (viz sekce 3.4.1)
  - `pdt-vallex/` – PDT-VALLEX, valenční slovník, viz sekce 3.8

  - `pdt1.0-update/` – aktualizace dat z CD-ROM PDT 1.0 (pouze na distribučním CD-ROM), viz sekce 3.9
  - `sample/` – malá ukázka anotovaných dat, viz sekce 3.7
  - `schemas/` – PML a RelaxNG schémata dat

- `doc/` – viz kapitola 5
  - `data-formats/` – dokumentace dat, viz sekce 3.4
  - `manuals/` – manuály (pokyny) pro anotátory, viz kapitola 2
  - `pdt-guide/` – tento průvodce PDT
  - `styles/` – kaskádové styly pro manuály a průvodce PDT
  - `tools/` – dokumentace nástrojů, viz kapitola 4

- `publications/` – publikace týkající se PDT 2.0, viz kapitola 6
- `tools/` – viz kapitola 4
  - `checks/` – makra pro hledání chyb v datech, viz sekce 4.7
  - `format-conversions/` – nástroje pro konverzi mezi různými formáty dat, viz sekce 4.4
  - `machine-annotation/` – nástroje pro vytvoření syntaktických stromů z prostého českého textu, viz sekce 4.5
  - `netgraph/` – `Netgraph`, nástroj pro vyhledávání v datech, viz sekce 4.1
  - `pml/` – Relax NG definice schématu PML a XSLT styl pro konverzi schématu PML do RelaxNG, viz sekce Nástroje ve Specifikaci PML.
  - `tred/` – `TrEd` a `btred/ntred`, nástroje pro prohlížení a zpracování dat, viz sekce 4.2, 4.3

- `visual-data/`
  - `pdt-vallex/` – PDT-VALLEX, valenční slovník ve formě webovských stránek, viz sekce 3.8
  - `sample/` – ukázková data ve formě webovských stránek, viz sekce 3.7

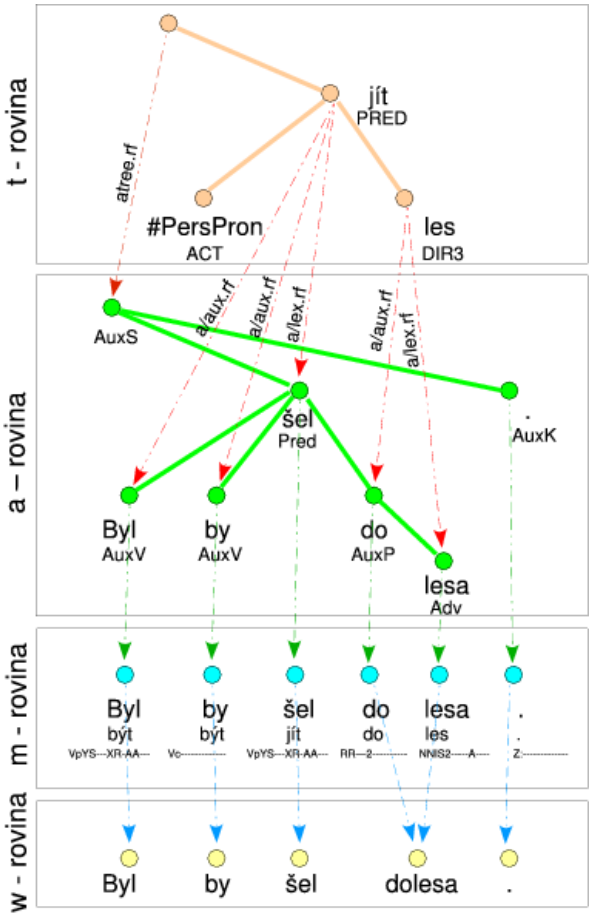
Kapitola 2

Roviny anotace

Data v PDT 2.0 jsou anotována na třech rovinách: na *morfologické rovině* (2.1), *analytické rovině* (2.2) a *tektogramatické rovině* (2.3). Ve skutečnosti existuje ještě jedna, neanotační rovina, reprezentující „surový text“. Na této rovině, zvané *slovní rovina*, je text rozdělen do dokumentů a odstavců. Jsou tu rozlišeny slovní jednotky (slova, čísla, interpunkce) a jsou opatřeny jednoznačnými identifikátory.

Slovní rovina je nazývána také *w-rovina*, morfologická *m-rovina*, analytická *a-rovina* a tektogramatická *t-rovina*. Podobně je uzel stromu reprezentujícího analytickou anotaci věty nazýván *a-uzel* atd.

Obrázek 2.1 znázorňuje vztah mezi sousedními rovinami, jak jsou anotovány a reprezentovány v datech. Zobrazená česká věta *Byl by šel dolesa.* obsahuje minulý čas podmiňovacího způsobu slovesa *jít* a tiskovou chybu.



Obrázek 2.1: Propojení rovin

2.1 Morfologická rovina

Tato sekce stručně popisuje morfologickou rovinu. Více informací najdete v Manuálu k morfologické anotaci.

2.1.1 Logická struktura

Na morfologické rovině je posloupnost slovních jednotek w-roviny rozdělena do vět. Anotace na této rovině spočívá v přiřazení několika atributů slovním jednotkám w-roviny, z nichž nejdůležitější jsou morfologické lemma a tag.

2.1.2 Fyzická realizace

Atribut `lemma` obsahuje lemma dané slovní jednotky. Reprezentuje jeho základní tvar a odpovídá jednoznačnému klíči příslušného záznamu v morfologickém slovníku. Atribut `tag` obsahuje morfologickou značku, která má 15 pozic a vyjadřuje slovní druh a hodnoty ostatních morfologických kategorií dané slovní jednotky. Atribut `id` obsahuje (v rámci PDT 2.0 jednoznačný) identifikátor této jednotky m-roviny, později používaný pro zpětnou referenci z analytické roviny (pro celkový přehled o propojení rovin, viz 2.1), a referenční atribut `w.rf` odkazuje zpět do w-roviny. Několik dalších atributů slouží k možným (ale vzácným) opravám a/nebo normalizacím týkajícím se w-roviny; nejdůležitější z nich je atribut `form`, který obsahuje správnou textovou podobu slovní jednotky (která se může lišit od textové podoby vyskytující se v původním textu z důvodu tiskových chyb, nesprávně rozdělených nebo spojených slov, špatného znaku pro desetinnou čárku v číslech nebo dalších technických problémů).

Příklad věty najdete v tabulce 2.2

2.1.3 Proces anotace

Morfologická rovina PDT byla anotována skupinou sedmi anotátorů. Anotace postupovala ve dvou oddělených fázích. Během první fáze byl každý text nejprve předzpracován automatickým morfologickým analyzátorem. Z jeho výstupu pak dva anotátoři nezávisle na sobě vybrali správné lemma a morfologický tag. Ve druhé, rozhodovací fázi byly všechny neshody těchto dvou anotátorů vyřešeny třetím anotátorem - rozhodčím.

Po oddělených kontrolách morfologické a syntakticko-analytické roviny byla provedena jejich společná revize. Soustředila se na vztah mezi analytickými funkcemi a morfologickými tagy, vztah mezi předložkami a pády závislých uzlů a nakonec na shodu v pádě, rodu a čísle mezi závislými a nadřízenými uzly.

2.2 Analytická rovina

Tato sekce stručně popisuje analytickou rovinu. Více informací najdete v textu Anotace na analytické rovině.

2.2.1 Logická struktura

Na analytické rovině je věta reprezentována orientovaným stromem s kořenem, s ohodnocenými hranami a uzly. Každý prvek morfologické roviny (viz sekce 2.1) odpovídá právě jednomu uzlu stromu a závislostní vztah mezi dvěma slovními jednotkami je vyjádřen hranou mezi příslušnými dvěma uzly. Typ vztahu je dán funkčním ohodnocením hrany. Většina hran reprezentuje závislostní vztah, ostatní odrážejí různé další lingvistické či technické jevy, např. koordinaci, apozici, interpunkci apod. Zaznamenáno je i lineární uspořádání uzlů, odpovídající pořadí slovních jednotek ve větě, což umožňuje „správné“ grafické zobrazení stromu.

2.2.2 Fyzická realizace

Každému uzlu je přiřazeno šest atributů (kromě technického kořene stromu, který jich má méně). Atribut `id` obsahuje identifikátor uzlu, jednoznačný v rámci PDT 2.0, na který se zpětně odkazuje z tektogramatické roviny (viz obrázek 2.1). Lineární uspořádání uzlů zachycuje atribut `ord`, obsahující pozici příslušné slovní jednotky ve větě. Z technických důvodů je analytická funkce hrany vyjádřena v atributu `afun` u uzlu na závislém konci hrany. Atributy `is_member` a `is_parenthesis_root` napomáhají

správné interpretaci koordinace, apozice a závorek. A konečně atribut `m.rf` spojuje uzel s odpovídajícím prvkem na morfologické rovině.

Příklad stromu najdete na obrázku 2.3

2.2.3 Proces anotace

Všechna analytická data byla anotována ručně týmem šesti anotátorů. Zpočátku museli anotátoři ručně vytvářet celý strom a rovněž ručně přiřazovat všechny analytické funkce. Později byly věty nejprve předzpracovány parserem a předběžné analytické funkce byly přiřazeny pravidlově založenou automatickou procedurou. Anotátoři však museli zkontrolovat a opravit výstup obou těchto automatických procedur, který byl často chybný.

Po skončení anotace byly na datech provedeny kontrolní testy. Příkladem takového testu je ověření platnosti tvrzení, že slovesný jmenný predikát (indikovaný analytickou funkcí `Pnom`) musí vždy přímo záviset na slovese *být*. Všechna porušení těchto pravidel/testů byla ručně prověřena a opravena.

2.3 Tektogramatická rovina

Tato sekce stručně popisuje tektogramatickou rovinu. Více informací najdete v textu Tektogramatická anotace PDT: Pokyny pro anotátory.

2.3.1 Logická struktura

Tektogramatická reprezentace věty zachycuje informace z následujících oblastí:

- **Tektogramatická struktura a funktory.** Každá věta je reprezentována jako orientovaný strom s kořenem, s ohodnocenými hranami a uzly. Strom zachycuje hloubkovou strukturu věty. Uzly zastupují pouze plnovýznamová slova (s několika výjimkami technické povahy). Narozdíl od analytické roviny, ne všechny morfologické prvky jsou na tektogramatické rovině reprezentovány jako uzly (např. tu chybějí předložky) a některé tektogramatické uzly neodpovídají žádnému morfologickému prvku (např. struktura obsahuje uzel reprezentující vynechaný subjekt v konstrukcích s nevyjádřeným podmětem (pro-drop constructions). K některým uzlům jsou připojeny *gramatémy* poskytující o uzlu informaci, kterou nelze odvodit ze struktury, funktoru či jiných atributů (např. číslo u podstatných jmen, modalitu a čas u sloves apod.). Hrany stromu reprezentují vztah mezi uzly, které spojují; typ vztahu je, podobně jako u analytické roviny, vyjádřen ohodnocením hrany. Ke každému uzlu reprezentujícímu sloveso nebo jistý typ podstatného jména je přiřazen valenční rámec (ve smyslu odkazu na prvek valenčního slovníku, viz sekce 3.8).

- **Aktuální členění (TFA, Topic–focus articulation).** Každému uzlu je na základě jeho kontextového zapojení přiřazena jedna ze tří hodnot: uzel může být kontextově zapojený, kontrastivně kontextově zapojený nebo kontextově nezapojený. Uzly v základové (topic) části věty jsou navíc seřazeny podle předpokládané výpovědní dynamičnosti.

- **Koreference.** V současné verzi anotace jsou zachyceny některé druhy koreferenčních vztahů mezi uzly, s rozlišením, o jaký druh vztahu se jedná (textový, gramatický nebo „druhá závislost“ doplňku).

2.3.2 Fyzická realizace

Každému nekořenovému uzlu tektogramatického stromu je přiřazeno 39 atributů; v závislosti na typu uzlu (určeného atributem `nodetype`) je však vyplněna jen určitá jejich podmnožina. Řada atributů je typu seznam nebo množina a obsahují více hodnot.

- **Tektogramatická struktura a funktory.** Podobně jako na analytické rovině, ke každému uzlu patří skupina atributů; na tektogramatické rovině je jich však mnohem více. Atribut `id` obsahuje v rámci PDT 2.0 jednoznačný identifikátor uzlu, atribut `functor` popisuje typ hrany vedoucí od uzlu k jeho předchůdci (hrana může reprezentovat jak závislostní vztah, tak i další technické jevy). Atribut `t_lemma` obsahuje tektogramatické lemma uzlu. Gramatémy jsou vyjádřeny skupinou 16 atributů, označených „předponou“ `gram` (např. `gram/verbmod` pro slovesnou modalitu). Další atributy slouží k zpětnému odkazování do analytické roviny (viz obrázek 2.1), jiné pro koordinaci a apozici, závorky, přímou řeč, citace apod.

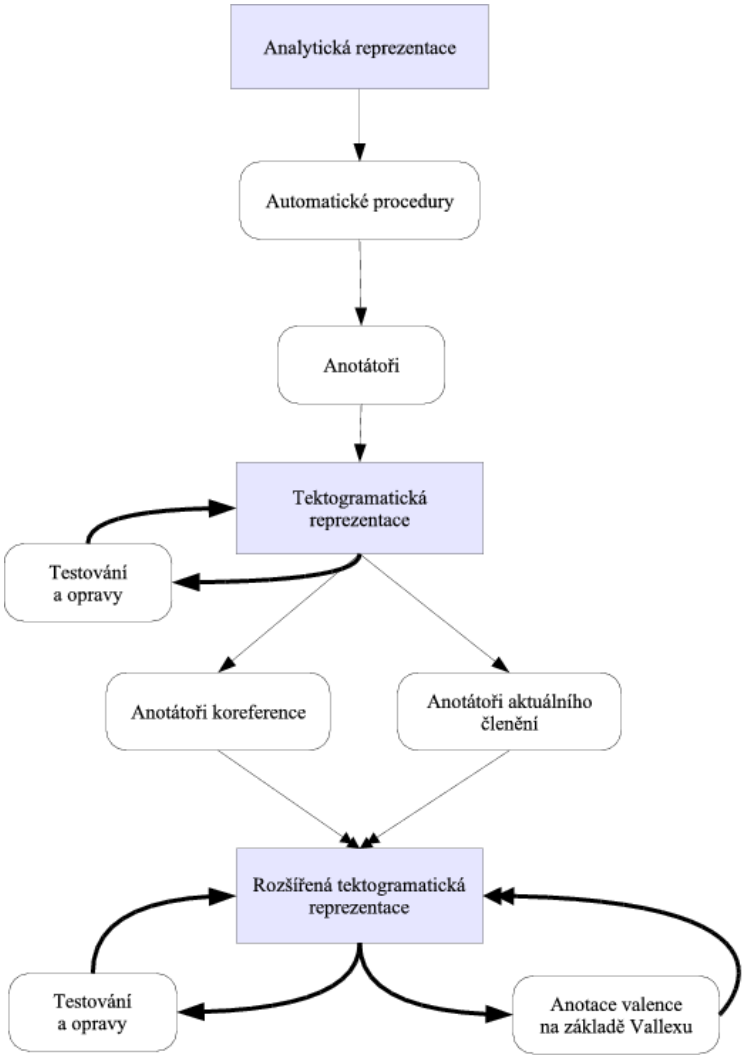
- **Aktuální členění.** Rozdělení uzlů na kontextově zapojené, kontrastivně kontextově zapojené a kontextově nezapojené je reprezentováno hodnotami atributu `tfa`. Číselný atribut `deepord` je použit pro hloubkové pořadí uzlů, založené na výpovědní dynamičnosti.

- **Koreference.** Atributy `coref_text.rf`, `coref_gram.rf` a `compl.rf` obsahují `id` koreferenčních uzlů příslušných typů. Atribut `coref_special` nese informaci o zvláštních případech koreference.

Příklad stromu najdete na obrázku 2.4.

2.3.3 Proces anotace

Jelikož je tektogramatická struktura rovněž založená na závislostních relacích, byly použity automatické postupy ke konverzi závislostních analytických stromů do provizorních stromů tektogramatického typu. Všechny vytvořené provizorní stromy pak byly zpracovány anotátory, kteří doplnili velké množství chybějících informací a opravili chyby. Koreference, aktuální členění a některé gramatémy byly anotovány odděleně. Všechna data pak byla zkontrolována množstvím poanotačních testů (viz sekce 4.7).



Obrázek 2.2: Schéma průběhu prací na datech a anotacích

Schéma průběhu prací na datech a anotacích je zobrazeno na obrázku 2.2. Silné šipky znamenají opakované operace, dvojité šipky značí procedury spojování, které byly použity, kdykoliv byla jedna data anotována na více podrovinách současně.

Tabulka 2.1: Ukázková věta

Některé	*kontury*	*problému*	*se*	*však*	*po*	*oživením*	*Havlovým*	*projevem*	*zdají*	*být*	*jasnější*	.

2.4 Ukázka anotace na třech rovinách

Ukázkovou větu vidíte v tabulce 2.1.

Anotace této věty na morfologické rovině je zachycena v tabulce 2.2. Všimněte si, že sedmý pád slova *oživení* byl změněn na šestý pád. Důvodem (jak je naznačeno elementem `form_change`) je tisková chyba.

