Prague Dependency Treebanks
A Family of Annotated Treebanks

Jiří Mírovský

Jarmila Panevová

Charles University in Prague

Institute of Formal and Applied Linguistics

Prague Dependency Treebanks
A Family of Annotated Treebanks

- Why "A Family"?
  - not a parallel corpus in the proper sense (with one exception)
- How to use it for building parallel corpora?
  - annotation scenario developed and used since 1996

ÚFAL

Prague Dependency Treebank 2.0 (PDT 2.0)

- published by LDC in 2006 (registration required)
- **(mostly) manually annotated** texts from 1990's
  - 40 % general newspaper articles
  - 20 % economic news and analyses
  - 20 % articles from a popular science magazine
  - 20 % information technology texts
- The annotation scenario was developed between 1996–2000.

ÚFAL

Prague Dependency Treebank 2.0
Three Layers of Annotation

- **morphological** layer (100 % of data)
  - 7,110 documents, 115,844 sentences, 1,957,247 tokens
- **analytical** (surface syntax) layer (75 % of data)
  - 5,330 documents, 87,913 sentences, 1,503,739 tokens
- **tectogrammatical** (deep syntax) layer (45 % of data)
  - 3,165 documents, 49,431 sentences, 833,195 tokens

ÚFAL

The Annotation Scenario

- **Dependency approach** – tree structures representing relations between governing (parent) node and its children nodes as governed by it. The nodes are labeled by their respective functions
- **Analytical layer**
  - surface syntactic functions – Pred, Sb, Obj, Adv, Atr, ...
  - special convention for coordination, apposition and parenthesis
  - The synsemantic words have their representation as special nodes (Prep, Conj, AuxV, graphic symbols as commas etc. have their own node).

The Annotation Scenario

- **Tectogrammatical layer**
  - Semantic labels (as counterparts of syntactic functions on the analytical layer) called **functors** are more grained, approx. 44 labels.
  - Some of them are further specified by **subfunctors**, e.g. types of local modification (functor LOC "where" is divided according to the position of the referred object by subfunctors *in, behind, under, along* etc.).

The Annotation Scenario

- **Tectogrammatical layer (cont.)**
  - conventions for **coordination/apposition** – represented by tree edges although not expressing dependency
    - algorithm for determining the proper parent – child relation (effective parentage)
  - **topic-focus articulation** – (contrastively) contextually bound and contextually non-bound elements, comunicative dynamism expressed in node order

The Annotation Scenario

- **Tectogrammatical layer (cont.)**
  - **coreference**
    - grammatical (given by grammatical rules)
    - pronominal textual

Prague Dependency Treebank Updates

- PDT 1.0 – published in 2001 (LDC)
- PDT 2.0 – published in 2006 (LDC)
- **PDT 2.5** – published in 2012 (downloadable from Lindat/Clarin repository, Creative Commons License)
- **PDT 3.0** – to be published in 2013

Prague Dependency Treebank 2.5 (PDT 2.5)

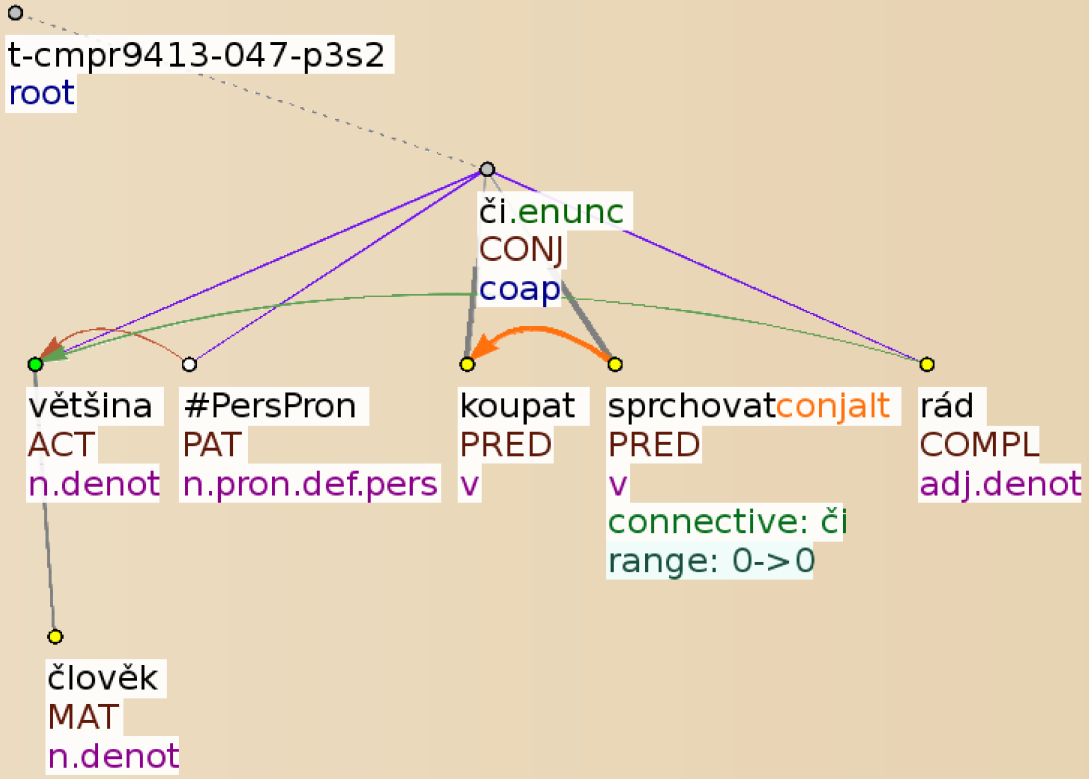
- published in 2012, same data as PDT 2.0
- corrections of errors
- additional annotation
  - multiword expressions
  - segmentation to clauses within sentences
  - new grammateme as complementation of noun number "pair-group" meaning

