Tree-based Machine Translation

using syntax and semantics

Jan Hajič

Charles University in Prague

Faculty of Mathematics and Physics

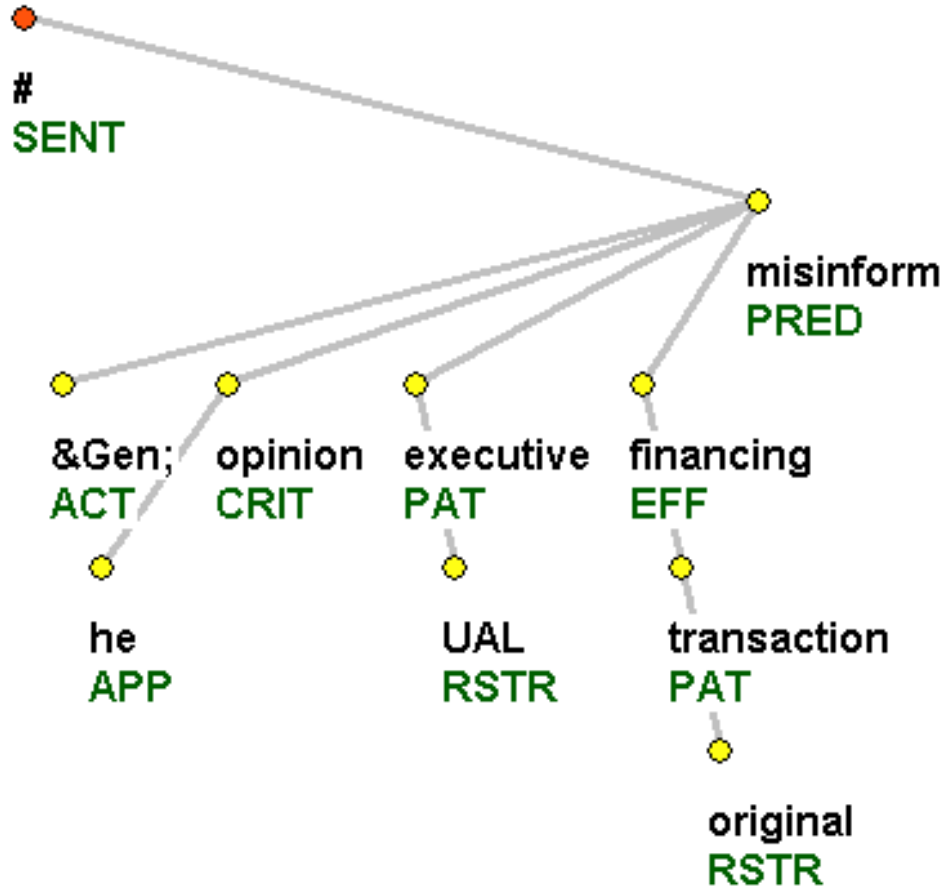
School of Computer Science

Institute of Formal and Applied Linguistics

Czech Republic

The Kinds of Trees We Grow

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



Meaning Representation

- Language-dependent:
  - Unit: lexical unit with lexical “meaning” (executive)
- Almost language-independent: PAT →
  - Dependency relations (executive misinform)
  - Semantic features (executive$_{PL}$, …)
    - Number, Tense, Modality, Mood, (In)definitness, ...
- Language-independent:
  - Dependency tree (as a formal object)
  - Information structure (topic,focus) (executive^{t}, misinform^{f})
  - Co-reference (anaphora resolution) (PERSON-NAME←he)

The Prague Dependency Treebank (PDT)

- Meaning ("tectogrammatical") representation
  - Layered approach
  - Language specific (...but specificity is "minimal")
  - Highest unit: sentence (utterance)
  - Syntax: dependency based
    - Combined syntactic and semantic representation
- Languages
  - Czech, English, Arabic, (German)
  - Slovak, Slovene, Greek, Latin, ... (other teams)

PDT annotation layers

- L0 (w) Words (tokens)
  - automatic segmentation and markup only
- L1 (m) Morphology
  - Tag (full morphology), lemma
- L2 (a) Analytical layer (surface syntax)
  - Dependency, analytical dependency function
- L3 (t) Tectogrammatical layer (“deep” syntax)
  - Dependency (labeled), sem. features, ellipsis resolution, co-reference, topic/focus, valency