Tabulka 2.2: Morfologická analýza ukázkové věty

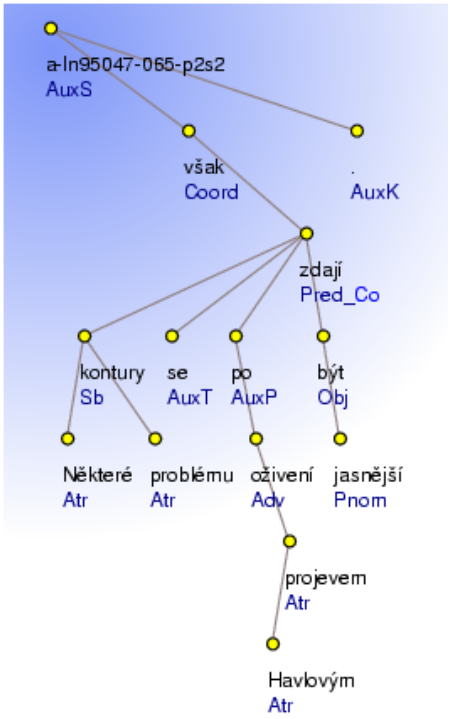
slovní forma	lemma	morfologický tag
Některé	*některý*	`PZFP1----------`
kontury	*kontura*	`NNFP1-----A----`
problému	*problém*	`NNIS2-----A----`
se	*se_^*(zvr._zájmeno/částice)	`P7-X4----------`
však	*však*	`J^------------`
po	*po*-1	`RR--6----------`
oživení	*oživení_^*(*3it)	`NNNS6-----A----`
Havlovým	*Havlův_;S_^*(*3el)	`AUIS7M---------`
projevem	*projev*	`NNIS7-----A----`
zdají	*zdát*	`VB-P---3P-AA---`
být	*být*	`Vf--------A----`
jasnější	*jasný*	`AAFP1----2A----`
.	.	`Z:------------`

Anotaci ukázkové věty na analytické rovině vidíte na obrázku 2.3. Všimněte si, že slovo *zdají* je označeno jako jediný člen koordinace. Tímto způsobem je na analytické rovině anotována koordinace s předchozí větou.

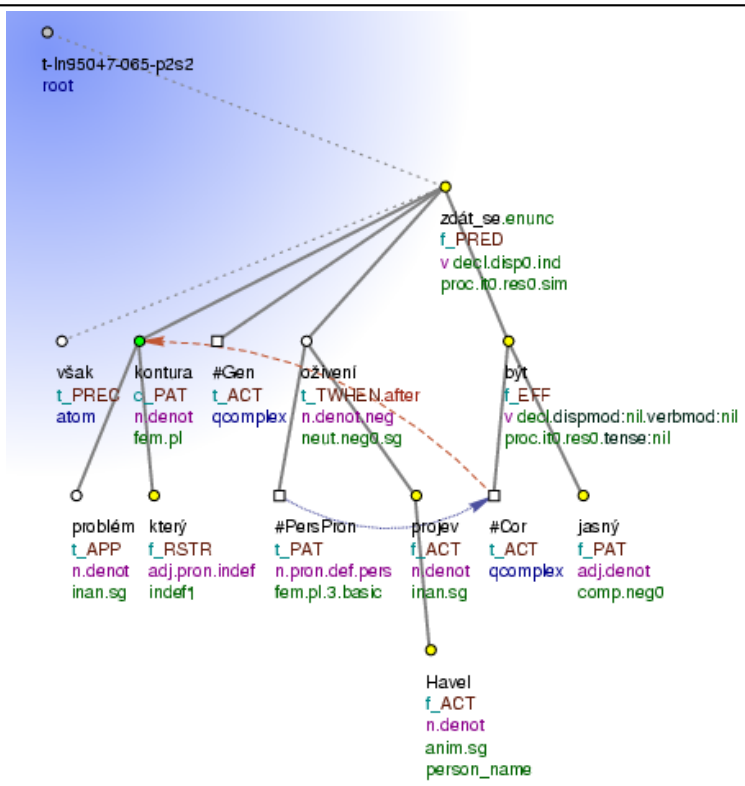
Anotaci ukázkové věty na tektogramatické rovině vidíte na obrázku 2.4.

Všimněte si, že slovo *však* už není koordinačním uzlem. Funktorem `PREC` je označeno jako slovo spojující tuto větu s větou předchozí. Dále si všimněte, že slovo *se* se stalo částí složené slovesné formy *zdát_se*, že zmizela předložka *po* (je však na ni odkazováno ze slova *oživení* a je základem hodnot funktoru a podfunktoru tohoto slova), že zájmeno *některý* má `t_lemma` *který* a jeho neurčitost je vyjádřena v hodnotách gramatémů `gram/sempos` a `gram/indeftype`, apod.

Více příkladů najdete v sekci 3.7.



Obrázek 2.3: Analytický strom ukázkové věty *Některé kontury problému se však po oživení Havlovým projevem zdají být jasnější.*



Obrázek 2.4: Tektogramatický strom ukázkové věty *Některé kontury problému se však po oživení Havlovým projevem zdají být jasnější.* (podrobné zobrazení)

Kapitola 3

Data

Vlastní data jsou jedinou částí PDT 2.0, kterou nelze stáhnout z webovských stránek PDT, <http://ufal.mff.cuni.cz/pdt2.0/>. Ke stažení je k dispozici jen část dat (ukázková data, viz sekce 3.7) a PDT-VALLEX (viz sekce 3.8). Chcete-li získat plná data (viz sekce 3.6), včetně aktualizace PDT 1.0 (viz 3.9), musíte si opatřit distribuční CD-ROM. Kapitola 7 popisuje, jak na to.

Data jsou umístěna v adresáři `data`.

3.1 Zdroje textů

Data v Pražském závislostním korpusu jsou anotované nezkrácené články z těchto novin a časopisů:

- Lidové noviny[1] (deník), ISSN 1213-1385, 1991, 1994, 1995
- Mladá fronta Dnes[2] (deník), 1992
- Českomoravský Profit[3] (ekonomický týdeník), 1994
- Vesmír[4] (populárně vědecký měsíčník), ISSN 1214-4029, Vesmír, s.r.o., 1992, 1993

Přehled množství dat z jednotlivých zdrojů najdete na obrázku 3.1.

Texty v elektronické podobě poskytl Ústav Českého národního korpusu.[5] Z původních zdrojů přicházely texty v různých podobách. Originální formátování bylo zachováno jen v některých případech, obecně bylo převzato jen rozdělení do dokumentů (článků) a odstavců.

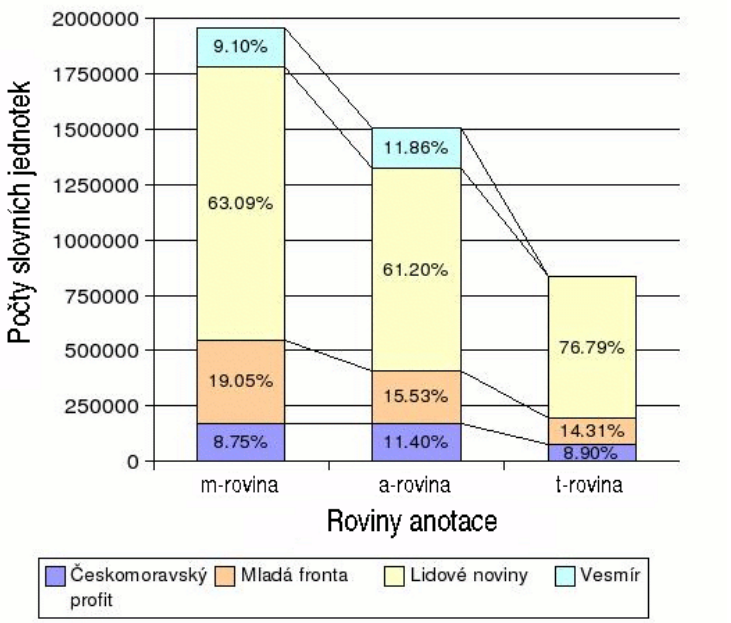
Originální data obsahovala z různých důvodů duplicity (většinou šlo o chybu). Pokud se opakovaly více než tři věty, byly odstraněny. Dále byla odstraněna téměř všechna vysoce četná neslovní data, jako přepisy šachových partií, tabulky výsledků sportovních utkání apod. Několik z nich jsme však zachovali, aby nám připomínaly svou existenci a abychom na nich předvedli navrhovaný (poněkud technický) způsob jejich anotace.

3.2 Rozdělení dat podle pokrytí anotacemi na jednotlivých rovinách

Anotace jednotlivých rovin nepokrývají celá data stejně. Čím vyšší rovina, tím méně dat na ní bylo anotováno. Důvod je zřejmý, anotace složitější roviny vyžaduje více lidské práce, a tedy více času a peněz. Existují ještě další technologické důvody: při určitém způsobu vývoje nástrojů pro vyšší roviny musí pro potřeby trénování existovat více dat na nižší rovině, jejíž anotace na vyšší rovině stejně nemůže být použita. Platí, že každý soubor, anotovaný na některé rovině, *je anotován rovněž* na všech rovinách nižších. Situaci ilustruje obrázek 3.2.

Další informace o rovinách najdete v kapitole 2. Informace o jmenné konvenci souborů, odrážející roviny anotace, najdete v sekci 3.5. Podrobné informace o množství dat najdete v sekci 3.6.

[1] <http://lidovky.centrum.cz/archivln/>
[2] <http://zpravy.idnes.cz/mfdnes.asp>
[3] <http://www.profit.cz/>
[4] <http://www.vesmir.cz/>
[5] <http://ucnk.ff.cuni.cz/>



Obrázek 3.1: Počet slovních jednotek (slov, čísel, interpunkce) z jednotlivých zdrojů

3.3 Rozdělení dat na trénovací a testovací

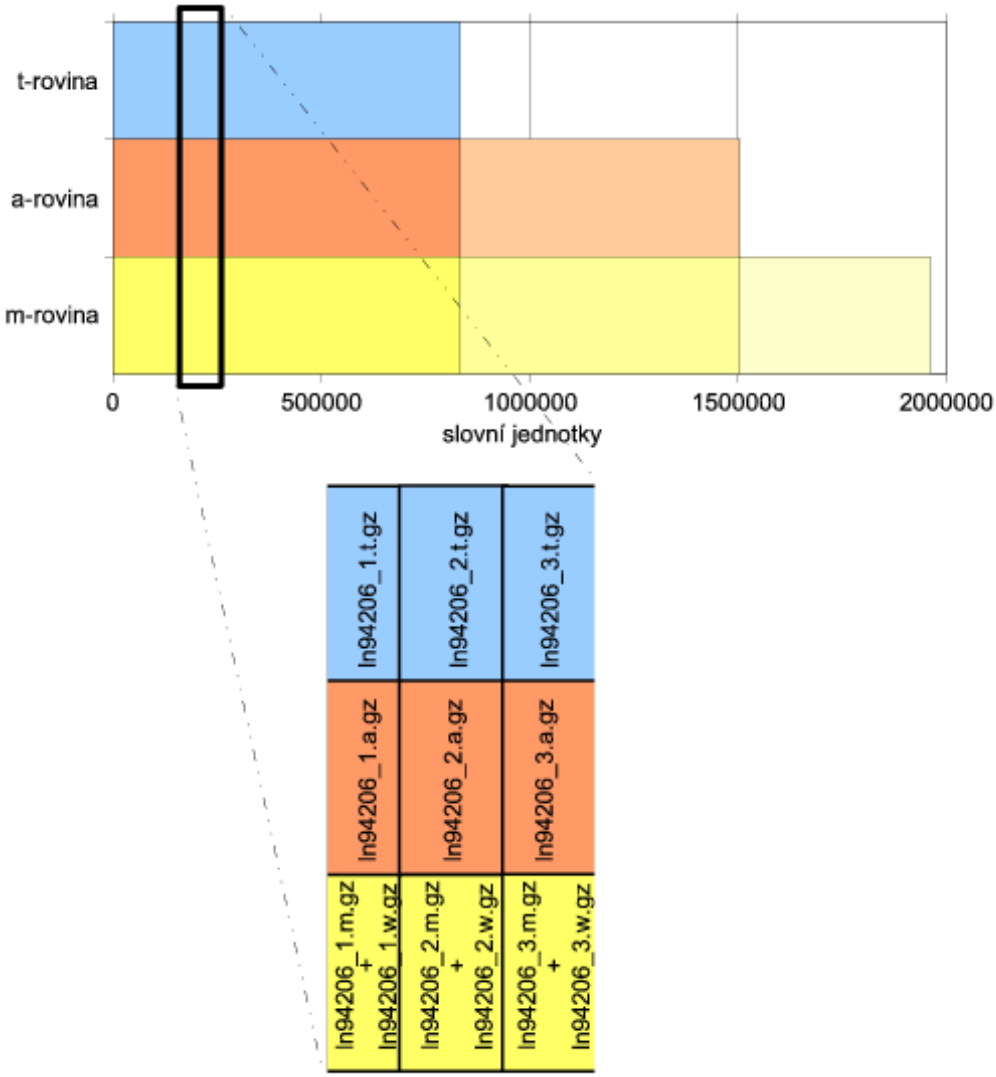
Data jsou rozdělena, jak je obvyklé, do tří skupin: trénovací data (*train*), vývojová testovací data (*dtest*) a evaluační testovací data (*etest*). Trénovací data tvoří přibližně 80% celkového množství dat, vývojová 10% a evaluační rovněž 10% (tento poměr platí na všech rovinách anotace).

Uživatelé mohou libovolně využívat trénovací data a prověřovat své hypotézy nebo nástroje na vývojových testovacích datech. Na evaluační testovací data by se neměli dívat nikdy, ta jsou určena *výhradně* pro evaluace. I tak by evaluační data měla být používána s rozvahou a co nejméně, neboť pozorování získaná opakovanými testy by mohla vést ke změně původní hypotézy či nástroje, a tak by evaluační data začala sloužit jako vývojová testovací data.

Poměr *train*/*dtest*/*etest* je přibližně stejný jako v PDT 1.0 (8:1:1), ale z různých důvodů nebylo zachováno staré rozdělení dat. Data v PDT 2.0 byla rozdělena následujícím způsobem: dokumenty morfologické roviny byly brány postupně a cyklicky rozdělovány, první byl vložen do množiny *train-1*, druhý do *train-2*, a tak dále až po *train-8*, devátý byl vložen do *dtest* a desátý do *etest*. Jedenáctý dokument připadl opět do *train-1* atd. (Rozdělení trénovací množiny do osmi podmnožin bylo provedeno proto, aby se zmenšil počet souborů v adresářích; existence deseti stejně velkých množin dat může navíc sloužit pro experimenty s křížovou validací.) Dokumenty anotované na ostatních rovinách připadly do stejných množin jako jejich morfologicky anotované verze. Díky sekvenčnímu výběru dokumentů pro anotaci tento algoritmus zaručuje, že poměr rozdělených dat zůstane i na vyšších rovinách téměř stejný (8:1:1), s malou odchylkou způsobenou rozdílem ve velikosti souborů.

Obrázek 3.3 ukazuje rozdělení dat. Algoritmus použitý k rozdělení zaručuje, že každý soubor patří do stejné množiny (*train*, *dtest*, *etest*) na všech rovinách, na kterých je anotován. (Podrobné informace o množství dat najdete v sekci 3.6.)

Poznamenejme, že uživatel, který provádí experiment např. na datech a-roviny a tento experiment se netýká t-roviny, by měl použít takové rozdělení dat, které nebere v úvahu, zda jsou daná data anotována na t-rovině či ne. Díky tomu je např. množina *etest* na a-rovině ve skutečnosti složena ze dvou částí, jak je vidět na obrázku 3.3 (dvě svisle šrafované oblasti ve středním sloupci). Podobně je množina *train-1* m-roviny složena ze tří částí. O těchto rozděleních pojednává rovněž sekce 3.6.



Obrázek 3.2: Rozdělení dat do rovin

3.4 Formáty dat

Hlavním formátem dat v PDT 2.0 je formát nazvaný *PML*, který je založený na XML[6]. Během vývoje PDT vznikly a byly používány ještě dva další formáty dat. Formát *FS* byl vytvořen pro vyhledávací program `Netgraph` (přísně vzato vlastně pro jeho předchůdce, editor `Graph`). Formát zvaný *CSTS*, založený na SGML, byl hlavním formátem dat v PDT 1.0. Nyní je používán jen jako přechodný formát pro kompatibilitu se staršími nástroji pro zpracovnání jazyka (taggery, parsery, ...).

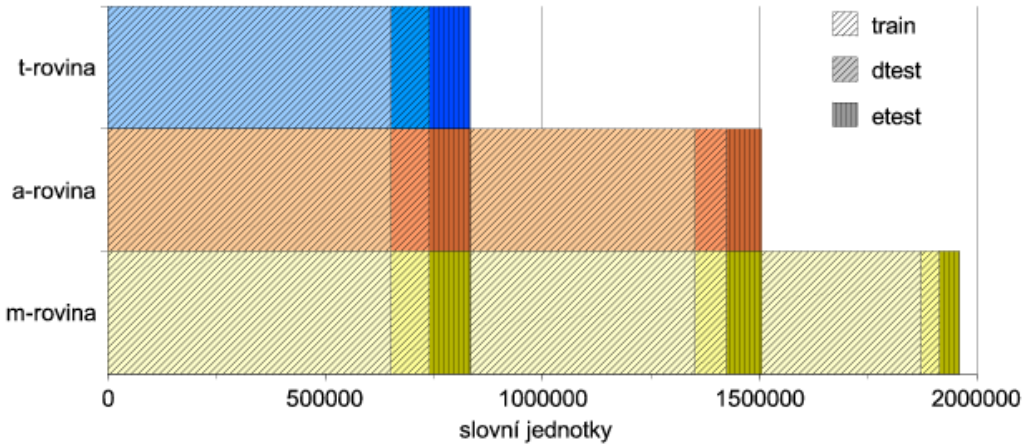
Informace o konverzích mezi těmito formáty najdete v sekci 4.4.1.