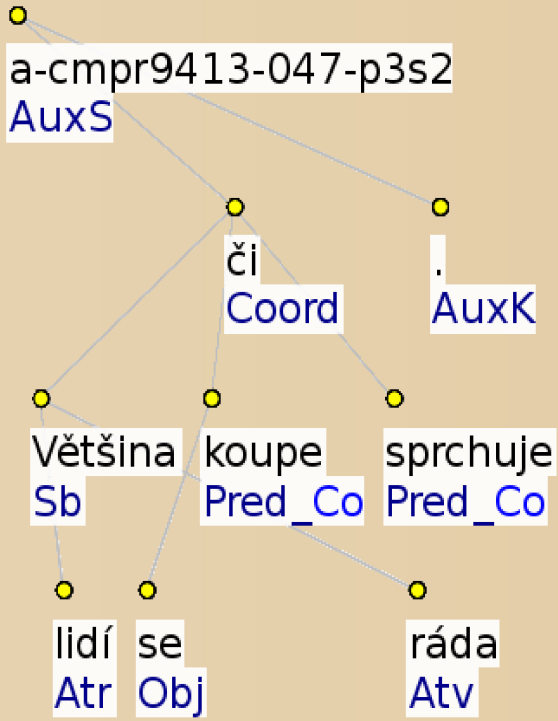
Prague Dependency Treebank 3.0 (PDT 3.0)

- to be published in 2013, same data as PDT 2.0
- corrections of errors
- additional or revised annotation
  - new or revised grammatemes (resultative, factmod, sentmod), revised ellipsis, analytical predicates with light verbs
  - manual checks of automatic annotation of grammatemes
  - extended textual coreference, bridging anaphora
  - discourse annotation

Prague Dependency Treebank

Většina lidí se koupe či sprchuje ráda.
(Most people enjoy taking a bath or shower.)





How Is PDT Used? Linguistic Research

- the database of examples for the study of particular syntactic and semantic topics (form and function relations, types of dependency relations, types of connection between sentences, clauses and syntagmas etc.) – *series of monographs printed at UFAL*
- discovering of gaps in the annotation scenario and introduction of the new (not yet subtle enough) distinctions – resulting in the more adequate description of the Czech syntax – *manuscript of Mluvnice češtiny 2. Syntax založená na anotovaném korpusu (Grammar of Czech 2. Syntax based on the annotated corpus (J. Panevová and others)*

How Is PDT Used?
NLP

- training and test data for any annotation task from the three layers of annotation
  - morphological tagging
  - parsing to the analytical or tectogrammatical layer (tree structure, attributes assignment)
  - anaphora resolution
  - topic-focus articulation
  - discourse relations
  - etc.

