Linguistic extraction for semantic annotation

Jan Dědek[1] Peter Vojtáš[1,2]

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

[2] Institute of Computer Science, Academy of Sciences of the Czech Republic

2nd International Symposium on Intelligent Distributed Computing, Catania, Italy, September 18-19, 2008

Outline

Information Extraction and the Semantic Web

- Information Extraction
  - Automatically find the information you're looking for.
  - Pick out the most useful bits.
  - Present it in preferred manner, at the right level of detail.
- Semantic Web
  - Web as universal medium for the exchange of information.
  - Not only for humans but also for software agents.
  - Main problem today: lack of semantic data on the Web.
- Extraction of information for the Semantic Web
  - Let's use information extraction to produce semantic data.

What about the data semantics?

We use semantic web ontologies to express the semantics.

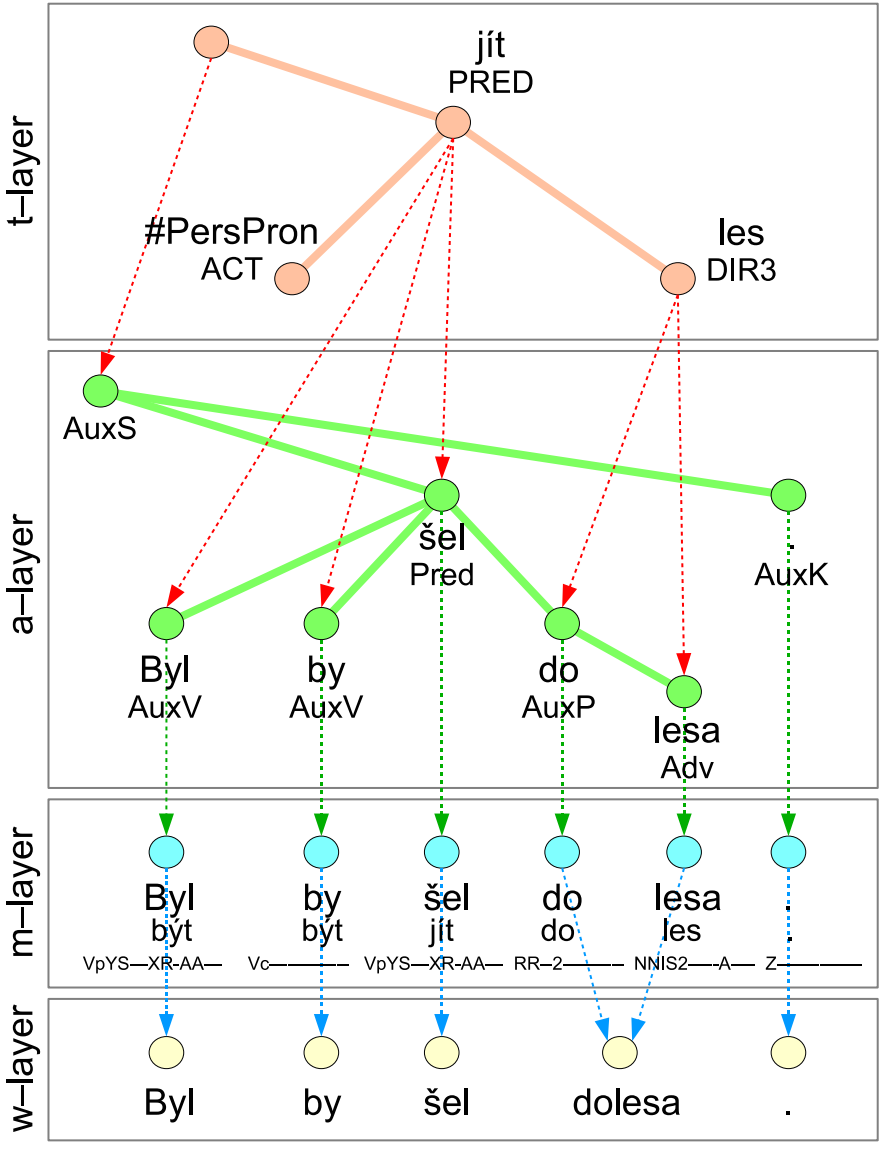
- RDF, OWL
- Motivated by description logics

