The Prague Dependency Treebanks

Morphology, Syntax, Semantics

Jan Hajič

Institute of Formal and Applied Linguistics

School of Computer Science

Faculty of Mathematics and Physics

Charles University, Prague

Czech Republic

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

PDT 1.0 (2001)
PDT 2.0 (2006)
PDT 2.5 (2011)

- 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

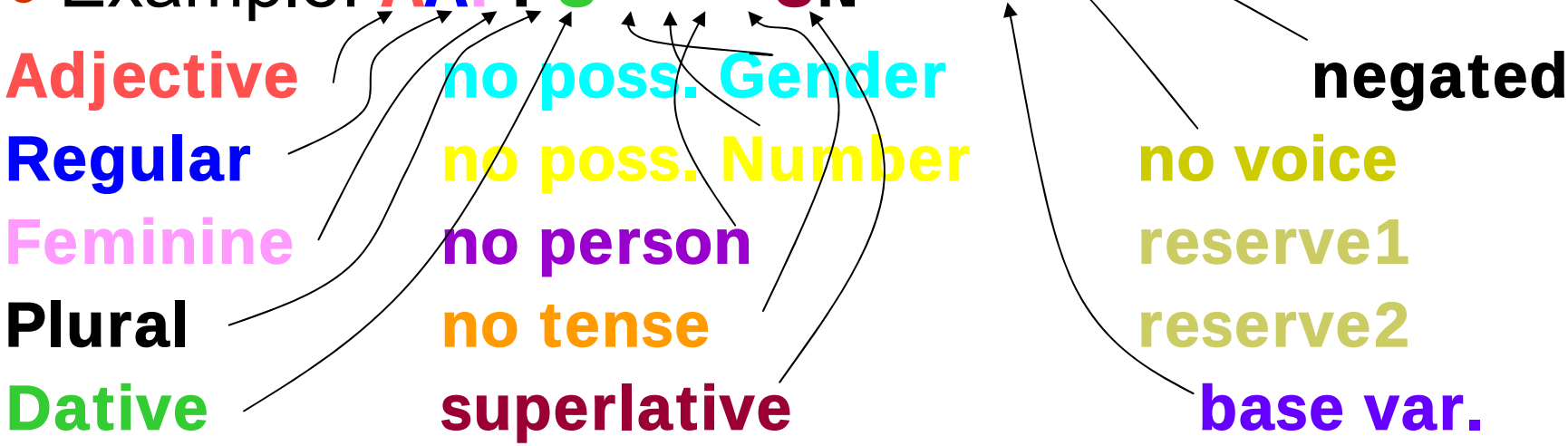
PDT Annotation Layers

- 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

Ex.: nejnezajímavějším
"(to) the most uninteresting"

- Tag: 13 categories
  - Example: AAFP3----3N----



- Lemma: POS-unique identifier

  Books/verb -> book-1, went -> go, to/prep. -> to-1

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)
    - EACL 2009 paper, http://ufal.mff.cuni.cz/compost

