Tree-based Translation with Tectogrammatical Representation

Jan Hajič

Institute of Formal and Applied Linguistics

School of Computer Science

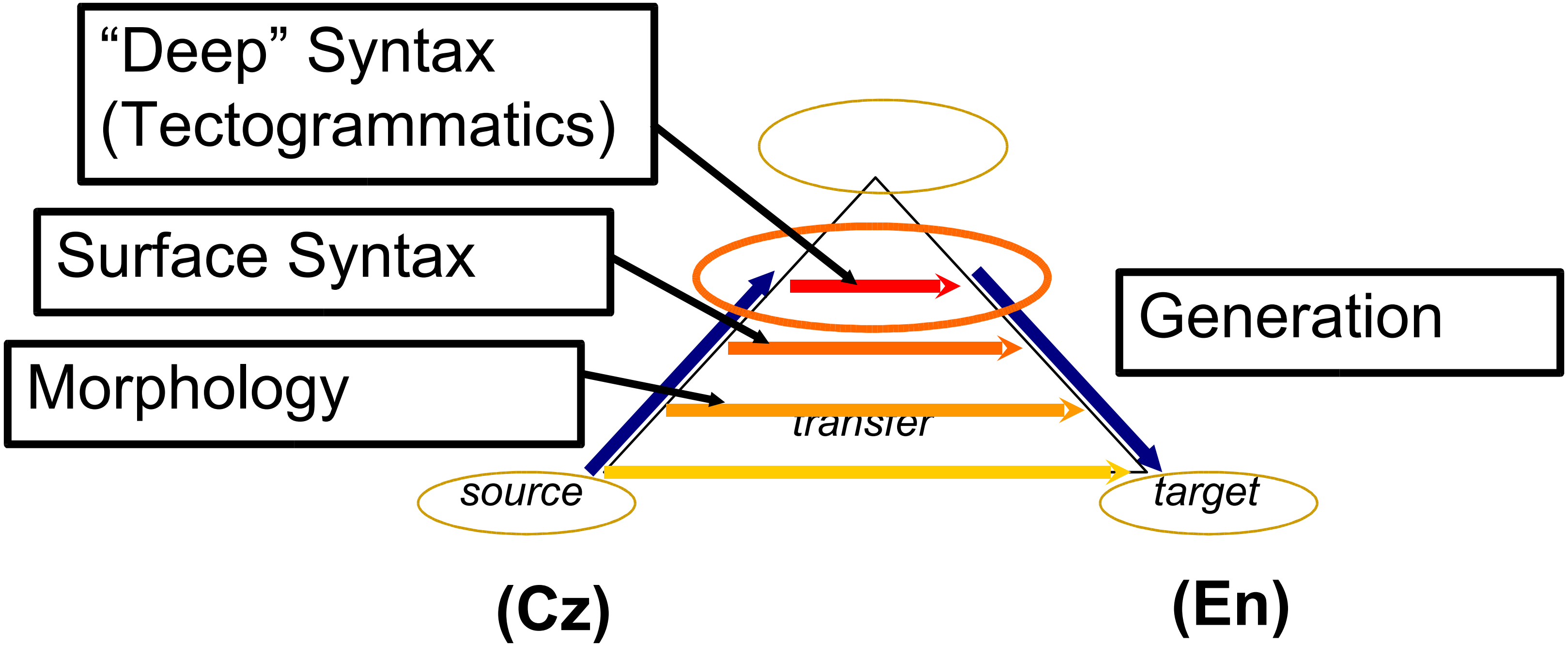
Faculty of Mathematics and Physics

Charles University, Prague

Czech Republic

Standard Scheme of Machine Translation

- The Translation ("Vauquois") triangle



"Deep" Syntax (Tectogrammatics)

Surface Syntax

Morphology

Generation

transfer

source

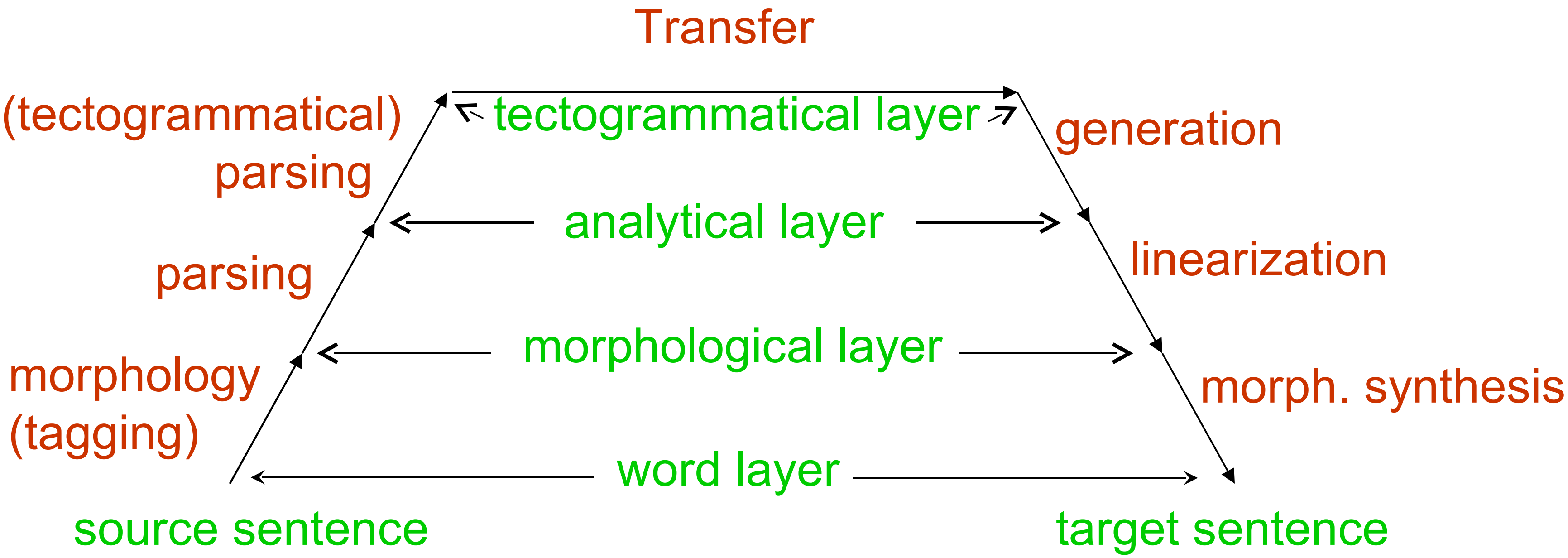
target

(Cz)

(En)

Machine Translation Architecture

- Tectogrammatical layer-based system:



Transfer

(tectogrammatical) parsing

tectogrammatical layer

generation

parsing

analytical layer

linearization

morphology (tagging)

morphological layer

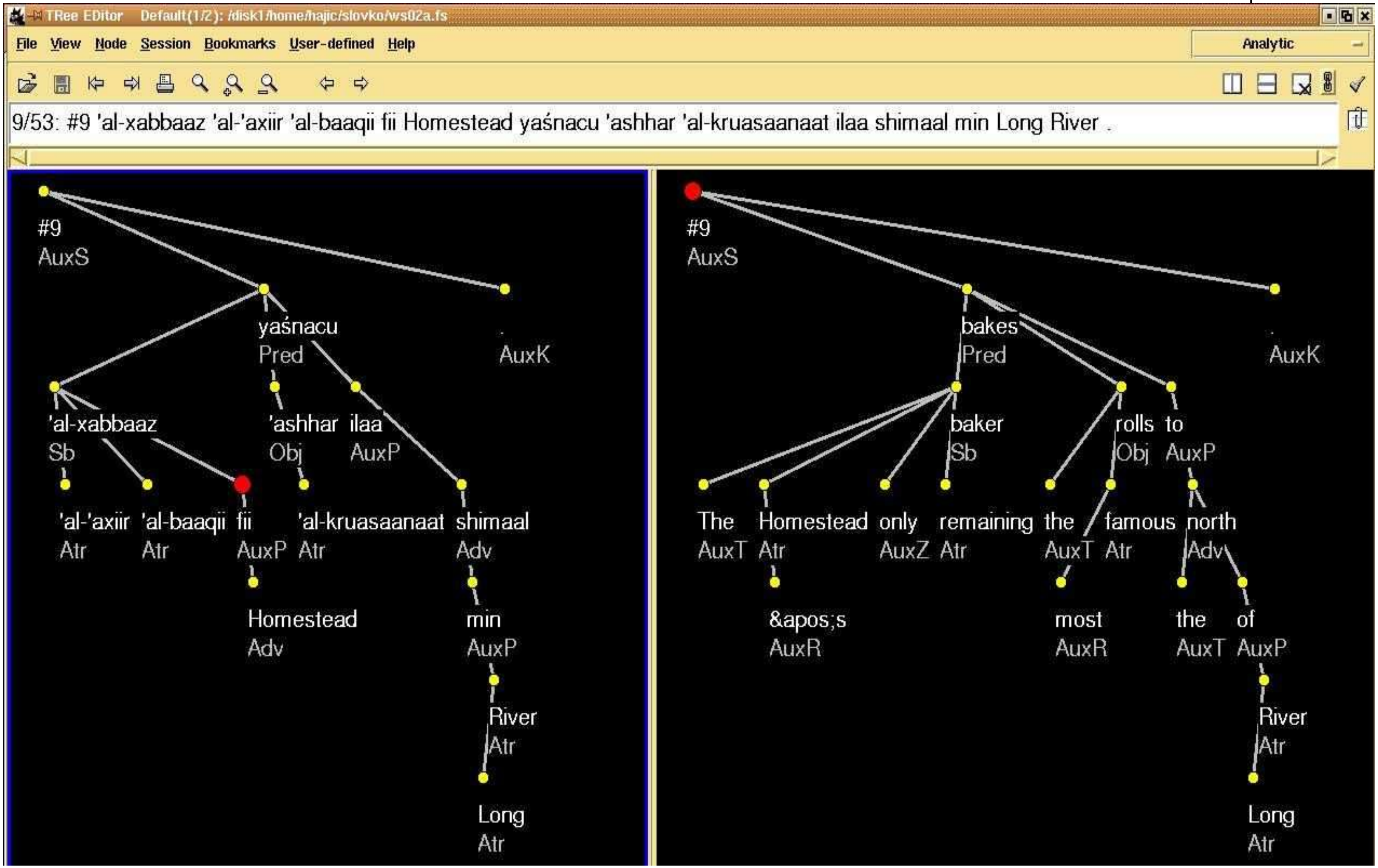
morph. synthesis

word layer

source sentence

target sentence

Analytical Layer Correspondence

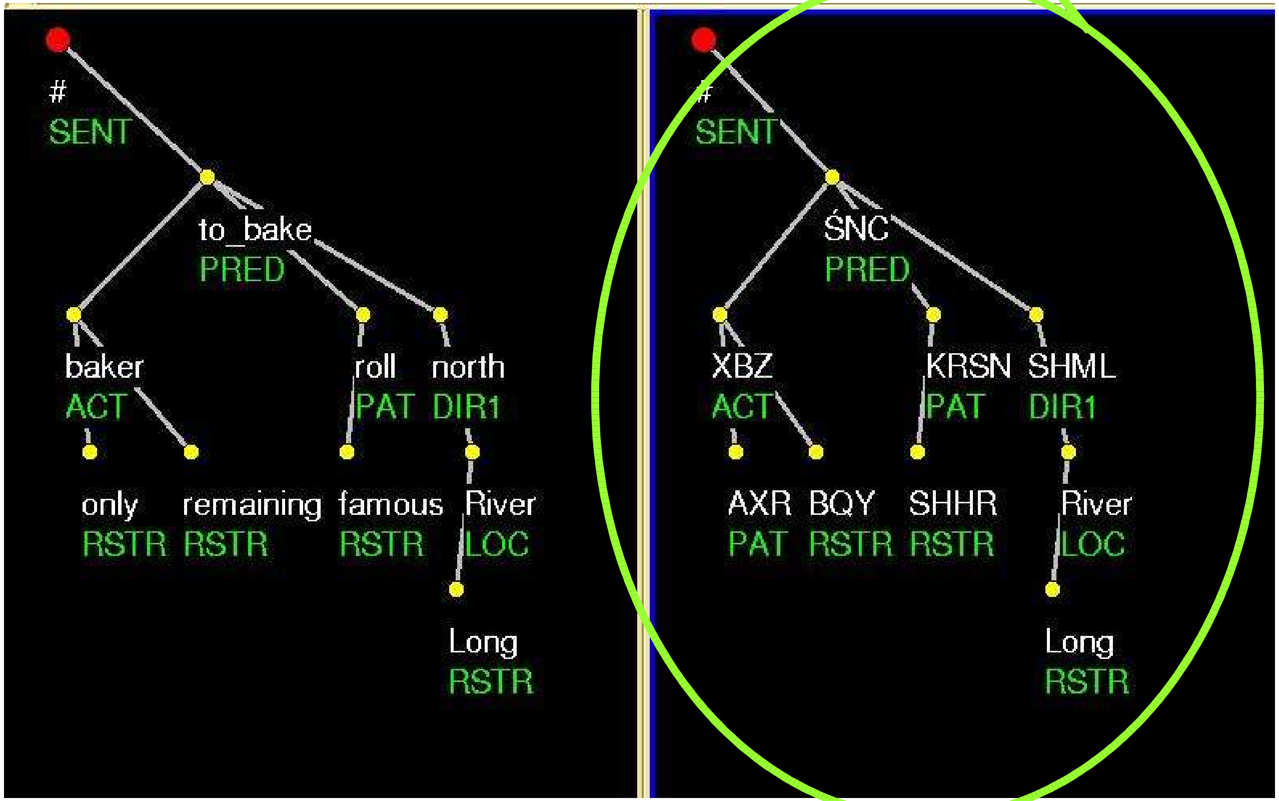


9/53: #9 'al-xabbaaz 'al-'axiir 'al-baaqii fii Homestead yaśnacu 'ashhar 'al-kruasaanaat ilaa shimaal min Long River .