3.4.1 PML

PML („Prague Markup Language“) je formát dat založený na XML, navržený pro reprezentaci bohaté lingvistické anotace textů, jako jsou morfologické značkování, závislotní stromy apod. PML je probíhající projekt ve své rané fázi. Přesto je již dostatečně pokročilý, aby umožnil přiměřenou a snadnou reprezentaci dat v PDT 2.0. Následující text obsahuje stručný přehled hlavních vlastností PML. Podrobné informace o tomto formátu najdete v dokumentaci PML. Informace o tom, jak jsou data PDT 2.0 reprezentována v PML, najdete v příručce anotačních značek PDT 2.0.

V PML se mohou jednotlivé oddělené roviny anotace překrývat a mohou být konzistentně propojeny jak mezi sebou, tak i s dalšími zdroji dat. Každá rovina anotace je popsána v souboru *PML schéma*, který je jakousi formalizací abstraktního anotačního schématu pro tu konkrétní rovinu anotace. PML schéma

[6] <http://www.w3.org/XML/>



Obrázek 3.3: Rozdělení dat na trénovací a testovací množiny

popisuje, které elementy se na dané rovině vyskytují, jak jsou spojovány, vnořovány a strukturovány, hodnoty jakého typu se v nich mohou vyskytovat a jakou roli hrají v anotačním schématu (tato informace o *PML-roli* může být využívána i aplikacemi ke správnému určení způsobu zobrazení PML dat uživateli). Z PML schématu mohou být automaticky generována další schémata, jako je Relax NG[7], díky čemuž může být konzistence dat ověřena pomocí běžných nástrojů pro XML (XSLT styl pro konverzi PML schématu do Relax NG je k dispozici v tools/pml/pml2rng.xsl).

Každý PML soubor začíná hlavičkou, odkazující na PML schéma souboru. V hlavičce jsou uvedeny všechny externí zdroje, na které je z tohoto souboru odkazováno, spolu s několika dalšími informacemi, potřebnými pro správné vyhodnocení odkazů. Zbytek souboru obsahuje vlastní anotaci.

Anotace je vyjádřena pomocí XML elementů a atributů, pojmenovaných a použitých v souladu s příslušným PML schématem. XML elementy všech souborů patří do vyhrazeného jmenného prostoru `http://ufal.mff.cuni.cz/pdt/pml/`. Formát PML poskytuje jednotnou reprezentaci většiny běžných anotačních konstrukcí, jako jsou struktury atribut-hodnota, seznam alternativních hodnot určitého typu (atomického nebo dále strukturovaného), odkazy v rámci PML souboru, odkazy mezi různými PML soubory (v PDT 2.0 použité k odkazům mezi rovinami) nebo do dalších externích zdrojů typu XML. V současné verzi nabízí PML i omezenou podporu XML elementů se smíšeným obsahem. Abychom se vyhnuli možné záměně s atributy XML, nazýváme obvykle atributy sktruktury atribut-hodnota *prvky*.

Anotace PDT 2.0 je rozdělena do čtyř rovin, naskládaných jedna na druhou, a to roviny slovní, morfologické, analytické a tektogramatické (viz kapitola 2). Každá z těchto rovin má vlastní PML schéma.

Tektogramatické a analytické stromy jsou v PML reprezentovány běžným způsobem jako vnořené struktury atribut-hodnota. Uzel stromu je reprezentován strukturou atribut-hodnota s PML-rolí `#NODE`. Každý uzel má prvek s PML-rolí `#CHILDNODES`, který obsahuje seznam přímých potomků daného uzlu. Technický kořen závislostních stromů v PDT 2.0 slouží zvláštním pomocným účelům, a proto je jeho struktura odlišná od ostatních uzlů (má jiné prvky).

Obsáhlé informace o reprezentaci čtyř anotačních rovin v PML najdete v Příručce anotačních značek PDT 2.0. PML a Relax NG schémata pro čtyři anotační roviny najdete v adresáři `data/schemas`.

3.4.2 Perl Storable Format

Formát PML, založený na XML, je primárním formátem dat v PDT 2.0. Při práci s ním však nástroje `TrEd` a `btred`, založené na Perlu, spotřebují mnoho času načítáním dat a jejich převodem do vnitřní paměťové reprezentace. Této časově náročné transformaci se lze vyhnout využitím formátu `pls.gz` (Perl Storable Format). Jde o binární datový formát, který přímo odráží vnitřní paměťovou reprezentaci dat v Perlu. Jeho ukládání a zpětné načítání je tedy mnohem rychlejší. Není ale založen na XML, a nelze jej tedy snadno použít jinými nástroji.

[7] <http://www.relaxng.org/>

3.4.3 FS

Formát FS („feature structure“) je formát souborů pro reprezentaci stromů, jejichž uzly jsou sktruktury atribut-hodnota. Může být chápán jako „meta formát“, podobně jako SGML nebo XML. Konkrétní použití tohoto formátu je plně specifikováno deklarací atributů v hlavičce FS souboru (hlavička FS souboru tak hraje podobnou roli jako DTD u SGML souboru).

FS soubor začíná deklarací atributů. Každá řádka deklarace sestává ze znaku `@`, vlastnosti atributu, mezery a jména atributu. Např. vlastnost `O`, „obligatory“, označuje povinný atribut, tedy atribut, jehož hodnota musí být u každého uzlu neprázdná. Vlastnost `L`, „list“, označuje výčtový atribut, tedy atribut, jehož hodnota u každého uzlu (pokud je neprázdná) musí být jednou z hodnot uvedených v seznamu následujícím za jménem atributu v hlavičce. Úplný popis najdete ve specifikaci FS formátu.

Deklarační hlavička končí prázdným řádkem, po němž následují popisy stromů anotace. Každý strom začíná na novém řádku. Stromy jsou popsány v obvyklé závorkové notaci, tj. po popisu uzlu následuje seznam jeho přímých potomků, uzavřený v závorkách. Jednotliví potomci jsou odděleni čárkou. Popis každého uzlu je uzavřen v hranatých závorkách a sestává ze seznamu dvojic *`atribut=hodnota`*, oddělených čárkou. Pokud je atribut v hlavičce deklarován jako poziční (`P`), může být u uzlu určen jen svou hodnotou a jeho jméno je odvozeno z předchozích známých atributů a z pořadí atributů v hlavičce.

3.4.4 CSTS

CSTS („Czech sentence tree structure“), formát založený na SGML, byl hlavním formátem dat v PDT 1.0. Ačkoliv byl v PDT 2.0 nahrazen PML, některé nástroje jej stále výhradně používají. CSTS může reprezentovat jen morfologickou a analytickou anotaci (abychom byli přesní, jeho definice obsahuje i několik elementů vztahujících se k tektogramatické anotaci, ale není schopen plného popisu t-roviny). Velmi doporučujeme používat místo něj PML (viz sekce 3.4.1), kdykoliv je to možné. To se týká zejména nových nástrojů. Více informací najdete v úplném popisu CSTS a jeho DTD souboru.

3.5 Konvence pojmenování souborů

Data v PDT 2.0 jsou distribuována ve formátu PML (viz popis PML v sekci 3.4.1). Každý datový soubor odpovídá jednomu anotovanému dokumentu. Základem jeho jména je identifikátor dokumentu (indikuje také zdroj dokumentu, viz sekce 3.1: `ln*` označuje Lidové noviny, `mf*` označuje Mladou frontu Dnes, `vesm*` označuje Vesmír a `cmpr*` označuje Českomoravský profit). Přípona souboru vyjadřuje rovinu anotace dokumentu (`.w` označuje w-rovinu, `.m` označuje m-rovinu, `.a` označuje a-rovinu a `.t` označuje t-rovinu). (Popis rovin najdete v kapitole 2.)

Každý soubor obsahující anotaci dokumentu na nějaké rovině odpovídá jedna ku jedné souborům obsahujícím anotace nižších rovin téhož dokumentu a obsahuje reference do těchto souborů. Z tohoto důvodu by soubory neměly být přejmenovány. Z nižších rovin anotace do vyšších rovin odkazy nevedou. Přehled propojení rovin najdete na obrázku 2.1.

Příklad: `cmpr9406_001.a.gz` označuje soubor (zkomprimovaný gzip-em) obsahující a-rovinu anotace dokumentu *cmpr9406_001* (pocházejícího z Českomoravského profitu). Ze souboru vedou odkazy do souborů `cmpr9406_001.m.gz` a `cmpr9406_001.w.gz`; z těchto údajů však nelze odvodit existenci souboru `cmpr9406_001.t.gz`.

Podle jména souboru se nepozná, zda soubor patří do trénovací nebo testovací množiny. To je dáno umístěním souboru v adresářové struktuře, viz sekce 3.3.

Ze jmen souborů jsou odvozena také jména identifikátorů vět a prvků vět, obsažených v těchto souborech. Každý identifikátor je jedinečný v rámci celého korpusu.

3.6 Plná data

Plná verze dat PDT 2.0 je k dispozici oprávněným uživetelům, kteří CD-ROM PDT 2.0 získali z Linguistic Data Consortium (viz kapitola 7). Malá ukázka dat může být volně stažena z internetu (viz sekce 3.7).

Plná verze dat PDT 2.0 sestává ze 7 110 ručně anotovaných textových dokumentů, obsahujících celkem 115 844 vět s 1 957 247 slovními jednotkami (slovy, čísly, interpunkcí). Všechny tyto dokumenty jsou anotovány na m-rovině. 75% dat m-roviny je anotováno rovněž na a-rovině (5 330 dokumentů,

87 913 vět, 1 503 739 slovních jednotek). 59% dat a-roviny je anotováno také na t-rovině (tj. 45% dat m-roviny; 3 165 dokumentů, 49 431 vět, 833 195 slovních jednotek).

Plná data ve formátu PML jsou uložena v adresáři `data/full` na CD-ROM PDT 2.0. (Pro rychlejší zpracování nástroji založenými na TrEdu jsou plná data, anotovaná alespoň na a-rovině, převedena rovněž do formátu Perl Storable Format; tato data jsou uložena v adresářích `data/binary/amw` a `data/binary/tamw`.) Datové soubory jsou rozděleny podle této dvoustupňové hierarchie:

- První větvení odpovídá nejvyšší vrstvě anotace (viz kapitola 2) dostupné pro daný dokument:
  - `data/full/tamw/` – dokumenty anotované na všech rovinách,
  - `data/full/amw/` – dokumenty anotované pouze na m-rovině a a-rovině,
  - `data/full/mw/` – dokumenty anotované pouze na m-rovině.
- Obsah každého z těchto tří adresářů je dále rozdělen do deseti přibližně stejně velkých částí (viz sekce 3.3). Osm z nich slouží pro trénovací účely (`train-1/` až `train-8/`), jedna pro vývojové testy (`dtest/`) a jedna pro evaluační testy (`etest/`).

Přestože jsou data takto rozdělena do třiceti adresářů, zůstává množství souborů v jednotlivých adresářích stále značné. To je způsobeno částečně tím, že počet fyzických souborů (v porovnání s počtem původních textových dokumentů) je v případě `tamw` dat násoben čtyřmi (pro každý dokument jsou v adresáři čtyři soubory, obsahující jeho anotaci na jednotlivých rovinách, viz sekce 3.5), třemi v případě `amw` dat a dvěma u `mw` dat. Tak se celkový počet datových souborů rovná 4 x 3 165 + 3 x 2 165 + 2 x 1 780 = 22 715. Například adresář `data/full/tamw/train-3/` obsahuje 4 x 317 = 1 268 datových souborů.

Poznamenejme, že se žádný datový soubor nevyskytuje v adresáři `data/full/` dvakrát (např. soubory `*.m` z `data/full/amw/` se již neobjeví v `data/full/mw/`). Všech třicet podadresářů má vzájemně se nepřekrývající obsah, soubory v těchto adresářích obsahují anotace různých textů.

Podrobný rozpis množství dat v jednotlivých adresářích, rozdělených podle výše uvedených zásad, najdete v tabulkách 3.1, 3.2 a 3.3.

Tabulka 3.1: Data anotovaná na všech vrstvách (`tamw`).

`tamw`	train	dtest	etest	celkem
Umístění na CD-ROM v `data/full/`	`tamw/train-1/` ... `tamw/train-8/`	`tamw/dtest/`	`tamw/etest/`	`tamw/*/`
# dokumentů	2 533 (80,0%)	316 (10,0%)	316 (10,0%)	3 165 (100,0%)
# vět	38 727 (78,3%)	5 228 (10,6%)	5 476 (11,1%)	49 431 (100,0%)
# slovních jednotek	652 544 (78,3%)	87 988 (10,6%)	92 663 (11,1%)	833 195 (100,0%)

Tabulka 3.2: Data anotovaná pouze na m-rovině a a-rovině (`amw`).

`amw`	train	dtest	etest	celkem
Umístění na CD-ROM v `data/full/`	`amw/train-1/` ... `amw/train-8/`	`amw/dtest/`	`amw/etest/`	`amw/*/`
# dokumentů	1 731 (80,0%)	217 (10,0%)	217 (10,0%)	2 165 (100,0%)
# vět	29 768 (77,4%)	4 042 (10,5%)	4 672 (12,1%)	38 482 (100,0%)
# slovních jednotek	518 647 (77,3%)	70 974 (10,6%)	80 923 (12,1%)	670 544 (100,0%)

Ti, kdo chtějí pracovat pouze s daty m-roviny nebo a-roviny bez ohledu na to, zda jsou dané dokumenty anotovány také na vyšších rovinách, by měli použít jiné rozdělení. Například při experimentech se všemi daty m-roviny by měla trénovací data sestávat ze všech souborů `data/full/{tamw,amw,mw}/train-[1-8]/*m.gz`.

Tabulka 3.3: Data anotovaná pouze na m-rovině (mw).

mw	train	dtest	etest	celkem
Umístění na CD-ROM v data/full/	mw/train-1/ ... mw/train-8/	mw/dtest/	mw/etest/	mw/*/
# dokumentů	1 422 (79,9%)	179 (10,1%)	179 (10,1%)	1 780 (100,0%)
# vět	22 333 (80,0%)	2 610 (9,3%)	2 988 (10,7%)	27 931 (100,0%)
# slovních jednotek	364 640 (80,4%)	42 689 (9,4%)	46 179 (10,2%)	453 508 (100,0%)

Počty všech dokumentů anotovaných na m-rovině (bez ohledu na to, zda existují jejich anotace na a-rovině a t-rovině) jsou sečteny v tabulce 3.4. Všechny dokumenty anotované na a-rovině (bez ohledu na to, zda existuje jejich anotace na t-rovině) jsou posčítány v tabulce 3.5.

Tabulka 3.4: Alternativní rozdělení: Všechny dokumenty anotované na m-rovině (sjednocení tamw, amw a mw).

all_m	train	dtest	etest	celkem
Umístění na CD-ROM v data/full/	*/train-1/ ... */train-8/	*/dtest/	*/etest/	*/*/
# dokumentů	5 686 (80,0%)	712 (10,0%)	712 (10,0%)	7 110 (100,0%)
# vět	90 828 (78,4%)	11 880 (10,3%)	13 136 (11,3%)	115 844 (100,0%)
# slovních jednotek	1 535 831 (78,5%)	201 651 (10,3%)	219 765 (11,2%)	1 957 247 (100,0%)

Tabulka 3.5: Alternativní rozdělení: Všechna data anotovaná na a-rovině (sjednocení tamw a amw).

all_a	train	dtest	etest	celkem
Umístění na CD-ROM v data/full/	*a*/train-1/ ... *a*/train-8/	*a*/dtest/	*a*/etest/	*a*/*/
# dokumentů	4 264 (80,0%)	533 (10,0%)	533 (10,0%)	5 330 (100,0%)
# vět	68 495 (77,9%)	9 270 (10,5%)	10 148 (11,5%)	87 913 (100,0%)
# slovních jednotek	1 171 191 (77,9%)	158 962 (10,6%)	173 586 (11,5%)	1 503 739 (100,0%)

Není jistě třeba dodávat, že každý zveřejněný experiment provedený na datech PDT 2.0 by měl obsahovat informaci o tom, jaká část dat byla pro jaký účel v experimentu použita.

Práci s velkým počtem datových souborů pomohou usnadnit předgenerované seznamy souborů, umístěné jako samostatné soubory v adresáři data/filelists/; jsou užitečné nejen při práci s programy tred/btred/ntred, ale i na příkazové řádce, kde odstraní problém s příliš velkým počtem argumentů. Připraveno je pouze několik základních seznamů souborů, uživatel má možnost snadno si vytvořit jakýkoliv jemu vyhovující další seznam souborů, odpovídající libovolné podmnožině všech dat (viz též tutoriál k btred/ntredu).