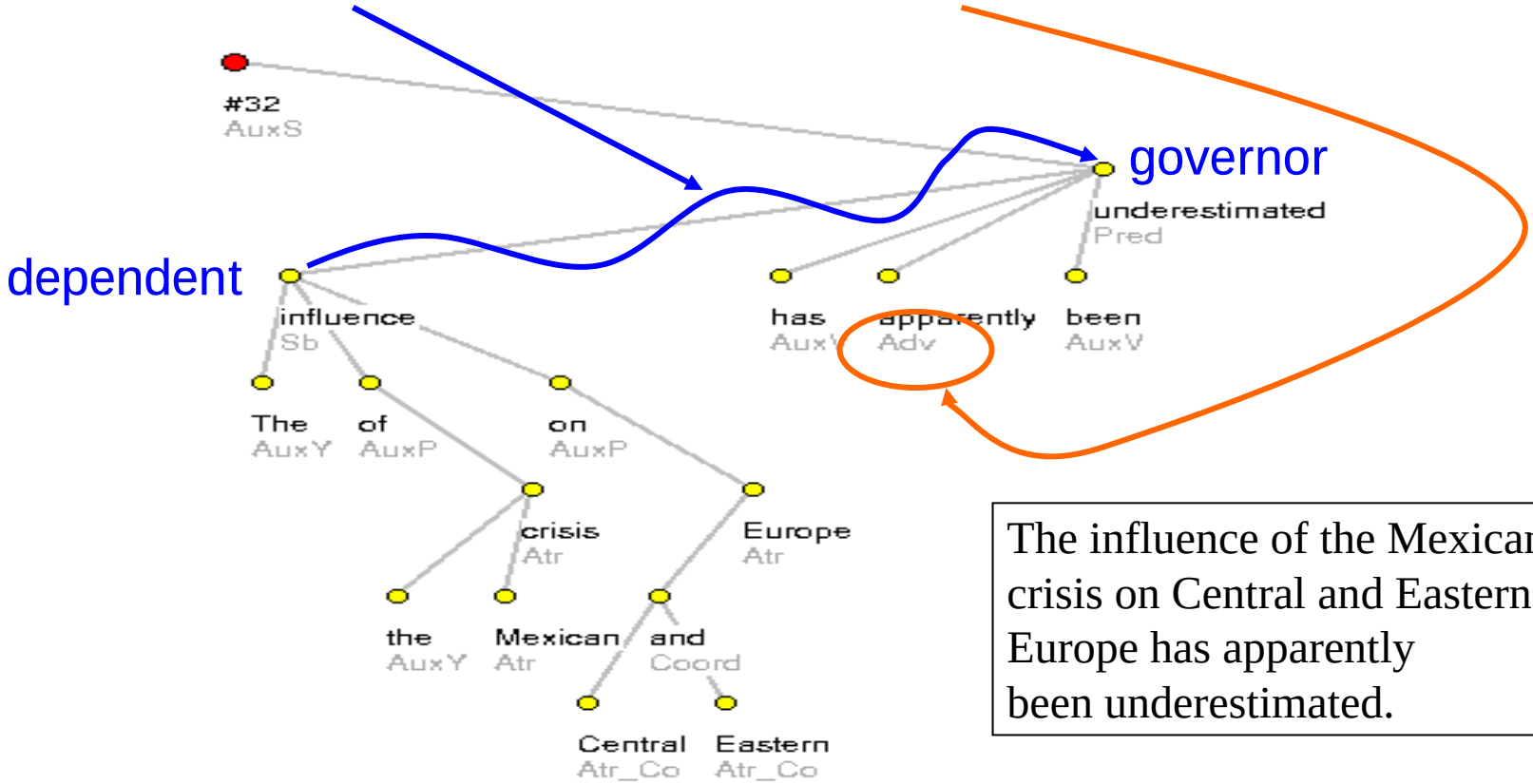
The Segmentation Problem: Arabic

- Tokenization / segmentation not always trivial

String ···· Token	Token Tag	Buckwalter Morph Tags	Token Form	Token Gloss
	F---------	FUT	sa-	will
سيخبرهم	VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I	yu-ḫbir-u	he-notify
	S----3MP4-	IVSUFF_DO:3MP	-hum	them
	P---------	PREP	bi-	about/by
بذلك	SD----MS--	DEM_PRON_MS	ḏālika	that
عن ·····	P---------	PREP	ʿan	by/about
طريق ·····	N-------2R	NOUN+CASE_DEF_GEN	ṭarīq-i	way-of
الرسائل ·····	N-------2D	DET+NOUN+CASE_DEF_GEN	ar-rasāʾil-i	the-messages
القصيرة ·····	A-----FS2D	DET+ADJ+NSUFF_FEM_SG+ +CASE_DEF_GEN	al-qaṣīr-at-i	the-short
	C---------	CONJ	wa-	and
والإنترنت	Z-------2D	DET+NOUN_PROP+ +CASE_DEF_GEN	al-ʾinternet-i	the-internet
	C---------	CONJ	wa-	and
وغيرها	FN------2R	NEG_PART+CASE_DEF_GEN	ġayr-i	other/not-of
	S----3FS2-	POSS_PRON_3FS	-hā	them

Layer 2 (a-layer): Analytical Syntax

- Dependency + Analytical Function



The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated.

