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 😊)
PDT: 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 support for Machine Learning
- Ergonomy of annotation
  - Graphical ("linguistic") presentation & editing
The Prague Dependency Treebank Project: Czech Treebank

- 1995 (Dublin) 1996-2006-2010-...
  - 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 (partial version) in 2004 (LDC)
  - PCEDT version 2.0 being prepared (2011)
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), grammattemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon
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
Morphological Attributes

- **Tag**: 13 categories
  - **Example**: AAFP\(^3\)----3N----
    - Adjective: no poss., Gender
    - Regular: no poss., Number
    - Feminine: no person
    - Plural: no tense
    - Dative: superlative
    - negated
    - no voice
    - reserve\(^1\)
    - reserve\(^2\)
    - base var.

- **Lemma**: POS-unique identifier
  - Books/verb -> book-1, went -> go, to/prep. -> to-1

Ex.: nejnezajímavějším
"(to) the most uninteresting"
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
  - Averaged Perceptron (M. Collins, EMNLP’02)
- All: ~ 94-96% accuracy (perceptron is best)
  - “COMPOST” (available for several languages)
### The Segmentation Problem: Arabic

- Tokenization / segmentation not always trivial

<table>
<thead>
<tr>
<th>String</th>
<th>Token</th>
<th>Token Tag</th>
<th>Buckwalter Morph Tags</th>
<th>Token Form</th>
<th>Token Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>سيخبرهم</td>
<td>F------</td>
<td>FUT</td>
<td></td>
<td>sa-</td>
<td>will</td>
</tr>
<tr>
<td>V1IA-3MS--</td>
<td>IV3MS+IV+IVSUFF_MOOD:1</td>
<td>yu-ḥbir-u</td>
<td>he-notify</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-----3MP4-</td>
<td>IVSUFF_DO:3MP</td>
<td>-hum</td>
<td>them</td>
<td></td>
<td></td>
</tr>
<tr>
<td>بذلك</td>
<td>P------</td>
<td>PREP</td>
<td></td>
<td>bi-</td>
<td>about/by</td>
</tr>
<tr>
<td>عن</td>
<td>SD-----MS--</td>
<td>DEM_PRON_MS</td>
<td></td>
<td>ḏālika</td>
<td>that</td>
</tr>
<tr>
<td>طريق</td>
<td>P------</td>
<td>PREP</td>
<td></td>
<td>ṭan</td>
<td>by/about</td>
</tr>
<tr>
<td>الرسائل</td>
<td>N------2R</td>
<td>NOUN+CASE_DEF_GEN</td>
<td></td>
<td>ṭarīq-i</td>
<td>way-of</td>
</tr>
<tr>
<td>القصيرة</td>
<td>N------2D</td>
<td>DET+NOUN+CASE_DEF_GEN</td>
<td>DET+ADJ+NSUFF_FEM_SG+</td>
<td>ar-rasāʾil-i</td>
<td>the-messages</td>
</tr>
<tr>
<td>والإنترنت</td>
<td>A------FS2D</td>
<td>DET+CASE_DEF_GEN</td>
<td>al-qaṣīr-at-i</td>
<td>the-short</td>
<td></td>
</tr>
<tr>
<td>وغيرها</td>
<td>C------</td>
<td>CONJ</td>
<td></td>
<td>wa-</td>
<td>and</td>
</tr>
<tr>
<td>Z------2D</td>
<td>DET+NOUN_PROP+</td>
<td>al-internet-i</td>
<td></td>
<td>wa-</td>
<td>and</td>
</tr>
<tr>
<td>C------</td>
<td>CONJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN------2R</td>
<td>NEG_PART+CASE_DEF_GEN</td>
<td>ǧayr-i</td>
<td>other/not-of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-----3FS2-</td>
<td>POSS_PRON_3FS</td>
<td>-ḥā</td>
<td>them</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
PDT-style
Arabic Surface Syntax

- Only several differences
  - (Sometimes) Separate nodes for individual segments (cf. tagging/segmentation)
  - Copula treatment (Czech: rare $\rightarrow$ treated as ellipsis; Arabic: systematic solution), Pred
  - (Added) analytic functions:
    - AuxM $\lambda m\check{a}$ (did-not)
    - Ante $m\check{a}$ (what)
- Work by Faculty of Arts (Arabic language) students
In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it.
English Analytic Layer

- By conversion from PTB
  - Extended analytic functions

- Head rules
  - Jason Eisner’s, added more for full conversion
    - Coordination, traces, etc.