Tectogrammatical Correspondence

The [Homestead's] only remaining baker bakes the most famous rolls to the north of Long River.

'al-xabaaz 'al-'axiir 'al-baaqii [fii Homestead] yaśmacu 'ashhar 'al-kruasaanaat ilaa shimaal min Long River.

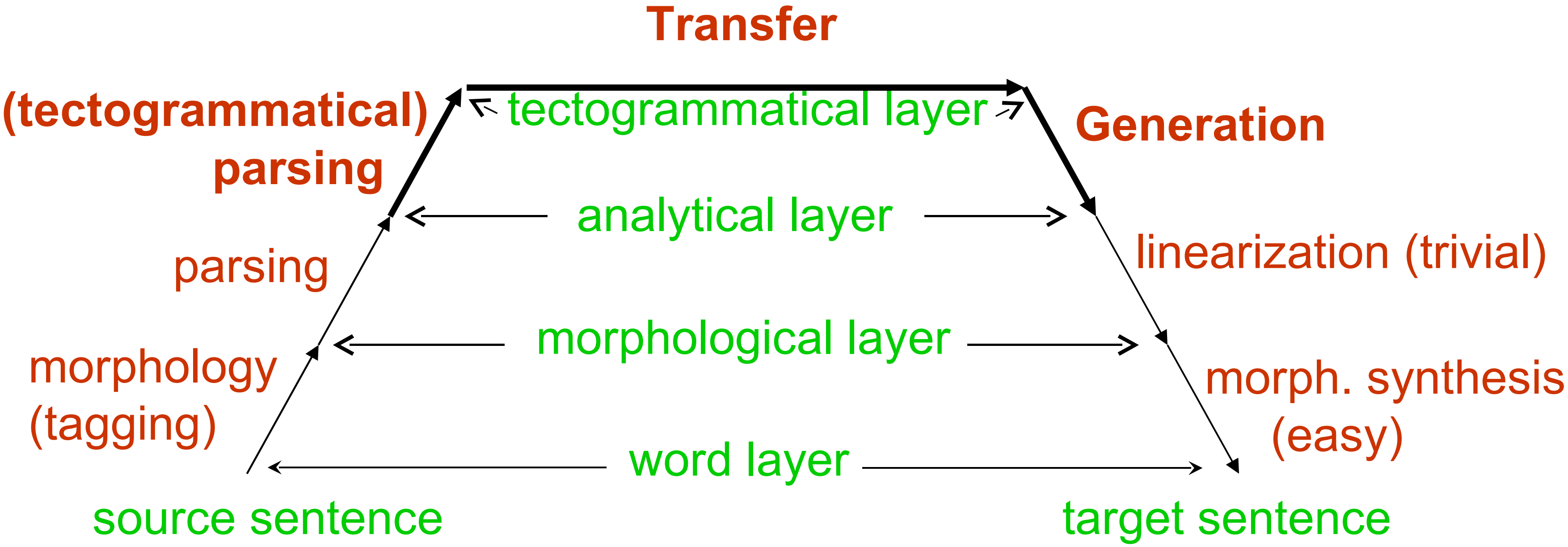


The Additional Steps

- 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)

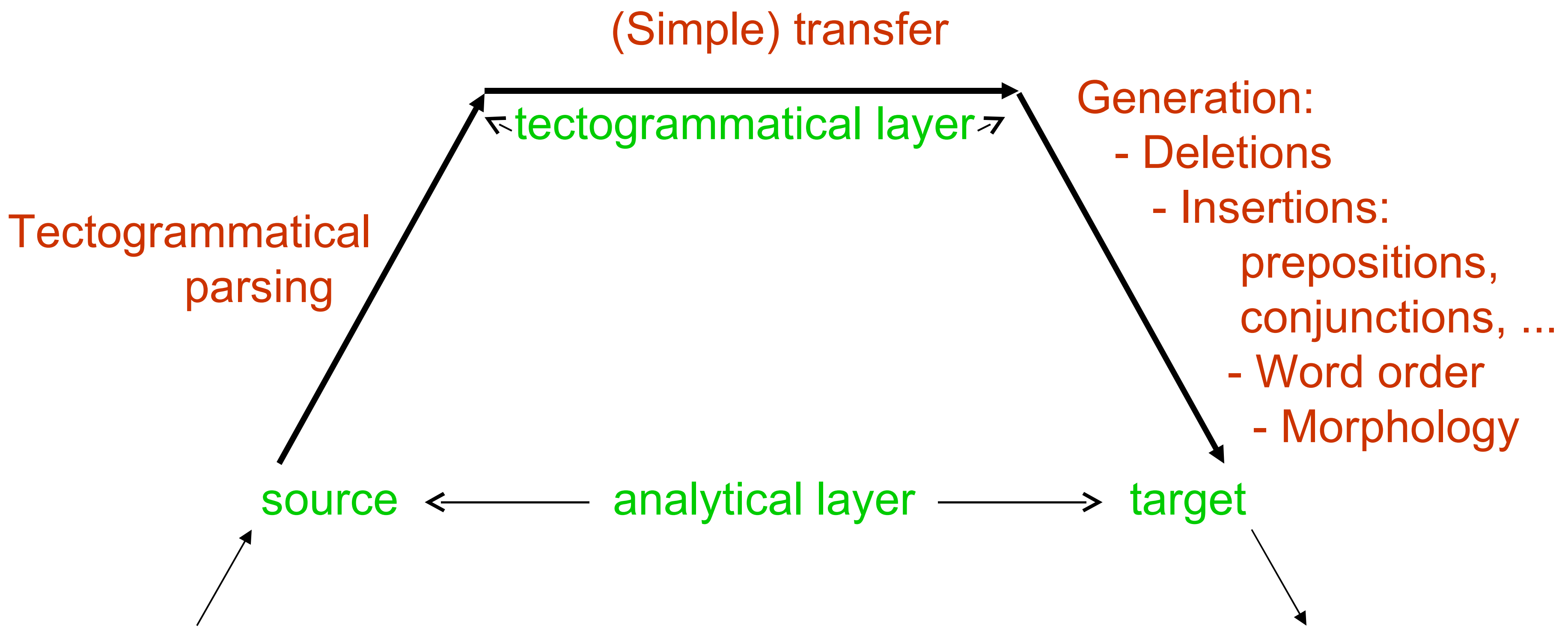
The Devil’s in ...

- The additional three steps:



Zooming In ...