Analytical Syntax: Functions

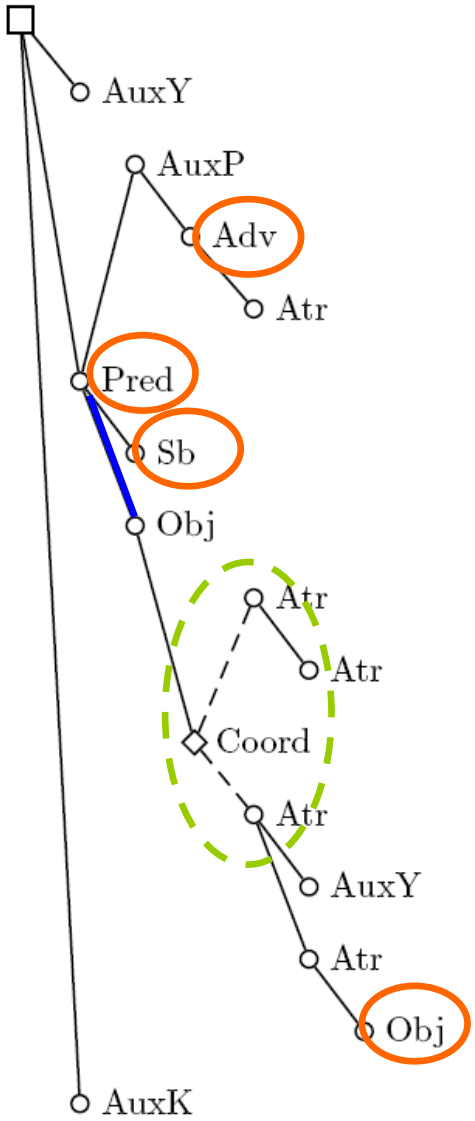
- Main (for [main] semantic lexemes):
  - Pred, Sb, Obj, Adv, Atr, Atv(V), AuxV, Pnom
  - “Double” dependency: AtrAdv, AtrObj, AtrAtr
- Special (function words, punctuation,...):
  - Reflefives, particles: AuxT, AuxR, AuxO, AuxZ, AuxY
  - Prepositions/Conjunctions: AuxP, AuxC
  - Punctuation, Graphics: 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 → treated as ellispsis; Arabic: systematic solution), Pred
  - (Added) analytic functions:
    - AuxM لَمْ *lam* (did-not)
    - Ante مَا *mā* (what)
- Work by Faculty of Arts (Arabic language) students

Arabic Surface Syntax Example

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



وَ *wa-*	and	C---------
فِي *fī*	in	P---------
مِلَفِّ *milaffi*	collection/file-of	N-------2R
اَلأَدَبِ *al-ʾadabi*	the-literature	N-------2D
طَرَحَت *ṭaraḥat*	it-presented	VP-A-3FS--
اَلمَجَلَّةُ *al-mağallatu*	the-magazine	N-----FS1D
قَضِيَّةَ *qaḍīyata*	issue-of	N-----FS4R
اَللُّغَةِ *al-luġati*	the-language	N-----FS2D
اَلعَرَبِيَّةِ *al-ʿarabīyati*	the-Arabic	A-----FS2D
وَ *wa-*	and	C---------
اَلأَخطَارِ *al-ʾaḫṭāri*	the-dangers	N-------2D
اَلَّتِي *allatī*	that	SR----FS--
تُهَدِّدُ *tuhaddidu*	they-threaten	VIIA-3FS--
هَا *-hā*	it	S----3FS4-
. .	.	G---------

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

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)

