

# **Towards Semantic Annotation Supported by Dependency Linguistics and ILP**

Jan Dědek

Department of Software Engineering, Faculty of Mathematics and Physics,  
Charles University in Prague, Czech Republic

ISWC 2010, Nov 9, 2010, Shanghai, China  
Doctoral Consortium

## Outline

### 1 Introduction

- Motivation
- Linguistics we have used
- Overview of the present work

### 2 Our Information Extraction Method

- Manually created rules
- Learning of rules
- Evaluation

### 3 Conclusion

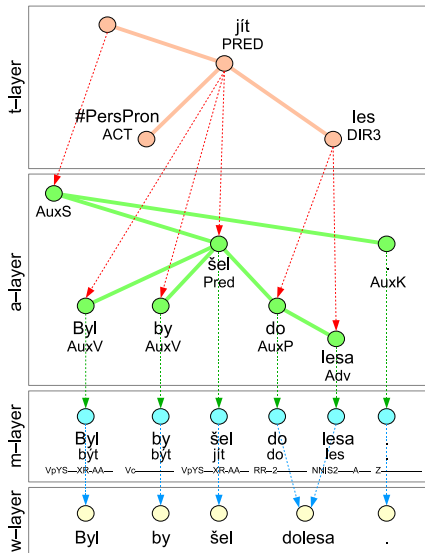
- Summary
- Future work

# Motivation

- ① The goal of the Semantic Web evolution
    - Bring structure to unstructured resources
  - ② Complex linguistic tools
    - The challenge of machine understanding of text
- 
- Using linguistics and Information Extraction for the Semantic Web
    - Semantic annotation

Linguistics we have used

# Layers of linguistic annotation in PDT



- Tectogrammatical layer  
“is supposed to represent the **semantic structure of a sentence**”

- Analytical layer (syntax)
- Morphological layer

- PDT 2.0 on-line:

<http://ufal.mff.cuni.cz/pdt2.0/>

*Sentence:*

Byl by šel dolesa.

He-was would went toforest.

## Tools for machine linguistic annotation

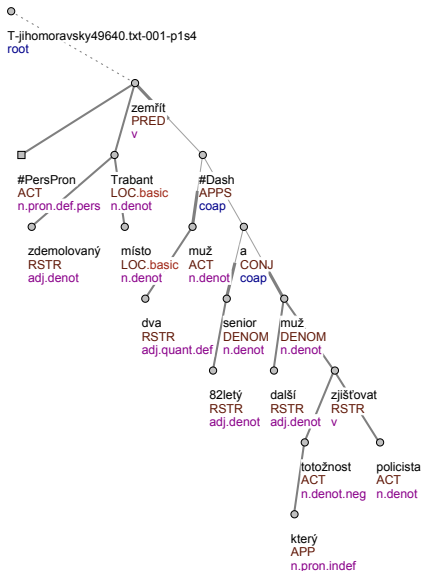
- ① Segmentation and tokenization
  - ② Morphological analysis
  - ③ Morphological tagging
  - ④ McDonald's Maximum Spanning Tree parser
    - Czech adaptation
  - ⑤ Analytical function assignment
  - ⑥ Tectogrammatical analysis
    - Developed by Václav Klimeš
- Available within the **TectoMT**<sup>1</sup> project

---

<sup>1</sup><http://ufal.mff.cuni.cz/tectomt/>

Linguistics we have used

## Example of tectogrammatical tree



- Lemmas
- Functors
- Semantic parts of speech

*Sentence:*

Ve zdemolovaném trabantu na místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost zjišťují policisté.

Two men died on the spot in demolished trabant – ...

## Introduction to the Presented Work

- Extraction of semantic information from **texts**.
  - In Czech language.
  - Coming from web pages.
- Exploiting of linguistic tools.
  - **Prague Dependency Treebank** project.
  - **TectoMT** project (ÚFAL MFF UK).
  - **GATE** project (The University of Sheffield).
  - Experiments with the **Czech WordNet**.
- **Rule based** extraction method.
  - Extraction rules  $\approx$  linguistic **tree queries**
  - **ILP learning** of extraction rules