- Concepts or Classes
- Predicates or Relations
- Individuals or Instances

Our work

- Extraction of semantic information form texts.
  - In Czech language.
  - Coming form web pages.
- Exploiting of linguistic tools.
  - Mainly from the Prague Dependency Treebank project.
  - Experiments with the Czech WordNet.
- Rule based extraction method.
  - Extraction rules $\approx$ tree queries of Netgraph application

Layers of linguistic annotation in PDT



- Tectogrammatical layer
- Analytical layer
- Morphological layer

Sentence:

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

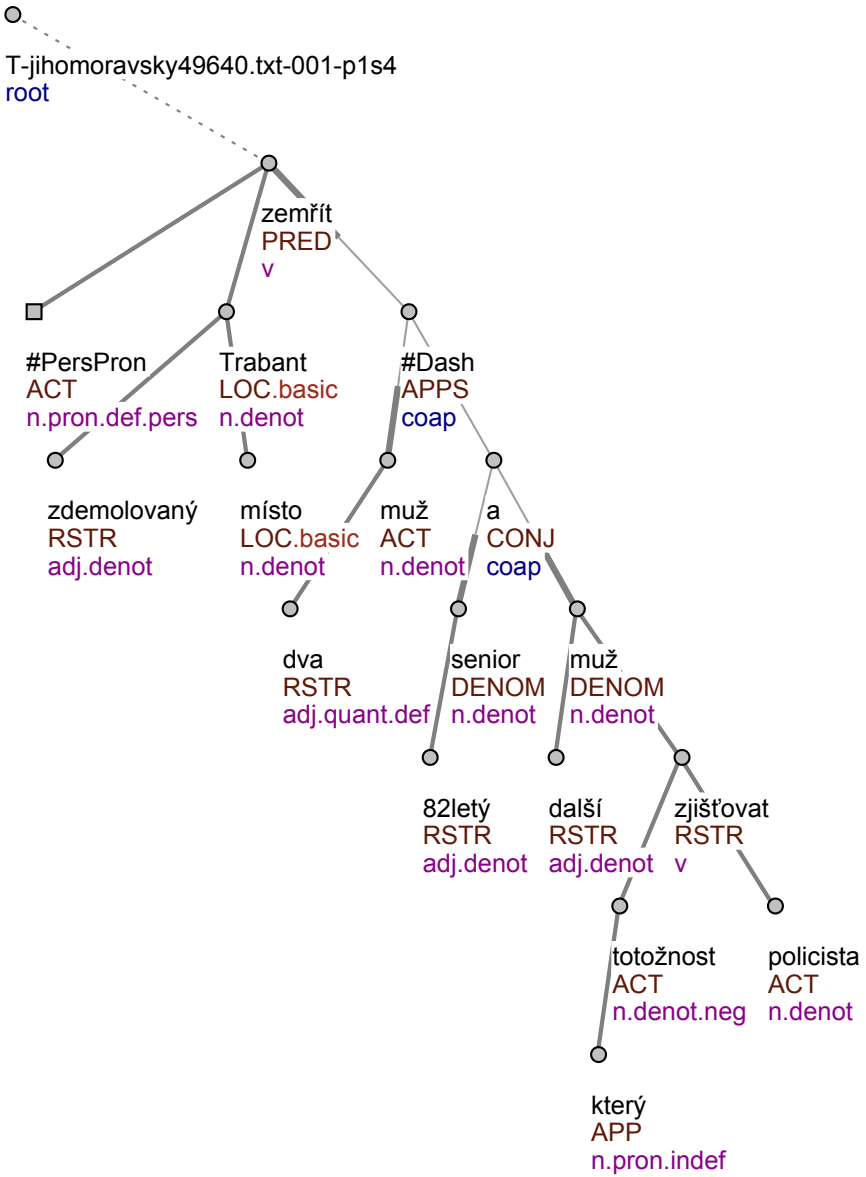
Tools for machine linguistic annotation

Available on the PDT 2.0 CD-ROM

1. Segmentation and tokenization
2. Morphological analysis
3. Morphological tagging
4. Collins' parser – Czech adaptation
5. Analytical function assignment

