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

Uživatelsko-webový seminář, 6. 10. 2010, MFF UK, Praha

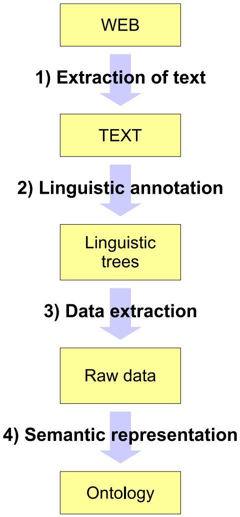
Outline

1. **Introduction**
   - Our Information Extraction System
   - Linguistics we have used
   - Domain of fire-department articles
2. **Our Information Extraction Method**
   - Manually created rules
   - Learning of rules
   - Evaluation
3. **Conclusion**

Introduction to Presented Work

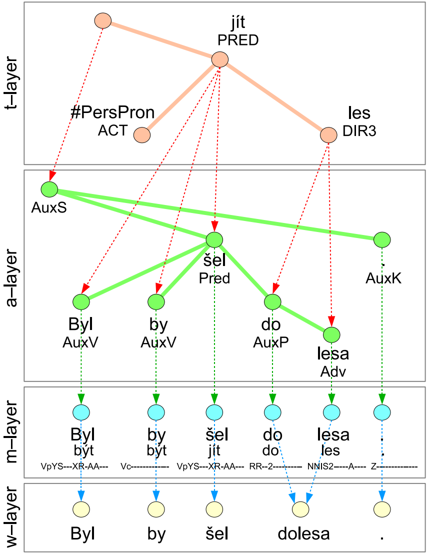
- Extraction of semantic information from texts.
  - In Czech language.
  - Coming from web pages.
- Using of Semantic Web ontologies.
  - RDF, OWL
- Exploiting of linguistic tools.
  - Mainly from the Prague Dependency Treebank project.
  - Experiments with the Czech WordNet.
- Rule based extraction method.
  - Extraction rules ≈ tree queries
  - ILP learning of extraction rules

Schema of the extraction process



1. Extraction of text
   - Using RSS feed to download pages.
   - Regular expression to extract text.
2. Linguistic annotation
   - Using chain of 6 linguistic tools (see on next slides).
3. Data extraction
   - Exploitation of linguistic trees.
   - Using extraction rules.
4. Semantic representation of data
   - Ontology needed.
   - Semantic interpretation of rules.
   - Far from finished in current state.

Layers of linguistic annotation in PDT



- Tectogrammatical layer
- Analytical layer
- Morphological layer

- PDT 2.0 on-line:

`http://ufal.mff.cuni.cz/pdt2.0/`

Sentence:

Byl by šel dolesa.
He-was would went toforest.

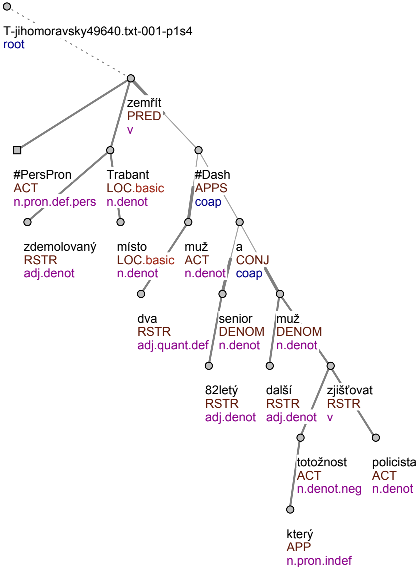
Tools for machine linguistic annotation

1. Segmentation and tokenization
2. Morphological analysis
3. Morphological tagging
4. McDonnald's Maximum Spanning Tree parser
   – Czech adaptation
5. Analytical function assignment
6. Tectogrammatical analysis
   – Developed by Václav Klimeš