3.7 Ukázková data

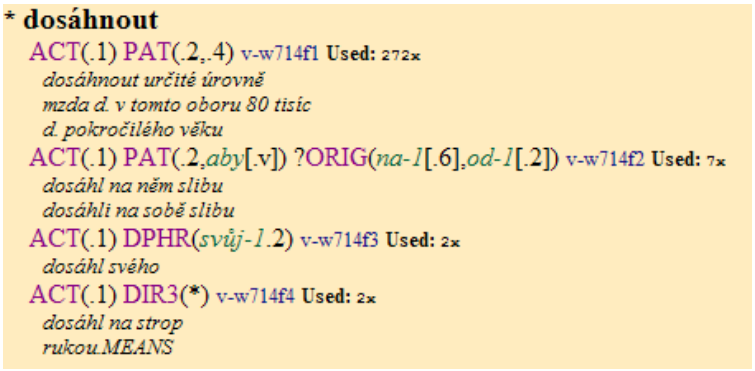
Malá část plných dat je k dispozici ke stažení na internetu (připomeňme, že postup k získání plné verze dat najdete v kapitole 7). Data jsou rozdělena do deseti skupin (*sample0* až *sample9*) přibližně

po 50 větách. Každá skupina sestává ze čtyř souborů (`sampleX.w.gz`, `sampleX.m.gz`, `sampleX.a.gz` a `sampleX.t.gz`); přípona souboru vyjadřuje rovinu anotace (viz sekce 3.5). Ukázková data jsou tvořena úseky vybranými náhodně z plných dat (viz sekce 3.6).

Ukázková data jsou umístěna v adresáři `data/sample`. Ve stejném adresáři najdete i archív všech ukázkových souborů. Pokud si nemůžete nebo nechcete nainstalovat nástroje pro práci s daty ve formátu PML (viz kapitola 4), můžete si ukázková data snadno prohlédnout v podobě webovských stránek.

3.8 PDT-VALLEX

PDT 2.0 obsahuje také omezenou lexikálně-sémantickou anotaci, která nově provazuje hloubkovou a povrchovou syntax a morfologii pomocí *valenčního slovníku*, zvaného *PDT-VALLEX*. Valenční slovník najdete v adresáři `data/pdt-vallex` ve formátu XML (viz jeho popis) nebo si ho můžete prohlédnout v podobě webovských stránek— viz zobrazení jedné jeho položky na obrázku 3.4.



Obrázek 3.4: Ukázka položky PDT-VALLEXu ve formátu pro zobrazení

Položky PDT-VALLEXu obsahují jednotlivé *významy* sloves a některých slovesných podstatných a přídavných jmen, které se vyskytují v korpusu. Každý význam obsahuje *valenční rámec* se sémantickou, syntaktickou a morfologickou informací o jeho sémanticky povinných a/nebo volitelných závislých členech.

Každý valenční rámec obsahuje nula nebo více *valenčních pozic*. Každá pozice má syntaktickou nebo sémantickou značku (např. `ACT`, `PAT`, `ADDR`, `LOC`, `AIM`, `CRIT`, `BEN` atd.; více obecných informací o tektogramatickém anotování najdete v textu Tektogramatická anotace PDT: pokyny pro anotátory), a je označena buď jako povinná (obligatorní) nebo jako volitelná (fakultativní). Pozice navíc obsahují povrchově syntaktickou a morfologickou informaci o své povrchové realizaci (výrazu), jako je morfologický pád, předložka, která má být použita s příslušnou lexikální jednotkou, nebo (v případě frázémů) celý syntaktický podstrom, který frazém na povrchu vytváří.

Nejdůležitější vlastností PDT-VALLEXu však je, že každý výskyt slovesa či slovesného podstatného jména v PDT 2.0 je provázán (s použitím zvlášního referenčního atributu) z korpusu na položku slovníku, čímž je vlastně provedena anotace významů těchto slov (*word sense annotation*). Položky slovníku, jejich značky, obligatornost/fakultativnost a povrchové morfologické formy byly zkontrolovány, aby plně souhlasily se všemi daty korpusu na všech rovinách anotace.

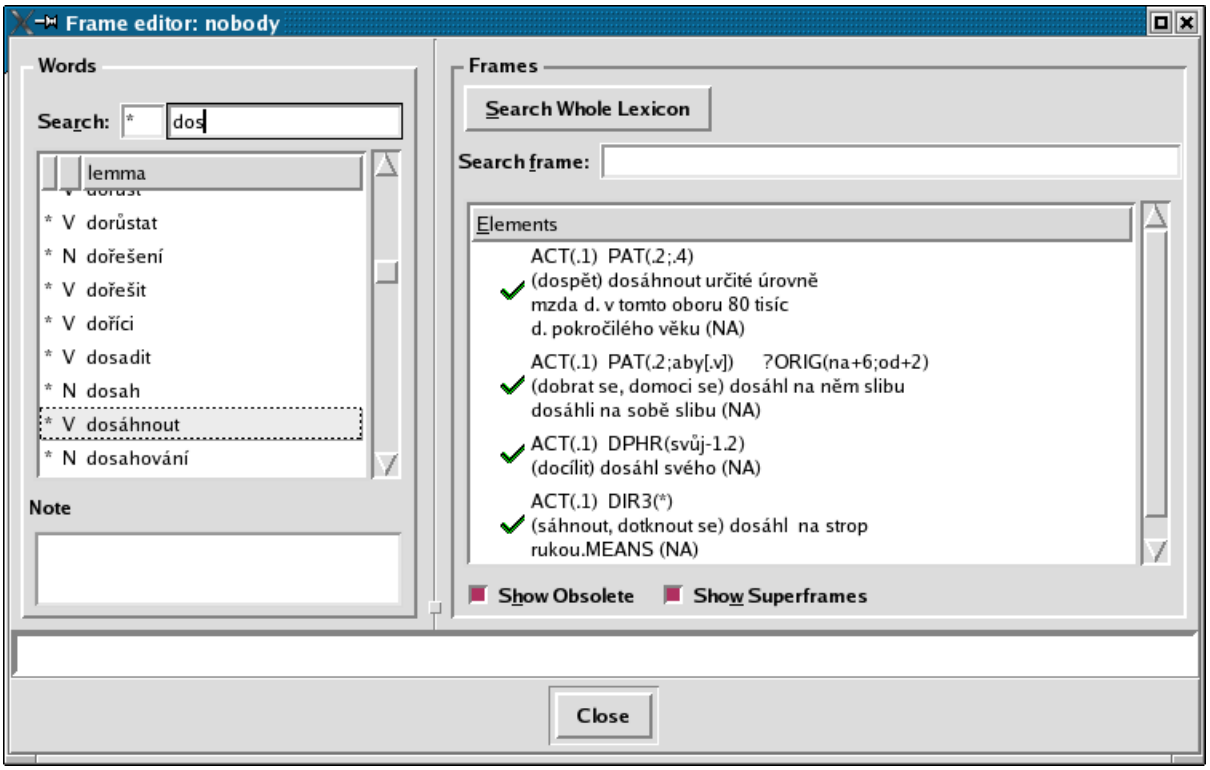
K dispozici jsou i nástroje, které umožňují spojení mezi korpusem a slovníkem využít (umožňují průběžné prohlížení, vyhledávání a editaci v editoru `TrEd`, viz obrázek 3.5).

3.9 Aktualizace PDT 1.0

Hlavní rozdíl mezi PDT 1.0 a PDT 2.0 spočívá v přítomnosti anotace na tektogramatické rovině (viz sekce 2.3). Mnoha změn však doznaly i nižší roviny. Pro uživatele PDT 1.0 jsme připravili aktualizaci dat, která k originálním datům přidává všechny změny a nové informace. Aktualizace je umístěna v adresáři `data/pdt1.0-update`. Aktualizační balík je určen pouze pro formát CSTS, staré FS soubory jím nemohou být aktualizovány.

Změny zahrnují:

- opravy různých chyb na morfologické a analytické rovině,



Obrázek 3.5: PDT-VALLEX v editoru `TrEd`

- opravy překlepů,
- přidání ruční morfologické anotace ve všech souborech.

Požadavky pro aplikaci aktualizačního balíku. Pro aktualizaci dat potřebujete dva GNU nástroje, `gunzip` a `patch`. *V linuxových distribucích* bývají tyto nástroje obvykle již instalovány. *Používáte-li MS Windows*, stáhněte si z internetu GNU `patch`[8] (jiné verze by nemusely fungovat). `gunzip` pro Windows můžete použít jak ve verzi z distribuce Cygwin[9], tak i ze stránek GNU[10]. Na CD-ROM PDT 2.0 najdete kopii nástroje `gunzip.exe` z distribuce Cygwin v adresáři `tools/tred/bin/`.

Aplikování aktualizačního balíku na všechny datové adresáře. PDT 1.0 CD-ROM obsahuje několik překrývajících se (ve smyslu pevných odkazů) podmnožin dat v podadresářích adresáře ***`PDT_1.0_CD-ROM`***`/Corpora/PDT_1.0/Data/`. Aktualizovány musejí být všechny kromě `fs/` a `fs-am/`. Pro současnou aktualizaci všech těchto podadresářů na *Linuxu* použijte skript `data/pdt1.0-update/linux-apply-patch.sh`. Skript spusťte a pokračujte podle instrukcí. V případě *MS Windows* nemůžeme poskytnout zaručený způsob, jak aktualizaci automaticky aplikovat na všechny datové adresáře. Postupujte podle instrukcí níže uvedených a aktualizujte jednotlivé adresáře postupně.

Aplikování aktualizačního balíku na jeden datový adresář.

1. Zkopírujte soubory z vybraného podadresáře `Corpora/PDT_1.0/Data/` (kromě podadresářů `fs/` a `fs-am/`) na disku PDT 1.0 do nějakého nového pracovního adresáře.
2. Přejděte do tohoto adresáře: **cd *`pracovní_adresář`***
3. Rozbalte všechny soubory: **gunzip *.gz**
4. Aplikujte aktualizační balík: **gunzip -c *`PDT_2.0_CD-ROM`*/data/pdt1.0-update/pdtpatch.gz | patch -p1 -t**

   Přepínač *`-t`* je vyžadován v případě aktualizace neúplných adresářů, tedy např. adresářů, které neobsahují všechny datové soubory PDT 1.0. Tento přepínač říká nástroji `patch`, že má přeskočit všechny neexistující soubory bez dotazování se uživatele. V *MS Windows* přidejte k příkazu **patch** přepínač *`--binary`*, jinak by aktualizace mohla selhat.

[8] <http://gnuwin32.sourceforge.net/packages/patch.htm>
[9] <http://cygwin.com>
[10] <http://gnuwin32.sourceforge.net/packages/gzip.htm>

Kapitola 4

Nástroje

Jedním z hlavních cílů PDT 2.0 (viz sekce 1.1) je poskytnout lingvistům velké množství skutečně reálných příkladů (nejen) jevů dříve popsaných v řadě teoretických prací zabývajících se závislostí, tektogramatickým popisem a přístupem funkčně-generativního popisu obecně. Využití takového korpusu by však bylo jen omezené, kdyby nebyl doplněn pohodlným nástrojem pro prohledávání.

Existuje přirozeně řada způsobů, jak korpus prohledávat. Velmi pokročilé vyhledávání umožňuje například nástroj `btred/ntred`, vyžaduje však jistou programátorskou dovednost (konkrétně znalost jazyka Perl a rozhraní `btred/ntredu`). Většině „běžných" uživatelů doporučujeme *`Netgraph`*, nástroj navržený a vytvořený právě pro snadné prohledávání PDT 1.0 a PDT 2.0.

4.1 Vyhledávání v korpusu: `Netgraph`

`Netgraph` je aplikace typu klient-server, která umožňuje prohledávat PDT 2.0 současně několika uživateli, připojenými přes internet. `Netgraph` je navržený tak, aby prohledávání bylo co nejjednodušší a intuitivní, při zachování vysoké síly dotazovacího jazyka.

Komunikace mezi dvěma částmi `Netgraphu`, klientem a serverem, probíhá přes internet. Server prohledává korpus, který je umístěn na stejném počítači či lokální síti jako server. Klient slouží jako grafické rozhraní pro uživatele a může být umístěn kdekoliv na internetu. Posílá serveru dotazy a přijímá zpátky výsledky. Server a klient mohou být samozřejmě umístěny i na jednom počítači.

`Netgraph` server je napsán v C a C++ a běží v operačním systému Linux, dalších systémech unixového typu a na Apple Mac OS. Existuje i experimentální verze pro MS Windows. Umožňuje nastavit uživatelská konta s různými přístupovými právy. Korpus, určený k prohledávání `Netgraphem`, musí být ve formátu FS a v kódování UTF-8.

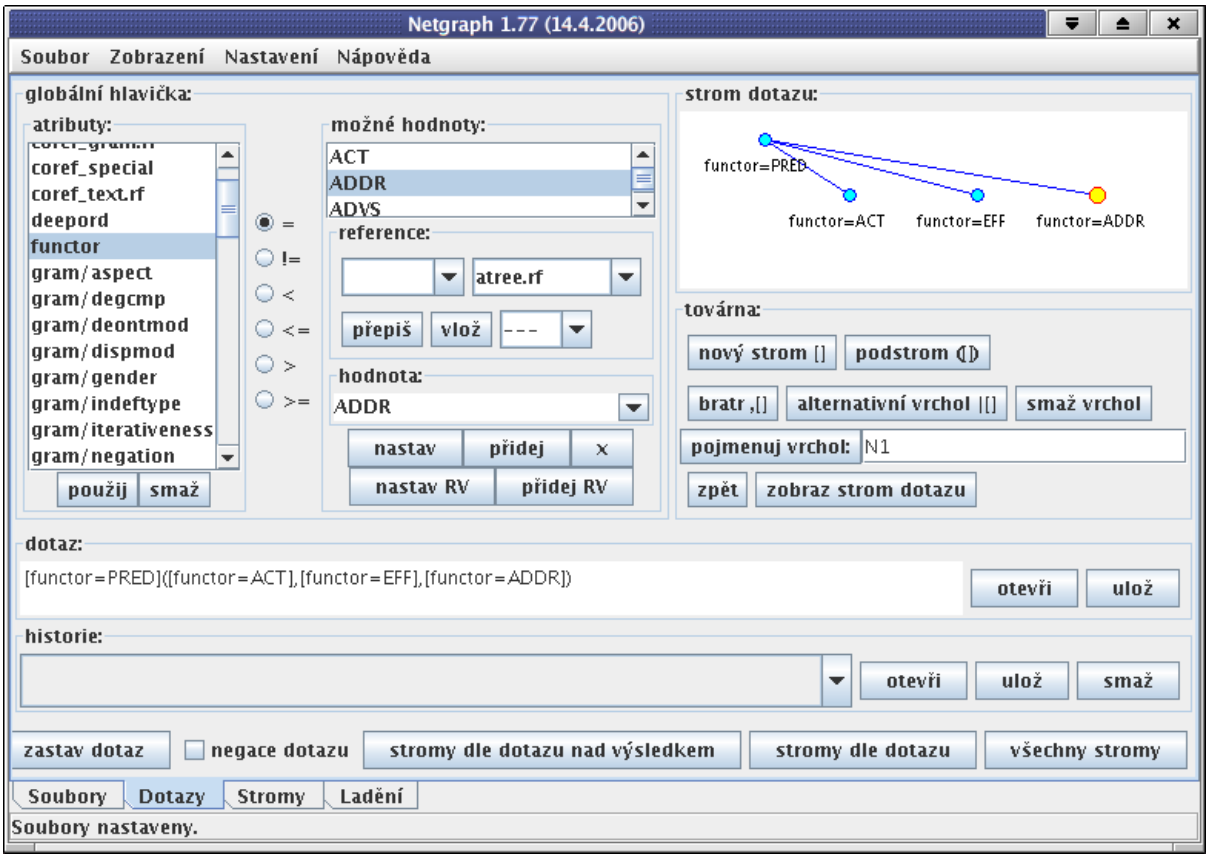
`Netgraph` klient je napsán v Javě a je nezávislý na platformě. Existuje ve dvou formách. První formou je samostatná javovská aplikace. V této podobě jsou dostupné všechny funkce klienta; musí však být nejprve nainstalován, spolu s Java 2 Runtime Environment. Druhou formou je javovský aplet. Ten, ač ochuzen o některé funkce, poskytuje plnou vyhledávací sílu a běží ve webovském prohlížeči bez předchozí instalace; vyžaduje ovšem, abyste měli ve svém prohlížeči nainstalovaný Java 2 plug-in.

Dotaz v `Netgraphu` je jeden uzel nebo strom s uživatelem definovanými vlastnostmi, který má být vyhledán v korpusu. Prohledání korpusu pak znamená hledat věty (samozřejmě ve formě anotovaných stromů), které obsahují dotaz jako svůj podstrom. Uživatel má možnost zadat dotazy nejrůznější složitosti, od těch nejjednodušších (jako je hledání všech stromů korpusu, které obsahují dané slovo), po velmi pokročilé (jako např. hledání všech vět, obsahujících sloveso rozvinuté adresátem, který není ve třetím pádě, *a* nejméně jedním příslovcem udávajícím směr, atd.). Dotazy mohou být dále rozšířeny tzv. *meta atributy*, které umožňují vyhledávat ještě složitější konstrukce. Meta atributy umožňují nastavení tranzitivních hran, volitelných uzlů, určení pozice dotazu v nalezených stromech, omezení velikosti nalezených stromů, nastavení pořadí uzlů, určení vztahů mezi hodnotami atributů u různých uzlů v nalezených stromech, negaci a mnoho dalších podmínek.

