The Prague Czech-English Dependency Treebank (part 8.1)

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

The Goal: 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 PCEDT

- One of “family” of PDT-like treebanks
- Texts:
  - 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
  - 1989-1991
  - Financial (most), news, arts, sports
  - 2499 documents in 25 sections
- Annotation
  - POS (Part-of-speech tags)
  - Syntactic “bracketing” + bracket (syntactic) labels
  - (Syntactic) Function tags

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)

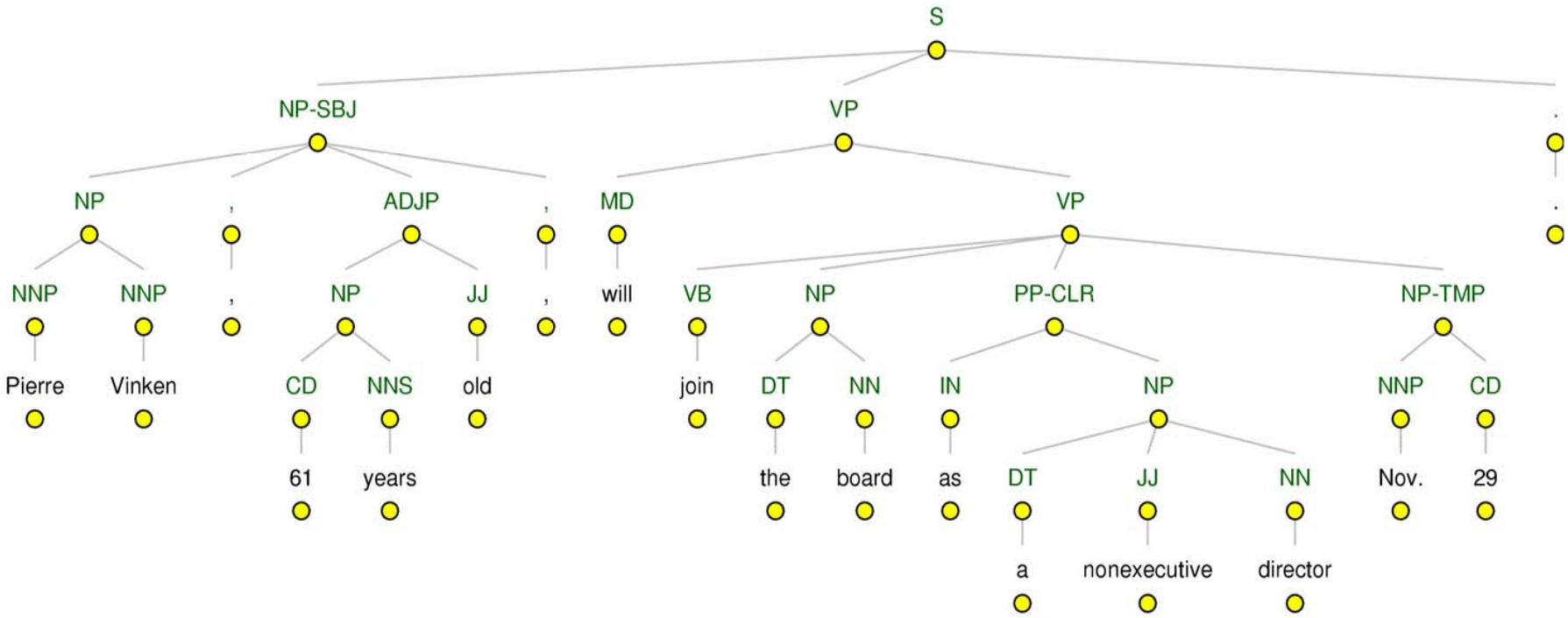
Noun Phrase

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:



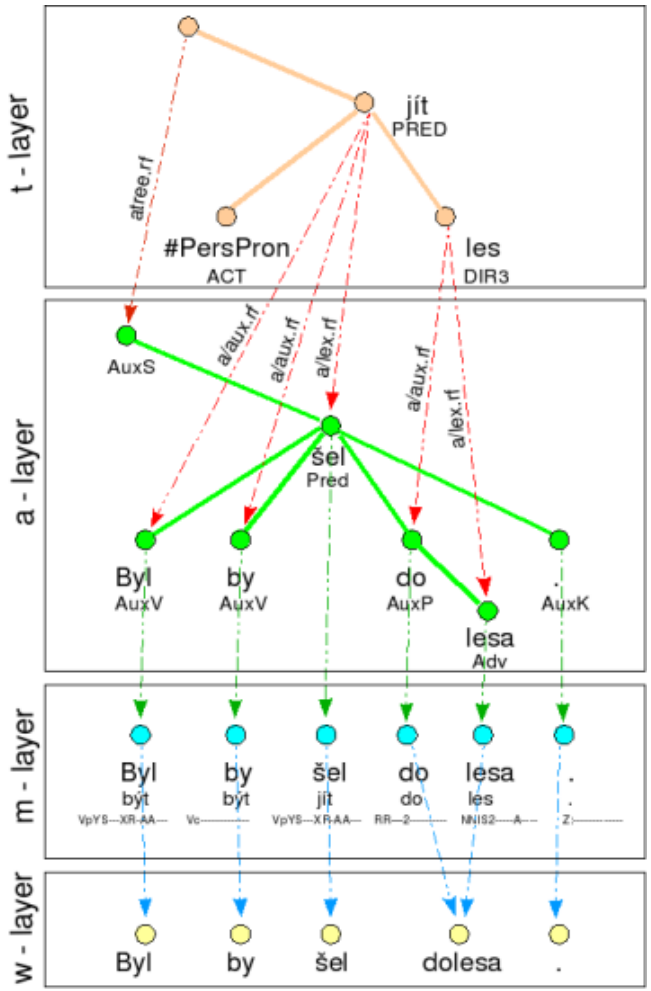
PDT Layers of Annotation

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: 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: 1173766
  - Translated to Czech:
    - Revised/PCEDT 1.0: 487929
    - (now: 1097471)