- Available within the TectoMT[1] project

[1] http://ufal.mff.cuni.cz/tectomt/

Example of tectogrammatical tree



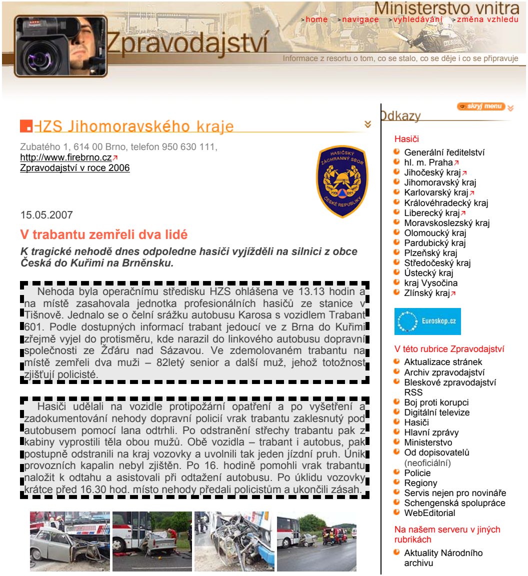
- Lemmas
- Functors
- Semantic parts of speech

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 – . . .

Example of the web-page with a report of a fire department



Ministerstvo vnitra

>home >navigace >vyhledávání >změna vzhledu

Zpravodajství

Informace z resortu o tom, co se stalo, co se děje i co se připravuje

HZS Jihomoravského kraje

Zubatého 1, 614 00 Brno, telefon 950 630 111,
http://www.firebrno.cz
Zpravodajství v roce 2006

15.05.2007

V trabantu zemřeli dva lidé

K tragické nehodě dnes odpoledne hasiči vyjížděli na silnici z obce Česká do Kuřimi na Brněnsku.

Nehoda byla operačnímu středisku HZS ohlášena ve 13.13 hodin a na místě zasahovala jednotka profesionálních hasičů ze stanice v Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 601. Podle dostupných informací trabant jedoucí ve z Brna do Kuřimi zřejmě vyjel do protisměru, kde narazil do linkového autobusu dopravní společnosti ze Žďáru nad Sázavou. Ve zdemolovaném trabantu na místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost zjišťují policisté.

Hasiči udělali na vozidle protipožární opatření a po vyšetření a zadokumentování nehody dopravní policií vrak trabantu zaklesnutý pod autobusem pomocí lana odtrhli. Po odstranění střechy trabantu pak z kabiny vyprostili těla obou mužů. Obě vozidla – trabant i autobus, pak postupně odstranili na kraj vozovky a uvolnili tak jeden jízdní pruh. Únik provozních kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabantu naložit k odtahu a asistovali při odtažení autobusu. Po úklidu vozovky krátce před 16.30 hod. místo nehody předali policistům a ukončili zásah.

Odkazy

skrýj menu

Hasiči

- Generální ředitelství
- hl. m. Praha
- Jihočeský kraj
- Jihomoravský kraj
- Karlovarský kraj
- Královéhradecký kraj
- Liberecký kraj
- Moravskoslezský kraj
- Olomoucký kraj
- Pardubický kraj
- Plzeňský kraj
- Středočeský kraj
- Ústecký kraj
- kraj Vysočina
- Zlínský kraj

Euroskop.cz

V této rubrice Zpravodajství

- Aktualizace stránek
- Archiv zpravodajství
- Bleskové zpravodajství RSS
- Boj proti korupci
- Digitální televize
- Hasiči
- Hlavní zprávy
- Ministerstvo
- Od dopisovatelů (neoficiální)
- Policie
- Regiony
- Servis nejen pro novináře
- Schengenská spolupráce
- WebEditorial

Na našem serveru v jiných rubrikách

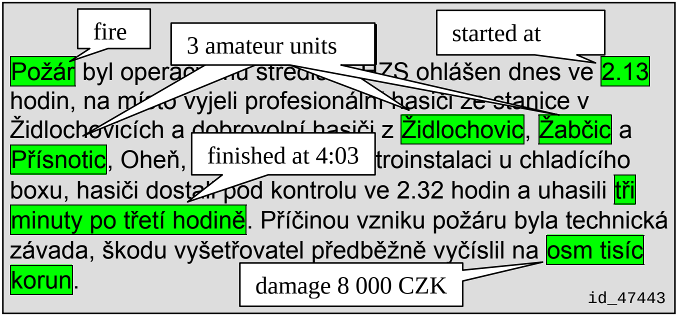
- Aktuality Národního archivu

Domain of our experiments

- Fire-department articles
- Published by The Ministry of Interior of the Czech Republic[2]
- Processed more than 800 articles
  from different regions of Czech Republic
- 1.2 MB of textual data
- Linguistic tools produced 10 MB of annotations,
  run time 3.5 hours
- Extracting information about injured and killed people
- 470 matches of the extraction rule,
  200 numeric values of quantity (described later)