6. Tectogrammatical analysis
   – Developed by Václav Klimeš

Example of tectogrammatical tree



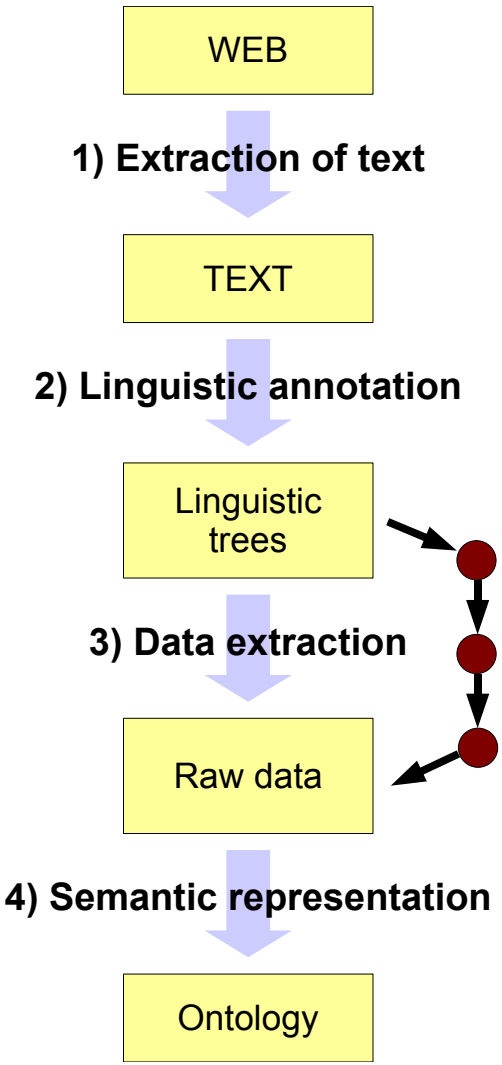
- Lemmas
- Functors
- Semantic parts of speech

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 – . . .

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 (already mentioned).
3. Data extraction
   - Made over tectogrammatical trees.
   - Supported by extraction rules.
4. Semantic representation of data
   - First step – choosing of ontology.
   - Supported by:
     - semantic interpretation of rules
     - or by additional data transformation
   - Not implemented yet :-(

Domain of our experiments

- Fire-department articles
- Published by The Ministry of Interior of the Czech Republic[1]
- 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)