- The additional three steps:



Tectogrammatical Parsing

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

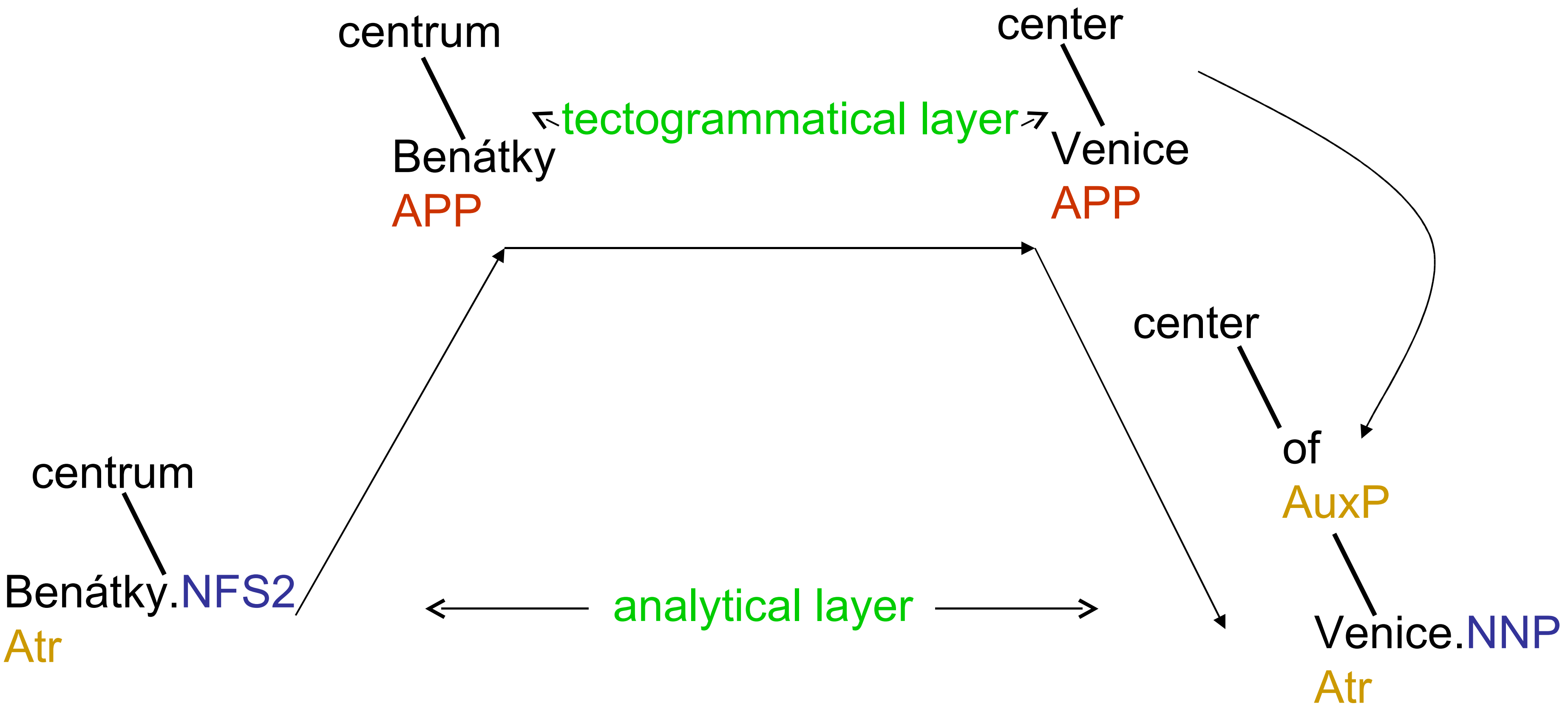
	m- and a-layer:	
Attribute	manual	auto
structure	89,3 %	76,4 %
`functor`	85,5 %	77,4 %
`val_frame.rf`	92,3 %	90,9 %
`t_lemma`	93,5 %	90,9 %
`nodetype`	94,5 %	92,6 %
`gram/sempos`	93,8 %	91,5 %
`a/lex.rf`	96,5 %	95,1 %
`a/aux.rf`	94,3 %	90,3 %
`is_member`	94,3 %	89,5 %
`is_generated`	96,6 %	95,2 %
`deepord`	68,0 %	66,7 %

