

# Obsah

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Úvod</b>                                       | <b>7</b>  |
| 1.1      | Motivace . . . . .                                | 7         |
| 1.2      | Přínosy práce . . . . .                           | 7         |
| 1.2.1    | Souhrn . . . . .                                  | 8         |
| 1.2.2    | Průzkum . . . . .                                 | 8         |
| 1.2.3    | Návrhy . . . . .                                  | 8         |
| 1.2.4    | Teorie . . . . .                                  | 8         |
| <b>2</b> | <b>Sémantická anotace</b>                         | <b>9</b>  |
| 2.1      | Rozdělení . . . . .                               | 9         |
| 2.2      | Teorie . . . . .                                  | 10        |
| <b>3</b> | <b>Lingvistická anotace</b>                       | <b>12</b> |
| 3.1      | Lingvistické značky . . . . .                     | 13        |
| 3.1.1    | Morfologická rovina . . . . .                     | 14        |
| 3.1.2    | Analytická rovina . . . . .                       | 14        |
| 3.1.3    | Tektogramatická rovina . . . . .                  | 15        |
| 3.1.4    | Příklady . . . . .                                | 19        |
| 3.2      | Jazyky pro zápis lingvistických anotací . . . . . | 20        |
| 3.2.1    | PML - The Prague Markup Language . . . . .        | 20        |
| 3.2.2    | FS - Feature structure . . . . .                  | 20        |
| 3.2.3    | CSTS - Czech sentence tree structure . . . . .    | 20        |
| 3.2.4    | Konverze mezi formáty PDT . . . . .               | 21        |
| 3.3      | The Prague Dependency Treebank . . . . .          | 21        |
| 3.4      | WordNet . . . . .                                 | 21        |
| 3.5      | Lingvistické nástroje . . . . .                   | 21        |
| 3.5.1    | NetGraph . . . . .                                | 21        |
| 3.5.2    | Tree Editor TrEd . . . . .                        | 21        |

|          |  |           |
|----------|--|-----------|
| 3.5.3    | Tools for machine annotation - PDT 2.0 . . . . . | 22        |
| <b>4</b> | <b>Experiment</b>                                | <b>24</b> |
| 4.1      | Vstupní data . . . . .                           | 24        |
| 4.1.1    | Hasiči . . . . .                                 | 25        |
| 4.1.2    | Úpadci . . . . .                                 | 25        |
| 4.2      | Výstupní data . . . . .                          | 25        |
| 4.3      | Software . . . . .                               | 25        |
| 4.3.1    | Modul XY . . . . .                               | 25        |
|          | <b>Literatura</b>                                | <b>26</b> |

Pracovní

XX

Hrubá osnova výzkumné zprávy:

1. Úvod charakterizující kontext problému
2. Přesné vymezení problému a cílů výzkumu v daném kontextu
3. Struktura práce ve vztahu k vytyčeným cílům
4. Analýza problému z různých hledisek
5. Návrh řešení s diskusí alternativ
6. Rozpracování vybraného přístupu
7. Zhodnocení splnění cílů, otevřené problémy, event. doporučení dalšího směru výzkumu
8. Srovnání s jinými přístupy známými z literatury s explicitním zdůrazněním původního přínosu
9. Závěr se stručným shrnutím výsledků a přínosu
10. Seznam citované literatury

# Kapitola 1

## Úvod

Sémantická anotace je .....

Sémantický web

### 1.1 Motivace

### 1.2 Přínosy práce

Komu?

Čím?

Srozumitelné pro nelingvistu.

Mimo jiné: Tato práce se snaží přiblížit možnosti, jak využít dostupné lingvistické nástroje analyzující český text především lidem, kteří se zabývají extrakcí informací z textu ale nejen jím. Práce na čtenáře neklade žádné nároky co se týká lingvistického vzdělání a sama základní znalosti z lingvistiky podněcuje. Těmito znalostmi se snaží pokrýt požadavky, které klade používání lingvistických nástrojů zde popisovaných. Zběžné znalosti zde poskytnuté jsou doplněny odkazy a referencemi na zdroje, kde se čtenář o dané problematice může dozvědět více.