## Example of processed text – a fire accident

fire

3 amateur units

started at

finished at 4:03

damage 8 000 CZK

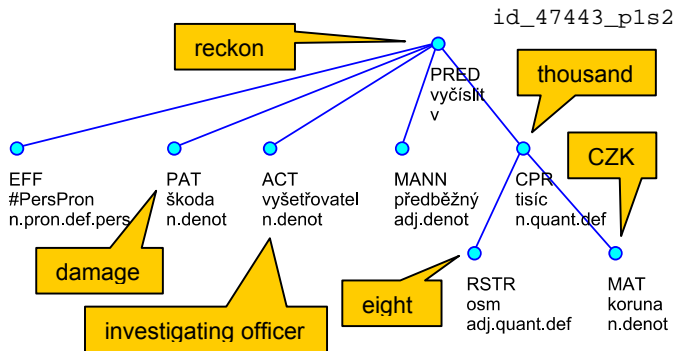
id\_47443

Požár byl operace na střední ŽS ohlášen dnes ve 2.13 hodin, na místo vyjeli profesionální hasiči ze stanice v Židlochovicích a dobrovolní hasiči z Židlochovic, Žabčic a Přísnotic, Oheň, troinstalaci u chladicího boxu, hasiči dostali pod kontrolu ve 2.32 hodin a uhasili tři minuty po třetí hodině. Příčinou vzniku požáru byla technická závada, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

- Information to be extracted is decorated.
- See the last sentence on the next slide.



## Example of a linguistic tree



..., škodu vyšetřovatel předběžně vyčísil na osm tisíc korun.

..., investigating officer preliminarily reckoned the damage to be 8 000 CZK.

- Our IE method uses **tree queries** (tree patterns)

## 1 Introduction

- Motivation
- Linguistics we have used
- Overview of the present work

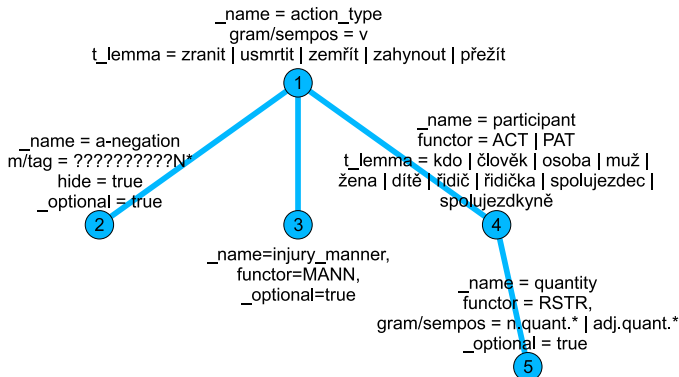
## 2 Our Information Extraction Method

- Manually created rules
- Learning of rules
- Evaluation

## 3 Conclusion

- Summary
- Future work

# Extraction rules – Netgraph queries



- Tree patterns on **shape** and **nodes** (on node attributes).
- Evaluation gives **actual matches** of particular nodes.
- **Names** of nodes allow use of references.

Manually created rules

# Raw data extraction output

```

<QueryMatches>
  <Match root_id="T-vysocina63466.txt-001-pls4" match_string="2:0,7:3,8:4,11:2">
    <Sentence>
      Při požáru byla jedna osoba lehce zraněna - jednalo se
      o majitele domu, který si vykloubil rameno.
    </Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zranit</Value>
      <Value variable_name="injury_manner" attribute_name="t_lemma">lehký</Value>
      <Value variable_name="participant" attribute_name="t_lemma">osoba</Value>
      <Value variable_name="quantity" attribute_name="t_lemma">jeden</Value>
    </Data>
  </Match>
  <Match root_id="T-jihomoravsky49640.txt-001-pls4" match_string="1:0,13:3,14:4">
    <Sentence>
      Ve zdemolovaném trabantu na místě zemřeli dva muži - 82letý senior
      a další muž, jehož totožnost zjišťují policisté.
    </Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zemřít</Value>
      <Value variable_name="participant" attribute_name="t_lemma">muž</Value>
      <Value variable_name="quantity" attribute_name="t_lemma">dva</Value>
    </Data>
  </Match>
  <Match root_id="T-jihomoravsky49736.txt-001-p4s3" match_string="1:0,3:3,7:1">
    <Sentence>Ctyřiatřicetiletý řidič nebyl zraněn.</Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zranit</Value>
      <Value variable_name="a-negation" attribute_name="m/tag">VpYS---XRⓃA---
      </Value>
      <Value variable_name="participant" attribute_name="t_lemma">řidič</Value>
    </Data>
  </Match>
</QueryMatches>

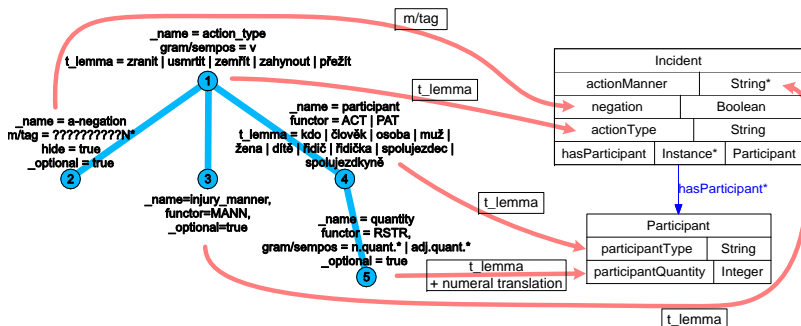
```