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



centrum

Benátky
APP

tectogrammatical layer

center

Venice
APP

center

of
AuxP

centrum

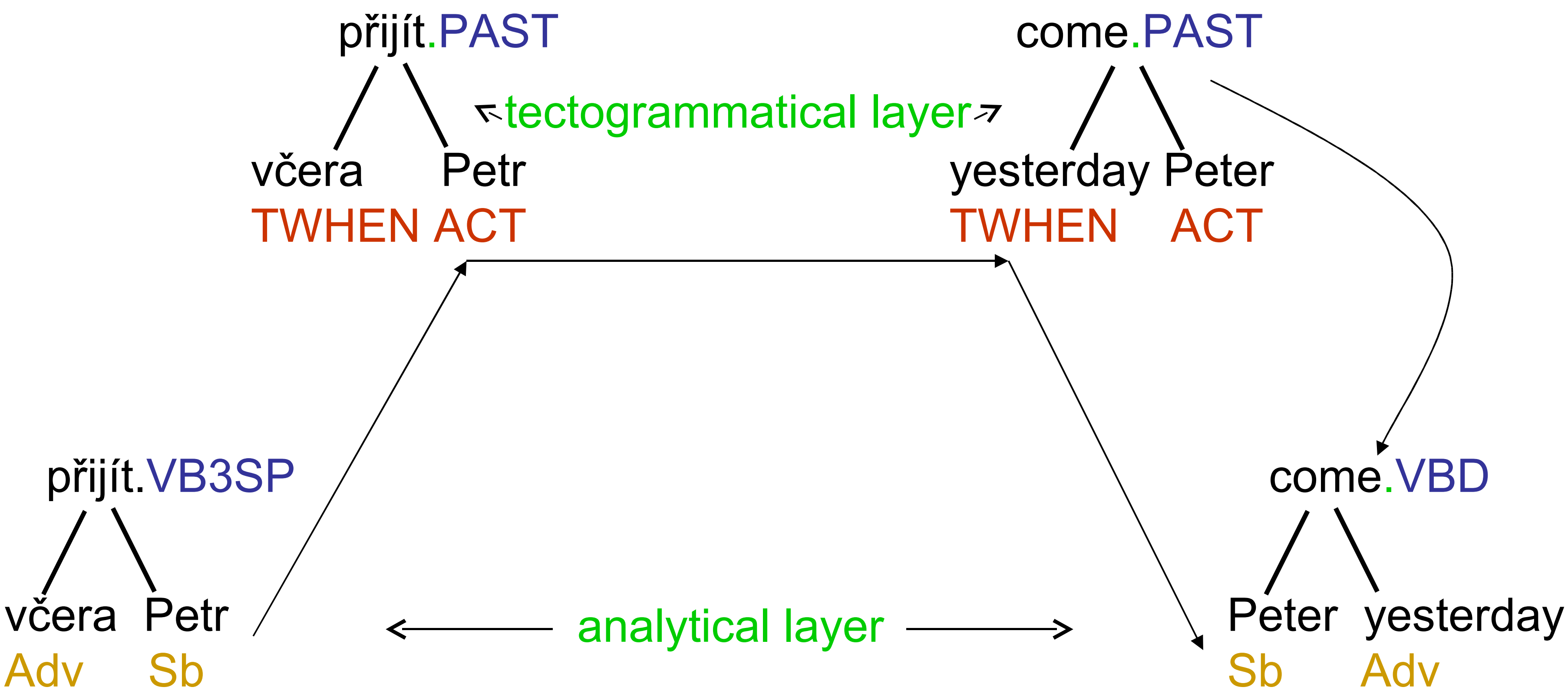
Benátky.NFS2
Atr

analytical layer

Venice.NNP
Atr

Example Translation

- Surface word order



přijít.PAST

včera TWHEN

Petr ACT

tectogrammatical layer

come.PAST

yesterday TWHEN

Peter ACT

přijít.VB3SP

včera Adv

Petr Sb

analytical layer

come.VBD

Peter Sb

yesterday Adv

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

Penn Treebank Example

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

- “Preterminal”
- POS tag (NNS)
- (noun, plural)

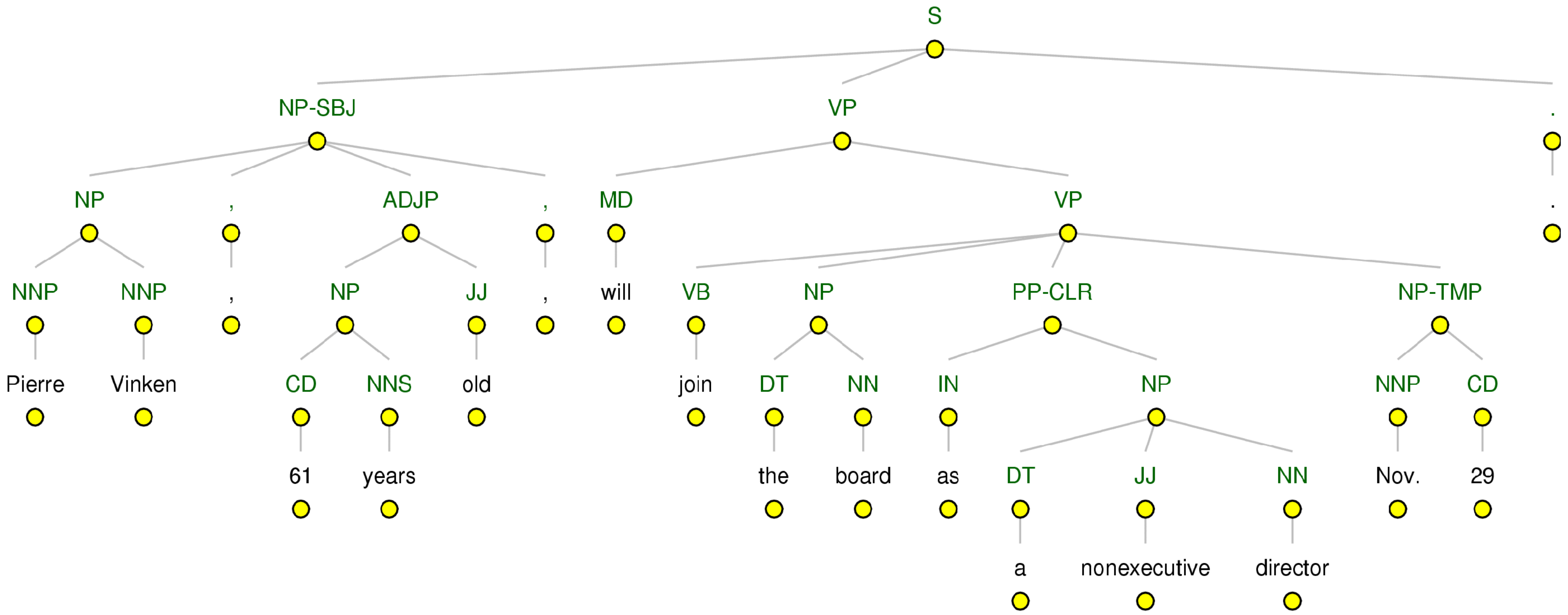
- Noun Phrase

- Phrase label (NP)

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

Penn Treebank Example: Sentence Tree

- Phrase-based tree representation:



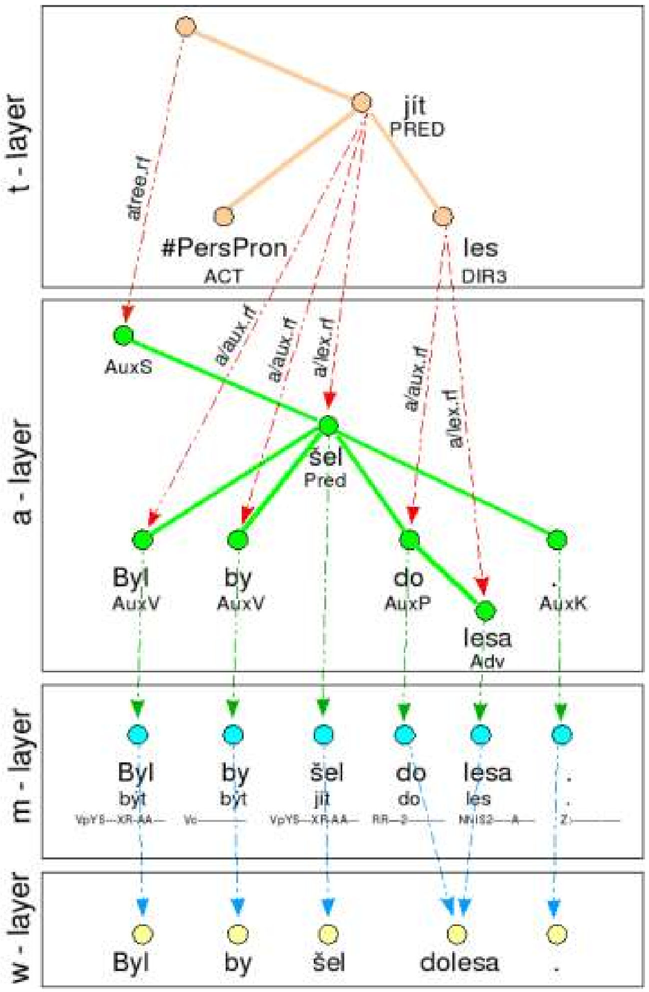
PDT Layers of Annotation

- 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)