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

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

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

skrýt menu

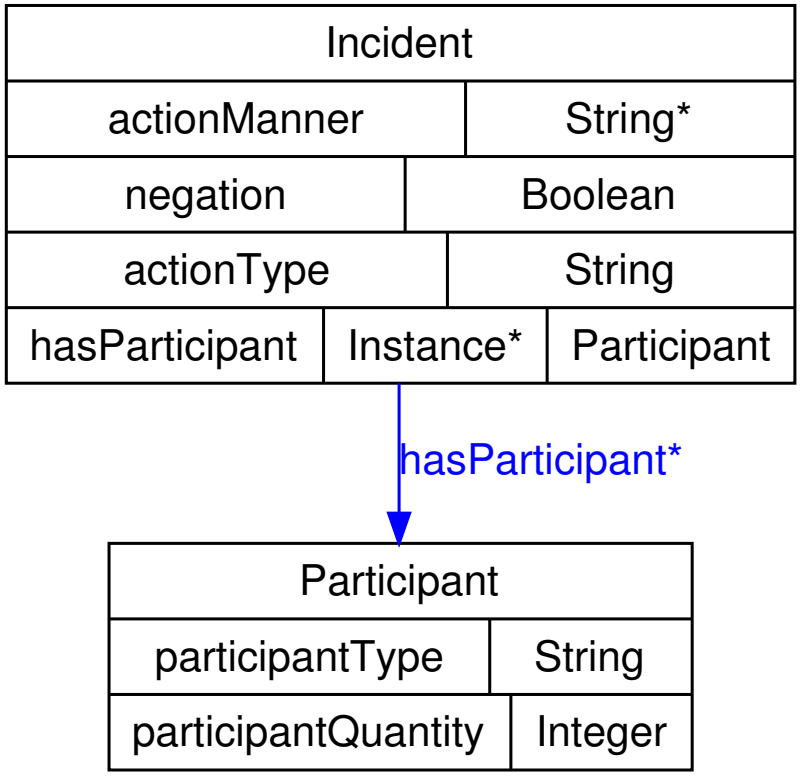
Hasiči
- Generální ředitelství
- hl. m. Praha
- Jihočeský kraj
- Jihomoravský kraj
- Karlovarský kraj
- Královéhradecký kraj
- Liberecký kraj
- Moravskoslezský kraj
- Olomoucký kraj
- Pardubický kraj
- Plzeňský kraj
- Středočeský kraj
- Ústecký kraj
- kraj Vysočina
- Zlínský kraj

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

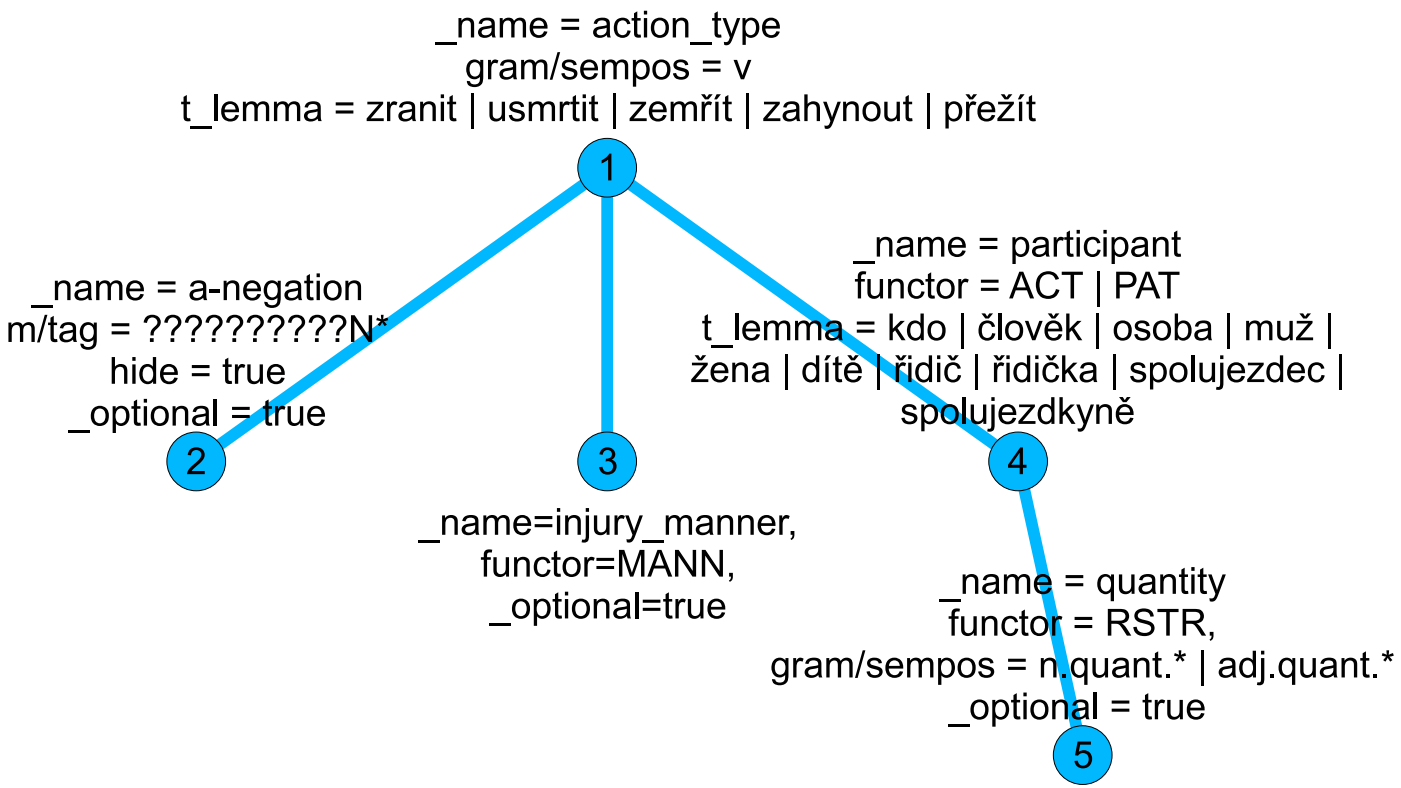
Na našem serveru v jiných rubrikách
- Aktuality Národního archivu

