Prague Dependency Treebank
(vs. Functional Generative Description)
and HamleDT Family

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz
Prague Dependency Treebank

~ application of the FGD theory on the large set of Czech data

http://ufal.mff.cuni.cz/pdt2.0/
https://ufal.mff.cuni.cz/pdt3.0/

1. data
2. tools
3. documentation:
   • manuals for individual layers
     https://ufal.mff.cuni.cz/pdt3.0/documentation
   • survey of data formats and tools
Prague Dependency Treebank (cont.)

4 layers:
- word layer (w-layer)
- morphological layer (m-layer)
- analytical layer (a-layer)
- tectogrammatical layer (t-layer)

<table>
<thead>
<tr>
<th>layers of description</th>
<th>t,a,m-layer</th>
<th>a,m-layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train</td>
<td>dtest</td>
</tr>
<tr>
<td># documents</td>
<td>2 536</td>
<td>316</td>
</tr>
<tr>
<td># sentences</td>
<td>38 737</td>
<td>5 228</td>
</tr>
<tr>
<td># tokens</td>
<td>652 700</td>
<td>87 988</td>
</tr>
</tbody>
</table>

PDT – FGD vs. PDT

Lopatková
Prague Dependency Treebank (cont.)

• stand-off annotation
• manual annotation
  with a massive post-annotation consistency checking
• formats and tools:
  – TrEd … tree editor and viewer (Pajas, xxxx)
  – PML data format (XML-based format)
  – PML-TQ … search tool
    http://ufal.mff.cuni.cz/~pajas/pmltq/
• more during the practical sessions
PDT: w-layer

  - Lidové noviny (daily newspapers)
  - Mladá fronta Dnes (daily newspapers)
  - Českomoravský Profit (business weekly)
  - Vesmír (scientific journal)

- part of the Czech National Corpus

- a sequence of tokens (word forms and punctuation marks)

- including errors, typing errors, bad segmentation, …
PDT: m-layer

• the sequence of tokens divided into sentences
• errors are corrected
• annotation:
  – morphological lemma
  – morphological tag
  – id
  – reference to w-layer
  – form (corrections: spelling errors, incorrectly split or joined words, …)
• manually annotated (parallel annotation)
Některé kontury problému se však po oživením Havlovým projevem zdají být jasnější.

[Some contours of the problem seem to be clearer after the resurgence by Havel's speech.]

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>Morphological tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Některé</td>
<td>některý</td>
<td>PZFP1-------------</td>
</tr>
<tr>
<td>kontury</td>
<td>kontura</td>
<td>NNFP1-----A----</td>
</tr>
<tr>
<td>problému</td>
<td>problém</td>
<td>NNIS2-----A----</td>
</tr>
<tr>
<td>se</td>
<td>se_^(zvr._zájmeno/částice)</td>
<td>P7-X4-----------</td>
</tr>
<tr>
<td>však</td>
<td>však</td>
<td>J^-------------</td>
</tr>
<tr>
<td>po</td>
<td>po-1</td>
<td>RR--6----------</td>
</tr>
<tr>
<td>oživení</td>
<td>oživení_^(3it)</td>
<td>NNNS6-----A----</td>
</tr>
<tr>
<td>Havlovým</td>
<td>Havlův_;S_^(3el)</td>
<td>AUIS7M-----------</td>
</tr>
<tr>
<td>projevem</td>
<td>projev</td>
<td>NNIS7-----A----</td>
</tr>
<tr>
<td>zdají</td>
<td>zdáť</td>
<td>VB-P---3P-AA---</td>
</tr>
<tr>
<td>být</td>
<td>být</td>
<td>Vf--------A----</td>
</tr>
<tr>
<td>jasnější</td>
<td>jasný</td>
<td>AAFP1----2A----</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>Z:-------------</td>
</tr>
</tbody>
</table>
PDT: a-layer

- dependency tree
- one token from m-layer ~ one node incl. prepositions, punctuation … plus technical root
- relations ~ edges
  dependency, coordination, punctuation, …
- linear ordering ~ surface word order
- annotation:
  - analytical function (afun)
  - linear order
    - is_member
    - is_parenthesis_root
  - id
  - reference to m-layer

PDT – FGD vs. PDT

Lopatková
Některé kontury problému se však po oživením Havlovým projevem zdají být jasnější.