Dotazy se v `Netgraphu` vytvářejí v uživatelsky přívětivém grafickém prostředí. Příkladem je dotaz na obrázku 4.1. V tomto jednoduchém dotazu hledáme všechny stromy, které obsahují uzel označený jako predikát, rozvitý nejméně třemi uzly, označenými jako aktor, efekt a adresát. Pořadí těchto uzlů v nalezených stromech není v dotazu nijak omezeno.

Jedním z výsledků, zaslaných zpět serverem, může být strom z obrázku 4.2.

Uzly výsledného stromu, které odpovídají uzlům dotazu, jsou zvýrazněny žlutou a zelenou barvou. Všimněte si, že predikát ve výsledném stromě má více synů, než jsme určili v dotazu. To je v souladu



Obrázek 4.1: Vytváření dotazu v `Netgraphu`

s definicí vyhledávání v `Netgraphu` - stačí, že strom dotazu je v nalezeném stromě obsažen jako podstrom. Všimněte si dále, že pořadí uzlů v dotazu a ve výsledku jsou odlišná. Meta atributy umožňují omezit jak skutečný počet synů uzlu ve výsledném stromě, tak i výsledné pořadí uzlů, pokud si tak uživatel přeje.

Informace o způsobu instalace `Netgraphu` najdete v instrukcích k rychlé instalaci `Netgraph` klienta a v instrukcích k rychlé instalaci `Netgraph` serveru. Důležité informace najdete též v Manuálu k `Netgraph` klientu a v Manuálu k instalaci `Netgraph` serveru.

Poznamenejme, že instalovat `Netgraph` server potřebujete pouze v případě, že chcete prohledávat svůj vlastní korpus. Pro prohledávání PDT 2.0 poskytuje Ústav formální a aplikované lingvistiky[1] výkonný server na adrese `quest.ms.mff.cuni.cz` a portu `2200`. Je přístupný přes internet pro anonymního uživatele *anonymous* a připojit se k němu můžete pomocí `Netgraph` klienta (viz instrukce k rychlé instalaci `Netgraph` klienta).

Více informací o `Netgraphu` najdete v Manuálu k `Netgraph` klientu. Máte-li zájem o plný neanonymní přístup k serveru či máte-li zájem o další informace, aktualizace a novinky, navštivte domovskou stránku `Netgraphu`[2].

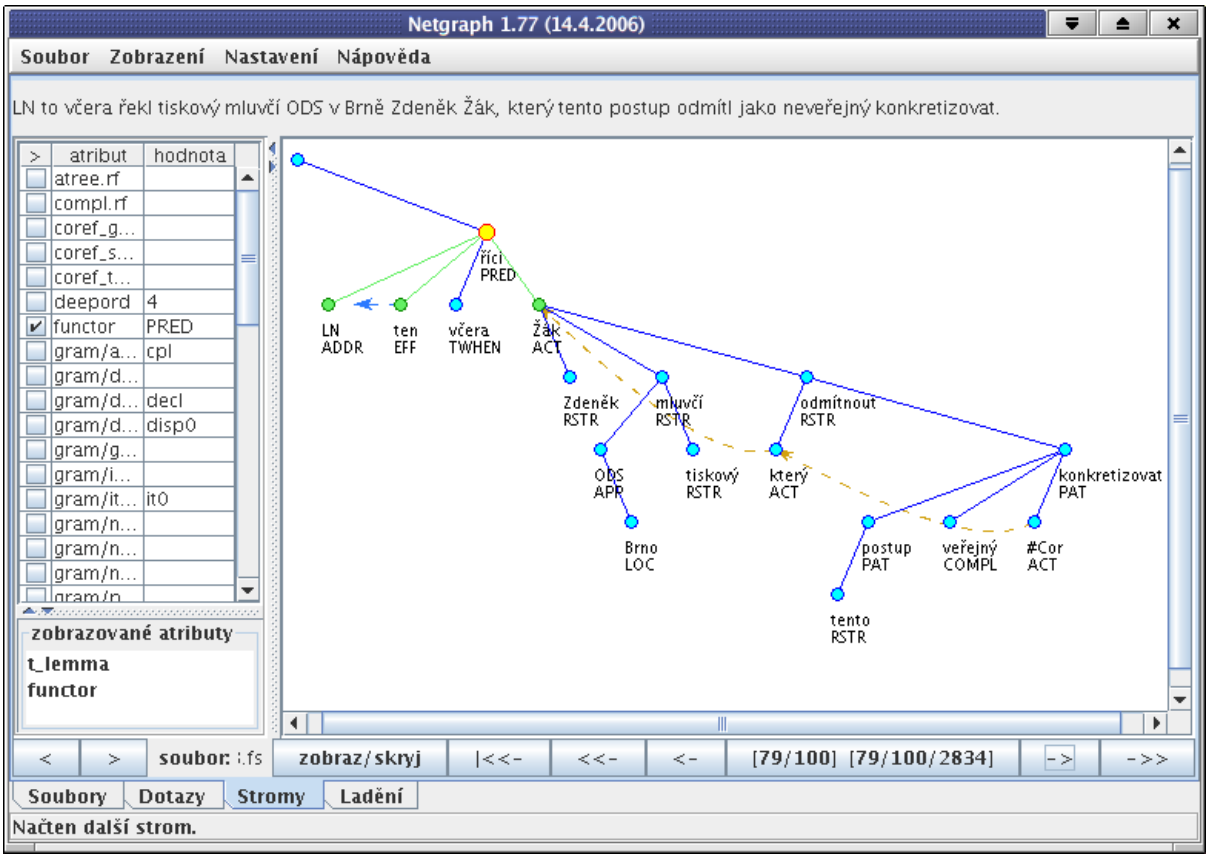
4.2 Prohlížení stromů: `TrEd`

Nejpřehlednější a nejpohodlnější zobrazení dat PDT 2.0 poskytuje `TrEd`. Prvotně sloužil jako hlavní anotační nástroj, ale může být použit i k prohlížení dat a obsahuje také několik druhů vyhledávacích funkcí. Instrukce k instalaci `TrEdu` najdete v dokumentaci k `TrEdu`.

Pro otevření souborů v `TrEdu` zvolte menu **File** a klikněte na položku **Open**. Vyberte jakýkoliv soubor `*.t.gz` (tj. soubor s tektogramatickou anotací nějakého dokumentu), `TrEd` jej otevře a ihned zobrazí strom pro první větu daného souboru.

Typický vzhled `TrEdu` vidíte na obrázku 4.3; jde o větu *Kde jsou auta, tam je kšeft.*

[1] <http://ufal.mff.cuni.cz>
[2] <http://quest.ms.mff.cuni.cz/netgraph>



Obrázek 4.2: Nalezený strom v `Netgraphu`

1. Na tomto místě vidíte jedno či více oken. Každé okno zobrazuje jeden strom.

2. V tomto poli vidíte prostou textovou formu věty zobrazené v právě vybraném okně.

3. Stavová řádka. Zobrazuje různé informace v závislosti na aktuálním kontextu.

4. Aktuální kontext. Kontext můžete změnit kliknutím na jméno aktuálního kontextu a následným výběrem nového kontextu ze zobrazeného seznamu (např. `PML_T_Edit`).

5. Aktuální zobrazovací styl. Může být změněn podobným způsobem jako kontext.

6. Sem klikněte pro editaci zobrazovacího stylu.

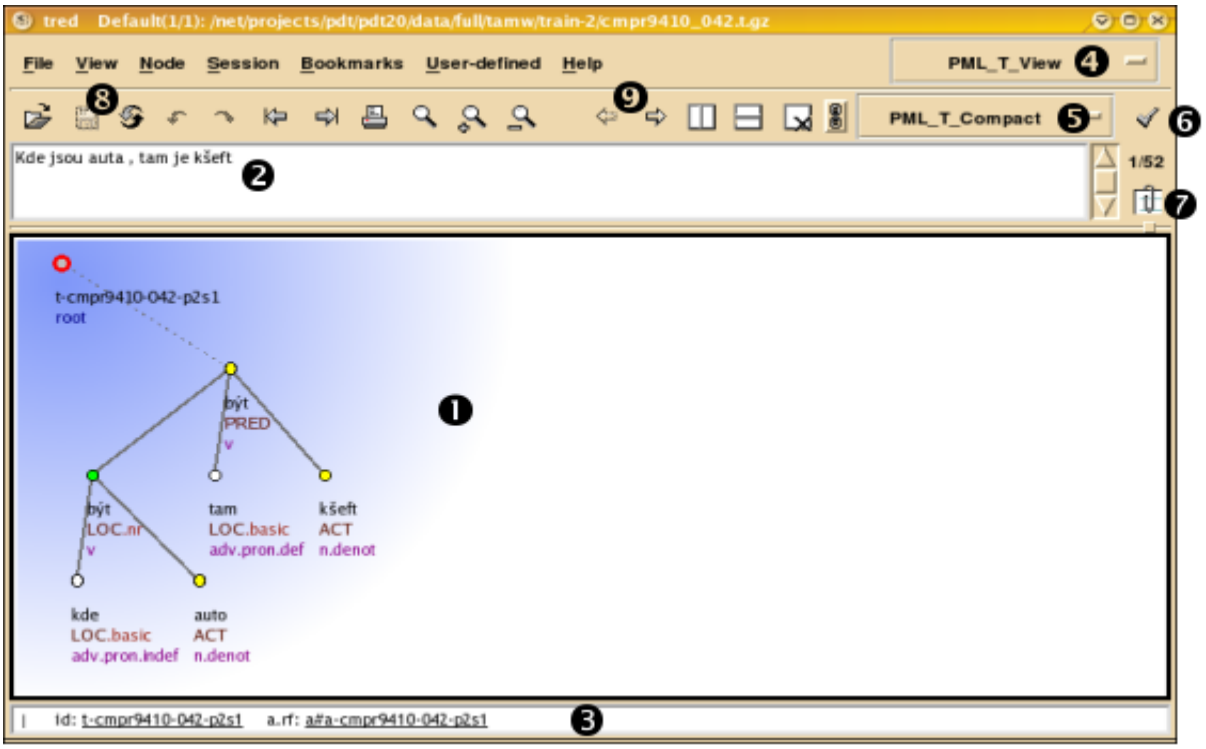
7. Kliknutím sem zobrazíte seznam všech vět aktuálního souboru. Nad tlačítkem je zobrazeno pořadí aktuálního stromu v aktuálním souboru.

8. Tlačítka pro otevření, uložení a opětovné otevření souboru. Ikony znamenají *Undo*, *Redo*, *Previous* a *Next File*, *Print*, *Find*, *Find Next*, *Find Previous*.

9. Tlačítka pro přesunutí na předchozí/následující strom v aktuálním souboru a pro správu oken.

Implicitně jsou tektogramatické soubory PDT 2.0 v PML formátu otevřeny v kontextu `PML_T_View`, který neumožňuje jejich editaci. Pokud si přejete soubory měnit, přepněte se do kontextu `PML_T_Edit`. V obou kontextech jsou k dispozici dva zobrazovací styly. Implicitní je `PML_T_Compact`, pro zobrazení více podrobností můžete použít `PML_T_Full`. Informace o kontextech a zobrazovacích stylech najdete v dokumentaci k makrům v `TrEdu`.

V libovolném kontextu můžete zobrazit seznam všech maker definovaných v daném kontextu a jejich klávesové zkratky, a to vybráním menu **View** → **List of Named Macros**.



Obrázek 4.3: Tektogramatický strom v `TrEdu`

4.3 Automatické zpracování stromů: `btred/ntred`

`Netgraph` (popsaný v sekci 4.1) umožňuje i neprogramátorům snadno a pohodlně vyhledávat stromy v PDT. Editor `TrEd` (popsaný v sekci 4.2) umožňuje rychlé, pohodlné a flexibilní procházení, prohlížení a úpravu jednotlivých stromů. Vývojáři nástrojů a programátoři obecně však potřebují plný přístup k datům. Můžete samozřejmě data zpracovávat přímo (koneckonců, jsou v XML), my ale doporučujeme k datům přistupovat pomocí perlovského rozhraní `btred/ntred`, ušitého datům PDT 2.0 na míru. *`btred`* je perlovský program, který umožňuje aplikaci jiného perlovského programu (zvaného *makro btredu*) na data uložená v jednom z formátů PDT. *`ntred`* je `btred` ve verzi klient-server a je vhodný pro paralelní zpracování dat na více strojích. (Mnemotechnika pro `btred/ntred`: „b“ znamená „batch processing“, dávkové zpracování, „n“ znamená „networked processing“, zpracování po síti.)

Budete-li postupovat podle uvedeného doporučení, získáte několik výhod:

- Objektově orientovaná reprezentace stromů, použitá v prostředí `btred/ntredu`, nabízí velké množství základních funkcí pro procházení stromů a pro mnoho dalších základních operací na stromech; k dispozici je i několik značně pokročilých funkcí, vhodných pro lingvisticky motivované procházení stromů (funkce, které berou v úvahu například vzájemné propojení mezi relacemi závislosti a koordinace).
- Technologie `btred/ntredu` byla široce používána několika programátory během vývoje PDT 2.0; tato dlouhodobá zkušenost vedla k mnoha vylepšením, díky nimž jsou tyto nástroje a přidružené knihovny rozumně stabilní.
- Máte-li k dispozici více počítačů, můžete použít `ntred` a zpracovávat data paralelně, což výpočet výrazně zrychluje. V závislosti na konkrétní situaci může být průchod celým PDT 2.0 zkrácen na pouhých několik sekund (s pouze přibližně 10 procesory přístupnými pro distribuovaný běh `btredu`).
- Programátoři mohou `btred/ntred` (v kombinaci s `TrEdem`) použít jako mocný a rychlý vyhledávací stroj. Napíšete makro, které v korpusu vyhledá pozice, o které se zajímáte, spustíte ho v `ntredu` a získané pozice si jednoduše prohlédnete v `TrEdu`.
- K tomu, abyste si osvojili psaní maker pro `btred/ntred`, potřebujete jen znát základy syntaxe jazyka Perl a zapamatovat si jména několika proměnných a funkcí, předdefinovaných v prostředí `btred/ntredu`.

- Jakmile si na práci s `btred/ntredem` zvyknete, budete moci všech jeho výhod využít i při zpracování dat dalších korpusů (ať už závislostních, nebo i bezprostředně složkových).

Pro úvodní seznámení si přečtěte tutoriál k `btred/ntredu`. Podívejte se také na manuálové stránky `btredu` a `ntredu`.

4.4 Konverze mezi různými formáty dat

4.4.1 Konverze mezi formáty PDT

Konverze mezi datovými formáty je velice obtížný úkol, pokud všechny formáty nemohou nést přesně stejné množství informací. Naneštěstí to je právě případ formátů, které vznikly během roků vývoje PDT. Z toho důvodu poskytujeme několik nástrojů, které usnadňují alespoň některé z konverzí. Mohou posloužit i jako příklady složitějších transformací, které mohou být potřebné pro některé úkoly. Úplný popis najdete v textu PDT 2.0: nástroje pro konverzi interních formátů.

V distribuci jsou skripty uloženy v adresáři `tools/format-conversions/pdt_formats`. Většina těchto skriptů potřebuje ke své činnosti `btred`, nástroj z balíku `TrEd`.

Podporovány jsou následující typy konverzí:

- konverze analytické anotace typu PDT 1.0 do PML,
- konverze a-dat PML do CSTS,
- konverze m-dat PML do CSTS,
- konverze dat PDT 2.0 do FS pro Netgraph,
- konverze dat PDT 2.0 do vnitřního binárního formátu Perlu (pro urychlení).

4.4.2 Konverze z formátů jiných korpusů

K dispozici jsou také skripty pro konverzi formátů Penn Treebanku a korpusu Negra do formátu FS. Konverzní skripty jsou umístěny v adresáři `tools/format-conversions/from_negra+ptb`. Jejich popis najdete v stručné dokumentaci.

Poznamenejme, že skripty neprovádějí *žádnou konverzi anotačních schémat*. Jinými slovy, složkové stromy zůstanou složkovými stromy, závislostní struktura není automaticky vytvářena.

4.5 Parsing češtiny: od prostého textu k závislostním stromům typu PDT

Společně s daty poskytujeme také nástroje, které provádějí automatickou anotaci. Ze surových českých vět vytvářejí závislostní stromy na analytické rovině. Nástroje jsou uloženy v adresáři `tools/machine-annotation`. Provádějí postupně tyto činnosti:

- rozpoznání slovních jednotek ve vstupním surovém textu a rozdělení textu na věty,
- morfologickou analýzu a tagging (morfologickou disambiguaci),
- závislostní parsing,
- přiřazení analytických (závislostních) funkcí všem uzlům zparsovaného stromu.

Nástroje pro následný parsing na tektogramatickou rovinu zatím neexistují. Prosíme, sledujte webovské stránky <`http://ufal.ms.mff.cuni.cz/pdt2.0update/`> obsahující aktualizace PDT 2.0 a nové nástroje.

Více informací najdete v podrobném popisu nástrojů.

4.6 Vytvoření dat pro vývoj parseru

Během vývoje nového parseru je důležité testovat jeho úspěšnost nejen na ručně anotovaných souborech m-roviny, ale také na souborech anotovaných automaticky. Pro nedostatek místa nebylo možné automaticky anotovaná data m-roviny umístit na CD-ROM. K dispozici je však nástroj pro generování dat vhodných pro vývoj parseru a jeho testování.