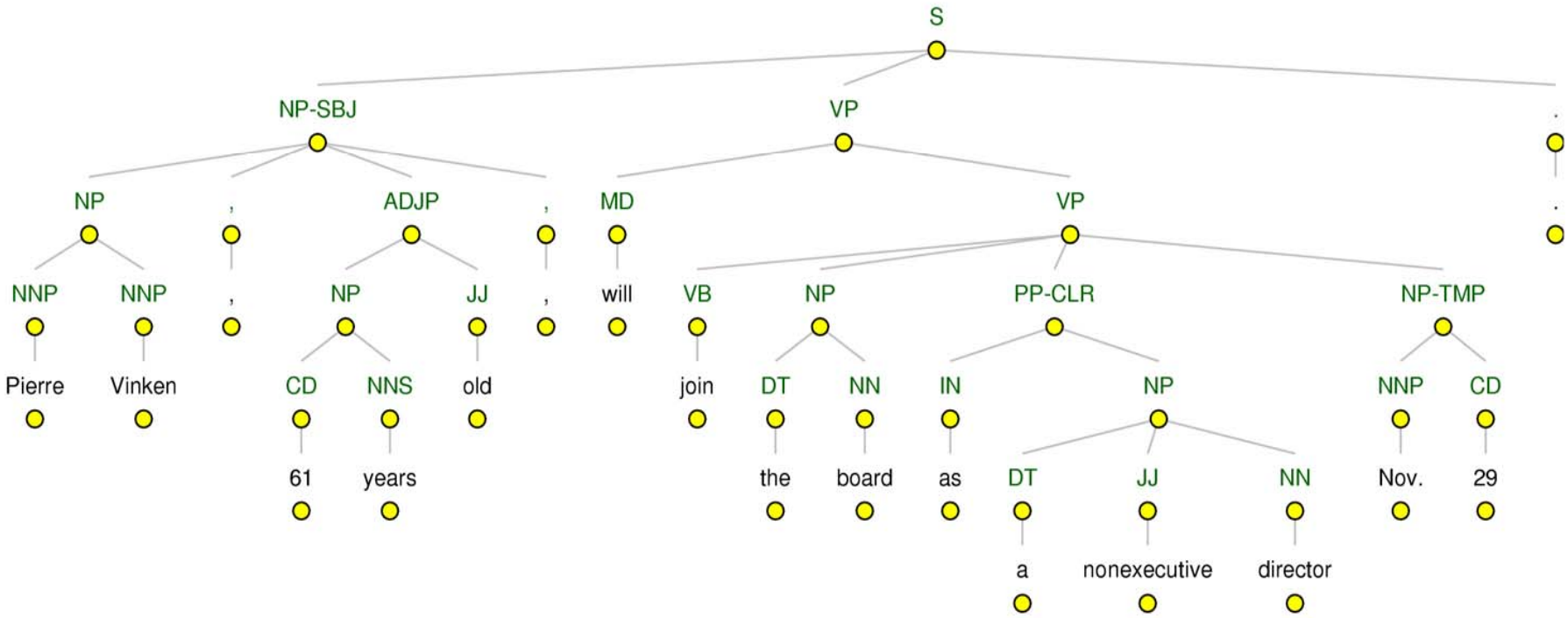
Noun Phrase

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:



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)

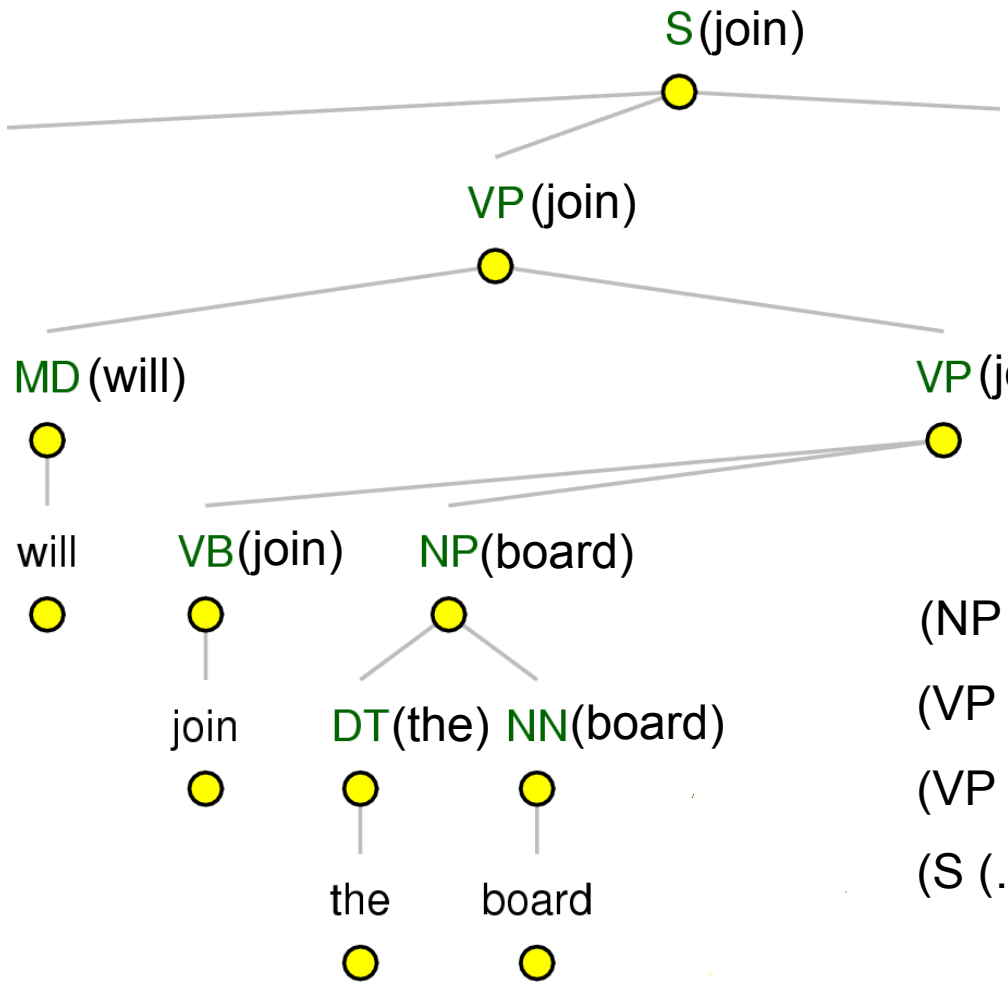
English Annotation POS and Syntax

- 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:

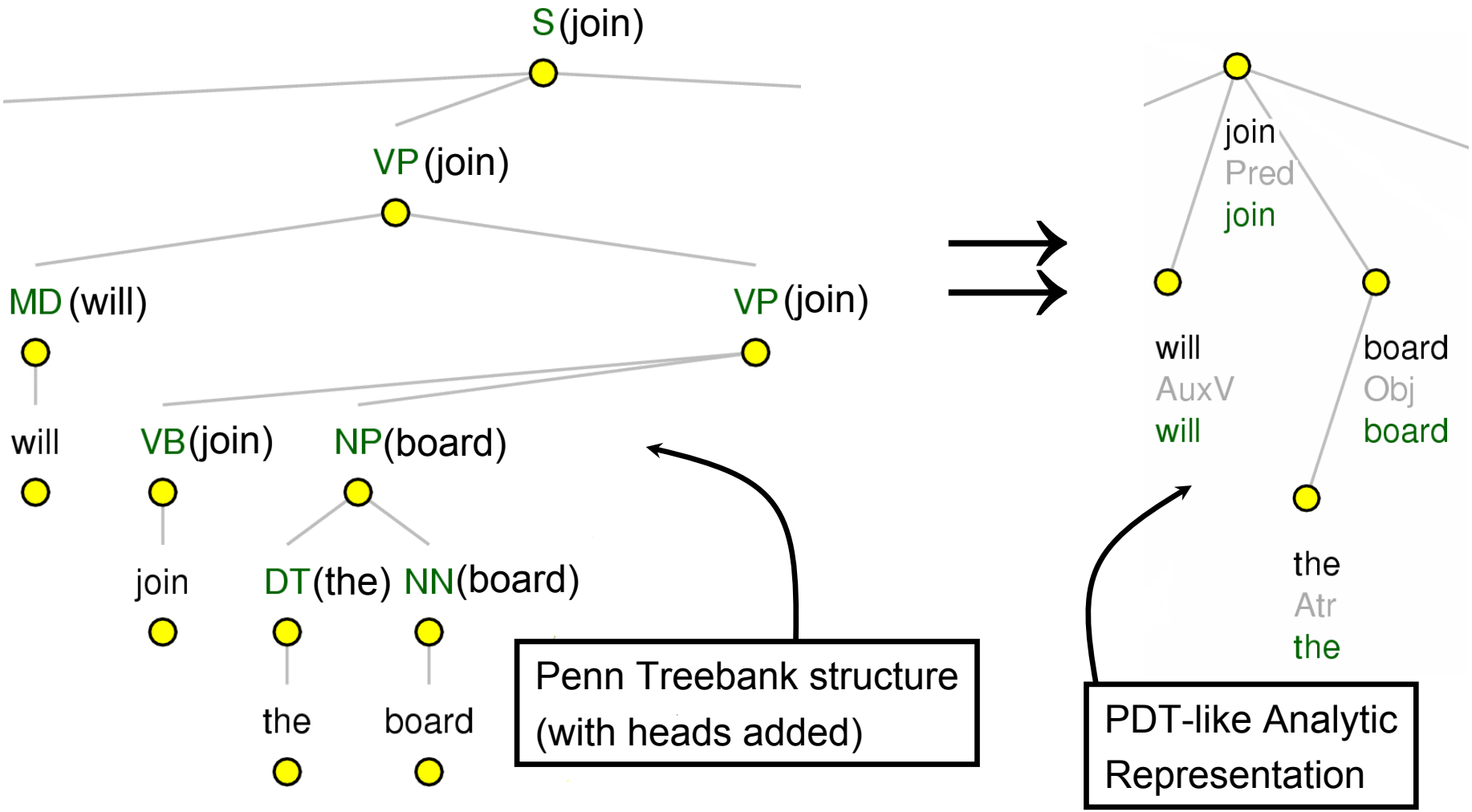
(NP (DT NN)) → NN

(VP (VB NP)) → VB

(VP (MD VP)) → VP

(S (… VP …)) → VP

Example: Analytical Structure, Functions



