M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

ÚPRAVY A DOPLŇKY PRAŽSKÉHO ZÁVISLOSTNÍHO KORPUSU (OD PDT 2.0 K PDT 3.0)

Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová,
Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský

ÚFAL Technical Report
TR-2013-53

ISSN 1214-5521

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czech Republic

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)

Marie Mikulová
Eduard Bejček
Jiří Mírovský
Anna Nedoluzhko
Jarmila Panevová
Lucie Poláková
Pavel Straňák
Magda Ševčíková
Zdeněk Žabokrtský

Poděkování

Dokument vznikl za finanční podpory projektů GA ČR P406/2010/0875, P406/12/P175 a P406/12/0658 a projektu MŠMT LINDAT-Clarin LM2010013.

Obsah

Úvod

V roce 2006 byla vydána a pro zájemce zpřístupněna druhá verze Pražského závislostního korpusu (Prague Dependency Treebank 2.0, dále PDT 2.0 [4]).[1] Tato verze obsahuje žurnalistické texty vybrané z Českého národního korpusu. Tyto texty byly anotovány na třech úrovních: 2 miliony výskytů slov byly anotovány na morfologické rovině, jejich část v rozsahu 1,5 milionu slov (odpovídajících 88 tis. větám) byla anotována na analytické rovině (rovině povrchové syntaxe) a část o rozsahu 0,8 milionu slov (49 tis. vět) byla anotována také na tektogramatické rovině (rovině hloubkové syntaxe). Tato textová databáze opatřená bohatou gramatickou anotací slouží především dvěma účelům:

- V komputační lingvistice byla použita k vývoji a zdokonalení nástrojů pro automatické zpracování češtiny, zejména pak pro strojový překlad. Anotovaná data slouží pro strojové učení z dat i pro další aplikace založené na přirozeném jazyce.
- Další využití PDT 2.0 spočívá v tom, že slouží jako materiálová databáze pro lingvistické studie ze současné češtiny. Na podkladě tohoto materiálu vznikly četné vědecké články, příspěvky na zahraničních konferencích a workshopech, ale také disertační a diplomové práce.

V obou uvedených směrech byli uživateli PDT 2.0 hlavně vědečtí pracovníci ÚFAL, jejich studenti a doktorandi zaměření jak informaticky, tak humanitně. Databáze PDT 2.0 je známa a také využívána v zahraničních vědeckých kolektivech (např. slovenština, slovinština, dánština, řečtina, arabština aj.) a je ceněna jako jeden z nejbohatěji anotovaných korpusů přirozených jazyků.

Od počátku bylo známo, že v první etapě anotování jsou některá přijatá řešení pouze předběžná; to platí zejména pro oblast hloubkové syntaxe (tektogramatiky). Bylo rovněž jasné, že i přes několikanásobné kontroly anotovaných dat jsou v nich obsaženy chyby lidských anotátorů, ovlivňující pak další kroky v procesu manuálního anotování.

Pro další využívání anotovaného počítačového korpusu PDT 2.0 bylo rozhodnuto, že tato procedura nebude pokračovat cestou extenze dat, ale spíše cestou jejich zkvalitňování a zapojení úrovně diskurzu a textových vztahů.

Jako mezikrok před vývojem PDT 3.0 byla v r. 2012 zveřejněna otevřená verze PDT 2.5 [2]. PDT 2.5 obsahuje stejná data (vstupní texty) jako PDT 2.0, ale anotování je rozšířeno o několik kroků:

- do tektogramatické anotace byl zaveden a soustavně anotován jeden nový substantivní gramatém (*typegroup*),
- byl aplikován slovník víceslovných pojmenování (pojmenovaných entit a jiných sousloví),
- na analytické rovině byl aplikován algoritmus segmentace souvětí na klauze,
- byla provedena oprava jednotlivých chyb, a to na všech třech úrovních.

Informace o PDT 2.5 včetně podrobné dokumentace jsou k dispozici (především díky péči Mgr. Jana Štěpánka, Ph.D.) na adrese http://ufal.mff.cuni.cz/pdt2.5. Zároveň byl v roce 2012 publikován korpus PDiT 1.0 (Prague Discourse Treebank [6]), ve kterém byla poprvé veřejnosti zpřístupněna anotace jmenné koreference, asociační anafory („bridging“) a mezivýpovědních významových vztahů („discourse relations“). Vstupními daty byla data korpusu PDT 2.5. Informace o PDiT 1.0 jsou k dispozici na stránce projektu

[1] První verze tohoto korpusu (Prague Dependency Treebank 1.0, PDT 1.0) byla vydána v roce 2001 a obsahovala morfologickou a analytickou anotaci.

ufal.mff.cuni.cz/pdit. Plná data korpusů PDT 2.5 a PDiT 1.0 jsou pod licencí Creative Commons 3.0-BY-NC-SA dostupná v repozitáři LINDAT CLARIN.

Důkladnější inovace byly provedeny ve verzi PDT 3.0 [1] (korpus zahrnuje i anotace obsažené v PDT 2.5 a PDiT 1.0). Jejich lingvistická podstata, zdůvodnění, pokyny pro anotaci a exemplifikace jsou obsahem této zprávy. Její jedna část je věnována nově zavedeným nebo upraveným jevům tektogramatické roviny. V její další části jsou představeny jevy textové koreference, mezivětných vztahů, žánrové klasifikace dat a způsob jejich anotování v PDT 3.0. Posledně jmenované jevy přesahují hranice tektogramatického anotování (hloubkové syntaxe), ale představují materiálovou základnu pro studium diskurzu, jeho stavby a koherence. Data korpusu PDT 3.0 budou opět přístupná v repositáři LINDAT CLARIN.

A. Úpravy a doplňky na tektogramatické rovině korpusu

1 Gramatémy a atribut *sentmod*

Jarmila Panevová a Magda Ševčíková

Popis

Ve schématu gramatémů byly v korpusu PDT 3.0 provedeny změny jak v souboru gramatémů substantivních, tak v gramatémech slovesných.

Substantivní gramatémy

Byl zaveden jeden nový gramatém těsně spojený s gramatémem čísla: gramatém souborovosti *typgroup*; jeho hodnoty spolu s kritérii a příklady jsou popsány níže (viz 1.1). Další substantivní gramatémy zůstávají beze změny.

Slovesné gramatémy

Úpravy slovesných gramatémů byly vyvolány snahou lépe teoreticky ukotvit (morfologické) významosloví slovesa. Teoretické zlepšení spočívá v zavedení nového gramatému diateze *diatgram*. Tento gramatém postihuje distinkce spojené se slovesným rodem. Jeho hodnoty spolu s kritérii a příklady budou popsány níže (viz 1.3). Zavedení atributu *diatgram* vedlo ke změně anotačního schématu pro slovesné kategorie, jak byly anotovány v PDT 2.0 a PDT 2.5.

Současné schéma pro anotaci slovesných gramatémů vypadá takto:

- Gramatém slovesné modality (*verbmod*; v manuálu [5] kap. 4.5.9) byl **nahrazen** gramatémem skutečnostní modality (*factmod*). K tomu viz dále 1.2.

Byly **zrušeny** slovesné gramatémy:

- dispoziční modality (*dispmod*) (v manuálu [5] kap. 4.5.11)
- rezultativu (*resultative*) (v manuálu [5] kap. 4.5.14)

Beze změny byly ponechány slovesné gramatémy:

- deontické modality (*deontmod*) (v manuálu [5] kap. 4.5.10)
- vidu (*aspect*) (v manuálu [5] kap. 4.5.12)
- času (*tense*) (v manuálu [5] kap. 4.5.13)
- iterativnosti (*iterativeness*) (v manuálu [5] kap. 4.5.15)

Byl **přidán** slovesný gramatém:

- diateze (*diatgram*)

1.1 Substantivní gramatém *typgroup*

Magda Ševčíková

Popis

Hodnotami gramatému *typgroup* je reprezentována sémantická opozice souborového významu vs. významu jednotlivých entit (hodnoty *group* vs. *single*; třetí hodnota *nr* se užívá pro nerozhodnutelné případy). Česká substantiva jako *ruce*, *boty* nebo *klíče* odkazují svými tvary množného čísla k páru nebo k většímu typickému souboru častěji než k většímu množství jednotlivých entit; např. plurálový tvar *ruce* označuje pár nebo několik párů rukou častěji než prosté větší množství horních končetin, tvar *boty* obvykle vyjadřuje pár nebo několik párů bot, tvar *klíče* odkazuje ke svazku nebo několika svazkům klíčů. Protože souborový význam vyjadřuje většina českých konkrétních substantiv a tento význam má vliv na spojitelnost substantiv s číslovkami (při označování souborů se substantivum spojuje pouze se souborovými číslovkami, při vyjadřování jednotlivých entit s číslovkami základními; srov. *dvoje boty* vs. *dvě boty*), je souborový význam považován za gramatikalizovaný význam českých substantiv.

V češtině je souborový význam vyjadřován formálně nepříznakovými plurálovými tvary substantiv. Protože souborový význam lze od jiných významů plurálových substantiv odlišit díky přítomnosti číslovky (k souvýskytu substantiva se souborovou číslovkou ovšem v korpusových datech dochází zřídka) nebo na základě kontextu nebo znalosti světa, identifikace tohoto významu u většiny substantiv v korpusu by byla úkolem pro ruční anotaci. Nicméně, vzhledem k tomu, že na základě pilotního anotačního experimentu byla očekávána nízká frekvence souborového významu, ve snaze o co nejefektivnější anotaci byly ručně anotovány pouze plurálové tvary těch substantiv, pro která je souborový význam považován za prototypický. Souborový význam je jako prototypický očekáván pro následující skupiny substantiv:

- substantiva označující párové nebo souborové části těla (př. *uši*, *prsty*, *vlasy*),
- oblečení a doplňky určené pro tyto tělesné části (př. *náušnice*, *rukavice*),
- rodinní příslušníci tvořící páry nebo skupiny, jako *rodiče*, *dvojčata*,
- předměty denní potřeby a potraviny prodávané nebo používané v typickém množství (e.g. *klíče*, *špagety, sirky*, *sušenky*).

Anotační procedura

V PDT 3.0 (původně v datech PDT 2.5) je gramatém *typgroup* poloautomaticky anotován u všech pojmenovacích sémantických substantiv (uzly se *sempos* = *n.denot*|*n.denot.neg*). Nejprve byla provedena ruční anotace: výskyty pro ruční anotaci byly vybrány na základě seznamu tektogramatických lemmat (t-lemmat) prototypických souborových substantiv. Do tohoto seznamu byla zařazena substantiva, která byla v datech PDT 2.0 a SYN2005 doprovázena souborovou číslovkou, dále substantiva uváděná v mluvnických příručkách a odborných statích týkajících se kategorie čísla, seznam byl doplněn také na základě introspekce. Pro t-lemmata z výsledného seznamu bylo v datech PDT 3.0 nalezeno více než 600 výskytů plurálových forem (většina z nich náležela k těmto t-lemmatům: *oko*, *rodič*, *ruka*, *bota*).

Tyto výskyty byly anotovány paralelně dvěma anotátory, s mezianotátorskou shodou 75.1% (Cohenovo kappa 0.67). Po ukončení ruční anotace byly výskyty, na jejichž anotaci se anototáři neshodli, rozhodnuty třetím anotátorem. Tímto anotátorem byly zkontrolovány rovněž výskyty anotované oběma anotátory shodně – s cílem zkontrolovat správnost a konzistenci anotace.

Souborový význam je úzce spjat s gramatickou kategorií čísla substantiv; v češtině je kategorie čísla konstituována opozicí singuláru a plurálu. V souvislosti s ruční anotací souborového významu byly hodnoty gramatému čísla *number* (hodnoty *sg, pl,* a *nr*) změněny oproti původní anotaci (PDT 2.0) následujícím způsobem: pokud byla plurálová forma identifikována jako vyjádření jednoho souboru (*typgroup=group*), hodnota gramatému *number* byla změněna na *sg*; pokud plurálová forma označovala více souborů (*typgroup=group*), hodnota gramatému *number* se neměnila (zůstala *pl*); pokud anotátoři nebyli s to rozhodnout mezi významem jednoho souboru a významem více souborů (*typgroup=group*), v gramatému *number* byla vyplněna hodnota *nr*.

U pojmenovacích sémantických substantiv, která nebyla zahrnuta do ruční anotace, byl gramatém *typgroup* anotován automaticky. Automatická anotace postupovala podle jednoduchého „algoritmu“ sestávajícího ze dvou kroků: v prvním kroku byla u substantiv, která byla doprovázena souborovou číslovkou *jedny* (s výjimkou pluralií tantum), vyplněna v gramatému *typgroup* hodnota *group*, v této souvislosti byla hodnota gramatému *number* změněna na *sg*; u substantiva doprovázeného souborovou číslovkou vyšší hodnoty (*dvoje, troje* atd.) byla v gramatému *typgroup* vyplněna hodnota *group*, zatímco gramatém *number* zůstal nezměněn (tj. *pl*). Ve druhém kroku byla všem zbývajícím substantivům v gramatému *typgroup* přiřazena hodnota *single*, hodnota gramatému *number* v těchto případech zůstala stejná jako ve výchozí anotaci (PDT 2.0).

V datech PDT 3.0 se vyskytují následující kombinace hodnot gramatémů *number* a *typgroup*:

- *sg.group* význam jednoho souboru, vyjadřován plurálovou formou substantiva,
- *pl.group* význam více souborů, vyjadřován plurálovou formou substantiva,
- *nr.group* případy, kdy plurálová forma substantiva označuje jeden nebo více souborů,
- *sg.single* význam jedné entity, vyjadřován singulárovou formou substantiva,
- *pl.single* význam více jednotlivých entit, vyjadřován plurálovou formou substantiva,
- *nr.single* u uzlů, u nichž nebyla rozpoznána hodnota čísla (*number=nr*), byla v gramatému *typgroup* defaultně vyplněna hodnota *single*,
- *nr.nr* tato kombinace je přiřazena víceznačným případům: u těchto případů nebylo možné vyloučit kombinaci *sg.group* ani *pl.group* ani *pl.single* (v kombinaci *nr.nr* ovšem není zahrnuta kombinace *sg.single*!).

Související literaturu viz publikace [7], [8], [9] a [10] v seznamu referencí na konci tohoto dokumentu.

Příklady

Pro každé pojmenovací sémantické substantivum jsou kurzivou uvedeny hodnoty gramatémů *number* a *typgroup*; substantiva, u kterých byla hodnota gramatému *typgroup* určena v rámci ruční anotace, jsou označena tučně.

(1) Navlékla bych si dvoje **ponožky**.*pl.group* a hrála bych naboso, dokud by mi někdo nesehnal nějaké **boty**.*sg.group*.
(2) Pro něho připravila firma.*sg.single* Lotto.*sg.single* speciální **kopačky**.*nr.group*.
(3) Sečíst pouhým okem.*sg.single* stranickou příslušnost.*sg.single* zvednutých **rukou**.*pl.single* bylo ve dvousetčlenné Poslanecké sněmovně.*sg.single* nemožné.
(4) ... je to také odpověď.*sg.single* na vzdělávací požadavky.*pl.single* **rodičů**.*nr.nr*, žáků.*pl.single*, ale i měnícího se trhu.*sg.single* práce.*sg.single*.
(5) Obsah PCB.*nr.single* ve vepřovém a drůbežím mase je již minimální.