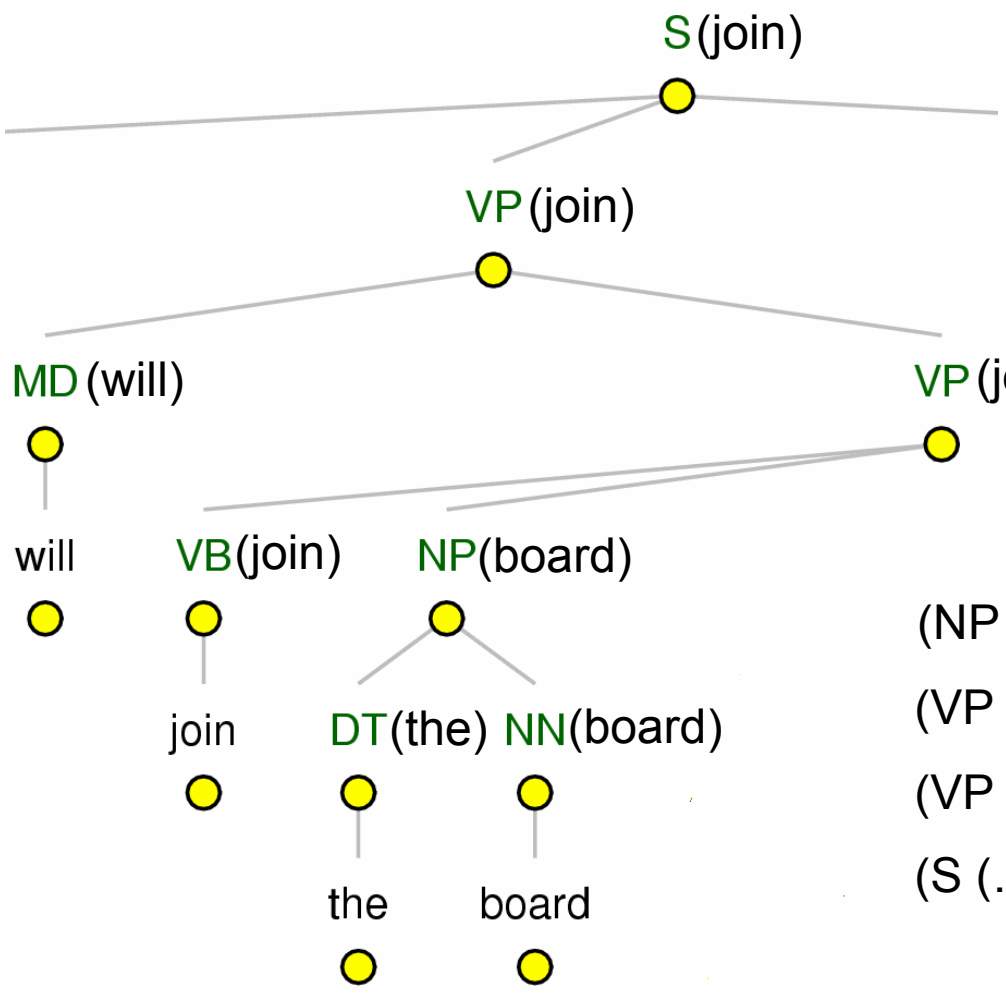
English Annotation
POS and Syntax

- 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