Tento nástroj je umístěn v adresáři `tools/machine-annotation/for_parser_devel/`. Spouští se příkazem

`run_for_parser_devel` *`vstupní_adresář výstupní_adresář`*

Vstupní adresář musí mít stejnou strukturu jako adresář `data/full/`, který bude typicky prvním argumentem nástroje. Nástroj kopíruje celou adresářovou strukturu vstupního adresáře do výstupního adresáře. Kopíruje rovněž všechny datové soubory kromě souborů m-roviny, které jsou nahrazeny soubory nově vytvořenými. Nové soubory m-roviny obsahují automaticky přiřazená lemmata a tagy. Upozorňujeme, že tyto nové soubory nejsou totožné s těmi, které by byly vytvořeny automatickou anotací použitou přímo na prostý text. Zachovávají totiž hranice vět a slovních jednotek a také identifikátory jednotek m-roviny obsažené v ručně anotovaných datech.

4.7 Makra pro detekce chyb

Přestože anotátoři viděli každý uzel každého stromu (a to často více než jednou), zůstaly v datech nějaké chyby. Některé byly způsobeny přehlédnutím, jiné tím, že se pravidla anotace během anotačního procesu vyvíjela a měnila, ale data nebyla přeanotována při každé změně. Z toho důvodu bylo během anotační a kontrolní fáze vytvořeno mnoho programů (maker pro `TrEd/btred/ntred`, viz sekce 4.2), které v datech hledaly porušení nějakého pravidla či invariantu nebo podezřelou anotaci a na každé takové místo upozorňovaly. Data pak byla ručně či automaticky opravena, v případě potřeby bylo makro dále upraveno.

POZNÁMKA

Jako pomůcka při psaní maker pro `TrEd` slouží dokumentace `TrEdu`.

Makra byla rozdělena do tří skupin: *find*, *fix* a *check*. Makra ze skupiny *find* pouze vyhledávala podezřelá místa v datech. Makra ze skupiny *fix* byla používána pro automatickou opravu dat, pokud byla možná (jako např. když uprostřed anotačního procesu došlo k jasné a jednoznačné změně anotačního pravidla). Poslední skupina (*check*) obsahovala makra podobná těm ve skupině *find*, ale zahrnovala seznam výjimek z obecného pravidla. (A existovala vlastně ještě další skupina, nazvaná *misc*, obsahující směs nejrůznějších dalších maker a skriptů.)

Makra byla dále rozdělena do skupin podle toho, pro kterou rovinu anotace byla určena (viz kapitola 2 pro další informace o rovinách).

Makra ze skupiny *check* jsou uložena v adresáři `tools/checks`.

VAROVÁNÍ

Tato makra již nejsou určena k použití na datech, protože formát dat se změnil, ale mohou posloužit k vytvoření jasnější představy, jaké druhy kontrol byly na data v PDT 2.0 aplikovány a jaká makra pro práci se stromy je možno psát.

Kapitola 5

Dokumentace

Toto je přehledný a strukturovaný seznam všech odkazů na dokumentace k nástrojům, formátům dat apod., vyskytujících se v celém tomto průvodci PDT.

- Průvodce PDT (to, co právě čtete)
  - verze HTML: `doc/pdt-guide/cz/html/index.html`
  - verze PDF: `doc/pdt-guide/cz/pdf/pdt-guide.pdf`
- Anotační manuály (viz též 2)
  - Manuál k morfologické anotaci
    - v angličtině
      - verze HTML: `doc/manuals/en/m-layer/html/index.html`
      - verze PDF: `doc/manuals/en/m-layer/pdf/m-man-en.pdf`
  - Manuál k analytické anotaci
    - v angličtině
      - verze HTML: `doc/manuals/en/a-layer/html/index.html`
      - verze PDF: `doc/manuals/en/a-layer/pdf/a-man-en.pdf`
    - v češtině
      - verze HTML: `doc/manuals/cz/a-layer/html/index.html`
      - verze PDF: `doc/manuals/cz/a-layer/pdf/a-man-cz.pdf`
  - Manuál k tektogramatické anotaci
    - v angličtině
      - verze HTML: `doc/manuals/en/t-layer/html/index.html`
      - verze PDF: `doc/manuals/en/t-layer/pdf/t-man-en.pdf`
    - v češtině
      - verze HTML: `doc/manuals/cz/t-layer/html/index.html`
      - verze PDF: `doc/manuals/cz/t-layer/pdf/t-man-cz.pdf`
- Data (viz též sekce 3.4)
  - CSTS
    - úplný popis: `doc/data-formats/csts/html/DTD-HOME.html`
    - DTD: `doc/data-formats/csts/csts.dtd`
  - FS - specifikace formátu: `doc/data-formats/fs/index.html`
  - PML
    - úplný popis:
      - verze HTML: `doc/data-formats/pml/index.html`

        - verze PDF: `doc/data-formats/pml/pml_doc.pdf`
      - schémata: `data/schemas`
  - systém značek PML (včetně popisu atributů uzlů): `doc/data-formats/pml-markup/index.html`
  - PDT-VALLEX - fyzická struktura: `doc/data-formats/pdt-vallex/pdt-vallex-struct.html`

- Nástroje (viz též kapitola 4)
  - `TrEd`, `btred/ntred`
    - `TrEd` - manuál: `doc/tools/tred/index.html`
    - `btred` - manuálová stránka: `doc/tools/tred/btred.html`
    - `ntred` - manuálová stránka: `doc/tools/tred/ntred.html`
    - `btred/ntred` - tutoriál: `doc/tools/tred/bn-tutorial.html`
    - Makra `TrEdu`: `doc/tools/tred/PML_mak.html`
  - `Netgraph:`
    - Rychlá instalace `Netgraph` klienta: `doc/tools/netgraph/README_QUICK_INSTALL_CLIENT`
    - Manuál k `Netgraph` klientu: `doc/tools/netgraph/netgraph_manual.html`
    - Rychlá instalace `Netgraph` serveru: `doc/tools/netgraph/README_QUICK_INSTALL_SERVER`
    - Manuál k instalaci `Netgraph` serveru: `doc/tools/netgraph/netgraph_server_install.html`
  - Konverzní skripty:
    - Z formátů Penn Treebanku a Negry: `doc/tools/format-conversions/from_negra+ptb/readme.txt`
    - Mezi formáty PDT: `doc/tools/format-conversions/pdt_formats/index.html`
  - Automatická anotace (rozpoznání slov, morfologie, parsing): `doc/tools/machine-annotation/index.html`

- Publikace (viz též kapitola 6)
  - záznamy BibTeXu: `publications/pdt.bib`

Kapitola 6

Publikace

Zde najdete seznam publikací, zabývajících se následujícími tématy:

- výzkum provedený především před začátkem projektu PDT a který byl klíčový pro vytvoření anotační strategie (sekce 6.1),
- úplný seznam publikací o vytváření PDT 2.0 (sekce 6.2),
- nástroje pro editaci, vyhledávací systémy, nástroje pro zpracování přirozeného jazyka (sekce 6.3).

Obecné publikace v sekci 6.2 jsou uspořádány podle data zveřejnění. Díky tomu lze získat přehled o postupu prací na PDT. Publikace v ostatních sekcích jsou seřazeny obvyklým způsobem, tj. v abecedním pořadí podle příjmení prvního autora.

Většina publikací je k dispozici v elektronické podobě (v souborech PDF i Postscript), jak je u každé publikace naznačeno. Elektronické verze jsou kopie autorů, poskytnuté na osobní žádost, a jako takové jsou určené *pouze pro osobní užití*. Ke všem zde uvedeným publikacím jsou k dispozici i záznamy BibTeX.

6.1 Teoretické pozadí PDT

- Eva Hajičová: *Issues of Sentence Structure and Discourse Patterns.* Charles University, Prague, Czech Republic, 1993. **K dispozici:** BibTeX
- Eva Hajičová, Jarmila Panevová: *Valency (case) frames.* V: P. Sgall (ed.): Contributions to Functional Syntax, Semantics and Language Comprehension, Prague:Academia, 1984, pp. 147–188. **K dispozici:** BibTeX
- Eva Hajičová, Barbara H. Partee, Petr Sgall: *Topic-focus articulation, tripartite structures, and semantic content.* Amsterdam:Kluwer, 1998. **K dispozici:** BibTeX
- Jarmila Panevová: *On verbal frames in Functional generative description I.* V: Prague Bulletin of Mathematical Linguistics, 22, MFF UK, Prague, Czech Republic, 1974, pp. 3–40. **K dispozici:** PDF, PS, BibTeX
- Jarmila Panevová: *On verbal frames in functional generative description II.* V: Prague Bulletin of Mathematical Linguistics, 23, MFF UK, Prague, Czech Republic, 1975, pp. 17–52. **K dispozici:** PDF, PS, BibTeX
- Jarmila Panevová: *Formy a funkce ve stavbě české věty.* Prague:Academia, 1980. **K dispozici:** BibTeX
- Vladimír Petkevič: *A new dependency based specification of underlying representations of sentences.* V: Theoretical Linguistics, 14, 1987, pp. 143–172. **K dispozici:** BibTeX
- Vladimir Petkevič: *A New Formal Specification of Underlying Representations.* V: Theoretical Linguistics, 21, 1995, pp. 7–61. **K dispozici:** BibTeX
- Petr Sgall: *Generativní popis jazyka a česká deklinace.* Prague:Academia, 1967. **K dispozici:** BibTeX
- Petr Sgall: *Contributions to Functional Syntax, Semantics and Language Comprehension.* Prague:Academia, 1984. **K dispozici:** BibTeX

- Petr Sgall: *Underlying Structure of Sentence and its Relation to Semantics.* V: T. Reuther (ed.): Wiener Slawisticher Almanach. Sonderband 33, 1992, pp. 273–282. **K dispozici:** BibTeX
- Petr Sgall: *Valency and Underlying Structure. An alternative view on dependency.* V: L. Wanner (ed.): Recent Trends in meaning-text theory, Amsterdam/Philadelphia: Benjamins, 1997, pp. 149–166. **K dispozici:** BibTeX
- Petr Sgall, Eva Hajičová, Jarmila Panevová: *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht:Reidel Publishing Company and Prague:Academia, 1986. **K dispozici:** BibTeX
- Vladimír Šmilauer: *Novočeská skladba.* Státní pedagogické nakladatelství, Prague, Czech Republic, 1969. **K dispozici:** BibTeX

6.2 PDT 2.0

6.2.1 Obecné informace

Motivace k vytvoření PDT

- Jan Hajič, Eva Hajičová, Alexander Rosen: *Formal Representation of Language Structures.* V: TELRI Newsletter, 3, 1996, pp. 12–19. **K dispozici:** PDF, PS, BibTeX

2000

- Jan Hajič, Alena Böhmová, Eva Hajičová, Barbora Vidová Hladká: *The Prague Dependency Treebank: A Three-Level Annotation Scenario.* V: A. Abeillé (ed.): Treebanks: Building and Using Parsed Corpora, Amsterdam:Kluwer, 2000, pp. 103–127. **K dispozici:** PDF, PS, BibTeX
- Jarmila Panevová: *Building an electronic language database nowadays: The Prague Dependency Treebank. 2000. **K dispozici:** PDF, PS, BibTeX*

2001

- Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, Barbora Vidová Hladká: *Prague Dependency Treebank 1.0 (Final Production Label).* V: CD-ROM, CAT: LDC2001T10, ISBN 1-58563-212-0, Linguistic Data Consortium, 2001. **K dispozici:** BibTeX
- Jan Hajič, Petr Pajas, Barbora Vidová Hladká: *The Prague Dependency Treebank: Annotation Structure and Support.* V: Proceedings of the IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA, 2001, pp. 105–114. **K dispozici:** PDF, PS, BibTeX
- Eva Hajičová, Jan Hajič, Martin Holub, Petr Pajas, Veronika Kolářová-Řezníčková, Petr Sgall, Barbora Vidová Hladká: *The Current Status of the Prague Dependency Treebank* V: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds.): Proceedings of the 5th International Conference on Text, Speech and Dialogue, Železná Ruda - Špičák, Czech Republic, Springer-Verlag Berlin Heidelberg New York, 2001, pp. 11–20. **K dispozici:** PDF, PS, BibTeX
- Eva Hajičová, Petr Sgall: *A reusable corpus needs syntactic annotations: Prague Dependency Treebank.* V: A rainbow of corpora–corpus linguistics and the languages of the world, Munich: Licom-Europa, 2001, pp. 37–48. **K dispozici:** PDF, PS, BibTeX

2002

- Eva Hajičová: *Theoretical description of language as a basis of corpus annotation: The case of Prague Dependency Treebank.* V: E. Hajičová, P. Sgall, J. Hana, T. Hoskovec (eds.): Prague Linguistic Circle Papers, (4), Amsterdam/Philadelphia:John Benjamins, 2002, pp. 111–127. **K dispozici:** BibTeX
- Jarmila Panevová, Eva Hajičová, Petr Sgall: *K nové úrovni bohemistické práce: Využití anotovaného korpusu. Část 1.* V: Slovo a slovesnost, 63, Czech Academy of Science, Prague, 2002, pp. 161–177. **K dispozici:** PDF, PS, BibTeX
- Jarmila Panevová, Eva Hajičová, Petr Sgall: *K nové úrovni bohemistické práce: Využití anotovaného korpusu. Část 2.* V: Slovo a slovesnost, 63, Czech Academy of Science, Prague, 2002, pp. 241–262. **K dispozici:** PDF, PS, BibTeX

- Barbora Vidová Hladká: *Pražský závislostní korpus aneb Co tady před padesáti lety nebylo.* V: Pokroky matematiky, fyziky a astronomie, 47, (4), Jednota českých matematiků a fyziků, 2002, pp. 298–306. **K dispozici:** PDF, PS, BibTeX

2003

- Alena Böhmová, Eva Hajičová: *Large Language Data and the Degrees of Automation.* V: E. Hajičová, A. Kotěšovcová, J. Mírovský (eds.): Proceedings of XVII International Congress of Linguists, CD-ROM, Matfyzpress, MFF UK, Prague, Czech Republic, 2003. **K dispozici:** PDF, PS, BibTeX

2004

- Jan Hajič: *Complex Corpus Annotation: The Prague Dependency Treebank.* Jazykovedný ústav Ľ. Štúra, SAV, Bratislava, Slovakia, 2004. **K dispozici:** PDF, PS, BibTeX

2005

- Petr Pajas, Jan Štěpánek: *A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague Dependency Treebank 2.0.* V: UFAL Technical Report, 29, MFF UK, Prague, Czech Republic, 2005. **K dispozici:** PDF, PS, BibTeX

6.2.2 Morfologická rovina

- Jan Hajič: *Disambiguation of Rich Inflection (Computational Morphology of Czech).* Karolinum, Charles University Press, Prague, Czech Republic, 2004. **K dispozici:** BibTeX
- Dan Zeman, Jiří Hana, Hana Hanová, Jan Hajič, Emil Jeřábek, Barbora Vidová Hladká: *A Manual for Morphological Annotation, 2nd edition (html).* V: ÚFAL Technical Report, 27, MFF UK, Prague, Czech Republic, 2005. **K dispozici:** PDF, PS, BibTeX

6.2.3 Analytická rovina

- Jan Hajič: *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank.* V: E. Hajičová (ed.): Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová, Karolinum, Charles University Press, Prague, Czech Republic, 1998, pp. 106–132. **K dispozici:** PDF, PS, BibTeX
- Jan Hajič, Eva Hajičová: *Syntactic tagging in the Prague Dependency Treebank.* V: R. Marcinkeviciene, N. Volz (eds.): Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe", TELRI, Kaunas, Lithuania, 1997, pp. 55–68. **K dispozici:** PDF, PS, BibTeX
- Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall: *Syntax v Českém národním korpusu.* V: Slovo a slovesnost, Czech Academy of Science, Prague, 1998, pp. 168–177. **K dispozici:** BibTeX
- Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, Alla Bémová: *Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory (html). 1999.* ***K dispozici:*** *PDF, PS, BibTeX*
- Eva Hajičová, Zdeněk Kirschner, Petr Sgall: "A Manual for Analytical Layer Annotation of the Prague Dependency Treebank (English translation) (html). 1999. **K dispozici:** PDF, PS, BibTeX "
- Roman Ondruška, Jarmila Panevová, Jan Štěpánek: *An Exploitation of the Prague Dependency Treebank: A Valency Case.* V: K. Simov, P. Osenova (eds.): Proceedings of the Workshop on Shallow Processing of Large Corpora, UCREL, Lancaster University, Lancaster, Great Britain, 2003, pp. 69–77. **K dispozici:** PDF, PS, BibTeX

6.2.4 Tektogramatická rovina

Struktura anotace

- Alena Böhmová: *Automatic Procedures in Tectogrammatical Tagging.* V: Prague Bulletin of Mathematical Linguistics, 76, MFF UK, Prague, Czech Republic, 2001, pp. 23–34. **K dispozici:** PDF, PS, BibTeX