The experimental ontology



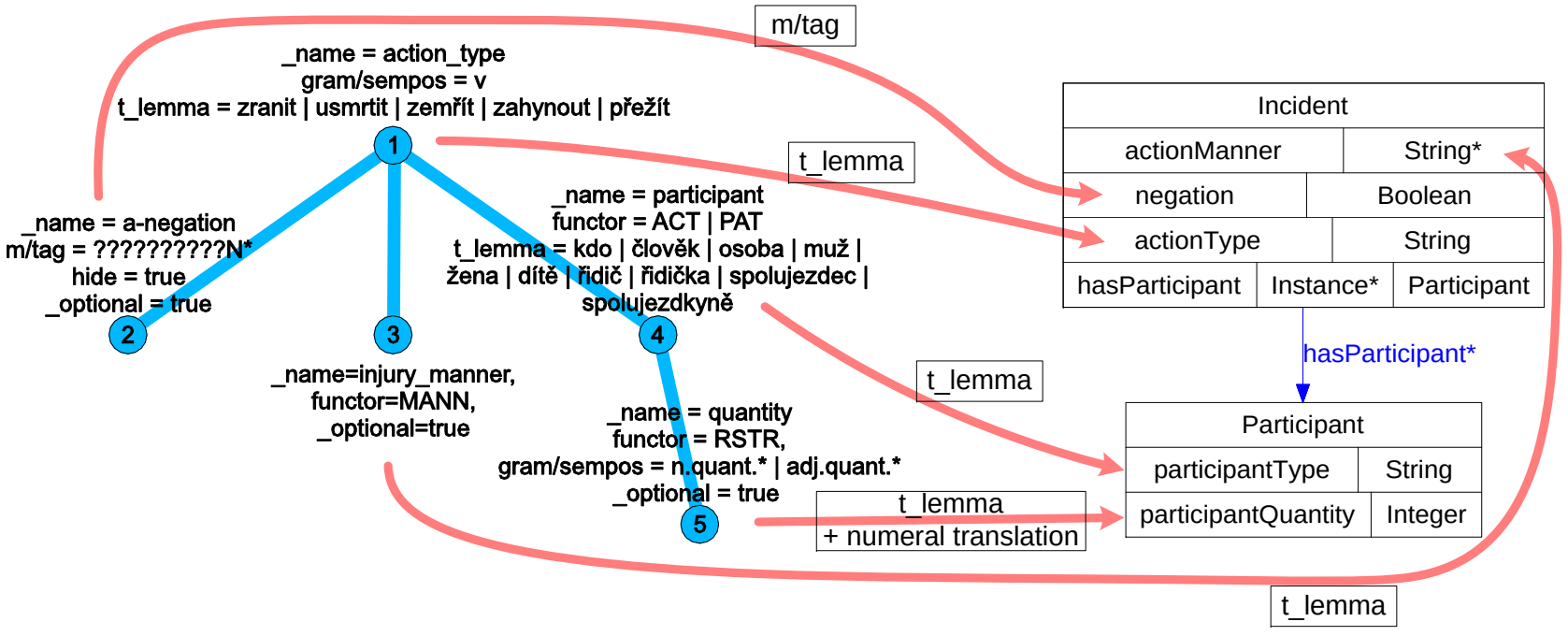
- Two classes
  - Incident and Participant
- One object property relation
  - hasParticipant
- Five datatype property relations
  - actionManner (light or heavy injury)
  - negation
  - actionType (injury or death)
  - participantType (man, woman, driver, etc.)
  - participantQuantity