1.2 Slovesný gramatém *factmod*

Magda Ševčíková

Popis

Gramatém *factmod* zachycuje, zda je děj (stav atd.) mluvčím ztvárněn jako daný nebo hypotetický (tzv. skutečnostní modalita v termínech *Mluvnice češtiny* [11]). Tyto modální významy jsou v povrchové stavbě věty vyjadřovány morfologickou kategorií slovesného způsobu. Děj je jako daný vyjádřen indikativním tvarem slovesa. Pokud jde o děje prezentované jako hypotetické, jsou rozlišovány dva typy hypotetických dějů v souladu s existencí dvou kondicionálových paradigmat českého slovesa: děje, které mohou nastat (potenciální děje), jsou vyjadřovány kondicionálem přítomným, děje, které nastat nemohou (ireálné děje), jsou jednoznačně vyjádřeny kondicionálem minulým (tyto tvary jsou ovšem v současné češtině často nahrazovány kondicionálem přítomným). Ačkoli imperativ je teoreticky považován za prostředek vyjadřující komunikační funkci výpovědi, děje vyjádřené (syntetickými i analytickými) tvary imperativu jsou v PDT 3.0 zachyceny hodnotou gramatému *factmod* jako další modální význam (jako děj předkládaný mluvčím jako požadovaný), důvodem tohoto rozhodnutí je především snaha reprezentovat všechny významy vyjadřované kategorií slovesného způsobu v anotovaných datech jediným prostředkem (gramatémem).

Pro gramatém *factmod* byly definovány následující čtyři hodnoty:

- ***asserted*** děje prezentované jako dané (tvrzené)
- ***potential*** děje prezentované jako potenciální
- ***irreal*** děje prezentované jako ireálné
- ***appeal*** děje prezentované jako požadované

Modální významy zachycované gramatémem *factmod* jsou vyjadřovány pouze finitními slovesnými tvary, u uzlů reprezentujících nefinitní slovesné formy (infinitivy, participia a přechodníky) byla v tomto gramatému vyplněna (pátá) hodnota *nil*.

Gramatémem *factmod* je nahrazen gramatém *verbmod*, který byl použit v tektogramatické anotaci PDT 2.0. Rozdíl mezi těmito dvěma gramatémy spočívá především v zachycení (obou typů) hypotetických dějů. V PDT 2.0 byla u uzlů reprezentujících slovesné tvary vyjadřující potenciální i ireálné děje vyplněna v gramatému *verbmod* hodnota *cdn* a tyto významy byly rozlišeny až hodnotou gramatému času *tense*; taková anotace byla v rozporu s teoretickým poznatkem, že kondicionál v češtině nevyjadřuje časové významy. Hodnoty *potential* a *irreal* rozlišované v rámci gramatému *factmod* umožňují v PDT 3.0 reflektovat rozdíl mezi významy kondicionálu přítomného a minulého teoreticky adekvátním způsobem. V souvislosti s touto změnou byla v PDT 3.0 hodnota gramatému *tense* u uzlů, jimž byla v gramatému *factmod* přiřazena hodnota *potential* nebo *irreal*, změněna na hodnotu *nil*.

Související literaturu viz publikace [12], [13], [14] a [15] v seznamu referencí na konci tohoto dokumentu.

Anotační procedura

Gramatém *factmod* byl vyplněn u uzlů reprezentujících finitní slovesnou formu, a to v rámci poloautomatické procedury. Během automatické části anotace byly intenzivně využívány informace z morfologické roviny. Následovala ruční kontrola zaměřující se na méně frekventované hodnoty a na identifikaci výjimek ve specifických kontextech.

Příklady

Slovesům v příkladech (1) až (5) byly přiděleny základní čtyři hodnoty gramatému *factmod*; př. (1) navíc obsahuje infinitiv ohodnocený hodnotou *nil*. Věty se syntetickými imperativními formami vyjadřují imperativní větnou modalitu (př. (4); viz gramatém *sentmod*, hodnota *imper*), zatímco analytické imperativní formy vystupují ve větách s přací modalitou (př. (5); hodnota *desid* gramatému *sentmod*). V př. (6) je ireálný děj vyjádřen tvarem kondicionálu přítomného místo jednoznačného kondicionálu minulého; tato substituce není hodnotou *factmod* zachycena, uzlu je na základě formálních rysů přidělena hodnota *potential*. Hodnota gramatému *factmod* je vždy zaznamenána u příslušného slovesného tvaru (jeho plnovýznamové části; celý slovesný tvar je zvýrazněn tučně).

(1) Pokud **dojde**.*asserted* k omylu, **lze**.*asserted* zpětně **požádat**.*nil* nového majitele, **aby poukázal**.*potential* peníze správnému majiteli cenných papírů.
(2) Uhrát tu remízu **by bylo**.*potential* úspěchem.
(3) Většina bangladéšského muslimského obyvatelstva **by** za normálních okolností inkriminované interview samozřejmě vůbec **bývala nezaznamenala**.*irreal*.
(4) **Zvedněte**.*appeal* telefon a **zavolejte**.*appeal*.
(5) **Ať si** provincie konečně **oddychne**.*appeal*.
(6) Svatý pijan Joseph Roth **by** dnes **oslavil**.*potential* rovnou stovku.

1.3 Slovesný gramatém *diatgram*

Jarmila Panevová

Popis

Teoretickým důvodem zavedení a uspořádání gramatému *diatgram* bylo propojení atributů: aktivum, pasivum, rezultativ (prostý a posesivní), recipientní pasivum, dispoziční diateze a (volněji připojený) reflexivní deagentiv do jedné kategorie diateze i s ohledem na to, že se jejich hodnoty navzájem nekombinují. Tyto kategorie jsou chápány jako morfologické významy slovesných forem, v případě dispoziční diateze a reflexivního deagentivu jsou vázány na jisté syntaktické podmínky. Některé z těchto kategorií jsou produktivní více (pasivum, deagentiv), jiné jsou produktivní méně (rezultativ, recipient), přesto je pokládáme za slovesné kategorie morfologické.

Pro každý určitý slovesný tvar musí být vybrána jedna z hodnot:

(a) ***act*** Karlovu univerzitu **založil**.*act* Karel IV.
(b) ***pas*** Karlova univerzita **byla založena**.*pas* Karlem IV.
(c) ***res1*** Obchod **je otevřen**.*res1* denně mimo neděli.
(d) ***res2.1*** Obchod **má otevřeno**.*res2.1*
res2.2 Firma už **má** smlouvu **podepsánu**.*res2.2*
(e) ***recip*** Horníci **dostanou** v lednu **přidáno**.*recip.*
(f) ***disp*** Tento produkt **se** dobře **prodává**.*disp*
(g) ***deagent*** **Čeká se**.*deagent* krutá zima.
Knihy **se** dnes **vydávají**.*deagent* i v elektronické podobě.

Pro hodnoty (a), (d), (e), (f) a (g) se opíráme o poměrně spolehlivou formální oporu (přítomnost pomocných sloves *mít* a *dostat* (v (d) a (e)), reflexivní podoba a hodnotící adverbium v (f), reflexivní podoba a všeobecný konatel v (g)). Značně nejasné je vedení hranice mezi (b) a (c), protože jsou formálně totožné.

Související literaturu viz publikace [16], [17] a [18] v seznamu referencí na konci tohoto dokumentu.

Anotační procedura

V PDT 3.0 byly hodnoty atributů *diatgram* anotovány podle nového uspořádání gramatémů semiautomaticky.

(d) Rezultativ posesivní (*res2.1* a *res2.2*) byl vyhledán pomocí skriptu (založeného na souvýskytu slovesa *mít* a *-n/-t* participia). Manuálně byla řešena syntaktická homonymie:

- v pozici subjektu je konatel (ACT) → hodnota gramatému *res2.1* (př. (2)).
- v pozici subjektu je adresát (ADDR) → hodnota gramatému *res2.2* (př. (1)). Tektogramatická struktura byla v tomto případě ručně upravena tak, aby v subjektu nebyl konatel, ale jiný aktant a byl doplněn *#Gen*.ACT.

(e) Případy **recipientní diateze** (*recip*) byly vyhledány automaticky pomocí skriptu založeného na souvýskytu slovesa *dostat* a –n/-t participia (př. (3)).

(f) Případy **dispoziční diateze** (*disp*) byly vyhledány a doplněny automaticky pomocí skriptu: V PDT 2.0 byl vyplněn gramatém *gram/disp*, pokud se jeho hodnota rovnala 1, šlo o případ této diateze, která byla přenesena do jiného atributu (př. (4)).

(g) Reflexivní deagentiv (*deagent*) byl vyhledán automaticky pomocí skriptu opírajícího se o reflexivum s všeobecným konatelem (př. (5) a (6)).

(b) (c) Kvůli časté homonymii mezi **pasivem** (*pas*) a **prostým rezultativem** (*res1*) bylo použito jak automatické vyhledávání pomocí několika skriptů, tak manuální kontrola

založená na lingvistických hypotézách. Skripty byly založeny na případech souvýskytu slovesa *být* a *–n/-t* participia, ty byly dále filtrovány podle dalších hledisek:

- V těch případech, kde byl přítomen ACT, byl vyplněn gramatém *diatgram* hodnotou *pas* (př. (7)).
- Konstrukcím s tvarem *–n/-t* participia ve středním rodě slovesa dokonavého v souvýskytu se všeobecným konatelem (*#Gen*) byl připsán gramatém *diatgram* s hodnotou *res1* (př. (8)).
- Dalším kritériem byla anotace na analytické rovině, kde už anotátoři PDT 2.0 řešili vztah děje (*pas*) a stavu (*res*). Tam, kde analytická struktura byla tvořena slovesem *být* a predikátem nominálním (*PNom*), byla v anotaci PDT 3.0 změněna struktura a gramatém *diatgram* byl vyplněn hodnotou *res1*. Tam, kde tvar slovesa *být* byl na analytické rovině zachycen jako *AuxV*, byla hodnota gramatému *diatgram* nastavena na *pas*.
- 750 případů bylo podrobeno ruční anotaci. Případy, kde se výsledky ručního anotování neshodovaly s výsledkem skriptu (227 případů), byly podrobeny dalšímu kolu ruční kontroly. Ruční kontrolou prošly i případy, které nebyly žádným z uvedených skriptů zachyceny (108 případů).

  (a) Zbývající případy dostávají nepříznakovou hodnotu *act*.

Celkové počty jednotlivých hodnot gramatému *diatgram* v korpusu PDT 3.0 jsou shrnuty v tab. 1.1.

diatgram		
(a)	*act*	81 257
(b)	*pas*	3 743
(c)	*res1*	967
(d)	*res2.1*	55
	res2.2	28
(e)	*recip*	0
(f)	*disp*	9
(g)	*deagent*	1 973

Tab. 1.1 Hodnoty gramatému *diatgram* v PDT 3.0

Příklady

(1) Já.ACT **nemám** vše **domyšleno**.*res2.2*, nejsem si jist, jestli…
(2) Klub.ADDR **má** na letošní rok financování **zajištěno**.*res2.1* ze státního rozpočtu.
(3) Výrobci **nedostanou zaplaceno**.*recip* dříve než v březnu.
(4) **Hrálo se**.*disp* mi.ACT výborně, vůbec se mi nechtělo střídat.
(5) Doplatili na to, že **se potvrdil**.*deagent* jejich optimistický odhad inflace.
(6) Na bezpečnost práce **se** mnoho **nehledí**.*deagent*
(7) Od té doby **byl** černý trh tímto opiátem.ACT **přehlcován**.*pas*
(8) Z vyšší daňové sazby **je vyňato**.*res1* ubytování a stravování při dětských rekreacích a táborech.

1.4 Atribut *sentmod*

Magda Ševčíková

Popis

Atributem *sentmod* je zachycována větná modalita. Sada hodnot i princip přidělování tohoto atributu (na základě pozice uzlu ve stromě) zůstávají v PDT 3.0 stejné jako ve verzi 2.0 (v manuálu [5] kap. 4.7). Specifikace nových pozic, ve kterých je stanovení hodnoty atributu *sentmod* teoreticky adekvátní (zvl. kořeny koordinovaných klauzí a kořeny identifikačních struktur), vedla k zásadnímu rozšíření množiny uzlů s hodnotou tohoto atributu.

Atributem *sentmod* je zachycována větná modalita věty, tedy zda je věta uváděna jako tvrzení, otázka, rozkaz atd. V psaných textech je větná modalita vyjadřována kombinací formálních prostředků, konkrétně slovesným způsobem, koncovou interpunkcí, slovosledem a příp. modálními částicemi *ať, kéž, nechť*.

Atribut *sentmod* byl k dispozici už v datech PDT 2. Protože se jednalo o zjednodušenou anotaci (pro koordinační struktury byla specifikována jediná hodnota atributu *sentmod* a byly tak vědomě zanedbány případy koordinovaných klauzí s různou modalitou), bylo nutné anotaci zrevidovat a v datech PDT 3.0 ji implementovat nově.

Hodnoty atributu *sentmod* přidělované v datech PDT 3.0 jsou totožné s hodnotami použitými v PDT 2.0:

- ***enunc*** oznamovací modalita (tvrzení)
- ***excl*** zvolací modalita (zvolání)
- ***desid*** přací modalita (přání)
- ***imper*** rozkazovací modalita (rozkaz / žádost)
- ***inter*** tázací modalita (otázka)

Rovněž princip, že hodnoty atributu *sentmod* jsou uzlu přidělovány na základě jeho pozice v tektogramatickém stromě, zůstal při anotaci tohoto atributu v datech PDT 3.0 stejný jako v PDT 2.0.

To, v čem se anotace atributu *sentmod* v PDT 2.0 liší od anotace tohoto atributu v PDT 3.0, je množina uzlů, kterým byla hodnota tohoto atributu přidělena. V PDT 2.0 byla hodnota atributu *sentmod* vyplněna u uzlů uvedených v následujícím seznamu pod body (a) až (c). Hlavní motivací pro revidování atributu *sentmod* bylo odstranění výše uvedeného zjednodušení v koordinačních strukturách. Hned po zahájení revize ovšem byly jako další typ struktur, které mohou (obdobně jako přímá řeč, viz (b), a parenteze (c)) v rámci věty vyjadřovat „vlastní“ větnou modalitu odlišnou od modality věty, jejíž jsou součástí, identifikovány podstromy reprezentující identifikační struktury (s funktorem *ID*), viz (d).

Vzhledem k tomu, že rozhodnutí specifikovat v PDT 3.0 hodnotu atributu *sentmod* pro každou klauzi v koordinaci zasáhlo každou ze skupin uvedených v bodech (a) až (c) (a týká se rovněž skupiny (d)), i vzhledem ke skutečnosti, že během systematické revize dat PDT 2.0 prováděné během posledních několika let byly opraveny chyby nejrůznějších typů, hodnoty atributu *sentmod* uvedené v datech PDT 2.0 byly smazány a množina uzlů, kterým tato hodnota náleží po navržených změnách, byla v datech PDT 3.0 vymezena nově. Tato množina kandidátských uzlů byla vymezena v krocích (a) až (c) (použitých už při anotaci PDT 2.0) doplněných o kroky (d) a (e):

(a) přímí potomci technického kořene stromu, tj. uzly reprezentující řídící sloveso nebo substantivum nebo kořen koordinační struktury,

(b) kořeny podstromů reprezentujících přímou řeč (identifikovány na základě atributu *is_dsp_root*),

(c) kořeny podstromů reprezentujících (syntakticky nezávislou) parentezi, jejíž efektivní kořen je ohodnocen funktorem PAR,
(d) kořeny podstromů reprezentujících identifikační struktury (ohodnocené funktorem ID),
(e) ze všech těchto kandidátů byly vyděleny kořeny koordinačních struktur a zpracovávány jako zvláštní skupina.