### 1.2.1 Souhrn

Práce poskytuje základní souhrn v oblasti sémantické anotace, takže si čtenář může udělat představu o tom, kterými směry se sémantická anotace ubírá, jaké metody byly využity, s jakou úspěšností atp.

XXXXX

Základní přehled ontologií.

Doporučení autorům stránek (DRFA, HTML-A)

### 1.2.2 Průzkum

Součástí práce je praktický experiment s lingvistickými nástroji: !!!!!!!!!!!!!!!dopsat!!!!!!!!!!!!. Čtenáři jsou poskytnuty zkušenosti z týkající se použití těchto nástrojů a z prací které s jejich použitím souvisely. Tyto zkušenosti se týkají především dostupnosti, zprovoznění, výkonnosti, přínosů a nedostatků těchto nástrojů. !!!!!!!!!!!!!!!!!!!!!!!!!!!!!rozvést: navržený postup X další možnosti

Postup experimentu se v jednotlivých fázích snaží kopírovat skutečné akce, které by bylo nutné provést v opravdovém projektu zaměřeném na sémantickou anotaci. V práci tak vzniká jednoduchá základní analýza tohoto typu projektů. Ve skutečném projektu pak bude možné ji přinejmenším jako inspiraci využít.

### 1.2.3 Návrhy

V práci je navržena metodika, jak by se při extrakci informací pomocí nástrojů v rámci této práce testovaných dalo postupovat.

Je zde navržený jednoduchý dotazovací jazyk pro lingvistické anotace.

Stručný návrh indukce vzorů.

Zamyšlení nad možnostmi lingvistické anotace pro indexaci dokumentů.

### 1.2.4 Teorie

Pokus o teoretický přínos v oblasti sémantické anotace.

## Kapitola 2

# Sémantická anotace

Sémantická anotace není v současné době přesně vymezený termín. V této práci budu s tímto pojmem pracovat poměrně volně až na kapitolu o teoretických otázkách sémantické anotace 2.2, kde bych se o takové vymezení chtěl pokusit. V knize [1] se tomuto termínu vyhýbají opisem *annotation for the Semantic Web* tedy anotace pro sémantický web. V tomto smyslu budu užívat pojem sémantické anotace ve zbytku práce. Tedy sémantickou anotací budu rozumět takovou anotaci, která je určená pro sémantický web. Pro přesnost ještě doplníme, že anotací jako hotovým dílem rozumíme výsledek, výstup případně výstupní data procesu anotace.

Člověk by intuitivně očekával, že výsledkem sémantické anotace bude nějak označovaný text, který je jejím vstupem. Případně se nemusí jednat o text ale o strukturovaná data - nejčastěji HTML případně XML. V našem způsobu chápání sémantické anotace, tedy jako anotace pro sémantický web, však jejím výstupem může být téměř libovolný datový celek. Příznačné pro sémantické anotace je to, že nějakým způsobem využívají ontologie a RDF, viz !!!!!!!!!!!!!doplnit!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!.

Autorská anotace.

Dodatečná anotace, strojová.

## 2.1 Rozdělení

Po praktické stránce: dostupnost, upravitelnost (jiná doména), stabilita - náchylnost na změny v datech

Po teoretické stránce: čísla (úspěšnost), metody

Labský (pouze extrakce informací, praktický, granty)

Xxxx <http://citeseer.ist.psu.edu/149407.html> složkovou lingvistickou anotace:

## 2.2 Teorie

Domnívám se, že nebylo mnoho teoretických otázek týkajících se sémantická anotace vůbec formulováno, natož uspokojivě vědecky vyřešeno.

XXX

rozdíl mezi konkrétním a abstraktním

XX

Zbytek rozpracováno

XX

V čem tkví sémantická anotace?

Co si představujeme pod ideální sémantickou anotací?

Jaký je vztah lingvistické a sémantické anotace? Jsou mezi nimi hranice? Kde přibližně?

Dal by se z toho odvodit nějaký univerzálnější návod/algoritmus, jak od lingvistické anotace k té sémantické přejít?

Pokusit se stanovit podmínky, které ideální anotaci brání, resp. předpoklady, které by ji umožnily.

Jaký je rozdíl mezi přirozeným jazykem a deskripční logikou?