Extraction rules – Netgraph queries



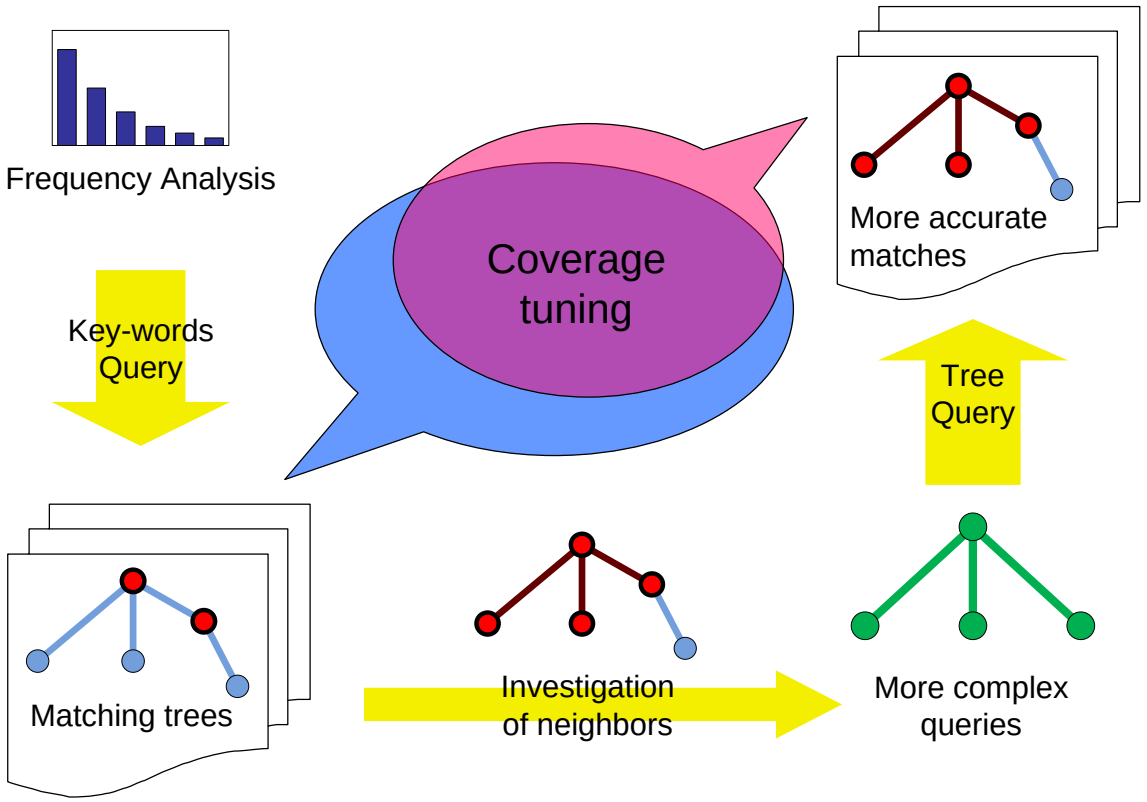
- Tree patterns on shape and nodes (on node attributes).
- Evaluation gives actual matches of particular nodes.
- Names of nodes allow use of references.