- Alena Böhmová, Silvie Cinková, Eva Hajičová: *A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank (English translation) (html). 2005.* ***K dispozici:*** *PDF, PS, BibTeX*
- Alena Böhmová, Petr Sgall: "Automatic procedures in tectogrammatical tagging." V: Proceedings of the Workshop on Linguistically Interpreted Corpora, 18th International Conference on Computational Linguistics, Saarbrücken, Germany, 2000, pp. 65–70. **K dispozici:** PDF, PS, BibTeX
- Eva Hajičová: *Prague Dependency Treebank: From analytic to tectogrammatical annotations.* V: P. Sojka, V. Matoušek, K. Pala, I. Kopeček (eds.): Proceedings of the 2nd International Conference on Text, Speech and Dialogue, Brno, Czech Republic, Springer-Verlag Berlin Heidelberg New York, 1998, pp. 45–50. **K dispozici:** PDF, PS, BibTeX
- Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, Daniel Zeman: *Issues of Projectivity in the Prague Dependency Treebank.* V: Prague Bulletin of Mathematical Linguistics, 81, MFF UK, Prague, Czech Republic, Prague, 2004, pp. 5–22. **K dispozici:** PDF, PS, BibTeX
- Eva Hajičová, Petr Pajas: *Evaluation of Tectogrammatical Annotation of PDT.* V: P. Sojka, I. Kopeček, K. Pala (eds.): Proceedings of the 3rd International Conference on Text, Speech and Dialogue, Brno, Czech Republic, Springer-Verlag Berlin Heidelberg New York, 2000, pp. 75–80. **K dispozici:** BibTeX
- Eva Hajičová, Petr Pajas, Kateřina Veselá: *Corpus Annotation on the Tectogrammatical Layer: Summarizing the First Stages of Evaluations.* V: Prague Bulletin of Mathematical Linguistics, 77, MFF UK, Prague, Czech Republic, Prague, 2002, pp. 5–18. **K dispozici:** PDF, PS, BibTeX
- Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová-Řezníčková, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský: *Anotace Pražského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory (html). 2005.* ***K dispozici:*** *PDF, PS, BibTeX*
- Jarmila Panevová, Alena Böhmová, Petr Sgall: "Syntactic Tagging: Procedure for the Transition from the Analytic to the Tectogrammatical Tree Structure." V: V. Matoušek, P. Mautner, J. Ocelíková, P. Sojka (eds.): Proceedings of the 2nd International Conference on Text, Speech and Dialogue, Plzeň, Czech Republic, Springer-Verlag Berlin Heidelberg New York, 1999, pp. 34–38. **K dispozici:** PDF, PS, BibTeX
- Jarmila Panevová, Eva Hajičová, Petr Sgall: *Tectogrammatics in corpus tagging.* V: I. Kenesei, R. M. Harnish (eds.): Perspectives on Semantics, Pragmatics, and Discourse; A Festschrift for Ferenc Kiefer (Pragmatics and Beyond new Series), (90), Amsterdam/Philadelphia:John Benjamins, 2001, pp. 294–299. **K dispozici:** PDF, PS, BibTeX
- Jarmila Panevová, Veronika Kolářová-Řezníčková, Zdeňka Urešová: *The Theory of Control Applied to the Prague Dependency Treebank (PDT).* V: R. Frank (ed.): Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), Universita di Venezia, Venezia, Italy, 2002, pp. 175–180. **K dispozici:** PDF, PS, BibTeX
- Veronika Kolářová-Řezníčková: *PDT: Two Steps in Tectogrammatical Annotation with respect to some Issues of Deletion.* V: Prague Bulletin of Mathematical Linguistics, 78, MFF UK, Prague, Czech Republic, Prague, 2002, pp. 37–52. **K dispozici:** PDF, PS, BibTeX
- Petr Sgall, Jarmila Panevová, Eva Hajičová: *Deep Syntactic Annotation: Tectogrammatical Representation and Beyond.* V: A. Meyers (ed.): Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 32–38. **K dispozici:** PDF, PS, BibTeX

Aktuální členění

- Eva Hajičová: *The Prague Dependency Treebank: Crossing the Sentence Boundary.* V: V. Matoušek, P. Mautner, J. Ocelíková, P. Sojka (eds.): Proceedings of the 2nd International Conference on Text, Speech and Dialogue, Plzeň, Czech Republic, Springer-Verlag Berlin Heidelberg New York, 1999, pp. 20–27. **K dispozici:** PDF, PS, BibTeX
- Eva Hajičová: *Dependency-Based Underlying-Structure Tagging of a Very Large Czech Corpus.* V: S. Kahane (ed.): Special issue of TAL journal, Grammaires de Dépendence / Dependency Grammars, Paris:Hermes, 2000, pp. 57–78. **K dispozici:** PDF, PS, BibTeX

- Eva Hajičová, Petr Sgall: *Degrees of Contrast and the Topic-Focus Articulation. (1),* Berlin:Walter de Gruyter, 2004, pp. 1–13. **K dispozici:** PDF, PS, BibTeX

- Eva Hajičová, Petr Sgall, Eva Buráňová: *Topic-Focus Articulation and degrees of salience in the Prague Dependency Treebank.* V: A. Carnie, H. Harley, M. Willie (eds.): Formal Approaches to Function in Grammar. In honor of Eloise Jelinek, Arizona, Amsterdam/Philadelphia:John Benjamins, Amsterdam/Philadelphia, 2003, pp. 165–177. **K dispozici:** PDF, PS, BibTeX

- Eva Hajičová, Petr Sgall, Eva Buráňová: *Tagging of very large corpora: Topic-Focus Articulation.* V: Proceedings of the 18th International Conference on Computational Linguistics, Saarbrücken, Germany, 2000, pp. 139–144. **K dispozici:** PDF, PS, BibTeX

- Lucie Kučová, Eva Hajičová, Kateřina Veselá, Jiří Havelka: *Topic-focus articulation and anaphoric relations: A corpus based probe.* V: (ed.): Prague Bulletin of Mathematical Linguistics, 84, MFF UK, Prague, Czech Republic, 2005, pp. 5–12. **K dispozici:** PDF, PS, BibTeX

- Petr Sgall: *Topic-Focus Articulation in Corpus Annotation.* V: W. Menzel, C. Vertan (eds.): Natural language processing between linguistic inquiry and system engineering, Editura Universitatii Alexandru Ioan Cuza, Iasi, 2003, pp. 95–101. **K dispozici:** PDF, PS, BibTeX

- Kateřina Veselá, Jiří Havelka: *Anotování aktuálního členění věty v Pražském závislostním korpusu.* V: ÚFAL Technical Report, 20, MFF UK, Prague, Czech Republic, 2003. **K dispozici:** PDF, PS, BibTeX

- Kateřina Veselá, Jiří Havelka, Eva Hajičová: *Annotators' Agreement: The Case of Topic-Focus Articulation.* V: Proceedings of the 4th International Conference on Language Resources and Evaluation, European Language Resources Association, Lisboa, Portugal, 2004, pp. 2191–2194. **K dispozici:** PDF, PS, BibTeX

Koreference

- Lucie Kučová, Eva Hajičová: *Coreferential Relations in the Prague Dependency Treebank.* V: (ed.): Proceedings of the 5th International Conference on Discourse Anaphora and Anaphor Resolution 2004,

  San Miguel, Azores, Sept. 23-24, 2004, 2005, pp. 94–102. **K dispozici:** PDF, PS, BibTeX

- Lucie Kučová, Veronika Kolářová-Řezníčková, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo: *Anotování koreference v Pražském závislostním korpusu.* V: ÚFAL Technical Report, 19, MFF UK, Prague, Czech Republic, 2003. **K dispozici:** PDF, PS, BibTeX

- Jarmila Panevová, Eva Hajičová, Petr Sgall: *Coreference in Annotating a Large Corpus.* V: M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer (eds.): Proceedings of the 2nd International Conference on Language Resources, (I), European Language Resources Association, Athens, Greece, 2000, pp. 497–500. **K dispozici:** PDF, PS, BibTeX

PDT-VALLEX

- Silvie Cinková, Veronika Kolářová-Řezníčková: *Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank.* V: Korpusy a korpusová lingvistika v zahraničí a na Slovensku, 2004. **K dispozici:** PDF, PS, BibTeX

- Jan Hajič, Václav Honetschläger: *Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation.* V: Prague Bulletin of Mathematical Linguistics, 79–80, MFF UK, Prague, Czech Republic, 2003, pp. 61–86. **K dispozici:** PDF, PS, BibTeX

- Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová-Řezníčková, Petr Pajas: *PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation.* V: J. Nivre, E. Hinrichs (eds.): Proceedings of The Second Workshop on Treebanks and Linguistic Theories, Vaxjo University Press, Vaxjo, Sweden, 2003, pp. 57–68. **K dispozici:** PDF, PS, BibTeX

- Jan Hajič, Zdeňka Urešová: *Linguistic Annotation: from Links to Cross-Layer Lexicons.* V: J. Nivre, E. Hinrichs (eds.): Proceedings of The Second Workshop on Treebanks and Linguistic Theories, Vaxjo University Press, Vaxjo, Sweden, 2003, pp. 69–80. **K dispozici:** PDF, PS, BibTeX

- Václav Honetschläger: *Using a Czech Valency Lexicon for Annotation Support.* V: V. Matoušek, P. Mautner (eds.): Proceedings of the 6th International Conference on Text, Speech and Dialogue, České Budějovice, Czech Republic, Springer-Verlag Berlin Heidelberg New York, 2003, pp. 120–126. **K dispozici:** PDF, PS, BibTeX
- Markéta Lopatková, Jarmila Panevová: *Recent developments of the theory of valency in the light of the Prague Dependency Treebank.* V: Mária Šimková (ed.):,

  Veda Bratislava, Slovakia,

  2005. **K dispozici:** PDF, PS, BibTeX
- Zdeňka Urešová: *The Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View.* Jazykovedný ústav Ľ. Štúra, SAV, Bratislava, Slovakia, 2004. **K dispozici:** PDF, PS, BibTeX
- Zdeněk Žabokrtský: *Valency Lexicon of Czech Verbs (PhD thesis).*

  UFAL MFF UK, Prague, Czech Republic, 2005. **K dispozici:** PDF, PS, BibTeX

6.3 Nástroje

6.3.1 Netgraph

- Jiří Mírovský, Roman Ondruška: *NetGraph System: Searching through the Prague Dependency Treebank.* V: Prague Bulletin of Mathematical Linguistics, 77, MFF UK, Prague, Czech Republic, Prague, 2002, pp. 101–104. **K dispozici:** PDF, PS, BibTeX
- Roman Ondruška, Jiří Mírovský, Daniel Průša: *Searching through Prague Dependency Treebank-Conception and Architecture.* V: Proceedings of The First Workshop on Treebanks and Linguistic Theories, LML, Bulgarian Academy of Sciences and SfS, Tuebingen University, Sofia, Bulgaria and Tuebingen, Germany, 2002, pp. 114–122. **K dispozici:** PDF, PS, BibTeX

6.3.2 Morfologická analýza a tagging

- Jan Hajič: *Disambiguation of Rich Inflection (Computational Morphology of Czech).* Karolinum, Charles Univeristy Press, Prague, Czech Republic, 2004. **K dispozici:** BibTeX
- Jan Hajič: *Morphological Tagging: Data vs. Dictionaries.* V: Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference, Seattle, Washington, USA, 2000, pp. 94–101. **K dispozici:** PDF, PS, BibTeX
- Jan Hajič, Barbora Vidová Hladká: *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset.* V: Proceedings of the COLING–ACL Conference, Montreal, Canada, 1998, pp. 483–490. **K dispozici:** PDF, PS, BibTeX
- Barbora Vidová-Hladká: *Czech Language Tagging.* PhD thesis, ÚFAL MFF UK, Prague, Czech Republic, 2000. **K dispozici:** PDF, PS, BibTeX

6.3.3 Parsing

- Jan Hajič, Barbora Hladká, Daniel Zeman, Michael Collins, Lance Ramshaw, Christoph Tillmann, Eric Brill, Douglas Jones, Cynthia Kuo, Ozren Schwartz: *Core Natural Language Processing Technology Applicable to Multiple Languages.* Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA, 1998. **K dispozici:** PDF, PS, BibTeX
- Vladislav Kuboň: *Problems of Robust Parsing of Czech.* PhD thesis, ÚFAL MFF UK, 2001. **K dispozici:** PDF, PS, BibTeX
- Ryan McDonald, Fernando Pereira, Kiril Ribarov, Jan Hajič: *Non-Projective Dependency Parsing using Spanning Tree Algorithms.* V: (ed.): Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP),

  Vancouver, BC, Canada, Oct. 6-8, 2005, pp. 523–530. **K dispozici:** PDF, PS, BibTeX
- Kiril Ribarov: *Automatic Building of a Dependency Tree–The Rule-Based Approach and Beyond.* PhD thesis, ÚFAL MFF UK, Prague, Czech Republic, 2004. **K dispozici:** PDF, PS, BibTeX

- Daniel Zeman: *Parsing with a Statistical Dependency Model.* PhD thesis, ÚFAL MFF UK, Prague, Czech Republic, 2005. **K dispozici:** PDF, PS, BibTeX

6.3.4 Automatické přiřazování funktorů

- Petr Sgall, Zdeněk Žabokrtský, Sašo Džeroski: *A Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank.* V: R. M. Rodríguez, C. Paz Suárez Araujo (eds.): Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, (5), European Language Resources Association, 2002, pp. 1513–1520. **K dispozici:** PDF, PS, BibTeX

- Zdeněk Žabokrtský: *Automatic Functor Assignment in the Prague Dependency Treebank.* V: P. Sojka, I. Kopeček, K. Pala (eds.): Proceedings of the 3rd International Conference on Text, Speech and Dialogue, Brno, Czech Republic, Springer-Verlag Berlin Heidelberg New York, 2000, pp. 45–50. **K dispozici:** PDF, PS, BibTeX

- Zdeněk Žabokrtský: *Automatic Functor Assignment in the Prague Dependency Treebank.* V: ÚFAL Technical Report, 10, MFF UK, Prague, Czech Republic, 2001. **K dispozici:** PDF, PS, BibTeX

Kapitola 7

Distribuce a licence

Abyste mohli používat PDT 2.0, musíte vyplnit licenční formulář a elektronicky ho podepsat (výjimka viz níže). Text licence najdete v sekci 7.1.

Jsou dvě možnosti, jak získat PDT 2.0. Standardní cestou je objednat si plnou distribuci PDT 2.0 z Linguistic Data Consortium (<http://www.ldc.upenn.edu>); během objednávání budete přesměrováni na formulářovou licenční stránku, kterou musíte vyplnit, aby mohla být objednávka dokončena.

Druhou možností je stáhnout si část PDT 2.0 přímo z našich webovských stránek (<http://ufal.mff.cuni.cz/pdt2.0>); jde o přesnou kopii distribuce poskytované LDC, obsahuje však jen malou ukázku anotovaných dat. Registrační licenční formulář (http://quest.mff.cuni.cz/pdt-lic/pdt20-reg.html[1]) můžete vyplnit předem nebo i potom, ale použít cokoliv ze stažených nástrojů, ukázkových dat apod. smíte až po jeho vyplnění. Jinými slovy, *tato licence je platná až po vyplnění formuláře*.

Některé části distribuce mohou být kryty licencí GPL (GNU Public License). U těchto nástrojů a dat je to vždy explicitně uvedeno (obvykle jsou dostupny i z jiných zdrojů, např. z osobních stránek autorů či ze standardních archívů Open Source a GNU softwaru, jako je sourceforge.net). V tom případě má GPL přednost před touto licencí. Pokud *všechny* části PDT, které jste si stáhli či se kterými pracujete, podléhají licenci GPL, *nemusíte* se registrovat pod touto licencí. Byli bychom ovšem rádi, kdybyste tak přesto učinili (ač se vás pravidla a podmínky určené licencí v takovém případě netýkají). Mít co nejvíce registrovaných uživatelů nám v budoucnu pomůže získat další finanční zdroje.

7.1 Licenční ujednání

Licenční ujednání pro výzkumné užití Pražského závislostního korpusu, verze 2.0

mezi

```
Institutem formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova v˜Praze
Malostranské náměstí 25
118 00  Praha 1
Česká republika
pdt@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz
```

(Vlastník)

a

```
Jméno:
Instituce:
Adresa (ulice, město, PSČ):
Země:
Telefon(y):
```

[1] <http://quest.ms.mff.cuni.cz/pdt-lic/pdt20-reg.html>

```
Fax(y):
E-mail:
```

(Uživatel)
kde

A Pražský závislostní korpus verze 2.0 (PDT 2.0) je kolekce textových dat a dokumentace, obsahující lingvistické anotace a softwarové nástroje pro zpracování těchto dat, jak je v dokumentaci popsáno, vytvořená Vlastníkem v rámci následujících projektů: Ministerstvo školství České republiky, projekty č. VS96151, LN00A063, 1P05ME752, MSM0021620838 a LC536, Grantová agentura České republiky, granty č. 405/96/0198, 405/96/K214 a 405/03/0913, výzkumné fondy Matematicko-fyzikální fakulty Univerzity Karlovy v Praze, Grantová agentura Akademie věd České republiky, granty č. 1ET101120503 a 1ET101120413, Grantová agentura Univerzity Karlovy v Praze, granty č. 489/04, 350/05, 352/05 a 375/05 a U.S. NSF Grant #IIS9732388.

B Vlastník je držitelem autorských práv PDT 2.0 a je oprávněn udělit licenci Uživateli.

C Uživatel je akademická, vzdělávací či výzkumná instituce nebo jiná organizace či fyzická osoba, která si přeje používat PDT 2.0 pro výzkumné a/nebo vzdělávací účely.

