Syntax Meets Semantics: the Family of Prague Dependency Treebanks

Jan Hajič
Institute of Formal and Applied Linguistics
School of Computer Science
Faculty of Mathematics and Physics
Charles University, Prague
Czech Republic

The Prague Dependency Treebank

- The idea
  - Apply the “old” Prague theory to real-word texts
  - Provide enough data for ML experiments
- “Old” Prague theory
  - Prague structuralism (1930s)
  - Stratificational approach
  - Centered on “deep syntax”
    - Separated from “surface form”
    - Dependency based (how else 😊)
Prague Dependency Treebank
The Methodology

- Manual annotation is PRIMARY
  - Some help from existing tools possible
- “No information loss, no redundancy”
  - Much formalization, but…
  - … original form always retrievable
- Dictionaries
  - In theory: “secondary”, side effect of annotation
  - In reality: help consistency
  - Links: data → dictionary(-ies)
- Extensive Machine Learning support
- Ergonomy of annotation
  - Graphical (“linguistic”) presentation & editing

The Prague Dependency Treebank Project (Czech Treebank)

- 1995 (Dublin!) 1996-2006-...
  - 1998 PDT v. 0.5 released (JHU workshop)
    - 400k words manually annotated, unchecked
  - 2001 PDT 1.0 released (LDC):
    - 1.3MW annotated, morphology & surface syntax
  - 2006 PDT 2.0 release
    - 0.8MW annotated (50k sentences) + PDT 1.0 corrected
    - the “tectogrammatical layer”
      - underlying (deep) syntax
Related Projects (Treebanks)

- Prague Czech-English Dependency Treebank
  - WSJ portion of PTB, translated to Czech (1.2 mil. words)
  - automatically analyzed
    - English side (PTB), too
    - Manual annotation started
- Prague Arabic Dependency Treebank
  - apply same representation to annotation of Arabic
  - surface syntax so far
- Both published in 2004 (LDC)
  - PCEDT version 2.0 being prepared (2008?)

PDT (Czech) Data

- 4 sources:
  - Lidové noviny (daily newspaper, incl. extra sections)
  - DNES (Mladá fronta Dnes) (daily newspaper)
  - Vesmír (popular science magazine, monthly)
  - Českomoravský Profit (economical journal, weekly)
- Full articles selected
  - article ~ DOCUMENT (basic corpus unit)
- Time period: 1990-1995
- 1.8 million tokens (~110 thousand sentences)
PDT Annotation Layers

- **L0 (w) Words (tokens)**
  - automatic segmentation and markup only
- **L1 (m) Morphology**
  - Tag (full morphology, 13 categories), lemma
- **L2 (a) Analytical layer (surface syntax)**
  - Dependency, analytical dependency function
- **L3 (t) Tectogrammatical layer ("deep" syntax)**
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon

PDT 1.0 (2001)
PDT 2.0 (2006)
Morphological Attributes

- Tag: 13 categories
- Example: \textit{AAdj P3 - - - 3N - - - negated}

- Adjective: no poss. Gender
- Regular: no poss. Number
- Feminine: no person
- Plural: no tense
- Dative: superlative

- Lemma: POS-unique identifier
  - Books/verb -> \textit{book-1}
  - went -> \textit{go}
  - to/prep. -> \textit{to-1}

Morphological Tagset

- 13 categories, 4452 plausible tags (combinations):

<table>
<thead>
<tr>
<th>Category</th>
<th># of values</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>10</td>
<td>N (noun), Z (punctuation)</td>
</tr>
<tr>
<td>SUBPOS</td>
<td>75</td>
<td>P (personal pron.), U (possessive adj.)</td>
</tr>
<tr>
<td>GENDER</td>
<td>8</td>
<td>I (masc. inanimate), X (any), - (N,A)</td>
</tr>
<tr>
<td>NUMBER</td>
<td>4</td>
<td>P (plural), D (dual)</td>
</tr>
<tr>
<td>CASE</td>
<td>9</td>
<td>1 (nominative), 6 (locative)</td>
</tr>
<tr>
<td>POSSGENDER</td>
<td>4</td>
<td>M (masc. animate), F (feminine)</td>
</tr>
<tr>
<td>POSSNUMBER</td>
<td>3</td>
<td>S (singular), P (plural)</td>
</tr>
<tr>
<td>PERSON</td>
<td>5</td>
<td>1 (first), ...</td>
</tr>
<tr>
<td>TENSE</td>
<td>4</td>
<td>P (present), M (past)</td>
</tr>
<tr>
<td>GRADE</td>
<td>5</td>
<td>3 (superlative)</td>
</tr>
<tr>
<td>NEGATION</td>
<td>3</td>
<td>A (affirmative), N (negative)</td>
</tr>
<tr>
<td>VOICE</td>
<td>3</td>
<td>A (active), P (passive)</td>
</tr>
<tr>
<td>VAR</td>
<td>11</td>
<td>1 (1st variant), 6 (colloq. style), 8 (abbrev.)</td>
</tr>
</tbody>
</table>
Morphological Analysis