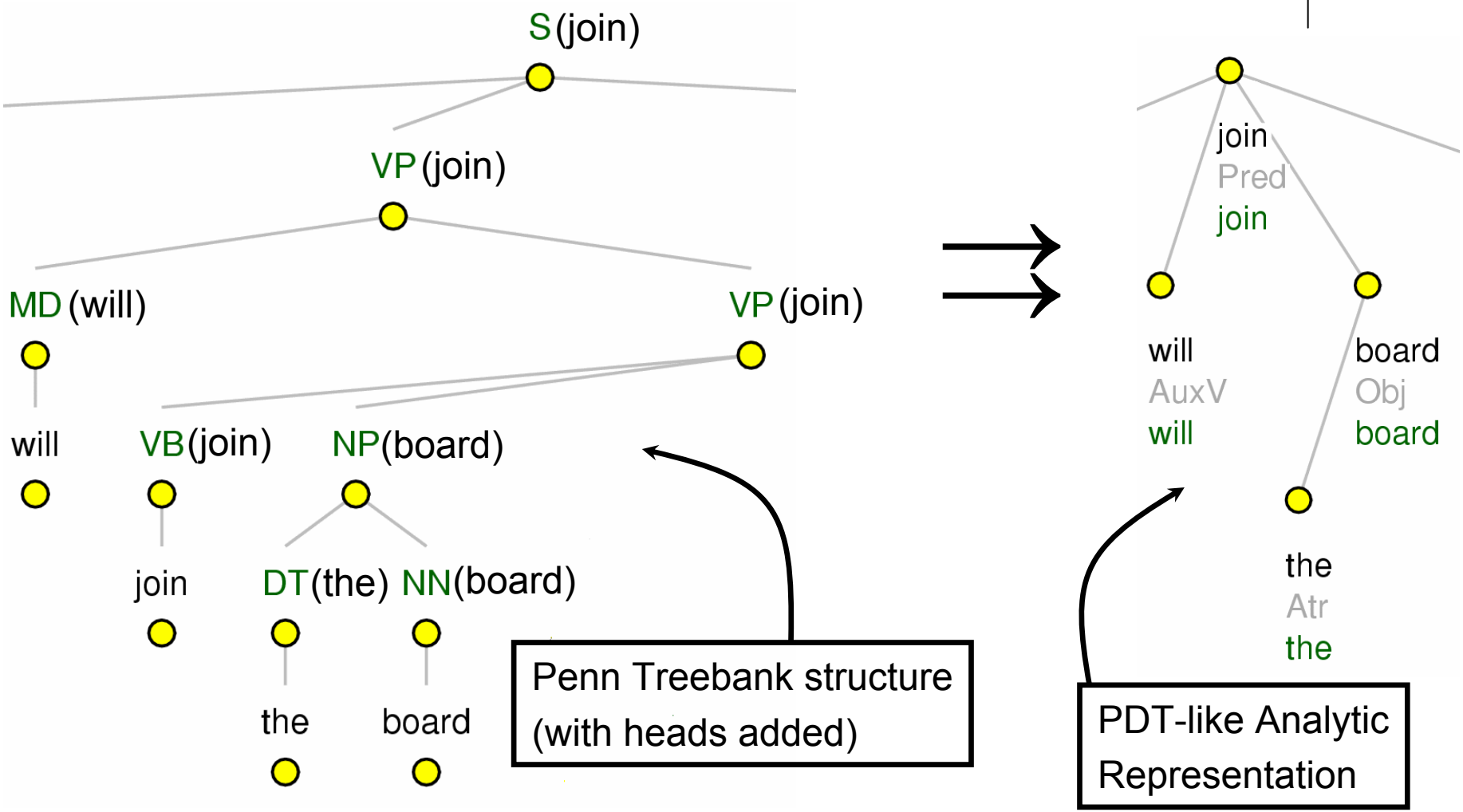
Conversion to Analytic Structure, Functions