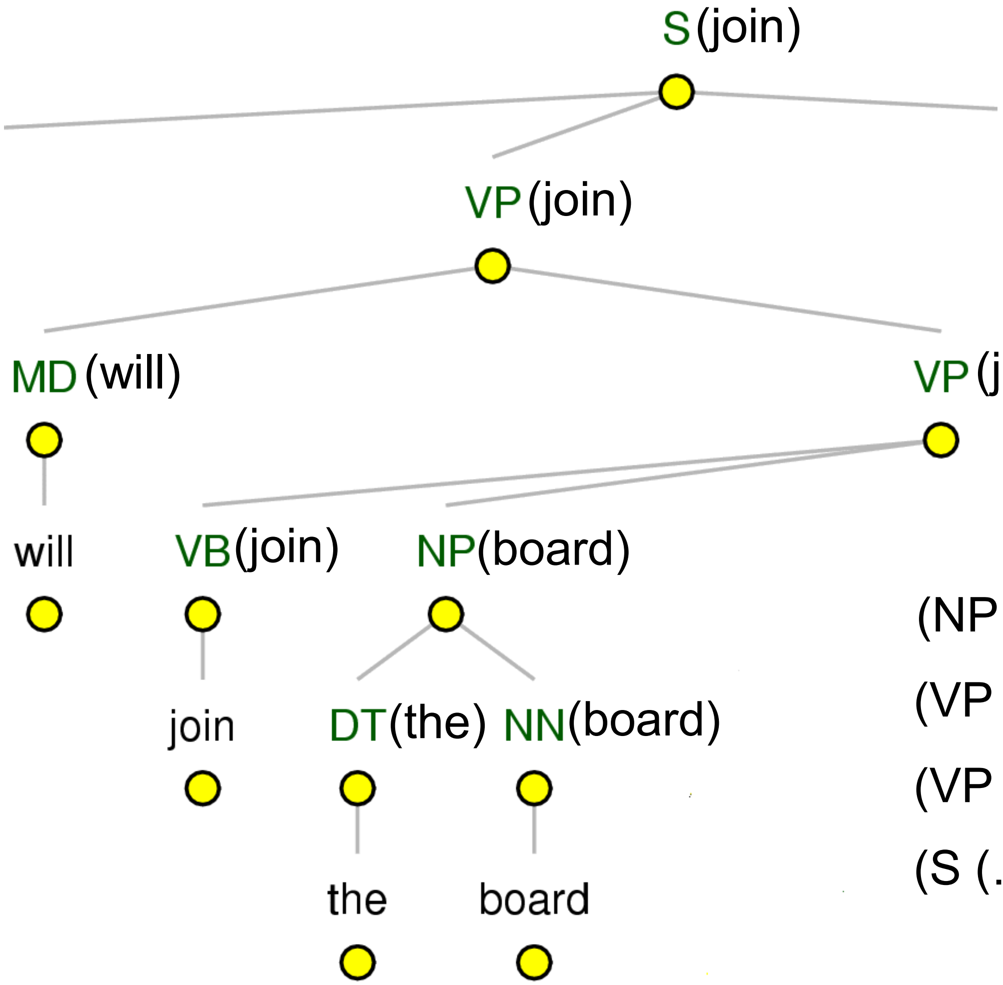
English Annotation POS and Syntax

- 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

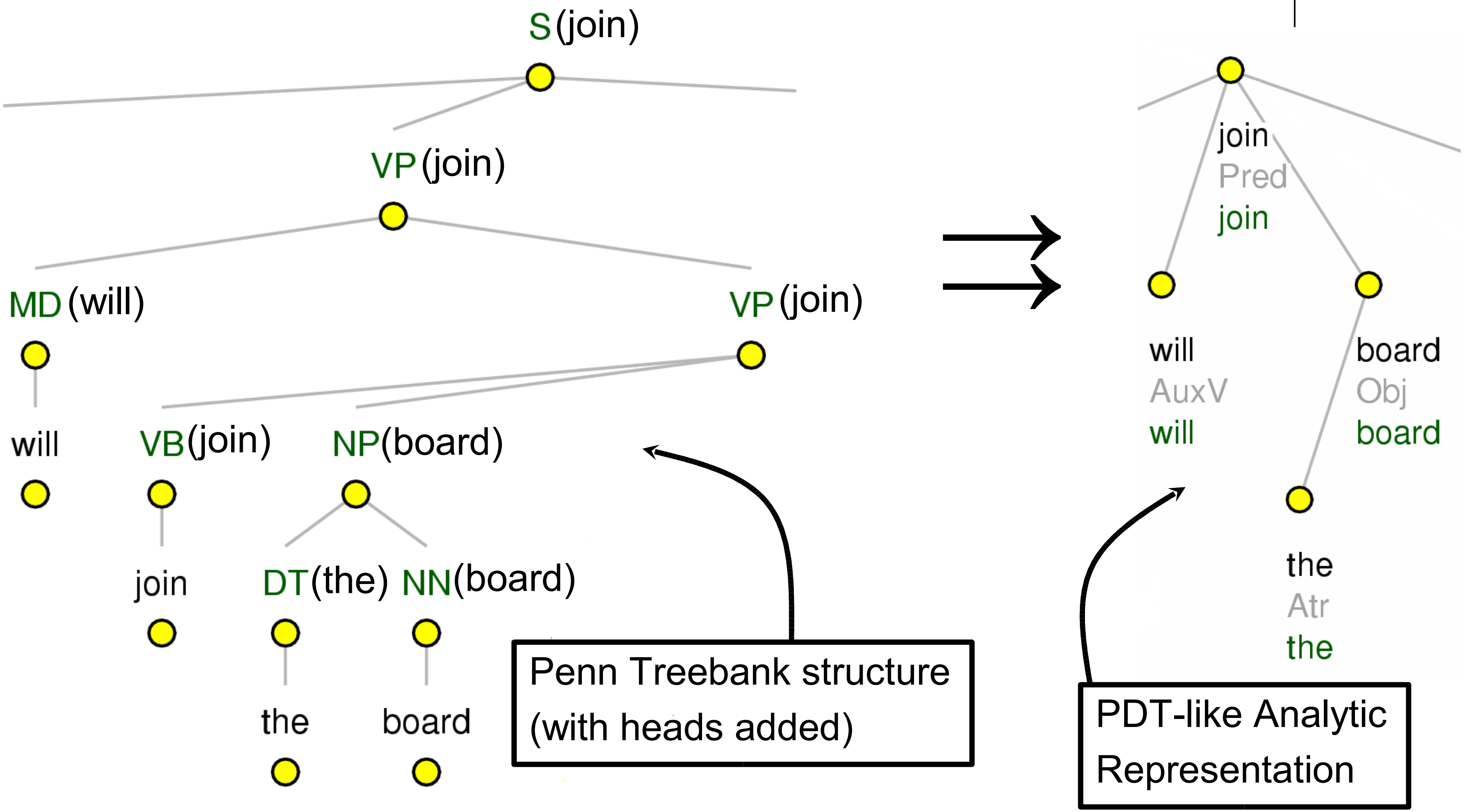
Conversion: Analytic Structure, Functions