SELECT **action\_type.t\_lemma**, **a-negation.mtag**, **injury\_manner.t\_lemma**,  
**participant.t\_lemma**, **quantity.t\_lemma** FROM \*\*\**extraction rule*\*\*\*

Manually created rules

# Semantic interpretation of extraction rules



- Determines how particular values of attributes are used.
- Gives semantics to extraction rule.
- Gives semantics to extracted data.

## 1 Introduction

- Motivation
- Linguistics we have used
- Overview of the present work

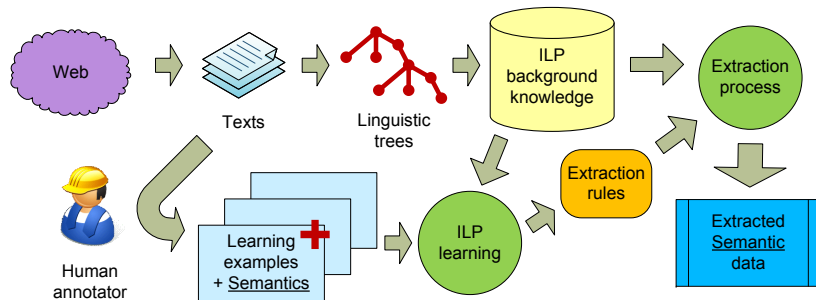
## 2 Our Information Extraction Method

- Manually created rules
- Learning of rules
- Evaluation

## 3 Conclusion

- Summary
- Future work

# Integration of ILP in our extraction process



- Transformation of trees to **logic representation**.

# Logic representation of linguistic trees

**Zpravodajství**

**ČZS Jihoomoravského kraje**

Zpravodajství v noci 2006

15.05.2007

**V trabantu zemřeli dva lidé**

K tragické nehodě dnes dopoledne hasiči vyjžděli na silnici z obce Čoska do Kuliší na Brněnsku.

Nehoda byla opančována zprávkou HZS občasna ve 13.15 hodin. Hasiči jedli se o čelní srážku autobusu Karosa s vozidlem Trabant. Před dostupných informací trabant jedoucí ve zřímě do Kuliší přeměnil na požár, kde narážel do linkového autobusu dopravce společnosti ze Zlína nad Sazavou. Ve zdemolovaném trabantu nalezli zemřeli dva muži – řidič a sedící muž, jehož totožnost zatím nezjistili.

Hasiči ušli na vozidlo protipožární opatření a po vyšetření zdemolovaného trabantu dopravní policie vtrahovala z ulice. Po vyšetření vozidla zjištěno, že řidič byl opilý. Po odstranění střechy trabantu pak byly vyprázdněny dva muži. Obě vozidla – trabant i autobus, po dopravní nehodě byly odvozeny a uvolněny tak jeden jízdní pruh. Uševních kapalin nebyly zjištěny. Po 18. hodině pomohli vrah trabant rozložit k odvozu a asistovali při odtažení autobusu. Po úklidu vozovky nastoupil 16.30 hod. místo nehody odevzdali policistům a kulišské zastávce.

**Logická reprezentace**

... two ...

Source web page

```

tree_root(node0_0). node(node0_0).
id(node0_0, t_jihomoravsky49640_txt_001_pls4).
***** node0_1 *****
node(node0_1).
functor(node0_1, pred).
gram_sempos(node0_1, v).
t_lemma(node0_1, zemrit).
***** node0_2 *****
node(node0_2).
functor(node0_2, act).
gram_sempos(node0_2, n_pron_def_pers).
t_lemma(node0_2, x_perspron).
***** node0_3 *****
node(node0_3). id(node0_3,
functor(node0_3, loc).
gram_sempos(node0_3, n_denot).
t_lemma(node0_3, trabant).
...
edge(node0_0, node0_1). edge(node0_1, node0_2).
edge(node0_1, node0_3). edge(node0_3, node0_4).
edge(node0_4, node0_5). edge(node0_3, node0_6).
edge(node0_3, node0_7). edge(node0_3, node0_8).
...

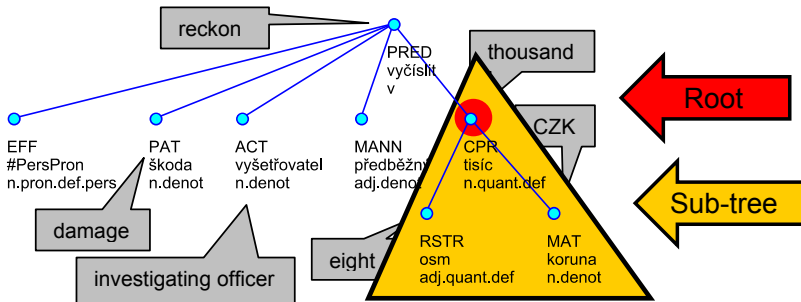
```