- Coordination handling
  - Same as in Czech/Arabic PDT
Penn Treebank

- University of Pennsylvania, 1993
  - Linguistic Data Consortium
- Wall Street Journal texts, ca. 50,000 sentences
  - 1989-1991
  - Financial (most), news, arts, sports
  - 2499 (2312) documents in 25 sections
- Annotation
  - POS (Part-of-speech tags)
  - Syntactic “bracketing” + bracket (syntactic) labels
  - (Syntactic) Function tags, traces, co-indexing
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Penn Treebank Example: Sentence Tree

- Phrase-based tree representation:
Parallel Czech-English Annotation

- English text -> Czech text (human translation)
- Czech side (goal): all layers manual annotation
- English side (goal):
  - Morphology and surface syntax: technical conversion
    - Penn Treebank style -> PDT Analytic layer
  - Tectogrammatical annotation: manual annotation
    - (Slightly) different rules needed for English
- Alignment
  - Natural, sentence level only (now)
Human Translation of WSJ Texts

- Hired translators / FCE level
- Specific rules for translation
  - Sentence per sentence only
    - …to get simple 1:1 alignment
  - Fluent Czech at the target side
  - If a choice, prefer “literal” translation
- The numbers:
  - English tokens: 1,173,766
  - Translated to Czech:
    - Revised/PCEDT 1.0: 487,929
    - Now finished (all 2312 documents)
English Annotation
POS and Syntax

- Automatic conversion from Penn Treebank
  - PDT morphological layer
    - From POS tags
  - PDT analytic layer
    - From:
      - Penn Treebank Syntactic Structure
      - Non-terminal labels
      - Function tags (non-terminal “suffixes”)
  - 2-step process
    - Head determination rules
    - Conversion to dependency + analytic function
Head Determination Rules

- Exhaustive set of rules
  - By J. Eisner + M. Cmejrek/J. Curin
  - 4000 rules (non-terminal based)
    - Ex.: (S (NP-SBJ VP .)) → VP
- Additional rules
  - Coordination, Apposition
  - Punctuation (end-of-sentence, internal)
- Original idea (possibility of conversion)
  - J. Robinson (1960s)
Example: Head Determination Rules (J.E.)

Rules:

\[(NP \ (DT \ NN)) \rightarrow NN\]
\[(VP \ (VB \ NP)) \rightarrow VB\]
\[(VP \ (MD \ VP)) \rightarrow VP\]
\[(S \ (\ldots \ VP \ \ldots)) \rightarrow VP\]
Example: Analytical Structure, Functions

Penn Treebank structure (with heads added)

PDT-like Analytic Representation
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

- 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

Deep function

Underlying verb + tense

Elided Actor in

Another ellipsis...

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
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...
Tectogrammatical Functors

syntactic semantic

- “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, ...; TWHEN, 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

Additional attributes (grammatemes):
- conditional + “allow”
Tectogrammatical Example

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

Disappeared

Added
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
Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...
Deep Word Order
Topic/Focus

- Example:

- Baker bakes rolls. vs. $Baker^{IC}$ bakes rolls.
Layer 3: Tectogrammatical

- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference
  - 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 ...

- John
- ACT
- promise
- PRED
- go
- PAT
- he
- ACT
- home
- DIR3
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
  - 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!)
The boundaries of some problems seem to be clearer after they were revived by Havel’s speech.
Arabic Example: Tectogrammatics

- In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it.
English PDT-style Annotation

- Morphology and Syntax
  - By conversion
- Tectogrammatical annotation
  - Manual (English TR: by S. Cinková)
  - Pre-annotation
    - Transformation from Penn Treebank & Propbank (Palmer, Kingsbury) by Z. Žabokrtský et al.
  - Valency
    - From Propbank Frame Files (Cinková, Šindlerová, Nedolužko, Semecký)
  - The annotation is finished now (Nov. 2010; 1 mil. words)
Valency in the PDT

Valency: *specific ability of a word to combine itself with other units of meaning*

---

**Valency**

- **dát (give)**
  - **Eva** (ACT)
  - **matka (mother)** (ADDR)
  - **dar (gift)** (PAT)
  - **neděle (Sunday)** (TWHEN)

- **pršet (rain)**
  - **zítra (tomorrow)** (TWHEN)

- **plakat (cry)**
  - **Adam** (ACT)
  - **noc (night)** (TWHEN)
Valency - Basic Principles

inner participants vs. free modifications
(arguments vs. adjuncts)

obligatory vs. optional modifications
(the dialogue test)
Inner Participant ...