[Some contours of the problem seem to be clearer after the resurgence by Havel's speech.]
PDT: t-layer

- tectogrammatically tree structure ~ dependency tree
  - nodes for auto-semantic/lexical words only
    syn-semantic/functional words as attributes of lexical words
    (plus technical root)
  - ellipses as nodes
  - edges ~ relations (dependency, coordination, others)
  - link to a valency lexicon for verbs and (certain types of) nouns

- topic-focus articulation (TFA)
  - linear ordering ~ deep word order
  - contextually bounded and unbounded nodes

- coreference
PDT: t-layer (basic attributes)

- tectogrammatical tree structure
  - t-lemma
  - functor
  - grammatemes (16 attributes starting with the prefix gram)
  - is_member
  - is_parenthesis_root
  - id
  - reference to a-layer
...
- topic-focus articulation (TFA)
  - deepord
  - tfa
- coreference
  - coref_text.rf
  - coref_gram.rf
...
Some contours of the problem seem to be clearer after the resurgence by Havel's speech.
Linking the layers

• references **from a higher layer to a lower layer**:  
  • t-layer $\rightarrow$ a-layer  
  • a-layer $\rightarrow$ m-layer  
  • m-layer $\rightarrow$ w-layer

• **1:1** correspondence between nodes of the *m-* and *a-*layers
PDT: Division of the data to layers

- t-layer
- a-layer
- m-layer

PDT – FGD vs. PDT
PDT: Division of the data into training and test sets
PDT: Number of tokens from the particular sources

![Bar chart showing the number of tokens from different sources across different layers of annotation.](chart.png)

- **m-layer**: 9.10% of tokens from Českomoravský, 63.09% from Mladá fronta, 19.05% from Lidové noviny, and 8.75% from Vesmír.
- **a-layer**: 11.86% of tokens from Českomoravský, 61.20% from Mladá fronta, 15.53% from Lidové noviny, and 11.40% from Vesmír.
- **t-layer**: 14.31% of tokens from Českomoravský, 76.79% from Mladá fronta, 8.90% from Lidové noviny, and 8.90% from Vesmír.
Návštěvy kin a divadel patří mezi méně časté aktivity mladých lidí v České republice. [Attending cinemas and theaters belongs to less frequent activities of young people in the Czech republic.]
Podle slov pražského primátora Jana Koukala by tato čtvrt měla vzniknout během roku a půl.

[In the words of the city's mayor Jan Koukal, this quarter should arise in a year and a half.]
Společnost vyrábí model Charade japonské automobilky Daihatsu, který je v Číně používán mimo jiné jako taxi.

[The company produces the Charade model of the Japanese car factory Daihatsu, which is used in China also as a taxi.]
Differences between FGD and PDT
# Differences between FGD and PDT

<table>
<thead>
<tr>
<th>FGD</th>
<th>PDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>tectogrammar/deep syntax</td>
<td>t-layer (tectogrammatical l.)</td>
</tr>
<tr>
<td>surface syntax</td>
<td>a-layer (analytical l.)</td>
</tr>
<tr>
<td>morphematics</td>
<td>m-layer (morphological l.)</td>
</tr>
<tr>
<td>morphonology</td>
<td>w-layer (word layer)</td>
</tr>
<tr>
<td>phonology</td>
<td></td>
</tr>
</tbody>
</table>

## reasons
- analysis vs. synthesis/generation  ➡ richer information
- technical reasons (financial, temporal restrictions, implementation)
Differences between FGD and PDT (cont.)

**morphematics** (FGD) vs. **m-layer** (PDT)

- morphemes for individual words are grouped
- grammatical categories ~ morphological tags
- annotated text is divided into sentences
Differences between FGD and PDT (cont.)

*structural layers*

- technical root
- connecting constructions for coordination and apposition in PDT
Differences between FGD and PDT (cont.)

1. **surface syntax** (FGD) vs. **a-layer** (PDT)
   - each token of m-layer is represented by a node
     (incl. prepositions, auxiliary verbs, punctuation, …)
     (vs. units corresponding to formemes)
   - edges for non-dependency relations
     (other than coordination/apposition)
     - function words (e.g., auxiliary verbs) usually below respective lexical words
     - exception: prepositions, subordinating conjunctions as parents of lexical words
Differences between FGD and PDT (cont.)

1. **surface syntax** (FGD) vs. **a-layer** (PDT)
   - each token of m-layer is represented by a node
     (incl. prepositions, auxiliary verbs, punctuation, …)
     (vs. units corresponding to formemes)
   - edges for non-dependency relations
     (other than coordination/apposition)
     - function words (e.g., auxiliary verbs) usually below respective lexical words
     - exception: prepositions, subordinating conjunctions as parents of lexical words
   - ellipses: elided words are not restored at a-layer
   - a word modifying an elided word as a child of the 'lowest' ancestor
2. **deep/tectogram** syntax (FGD) vs. **t-layer** (PDT)

- core vs. periphery
  - specific constructions (direct speech, comparison)
- edges for non-dependency relations
  - syntactically unclear expressions
  - list structures
  - phrasemes
- info on the (non)realization in the surface sentence (is_generated)
2. **deep/tectogram**. syntax (FGD) vs. **t-layer** (PDT)