XX

Při hledání nějaké věty k ukázkové anotaci jsem narazil na tuhle: (je to z přihlášky na DS)

Uchazeč vyplní obor studia, výzkumné téma, škálu pracoviště, zajistí podpis školitele a podpis předsedy ...

Tahle věta je zvláštní v tom, že je formulována obecně: pro všechny uchazeče, pro libovolné téma, pracoviště, podpis libovolného školitele, předsedy. Avšak školitel je pevně spojen s tématem práce a předseda je spojen s pracovištěm.

Jak anotovat takovou vetu? Jaka je její semantika? Vznikne nějaký Abox? Nebo v anotaci použijeme nějaké volné proměnné pro individua. Nebo budeme anotovat pouze pomocí názvu tříd a instance nějakým způsobem vynecháme?



## Kapitola 3

# Lingvistická anotace

Lingvistickou anotací budeme v této práci označovat činnost, při které se text přirozeného jazyka obohacuje o lingvistickou informaci o slovech, větách, vztazích mezi slovy, mezi větami, o typu a původu textu atp. Lingvistická anotace nebo též značkování korpusu je jedna z činností korpusové lingvistiky. Korpusem rozumíme soubor textů spolu s lingvistickou informací k nim dodanou. Korpusová lingvistika je poměrně novou disciplínou, jejíž vznik, stejně jako vznik celé počítačové lingvistiky vůbec, umožnil rozvoj výpočetní techniky posledních let. Korpusová lingvistika se zabývá zkoumáním a shromažďováním textů přirozeného jazyka (vytvářením korpusu). Texty se anotují za velké podpory počítače - například morfologická desambiguace (viz 3.5.3) by byla bez softwarové podpory nadlidský úkol. Avšak pro anotaci korpusu je stále nutná spousta lidské „ruční“ práce. Takto anotované texty představují velmi cenná data, ze kterých se především pomocí statistických metod dají vyvodit nové poznatky o jazyce. Díky ručně anotovaným korpusům vynikla a stále vyniká většina softwarových nástrojů pro počítačové zpracování přirozeného jazyka.

Korpusová lingvistika dnes nemalou mírou přispívá k jazykovému výzkumu. Na stránkách Českého národního korpusu<sup>1</sup> se dokonce uvádí, že přináší natolik nové poznatky o jazyce, že do dosavadního vývoje jazykovědy vnáší radikální převrat. Toto tvrzení nemusí působit překvapivě, pokud například srovnáme původní latinskou lingvistickou terminologii s tou, která vzniká v počítačové lingvistice v poslední době.

---

<sup>1</sup><http://ucnk.ff.cuni.cz/>

Korpusy se v zásadě značkují třemi druhy značek. 1) Značky správně zachycují identifikační údaje o každém textu - informace o jeho původu a zdroji. 2) Značky strukturní zachycují hierarchickou strukturu textu tj rozdělení textu do kapitol, odstavců, vět, slov a interpunkčních znamének (tokenů). 3) Značky lingvistické jsou přiřazeny k jednotlivým slovům a nesou informaci o lingvistických kategoriích, které dané slovo nese.

Samostatnou kapitolou lingvistické anotace je potom zachycení gramatické stavby věty. K tomu se používají dva typy gramatik - složková a závislostní gramatika. Závislostní gramatika má dlouholetou tradici v popisu jazyků evropského kontinentu a zdá se, že má určité výhody i pro popis angličtiny, která bývá častěji zpracovávána gramatikou složkovou. Složkový popis je blíže Chomského pojetí jazyka. Věta je podle složkové gramatiky rekurzivně dělena do menších a menších složek. Postup začíná rozdělením věty na část podmětnou a přísudkovou a postupuje dělením těchto složek na podsložky až dojde k jednotlivým slovům. Závislostní přístup naproti tomu vezme jednotlivá slova a ta pospojuje závislostními hranami do takzvaného závislostního stromu. Velmi podrobně je problematika gramatik a větné syntaxe popsána v [3].