The Annotation Layers

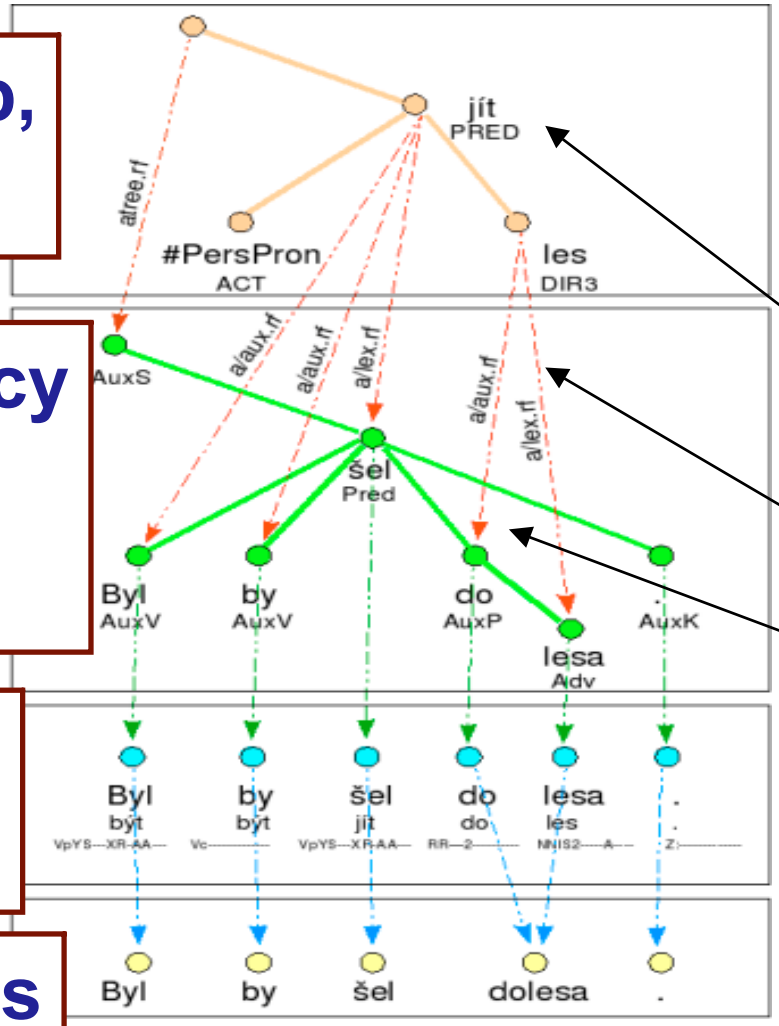
- Interlin
  top-down links
- API for cros
  access
  (programmi
- XML
- PML Schema /
  Relax

Meaning (deep, "rich" syntax)

Dependency Surface Syntax

Morphology, Lemmatization

Words



LFG analogy:

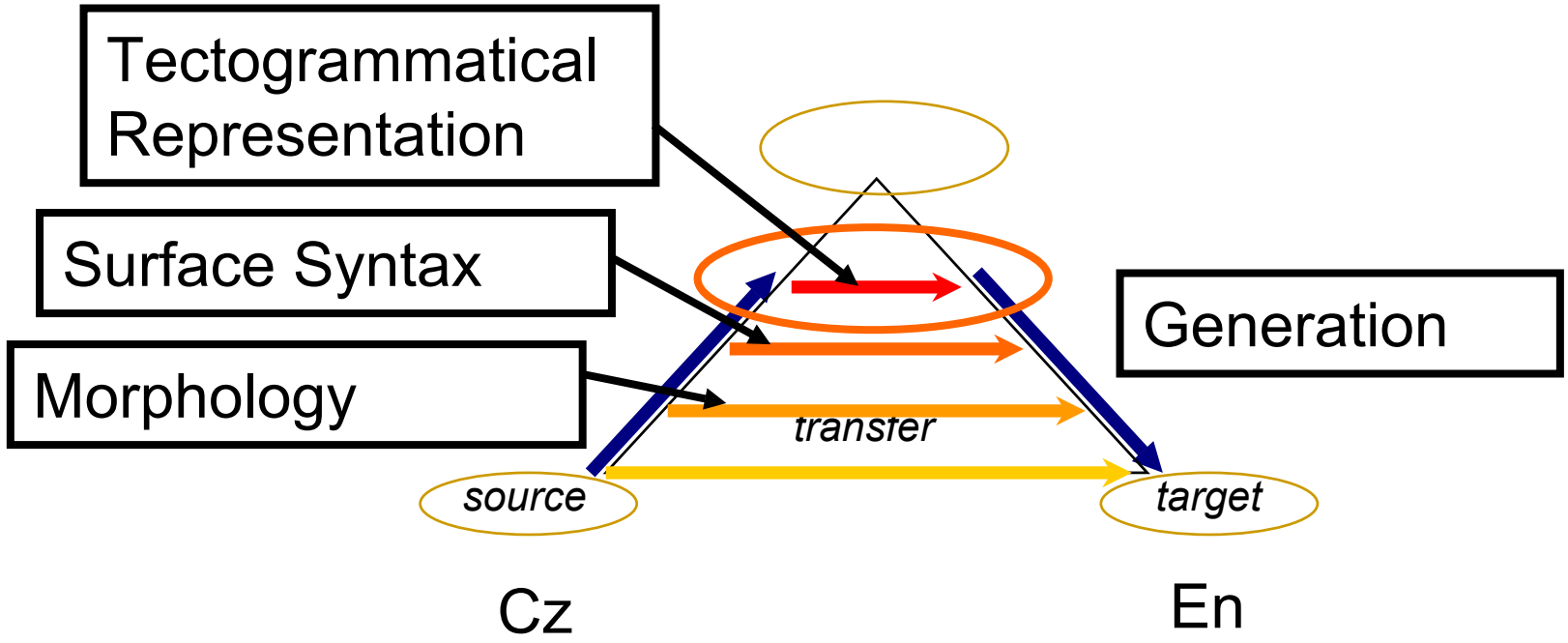
f-struct

ϕ

c-struct

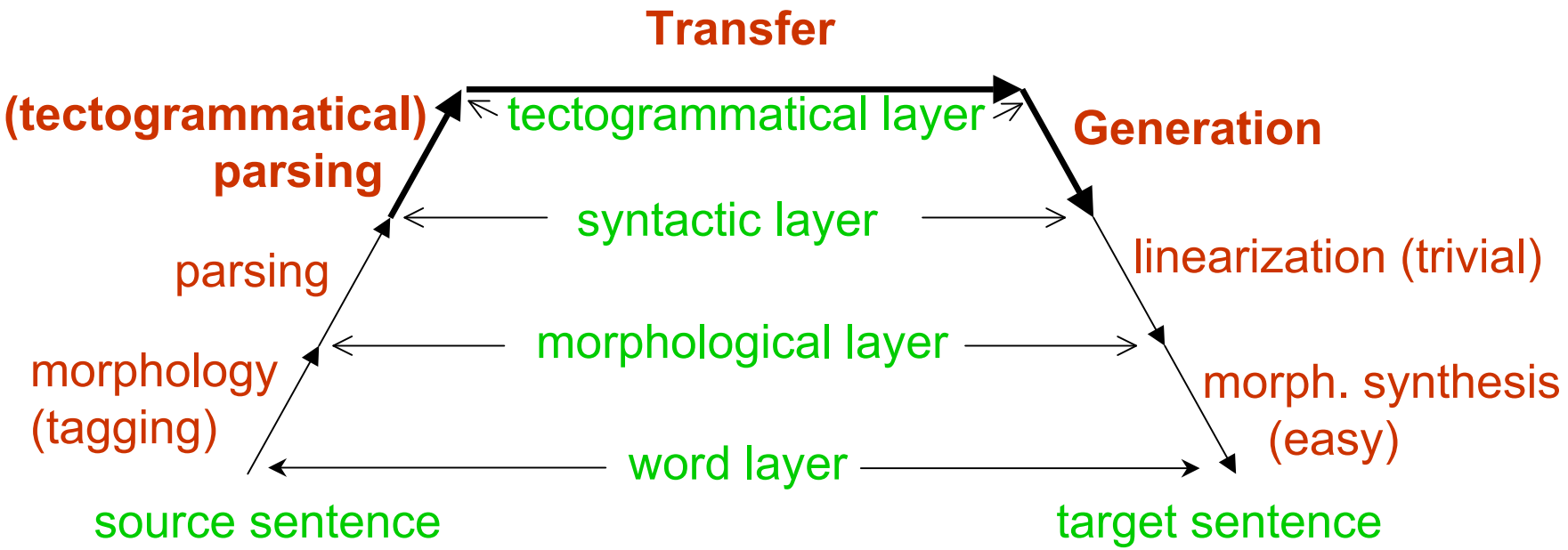
Machine Translation Scheme

- The Translation ("Vauquois") triangle



Tectogrammatical Layer in Machine Translation

- The additional three steps:

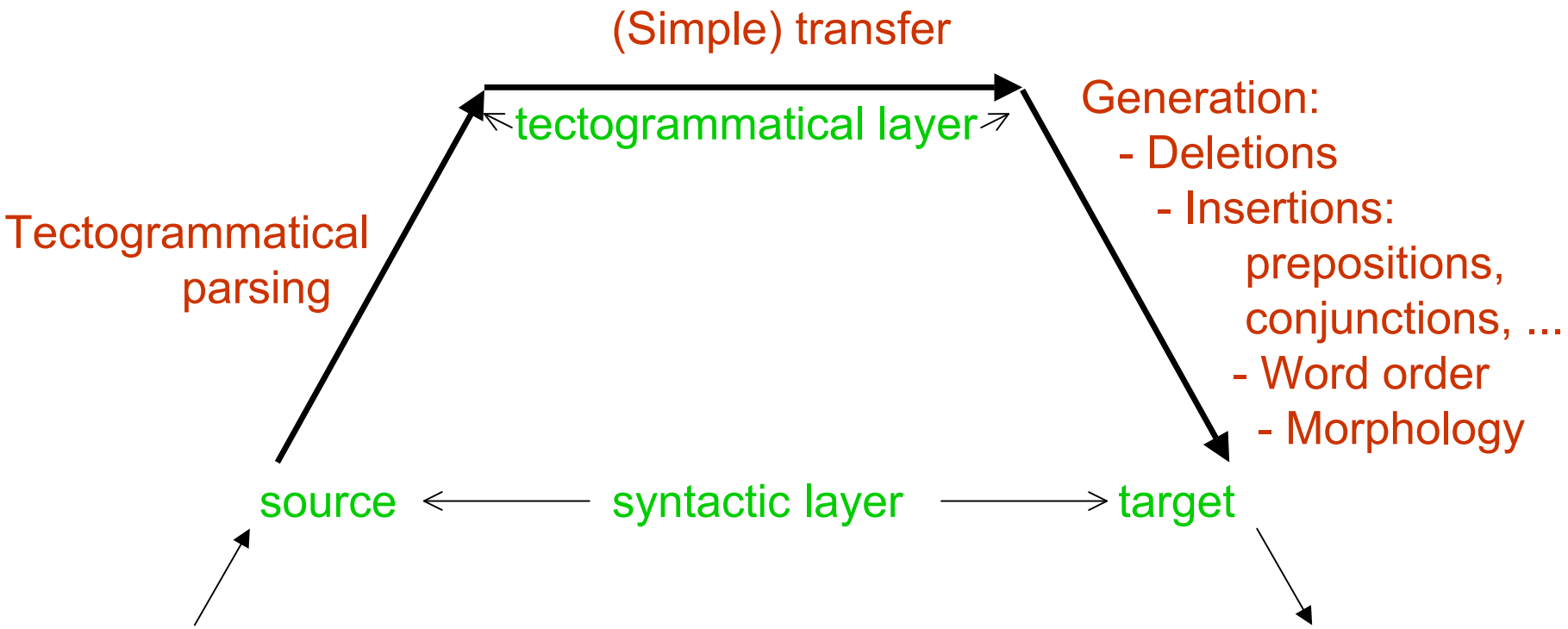


The Additional Steps

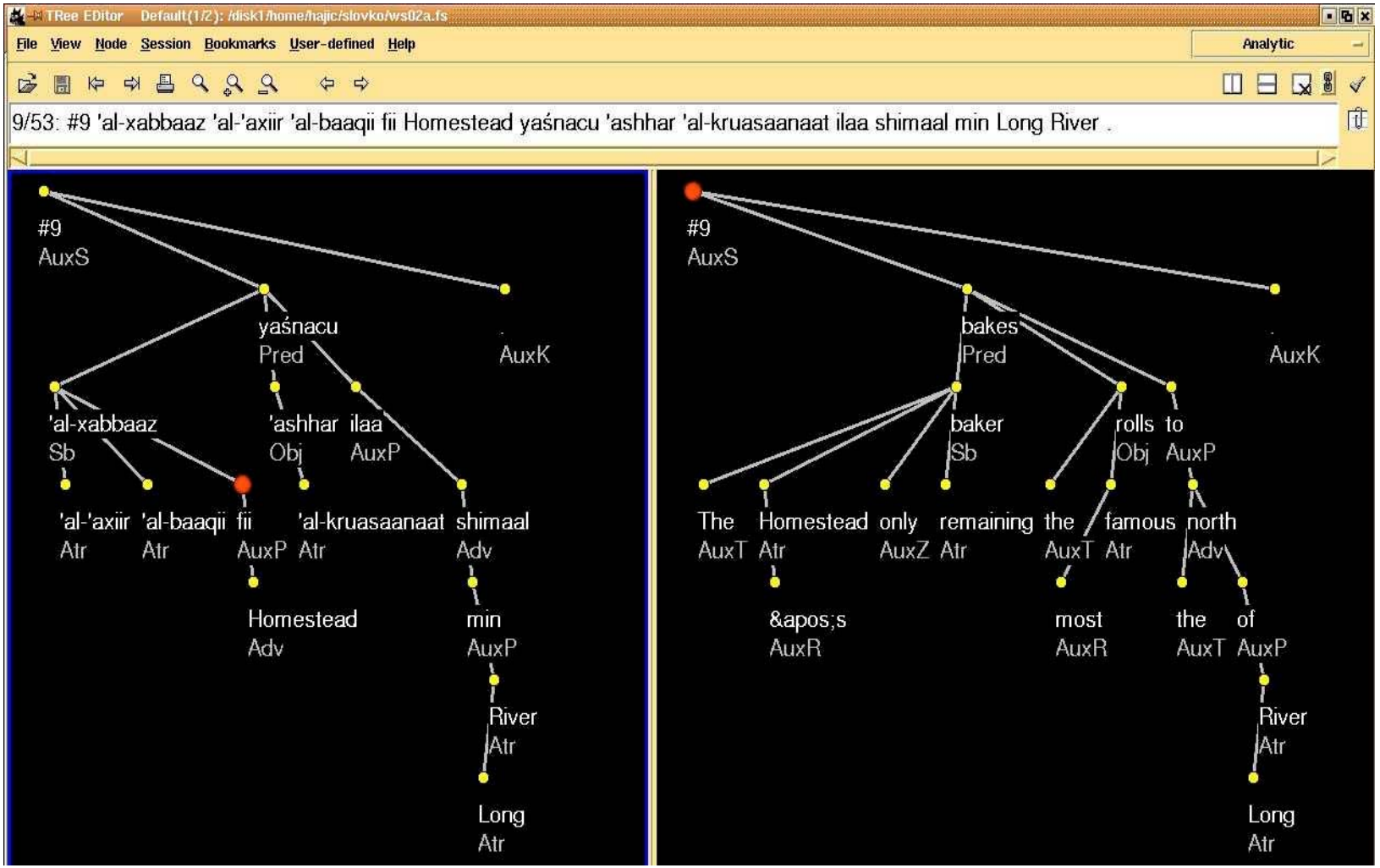
- Analytical (surface) → Tectogrammatical
  - additional parsing required
- Transfer
  - minimal effort: only “true”, non-1:1 transformations (*like swimming* ~ *schwimmen gern*)
- Generation
  - back from Tectogrammatical representation to Analytical (surface syntax)

Zooming In ...