[2] `http://www.mvcr.cz/rss/regionhzs.html`

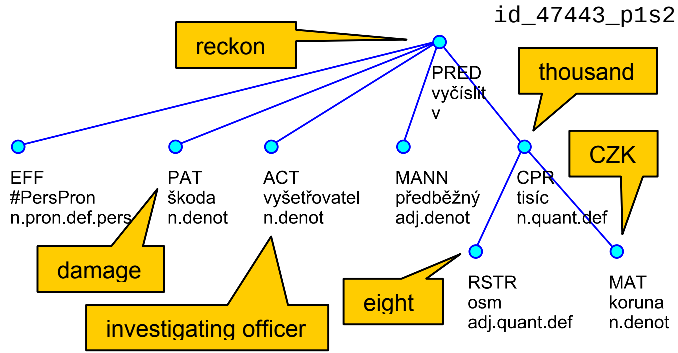
Example of processed text



Požár byl operac nu stredi ZS ohlášen dnes ve 2.13 hodin, na m to vyjeli profesionální hasiči ze stanice v Židlochovicích a dobrovolní hasiči z Židlochovic, Žabčic a Přísnotic, Oheň, troinstalaci u chladícího boxu, hasiči dosta 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

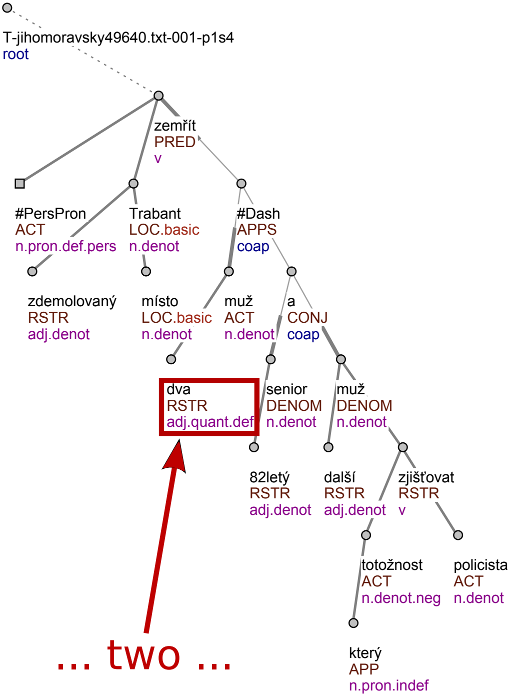


…, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

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

- Our IE method uses tree queries (tree patterns)

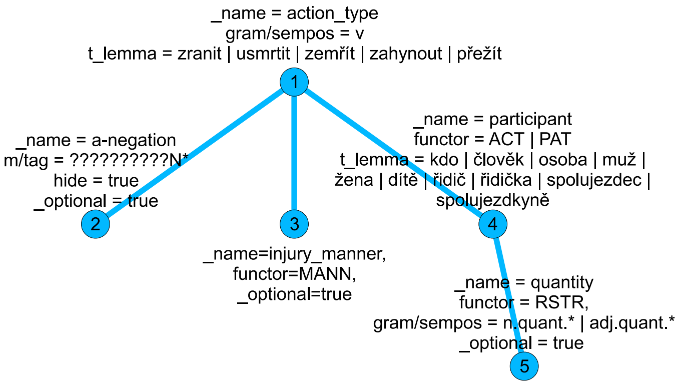
1. **Introduction**
   - Our Information Extraction System
   - Linguistics we have used
   - Domain of fire-department articles
2. **Our Information Extraction Method**
   - Manually created rules
   - Learning of rules
   - Evaluation
3. **Conclusion**



... two ...

- How to extract the information about two dead people?