Kromě toho, že je možné strukturu věty přirozeného jazyka zapisovat pomocí různých typů gramatik, je též možné zapisovat tuto strukturu na různých významových rovinách. Lingvistický výzkum různých jazyků ukazuje vhodnost takového vícevrstvého popisu. Na různých rovinách je totiž možné přehledněji postihnout různé jazykové jevy. !!!!!!!doplnit podrobnosti?!!!!!!! V práci [2] je představena tradiční trojice rovin pro popis struktury věty. Jsou to 1) rovina morfématická (též morfologická) - tvarosloví 2) rovina povrchové syntaxe (též rovina mluvnické stavby věty či analytická rovina) 3) rovina tektogramatická neboli významová stavba věty nebo též hloubková syntax. Stejně jsou roviny lingvistické anotace zpracovány v projektu PDT (viz oddíl 3.3), z jehož popisu budeme dále vycházet.

### 3.1 Lingvistické značky

Lingvistické značky rozdělíme do tří kategorií. Podle roviny lingvistické anotace, tak jak jsou rozděleny v projektu PDT (viz oddíl 3.3). Značky morfologické roviny jsou nezávisle přiřazovány jednotlivým slovům. Naproti tomu značky analytické a tektogramatické roviny popisují strukturu věty a jejich

značky popisující vztahy mezi jednotlivými slovy se mohou týkat více slov najednou.

Následuje stručný popis jednotlivých značek každé roviny. Podrobnější popis lingvistických značek je možné najít například v [5], [6], [7].

### **3.1.1 Morfologická rovina**

#### **Slovní tvar**

Tato značka obsahuje tvar, v jakém se dané slovo vyskytuje v původním textu včetně zápisu malých a velkých písmen. Od původního výskytu se liší se jen ve výjimečných případech, kdy například původní slovní tvar byl číslice s desetinnou čárkou (snaha o jednotný zápis čísel) nebo se jednalo o překlep.

#### **Lemma**

Lemma je takzvaný základní tvar slova. Jednoznačně slovo identifikuje. V tomto tvaru je dané slovo obvykle uváděno ve slovnících.

#### **Morfologická značka**

Morfologická značka v sobě spojuje informaci o morfologických kategoriích, které dané slovo nese. Z morfologické značky je možné zjistit slovní druh, jmenný rod, číslo, pád, osobu, čas, atd.

### **3.1.2 Analytická rovina**

Analytická rovina je první úroveň pro strukturní anotaci. Opouští se zde lineární anotace, kdy je každé slovo bráno samostatně bez ohledu na kontext, a do anotace textu se zavádí větná struktura. Všechna původní slova textu zůstávají zachována a dostávají ve výsledné struktuře svou funkci.

Na analytické rovině se vytváří stromová struktura věty (stromem rozumíme orientovaný acyklický graf s jedním kořenem). Uzly stromu jsou tvořeny jednotlivými slovy respektive tokeny. Hrany stromu reprezentují vztahy závislosti. Do kořene stromu je umístěno řídicí sloveso věty, na toto sloveso

se pak zavěšují ostatní slova. V případě, že se jedná o souřadné souvětí, kořenem stromu je spojka případně čárka, která jednotlivé věty souvětí odděluje. Základním cílem je korektní zachycení struktury věty a označení typu závislosti. Typ závislosti je uložen uvnitř lingvistické značky *analytická funkce*.

### Analytická funkce

Analytická funkce je poměrně dobře známý pojem, který se používá na českých základních a středních školách při takzvaném větném rozboru. Tam se ale většinou neoznačuje jako analytická funkce ale jako *větný člen*.

V závislostním stromu analytické roviny anotace, se analytickou funkcí označí každá závislostní hrana. Analytická funkce označuje typ této závislosti.

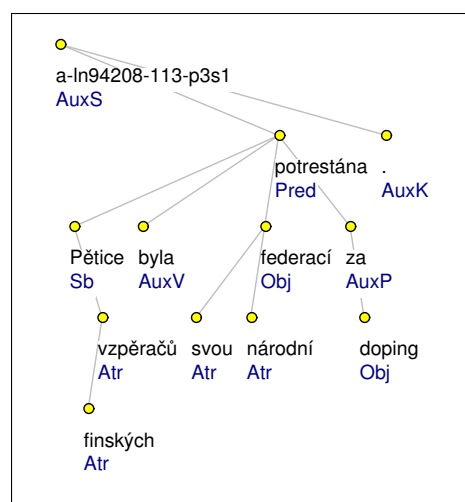
Příklady analytických funkcí:

- Subjekt (podmět)
- Objekt (předmět)
- Atribut (přívlastek)
- Adverbiale (přísluvečné určení)
- ...