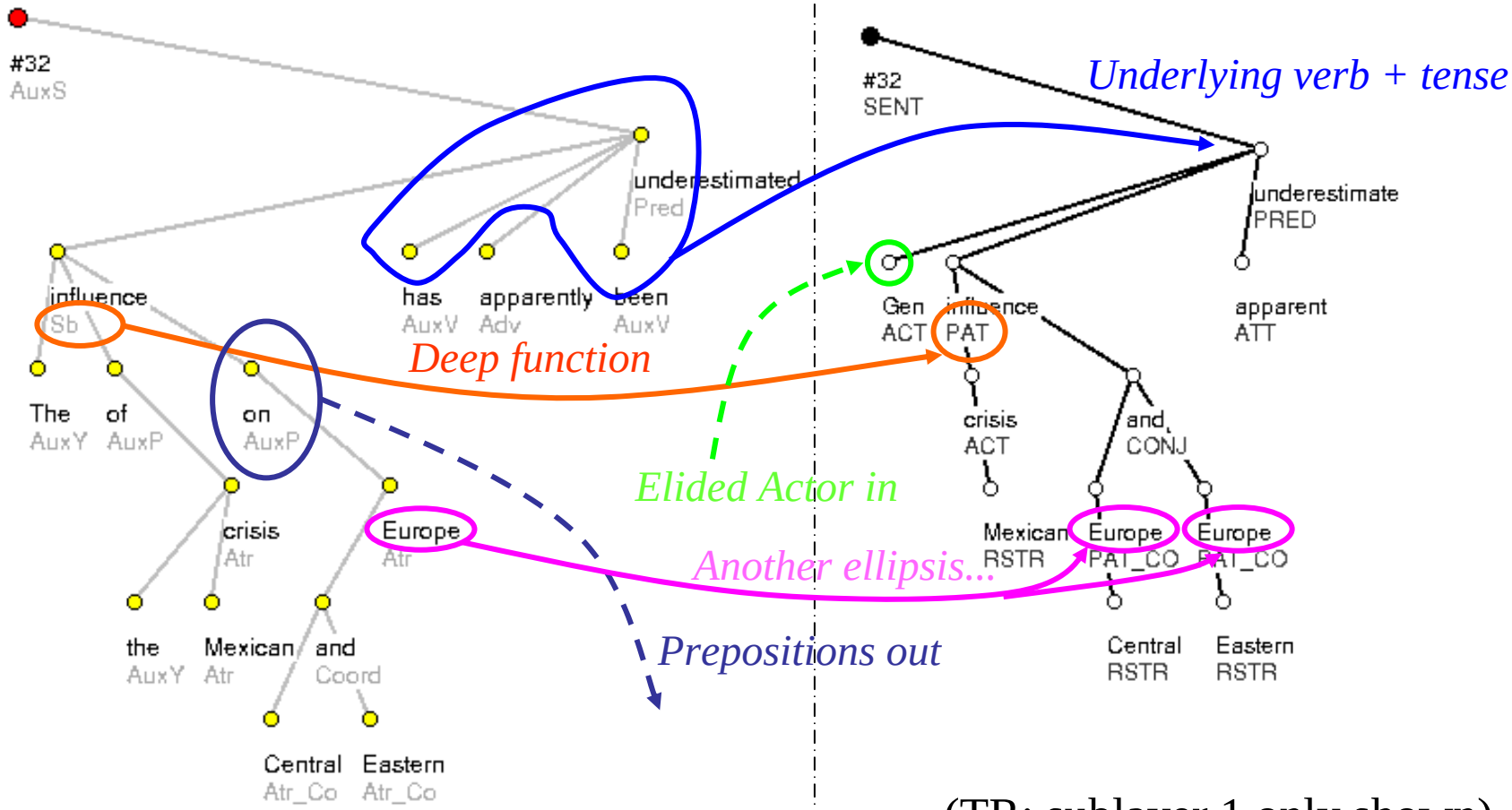
PDT Annotation Layers

- 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



(TR: sublayer 1 only shown)

