Tree-based Translation with Tectogrammatical Representation

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

- The Translation ("Vauquois") triangle

- "Deep" Syntax (Tectogrammatics)
- Surface Syntax
- Morphology

(source) (transfer) (target)

(Cz) (En)
Machine Translation Architecture

- Tectogrammatical layer-based system:

Transfer

(tectogrammatical) parsing

parsing

morphology (tagging)

source sentence

morphological layer

morph. synthesis

morphology (tagging)

word layer

linearization

target sentence

analytical layer

morph. synthesis

source sentence

analytical layer

Transfer

(tectogrammatical) parsing
Analytical Layer Correspondence
The [Homestead’s] only remaining baker bakes the most famous rolls to the north of Long River:

The Additional Steps

- Analytical (surface) → Tectogrammatical
  - additional parsing required

- Transfer
  - minimal effort: only “true” transformations needed (like swimming ~ schwimmen gern)

- Generation
  - back from Tectogrammatical representation to Analytical (surface syntax)
The Devil’s in ...

- The additional three steps:

  - Source sentence
  - Morphological layer
  - Analytical layer
  - Tectogrammatical layer
  - Generation
    - Linearization (trivial)
    - Morph. synthesis (easy)
  - Transfer
    - Parsing
    - Morphology (tagging)
    - Tectogrammatical parsing
Zooming In ...

• The additional three steps:

- Tectogrammatical parsing
- (Simple) transfer
- Tectogrammatical layer
- Generation:
  - Deletions
  - Insertions: prepositions, conjunctions, ...
  - Word order
  - Morphology
- Source
- Analytical layer
- Target
Tectogrammatical Parsing

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>manual</th>
<th>auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>structure</td>
<td>89.3%</td>
<td>76.4%</td>
</tr>
<tr>
<td>functor</td>
<td>85.5%</td>
<td>77.4%</td>
</tr>
<tr>
<td>val_frame.rf</td>
<td>92.3%</td>
<td>90.9%</td>
</tr>
<tr>
<td>t_lemma</td>
<td>93.5%</td>
<td>90.9%</td>
</tr>
<tr>
<td>nodetype</td>
<td>94.5%</td>
<td>92.6%</td>
</tr>
<tr>
<td>gram/sempos</td>
<td>93.8%</td>
<td>91.5%</td>
</tr>
<tr>
<td>a/lex.rf</td>
<td>96.5%</td>
<td>95.1%</td>
</tr>
<tr>
<td>a/aux.rf</td>
<td>94.3%</td>
<td>90.3%</td>
</tr>
<tr>
<td>is_member</td>
<td>94.3%</td>
<td>89.5%</td>
</tr>
<tr>
<td>is_generated</td>
<td>96.6%</td>
<td>95.2%</td>
</tr>
<tr>
<td>deepord</td>
<td>68.0%</td>
<td>66.7%</td>
</tr>
</tbody>
</table>
Generation

- Components:
  - Deletions of nodes [rare if going into English]
  - Insertions of nodes
    - prepositions, conjunctions, punctuation
    - splitting phrases/idioms/named entities
  - Tree reorganization (numeric expressions)
  - Surface word order (analytical tree: defined w.o.)
  - Morphology (agreement, cases based on subcat)
  - English, Czech
Example Translation

- Insertion of Prepositions

[Diagram showing the analytical and tectogrammatical layers with nodes and arrows indicating the insertion of prepositions.]
Example Translation

- Surface word order

**TWHEN ACT**

- **analytical layer**

- **tectogrammatical layer**
The Data: Parallel, Annotated Treebank

- Parallel corpora
  - Comparative/contrastive and translation studies
  - Semantics
  - Other “linguistic research goals”

- Machine Translation
  - “Training” material
    - Human-translated texts
  - Testing material
    - Evaluation – human, automatic
The Prague Czech-English Dependency Treebank

- “PCEDT”
- One of “family” of PDT-like treebanks
  - Wall Street Portion of the Penn Treebank, ver. III
  - Czech translation (manual) of the above
- Size
  - 1.2 million words, ~50,000 sentences
- Annotation
  - All 4 layers as in PDT: tokens, morphology, syntax, tectogrammatical representation
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:
PDT Layers of Annotation

- Tectogrammatical structure
- Surface syntax
- Morphology
- Tokens (words)
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

```
Rules:
(NP (DT NN)) → NN
(VP (VB NP)) → VB
(VP (MD VP)) → VP
(S (… VP …)) → VP
```
Conversion: Analytic Structure, Functions

- Analytic Function assignment (conversion)
- Rules
  - based on functional tags:
    - SBJ Sb
    - BNF Obj
    - LGS Obj
    - DIR Adv
    - LOC Adv
    - PRP Adv
    - TMP Adv
    - PRD Pnom
    - DTV Obj
    - ADV Adv
    - EXT Adv
    - MNR Adv
    - PUT Adv
  - Ad-hoc rules (if functional tags missing)
  - Lemmatization (years → year)
Example: Analytical Structure, Functions

Penn Treebank structure (with heads added)

PDT-like Analytic Representation
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)

- Volume annotation starting now
Czech PDT-style Annotation

- All layers
  - (morphology, analytic, tectogrammatical)
- So far…
  - Automatic (many tools by many authors)
- Manual annotation
  - Started
    - revised guidelines: M. Mikulová, J. Štěpánek
  - Top-down
    - Tectogrammatical first (lower layers automatically)
    - … then analytic structure and morphology
Those current holders would also receive minority interests in the new company.
According to his opinion UAL's executives were misinformed about the financing of the original transaction.
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.
PCEDT 1.0 – The CD

- Published 2004 by the LDC (LDC2004T25)
- Texts, size of data:
  - 480,000 words: parallel annotated WSJ treebank (Cz: auto)
    - 21,600 sentences
  - 2 mil. words (53,000 sent.): Reader’s Digest short stories
  - Evaluation data (5 reference translations, 500 sent.)
- Tools
  - GIZA++ (Statistical Machine Translation Toolkit)
  - Scripts for easy training (“SMT Quick Run”)
  - Probabilistic dictionary (46,150 words, lemmatized)
    - Czech – English (WSJ and other sources)
- Euromatrix & other projects: PCEDT 2.0 (2008)
PCEDT – some pointers

- **PCEDT 1.0**
  - [http://www.ldc.upenn.edu](http://www.ldc.upenn.edu) catalog No. LDC2004T25

- **PDT 2.0 (Czech annotation - documentation)**
  - [http://www.ldc.upenn.edu](http://www.ldc.upenn.edu) catalog No. LDC2006T01
  - [http://ufal.mff.cuni.cz/pdt2.0](http://ufal.mff.cuni.cz/pdt2.0)

- **Cinkova: English Tectogrammatics**
  - [http://acl.ldc.upenn.edu/W/W06/W06-0612.pdf](http://acl.ldc.upenn.edu/W/W06/W06-0612.pdf)