### 3.1.3 Tektogramatická rovina

Tektogramatická rovina je poměrně rozsáhlý koncept s hlubokou lingvistickou teorií v pozadí. Byla popsána už v roce 1961 v článku [4].

Tektogramatická rovina slouží k zachycení významové struktury věty. Struktura reprezentace zůstává stejná jako na analytické úrovni, avšak některé uzly se vypouští, některé se přidávají, a struktura věty může být obecně jiná, než na analytické úrovni. U vět které připouštějí více různých významů (víceznačné věty) je teoreticky možné vytvořit více tektogramatických stromů. V případě synonymie může naopak různým větám odpovídat tentýž tektogramatický strom. Tedy zatímco na morfologické rovině jsou každému slovu věty přiřazeny jeho lema a tag (morfologická značka) a na



Obrázek 3.1: Příklad notace na analytické rovině.

analytické rovině každému slovu věty odpovídá uzel v analytickém stromě s příslušnou analytickou funkcí, tektogramatická rovina už tento těsný vztah k povrchovému zápisu věty nemá. Uzly tektogramatické roviny v sobě nesou informaci rozdělenou do několika atributů. Základními atributy uzlu tektogramatického stromu jsou tektogramatické lema, gramatémy a funktor. Vztah mezi uzly tektogramatické a analytické roviny (který je obecně typu M:N), je též zachycen v několika attributech uzlů tektogramatického stromu. Následuje podrobnější popis některých atributů

### Tektogramatické lemma

Tektogramatické lemma (t-lemma) zachycuje lexikální význam uzlu. U jednoduchých uzlů odpovídá lemmatu, které bylo řídícímu slovu tektogramatického uzlu přiřazeno na morfologické rovině. Uzlům na tektogramatické rovině nově vytvořeným je přiřazeno zástupné t-lemma speciálního tvaru.

### Gramatémy

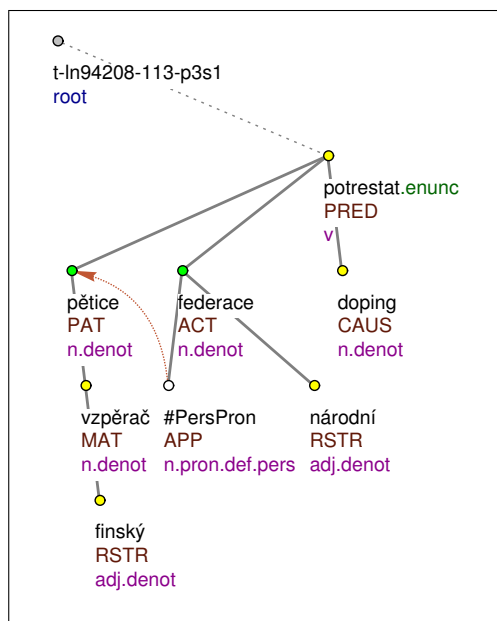
Gramatémy jsou tektogramatickým rozšířením morfologických značek. Gramatémy nalezneme pouze mezi atributy uzlů u kterých to má smysl, tedy u uzlů které se vztahují k nějakému významovému slovu věty.

## Funktor

Funktory jsou velkým přínosem tektogramatické roviny po praktické stránce. Funktory chápeme jako sémantické ohodnocení hran mezi uzly tektogramatického stromu. Tektogramatické funktory můžeme též chápat jako ekvivalent analytických funkcí. Rozdíl mezi tektogramatickými funktory a analytickými funkcemi je v tom, že funktory se snaží postihnout sémantiku vztahu, zatímco analytické funkce se zaměřují na jeho syntaktickou roli. Následuje popis některých důležitých funktorů vždy s několika příklady jejich výskytu ve větě. Vyčerpávající seznam je možné nalézt například v [7].