- Formally: \( MA: A^+ \rightarrow \text{Pow}(L \times T) \)
  - \( MA(f) = \{ [ l, t ] \} \);
    - \( f \in A^+ \) (the token),
    - \( l \in L \) (lemma),
    - \( t \in T \) (tag)
- tokens taken in isolation
- no attempt to solve e.g. auxiliaries vs. full verbs
- Ex.: \( MA(\text{"má"}) = \{ [\text{mit}, VB-S---3P-AA---], \text{ lit. "to have" }\}
  - \( [\text{můj}, PSFS1-S1------1], \text{ lit. "my" }\)
  - \( [\text{můj}, PSFS5-S1------1], \text{ lit. "my" }\)
  - \( [\text{můj}, PSNP1-S1------1], \text{ lit. "my" }\)
  - \( [\text{můj}, PSNP4-S1------1], \text{ lit. "my" }\)
  - \( [\text{můj}, PSNP5-S1------1] \}

Using the Results: Morphological Disambiguation

- Full morphological disambiguation
  - more complex than (e.g. English) POS tagging
- Several full morphological taggers:
  - (Pure) HMM
  - Feature-based (MaxEnt-like)
    - used in the PDT distribution
  - Voted Perceptron, (M. Collins, EMNLP’02)
- All: ~ 94-5% accuracy (perceptron is best)
  - rule & statistic combination: tiny improvement
    (Hajič et al., ACL 2001, Spoustova et al., 2007: > 96%)
The Segmentation Problem: Possible solution (Arabic)

- Tokenization / segmentation not always trivial
  - Arabic, German, Chinese, Japanese
- Find max. no. of segments
  - 4 for Arabic
- Expand every solution (morph. analysis) to the same number of segments, adding “blank” segments to the end
- Concatenate tags (→ same length)
- Concatenate “lemmas” (roots, ...)
- Result:
  - The same formal definition; can be converted back to segments trivially
  - Tagging solves segmentation (98-99%)! (tagging: 89%)

PDT Annotation Layers

- L0 (w) Words (tokens)
  - Automatic segmentation and markup only
- L1 (m) Morphology
  - Tag (full morphology, 13 categories), lemma
- L2 (a) Analytical layer (surface syntax)
  - Dependency, analytical dependency function
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon
Layer 2 (a-layer): Analytical Syntax

- Dependency + Analytical Function

The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated.

Analytical Syntax: Functions

- Main (for [main] semantic lexemes):
  - Pred, Sb, Obj, Adv, Atr, Atv(V), AuxV, Pnom
  - "Double" dependency: AtrAdv, AtrObj, AtrAtr
- Special (function words, punctuation,...):
  - Reflexives, particles: AuxT, AuxR, AuxO, AuxZ, AuxY
  - Prepositions/Conjunctions: AuxP, AuxC
  - Punctuation, Graphics: AuxX, AuxS, AuxG, AuxK
- Structural
  - Elipsis: ExD, Coordination etc.: Coord, Apos
Surface Syntax Example

- Complete sentence: Sb, Pred, Obj
  - The-baker bakes rolls.
  - Pekař peče housky.