Prague Family of Treebanks

- Prague Dependency Treebank
  (PDT 2.0, PDT 2.5, PDT 3.0)
- Prague Czech-English Dependency Treebank
  (PCEDT 2.0)
- Prague Dependency Treebank of Spoken Czech
  (PDTSC 1.0)

ÚFAL

Prague Czech-English Dependency Treebank (PCEDT 2.0)

- published in 2012 (LDC)
  - version 1.0 published in 2004 (LDC)
- Penn Treebank, Wall Street Journal texts
  - translated to Czech
  - 1.2 million tokens, almost 50 thousand sentences

Prague Czech-English Dependency Treebank (PCEDT 2.0)

- Czech part
  - morphological and analytical layer annotated automatically (Collin's parser for a-layer)
  - tectogrammatical layer annotated manually (limited)
- English part
  - morphological and analytical layer transformed from Penn Treebank
  - tectogrammatical layer annotated manually (limited), using additional sources (PropBank and others)

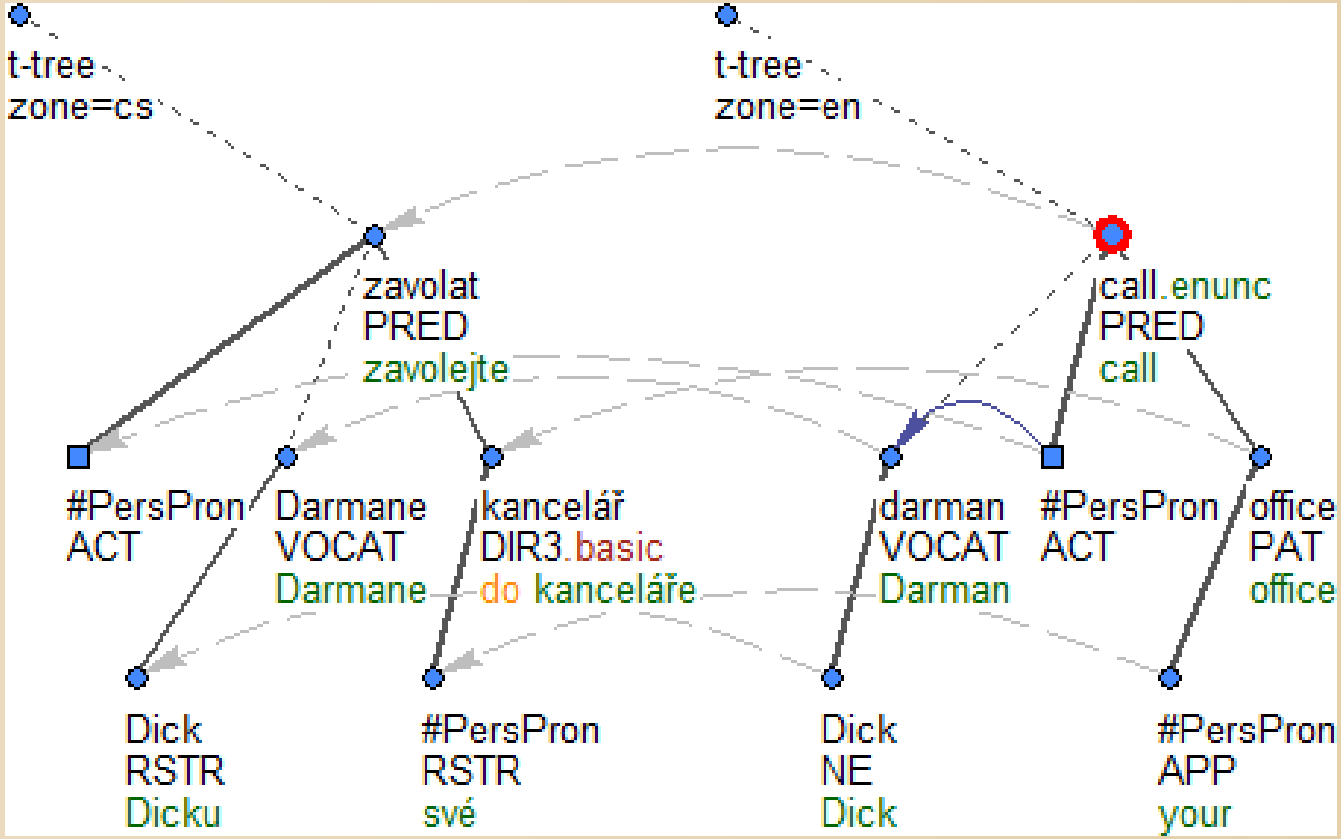
ÚFAL

Prague Czech-English Dependency Treebank (PCEDT 2.0)

- Parallel texts alligned on sentence level and (automatically) also on word/node level (separately for each layer)

- On both language sides:
  - grammatemes missing
  - additional on-going annotations:
    - topic-focus articulation (on data sample)
    - textual coreference

Prague Czech-English Dependency Treebank (PCEDT 2.0)



Dick Darman, call your office.
Dicku Darmane, zavolejte do své kanceláře.

How Is PCEDT Used?