Anotační procedura

Nekoordinovaným uzlům, které zůstaly po aplikaci kroku (e), byla hodnota atributu *sentmod* přidělena poloautomaticky v rámci následující procedury, při níž byly využívány odkazy mezi tektogramatickou, analytickou a morfologickou anotací:

(i) pokud uzel reprezentoval syntetickou imperativní formu slovesa (technicky: pokud jeden z morfologických tokenů, ke kterému vedly odkazy z daného tektogramatického uzlu, měl morfologickou značku začínající *Vi.** (imperativní tvar slovesa)), uzlu byla v atributu *sentmod* přidělena hodnota *imper*;
(ii) pokud syntaktická struktura, do které daný uzel náležel, končila otazníkem (technicky: pokud daný tektogramatický uzel odpovídal analytickému uzlu, mezi jehož dětmi byl otazník), byla u tohoto uzlu v atributu *sentmod* vyplněna hodnota *inter*;
(iii) ze zbývajících uzlů byly vybrány ty, které byly součástí věty uvozené některou z částic *ať, kéž* nebo *nechť* a/nebo zakončené vykřičníkem, a v atributu *sentmod* jim byla manuálně přidělena některá z hodnot *desid*, *excl* nebo *imper*;
(iv) u ostatních uzlů byla v atributu *sentmod* vyplněna hodnota *enunc*.

Koordinace byly zpracovávány jako homogenní skupina bez ohledu na to, ke kterému z bodů (a) až (d) patřily. Hodnota atributu *sentmod* byla v koordinacích přidělována kořenům koordinovaných klauzí, tyto uzly byly identifikovány na základě seznamu kořenů koordinačních struktur.

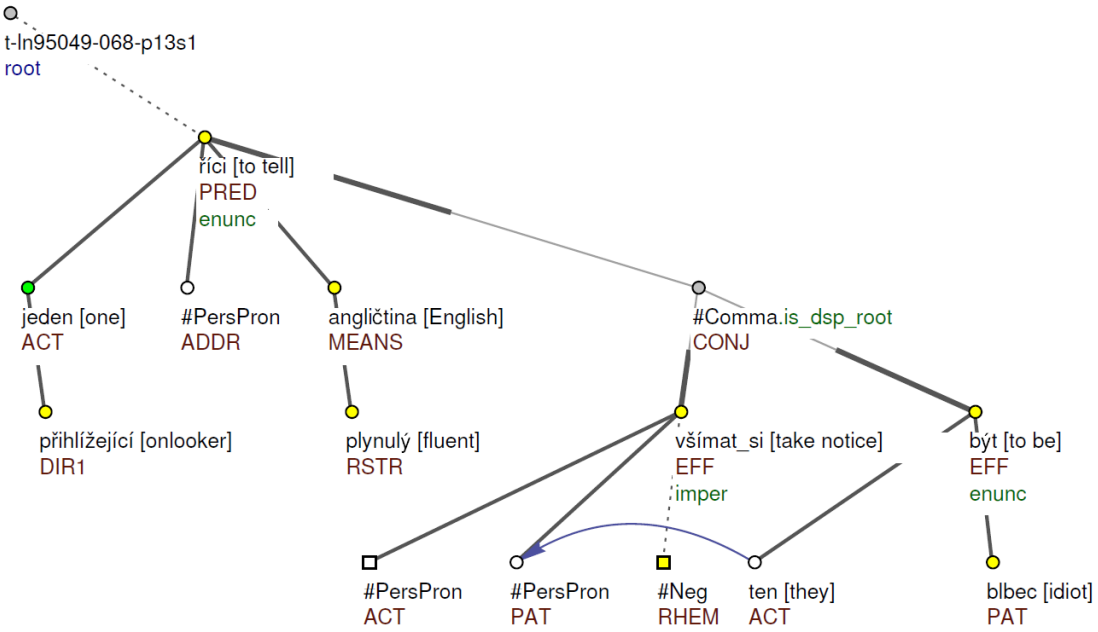
Při určování hodnoty atributu *sentmod* u kořene každé koordinované klauze bylo možné využít krok (i), ale pouze „lokálně“, tj. pouze pro danou klauzi, nikoli pro ostatní klauze s touto klauzí koordinované: hodnota *imper* byla vyplněna u kořene koordinované klauze, který reprezentoval imperativní slovesný tvar.

Klauzím koordinovaným s takovou imperativní klauzí byla hodnota atributu *sentmod* přidělena manuálně. Další skupinou uzlů pro manuální anotaci byly kořeny koordinovaných klauzí, které vystupovaly jako součást koordinačních struktur zakončených otazníkem. Předpoklad, že otazník použitý jako koncová interpunkce koordinační struktury může být interpretován jako signál větné modality pouze pro poslední koordinovanou klauzi (tedy že nemusí korespondovat s větnou modalitou koordinovaných klauzí vystupujících v koordinaci na jiném než posledním místě), se během anotace potvrdil. Třetí skupinu uzlů pro manuální anotaci tvořily kořeny koordinovaných klauzí, které byly součástí koordinačních struktur zakončených vykřičníkem a/nebo obsahovaly částice *ať, kéž* nebo *nechť*. Manuální anotaci prováděli paralelně dva anotátoři, a to s vysokou mezianotátorskou shodou (93,7 %; Kohenovo kappa 0,89).

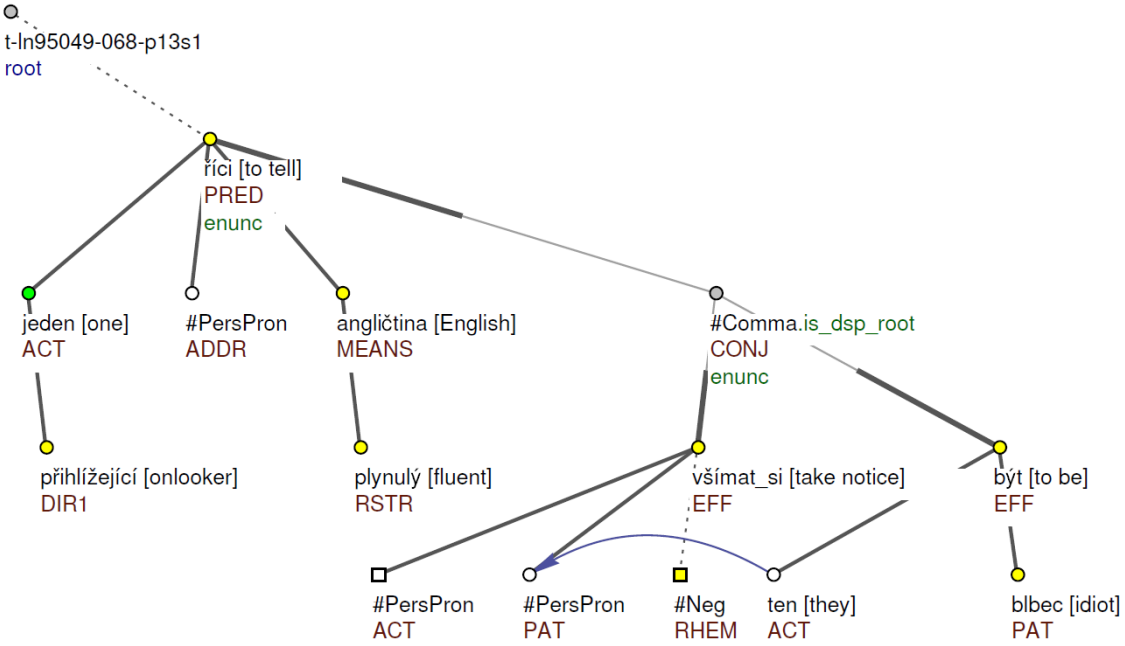
Všechny zbývající koordinační struktury byly zakončeny tečkou (popř. třemi tečkami apod. nebo byly bez koncové interpunkce) a skládaly se pouze z klauzí s indikativním nebo kondicionálovým slovesem. Ruční inspekcí 100 koordinačních struktur náhodně vybraných z této skupiny se ukázalo, že koordinované klauze mají vždy oznamovací modalitu, klauzím v těchto strukturách proto byla v atributu *sentmod* automaticky přidělena hodnota *enunc*.

Vizualizace

Výsledná anotace hodnot atributu *sentmod* v datech PDT 3.0 je porovnána s anotací v PDT 2.0 na obr. 1.1 a 1.2. Obr. 1.1 a obr. 1.2 představuje tektogramatický strom věty *„Nevšímejte si jich, jsou to blbci," řekl mi plynulou angličtinou jeden z přihlížejících.*, v níž jsou dvě klauze s různou větnou modalitou koordinovány v přímé řeči (kořen koordinační struktury má přidělen funktor CONJ a atribut *is_dsp_root*). V PDT 3.0 (obr. 1.1) byla hodnota atributu *sentmod* specifikována pro každou z těchto klauzí přímé řeči (hodnoty *imper* a *enunc*) a také pro řídící sloveso uvozovací klauze (hodnota *enunc*). V PDT 2.0 (obr. 1.2) byla uzlu s funktorem CONJ přidělena hodnota *enunc*, imperativní modalita první klauze přímé řeči nebyla v datech PDT 2.0 zachycena.



Obr. 1.1: *„Nevšímejte si jich, jsou to blbci," řekl mi plynulou angličtinou jeden z přihlížejících.* (PDT 3.0)



Obr. 1.2: *„Nevšímejte si jich, jsou to blbci," řekl mi plynulou angličtinou jeden z přihlížejících.* (PDT 2.0)

Příklady

Příklady (1) až (5) ilustrují po řadě jednotlivé hodnoty atributu *sentmod*. V příkladu (6) je uvedena věta skládající se z řídící (syntakticky nezávislé) a závislé klauze, které jsou z hlediska větné modality chápány jako jeden celek – hodnota atributu *sentmod* je přidělena slovesu řídící klauze. Oproti tomu koordinované (syntakticky nezávislé) klauze mohou každá vyjadřovat jinou větnou modalitu, př. (7). Obdobně přímá řeč, parenteze a identifikační struktura mohou vyjadřovat vlastní větnou modalitu odlišnou od modality věty, jejíž jsou součástí, srov. př. (8) až (10). V případě konfliktu mezi způsobem slovesného tvaru (syntetická imperativní forma v řídící klauzi v př. (11)) a koncovou interpunkcí (otazník; srov. kroky (i) a (ii) popisované výše) byla hodnota atributu *sentmod* zvolena podle koncové interpunkce (*inter* v př. (11)). Hodnota atributu *sentmod* je v příkladech připsána řídícímu slovesu příslušné klauze (vyznačeno tučně).

(1) Ekonomika **jde**.*enunc* do vzestupu už letos.
(2) Jaká **je**.*inter* nezaměstnanost v této zemi?
(3) **Podívej se**.*imper* na mě!
(4) Ať **si** provincie konečně **oddychne**.*desid*.
(5) To **nejsou**.*excl* špatně rozdané karty!
(6) **Neptejte se**.*imper* mě, proč jsem přijel do Prahy.
(7) Poprvé **jste nastoupil**.*enunc* v závěru zápasu v Benešově, jaké to **bylo**.*inter*?
(8) Kam **se poděla**.*inter* má bojovnost? **ptala**.*enunc* se sama sebe po utkání Martinezová.
(9) Pane kolego, **věřte**.*imper* **nevěřte**.*imper*, počítač **nelže**.*enunc*.
(10) Zítra **bude** u příležitosti III. výročí české a slovenské edice Playboy **otevřena**.*enunc* výstava **Pohlaďte si**.*imper* králíčka sestavená z ilustrací pro časopis Playboy.
(11) **Hádejte**.*inter*, kde se toto menu ve Windows najde?

Anotaci větné modality v PDT 3.0 se věnuje i článek [19] v seznamu referencí na konci tohoto dokumentu.

2 Zrušení lemmatu *#Benef*

Marie Mikulová a Jarmila Panevová

Popis

Zástupné t-lemma *#Benef* náleželo v datech PDT 2.0 nově vytvořenému uzlu (*is_generated=1*), který reprezentoval v povrchové podobě věty nevyjádřené volné doplnění s významem benefaktoru (*functor*=BEN). Uzel se doplňoval na pozici kontrolujícího členu (controller) v následujících typech konstrukcí s kontrolou, v nichž infinitiv má roli subjektu nebo atributu:

- konstrukce *být* + přísudkové substantivum, kde infinitiv má roli subjektu, např.: *Transformovat bezpečnostní složky je hračkou* [pro kohokoli], *Je nutností* [pro kohokoli] *pořídit vybavení.; Je radost* [pro kohokoli] *dostávat dárky.* (v manuálu [5] kap. 8.2.4.4.4.2, 8.2.4.4.4.3 a 8.2.4.4.4.5),
- konstrukce *být* + přísudkové adjektivum, kde infinitiv má roli subjektu, např.: *Je nutné* [pro kohokoli] *přijít.*; *Je trapné* [pro kohokoli] *přijít pozdě.* (v manuálu [5] kap. 8.2.4.4.4.4 a 8.2.4.4.4.5),
- konstrukce *být* + predikativní adverbium, např.: *Je škoda* [pro kohokoli] *se ochudit o tolik vzácných látek.* (v manuálu [5] kap. 8.2.4.4.4.6),
- infinitiv závisí na predikátu *lze*, např.: *Lze* [pro kohokoli] *tam přijít kdykoli.* (v manuálu [5] kap. 8.2.4.4.5),
- kontrola u typu *Je vidět Sněžku.* (v manuálu [5] kap. 8.2.4.4.5),
- konstrukce odvozené od výše uvedených (v manuálu [5] kap. 8.2.4.5.1 a 8.2.4.7.1).

V datech PDT 3.0 byl uzel s t-lemmatem *#Benef* nahrazen:

- uzlem s t-lemmatem *#Gen* (*functor*=BEN; *is_generated=1*) v případě všeobecného benefaktoru.
  Např.: *Je dobré chodit brzo spát*. = pro kohokoliv je to dobré
- uzlem s t-lemmatem *#PersPron* (*functor*=BEN; *is_generated=1*) v případě aktuální elipsy. V těchto případech byla také vyznačena příslušná textová koreference a uzlu byly vyplněny příslušné gramatémy.
  Např.: *Pavel přišel včera pozdě. Bylo by dobré jít dnes brzo spát.* = pro Pavla by bylo dobré jít brzo spát

Motivace změny

Empirickým výzkumem bylo zjištěno, že mezi konstrukce s kontrolou patří i takové infinitivní konstrukce, v nichž jejich subjekt je kontrolován volným určením prospěchu (BEN, benefaktor); toto určení je buď explicitně vyjádřeno (*Povinnost starat se o zámek plyne pro majitele ze zákona.),* případně se jedná o aktuální elipsu (*Čím větší odchylka, tím víc čeká firmu práce navíc, protože je třeba* [pro firmu.BEN] *výpadek kompenzovat jiným zbožím.*) nebo je zevšeobecněno (*Je dobré chodit brzo spát.)*. Prozatímní řešení aplikované v PDT 2.0 a 2.5, kde se doplňovalo umělé t-lemma *#Benef*, bylo zrušeno, a anotace nabyla obdobného tvaru, jako u jiných konstrukcí s kontrolorem vyjádřeným nebo t-lemmatem pro všeobecnost (*#Gen*) s funktorem BEN.