- 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: Conversion to Analytical Structure, Functions



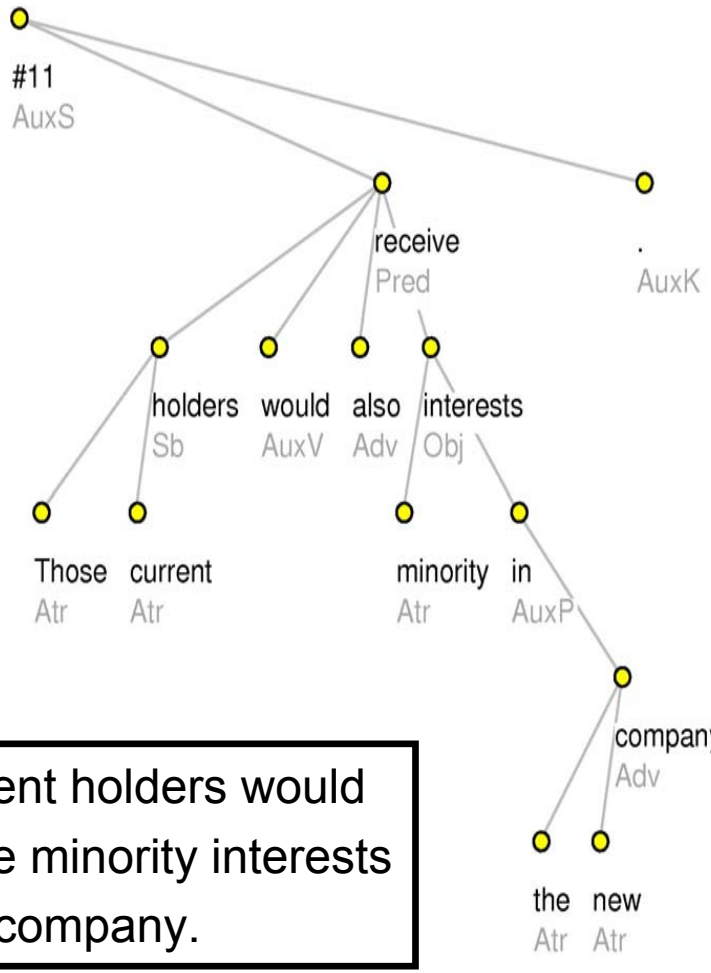
English PDT-style Annotation

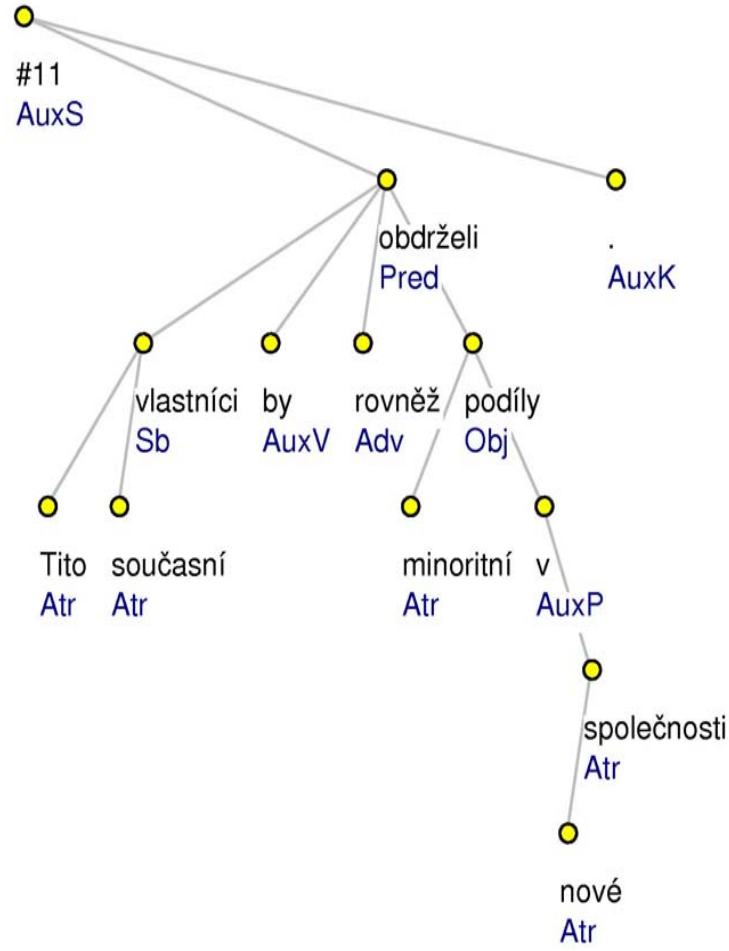
- Morphology and Syntax
  - By conversion
- Tectogrammatical annotation
  - Manual
  - Pre-annotation
    - Transformation from Penn Treebank & Propbank (Palmer, Kingsbury)
  - Valency
    - From Propbank Frame Files
- Starting now

Czech PDT-style Annotation

- All layers
  - (morphology, analytic, tectogrammatical)
- So far…
  - Automatic
- Manual annotation
  - Starting now
  - Top-down
    - Tectogrammatical first (lower layers automatically)
    - … then analytic structure and morphology