... Free Modification

- ACT(or), PAT(ient), ADDR(essee), EFF(ect), ORIG(in) (5)
  - each occurs just with particular verbs
  - each modifies the verb only once (in a clause)

- Location (LOC, DIR1, ...), Time (TWHEN, TTILL, ...), Manner, Intention, ... (70)
  - can modify in principle any verb
  - can be repeated (within the same clause)
Obligatory … Optional

The Dialogue Test

Answering a question about a semantically obligatory modification, the speaker cannot say: *I don't know.*

A: John left.
B: From where?
A: *I don't know.

A: John left.
B: To where?
A: *I don't know.*

„from where“
⇒ obligatory modification

„to where“
⇒ optional modification
Valency frame

Structure:

<table>
<thead>
<tr>
<th></th>
<th>obligatory</th>
<th>optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>argument</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adjunct</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Contents:

- functor
- obligatoriness
- surface form

one meaning of the word → one valency frame

word: leave
meaning 1: sb left sth
meaning 2: sb left from somewhere

frame1: ACT PAT
frame2: ACT DIR1
Valency lexicon: PDT-VALLEX

- 8500 verb senses / valency frames
- 9000 noun sense / valency frames
- some adjectives and adverbs

PDT-VALLEX Entry
verb: dosáhnout
meaning 1: to reach sth
meaning 2: to get sb to do sth
meaning 3: ...
meaning 4: ...

* dosáhnout
ACT(1) PAT(2,4) v-w714f1 Used: 272x
dosáhnout určité úrovně
mzdá v tomto oboru 80 tisíc
d. pokročilého věku

ACT(1) PAT(2,aby[v]) ?ORIG(na-1.[6],od-1.[2]) v-w714f2 Used: 7x
dosáhl na něm sľubu
dosáhl na sobě sľub

ACT(1) DPHR(svůj-1.2) v-w714f3 Used: 2x
dosáhl svého
dosáhl na strop

ACT(1) DIR3(*) v-w714f4 Used: 2x
růku.MEANS
The PDT-VALLEX editor

'stress' senses:

lay down
resign
win
ask
Valency Lexicon and TrEd

to write sth (about sth)
**Corpus** – occurrences of „uzavřít“ (to close):

**Lexicon:**

**ENTRY:** uzavřít

\[ \text{vf}_1: \text{ACT(.1) CPHR(\{smlouva\}.4)} \]

ex: u. dohodu (*close a contract*)

\[ \text{vf}_2: \text{ACT(.1) PAT(.4)} \]

ex.: u. pokoj (*close a room, house*)
Valency and Text Generation

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

```
starat_se
PRED
Martin
ACT

tygr
PAT
“tiger”
```

**Example:**

```
staratV

o

tygr
“to take care of”

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

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

```
“Martin takes care of tigers.”
```

"Martin se stará o tygry."

“Martin takes care of tigers.”
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
The Annotation Scheme

- XML + principles of linear- and tree-based standoff annotation

  $$\Rightarrow \text{PML}$$

  (Prague Markup Language)

- Layer schemes (Relax NG)
  - PDT/PADT: t(ecto), a(nalytic), m(orphology), …
  - English: + phrase-based (p-layer)
PML/XML Annotation Layers

- Strictly top-down links
- w+m+a can be easily "knitted"
- API for cross-layer access (programming)
- PML Schema / Relax NG
- [z and audio layers: used for spoken data (audio as layer "-1")]

LFG analogy:

- f-struct
- c-struct

audio

BYL BYS ČELO LESA ...
PDT 2.0: The Data

- Data sizes
The Translation ("Vauquois") triangle

- Tectogrammatical Representation
- Surface Syntax
- Morphology
- Generation

Transfer from source to target language (Cz to En)
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.
Valency and Translation

- leave-1 → nechat-3
  - ACT() PAT() LOC() → ACT(.1) PAT(.4) LOC()

- leave-2 → odjet-1
  - ACT() DIR1(from.) → ACT(.1) DIR1(z.[.2])
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, LDC2006T01
  - all three levels, 1.9/1.5/0.8 Mwords
  - http://ufal.mff.cuni.cz/pdt2.0

- http://ufal.mff.cuni.cz
  - Research -> Corpora (Treebank(s))

- http://ufal.mff.cuni.cz/pedt
  - Deep syntax (TR) of Penn Treebank texts

- 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