Extraction rules – Netgraph queries



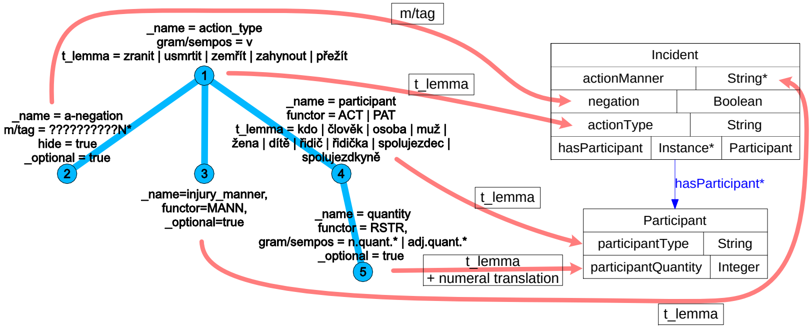
- Tree patterns on shape and nodes (on node attributes).
- Evaluation gives actual matches of particular nodes.
- Names of nodes allow use of references.

Raw data extraction output

```
<QueryMatches>
  <Match root_id="T-vysocina63466.txt-001-p1s4" 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-p1s4" 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>Čtyřiatřicetiletý 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-NA---
      </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****

Semantic interpretation of extraction rules



_name = action_type
gram/sempos = v
t_lemma = zranit | usmrtit | zemřít | zahynout | přežít

_name = a-negation
m/tag = ??????????N*
hide = true
_optional = true

_name=injury_manner,
functor=MANN,
_optional=true

_name = participant
functor = ACT | PAT
t_lemma = kdo | člověk | osoba | muž |
žena | dítě | řidič | řidička | spolujezdec |
spolujezdkyně

_name = quantity
functor = RSTR,
gram/sempos = n.quant.* | adj.quant.*
_optional = true

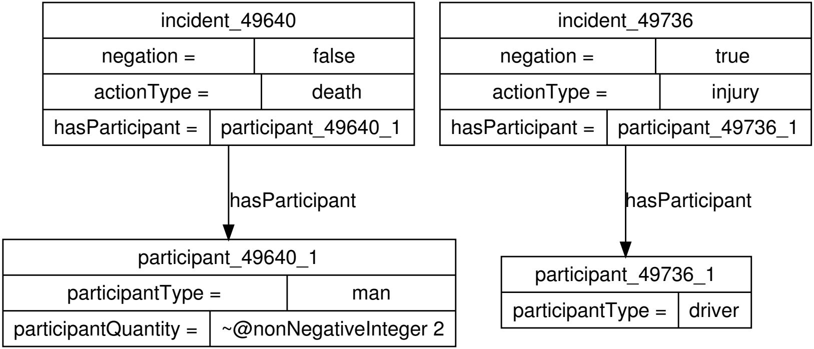
Incident		
actionManner	String*	
negation	Boolean	
actionType	String	
hasParticipant	Instance*	Participant

hasParticipant*

Participant	
participantType	String
participantQuantity	Integer

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

Semantic data output



incident_49640
negation = false
actionType = death
hasParticipant = participant_49640_1

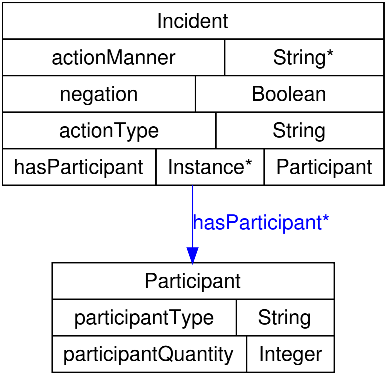
incident_49736
negation = true
actionType = injury
hasParticipant = participant_49736_1

participant_49640_1
participantType = man
participantQuantity = ~@nonNegativeInteger 2

participant_49736_1
participantType = driver

- Two instances of two ontology classes.

The experimental ontology

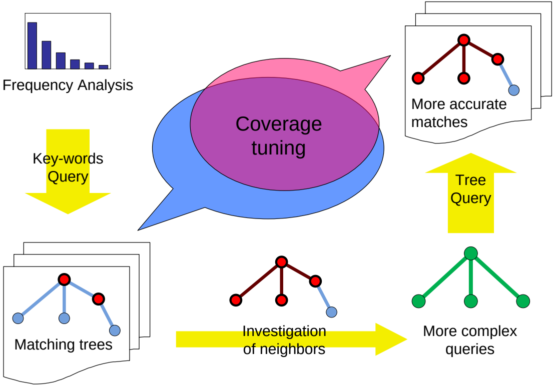


Incident		
actionManner	String*	
negation	Boolean	
actionType	String	
hasParticipant	Instance*	Participant

Participant	
participantType	String
participantQuantity	Integer

- Two classes
  - Incident and Participant
- One object property relation
  - hasParticipant
- Five datatype property relations
  - actionManner (light or heavy injury)
  - negation
  - actionType (injury or death)
  - participantType (man, woman, driver, etc.)
  - participantQuantity

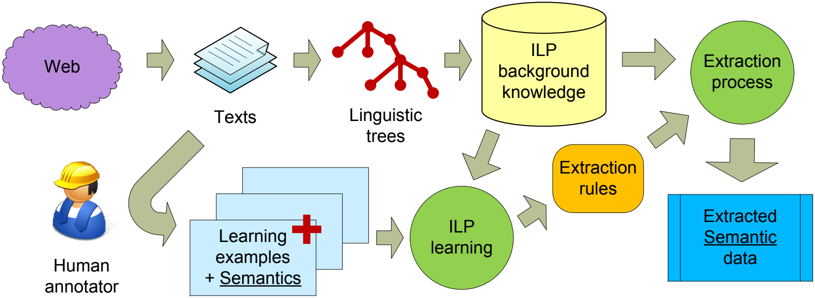
Design of extraction rules – iterative process



1. Frequency analysis → representative key-words.