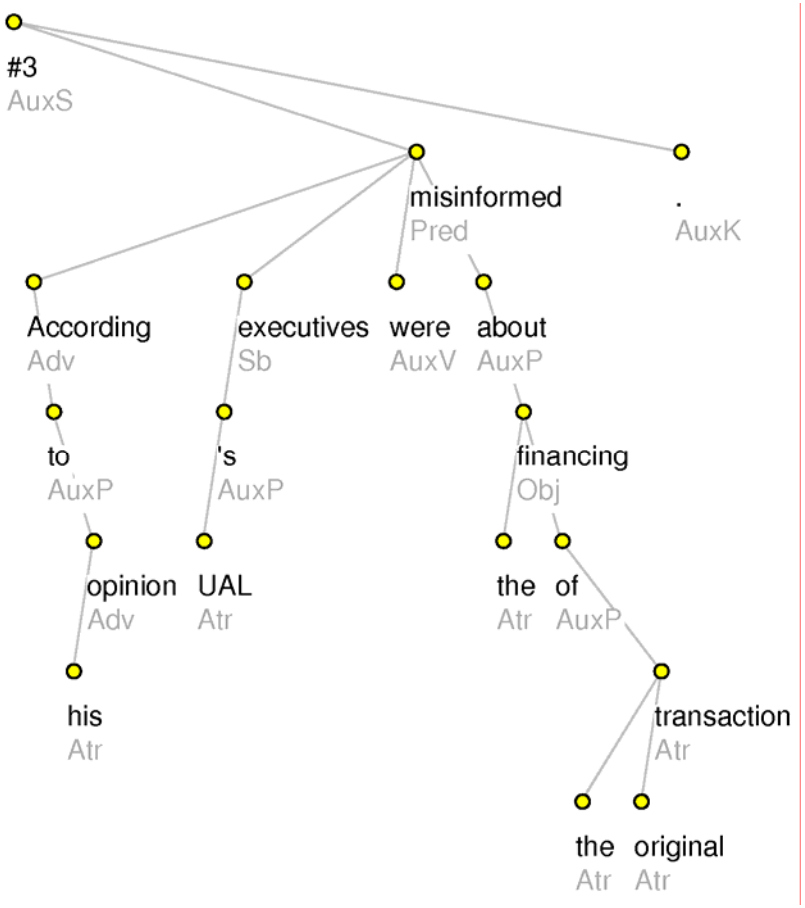
Analytical Pair En – Cz

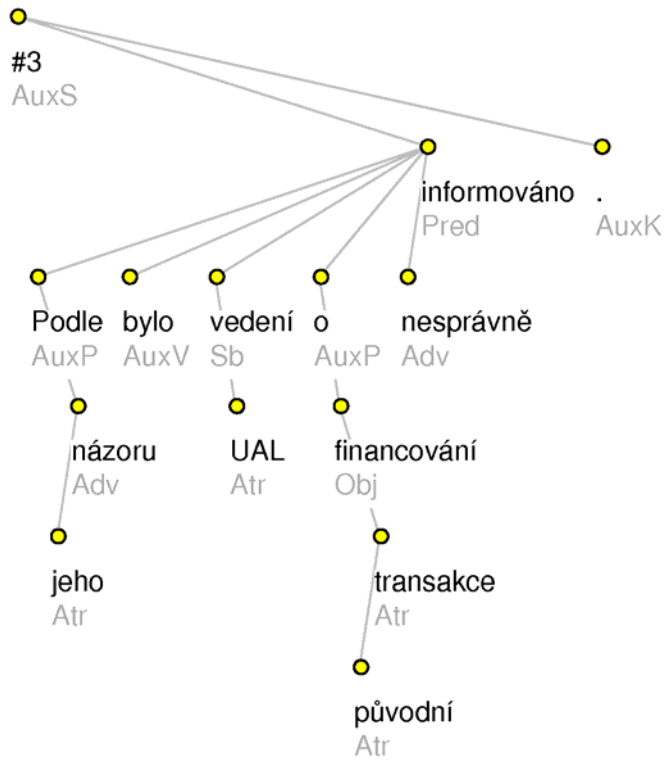




Those current holders would also receive minority interests in the new company.