- The additional three steps:



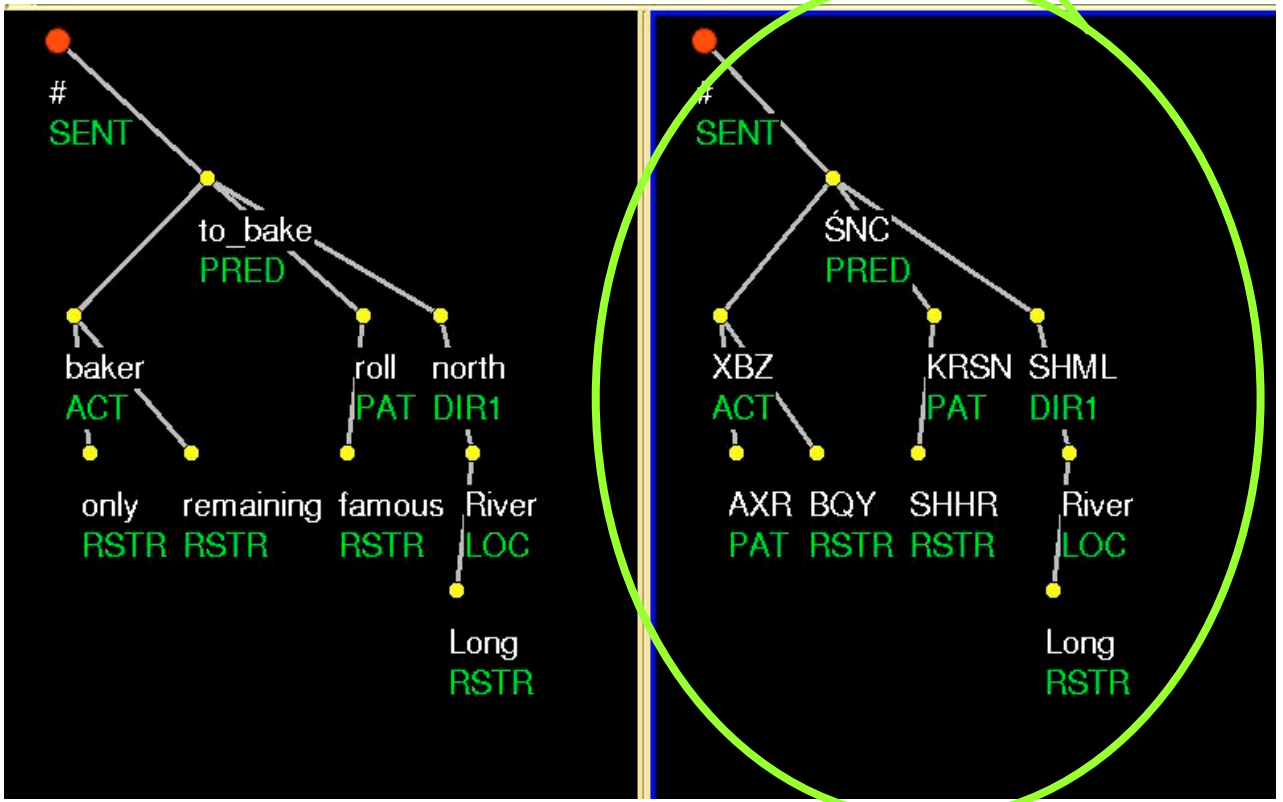
Analytical Layer Correspondence (Ar-En)