Surface Syntax Example

- Incomplete phrases
  - Peter works well, but Paul badly
  - Petr pracuje dobře, ale Pavel špatně
Surface Syntax Example

- Variants (equal meaning)
  - (he) bought shoes for boy
  - koupil boty pro kluka

PDT Annotation Layers

- L0 (w) Words (tokens)
  - automatic segmentation and markup only
- L1 (m) Morphology
  - Tag (full morphology, 13 categories), lemma
- L2 (a) Analytical layer (surface syntax)
  - Dependency, analytical dependency function
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency, functor (detailed), grammatemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon
Layer 3 (t-layer):
Tectogrammatical Annotation

- Underlying (deep) syntax
- 4 sublayers (integrated):
  - dependency structure, (detailed) functors
  - valency annotation
  - topic/focus and deep word order
  - coreference (mostly grammatical only)
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...
- Total
  - 39 attributes (vs. 5 at m-layer, 2 at a-layer)

Analytical vs. Tectogrammatical annotation (TR: sublayer 1 only)

Underlying verb + tense
Deep function
Elided Actor in
Prepositions out

(TR: sublayer 1 only shown)
Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference (mostly grammatical only)
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

Tectogrammatical Functors

- “Actants”: ACT, PAT, EFF, ADDR, ORIG
  - modify: verbs, nouns, adjectives
  - cannot repeat in a clause, usually obligatory
- Free modifications (~ 50), semantically defined
  - can repeat; optional, sometimes obligatory
  - Ex.: LOC, DIR1, ...; WHEN, TTILL,...; RSTR; BEN, ATT, ACMP, INTT, MANN; MAT, APP; ID, DPHR, ...
- Special
  - Coordination, Rhematizers, Foreign phrases,...
Tectogrammatical Example

- Analytical verb form:
  - (he) allowed would-be to-be enrolled
  - směl by být zapsán

- Passive construction (action)
  - (The) book has been translated [by Mr. X]
  - Kniha byla přeložena
Tectogrammatical Example

- **Object**
  - (he) gave him a-book
  - dal mu knihu

Obj goes into ACT, PAT, ADDR, EFF or ORIG based on governor’s valency frame

Tectogrammatical Example

- **Incomplete phrases**
  - Peter works well, but Paul badly
  - Petr pracuje dobře, ale Pavel špatně

Added
Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference (mostly grammatical only)
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

Deep Word Order

- Example:

  - Baker bakes rolls. vs. BakerIC bakes rolls.

Analytical dep. tree:
Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference (mostly grammatical only)
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

Coreference

- Grammatical (easy)
  - relative clauses
    - which, who
      - Peter and Paul, who ...
  - control
    - infinitival constructions
      - John promised to go ...
  - reflexive pronouns
    - {him, her, thme}self(-ves)
      - Mary saw herself in ...
Coreference

- Textual
  - Ex.: Peter moved to Iowa after he finished his PhD.

Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference (mostly grammatical only)
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...
Grammatemes

- Detailed functors (subfunctors)
  - only for some functors:
    - TWHEN: before/after
    - LOC: next-to, behind, in-front-of, ...
    - also: ACMP, BEN, CPR, DIR1, DIR2, DIR3, EXT

- Lexical (underlying)
  - number (SG/PL), tense, modality, degree of comparison, ...
  - strictly only where necessary (agreement!)

Example - simplified view

Se zuby jsem měl v minulosti jen problémy.
With teeth I-have had in the past only problems.
The boundaries of some problems seem to be clearer after they were revived by Havel’s speech.

Valency

- Valency: ability ("desire") of words (verbs, nouns, adjectives) to combine themselves with other units of meaning
- Properties of valency:
  - Specific for every word meaning (in general)
    - leave: sb left sth for sb vs. sb left from somewhere
    - same as in PropBank leave.02 vs. leave.01
  - Typically strongly correlates with surface form
    - morphological case (~ ending), preposition+case, ...
- Semantic constraints
Structure of Valency

- word (lemma)
  - word sense group 1
    - valency frame:
      - slot₁, slot₂, slot₃
      - surface expression
    - word sense group 2
      - ...

vyměnit (to replace)
  - vyměnit₁
  - ACT PAT EFF
  - Nom. Acc. za+Acc.
  - vyměnit₂
  - ...