Související literaturu viz publikace [20], [21], [22] a [23] v seznamu referencí na konci tohoto dokumentu.

Anotační procedura

V datech PDT 2.0 bylo 1 394 uzlů s t-lemmatem *#Benef* (*functor*=BEN; *is_generated=1*). 100 výskytů bylo nahrazeno ručně. Zbývající 1 294 uzlů bylo automaticky převedeno na uzel s t-lemmatem *#Gen* (*functor*=BEN; *is_generated=1*). K tomuto kroku bylo přistoupeno z toho důvodu, že nedokončená anotace valence substantiv a adjektiv (viz k tomu manuál [5] kap. 5.2.4) neumožňuje v mnoha případech anotovat konstrukce s aktuální elipsou benefaktoru správně (není kam vést koreferenční šipku od doplněného uzlu s t-lemmatem *#PersPron*).

Příklady

Příklady s všeobecným benefaktorem v pozici kontrolujícího členu

(1) Je-li vypovídání smluv legální, je nutné **[*#Gen*.BEN]** novelizovat zákony. (Viz obr. 2.1)

(2) Česká republika, která je toho času nestálým členem Rady bezpečnosti, má možnost zaujmout ke vzniklé realitě jednoznačné stanovisko, neboť je třeba **[*#Gen*.BEN]** podívat se pravdě do očí.

Příklady s aktuální elipsou benefaktoru v pozici kontrolujícího členu

(1) Rady **dikům**

Znovu je tady čas, kdy je třeba **[*#PersPron*.BEN]** se rozhodnout.

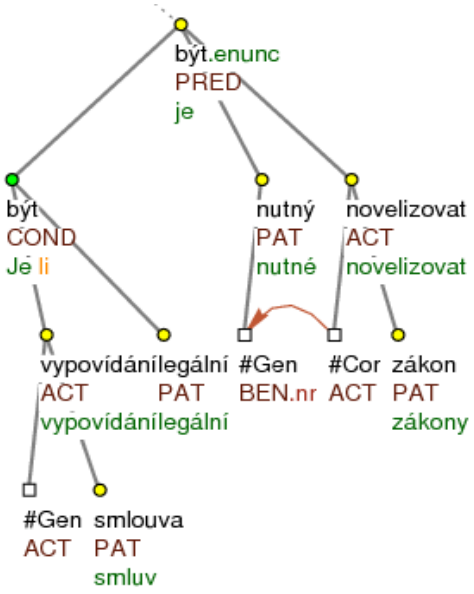
Na majitele kuponových knížek dotírají otázky - kam vložit své body, jaký obor si vybrat, raději investovat do velkého podniku, nebo do neznámého podničku?

Podobných otázek, na něž samotní dikové, bez patřičných informací jen těžko hledají odpověď, je daleko víc.

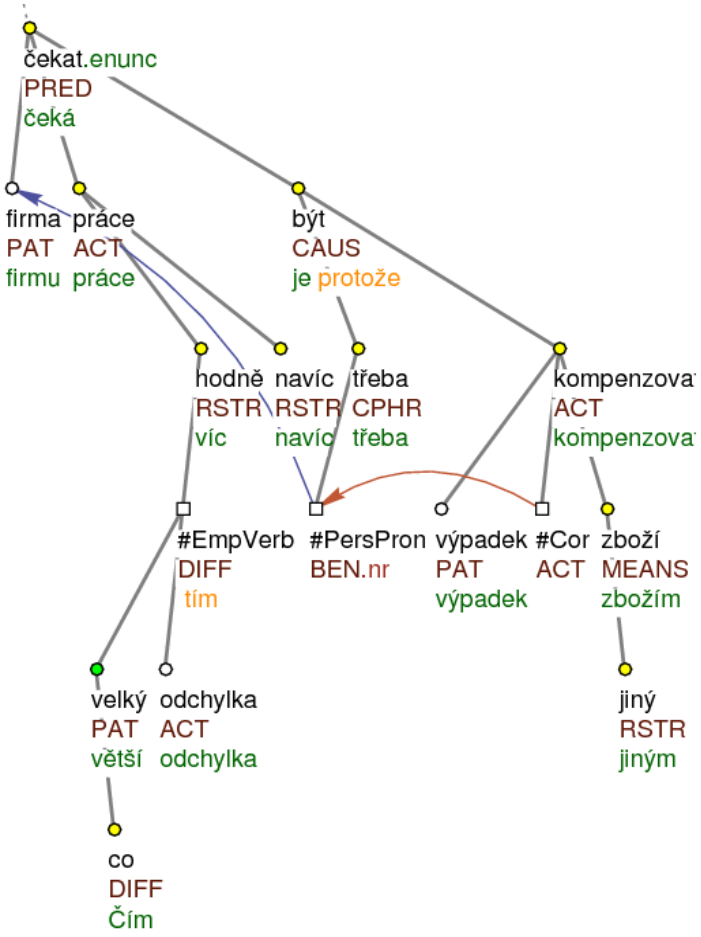
(2) Čím větší odchylka, tím víc čeká **firmu** práce navíc, protože je třeba **[*#PersPron*.BEN]** výpadek kompenzovat jiným zbožím. (Viz obr. 2.2)

(3) **Hráč** musí sám vědět, co to znamená **[*#PersPron*.BEN]** být profesionálem. (Viz obr. 2.3)

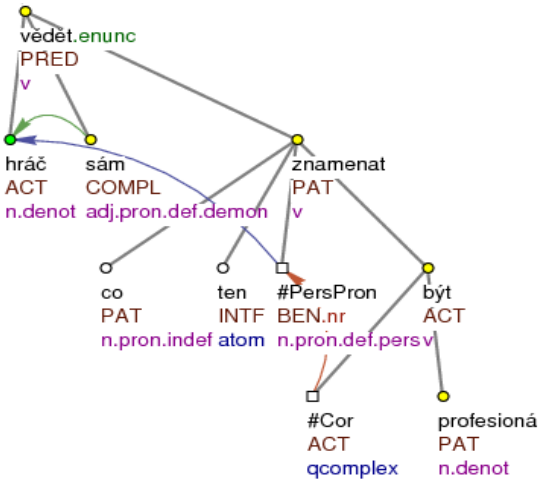
Vizualizace



Obr. 2.1: *Je-li vypovídání smluv legální, je nutné novelizovat zákony.*



Obr. 2.2.: *Čím větší odchylka, tím víc čeká firmu práce navíc, protože je třeba výpadek kompenzovat jiným zbožím.*



Obr. 2.3: *Hráč musí sám vědět, co to znamená být profesionálem.*

3 Koreference a asociační anafora

Anna Nedoluzhko

Popis

Tektogramatická rovina PDT 2.0 zahrnuje ruční anotaci koreferenčních vztahů dvou druhů: gramatické koreference (kdy je možné určit koreferující členy na základě gramatických pravidel; v manuálu [5] kap. 8.2) a textové koreference (kde odkazování není realizováno pouze gramatickými prostředky, ale plyne nejčastěji z kontextu; v [5] kap. 8.3). Textová koreference byla anotovaná u osobních a přivlastňovacích zájmen pro 3. osobu, u ukazovacích zájmen *ten, ta, to* a při aktuální elipse zájmena/jména.

V PDT 3.0 byla anotace tohoto jevu rozšířena následujícím způsobem:

- anotace textové koreference byla rozšířena o anotaci dalších typů výrazů (viz 3.1);
- byla provedena ruční anotace některých typů asociační anafory (viz 3.2);
- vztahy koreference a asociační anafory byly anotovány rovněž pro zájmena v první a druhé osobě (viz 3.3).

Při anotaci koreference a asociační anafory platí princip maximální velikosti koreferujícího výrazu. To znamená, že za koreferenční výraz se považuje celý podstrom uzlu propojeného koreferenčním vztahem. Technicky však koreferenční šipka vede od řídícího uzlu koreferenčního výrazu (resp. na něj).

Při anotaci koreferenčních vztahů se dodržuje zásada udržovat jednoduchou linearitu koreferenčního řetězce. Pro asociační anaforu tento princip neplatí.

Z teoretického hlediska jsou vztahy koreference a asociační anafory součástí jiné roviny popisu jazyka, tzv. roviny „nadvětné“, textové, jež popisuje lingvistické jevy z perspektivy struktury textu a textové koherence. Technicky však anotace probíhala v rámci tektogramatické reprezentace. Tento přístup umožnil použití syntaktických jevů, které již byly anotovány dříve, jako např. funktory, typy uzlů, gramatémy apod. Výhodou je rovněž možnost využití syntaktické analýzy jako takové, např. elegantního řešení eliptických struktur, parentezí, predikativních vztahů, apozic apod.

Související literaturu viz publikace [24], [25], [26], [27] a [28] v seznamu referencí na konci tohoto dokumentu.

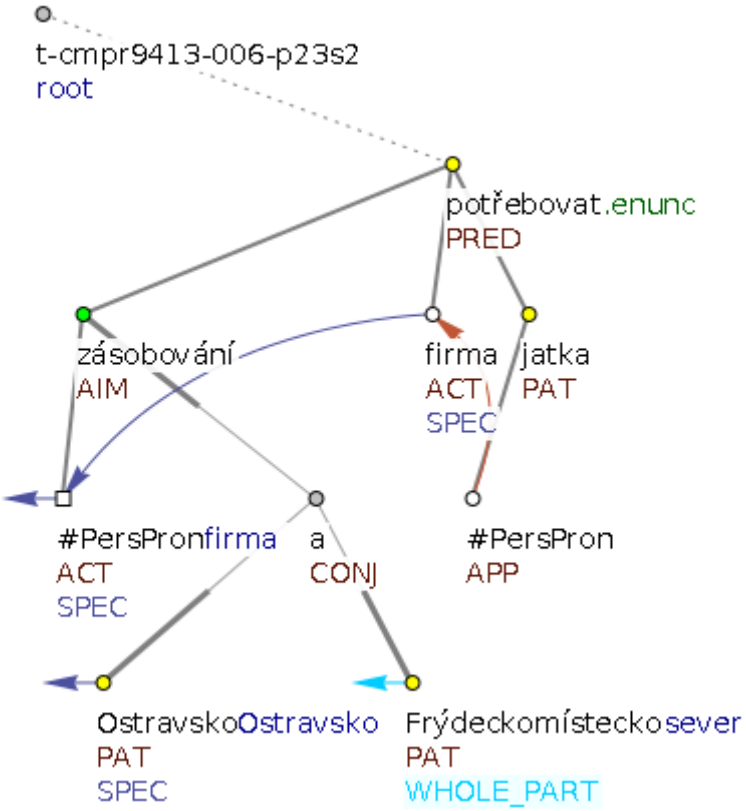
Anotační procedura

Anotace rozšířené textové koreference a asociační anafory sestávala z následujících kroků:

- automatická předanotace (např. propojení koreferencí pojmenovaných entit)
- aplikace některých užitečných pomůcek, které pomáhaly anotátorům hledat správné antecedenty (zvýrazňování uzlů ve stromě, které jsou již propojeny koreferencí nebo asociační anaforou, podtržení výrazů se stejným lemmatem apod.)
- ruční anotace
- automatická kontrola některých aspektů koreferenčních vztahů (hledání nejbližšího antecedentu, udržování koreferenčního řetězce, napojování vztahů asociační anafory na koreferenční řetězce apod.)

Vizualizace

Obr. 3.1 znázorňuje základní rysy anotace koreference a asociační anafory. Vztahy koreference a asociační anafory jsou značeny mezi podstromy tektogramatického stromu pomocí šipek různých barev (tmavě červená šipka pro gramatickou koreferenci, tmavě modrá šipka pro textovou koreferenci a bledě modrá pro asociační anaforu). Šipky vedou od anaforického výrazu k antecedentu. Pokud je antecedent v jedné z předcházejících vět, jeho lemma je napsáno tmavě modrou barvou vedle lemmatu anaforického členu.



Obr. 3.1: *Pro zásobování Ostravska a Frýdeckomístecka potřebuje firma svá jatka.*

3.1 Rozšířená textová koreference

Anna Nedoluzhko

Popis

V PDT 3.0 byly koreferenční vztahy anotovány pro:

- jmenné fráze,
- adjektiva odvozená od pojmenovaných entit (*pražský, český* apod.),
- zájmenná příslovce (*tak, tam, tehdy* apod.), které mají explicitní antecedenty v předchozím (vzácně také následujícím) kontextu.

Anotace textové koreference se nyní skládá z pronominální koreference a koreference elips převzaté z PDT 2.0 a z nově zpracované rozšířené textové koreference. Technicky je tento vztah zaznamenán ve strukturovaném atributu *coref_text* u počátečního uzlu vztahu, který se skládá z identifikátoru antecedentu a typu vztahu.

V PDT 3.0 jsou anotovány následující typy textové koreference:

- ***SPEC*** koreference jmenných frází se specifickou referencí; např: *Německo – země*; viz i př. (1).
- ***GEN*** koreference jmenných frází s generickou referencí (př. (2)).

U každého textově koreferenčního vztahu je zaznamenáno, zda jde o koreferenci specifických nebo generických výrazů. Osobní, přivlastňovací, ukazovací zájmena a rekonstruované uzly s t_lemmatem *#PersPron* mají přednastavený typ SPEC. V případě nutnosti byly tyto hodnoty ručně opraveny. V případě nejednoznačnosti výběru je preferován typ SPEC.

Anotace rozšířené textové koreference je zaměřena na zachycení koreference, čili vztahů identity referentů, nikoli na anaforický vztah.

Textová koreference byla anotována na vzdálenost maximálně 20 vět. Anotace na větší textovou vzdálenost byla umožněna pouze v případech automatické předanotace koreference u pojmenovaných entit. Z jednoho uzlu může vycházet (resp. k němu může ukazovat) maximálně jedna koreferenční šipka.

Kromě již zmíněného byly v PDT 2.0 anotovány dva specifické případy textové (ko)reference: exoforické odkazování (*coref_special*, typ ***exoph***) a odkazování k segmentu, k úseku textu většímu než jedna věta (*coref_special*, typ ***segm***). V PDT 3.0 byla anotace těchto jevů doplněna o případy, kdy referujícími výrazy nebyly ani zájmena, ani elipsy:

- Exoforické odkazování (typ *exoph*) je anotováno v případě časové a místní deixe a deixe zájmenných příslovcí; př. (3).
- Reference k většímu úseku textu (discourse deixis) je značena v případě, kdy jmenná fráze buď odkazuje k textovému úseku většímu než jedna věta, nebo k úseku věty, který technicky není možné vyčlenit.

Příklady

(1) Jeho dojetí znásobila při vyhlašování přítomnost pořadatelů **soutěže** – Českého manažerského centra v Čelákovicích. . Na letošním ročníku **soutěže.***SPEC* se spolupodílí i Profit.

(2) **Droga** je tak účinná, že ten, kdo **ji.***GEN* užívá, se snadno dostane do „pohody“ kouřením nebo šňupáním.

(3) Dokončeny by měly být do 31. prosince 1995, a to i přes jisté zdržení způsobené opožděným stěhováním nájemníků z domů čp. 8 a 518 do náhradních bytů na sídlišti Barrandov **v těchto dnech.***exoph*

3.2 Asociační anafora

Anna Nedoluzhko

Popis

Kromě rozšířené textové koreference jsou v PDT 3.0 anotovány nekoreferenční asociační vztahy v rámci tzv. asociační anafory (bridging relations). Asociační anafora v našem pojetí zahrnuje několik typů lexikálně-sémantických a pragmatických vztahů přispívajících ke koherenci textu. Technicky jsou tyto vztahy zaznamenány ve strukturovaném atributu *bridging* u počátečního uzlu vztahu, který se skládá z identifikátoru antecedentu a typu vztahu.