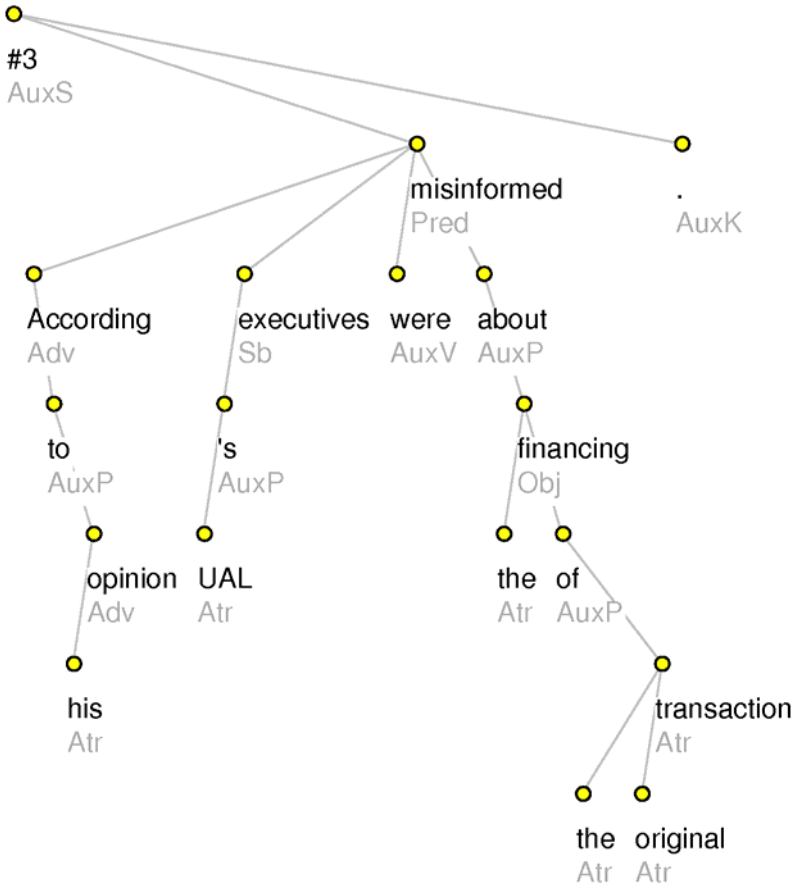
Tectogrammatical Correspondence (En-Ar)

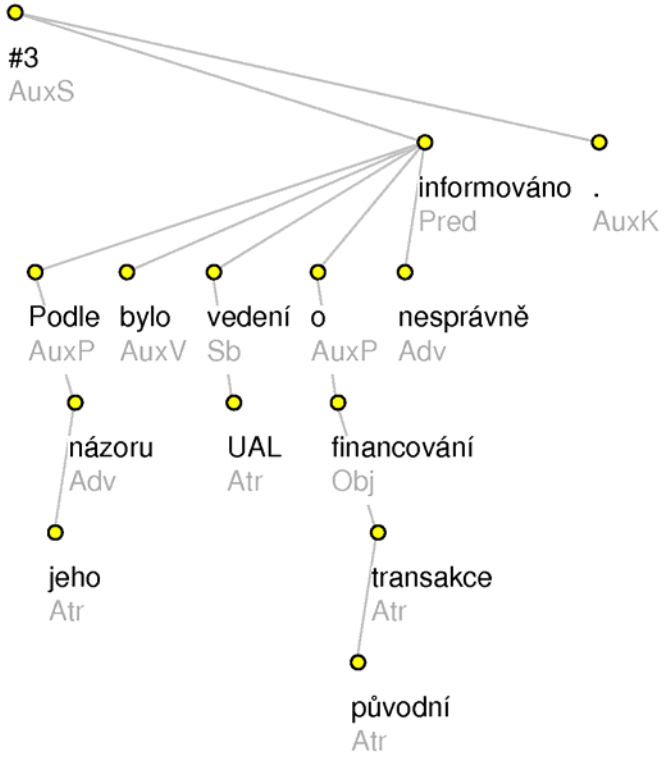
The [Homestead's] only remaining baker bakes the most famous rolls to the north of Long River.