The Valency Lexicon
PDT-VALLEX

- Valency frames
  - each verb, some nouns, adjectives
- Basic set prepared in advance, annotators add entries on-the-go, checking and approval process follows (consistency)
- VALLEX
  - more detailed and complex annotation of valency
  - Žabokrtský, Lopatková (2005), VALLEX 1.0
  - All about valency:
    http://ckl.ms.mff.cuni.cz/~semecky/vallex/
**PDT-VALLEX Entry**

- dosáhnout: “to reach”, “to get [sb to do sth]”
- browser/user-formatted example:

<table>
<thead>
<tr>
<th>dosáhnout</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT(1) PAT(2, 4) v-w 714#1</td>
</tr>
<tr>
<td>dosáhnout správně zavřít</td>
</tr>
<tr>
<td>ACT(1) PAT(2, aby[v])</td>
</tr>
<tr>
<td>dosáhnout na něm zlibu</td>
</tr>
<tr>
<td>ACT(1) DPHR(stvýr: [- I 2])</td>
</tr>
<tr>
<td>dosáhnout odvahu</td>
</tr>
<tr>
<td>ACT(1) DIR3(*)</td>
</tr>
<tr>
<td>dosáhnout na strop</td>
</tr>
</tbody>
</table>

**Corpus <-> Valency Lexicon**

- Corpus:

  - Sentence 2035:
    - uzavřít
    - PRED
    - on
    - nedávno
    - on
    - plékáka
    - CAUS
  
  - Sentence 15345:
    - uzavřít
    - PRED
    - &Gen: také
    - ACT
    - továrna
  
  - Sentence 51042:
    - uzavřít
    - PRED
    - on srpen
    - smlouva

- Lexicon:

  - ENTRY: uzavřít
    - V1: ACT(1) CPHR(smlouva).4
      - ex.: dohodu (close a contract)
    - V2: ACT(1) PAT(4)
      - ex.: u. pokoj (close a room, house)
Valency and Text Generation

- Tectogrammatical Representation
  - has *all* the information to (re)generate the surface form of the sentence:
    - in a “generalized” form
    - non-redundant (almost... but for generation, it is o.k.)
  - ...except the links to a-layer, however
    - links used only for *training* [statistical models for] parsing/generation modules
    - not present when e.g. doing text planning, translation, ...
  - valency dictionary: form of “learned” knowledge

Using valency for...
- ...getting the correct (lemma, tag) of verb arguments
- Example:

```
starat_se
PRED
Martin
ACT
tygr
PAT

"to take care of"

VALLEX entry: starat (se) ACT(.1) PAT(o.[.4])
```

"Martin takes care of tigers."

July 4, 2007
Brno Summer School 2007
The Annotation Process

- 4 sublayers
  - work on structure first, rest in parallel
- Structure
  - automatic preprocessing - programmed conversion from analytical layer annotation
- Grammatemes
  - mostly automatically (based on lower layers’ annotation), manual checking, corrections
- Cross-sublayer/cross-layer checking
  - partly automatic, then manual
Tectogrammatical Annotation Tools

- Manual annotation
  - 4 groups of annotators ~ 4 sublayers
  - Special graphical tool (TrEd)
    - Customizable graphical tree editor
- Preprocessing
  - Data from analytical layer, preprocessed
  - Online dependency function preassignment

The [Manual] Annotation Tool

- Perl/PerlTk based, platform-independent
  - Linux, Windows 95/98/2000, Solaris, ...
- Perl as the “macro” language
  - “unlimited” online processing capability
- Flexibility for interactive checking
  - split screen, graphical “diff” function
- Customization, printing, “plugins”, ...
The “TrEd” Tree Editor

- Graphical tool: TrEd
- Main screen:
  
  **Original sentence:**
  [This year’s flu season is still quiet in Europe.]

- Editing window customization
- Run a macro
- Multiframe editing/compare

Valency Lexicon in TrEd

- [example of valency lexicon in TrEd]
- to write sth (about sth)
**XML Annotation Layers**

- Strictly top-down links
- \( w + m + a \) can be easily "knitted"
- API for cross-layer access (programming)
- PML Schema / Relax NG
- [With slight modification, can be used for spoken data (audio as layer "-1")]

**The Prague Markup Language Example**

- m-layer data, linked to w-layer:

```xml
<m id="m-tr/_12941_01_00013.fs-s1w4">
  <src.rf>manual</src.rf>
  <w>
    <dest.rf>w#tr/_12941_01_00013.fs-s1w4</dest.rf>
    <trans>basic</trans>
  </w>
  <form>pocházela</form>
  <lemma>pocházet: T</lemma>
  <tag>VpQW---XR-AA---</tag>
</m>
```

```xml
<m id="m-tr/_12941_01_00013.fs-s1w5">
  ...
</m>
```
PDT 2.0: The Data