V rámci asociační anafory se v PDT 3.0 anotují následující typy vztahů:

- ***PART_WHOLE*** a ***WHOLE_PART*** metonymický vztah části a celku; např.: *pokoj - strop, Německo - Bavorsko - Mnichov*;
- ***SET_SUB*** a ***SUB_SET*** vztah mezi množinou a podmnožinou/prvkem množiny; např.: *studenti – několik studentů – student;*
- ***P_FUNCT*** a ***FUNCT_P*** vztah mezi objektem a definovanou na něm unikátní funkcí; např.: *trenér – mužstvo; premiér – vláda; firma – ředitel; akce – organizátor;*
- ***CONTRAST*** vztah sémantického a pragmatického kontrastu; př. (1);
- ***ANAF*** nekoreferenční anaforický vztah v případě explicitního anaforického odkazování (pomocí identifikátorů) na nekoreferenční antecedent; př. (2);
- ***REST*** blíže neupřesněná kategorie , do které jsou zahrnuty vztahy rodinné příslušnosti (*otec – syn*), místo – obyvatel (*Praha – Pražáci*), autor – dílo (př.(3)), věc – majitel, vztah mezi stejně vyjádřenými nebo synonymními nekoreferenčními výrazy, událost – argument (*podnikání – podnikatel*) a objekt – velmi typický instrument (*provazochodec – lano*).

Vztahy typu *PART, SUBSET* a *FUNCT* jsou dále členěny podle lineárního pořadí antecedentu a koreferujícího členu v textu, např. vztah *PART_WHOLE* se používá, pokud antecedent referuje k části a anaforický výraz k celku, zatímco vztah *WHOLE_PART* se používá pro obrácené pořadí.

Odkaz k více antecedentům, který byl v PDT 2.0 povolen jako specifický případ textové koreference, je v PDT 3.0 anotován jako asociační anafora typu *SUB_SET*. Srov. např. vztah asociační anafory, který směřuje od zájmena *ně* k antecedentům *Marie* a *Vlasta* ve větě ***Marie*** *vzala* ***Vlastu*** *do divadla, kde* ***na ně*** *čekal Marek.*

Příklady

(1) Dnes, po rozdělení ČSFR, je jasné, že **osud ČR** bude stále více spojený s Německem a přes něj s Evropskou unií a **osud Slovenska**.*CONTRAST* s Ruskem.

(2) A přesvědčen jsem ještě o jednom – je třeba mít **vysoké cíle** a s **malými [cíli**.*ANAF***]** se nespokojit.

(3) Při výběru **obrazu** bude hrát určitě velkou roli **autor**.*REST*

3.3 Koreference a asociační anafora u zájmen v 1. a 2. osobě

Anna Nedoluzhko

Popis

Dodatečně byla provedena anotace koreference a asociační anafory pro osobní a přivlastňovací zájmena v první a druhé osobě. Při anotaci byla použita pravidla pro anotaci rozšířené textové koreference a asociační anafory.

Všechny výskyty koreference u zájmen první a druhé osoby se anotovaly nezávisle na tom, jestli vystupují v anaforické funkci či nikoliv (př. (1)).

Zájmena první a druhé osoby mají často generickou interpretaci. Koreference v takových případech může být problematická, což bývá příčinou nízké mezianotátorské shody. Generická užití zájmen první a druhé osoby jsou typická pro firmy, země, sportovní týmy, zájmové skupiny apod. V jasných případech se anotuje jejich koreference s nezájmennými antecedenty včetně antecedentů nulových (př. (2)).

Kataforické užití zájmen první a druhé osoby se anotuje jako exoforická reference (*coref_special*, typ *exoph*). Jmenné fráze, které se vyskytují dále v textu, jsou-li s nimi koreferenční, na ně anaforicky odkazují (př. (3)).

Příklady

(1) Potřebu dalších investic **[*#PersPron*.ACT]** odhaduji do roku dva tisíce na více jak dvě miliardy korun, říká **ředitel Nováček**.

(2) **Slévárně Škoda** v Českých Budějovicích dluží plzeňská Škoda 61 miliónů Kčs. **[*#PersPron*.ACT]** Potřebujeme je hned a na stůl. Situace je vážná a z **naší** strany téměř neřešitelná. Bez finančních prostředků se už **[*#PersPron*.ACT]** neobejdeme," řekl včera Milan Fučík.

(3) Ačkoliv **naše** produkty se běžně prodávají v různých evropských zemích, **Česká republika** ještě není plnoprávným partnerem na evropském trhu.

4 Mezivýpovědní významové vztahy

Lucie Poláková

Popis

Anotace mezivýpovědních významových vztahů (textových vztahů, „discourse relations“) v PDT je inspirována filadelfským projektem anotace textových vztahů Penn Discourse TreeBank 2.0 [32], zčásti také využívá scénáře tektogramatické reprezentace PDT. Česká data s touto anotací byla poprvé vydána v rámci korpusu Prague Discourse Treebank 1.0 (PDiT 1.0; [6], [31]). PDT 3.0 zahrnuje tuto anotaci s drobnými opravami a obohacenou o několik nových textových jevů: anotace žánrů korpusových textů, anotace některých aktualizačních částic (tzv. rematizátorů) jakožto textových konektorů, anotace druhých vztahů (textové vztahy s více než jedním sémantickým typem) a zavedení nového atributu *discourse_special*.

Textové konektory, textové jednotky

Anotace mezivýpovědních významových vztahů (diskurzu) v PDT 3.0 je zaměřena na analýzu konektivních prostředků (textových konektorů, „discourse connectives“) a významových vztahů jimi vyjádřených. Jako nejmenší jednotka vstupující do textového významového vztahu je chápána výpověď obsahující finitní sloveso (finitní klauze). Textový konektor je definován jako predikát binární relace – otvírá dvě pozice pro dva textové úseky (argumenty), které spojuje a zároveň signalizuje typ významového vztahu mezi nimi. Konektory jsou neohebné (s výjimkou konektivního prostředku *což*) a neúčastní se větné stavby (nejsou větnými členy). Jsou reprezentovány souřadicími (*a, ale, kdežto*) a podřadicími spojkami (*protože, když, zatímco*), některými částicemi (*jen, také*), některými příslovci (*poté, dále*) a okrajově též některými dalšími slovními druhy – zejména v ustálených konstrukcích jako *jinými slovy* či *na druhé straně*. V PDT 3.0, stejně jako v PDiT 1.0, je anotace textových vztahů zaměřena pouze na vztahy vyjádřené povrchově přítomnými (tzv. explicitními) textovými konektory – vztahy bez takového konektoru nebyly v této fázi projektu značeny.

Kromě textových vztahů signalizovaných konektory obsahuje textová anotace v PDT 3.0 také značení seznamových struktur, nadpisů, popisů obrázků, tabulek a grafů, nekoherentních textů jako např. souborů krátkých zpráv apod.

Anotace rematizátorů jakožto textových konektorů

Rematizátory (výrazy s tektogramatickým funktorem RHEM) jsou v celé anotaci za konektory mezivýpovědních významových vztahů považovány pouze v případě, že jako konektory fungují – tj. v daném kontextu otevírají dvě pozice, které jsou obsazeny částmi textu obsahujícími každá alespoň jeden určitý slovesný tvar. Ukázka takového kontextu je v př. (1) a (2), rematizátory fungující jako konektory jsou vyznačeny tučně.

V případě, že dochází ke spojení pouhých jmenných frází nebo jsou v kontextu v obou částech významově podobná/stejná slovesa (viz př. (3) a (4)), se rematizátor za konektor nepovažuje, a to ani v případě, že stojí v iniciální pozici, jako je tomu v př. (5).

Oproti PDiT 1.0 byly v nové fázi do anotace zahrnuty také RHEM, které tvoří konektor společně s konektory konjunktivního vztahu v souvětí. Příklad takového konektoru je v př. (6).

Anotace druhých vztahů

Vztahy s více než jedním sémantickým typem jsou nyní nově anotovány oběma typy – ve dvou nezávislých vztazích reprezentovaných dvěma atributy discourse; každý z nich označuje příslušný konektor a sémantický typ. (To znamená, že mezi dvěma uzly reprezentujícími argumenty vztahu vedou dvě šipky, každá s jiným sémantickým typem a konektorem.)

Nový atribut *discourse_special*

Nově zavedený atribut *discourse_special* zachycuje tři speciální role fráze reprezentované uzlem a jeho podstromem; má tři možné hodnoty: *heading* (nadpis či podnadpis, nahrazuje atribut *is_heading* z PDiT), *metatext* (text nepatřící do originálního novinového textu, vzniklý při korpusovém zpracování) a *caption* (popisek obrázku, grafu atd.).

Příklady

(1) Děti se s některými záležitostmi nechtějí svěřit rodičům, i když žijí v normálně fungující rodině. […] Dnes mají **také** mnozí rodiče méně času na své ratolesti než dřív.

(2) Povinností budoucího nájemce tohoto areálu o rozloze 103 tisíc metrů čtverečních bude mj. péče o všechny nemovitosti včetně jejich údržby a oprav. Nájemce bude **také** muset vyřešit podmínky parkování pro návštěvníky tržnice a splnit podmínky Pražského ústavu památkové péče při úpravách objektů vzhledem k tomu, že jde o kulturní památku.

(3) Podle Mandíkových slov lze komerčně využít zhruba deset hektarů pozemků v železniční stanici Praha- Žižkov. Využít lze **také** prostory stanice Praha- Smíchov.

(4) Vyrábějí se zde především tresti do lihovin, limonád, sirupů a pečiva. Firma **také** produkuje cukrářské pasty.

(5) V okolí Brna a Kyjova se hojně vyskytují muchomůrky zelené. **Také** v Hostivaři a v dalších pražských lesoparcích byl nyní výskyt této houby zaznamenán.

(6) Taková odměna může mít skutečně silný motivační účinek pro účastníky **a** může být **také** užitečným přínosem pro firmu, která náklady plně hradí.

Anotační procedura

Oproti podobně zaměřeným korpusovým projektům (např. výše zmíněný Penn Discourse Treebank 2.0 [32]) byly textové jevy v českém projektu anotovány přímo na stromových strukturách tektogramatické reprezentace. Tento přístup umožnuje využít v anotaci mezivýpovědních významových vztahů také určité relevantní informace zachycené již v dřívější analýze věty (například vztahy vnitrovětné realizované hypotakticky) a těžit ze samotné syntaktické struktury (reprezentace eliptických struktur, vsuvek, apozic atd.). Z hlediska vyhledávání v datech a vizualizace je tak uživateli umožněno získat a zobrazit různé typy lingvistických informací najednou, od morfologie až po textové jevy.

Vzhledem k tomu, že řada vnitrovětných mezivýpovědních vztahů byla dostatečně zachycena již v rámci anotace tektogramatiky, probíhala anotace pražských dat ve dvou fázích. V první fázi byly ručně označeny všechny vztahy mezivětné, a dále takové vnitrovětné, které bylo třeba z hlediska textu interpretovat jinak, než jak byly interpretovány na tektogramatické rovině. V druhé fázi pak bylo možno automatickou procedurou extrahovat zbylé vztahy vnitrovětné (přítomnost vztahu, rozsah argumentů, konektory) a přiřadit jim textovou značku odpovídající jejich značce tektogramatické (funktoru). V obou částech anotace byla provedena podrobná kontrola konzistence anotovaných dat (více [29]).

Tab. 4.1 ukazuje distribuce typů mezivýpovědních významových vztahů v PDT 3.0 a počty jejich vnitro- a mezivětných realizací.

Anotátoři v ruční fázi anotací postupovali od analýzy prostého textu (nalézání a označování možných konektorů v prostém textu) ke značení mezivýpovědních vztahů přímo na tektogramatických stromech. Syntaktické stromové struktury, které daný konektor spojuje, jsou v datové reprezentaci propojeny silnou oranžovou šipkou, vztah mezi nimi je sémanticky interpretován a daný konektor je připojen k šipce (viz obr. 4.1).

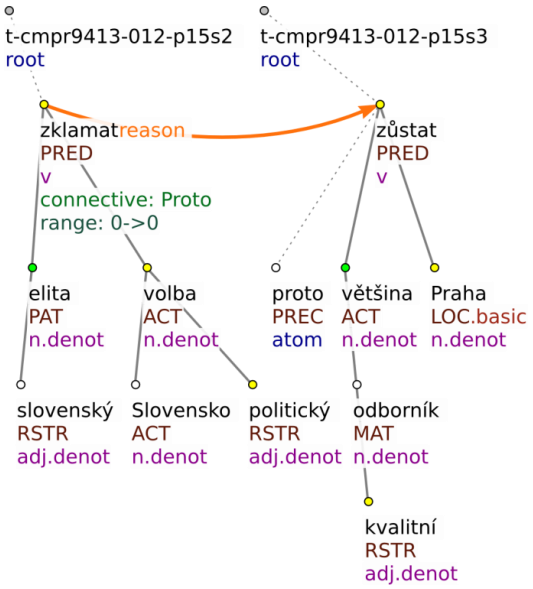
Bližší informace k procesu anotací jsou zpracovány v anotačním manuálu [30]).

Mezivýpovědní významové vztahy v PDT 3.0 – distribuce				
Typ vztahu	**Zkratka**	**vnitrovětné**	**mezivětné**	**Celkem**
concession – textová přípustka	***conc***	617	263	880
condition – textová podmínka	***cond***	1 350	19	1 369
confrontation – konfrontace	***confr***	345	308	653
conjunction – konjunkce	***conj***	6 109	1 389	7 498
conjuctive alternative – konjunktivní alternativa	***conjalt***	69	21	90
correction – rektifikace	***corr***	322	123	445
disjunctive alternative – disjunktivní alternativa	***disjalt***	257	15	272
equivalence – ekvivalence	***equiv***	41	64	105
exemplification – exemplifikace	***exempl***	28	120	148
explication – explikace	***explicat***	100	130	230
pragmatic condition – nepravá podmínka	***f_cond***	15	1	16
pragmatic contrast – nepravý kontrast	***f_opp***	23	27	50
pragmatic reason+result – nepravá příčina+následek	***f_reason***	12	28	40
generalization – generalizace	***gener***	9	97	106
gradation – gradace	***grad***	241	204	445
opposition – opozice	***opp***	1 396	1 800	3 196
other – jiné	***other***	1	1	2
precedence+succession – předčasnost+následnost	***preced***	591	249	840
purpuse – účel	***purp***	413	1	414
reason+result – příčina+následek	***reason***	1 601	1 031	2 632
restrictive opposition – restriktivní opozice	***restr***	97	172	269
specification – specifikace	***spec***	519	111	630
synchrony – současnost	***synchr***	174	52	226
Celkem		**14 330**	**6 226**	**20 556**
			Celkem seznamů:	83

Tab. 4.1: Distribuce mezivýpovědních významových vztahů v PDT 3.0

Vizualizace

Obr. 4.1 ukazuje anotaci vztahu mezi syntaktickými strukturami uvedenými výše v př. (1). Šipka nese sémantickou značku *reason* – jedná se o vztah příčiny a následku, je s ní asociován konektor *proto*. Zároveň je také označen rozsah obou textových úseků, které daný vztah spojuje (*range: 0 > 0*) – v tomto případě jsou součástí vztahu pouze oba uvedené stromy.