Layer 3: Tectogrammatical

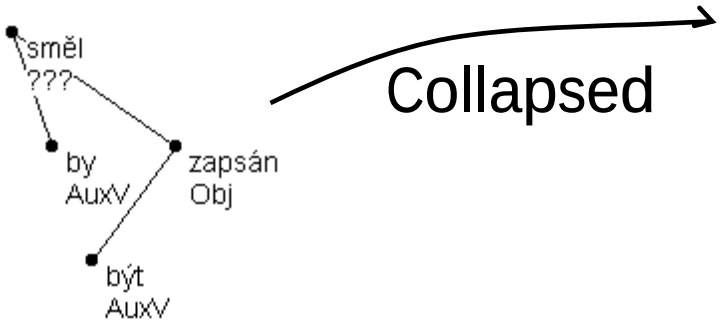
- Underlying (deep) syntax
- 4 sublayers:
  - dependency structure, (detailed) functors
  - topic/focus and deep word order
  - coreference
  - all the rest (grammatemes):
    - detailed functors
    - underlying gender, number, ...

Tectogrammatical Functors

- "Actants": syntactic ACT, PAT, semantic 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

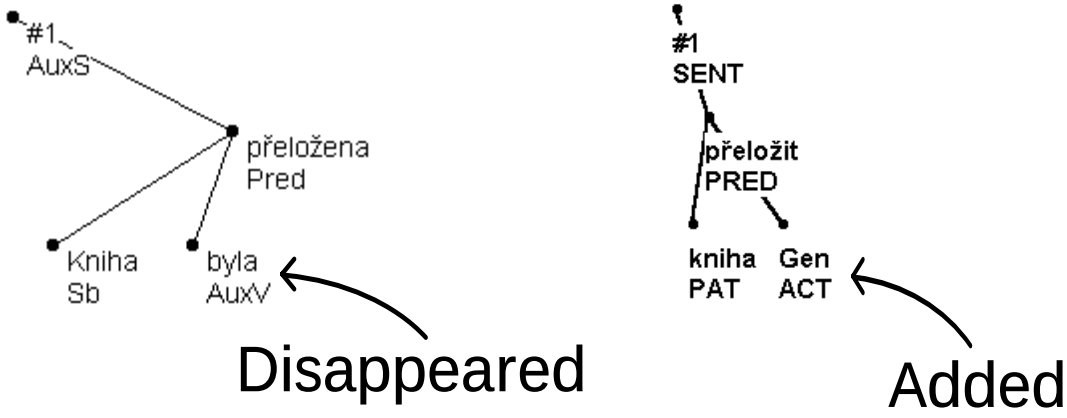
- Analytical verb form:
  - (he) allowed would-be to-be enrolled
  - směl by být zapsán



Additional attributes (grammatemes): conditional + “allow”

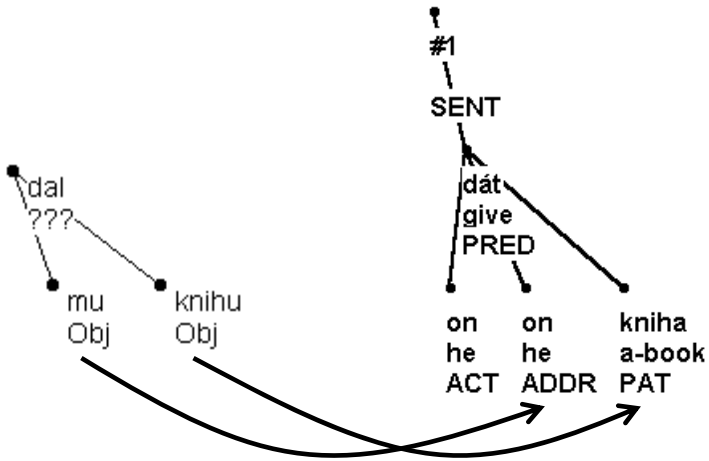
Tectogrammatical Example

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



Tectogrammatical Example

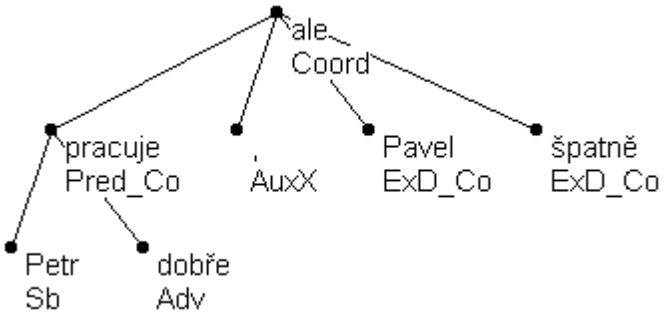
- Object
  - (he) gave him a-book
  - dal mu knihu