Semantic interpretation of extraction rules



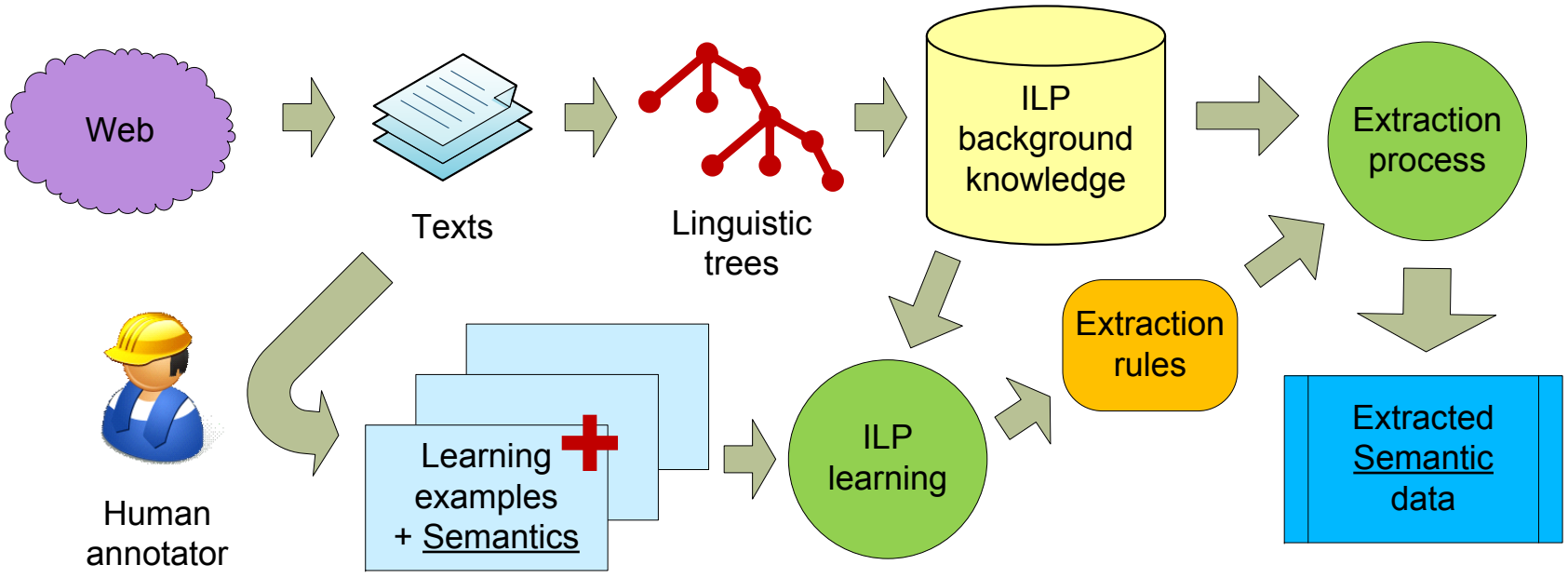
- Determines how particular values of attributes are used.
- Gives semantics to extraction rule.
- Gives semantics to extracted data.

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.