- Funktor ACT (actor) označuje původce děje, nositele děje nebo vlastnosti.
  - Její *manžel*.ACT tam však pracuje dál.
  - Ten *román*.ACT mě oslovil.
  - Českým *skokanům*.ACT se dařilo dobře.
  - Je *mi*.ACT smutno.
- Funktor ADDR (addressee) odpovídá roli příjemce děje.
  - Dal *dítěti*.ADDR hračku.
  - Učí *děti*.ADDR angličtinu.
  - Obrátil se na *soud*.ADDR s problémem.
- Funktor PAT (patiens) označuje předmět dějem zasažený.
  - Snědl *polévku*.PAT
  - Neubližujte *zvířatům*.PAT
  - Učil se *kominíkem*.PAT
  - Mít dost *peněz*.PAT
- Funktor MANN (manner) vyjadřuje, hodnotí způsob provedení děje.
  - Pracuje *pomalu*.MANN
  - *Nějak*.MANN to uděláme.
  - *Prudce*.MANN se zvýšily mezibankovní úrokové míry.

- Funktor TWHEN (temporal : when) vyjadřuje časové určení odpovídající na otázku "kdy?".
  - *Zítřka*.TWHEN má být už hezky.
  - *Hned*.TWHEN se vrátím.
  - Součástka se *časem*.TWHEN opotřebuje.
- Funktor LOC (locative) označuje místo, do kterého je děj nebo stav vyjádřený řídicím slovem lokalizován.
  - Zůstaň *doma*.LOC
  - *Nalevo*.LOC stál pěkný dům.
  - *Místy*.LOC ležel v ulicích ještě sníh.
- Funktor DIR1 (directional: from) vyjadřuje určení místa odpovídající na otázku "odkud?".
  - Přijel z *Prahy*.DIR1
- Funktor DIR2 (directional: which way) vyjadřuje určení místa odpovídající na otázku "kudy?".
  - Jdou *lesem*.DIR2
- Funktor DIR3 (directional: to) vyjadřuje určení místa odpovídající na otázku "kam?".
  - Přišel *domů*.DIR3
- Funktor RSTR volně doplňuje blíže specifikující řídicí substantivum.
  - *velký*.RSTR dům
- Funktor CONJ (conjunction) je kořen souřadné struktury (tektogramatického podstromu), která reprezentuje spojení dvou a více obsahů.
  - Jezte ovoce *a*.CONJ zeleninu.



Obrázek 3.2: Příklad notace na tektogramatické rovině.

### 3.1.4 Příklady

Pro ilustraci předkládáme dva obrázky závislostních stromů, které vznikli ruční anotací věty:

*Pětice finských vzpěračů byla svou národní federací potrestána za doping.*

Na obrázku 3.1 je strom analytické roviny a na obrázku 3.2 je strom tektogramatické roviny. Za povšimnutí stojí rozdíl v počtu uzlů obou stromů, který je na tektogramatické rovině o něco menší. Oba tyto příklady pocházejí z PDT 2.0 - sample data. Obrázky jsou vygenerované pomocí editoru TrEd. Kromě struktury jednotlivých stromů v této kapitole popisovaných jsou na obrázcích vidět i další především technické jevy spojené s konkrétní reprezentací lingvistické anotace. Například tektogramatické funktoři a analytické funkce nejsou přiřazeny k hranám stromu ale k závislému uzlu, každý strom má navíc takzvaný *technický kořen*, na kterém teprve skutečný kořen lingvistického stromu visí. U analytického stromu též stojí za povšimnutí



umístění tečky na konci věty, která je potomkem umělého kořene. Podrobnější vysvětlení jednotlivých značek na těchto obrázcích je možné nalézt v kapitole 3.5.2 věnované editoru TrEd.

## **3.2 Jazyky pro zápis lingvistických anotací**

### **3.2.1 PML - The Prague Markup Language**

Hlavním formátem dat v PDT 2.0 je formát nazvaný PML, který je založený na XML.

<http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pml/index.html>

### **3.2.2 FS - Feature structure**

Formát FS byl vytvořen pro vyhledávací program Netgraph (přísně vzato vlastně pro jeho předchůdce, editor Graph).

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

### **3.2.3 CSTS - Czech sentence tree structure**

Formát zvaný CSTS, založený na SGML<sup>2</sup>, byl hlavním formátem dat v PDT 1.0. Nyní je používán jen jako přechodný formát pro kompatibilitu se staršími nástroji pro zpracování jazyka (taggery, parsers, ...). CSTS může reprezentovat jen morfologickou a analytickou anotaci, není schopen plného popisu tektogramatické roviny.

