The Prague Czech-English Dependency Treebank

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popeňka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, Zdeněk Zabokrtský

Charles University in Prague
Institute of Formal and Applied Linguistics
Czech Republic
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

Parallel treebank
The Prague Czech-English Dependency Treebank (PCEDT) 2.0
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

Aligned trees

Aligned nodes
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

Parallel treebank
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
Parallel tree (surface) &
Dependence
- (surface)
- syntax & semantics
  = “tectogrammatics”
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
- Penn Treebank translation into Czech
The Prague Czech-English Dependency Treebank (PCEDT) 2.0
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

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

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
- Penn Treebank translation into Czech
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
- Penn Treebank translation into Czech
- 1 million words
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>

<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>
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
- Penn Treebank translation into Czech
- 1 million words
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
- Penn Treebank translation into Czech
- 1 million words
- Getting ready at LDC
  - Also available through LINDAT-Clarin and META-SHARE
PCEDT 2.0
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 (in part)
Incorrected (in part)

Due to translation

Corrected (in part)
interaction(s) (due to translation)
PCEDT 2.0
The Alignment(s)

- Czech-English alignments
  - Sentence-level (manual, natural due to translation)
  - Word (node) level
    - automatic, test section manually corrected (in part)

- Between annotation levels
  - Tectogrammatics to surface syntax
    - \( m \rightarrow n, \text{incl. } 1 \rightarrow 0 \)
  - Surface syntax to word level (1 \( \rightarrow \) 1)
Tectogrammatics

Surface syntax

She and their families were analyzed by the arrangements.

May 25, 2012
T 2.0
department(s)
surface syntax
PTB syntax

She analyzed their families
by

arranged

analyzed

family

their

sleeping

arrangements

arrangements

She

analyzed

families

by

sleeping

arrangements

arrangements

arrangements

May 25, 2012
Surface syntax annotation

- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + "pure" POS tags from PTB
  - Automatic (from PTB) + a few manual corrections
Surface syntax annotation

- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + „pure“ POS tags from PTB
  - Automatic (from PTB) + a few manual corrections
Surface syntax annotation

- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + „pure“ POS tags from PTB
  - Automatic (from PTB) + a few manual corrections

- Czech
  - PDT style, no change
  - Syntax: automatic, 2000 sentences fully manual for testing
  - Lemmatization and tagging: auto
    - 99%/96%, Spoustová et al. EACL 2009 (COMPOST tagger)
    - http://ufal.mff.cuni.cz/compost (Czech, English & other)
  - No p-level (of course 😊)
Surface syntax annotation

- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + „pure“ POS tags from PTB
  - Automatic (from PTB) + a few manual corrections

- Czech
  - PDT style, no change
  - Syntax: automatic, 2000 sentences fully manual for testing
  - Lemmatization and tagging: auto
    - 99%/96%, Spoustová et al. EACL 2009 (COMPOST tagger)
    - http://ufal.mff.cuni.cz/compost (Czech, English & other)
  - No p-level (of course 😊)
- PDT style, no change
- Syntax: automatic, 2000 sentences fully manual for testing
- Lemmatization and tagging: auto
  - 99%/96%, Spoustová et al. EACL 2009 (COMPOST tagger)
- No p-level (of course 😊)
Tectogrammatical annotation

- Manual (both languages)
- Major features
  - Nodes with “autosemantic” words only (no function words)
    - Ellipsis “restored” (new node for verbal arguments)
  - (Semantic) function (dependent → head relation)
    - Verb arguments + ca 50 functions for other relations
  - Valency lexicons attached (Eng: links to PropBank)
  - “Formemes”: prep+case style label (useful in MT and search)
  - Co-reference integrated (Eng: BBN + more), Czech: manually
- Alignment
  - To surface syntax & between Czech and English
This temblor-prone city dispatched inspectors, firefighters and other earthquake-trained personnel *-1 to aid San Francisco.
This temblor-prone city dispatched inspectors, firefighters and other earthquake-trained personnel *-1 to aid San Francisco.
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 annotations, 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
FAQ
(with ‘take-home’ answers)

What is different from PCEDT 1.0 (2004)?
- Size: full PTB & translation (1 mil.) vs. ~400k
- Full tecto-manual annotation both Czech and English
- Integrated co-reference, named entities
- Valency lexicons on both sides, links to PropBank, VerbNet
- “Formemes”: prep+case label
- English lemmatization (m-level, t-level)

Next?
- Manual “native” annotation of surface dependencies
- Extended semantic annotation
  - modalities, tense, number, etc.
- Aligned valency lexicon entries (Cze ↔ Eng)
- Remote access and „EZ“ search for data and lexicons
FAQ
(with ‘take-home’ answers)

What is different from PCEDT 1.0 (2004)?
- Size: full PTB & translation (1 mil.) vs. ~400k
- Full tecto-manual annotation both Czech and English
- Integrated co-reference, named entities
- Valency lexicons on both sides, links to PropBank, VerbNet
- “Formemes”: prep+case label
- English lemmatization (m-level, t-level)

Next?
- Manual “native” annotation of surface dependencies
- Extended semantic annotation
  - modalities, tense, number, etc.
- Aligned valency lexicon entries (Cze ↔ Eng)
- Remote access and „EZ“ search for data and lexicons
This is only a small sample of the corpus. You need to order and properly license the corpus to browse it in its entirety.

Section: 00  File: 01  Sentence: 1

Pierre Vinken, 61 years old, will join the board as a non-executive director Nov. 29.

Jednašedesátiletý Pierre Vinken se připojí ke správní radě jako nevykonné ředitel dne 29. listopadu.
Explore & enjoy!

This is only a small sample of the corpus. You need to order and properly license the corpus to browse it in its entirety.

Section: 00  File: 01  Sentence: 1

[en] Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Explore & enjoy!

Get the DVD ROM through LDC or browse a sample:
ufal.mff.cuni.cz/pcedt2.0