- core vs. periphery
  - specific constructions (direct speech, comparison)
- edges for non-dependency relations
  - syntactically unclear expressions
  - list structures
  - phrasemes
- info on the (non)realization in the surface sentence (is_generated)
- topic-focus articulation
- coreference
  - relative/interrogative pronouns, personal pronouns (3rd person)
  - grammatical control, complement
Other treebanks: Prague dependency family

Prague Dependency Treebank
Other treebanks: Prague dependency family

Czech:

Prague Dependency Treebank
1.0 (2001); 2.0 (2006); 2.5 (2011); 3.1 (2013)
http://ufal.mff.cuni.cz/pdt2.5/

Czech Academic Corpus 1.0 (2006), 2.0 (2008)
- morphological annotation (652 000 tokens, 32 000 sentences)
- analytical annotation (493 000 tokens, 25 000 sentences)
- both written and spoken language
- manually annotated

Prague Dependency Treebank of Spoken Czech
http://ufal.mff.cuni.cz/pdtsl/
Other treebanks: Prague dependency family

Prague English Dependency Treebank 1.0 (2009)

- texts from the Wall Street Journal (Penn Treebank III)
- adaptation of the PDT-like annotation scheme to English
- tectogrammatical annotation
- 12 440 annotated and checked trees

Whether desirable or not, this is a child-care program, not an educational program. (Wall Street Journal 1286/49)
Other treebanks: Prague dependency family

Prague **Czech-English** Dependency Treebank 1.0 (2004)

- Penn Treebank data (Wall Street Journal, 21,600 English sentences)
- Human translators
- Automatic conversions of Penn Treebank annotation into PDT-like annotation scheme (m-, a- and t-layers)
- Plain text from Reader's Digest 1993-1996 (50,000 sentences)

Test data:
- 515 sentence pairs
- Manually annotated on tectogrammatical level, Czech and English
- Retranslated from Czech to English by 4 different translation companies
Other treebanks: Prague dependency family

Prague Czech-English Dependency Treebank 2.0

- Penn Treebank data
- manually annotated data (49 000 sentences)

*But the strategy isn’t helping much this time.*

*Tato strategie však tentokrát příliš nepomáhá.*
Ale musíte uznat, že se tyto události odehrály před 35 lety.

But you have *-1 to recognize that these events took place 35 years ago.
In the new position he will oversee Mazda's U.S. sales, service, parts and marketing operations.

Vitulli bude ve své nové funkci dohlížet na americký prodej, služby, součásti a marketing společnosti Mazda.
Pětapadesátiletý Rudolf Agnew, bývalý předseda společnosti Consolidated Gold Fields PLC, byl jmenován nevýkonným ředitelem tohoto britského průmyslového konglomerátu.

Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate.
Other treebanks: Prague dependency family

**Czech-English** Parallel Corpus 1.0
(~15.0 M parallel sentences )
http://ufal.mff.cuni.cz/czeng/
- collected automatically
- annotated automatically
- European laws, subtitles, technical documentation, electronic books, newspapers, ...

*It is extremely important that Iraq held elections to a constitutional assembly.*

PDT – FGD vs. PDT
Other treebanks: Prague dependency family

Prague Arabic Dependency Treebank 1.0 (2004)
http://ufal.mff.cuni.cz/padt/PADT_1.0/docs/index.html

- Functional Arabic Morphology
- analytical layer
  (about 130 000 tokens)
- tectogrammatical layer
Other treebanks: Prague dependency family

**HamleDT** ~ a compilation of existing dependency treebanks (or dependency conversions of other treebanks), transformed so that they all conform to the same annotation style

36 languages, 42 treebanks in HamleDT 3.0 (2015)

http://ufal.mff.cuni.cz/hamledt/

PDT – FGD vs. PDT

Lopatková
Other treebanks: Prague dependency family

HamleDT ~ a compilation of existing dependency treebanks (or dependency conversions of other treebanks), transformed so that they all conform to the same annotation style

PDT-like tree

Universal Dependencies

PDT – FGD vs. PDT

Lopatková
How to access / obtain dependency treebanks

- **as a web service**
  
  [https://lindat.mff.cuni.cz/services/pmltq/#!/home](https://lindat.mff.cuni.cz/services/pmltq/#!/home)

  LINDAT/Clarin Repository
  PML-TQ search tool
How to access / obtain dependency treebanks

• **as a web service**
  
  https://lindat.mff.cuni.cz/services/pmltq/#!/home

  LINDAT/Clarin Repository

  PML-TQ search tool

  more stable, quick

• **via Tred instalation**

  PML-TQ search tool

  graphical interface for creating queries (practical lectures)
References

• PDT guide http://ufal.mff.cuni.cz/pdt2.0/
• ...