---

<sup>2</sup><http://www.w3.org/MarkUp/SGML/>

### 3.2.4 Konverze mezi formáty PDT

## 3.3 The Prague Dependency Treebank

Pražský závislostní korpus (PDT) je probíhající projekt Centra počítačové lingvistiky Ústavu formální a aplikované lingvistiky (ÚFAL) v Praze<sup>3</sup>

Náplní projektu je především ruční lingvistická anotace velkého množství českých textů. Projekt se vyznačuje velkou hloubkou anotace, která sahá až po tektogramatickou rovinou. Kromě velkého množství anotovaných textů bylo v souvislosti s projektem vyvinuto i množství užitečných nástrojů pro práci s anotacemi a nástroje, které umožňují automatickou lingvistickou anotaci českého textu.

dopsat

PDT 1.0

PDT 2.0

## 3.4 WordNet

## 3.5 Lingvistické nástroje

### 3.5.1 NetGraph

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

<http://quest.ms.mff.cuni.cz/netgraph/>

### 3.5.2 Tree Editor TrEd

<http://ufal.mff.cuni.cz/pajas/tred/>

---

<sup>3</sup><http://ufal.mff.cuni.cz/>

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

Editor TrEd umožňuje rychlé, pohodlné a flexibilní procházení, prohlížení a úpravu jednotlivých stromů.

Na obrázku 3.1 je strom analytické roviny a na obrázku 3.2 je strom tektogramatické roviny. Oba tyto příklady

### 3.5.3 Tools for machine annotation - PDT 2.0

Jedná o skupinu nástrojů, které provádějí plně automatickou lingvistickou analýzu českého textu. Ze surových českých vět vytvářejí závislostní stromy na analytické rovině. Proces anotace se skládá z následujících funkcí.

1. Rozpoznání slovních jednotek ve vstupním surovém textu a rozdělení textu na věty.
2. Morfologická analýza a tagging (morfologická desambiguace).
3. Závislostní parsing.
4. Přiřazení analytických (závislostních) funkcí všem uzlům zparsovaného stromu.

Tyto funkce jsou implementovány v celkem šesti oddělených nástrojích. Vstupem každého nástroje je vždy výstup předchozího s výjimkou prvního, jehož vstupem je prostý text. Nástroje jsou napsány z části v Perlu, zbytek tvoří přeložený kód (C++) pro Linux běžící na i386 architektuře.

Celý řetěz nástrojů se dá spustit jediným skriptem *run\_all*.

Tyto nástroje a jejich podrobný popis<sup>4</sup> (včetně naměřené chybovosti) jsou k dispozici jako součást PDT 2.0 CD-ROM.

Následuje podrobnější popis jednotlivých nástrojů.

#### Segmentation and tokenization

Provádí rozdělení textu na slova a interpunkční znaménka (tokenizace) a rozdělí tyto tokeny do vět (segmentace).

---

<sup>4</sup><http://ufal.mff.cuni.cz/pdt2.0/doc/tools/machine-annotation/>

## Morphological analysis

Pro každé slovo vyhledá všechna možná lemmata a morfologické značky, která by mu mohly odpovídat.

## Morphological tagging

Ze všech možných alternativ získaných v předchozím kroku pro každé slovo vybere jedno lemma a morfologickou značku. Tento proces se často nazývá desambiguace. Tagging pro Češtinu je poměrně zajímavý vědecký problém, který je podrobně rozpracován v mnoha publikacích<sup>5</sup>.

## Parsing

Morfologicky označovaná slova v každé větě uspořádá do závislostního stromu. Problém automatického závislostního parsingu<sup>6</sup> je stále poměrně živý. Aktuálně nejlepší parser [8] dosahuje přesnosti přibližně 86%

## Analytical function assignment

Jednotlivým hranám závislostního stromu, které vznikly v předchozím kroku, přiřadí funktory analytické roviny. Nástroj pracuje jako klasifikátor založený na rozhodovacím stromu. Řídící rozhodovací strom byl vytvořen pomocí Quinlanova C5 klasifikátoru z dat PDT 1.0.

## Conversion into PML

Zapiše výstup předchozího nástroje v PML jazyce. Pro podrobnější informace o PML viz oddíl 3.3.

