From Morphology to Semantics: the Prague Dependency Treebank Family

Jan Hajič

Charles University in Prague
Institute of Formal and Applied Linguistics
LINDAT-Clarin and META-NET (CZ)

Czech Republic
History

- **LDC: Penn Treebank I (1993)**
  - We want it too!

- **But:**
  - LDC’s unlikely to do Czech (soon 😊)
  - Prague (old time structuralist) tradition: dependency

- **1995: decision to build our own treebank**
  - Started 1996 with a specification grant
  - Tool development, annotation since 1997
  - First PDT (1.0) published in 2001 (LDC2001T10)
    - Morphology and syntax only, but > 1M words
  - PDT 2.0 2006 (LDC2006T01)
    - Full annotation & correction of 1.0
  - Other treebanks: 2004, 2012 (more to come, also by other groups)
Prague Dependency Treebanks
the Basics

- General Features
  - Multilayered annotation, interlinked layers
  - Dependency-based syntax (both surface and deep)
    - Includes semantic functions, valency dictionary(-ies)
  - Information structure of the sentence (topic/focus)
  - Grammatical and textual co-reference, new: bridging
  - New: discourse relations (not published yet)

- Languages: Czech, English (also parallel), Arabic:
  - Indonesian, Urdu, Russian, … (Student work on samples)
  - (Auto) conversion from other treebanks (25 so far; experimental)
  - Spoken: Czech and English (non-parallel, dialogs)
The Layers

- Three basic layers
  - Morphological layer
  - Surface syntax (“a”) layer
  - “Tectogrammatical” layer: underlying syntax, semantic roles (valency), inf. structure, co-reference (anaphora)

- Format
  - Prague Markup Language (XML + Schema)

(Speech:
  - Additional layers:
    - audio, transcript )
Tectogrammatical vs. Analytical (Surface) Syntax

In practice, that procedure will require making of certified copies.
PDT-style Treebanks (written language)

- **Czech**
  - Prague Dependency Treebank
    - Complex annotation, all levels, additional annotation
  - Translation of Penn Treebank, aligned
    - Tectogrammatical layer only, no information structure
    - Analytical, morphology: automatic tools
    - Will be manually revised later

- **English**
  - Re-annotation of Penn Treebank, TR only so far

- **Arabic**
  - New morphology, analytical syntax, sample TR only
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

Aligned trees

Aligned nodes
Názory na její tříměsíční perspektivu se různí.
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td></td>
<td>49,208</td>
</tr>
<tr>
<td>a-nodes (automatic)</td>
<td>1,151,150</td>
<td>1,173,766</td>
</tr>
<tr>
<td>t-nodes (manual)</td>
<td>931,846</td>
<td>838,212</td>
</tr>
</tbody>
</table>

- Published June 2012 (LDC2012T08)
- Also available through LINDAT-Clarin (with browsing and search tools) and META-SHARE

<table>
<thead>
<tr>
<th>Alignment links</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a-layer</td>
<td>1,214,441</td>
</tr>
<tr>
<td>t-layer</td>
<td>727,415</td>
</tr>
</tbody>
</table>
Czech-English alignments

- Sentence-level (manual, natural due to translation)
  - At both syntactic levels
- Word (node) level
  - automatic, test section manually corrected (in part)
The Alignment(s)

- Czech-English alignments
- Sentence-level (manual, natural due to translation)
- At both syntactic levels:
  - Word (node) level: automatic, test section manually corrected, $m \rightarrow n$
- Between annotation levels:
  - Tectogrammatics to surface syntax $m \rightarrow n$, incl. $1 \rightarrow 0$
  - Surface syntax to word level ($1 \rightarrow 1$)
This temblor-prone city dispatched inspectors, firefighters and other earthquake-trained personnel *-1 to aid San Francisco.
PDT-style Treebanks (spoken language)

- Specifics of spoken language
  - Short sentences but unclear segmentation
    - Sentence breaks must be (re)annotated
  - Ungrammatical (esp. for Czech – coll.)
    - Annotation based on written-language rules difficult if not impossible

- ...additional decisions:
  - Change annotation?
  - Change the input? (but original must be kept)
Spoken corpora

- Solution: “Speech reconstruction”
  - Keep audio, word-for-word transcript
    - Adds two layers to the annotation scheme: audio, transcript
  - Add edited text: **LINKS** to original transcript / audio

- Annotate edited text (using usual guidelines)
Accompanying Tools

- **TrEd** ([http://ufal.mff.cuni.cz/tred](http://ufal.mff.cuni.cz/tred))
  - Annotation, View/Browse and Search environment
  - Open source, perl
  - Search and visualization: **PML-TQ**
    - Powerful query language for complex NLP annotation, esp. tree-based

- **Treex** ([http://ufal.mff.cuni.cz/treex](http://ufal.mff.cuni.cz/treex))
  - Modular NLP processing environment
  - Easy handling of complex NLP-annotated data
  - Modules exists for Czech, English data processing
    - incl. 3rd-party tools integrated into Treex
  - CPAN-distributed
Lessons Learned (1)

Positive experience
- Dependency style
- Separate layers of annotation
  - Most importantly: separate surface syntax vs. deep syntax
- Specific format and specific graphical tools (TrEd et al.)
  - Stand-off annotation
- Spoken annotation “trick” with speech reconstruction
  - Still, additional guidelines needed

Negative experience
- Lots of time spent on consistency checking
  - Annotator training: guidelines too detailed
  - Prevents crowdsourcing
- Lots of time goes to final quality checking and corrections
  - min. 3 PY for PDT, PCEDT
Lessons Learned (2)

- For future projects
  - Annotation in small teams
    - “Phenomenon-by-phenomenon”
  - Ongoing quality checking, time allotted for final QC
    - Error discovered at annotation time *much* cheaper to correct
    - Consequences for tool selection (“intelligent” annotation SW)
  - Need for excellent software and annotator’s support
    - Programmers’ efforts always underestimated
    - “helpdesk” for annotators important (usually former annotator)
    - Organization, statistics, watchdog
      - Single repository for annotated data
  - Payment
    - Annotator’s incentives work (for speed of annotation)
  - Speed of annotation vs. quality
    - Almost no correlation

Acknowledgements:
LC536, MSM0021620838

Acknowledgements:
Czech Science Foundation
GAP406/10/0875

Acknowledgements:
Czech Science Foundation
GA405/09/0729

Acknowledgements:
“Information Society” Programme
1ET101120503

Acknowledgements:
Charles University research funds (“PRVOUK”)
Happy 88th Birthday!