- Data sizes

![Diagram of data sizes](image)

---

TrEd

![Diagram of TrEd](image)
Using the Results: Parsing

- Several parsers of Czech
  - Analytical layer dependency syntax
  - Trained on PDT 1.0 data, 1.2 mil. words
- Collins(98), Charniak(00), Žabokrtský(02), Ribarov(04), Nivre(05), Zeman(05), McDonald(05), CoNLL’06 (19 parsers)
- Best results
  - accuracy: percent of correct dependencies:
  - 84-85% for a single parser, > 86% for a combination

Tectogrammatical Parsing

- Newest results:
  - 4 phases
  - Transformation-based learning
  - FnTBL
  - Largely language independent
  - Coreference: >90%

<table>
<thead>
<tr>
<th>Attribute</th>
<th>m- and a-layer:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>manual</td>
<td>auto</td>
<td></td>
</tr>
<tr>
<td>structure</td>
<td>89,3 %</td>
<td>76,4 %</td>
<td></td>
</tr>
<tr>
<td>functor</td>
<td>85,5 %</td>
<td>77,4 %</td>
<td></td>
</tr>
<tr>
<td>val_frame.rf</td>
<td>92,3 %</td>
<td>90,9 %</td>
<td></td>
</tr>
<tr>
<td>t_lemma</td>
<td>93,5 %</td>
<td>90,9 %</td>
<td></td>
</tr>
<tr>
<td>nodetype</td>
<td>94,5 %</td>
<td>92,6 %</td>
<td></td>
</tr>
<tr>
<td>gram/sempos</td>
<td>93,8 %</td>
<td>91,5 %</td>
<td></td>
</tr>
<tr>
<td>a/lex.rf</td>
<td>96,5 %</td>
<td>95,1 %</td>
<td></td>
</tr>
<tr>
<td>a/aux.rf</td>
<td>94,3 %</td>
<td>90,3 %</td>
<td></td>
</tr>
<tr>
<td>is_member</td>
<td>94,3 %</td>
<td>89,5 %</td>
<td></td>
</tr>
<tr>
<td>is_generated</td>
<td>96,6 %</td>
<td>95,2 %</td>
<td></td>
</tr>
<tr>
<td>deepord</td>
<td>68,0 %</td>
<td>66,7 %</td>
<td></td>
</tr>
</tbody>
</table>
Tectogrammatical Layer in Machine Translation

- The Translation (“Vauquois”) triangle

Tectogrammatical Representation

Surface Syntax

Morphology

Generation

Dependency trees in MT

According to his opinion UAL’s executives were misinformed about the financing of the original transaction.

Podle jeho názoru bylo vedení UAL o financování původní transakce nesprávně informováno.

Transfer:
- structure (~0)
- lexical
- functions
- grammatical
Valency and Translation

- leave:
  - leave-1
    - to leave [from] somewhere
  - leave-2
    - to leave sth for sb
- Translating (from English into Czech):
  - which equivalent to chose?
    - nechat vs. odjet/opustit
  - which prepositions, cases, ... to use?
    - accusative vs. “z” (“from”) with genitive vs. …?
To summarize…

- PDT is/has (a)…
  - Dependency-based treebanking project
    - Czech (other languages in the works – Eng, Ar)
  - ~ 1mil. words
    - sufficient size for ML experiments
  - 4 layers of annotation
    - token, morphology, syntax, deep syntax/semantics++
    - independent and full information at all levels, but...
    - interlinked (for the development of parsers/generators)
  - Valency dictionary integrated (links from data)

Some pointers

- Current version of PDT: v2.0 beta
  - all three levels, 1.9/1.5/0.8 Mwords
  - http://ufal.mff.cuni.cz/pdt2.0

- http://ufal.mff.cuni.cz
  - Projects -> Treebank

- http://www.ldc.upenn.edu
  - LDC2001T10 (PDT v1.0), LDC2004T23 (PADT 1.0), LDC2004T25 (PCEDT 1.0), LDC2006T01 (PDT 2.0)

  - Using TL for MT Generation