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



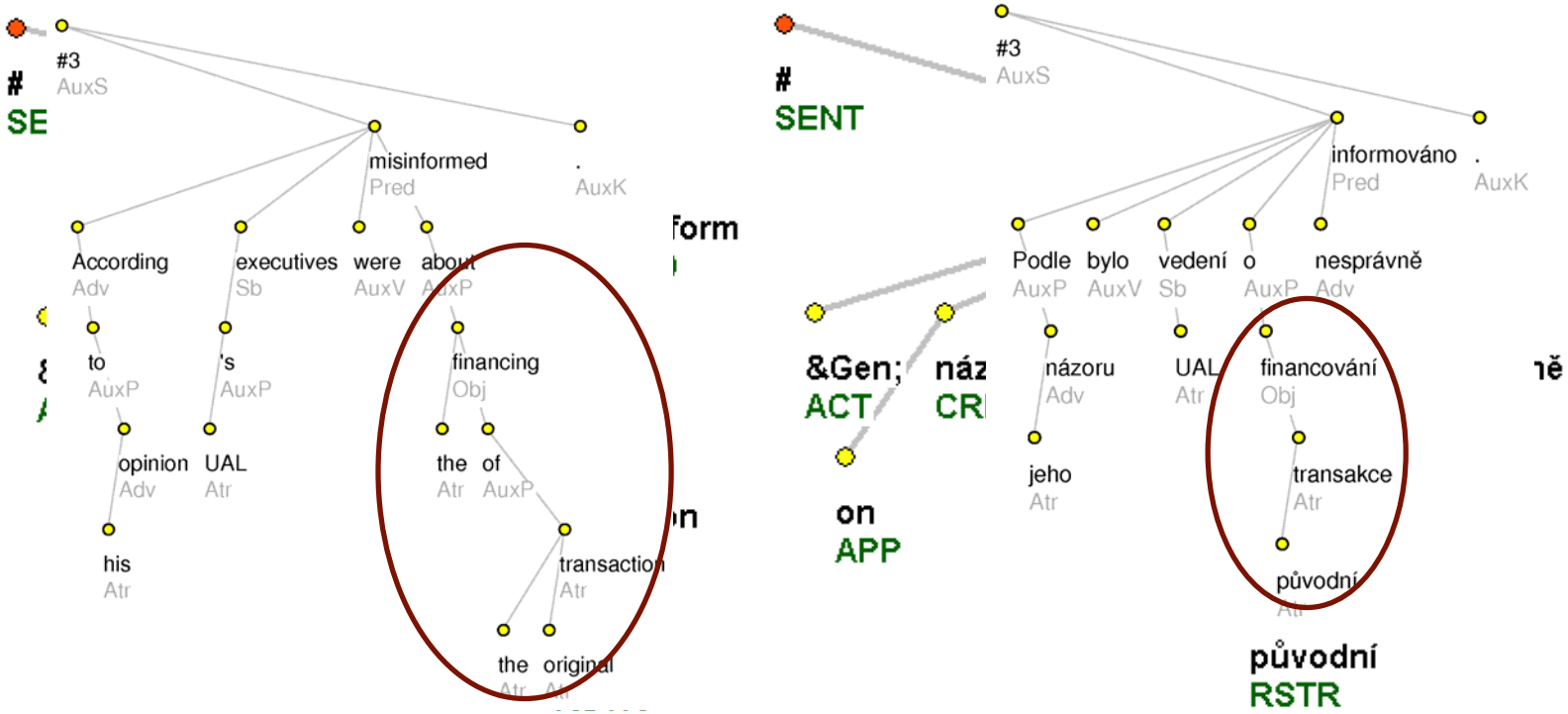
Depedency Syntax En-Cz





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

Meaning Level
En-Cz Correspondence



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

Parallel Czech-English Annotation: Penn Treebank

- English text -> Czech text (human translation)
- Czech side: all layers manual annotation
- English side:
  - Morphology and surface syntax: technical conversion
    - Penn Treebank style -> PDT surface dep. syntax layer
  - Tectogrammatical annotation: manual annotation
    - Auto pre-annotation
    - Many other resources merged in:
      - NP structure, BBN corpus (coreference, NE), Prop- &NomBank
- Alignment: natural, sentence level

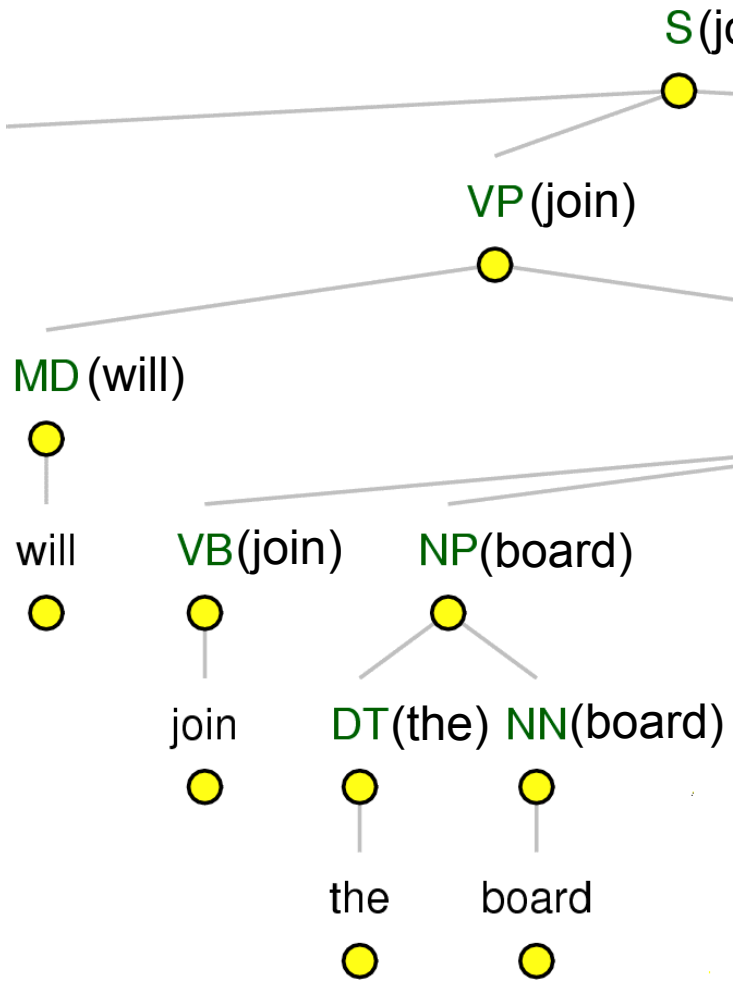
Human Translation of WSJ Texts

- 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 - “literal” translation preferred
- The numbers:
  - English tokens: 1173766
  - Documents (all of WSJ): 2312