2. Investigating of matching trees → tuning of tree query.
3. Complexity of the query ≅ complexity of extracted data.

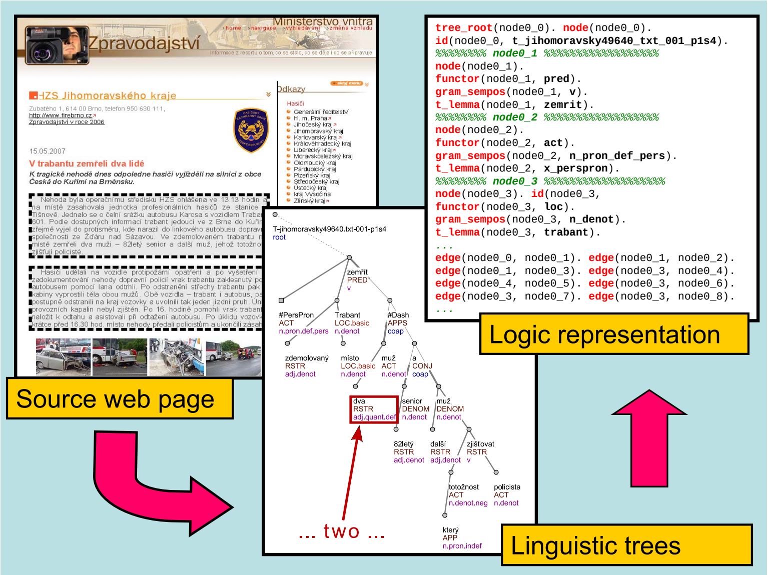
1. **Introduction**
   - Our Information Extraction System
   - Linguistics we have used
   - Domain of fire-department articles
2. **Our Information Extraction Method**
   - Manually created rules
   - Learning of rules
   - Evaluation
3. **Conclusion**

Integration of ILP in our extraction process



- Transformation of trees to logic representation.
- Today: just first promising experiments.

Logic representation of linguistic trees



Ministerstvo vnitra
› home › navigace › vyhledávání › změna vzhledu
Zpravodajství
Informace z resortu o tom, co se stalo, co se děje i co se připravuje

HZS Jihomoravského kraje

Zubatého 1, 614 00 Brno, telefon 950 630 111,
http://www.firebrno.cz
Zpravodajství v roce 2006

15.05.2007

V trabantu zemřeli dva lidé

K tragické nehodě dnes odpoledne hasiči vyjížděli na silnici z obce Česká do Kuřimi na Brněnsku.

Nehoda byla operačnímu středisku HZS ohlášena ve 13.13 hodin a na místě zasahovala jednotka profesionálních hasičů ze stanice Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabar 601. Podle dostupných informací trabant jedoucí ve z Brna do Kuřin zřejmě vyjel do protisměru, kde narazil do linkového autobusu dopravr společnosti ze Žďáru nad Sázavou. Ve zdemolovaném trabantu n místě zemřeli dva muži – 82letý senior a další muž, jehož totožnos zjišťují policisté.

Hasiči udělali na vozidle protipožární opatření a po vyšetření zadokumentování nehody dopravní policií vrak trabantu zaklesnutý po autobusem pomocí lana odtrhli. Po odstranění střechy trabantu pak kabiny vyprostili těla obou mužů. Obě vozidla – trabant i autobus, pa postupně odstranili na kraj vozovky a uvolnili tak jeden jízdní pruh. Ún provozních kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabant naložit k odtahu a asistovali při odtažení autobusu. Po úklidu vozovk krátce před 16.30 hod. místo nehody předali policistům a ukončili zásah

Odkazy
skryj menu

Hasiči
- Generální ředitelství
- hl. m. Praha
- Jihočeský kraj
- Jihomoravský kraj
- Karlovarský kraj
- Královéhradecký kraj
- Liberecký kraj
- Moravskoslezský kraj
- Olomoucký kraj
- Pardubický kraj
- Plzeňský kraj
- Středočeský kraj
- Ústecký kraj
- kraj Vysočina
- Zlínský kraj

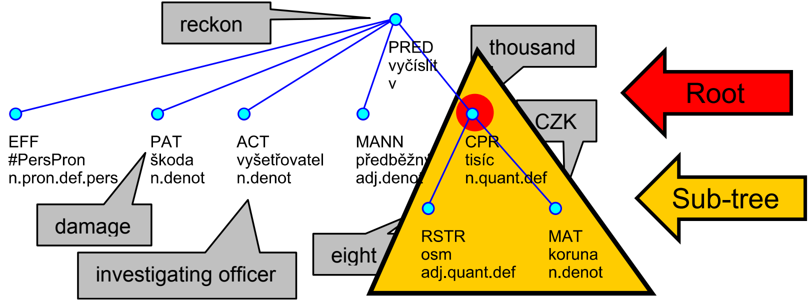
Source web page

Linguistic trees