- Analytic Function assignment (conversion)
- Rules
  - based on functional tags:

-SBJ	Sb	-PRD	Pnom
-BNF	Obj	-DTV	Obj
-LGS	Obj	-ADV	Adv
-DIR	Adv	-EXT	Adv
-LOC	Adv	-MNR	Adv
-PRP	Adv	-PUT	Adv
-TMP	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

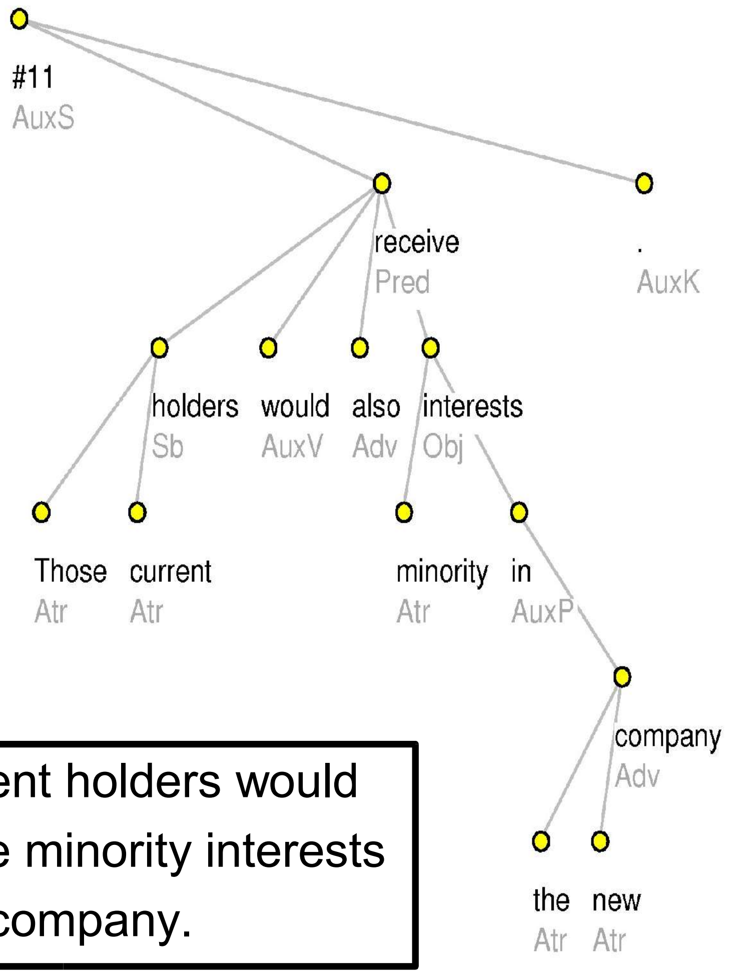
English PDT-style Annotation

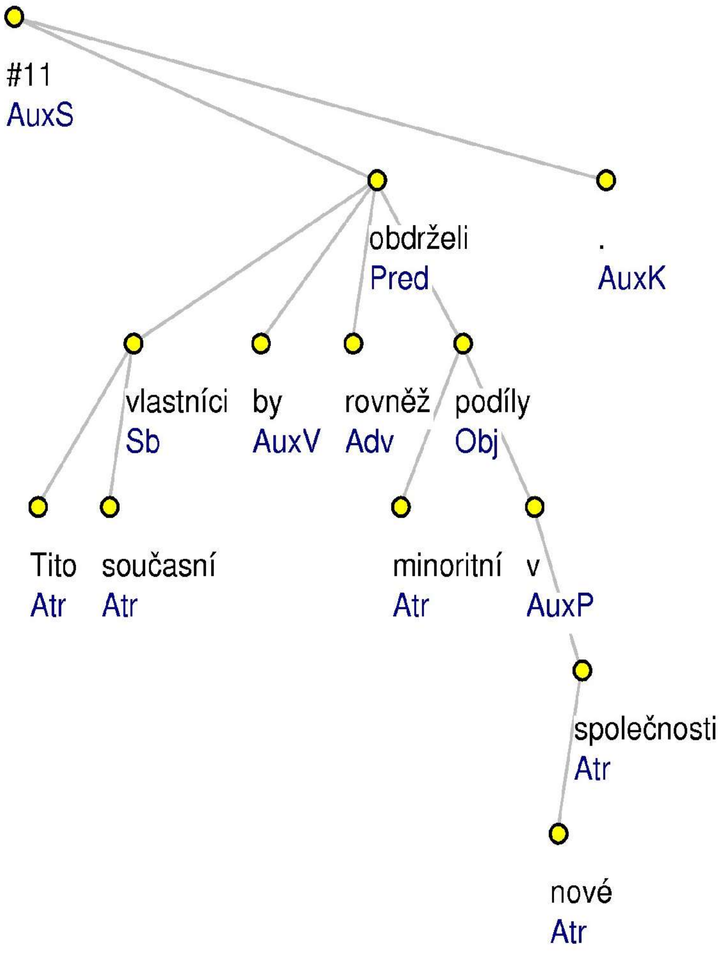
- Morphology and Syntax
  - By conversion
- Tectogrammatical annotation
  - Manual (English TR: by S. Cinková)
  - Pre-annotation
    - Transformation from Penn Treebank & Propbank (Palmer, Kingsbury) by Z. Žabokrtský et al.
  - Valency
    - From Propbank Frame Files (Cinková, Šindlerová, Nedolužko, 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