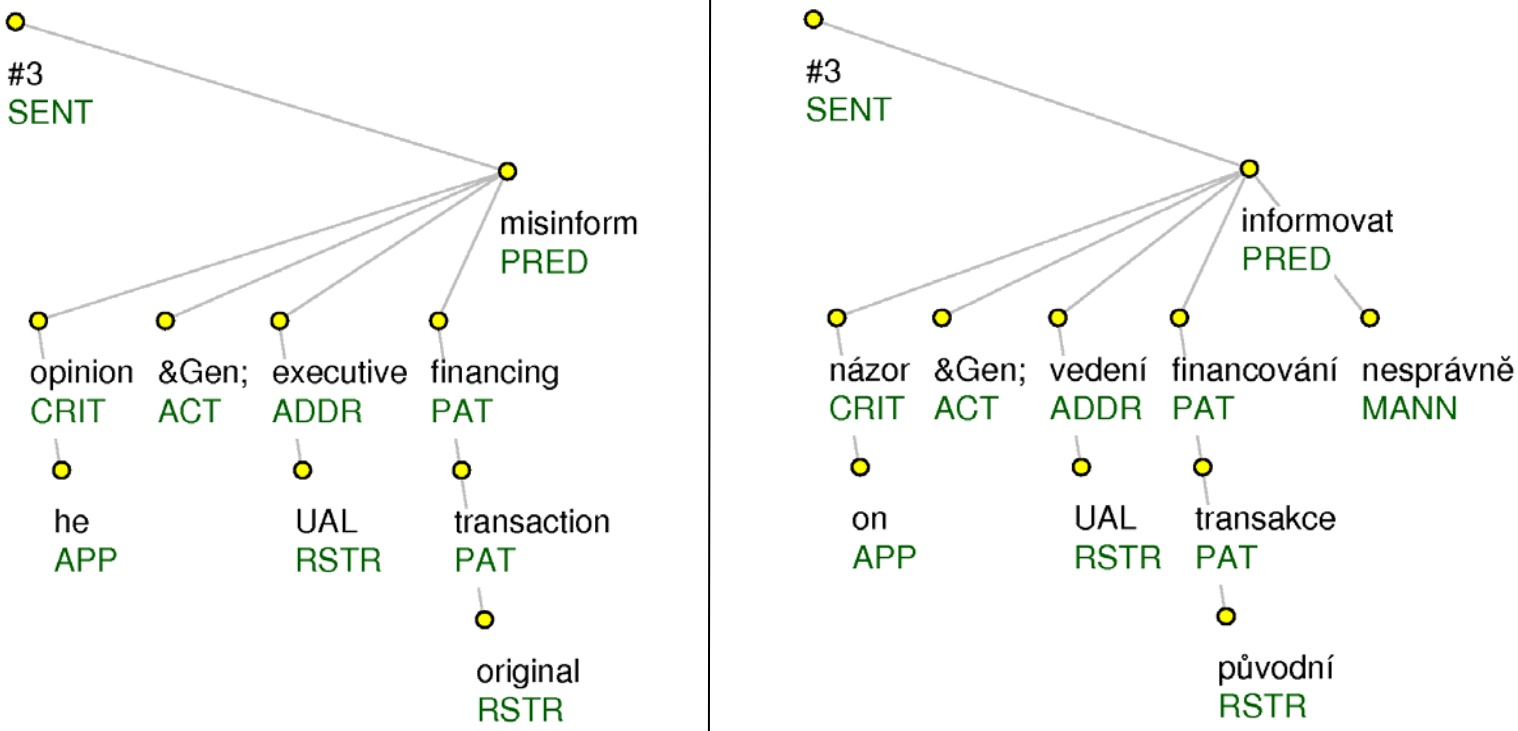
Analytical Pair En - Cz





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

Tectogrammatical Pair En - Cz



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.

Using Parallel Treebanks

- Word-based alignment
  - Phrasal elignment
- Dictionary extraction
  - From word/phrasal alignment
  - Probabilistic
- Machine translation
  - Statistical models
  - Evaluation/testing of systems

PCEDT 1.0 – The CD

- Published 2004 by the LDC (LDC2004T25)
- Texts, size of data:
  - 480,000 words: parallel annotated WSJ treebank
    - 21,600 sentences
  - 2 mil. words (53,000 sent.): Reader's Digest short stories
- 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)
- And more…

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
- Semecky, Cinkova:
  - Constructing an English Valency Lexicon
  - http://acl.ldc.upenn.edu/W/W06/W06-0612.pdf