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

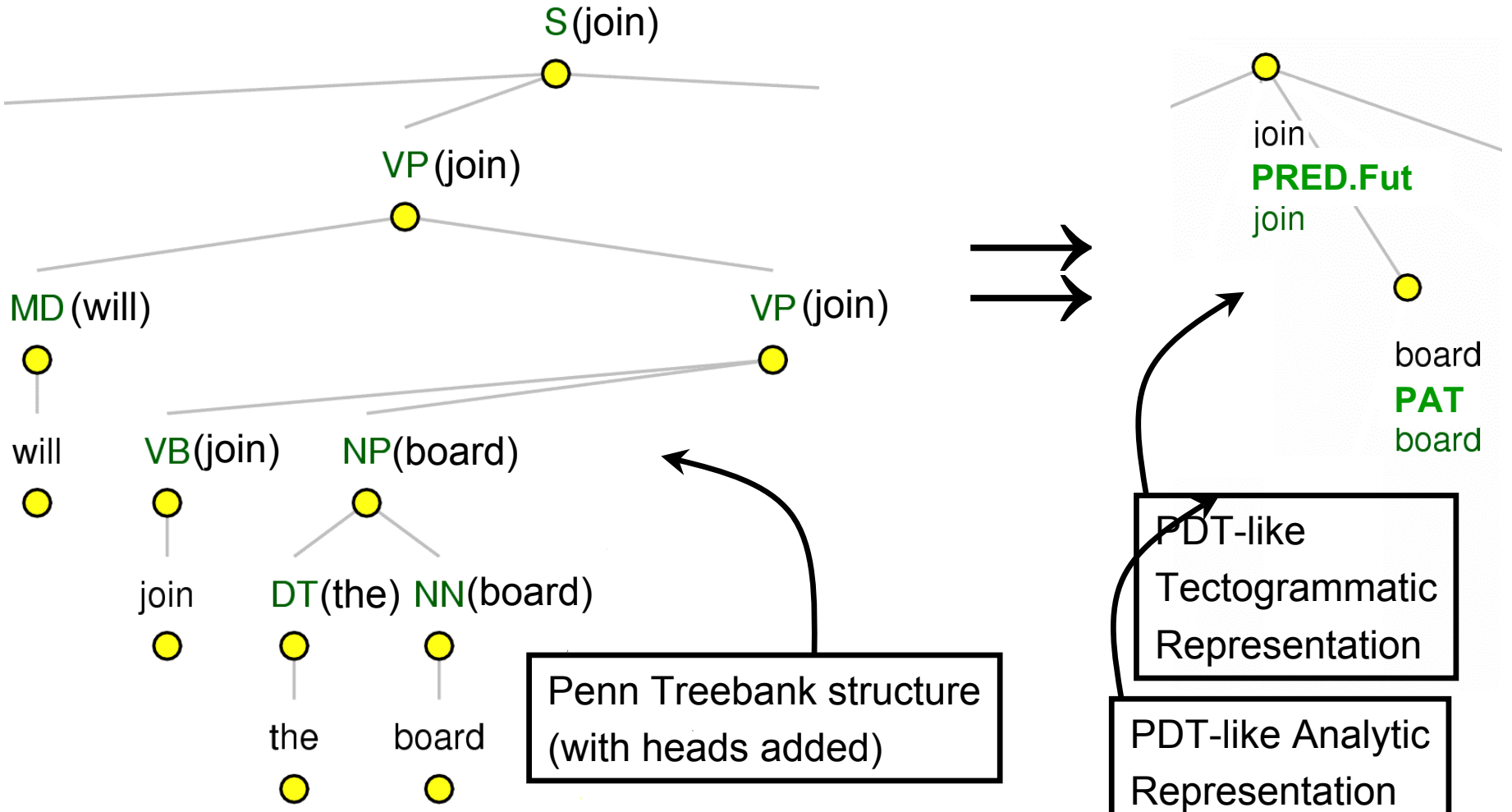
Conversion: Analytic Structure, Functions

- Syntactic 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)

Syntactic Structure, Functions: PTB to PDT



Czech PDT-style Annotation

- All layers
  - (morphology, analytic, tectogrammatical)
- So far…
  - Automatic (many tools by many authors)
- Manual annotation
  - In progress
  - Top-down
    - Tectogrammatical first (lower layers automatically)
    - … then syntactic structure and morphology

To summarize:

- PDT is/has (a)…
  - (Family of) dependency-based treebanking project(s)
    - Czech (English, Arabic, ...)
  - ~ 1mil. words
    - sufficient size for ML experiments
  - 4 interlinked layers of annotation
    - token, morphology, syntax, **deep syntax/semantics++**)
    - independent and “full” information at all levels
    - interlinked (for the development of parsers/generators)
  - Parallel corpus Cze <-> Eng -> *Machine Translation*

Some pointers

- Current version of PDT: v2.0, LDC2006T01
  - http://ufal.mff.cuni.cz/pdt2.0
- http://ufal.mff.cuni.cz
  - Research -> Corpora (Treebanks)
- http://www.ldc.upenn.edu
  - LDC2001T10 (PDT v1.0), LDC2004T23 (PADT 1.0), LDC2004T25 (PCEDT 1.0), LDC2006T01 (PDT 2.0)
- http://ufal.mff.cuni.cz/pedt
  - Penn Treebank in PDT style annotation (1/3)