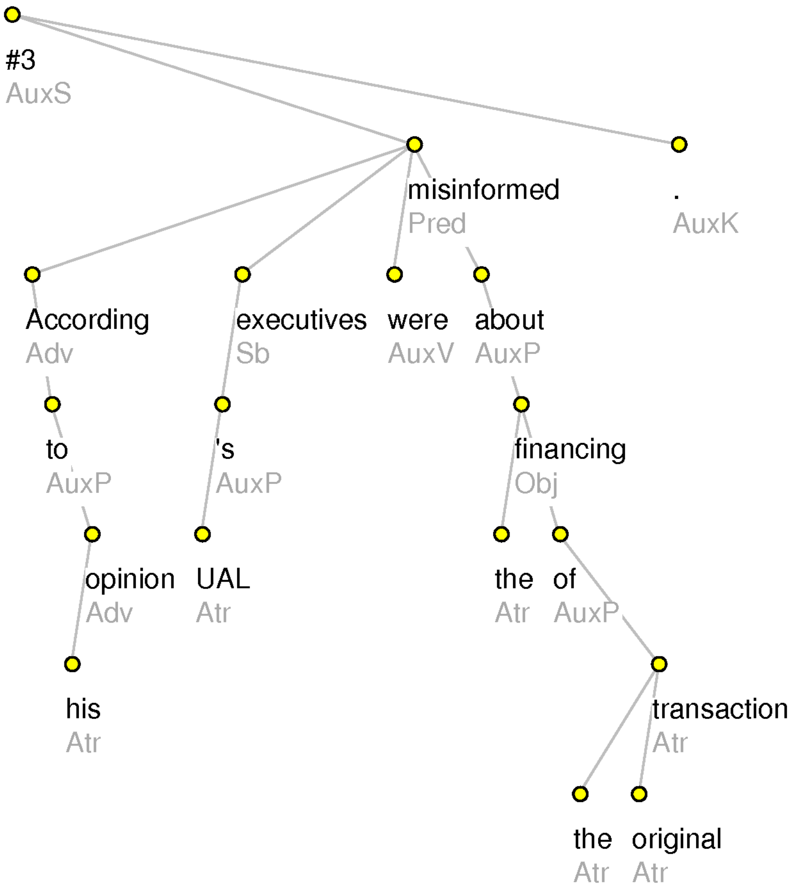
Analytical Pair En – Cz

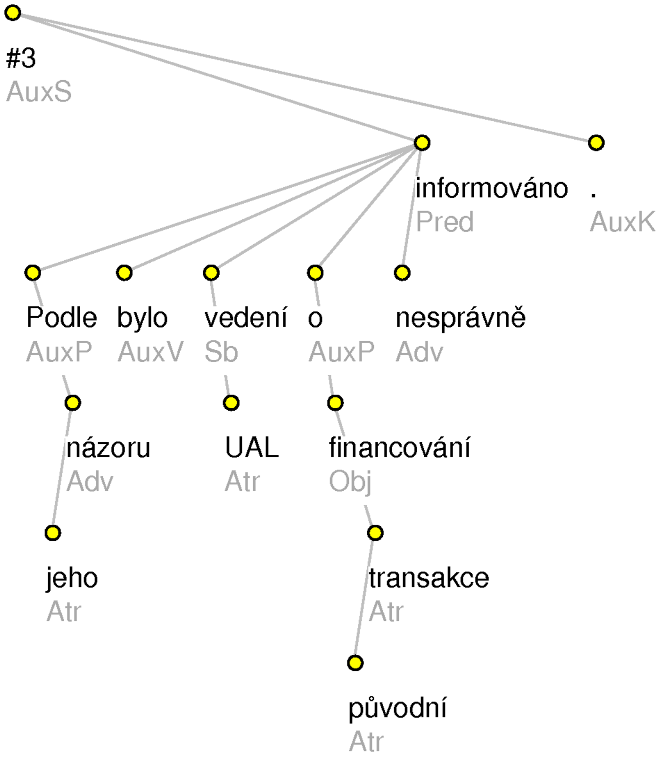




- Those current holders would
- also receive minority interests
- in the new company.

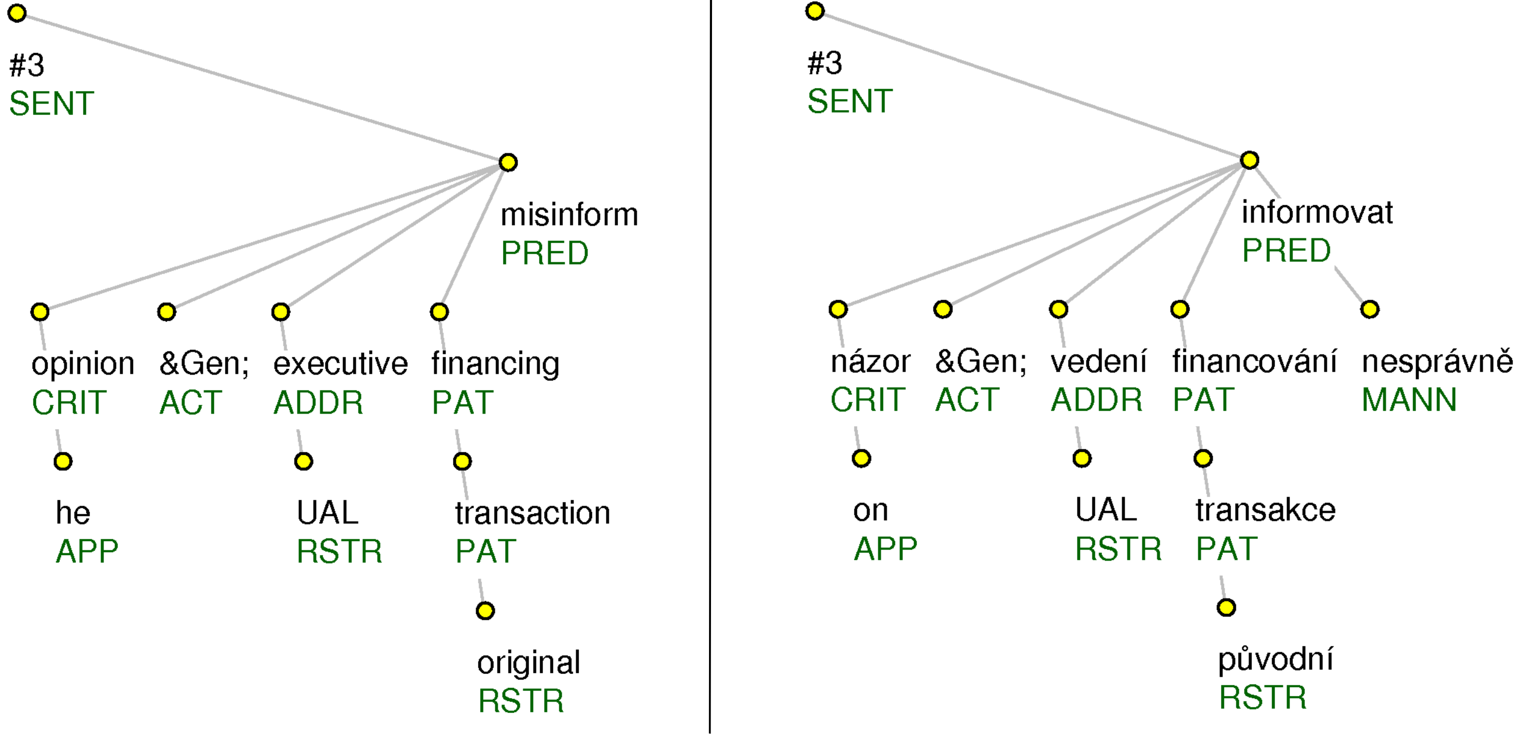
Analytical Pair En - Cz





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

Tectogrammatical Pair



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 catalog No. LDC2004T25
  - http://ufal.mff.cuni.cz/pcedt
- PDT 2.0 (Czech annotation - documentation)
  - http://www.ldc.upenn.edu catalog No. LDC2006T01
  - http://ufal.mff.cuni.cz/pdt2.0
- Cinkova: English Tectogrammatics
  - http://ufal.mff.cuni.cz/~cinkova/TR_En.zip
  - http://acl.ldc.upenn.edu/W/W06/W06-0612.pdf