```
tree_root(node0_0). node(node0_0).
id(node0_0, t_jihomoravsky49640_txt_001_p1s4).
%%%%%%%% 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

Root/Subtree Preprocessing/Postprocessing (Chunk learning)



…, škodu vyšetřovatel předběžně vyčíslil 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**
   - Our Information Extraction System
   - Linguistics we have used
   - Domain of fire-department articles
2. **Our Information Extraction Method**
   - Manually created rules
   - Learning of rules
   - Evaluation
3. **Conclusion**

Evaluation results

task/method	matching	missing	excess	overlap	prec.%	recall%	F1.0%
damage/ILP	14	0	7	6	51.85	70.00	59.57
damage/ILP – lenient measures					74.07	100.00	85.11
dam./ILP-roots	16	4	2	0	88.89	80.00	84.21
damage/Paum	20	0	6	0	76.92	100.00	86.96
injuries/ILP	15	18	11	0	57.69	45.45	50.85
injuries/Paum	25	8	54	0	31.65	75.76	44.64
inj./Paum-afun	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
   - Our Information Extraction System
   - Linguistics we have used
   - Domain of fire-department articles
2. Our Information Extraction Method
   - Manually created rules
   - Learning of rules
   - Evaluation
3. **Conclusion**

Summary

- Proposed a system for extraction of semantic information
- Based on linguistic tools for automatic text annotation
- Extraction rules adopted from Netgraph application.
- ILP used for learning rules.
- All methods integrated inside GATE[3].

- Our future research will concentrate on:
  - Learning of extraction rules.
  - Extension of the method with WordNet technology.
  - Adaptation of this method on other languages.
  - Evaluation of the method.

[3] http://gate.ac.uk/