- contrastive studies
  - passive in English and Czech contrasting with the word-order in both languages
  - verb patterns in English (contrastive valency studies)

- machine translation
  - training and test data

Prague Dependency Treebank of Spoken Czech PDTSC 1.0

- planned for publication in LDC in 2013/2014
- 742,257 word-tokens, 73,835 sentences
- data – spoken language
  - Czech testimonies from MALACH
  - dialogs collected within the COMPANION project (dialogues with older people about their personal photos)

Prague Dependency Treebank of Spoken Czech PDTSC 1.0

- speech **recognition** and **reconstruction**

Prague Dependency Treebank of Spoken Czech PDTSC 1.0

- **speech reconstruction**
  - inspired by F. Jelinek and E. Fitzgerald
  - phenomena typical for the spoken language and non-verbal means are deleted, segmentation to sentences
  - the result can be further processed as normal written texts
- **the morphological, analytical and tectogrammatical layers**
  - morphology and analytical structure obtained automatically
  - tectogrammatic annotation in progress
  - attributes annotated manually according to the same rules as PDT

ÚFAL

How Is PDTSC Used?

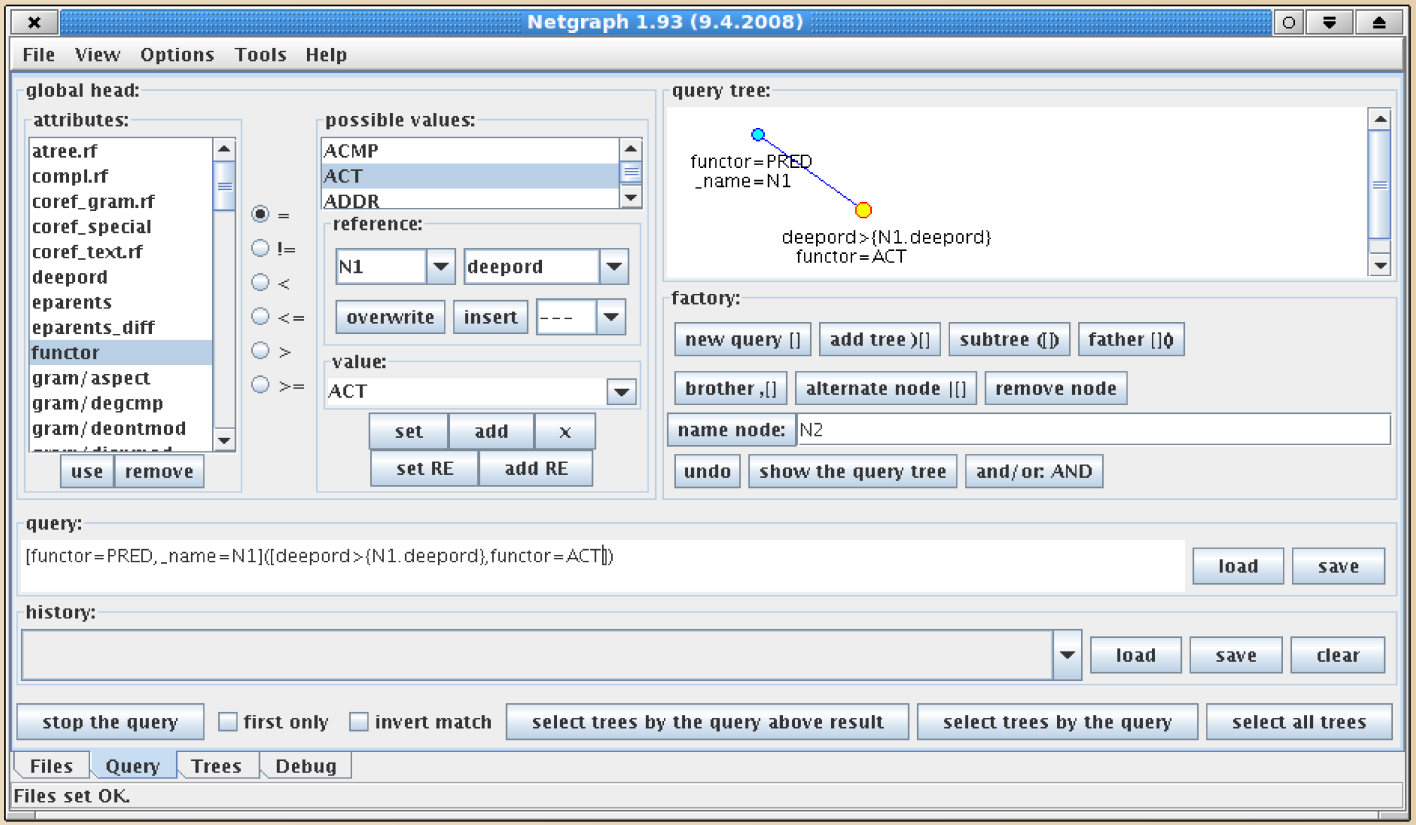
- NLP – spoken language in modern technologies (smart phones, GPS, tools for handicapped persons)
- language research comparing the phenomena in spoken and written language

What Else Do We Have?