**Logic representation**

**Linguistic trees**



# Root/Subtree Preprocessing/Postprocessing (Chunk learning)



..., škodu vyšetřovatel předběžně vyčísлил na osm tisíc korun.

..., investigating officer preliminarily reckoned the damage to be eight thousand Crowns (CZK).

## Examples of learned rules, Czech words are translated.

### Example

[Rule 1] [Pos cover = 14 Neg cover = 0]

```
damage_root(A) :- lex_rf(B,A), has_sempos(B,'n.quant.def'),
    tDependency(C,B), tDependency(C,D),
    has_t_lemma(D,'investigator').
```

[Rule 2] [Pos cover = 13 Neg cover = 0]

```
damage_root(A) :- lex_rf(B,A), has_functor(B,'TOWH'),
    tDependency(C,B), tDependency(C,D), has_t_lemma(D,'damage').
```

[Rule 1] [Pos cover = 7 Neg cover = 0]

```
injuries(A) :- lex_rf(B,A), has_functor(B,'PAT'),
    has_gender(B,anim), tDependency(B,C), has_t_lemma(C,'injured').
```

[Rule 8] [Pos cover = 6 Neg cover = 0]

```
injuries(A) :- lex_rf(B,A), has_gender(B,anim), tDependency(C,B),
    has_t_lemma(C,'injure'), has_negation(C,neg0).
```

## 1 Introduction

- Motivation
- Linguistics we have used
- Overview of the present work

## 2 Our Information Extraction Method

- Manually created rules
- Learning of rules
- Evaluation

## 3 Conclusion

- Summary
- Future work

## Evaluation results

| task/method                          | matching | missing | excess | overlap | prec.% | recall% | F1.0%        |
|--------------------------------------|----------|---------|--------|---------|--------|---------|--------------|
| <b>damage/ILP</b>                    | 14       | 0       | 7      | 6       | 51.85  | 70.00   | 59.57        |
| <b>damage/ILP – lenient measures</b> |          |         |        |         | 74.07  | 100.00  | 85.11        |
| <b>dam./ILP-roots</b>                | 16       | 4       | 2      | 0       | 88.89  | 80.00   | <b>84.21</b> |
| <b>damage/Paum</b>                   | 20       | 0       | 6      | 0       | 76.92  | 100.00  | 86.96        |
| <b>injuries/ILP</b>                  | 15       | 18      | 11     | 0       | 57.69  | 45.45   | <b>50.85</b> |
| <b>injuries/Paum</b>                 | 25       | 8       | 54     | 0       | 31.65  | 75.76   | 44.64        |
| <b>inj./Paum-afun</b>                | 24       | 9       | 38     | 0       | 38.71  | 72.73   | 50.53        |

- 10-fold cross validation
- Two tasks: ‘damage’ and ‘injuries’
- Root/subtree preprocessing/postprocessing used for ‘damage’ task

## 1 Introduction

- Motivation
- Linguistics we have used
- Overview of the present work

## 2 Our Information Extraction Method

- Manually created rules
- Learning of rules
- Evaluation

## 3 Conclusion

- Summary
- Future work

# Summary

- Implemented a system for extraction of semantic information
- Based on third party linguistic tools (**TectoMT**<sup>2</sup>)
- Extraction rules adopted from **Netgraph**<sup>3</sup> application.
- **ILP** used for learning rules.
- All methods integrated inside **GATE**<sup>4</sup>.
- Main advantages:
  - Automated selection of learning features
  - “Language independent”
  - Rule based

---

<sup>2</sup><http://ufal.mff.cuni.cz/tectomt/>

<sup>3</sup><http://quest.ms.mff.cuni.cz/netgraph/>

<sup>4</sup><http://gate.ac.uk/>

## Future work

- Use some **Knowledge Base** (e.g. WordNet).
- Adaptation of this method on **other languages**.
- Evaluation of the method on **other datasets**.
- Be able to provide **more semantics**.
  - e.g. sophisticated semantic interpretation of extracted data