Obr. 4.1: *Slovenská elita byla zklamána politickou volbou Slovenska. Proto většina kvalitních odborníků zůstala v Praze.*

Opravy oproti PDiT 1.0

Korpus PDT 3.0 zahrnuje oproti PDiT následující typy oprav:

- opravy původních šipek v místech, kde byla manuální (dobře) i automatická (špatně, navíc) šipka.
- odstranění zbytečných skupin (zapomenutých kvůli dvěma chybám v odstraňovacím skriptu)
- oprava hodnot atributu start/target_range a start/target_group_id v některých případech odstraněné skupiny
- změna směru automatických vnitrovětných šipek odvozených z funktoru CSQ
- řada oprav jednotlivých šipek

Seznam anotačních atributů pro textovou anotaci v PDT 3.0

Mezivýpovědní významové vztahy jsou technicky zaznamenány především ve strukturovaném atributu *discourse* u počátečního uzlu vztahu, pro zaznamenání doplňkových informací slouží atributy *discourse_groups* a *discourse_special*.

- ***discourse/target_node.rf*** – *id* cílového uzlu, nebo nedefinováno v případě, že není cílový uzel (např. chybějící hypertéma u seznamu)
- ***discourse/type*** – druh šipky; dvě možné hodnoty: *discourse* (mezivýpovědní významový vztah), *list* (položka seznamu)
- ***discourse/start_range*** – rozsah začátku šipky; možné hodnoty: *n*, kdy *n* (celé nezáporné číslo) = počet stromů vpravo od aktuálního náležejících k argumentu (navíc k danému uzlu a jeho podstromu; *0* značí argument sestávající pouze z daného uzlu a jeho podstromu), *group* (libovolná množina uzlů; viz níže atributy *discourse/start_group_id* a *discourse_groups*), *forward* (daný uzel se svým podstromem a blíže nespecifikované

množství následujících stromů), *backward* (daný uzel se svým podstromem a blíže nespecifikované množství předchozích stromů)

- ***discourse/target_range*** – rozsah cíle šipky; možné hodnoty jako u *start_range*
- ***discourse/start_group_id*** – identifikátor skupiny uzlů (celé kladné číslo), pokud rozsah začátku šipky je nastaven na *group*; jednotlivé uzly náležející do skupiny mají identifikátor skupiny uveden v atributu *discourse_groups*.
- ***discourse/target_group_id*** – identifikátor skupiny uzlů (celé kladné číslo), pokud rozsah konce šipky je nastaven na *group*; jednotlivé uzly náležející do skupiny mají identifikátor skupiny uveden v atributu *discourse_groups*.
- ***discourse/discourse_type*** – typ významového vztahu, například *cond* (textová podmínka); možné hodnoty jsou uvedeny ve sloupci *zkratka* v tab. 4.1.
- ***discourse/t-connectors.rf*** – seznam *id* uzlů tektogramatické roviny, které tvoří konektor
- ***discourse/a-connectors.rf*** – seznam *id* uzlů analytické roviny, které tvoří konektor
- ***discourse_groups*** – seznam identifikátorů skupin, ke kterým daný uzel patří
- ***discourse_special*** – tři možné hodnoty pro tři speciální role fráze reprezentované uzlem a jeho podstromem: *heading*, *metatext* a *caption*

5 Žánrová specifikace textů

Lucie Poláková

Popis

Texty Pražského závislostního korpusu pocházejí ze dvou velkých českých deníků, Mladá Fronta a Lidové noviny, dále z ekonomického týdeníku Českomoravský Profit a z populárně vědeckého měsíčníku Vesmír. Jde tedy o korpus žurnalistického stylu. Přesto byla v průběhu různých anotačních projektů zjištěna poměrně velká diverzita dat v tomto ohledu – korpus obsahuje textové útvary od televizních programů po kulturní recenze a také určité množství nekoherentních textů jako jsou soubory krátkých zpráv a podobně.

Ruční klasifikace žánrů textů v PDT (tj. 3 165 dokumentů), která je nově zahrnuta ve vydání PDT 3.0, má za cíl sloužit zejména následujícím účelům:

- vyloučení krátkých a nesouvislých textů z trénovacích sad pro modelování jakéhokoliv typu textové koherence,
- efektivnější shlukování podobných typů textů (nebo způsobů výstavby textu) pro nejrůznější experimenty ve zpracovávání přirozených jazyků (NLP), zejména pro takové, které pracují s větami a většími celky (rozlišení anafory, tematické struktury a salience, zpracování textu, analýza sentimentu atd.),
- získání zlatých dat pro automatickou klasifikaci textů/žánrů.

Žánr dokumentu je zachycen v atributu *genre* , který náleží celému dokumentu. Možné hodnoty jsou uvedeny ve sloupci „Zkratka“ v tab. 5.1.

Anotační procedura

Za pomoci předchozích zkušeností anotátorů s texty PDT byla vytvořena taxonomie dvaceti žánrových kategorií ve třech hlavních třídách: monologické žánry, dialogické žánry a ostatní, okrajové žánry (viz tabulka). Pro maximální jednoduchost anotace je tato taxonomie jednoúrovňová, tj. anotátoři museli zvolit jednu z dvaceti kategorií, ne pouze jejich nadřazenou třídu. Každému dokumentu je prozatím přiřazena právě jedna značka, přestože zvolená kategorizace kombinuje některé formální a obsahové rysy (např. *rozhovor* – důležitá je formální struktura textu, a *sport* – důležitý je obsah. Např. u rozhovorů se sportovci se pak užívala značka pro převládající žánr: pokud je celý text rozhovorem, je tak i značen, pokud jde o sportovní reportáž s vloženým rozhovorem se sportovcem, jde o dokument značený jako sportovní zpráva.). Typy žánrů v textech jsou souhrnně uvedeny v tab. 5.1.

Samotné ruční anotaci žánrů předcházela automatická předanotace, jež využila informace z ruční anotace mezivýpovědních významových vztahů. V té anotátoři značili mimo jiné nekoherentní dokumenty sestávající ze sady krátkých nesouvisejících textů, případně i různých žánrů (tyto případy byly předznačeny jako *collections*), dále popisky fotografií, grafů či tabulek (dokumenty sestávající pouze z jedné věty a zároveň označené dříve jako popisky (*captions*) byly předznačeny žánrem *caption*). Anotace zbývajících dokumentů byla provedena osmi anotátory. 1/5 korpusu (vývojová a evaluační testovací data) byla anotována paralelně dvěma anotátory. Nesrovnalosti byly posléze řešeny rozhodcem. V případě systémových rozporů byly problematické žánry zkontrolovány rozhodcem ve všech datech příslušných anotátorů.

Související literaturu viz publikace [30], [31] a [33] v seznamu referencí na konci tohoto dokumentu.

Typy žánrů v textech PDT 3.0		
Český název	**Zkratka**	**Poznámka**
monologické žánry		
recenze	***review***	knihy, filmu, výstavy, koncertu, divadelního představení apod.
pozvánka	***invitation***	na koncert, na výstavu apod.
dopisy čtenářů	***letter***	
poradna / rady pro čtenáře	***advice***	rada, výklad jevu, příp. návod (jak podat trestní oznámení, jak postupovat v případě závěti, škola šachu, odpovědi na dopisy čtenářů)
program	***program***	televizní, rádia, výstavy atd.
popis filmu	***plot***	či televizního programu
sportovní zpráva	***sport***	+ přehled sportovních výsledků
komentář	***comment***	komentář něčeho, co se děje aktuálně (**kratší** rozsahem), vyjadřuje se subjektivní názor
zpráva	***news***	zpráva o něčem aktuálním, neobsahuje hodnocení, popisuje se situace. Patří sem i výsledky hospodaření apod.
úvaha	***essay***	větší zpráva/komentář, rozsahem **delší,** autor se **subjektivně** věnuje nějakému aktuálnímu nebo obecnému tématu
přehled	***overview***	např. přehled různých kurzů, přehled školení apod.
popis	***description***	produktu, firmy, služby apod.
předpověď počasí	***weather***	
výsledky ankety	***survey***	anketa a její výsledky
dialogické žánry		
rozhovor na určité téma	***topic_interv***	„aktuální rozhovor“, tj. rozhovor s odborníkem na určité aktuální téma
rozhovor se známou osobností	***person_interv***	obsahuje více různých témat, čtenáři se dozvídají informace ne o aktuálním tématu, ale o dané osobnosti
okrajové žánry		
kolekce	***collection***	kolekce různých textů v jednom dokumentu
popisek fotky	***caption***	popisky obrázků, grafů, tabulek apod.
metatext	***metatext***	text vzniklý v korpusu omylem, např. „konec podnadpisu“
jiné	***other***	nelze určit, co to má být – zejména u osamocených vět

Tab.5.1: Typy žánrů v PDT 3.0

Příklady

ln95046_021.t., genre = "sport"

(1) Další Jágrova branka
(2) New York -
(3) Český hokejista Jaromír Jágr vsítil svůj čtrnáctý gól této sezóny NHL a rozhodl jím o výsledku utkání Pittsburgh - Quebec (5:4).
(4) Závěrečná třetina byla nesmírně dramatická, padlo v ní šest branek, přičemž poslední slovo měl právě Jágr, který rozhodl zápas pouhých 22 sekund poté, co Nolan z Quebeku srovnal skóre na 4:4.
(5) Po čtyřzápasové pauze zaviněné chřipkou nastoupil i Martin Straka a vstřelil jeden gól.
(6) V Miami podlehla Florida týmu New Yorku Rangers 3:5.
(7) Za stavu 3:3 v závěrečné třetině prolomil nerozhodný stav Karpovcev, když puk z jeho hokejky skončil po odrazu v soupeřově brance.
(8) O konečném vítězství Jezdců 5:3 rozhodl Olczyk.

ln95045_056.t, genre = "collection"

(1) Krátce
(2) Návrhy britského premiéra J. Majora a jeho irského partnera J. Burtona na budoucí uspořádání Severního Irska získaly včera podporu britské vlády.
(3) Dokument se stane v příštích týdnech předmětem diskusí konstitučních severoirských politických stran.
(4) Dvěma hlavními cíli české zahraniční politiky jsou členství v Evropské unii a Severoatlantické alianci, řekl včera český ministr zahraničí Josef Zieleniec ve výboru pro zahraniční věci a zahraniční obchod Poslanecké sněmovny kanadského parlamentu.
(5) Dohodu o zastavení palby porušil další ozbrojený konflikt mezi armádou a povstaleckou organizací UNITA, ke kterému došlo u severoangolského města Uige.
(6) Irácká vláda nadále v "děsivé" míře a "bez jakýchkoli známek zlepšení" pošlapává lidská práva, konstatuje zvláštní zpravodaj OSN pro Irák Max van der Stoel ve zprávě, která byla včera zveřejněna v ženevském sídle OSN.
(7) Zatím nelze říci, kdy bude sestavena nová polská vláda, řekl po setkání představitelů polské vládní koalice, Polské lidové strany a Svazu demokratické levice koaliční kandidát na křeslo premiéra, maršálek Sejmu J. Oleksy.

ln94211_77.t, genre = "caption"

(1) Bývalého generála sovětského strategického letectva nezapře Džochar Dudajev vzorně salutující na slavnostní přehlídce uspořádané při příležitosti třetího výročí vyhlášení nezávislosti Čečenska na Rusku
(2) Foto Reuter

6 Víceslovné výrazy

Eduard Bejček a Pavel Straňák

Popis

Všechny víceslovné výrazy dané věty jsou uloženy v atributu *mwes* kořene jejího tektogramatického stromu. Atribut *mwes* je tvořen seznamem, jehož prvky reprezentují jednotlivé víceslovné výrazy v dané větě. Každý prvek seznamu *mwes*, tedy každá reprezentace víceslovné jednotky obsahuje *ID*, základní tvar výrazu v elementu *basic_form*, typ výrazu v elementu *type* a *seznam ID t-uzlů*, z nichž se víceslovný výraz skládá.

Anotované víceslovné výrazy jsou buďto víceslovné lexémy, nebo pojmenované entity, které nemusí mít charakter lexému, ale přesto uvnitř nich neplatí běžné syntaktické vztahy. U pojmenovaných entit určujme jejich druh. Typ víceslovné jednotky tedy může nabývat následujících hodnot:

- ***lexeme*** víceslovný lexém
- ***person*** jméno osoby nebo zvířete
- ***institution*** název instituce
- ***location*** zeměpisný název
- ***object*** název knihy, měrné jednotky, přírodovědné a chemické názvosloví
- ***address*** adresa
- ***time*** časový údaj či datum
- ***biblio*** bibliografický údaj
- ***foreign*** cizojazyčný výraz
- ***number*** číselný výraz, obvykle rozsah

Anotované víceslovné výrazy se v TrEdu zobrazují jako barevně šrafované množiny uzlů v tektogramatickém stromě (jehož topologie je až na případné opravy shodná s PDT 2.0), nebo jako jeden uzel. V tomto zkolabovaném zobrazení jsou děti t-uzlů, jež se staly prvky víceslovného výrazu, převěšeny přímo na tento nový uzel víceslovného výrazu. V „rozbaleném“ zobrazení jsou množiny různých druhů víceslovných výrazů (viz výše) zobrazeny různými barvami.

Anotační procedura

Byly anotovány všechny výskyty víceslovných výrazů (včetně pojmenovaných entit, viz níže) v části PDT, která obsahuje tektogramatickou rovinu. Značná část dat byla anotována paralelně. Tab. 6.1 udává, kolik dat bylo anotováno jedním anotátorem, dvěma či třemi anotátory paralelně, a jaký je poměr paralelních anotací ve vztahu k celé t-rovině. Čísla jsou uvedena v počtu souborů i v počtu tektogramatických uzlů v těchto souborech.

	Anotovaná data				
Paralelní anotace	1	2	3	PDT	2+3/PDT
t-soubory	1 288	1 412	465	3 165	59%
t-uzly	248 448	343 834	82 683	674 965	63%

Tab. 6.1: Anotace víceslovných výrazů

Soubory anotované jednotlivými anotátory nejsou součástí PDT 2.5, ale jsou dostupné pod licencí Creative Commons na stránce projektu (http://ufal.mff.cuni.cz/lexemann/mwe/).

Pro vydání v PDT 2.5 byly všechny neshody anotátorů jednoznačně rozhodnuty a vznikla tak *gold standard* data. Pokud se anotátoři shodli, víceslovný výraz byl zachován. Případy, kdy se neshodli, byly řešeny následovně:

- Pokud byl víceslovný výraz označen pouze jedním anotátorem, ponechali jsme anotaci, neboť se ukázalo, že mnohem spíše anotátor víceslovný výraz přehlédne, než by chybně označil něco navíc.
- Pokud jeden anotátor anotoval víceslovný výraz, který tvoří podmnožinu uzlů výrazu anotovaného druhým anotátorem, ponechali jsme výraz rozsáhlejší.
- Naopak jestliže jeden anotátor anotoval několik podmnožin výrazu druhého anotátora tak, že svou anotací plně pokryl daný rozsáhlejší výraz, zvolili jsme kratší výrazy.
- Případy průniku anotace dvou anotátorů byly individuálně rozhodnuty třetím anotátorem.
- Případy, kdy jeden anotátor anotoval podmnožiny výrazu anotovaného druhým anotátorem, ale nepokryl výraz celý (tedy se anotátoři neshodli, které t-uzly jsou součástí víceslovného výrazu) byly také rozhodnuty ručně třetím anotátorem.