- CzEng 1.0 (Czech-English Parallel Corpus)
  - 15 million parallel sentences annotated up to the tectogrammatical layer
  - everything annotated automatically
  - used for machine translation

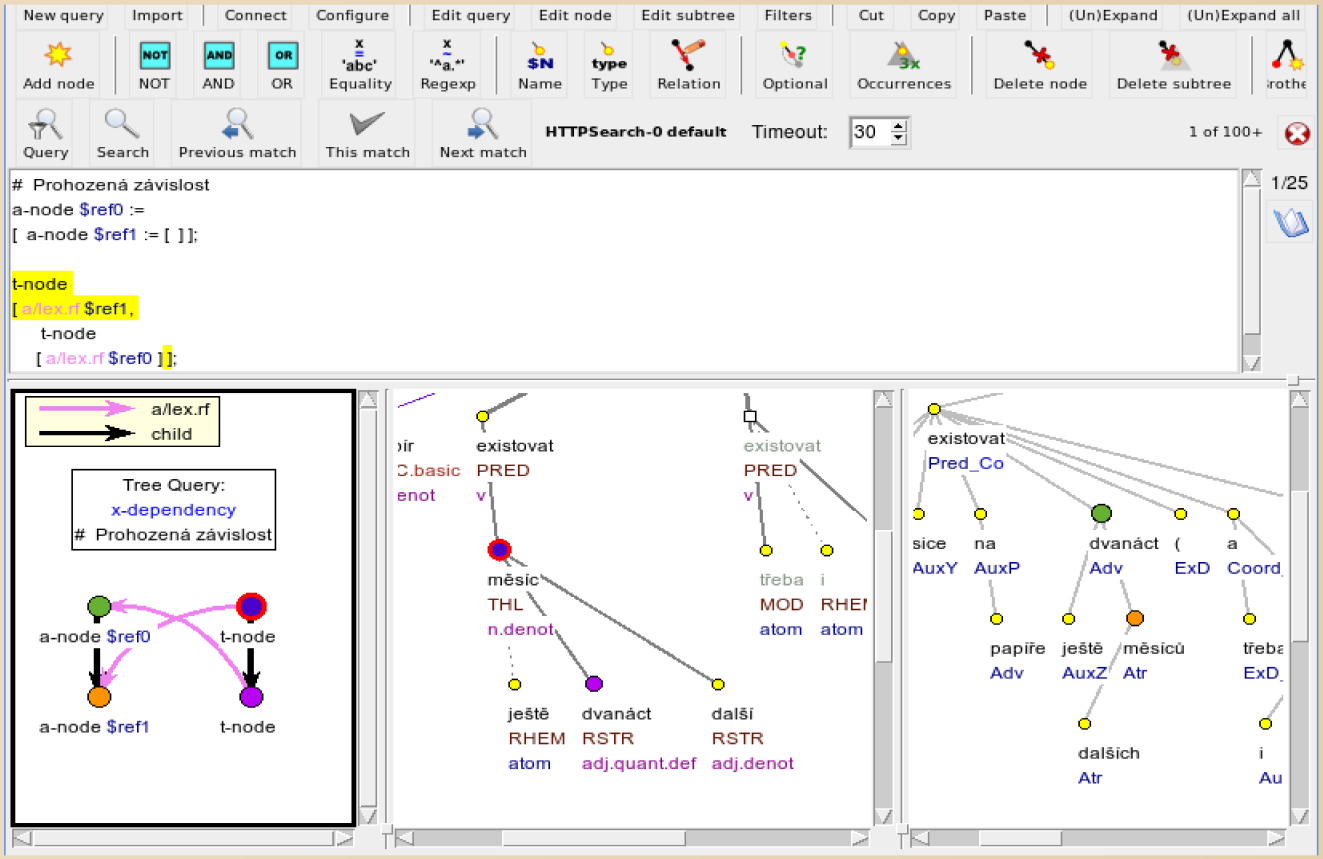
Searching Tools

- NetGraph
  - easy to use but outdated, limited power



Searching Tools

- PML-TQ
  - a little bit more complex but very powerful



Prague Dependency Treebanks
A Family of Annotated Treebanks

Thank you for you attention!

Jiří Mírovský

Jarmila Panevová

ÚFAL