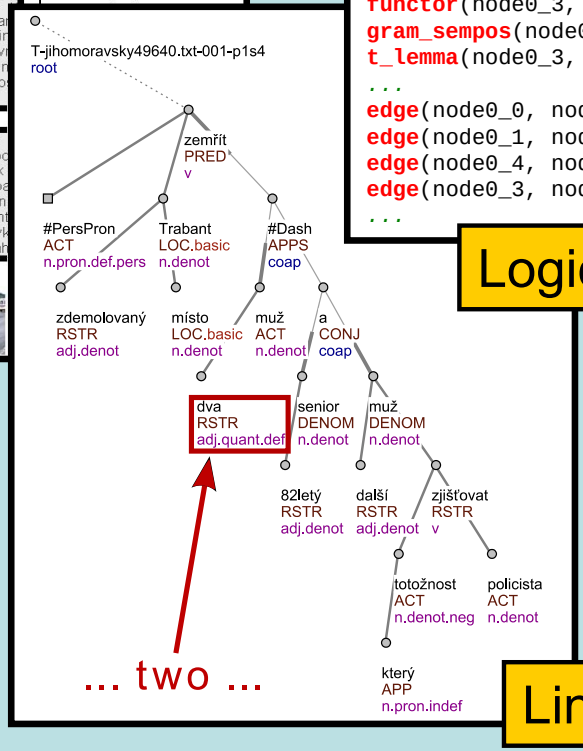
Integration of ILP in our extraction process



- Today: just first experiments.
- Transformation of trees to logic representation.
- Not presented in this paper.
- Published as late breaking paper at ILP2008 conference.

Logic representation of linguistic trees

Source web page

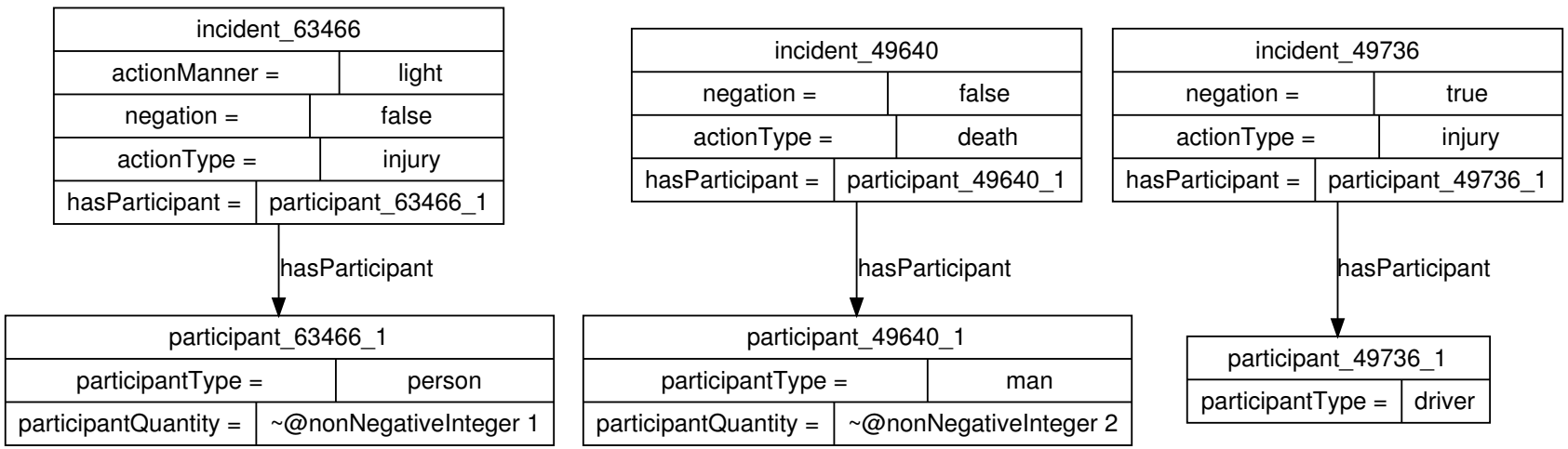


Linguistic trees

```
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

Semantic data output



- Three instances of two ontology classes.

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-N 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****

Summary

- Proposed a system for extraction of semantic information
- Based on linguistic tools for machine annotation
- Extraction rules adopted from Netgraph application.
- 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.