Související literaturu viz publikace [34], [35] a [36] v seznamu referencí na konci tohoto dokumentu.

Příklady

(1) Prezident Havel by měl **15. července*** **na Pražském hradě**** jmenovat třináct soudců **Ústavního soudu*****.
* – *date, basic_form* "15. července"
** – *location, basic_form* "Pražský hrad"
*** – *institution, basic_form* "Ústavní soud"

(2) Funkce **ústavního soudce*** je neslučitelná s členstvím **v politických stranách****.
* – *lexeme, basic_form*"ústavní soudce"
** – *lexeme, basic_form* "politická strana"

7 Valenční slovník PDT-Vallex 3.0

Marie Mikulová

Popis

Současně s korpusem PDT 3.0 je vydána nová verze valenčního slovníku PDT-Vallex 3.0. Slovník PDT-Vallex vzniká paralelně se sémanticko-syntaktickou anotací vět, obsahuje téměř výlučně ta slovesa a ty jejich významy, které se vyskytly v anotovaných datech, tedy ty, jejichž valenční rámce musel anotátor znát, aby dokázal správně anotovat jednotlivá valenční doplnění a další doplnění slovesa v anotované větě. První verze slovníku PDT-Vallex (verze 1.0) vznikla při anotaci korpusu PDT 2.0. V rámci dalších anotačních projektů byl slovník dále rozšiřován.

Rozšiřování valenčního slovníku

První rozšíření slovníku přinesla anotace české části Pražského česko-anglického závislostního korpusu (Hajič et al., 2011; dále PCEDT 2.0; zkratka z anglického *The Prague Czech-English Dependency Treebank 2.0;* [37]). Korpus PCEDT 2.0 obsahuje články z deníku Wall Street Journal (z roku 1989), které byly pro českou část korpusu přeloženy do češtiny. Jedná se převážně o texty s ekonomickou tematikou. PDT-Vallex byl tudíž hojně rozšířen o slovesa a významy z této oblasti (např.: *nakonfigurovat, podhodnocovat, porcovat medvěda, prát peníze, segmentovat trh, seškrtnout finanční prostředky, srovnat se s riziky*).

Další velké rozšíření slovníku přinesla anotace Pražského závislostního korpusu mluvené češtiny (dále PDTSC 2.0; zkratka z anglického *The Prague Dependency Treebank of Spoken Czech 2.0;* [38]). Korpus PDTSC 2.0 obsahuje nahrávky dvojího typu: (i) českou část korpusu, který vznikl v rámci mezinárodního projektu Malach, (jedná se o lehce moderované dialogy s lidmi, kteří přežili holocaust) a (ii) dialogy, které byly nahrány v rámci projektu Companions (tématem dialogů je konverzace nad osobní sbírkou fotografií jednoho z účastníků dialogu). Do valenčního slovníku přibyla hesla z oblasti běžného (rodinného) života jako *háčkovat, houbařit, koledovat, pošťuchovat se, přebalit dítě, přivdat se, sáňkovat, zavařovat*, ale též i hesla z autentických výpovědí pamětníků holocaustu jako *prováličit, vybombardovat, odvlíknout, přežít, srocovat se*.

V rámci nových anotací nad daty korpusu PDT 2.0, které jsou nyní vydávány jako korpus PDT 3.0 byly ve valenčním slovníku provedeny jen drobné úpravy. Největší změnou bylo přidání nového rámce pro verbonominální predikáty (*být* + adjektivum, substantivum), jejichž infinitiv v pozici aktoru je kontrolován benefaktorem závislým na jmenné části predikátu: ACT(.f;aby[.v];že[.v]) PAT(.a1;.a7;.d); např.: *Je možné odejít. Je možno odejít. Je pro nás*.BEN *důležité přijít včas.* Rámec byl přiřazen 456 verbonominálním predikátům, jejichž jmennou část tvoří lemmata *možný, nutný, možno, nutno*. V další etapě práce bude seznam adjektiv v této funkci (PAT) rozšířen o další typy (např. *obtížný, snadný, zajímavý, ideální* atd.)

V tab. 7.1 jsou jednotlivá rozšíření valenčního slovníku vyjádřena v číslech. Po anotaci korpusu PDT 2.0 obsahoval valenční slovník 5 510 slovesných hesel a 9 191 valenčních rámců. Anotace dalších korpusů, které jsou s korpusem PDT 2.0 více méně srovnatelné (korpus PCEDT 2.0 obsahuje téměř stejný počet vět, tyto věty jsou však v průměru delší; korpus PDTSC 2.0 pak obsahuje velké množství krátkých vět s mnoha slovesy), obohatila slovník vždy o dalších cca 1 500 nových hesel a 2 500 nových valenčních rámců. (Zatím) poslední verze slovníku obsahuje téměř 8 500 slovesných hesel a 14 500 valenčních rámců.

		PDT 2.0	PCEDT 2.0	PDTSC 2.0
Data	Počet tokenů	833 195	1 151 150	742 221
	Počet vět	49 431	49 208	73 835
	Počet výskytů sloves	88 103	118 035	125 271
	Počet přiřazených hesel	5 376	4 880	4 628
	Počet přiřazených rámců	7 674	8 285	7 582
Slovník	Počet hesel ve slovníku	5 510	7 104	8 459
	Počet rámců ve slovníku	9 191	11 933	14 517
		PDT-Vallex 1.0	**PDT-Vallex 2.0**	**PDT-Vallex 3.0**

Tab. 7.1: Rozšiřování valenčního slovníku

Zachycení nestandardních jevů ve valenčním slovníku

Anotace mluveného korpusu PDTSC 2.0 si vyžádala nové úpravy v zápisu valenčních hesel. Pro označení různého stupně nestandardních jevů byl do zápisu valenčních hesel zaveden znak procenta (%). Tento znak lze použít v následujících kontextech:

- za lemmatem, kde značí nestandardní lemma. Jeden znak % se používá pro nespisovná, expresivní či jinak „podivná“ lemmata (př. (1)). Dva znaky % mají vulgární lemmata (př. (2)).
- za celým rámcem. Zde % označuje nestandardní slovesný rámec, nějaký méně obvyklý význam daného slovesného lemmatu (př. (3)). Dva znaky % používáme pro vulgární slovesné významy (př. (4)).
- za značkou pro funkci valenčního členu, kde označuje nestandardní valenční člen, který se v daném významu obvykle neužívá, a působí proto nepatřičně (př. (5)).
- za formou, kde označuje nestandardní formální realizaci daného valenčního členu, která se obvykle neužívá a v psaném textu by působila nepatřičně, stylisticky neobratně (př. (6)).

Různé kontexty užití znaku % lze v rámci zápisu valenčního hesla kombinovat. Pomocí znaku % u lemmatu a zároveň u valenčního rámce zachycujeme případy, kdy u nespisovné podoby (např. *píct*) jinak běžného spisovného slovesa (*péci*) označuje jeden z valenčních rámců příznakový význam, zatímco další valenční rámce označují nepříznakové významy (př. (7)).

Příklady

(1) čumět **%** ACT(.1) DIR3(*) *Čuměla dvě hodiny na obraz.*
(2) chlastat **%%** ACT(.1) PAT(.4) *Začal chlastat alkohol.*
(3) bruslit ACT(.1) PAT(v+6) **%** *Bruslil jsem v chemii.*
(4) držet ACT(.1) DPHR (hubu) **%%** *Drž hubu!*
(5) dobýt ACT (.1) PAT(.4) ?ORIG**%**(od+2) *Angličané dobyli Palestinu od Turků.*
(6) dráždit ACT (.1) ADDR(.4) ?PAT(k+3;na+4**%**) *Dráždí mě to na kašel.*
(7) píct **%** ACT(.1) PAT(s+7) **%** *Mohl bych píct s jinou.*
ACT(.1) PAT(.4) *Budeme píct koláče.*

Související literaturu viz publikace [39], [40] a [41] v seznamu referencí na konci tohoto dokumentu.

B. Úpravy a doplňky na analytické rovině korpusu

8 Klauze

Zdeněk Žabokrtský

Popis

Analytické stromy v PDT 3.0 (původně v PDT 2.5) jsou obohaceny o anotaci klauzí. Klauze obsahuje jedno určité sloveso. Klauze může být součástí souvětí. Anotaci klauzí lze použít pro trénovaní rozpoznávačů hranic vět, které mohou být užitečné v řadě dílčích úkolů při zpracování přirozeného jazyka, mj. při syntaktické analýze nebo strojovém překladu.

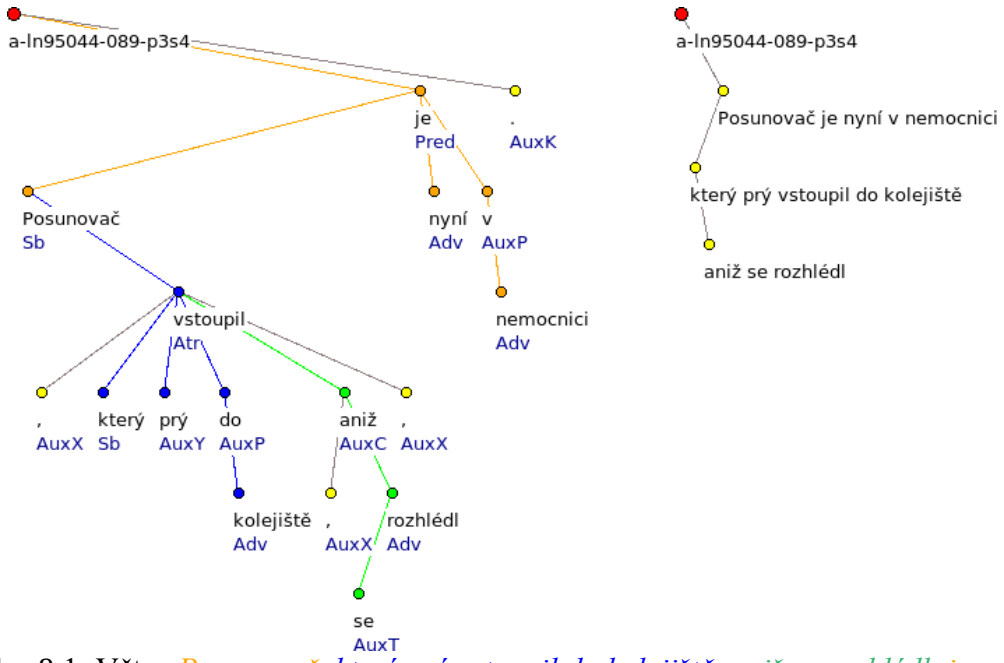
Očekávali jsme, že anotaci klauzí lze provést s vysokou úspěšností automaticky, máme-li už k dispozici ruční anotaci morfologie a závislostní syntaxe. Anotace klauzí v PDT 3.0 je tedy z větší části založena na aplikaci automatických pravidel. Ruční anotace byly provedeny pouze pro malou část dat, tato data byla použita pro vývoj procedury pro automatické značkování. F-skóre této procedury dosahovalo 97.51 %. Aby byla anotace klauzí v PDT konzistentní, tato procedura byla aplikována na všechna data v PDT a původní ruční anotace nejsou do PDT 3.0 zahrnuty.

Technicky jsou hranice klauzí reprezentovány novým atributem ***clause_number*** přiřazeným k analytickým uzlům. Jestliže dva analytické uzly ve stromě sdílejí stejnou nenulovou hodnotu tohoto atributu, potom patří do stejné klauze. Nulová hodnota tohoto atributu je rezervovaná pro hraniční tokeny (tj. tokeny umístěné na hranici dvou klauzí, které nemohou být jednoznačně přiřazeny ani k jedné z nich). Hraničními tokeny jsou typicky interpunkční značky (značkované jako `Z:`) nebo koordinační spojky (`J^`). Podřadicí spojky (`J,`) jsou systematicky zařazovány do příslušné závislé klauze.

Vizualizace

Segmentaci klauzí lze pohodlně zobrazit v TrEdu (viz obr. 8.1). Nové rozšíření TrEdu pro zobrazování dat PDT 3.0 nabízí dvě další makra spojená s anotací klauzí:

- **Sbalování/rozbalování klauzí (f)** – Pokud je sbalování klauzí zapnuto, všechny tokeny reprezentující stejnou klauzi jsou reprezentovány jediným uzlem. V takovém zobrazení nejlépe vynikne struktura souvětí.
- **Barvení klauzí (c)** – Pokud je barvení klauzí zapnuto, uzly, které patří do stejné klauze, jsou zobrazeny stejnou barvou.



Obr. 8.1: Věta „*Posunovač, který prý vstoupil do kolejiště, aniž se rozhlédl, je nyní v nemocnici*.“ reprezentovaná dvěma stromy: úplný strom na levé straně a strom s klauzemi zredukovanými do uzlů na pravé straně.

Příklady

(1) U sochy básníka seděl vlasatý mladík a* hrál Vysockého písně.**
* – hranice klauzí, souřadicí spojka spojuje dvě klauze
** – interpunkce na konci věty, hranice věty

(2) Pokud jde o kupní smlouvu a* všechny náležitosti s ní spojené,** musí si to zařídit a* zaplatit strany samy.
* – koordinační spojky spojující větné členy uvnitř klauze
** – hranice klauzí, interpunkce

(3) Lidé na nás tehdy chodili, aby* se odreagovali od přítomného režimu.
* – podřadicí spojka

(4) Posunovač, který prý vstoupil do kolejiště, aniž se rozhlédl, je nyní v nemocnici*.
* – hlavní klauze rozdělená do dvou částí závislou vztažnou klauzí (která je dále modifikována závislou klauzí)

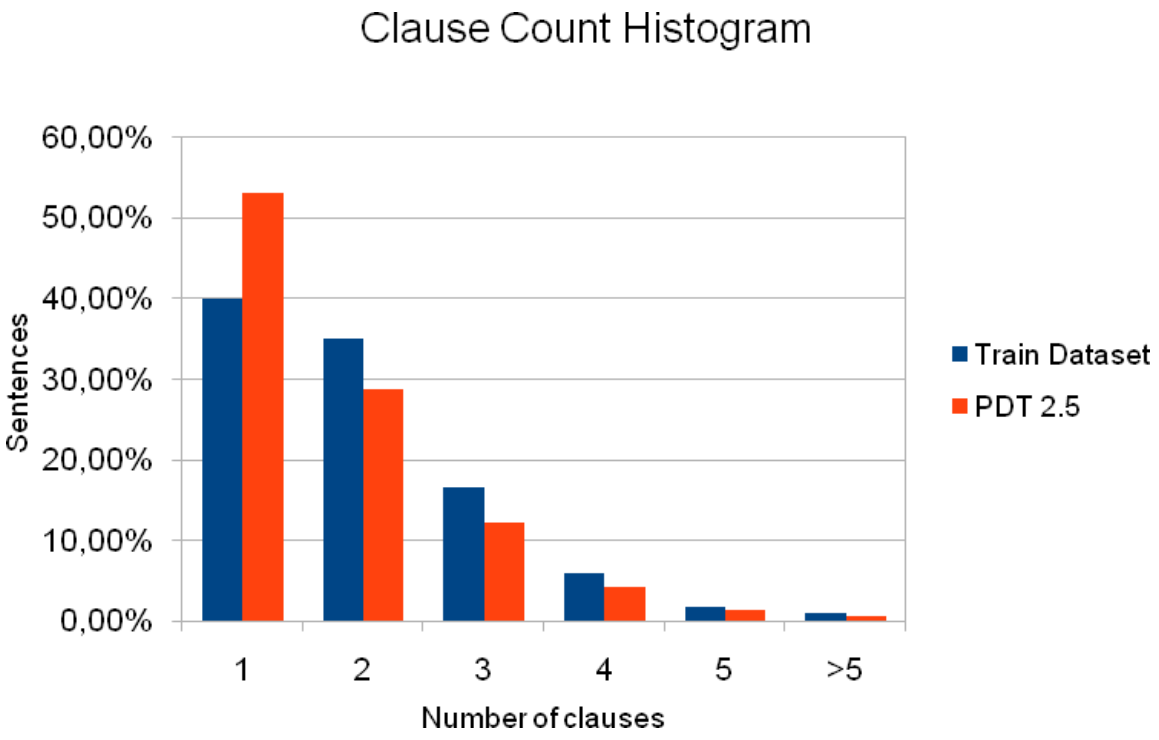
Anotační procedura

Automatická procedura pro rozpoznávání klauzí pracuje následovně:

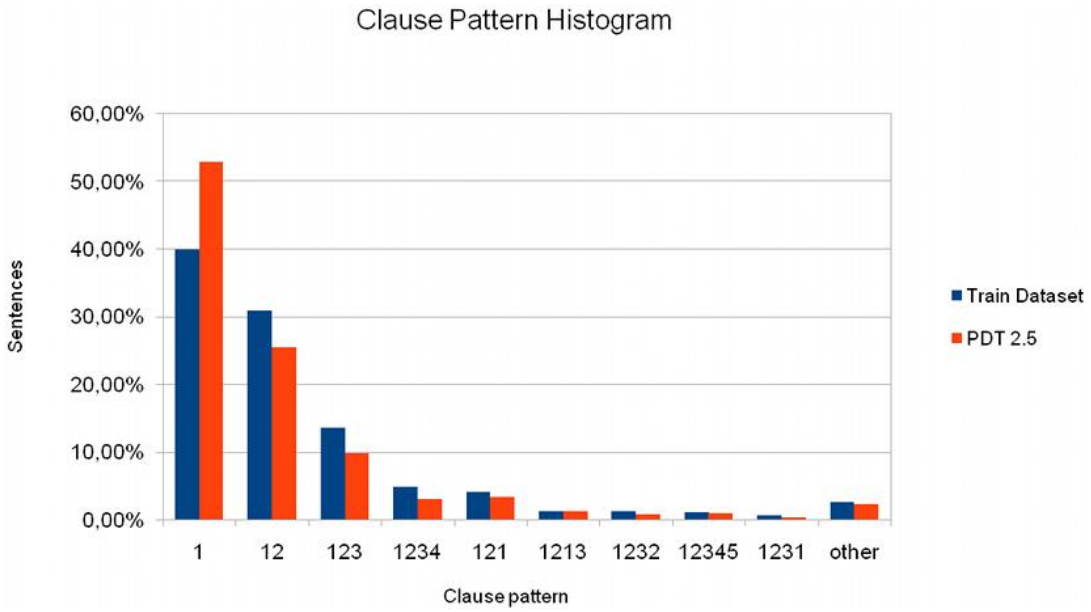
- Identifikujeme "jádra" klauzí. Každý výskyt určitého slovesa je označen jako nové jádro klauze.
- Jádra tvořící složenou slovesnou formu jsou spojena. Jádra s analytickou funkcí pomocného slovesa (*AuxV*) nemohou samy tvořit klauzi.
- Strom je rekurzivně procházen (v pořadí post-order) a každá hlava koordinace je dočasně připojena ke klauzi svého nejvíce vpravo umístěného členu.
- Dokončení klauzí. Strom je rekurzivně procházen (v pořadí pre-order) a každý uzel je zpracován se svými potomky. Uzly, které nepatří zatím do žádné klauze, jsou typicky přiřazeny ke klauzi rodiče. Speciální zpracování ovšem vyžadují koordinační struktury.
- Náležitost k příslušné klauzi je znovu přepočítána pro všechny potenciální hraniční uzly.

Souhrnné statistické údaje

Segmentace na jednotlivé klauze je v PDT 3.0 k dispozici pro 87 913 vět, které obsahují celkem 153,434 klauzí. Obr. 8.2 zobrazuje rozložení počtu klauzí na větu, obr. 8.3 zobrazuje nejčastější typy struktury souvětí.



Obr.8.2: Rozložení počtu klauzí na větu



Obr. 8.3: Pro jednoduchost jsou zde klauze označené číslicemi. Vzorec "12" odpovídá souvětí tvořenému dvěma klauzemi, vzorec "121" odpovídá také souvětí se dvěma klauzemi, přičemž jedna z nich je vnořená uprostřed druhé.

Reference

[1] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., Zikánová, Š.: *Prague Dependency Treebank 3.0*. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, 2013.
http://ufal.mff.cuni.cz/pdt3.0/

[2] Bejček, E., Panevová, J., Popelka, J., Smejkalová, L., Straňák, P., Ševčíková, M., Štěpánek, J., Toman, J., Žabokrtský, Z., Hajič, J.: *Prague Dependency Treebank 2.5*. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, 2011.
http://ufal.mff.cuni.cz/pdt2.5/; https://ufal-point.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0006-DB11-8

[3] Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., Žabokrtský, Z.: Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, Coling 2012 Organizing Committee, Mumbai, India, pp. 231-246, 2012.

[4] Hajič a kol.: *Prague Dependency Treebank 2.0*. Data/software, Linguistic Data Consortium, Philadelphia, PA, USA, 2006. ISBN 1-58563-370-4
www.ldc.upenn.edu, 2006.

[5] Mikulová a kol.: *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka.* Technical report no. 2005/TR-2005-28, Univerzita Karlova v Praze, MFF, ÚFAL, Praha, 2005. ISSN 1214-5521.
http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/index.html

[6] Poláková, L., Jínová, P., Zikánová, Š., Hajičová, E., Mírovský, J., Nedoluzhko, A., Rysová, M., Pavlíková, V., Zdeňková, J., Pergler, J., Ocelák, R.: *Prague Discourse Treebank 1.0*. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, 2012.
http://ufal.mff.cuni.cz/discourse/; https://ufal-point.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0008-E130-A

Reference k *1.1 Substantivní gramatém typgroup*

[7] Panevová, J. – Ševčíková, M.: Delimitation of information between grammatical rules and lexicon. In: *Proceedings of the International Conference on Dependency Linguistics (Depling 2011)*, Universitat Pompeu Fabra, Barcelona, 2011, pp. 173–182. ISBN 978-84-615-1834-0.

[8] Panevová, J. – Ševčíková, M.: Jak se počítají substantiva v češtině: poznámky ke kategorii čísla. *Slovo a slovesnost*, 72, 2011, s. 163–176. ISSN 0037-7031.

[9] Ševčíková, M. – Panevová, J. – Smejkalová, L.: Specificity of the number of nouns in Czech and its annotation in Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 96, pp. 27–47, 2011. ISSN 0032-6585.

[10] Ševčíková, M. – Panevová, J. – Žabokrtský, Z.: Grammatical number of nouns in Czech: linguistic theory and treebank annotation. In: *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT 2010)*, NEALT Proceedings Series, Vol. 9. Tartu, Estonia, 2010, pp. 211–222. ISSN 1736-8197.

Reference k *1.2 Slovesný gramatém factmod*

[11] *Mluvnice češtiny II.* Academia, Praha, 1986.

[12] Panevová, J. – Ševčíková, M: Annotation of Morphological Meanings of Verbs Revisited. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta: ELRA, 2010, pp. 1491–1498. ISBN 2-9517408-6-7.

[13] Ševčíková, M.: *Funkce kondicionálu z hlediska významové roviny.* UFAL MFF UK, Praha, 2010, 179 pp. ISBN 978-80-904175-2-6.

[14] Ševčíková, M.: The meaning of the conditional mood within the tectogrammatical annotation of Prague Dependency Treebank 2.0. In: *Proceedings of the Slovko 2009 Conference: NLP, Corpus Linguistics, Corpus Based Grammar Research*. Bratislava: Slovenská akadémia vied, pp. 321–330, 2009. ISBN 978-80-7399-875-2.

[15] Ševčíková, M.: Kondicionál přítomný jako součást explicitních performativních formulí. *Korpus – gramatika – axiologie*, Vol. 1, No. 1, pp. 41–62, 2010. ISSN 1804-137X.

Reference k *1.3 Slovesný gramatém diatgram*

[16] Panevová, J.: O rezultativnosti (zejména) v češtině. In: *Gramatika i leksika u slovenskim jezicina*. Novi Sad, Beograd: Matica Srbska, Institut za srpski jezik. pp. 165 – 176, 2011.

[17] Panevová, J. – Ševčíková, M.: Delimitation of Information between Grammatical Rules and Lexicon. In: *Linguistic Aspects of Dependency* (Wanner, L., Gerdes, K, eds.). John Benjamins Publ. House, Amsterdam/the Netherland, pp.1- 20, 2013.

[18] Panevová, J. – Ševčíková, M.: The Role of Grammatical Constraints in Lexical Komponent in Functional Generative Description. In: *Proceedings of the 6th International Konference on Meaning-Text Tudory*. Praha, pp. 134-143, 2013.

Reference k *1.4 Atribut sentmod*

[19] Ševčíková, M. – Mírovský, J.: Sentence Modality Assignment in the Prague Dependency Treebank. In: *Proceedings of the 15th International Conference Text, Speech and Dialogue (TSD 2012)*. Springer, Berlin, pp. 56–63, 2012. ISBN 978-3-642-32789-6, ISSN 0302-9743.

Reference k *2 Zrušení lemmatu #Benef*

[20] Panevová, J.: „Být posel dobrých zpráv je mi příjemné“ (Několik poznámek k infinitivním konstrukcím). In: *Karlík a továrna na lingvistiku. Prof. Petru Karlíkovi k 60. Narozeninám*. (eds. A. Bičan, J. Klaška, P. Macurová, J. Zmrzlíková). Host/Masarykova univerzita, Brno, s. 345 – 354, 2010.

[21] Panevová, J.: On the Syntax and Semantics of Czech Infinitival Constructions: A Case Study. In: *Slovo i jazyk. Sbornik statej k vosmidesjatiletiju akademika Ju. D. Apresjana.* Jazyki slavjanskich kul'tur, Moskva, pp. 541 – 551, 2011.

[22] Panevová, J.: Infinitiv ve funkci atributu. In: *Kapitoly z české gramatiky* (ed. F. Štícha), Academia., Praha, s. 945 – 960, 2011.

[23] Panevová, J. a kol.: *Mluvnice současné češtiny 2. Syntax na základě anotovaného korpusu (kap. 5).* Karolinum, Praha (v tisku).

Reference k *3 Koreference a asociační anafora*

[24] Nedoluzhko, A.: *Rozšířená textová koreference a asociační anafora. Koncepce anotace českých dat v Pražském závislostním korpusu*. UFAL MFF UK, Praha, 2011.

[25] Nedoluzhko, A., Mírovský, J.: *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank. Annotation manual.* Technical report No. 44, UFAL MFF UK, Prague, 2011.

[26] Nedoluzhko, A., Mírovský, J.; Novák, M.: A Coreferentially annotated Corpus and Anaphora Resolution for Czech. In: *Computational Linguistics and Intellectual Technologies*. ABBYY, Moscow, Russia, pp. 467-475, 2013. ISBN 978-1-937284-58-9

[27] Nedoluzhko, A.: Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In: *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*. Omnipress, Inc, Sofia, Bulgaria, pp. 103-111, 2013. ISBN 978-1-937284-58-9

[28] Nedoluzhko, A., Mírovský, J.: How Dependency Trees and Tectogrammatics Help Annotating Coreference and Bridging Relations in Prague Dependency Treebank. In: *Proceedings of the Second International Conference on Dependency Linguistics, Depling 2013*. Matfyzpress, Charles University in Prague, Prague, pp. 244-251, 2013. ISBN 978-80-7378-240-5

Reference k *4 Mezivýpovědní významové vztahy* a *5 Žánrová specifikace textů*

[29] Jínová, P., Mírovský, J., Poláková, L.: Analyzing the Most Common Errors in the Discourse Annotation of the Prague Dependency Treebank. In: *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, Edicoes Colibri, Lisboa, Portugal, pp. 127-132, 2012. ISBN 978-989-689-274-6

[30] Poláková, L., Jínová, P., Zikánová, Š., Bedřichová, Z., Mírovský, J., Rysová, M., Zdeňková, J., Pavlíková, V., Hajičová, E.: *Manual for Annotation of Discourse Relations in Prague Dependency Treebank.* Technical report no. 2012/47, UFAL MFF UK, Praha, 2012.

[31] Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., Hajičová, E.: Introducing the Prague Discourse Treebank 1.0. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Asian Federation of Natural Language Processing, pp. 91-99, 2013. ISBN 978-4-9907348-0-0

[32] Prasad, R., Dinesh, N., Lee, N., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse Treebank 2.0. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).* Marrakech, Morocco,2008.

[33] Webber, B.: Genre distinctions for Discourse in the Penn TreeBank. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*. Singapore, 2009.

Reference k *6 Víceslovné výrazy*

[34] Bejček, E., Straňák, P.: Annotation of Multiword Expressions in the Prague Dependency Treebank. In: *Language Resources and Evaluation*, Vol. 44, No. 1-2, Springer Netherlands, pp.7-21, 2010. ISSN 1574-020X.

[35] Bejček, E., Straňák, P., Hajič, J.: Finalising Multiword Annotations in PDT. In: *Proceedings of 8th Treebanks and Linguistic Theories Workshop (TLT)*. Universita Cattolica del Sacro Cuore, Milano, Italy, pp. 17-25, 2009. ISBN 978-88-8311-712-1.

[36] Straňák, P.: Annotation of Multiword Expressions in The Prague Dependency Treebank. Ph.D. thesis, Univerzita Karlova v Praze, Praha, 2010.

Reference k *7 Valenční slovník PDT-Vallex 3.0*

[37] Hajič, J. et al.: *Prague Czech-English Dependency Treebank 2.0*. Data/software. UFAL MFF UK, Prague, 2011.
http://ufal.mff.cuni.cz/pcedt2.0/.

[38] Hajič, J. et al.: *Prague Dependency Treebank of Spoken Czech 2.0*. Data/software. UFAL MFF UK, Prague, (připravuje se k vydání v 2014).
http://ufal.mff.cuni.cz/pdtsc1.0/.

[39] Hajič, J. et al.: PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*. Vaxjo University Press, Vaxjo, pp. 57-68, 2003.

[40] Mikulová, M. – Urešová, Z.: Liší se mluvené a psané texty ve valenci? *Korpus – gramatika – axiologie,* No. 8, pp. 36-46, 2013.

[41] Urešová, Z.: *Valenční slovník Pražského závislostního korpusu (PDT-Vallex).* UFAL MFF UK, Praha, 2011.

Reference k *8 Klauze*

[42] Lopatková, M., Homola, P., Klyueva, N.: Annotation of sentence structure: Capturing the relationship between clauses in Czech sentences. In: *Language Resources and Evaluation*, Vol. 46, No. 1, Springer Netherlands, pp. 25-36, 2011. ISSN 1574-020X

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2004-26 Martin Holub, Jiří Diviš, Jan Pávek, Pavel Pecina, Jiří Semecký, *Topics of Texts. Annotation, Automatic Searching and Indexing*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Nguy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová, *Manual for annotation of discourse relations in the Prague Dependency Treebank*

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)*