Smluvní strany se dohodly na následujícím:

1. Tato dohoda je uzavřena *dnem odeslání* a vstupuje v platnost okamžitě.

2. Uživatel získává nevýhradní právo k používání, pozměňování, zvětšování či obohacování PDT 2.0 za účelem příméno nebo nepřímého získání informací v jakékoliv formě a množství, za podmínky, že PDT 2.0 samotný či jakýkoliv z něho odvozený produkt je použit pouze Uživatelem nebo jeho přímými spolupracovníky, zaměstnanci, manažery a/nebo jeho studenty z té samé instituce, a to výhradně pro výzkumné účely, a za podmínky, že tento soustavně dodržuje všechna ujednání a podmínky obsažené v této dohodě. Pokud jakákoliv část PDT 2.0 obsahuje svoji vlastní licenci či další omezení, platí více omezující verze licence, není-li v té které části specifikováno jinak. Všechna dokumentace, která je k dispozici v některém z formátů RTF, PDF nebo PostScript, musí být považována za osobní kopie příslušných autorů a jako s takovými s nimi musí být zacházeno.

3. Uživatel nepoužije PDT 2.0 samotný či jakýkoliv odvozený produkt (což zahrnuje mimo jiné i získané statistiky) na něm založený (jakkoliv malý může příspěvek PDT 2.0 k takovému odvozenému produktu být) žádným způsobem pro jakékoliv komerční účely, ani jako součást jakékoliv běžně používané aplikace, bez ohledu na to, zda je komerčního druhu.

4. Uživatel vloží následující poznámku do všech publikací či veřejně přístupných materiálů, bez ohledu na jejich formu (tištěných, elektronických nebo jiných), popisujících práci, ve které byl použit PDT 2.0: „Pražský závislostní korpus, verze 2.0 byl vytvořen Ústavem formální a aplikované lingvistiky, `http://ufal.mff.cuni.cz`.“ Do tištěných materiálů, jako jsou články, příspěvky v časopisech apod., by měla být vložena jedna publikace ze seznamu dokumentace PDT 2.0, nejvhodnější pro odkaz týkající se práce Uživatele. Do elektronických publikací umístěných na internetu by měl být vložen webovský odkaz na výše uvedenou webovskou stránku. Kvůli závazkům Vlastníka vůči držitelům autorských práv textů jsou textové příklady či citace z PDT 2.0 nebo jakéhokoliv odvozeného produktu (bez ohledu, zda obsahují nějaké anotace) omezeny na maximálně 200 slov na publikaci či sérii publikací na stejné téma (ať už tištěných, elektronických či v jakékoliv jiné formě).

5. Uživatel souhlasí s tím, že nebude dále šířit ani jakkoliv činit veřejně přístupným ani PDT 2.0, ani jakékoliv odvozené produkty na něm založené, jak je popsáno v odstavci 3, třetím stranám bez předchozího písemného svolení Uživatele, s výjimkou příkladů a citací, jak je uvedeno v odstavci 4.

6. Uživatel je zodpovědný za to, že přijme všechna bezpečnostní opatření potřebná k ochraně autorských práv Vlastníka na PDT 2.0 a zavazuje se podniknout všechny rozumné kroky k zajištění, že nedojde k neoprávněnému použití PDT 2.0 a jeho kopií, odvozených produktů ani jejich částí.

7. Jakékoliv použití PDT 2.0, které by se neřídilo specifikacemi uvedenými ve 3. odstavci této dohody (jako je např. komerční použití PDT 2.0), je předmětem samostatných jednání a písemných smluv mezi Uživatelem a Vlastníkem a/nebo dalších stran. Vlastník není obecně povinen přistoupit na taková jednání.

8. PDT 2.0 je poskytován bez jakýchkoliv záruk. Vlastník nezaručuje použitelnost PDT 2.0 pro žádný účel, bez ohledu na formulace, které se mohou vyskytovat na některých místech v doprovodné dokumentaci, vyjadřující zamýšlený účel a použití PDT 2.0.

9. Nahlásí-li Uživatel Vlastníkovi objevené chyby, nekonzistence nebo návrhy na opravy či vylepšení PDT 2.0, Vlastník se zavazuje: (a) zachovat tyto komentáře v tajnosti a použít je jen pro účely zdokonalení, opravení a/nebo údržby PDT 2.0, (b) nepředat tyto komentáře nikomu kromě důvěrně těm ze svých zaměstnanců nebo vedoucích pracovníků, kteří je potřebují znát pro výše uvedené účely.

10. Pokud Uživatel sám či kdokoliv, kdo jedná v jeho jménu, poruší některou z podmínek této dohody (a nemá k tomu písemné svolení Vlastníka), tato dohoda okamžitě pozbývá platnosti a Uživatel bezodkladně odstraní PDT 2.0, všechny jeho kopie a na na něm založené odvozené produkty ze svých zdrojů a všech zdrojů, které má pod kontrolou. Toto ukončení dohody nemá žádný vliv na nároky Vlasníka týkající se finančních dluhů a/nebo způsobených škod a/nebo další nároky.

11. Opominutí či nezdar Vlastníka vykonat nebo uplatnit práva vyplývající z této dohody nemůže být považováno za zřeknutí se těchto práv a nemůže zabránit vykonání či uplatnění těchto práv kdykoliv v budoucnu.

12. Tato dohoda končí, pokud (a) Uživatel odstraní všechny kopie PDT 2.0 a všech z něj odvozených produktů, (b) Uživatel či jeho instituce přestane existovat a nedošlo k přenesení všech jeho závazků na nový subjekt, který je v takovém případě považován za vázaného touto dohodou. Uživatel či jeho následník informuje Vlastníka o každém takovém přenesení závazků či následnictví; pokud tak neučiní, tato dohoda končí jeden měsíc po takovém přenesení závazků či následnictví. (c) Vlastník přestane existovat bez oficiálního následníka.

13. Vlastník bude považovat veškeré informace poskytnuté Uživatelem v rámci odeslání této dohody za důvěrné a neodhalí je dalším stranám, s výjimkou ve formě souhrnných informací, ze kterých nebude možno identifikovat jednotlivé uživatele. Ke svému zveřejnění může Vlastníka oprávnit pouze Uživatel písemným prohlášením.

14. Tato dohoda podléhá zákonům České republiky a veškeré spory týkající se této dohody budou řešeny jejím právním systémem.

Kapitola 8

Instalace

Pro usnadnění instalace PDT jsme připravili instalační programy pro operační systémy Linux a MS Windows. Poznamenejme však, že většinu částí PDT 2.0 lze používat přímo z distribučního CD-ROM nebo z jeho kopie; některé části mohou být instalovány samostatně pomocí svých vlastních instalačních programů.

Instalace na Linuxu. V kořenovém adresáři distribuce spusťte program **./Install-Linux.pl**. Budete vyzváni k výběru komponent, které si přejete nainstalovat, a k určení cílového adresáře na vašem systému; zvolené komponenty budou zkopírovány (a v některých případech i rozbaleny). Na závěr zobrazí instalační program informaci o tom, jak provést instalaci editoru stromů `TrEd`.

Instalace na MS Windows. Instalační program spusťte poklepem na ikonu **Install-Windows** v kořenovém adresáři distribuce. Instalační program nejprve ověří, zda je na vašem systému nainstalován Active State Perl ve správné verzi (nutný pro práci editoru stromů `TrEd`). Pokud není, bude zobrazena informace, odkud je možno tento Perl stáhnout a nainstalovat. Instalační program vám pak umožní vybrat komponenty PDT 2.0, které si přejete nainstalovat, a určit cílový adresář. Tyto komponenty jsou pak do zvoleného adresáře na vašem systému zkopírovány. (Upozorňujeme, že instalační program pro MS Windows nenabízí instalaci nástrojů pro automatickou anotaci textů; tyto nástroje jsou k dispozici jen pro Linux.) Na závěr je spuštěn samostatný instalační program editoru stromů `TrEd`.

Připravené instalační programy pro Linux a MS Windows nezahrnují instalaci `Netgraphu`, nástroje pro vyhledávání v korpusu. Pokud si přejete `Netgraph` nainstalovat, postupujte podle těchto návodů:

- Rychlá instalace Netgraph klienta: `doc/tools/netgraph/README_QUICK_INSTALL_CLIENT`
- Rychlá instalace Netgraph serveru: `doc/tools/netgraph/README_QUICK_INSTALL_SERVER`
- Manuál k Netgraph klientu: `doc/tools/netgraph/netgraph_manual.html`
- Instalační manuál Netgraph serveru: `doc/tools/netgraph/netgraph_server_install.html`

Kapitola 9

Zásluhy

Následující lidé tím či oním způsobem přispěli k vytvoření a vývoji Pražského závislostního korpusu, verze 2.0. Uvedeni jsou v abecedním pořadí (podle příjmení), s výjimkou publikací (jako jsou Pokyny pro anotátory), u kterých je zachováno publikované pořadí autorů.

- **PDT 2.0**
  - **Morfologická rovina**
    - *Koordinátor*: Barbora Hladká
    - *Odborný garant*: Jan Hajič
    - *Anotační manuál*
      - *Anglická verze*: Daniel Zeman, Jan Hajič, Jiří Hana, Hana Hanová, Barbora Hladká, Emil Jeřábek
    - *Anotátoři*: Martin Buben, Jiří Hana, Hana Hanová, Emil Jeřábek, Lenka Kebortová, Kristýna Kupková, Pavel Květoň, Jiří Mírovský, Andrea Pfimpfrová
    - *Poanotační kontrola*: Jiří Hana, Hana Hanová, Barbora Hladká, Emil Jeřábek
    - *Kontrola po vydání PDT 1.0*: Pavel Květoň, Petr Pajas, Pavel Pecina, Jan Štěpánek, Daniel Zeman, Zdeněk Žabokrtský
    - *Software a technická podpora*: Jan Hajič, Jiří Hana, Karel Skoupý
  - **Syntakticko-analytická rovina**
    - *Koordinátor*: Jan Hajič
    - *Odborný garant*: Jarmila Panevová
    - *Anotační manuál*
      - *Česká verze*: Alla Bémová, Eva Buráňová, Jan Hajič, Jiří Kárník, Petr Pajas, Jarmila Panevová, Zdeňka Urešová, Jan Štěpánek
      - *Překlad do angličtiny*: Eva Hajičová, Zdeněk Kirschner, Petr Sgall
    - *Anotátoři*: Alla Bémová, Eva Buráňová, Jiří Kárník, Petr Pajas, Jan Štěpánek, Zdeňka Urešová
    - *Poanotační kontrola*: Eva Buráňová, Jakub Dotlačil, Petr Pajas, Jan Štěpánek
    - *Kontrola po vydání PDT 1.0*: Petr Pajas, Jan Štěpánek, Zdeněk Žabokrtský
    - *Software a technická podpora*: Jan Hajič, Jiří Havelka, Michal Křen, Petr Pajas, Jan Štěpánek, Daniel Zeman
  - **Tektogramatická rovina**
    - *Koordinátor*: Jan Hajič
    - *Odborný garant*: Eva Hajičová, Jarmila Panevová, Petr Sgall
    - *Anotační manuál*
      - *Česká verze*: Marie Mikulová, Alla Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová-Řezníčková, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský
      - *Překlad do angličtiny*: Alena Böhmová, Silvie Cinková, Eva Hajičová, Pavel Straňák

    - *Výcvik anotátorů*: Veronika Kolářová-Řezníčková, Ivona Kučerová
    - *Struktura tektogramatické anotace, přiřazování funktorů a valenčních rámců*
        - *Koordinátor*: Jan Hajič
        - *Anotátoři*: Alla Bémová, Eva Buráňová, Jakub Dotlačil, Marie Mikulová, Magda Razímová, Kateřina Součková, Zdeňka Urešová, Jana Vejvodová
        - *Poanotační kontrola*: Václava Benešová, Ondřej Bojar, Jan Hajič, Markéta Lopatková, Petr Pajas, Jan Štěpánek, Zdeňka Urešová, Jana Vejvodová, Šárka Zikánová-Lešnerová, Zdeněk Žabokrtský
        - *Software a technická podpora*: Alena Böhmová, Petr Pajas, Jan Štěpánek, Zdeněk Žabokrtský
    - *Aktuální členění*
        - *Koordinátor*: Jiří Havelka
        - *Pokyny pro anotátory*: Kateřina Veselá
        - *Anotátoři*: Eva Buráňová, Anna Dostálová, Barbora Smrčková, Kateřina Veselá, Šárka Zikánová-Lešnerová
        - *Poanotační kontrola*: Jakub Dotlačil, Jiří Havelka, Barbora Smrčková, Kateřina Součková, Kateřina Veselá, Šárka Zikánová-Lešnerová
        - *Software a technická podpora*: Jiří Havelka
    - *Koreference*
        - *Koordinátor*: Zdeněk Žabokrtský
        - *Pokyny pro anotátory*: Veronika Kolářová-Řezníčková, Lucie Kučová
        - *Anotátoři*: Kateřina Černá, Lucie Kučová, Jana Vejvodová
        - *Poanotační kontrola*: Lucie Kučová, Petr Pajas, Magda Razímová, Jiří Semecký, Jan Štěpánek, Zdeněk Žabokrtský
        - *Software a technická podpora*: Oliver Čulo, Petr Pajas, Zdeněk Žabokrtský
    - *Gramatémy*
        - *Koordinátor*: Zdeněk Žabokrtský
        - *Pokyny pro anotátory*: Magda Razímová
        - *Anotátoři*: Kateřina Marková, Kamila Pacovská, Magda Razímová
        - *Software a technická podpora*: Daniel Zeman
    - *PDT Vallex*
        - *Koordinátor*: Petr Pajas
        - *Anotátoři*: Alla Bémová, Veronika Kolářová-Řezníčková, Markéta Lopatková, Zdeňka Urešová
        - *Poanotační kontrola*: Alla Bémová, Jan Hajič, Veronika Kolářová-Řezníčková, Markéta Lopatková, Petr Pajas, Zdeňka Urešová
        - *Software a technická podpora*: Petr Pajas, Zdeněk Žabokrtský

- **NÁSTROJE**
    - **TrEd** Petr Pajas
    - **NTrEd** Petr Pajas, Zdeněk Žabokrtský
    - **Netgraph** Jiří Mírovský, Roman Ondruška
    - **Segmentace a tokenizace českých textů** Jan Hajič, Michal Křen
    - **Morfologický analyzátor češtiny** Jan Hajič, Jaroslava Hlaváčová
    - **Tagger** Jan Hajič
    - **Parser** Michael Collins, Václav Honetschläger
    - **Přiřazování analytických značek PDT** Petr Pajas, Zdeněk Žabokrtský

- **PUBLIKACE**
    - *Sběr, formátování*: Barbora Hladká, Petr Homola, Jiří Semecký

- **CD-ROM, webovské stránky**

- *Adresářová struktura*: Václav Honetschläger, Zdeněk Žabokrtský
- *Instalační skript*: Ondřej Bojar
- *Validace*: Petr Podveský
- *Editoři Průvodce PDT*: Václav Honetschläger, Zdeněk Žabokrtský
- *Obal*: Alena Böhmová
- *Webovské stránky*: Václav Honetschläger

Kapitola 10

Poděkování

Vývoj Pražského závislostního korpusu, verze 2.0 byl podporován těmito organizacemi, projekty a sponzory:

- Ministerstvo školství a mládeže České republiky[1], projekty č. VS96151, LN00A063, 1P05ME752, MSM0021620838 a LC536,
- Grantová agentura České republiky[2], granty č. 405/96/0198, 405/96/K214 a 405/03/0913,
- výzkumné fondy Matematicko-fyzikální fakulty[3],
- Univerzita Karlova v Praze[4], Česká republika,
- Grantová agentura Akademie věd České republiky, Praha, Česká republika[5], projekty č. 1ET101120503, 1ET101120413 a 1ET201120505,
- Grantová agentura Univerzity Karlovy v Praze[6], granty č. 489/04, 350/05, 352/05 a 375/05,
- a U.S. NSF[7] Grant #IIS9732388.

Děkujeme našim partnerům v uvedených projektech, jmenovitě Ústavu Českého národního korpusu[8] a Ústavu teoretické a komputační lingvistiky[9] Filozofické fakulty Univerzity Karlovy v Praze, prvnímu z nich za poskytnutí původních surových dat a oběma za vhled do problematiky během diskusí v devítiletém období, které vyvrcholilo zveřejněním PDT 2.0. Jsme vděční také poskytovatelům textů, jedná se o Lidové Noviny Publishers, Mladou Frontu Dnes, původní vydavatele časopisu Českomoravský Profit a Vesmír s.r.o., za svolení k vložení jejich textů do distribuce PDT. Rádi bychom poděkovali také mnoha dalším, jejichž práce byla využita během vytváření PDT 2.0, zvláště autorům obrovského množství úžasných nástrojů dostupných pod licencí GPL nebo jiným způsobem nám přístupných, od Linuxu přes Perl po všechny drobné, ale čas šetřící programátorské klenoty.

[1] <http://www.msmt.cz>
[2] <http://www.gacr.cz>
[3] <http://www.mff.cuni.cz>
[4] <http://www.cuni.cz>
[5] <http://www.cas.cz>
[6] <http://www.cuni.cz/UK-33.html>
[7] <http://www.nsf.gov>
[8] <http://ucnk.ff.cuni.cz/>
[9] <http://utkl.ff.cuni.cz/>