---

<sup>5</sup><http://ufal.mff.cuni.cz/czech-tagging/>

<sup>6</sup><http://ufal.mff.cuni.cz/czech-parsing/>

# Kapitola 4

## Experiment

Experiment provedený v rámci této práce spočíval v otestování některých dostupných nástrojů pro lingvistickou anotaci českých textů. Jmenovitě se jednalo o tyto nástroje: !!!!!!!!!!!!!doplnit!!!!!!!!!!!!!!!. Byly prozkoumány možnosti využití těchto nástrojů pro extrakci informací a sémantickou anotaci.

!!!!!!Pzor následující odstavec jsem okopíroval i do kapitoly přínosy - vyřešit

Postup experimentu se v jednotlivých fázích snaží kopírovat skutečné akce, které by bylo nutné provést v opravdovém projektu zaměřeném na sémantickou anotaci. V práci tak vzniká jednoduchá základní analýza tohoto typu projektů. Ve skutečném projektu pak bude možné ji přinejmenším jako inspiraci využít.

Pro experiment byla vybrána a použita data ze dvou poměrně odlišných zdrojů.

Extrakce a čištění (zamyšlení nad různými formáty zdroje PDF, DOC, HTML, XML, částečné řešení v GATE softu)

Lingvistická nanotace

Extrkace informací

Sémantické anotace - víceméně jen teoreticky

### 4.1 Vstupní data

Volba zdrojových textů

Proč hasiči?

**Proč ne korpus PDT?** — Důraz byl kladen na co možná největší se přiblížení k podmínkám a problémům skutečného projektu. V takovém případě bychom se těžko mohli opřít o to, že by nám data která chceme analyzovat někdo ručně lingvisticky anotoval.

+ pokusy s PDT sample data. Nicméně pokusy nad ručními lingvistickými anotacemi dat PDT<sup>1</sup> proběhly.

#### 4.1.1 Hasiči

#### 4.1.2 Úpadci

### 4.2 Výstupní data

Vzhledem k tomu, že pojem sémantické anotace, jak ho zmiňuji v kapitole 2, je velmi široký, není ani přesně určeno, jaká data by při procesu sémantické anotace měla vzniknout.

### 4.3 Software

#### 4.3.1 Modul XY

---

<sup>1</sup>Jedná se o *sample data* PDT 2.0, <http://ufal.mff.cuni.cz/pdt2.0/data/sample/>

# Literatura

- [1] S. Handschuh, S. Staab (edited by). Annotation for the Semantic Web. Volume 96 Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam, The Netherlands, 2003. ISBN 1-58603-345-x.
- [2] E. Hajičová, M. Plátek, P. Sgall: Komunikace s počítačem v češtině, Sborník referátů seminára SOFSEM 81, Výzkumné výpočtové středisko Bratislava, 1981. 85–114.
- [3] P. Sgall a kolektiv: Úvod do syntaxe a sémantiky, Academia, Praha, 1986.
- [4] Curry, H. B.: Some logical aspects of grammatical structure, Structure of Language and Its Mathematical Aspects (red. R. Jakobson), Proceedings of Symposia in Applied Mathematics 12. American Math. Society, Providence, RI 1961.
- [5] Dan Zeman, Jiří Hana, Hana Hanová, Jan Hajič, Barbora Hladká, Emil Jeřábek. A Manual for Morphological Annotation, 2nd edition. Technical Report 27, ÚFAL MFF UK, Prague, Czech Republic, 2005. URL <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/pdf/m-man-en.pdf>.
- [6] Eva Hajičová, Zdeněk Kirschner, Petr Sgall. A Manual for Analytic Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL, MFF UK, Prague, Czech Republic, 1999. URL <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf>.
- [7] Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářova-Řezníčková, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr

Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský. Anotace Prazžského závislostního korpusu na tekto-gramatické rovině: pokyny pro anotátory [A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank]. Technical report, ÚFAL, MFF UK, Prague, Czech Republic, 2005. URL <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/pdf/t-man-cz.pdf>.

- [8] Daniel Zeman, Zdeněk Žabokrtský. Improving Parsing Accuracy by Combining Diverse Dependency Parsers. In: Proceedings of the International Workshop on Parsing Technologies (IWPT 2005). Association for Computational Linguistics, Vancouver, British Columbia.