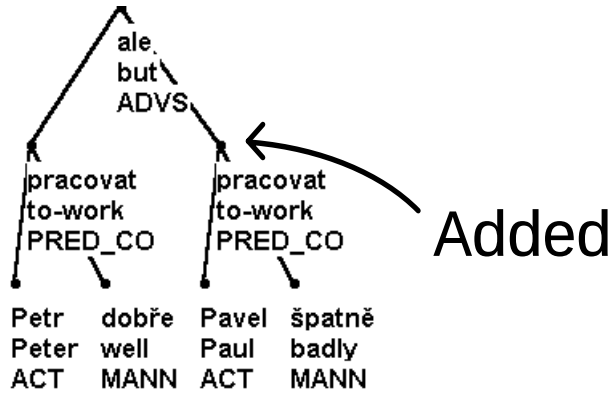


Obj goes into ACT, PAT, ADDR, EFF or ORIG based on governor's valency frame

Tectogrammatical Example

- Incomplete phrases
  - Peter works well , but Paul badly
  - Petr pracuje dobře, ale Pavel špatně





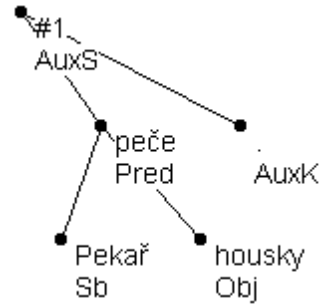
Layer 3: Tectogrammatical

- 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, ...

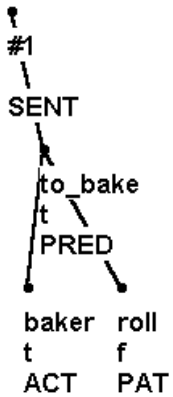
Deep Word Order Topic/Focus

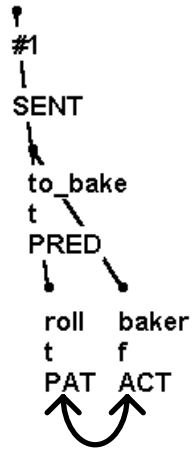
- Example:

Analytical dep. tree:



- Baker bakes rolls. vs. *Baker*^{IC} bakes rolls.



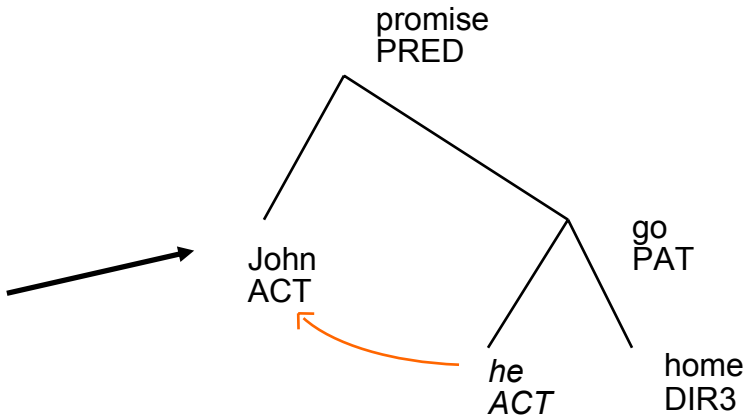


Layer 3: Tectogrammatical

- 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, ...

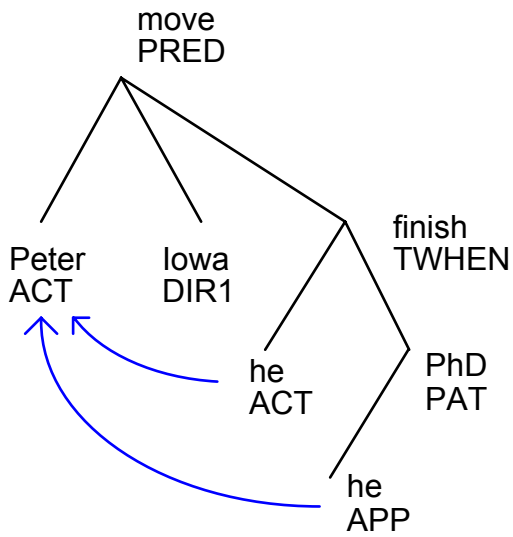
Coreference

- 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 ...



Coreference

- Textual
  - Ex.: Peter moved to Iowa after he finished his PhD.



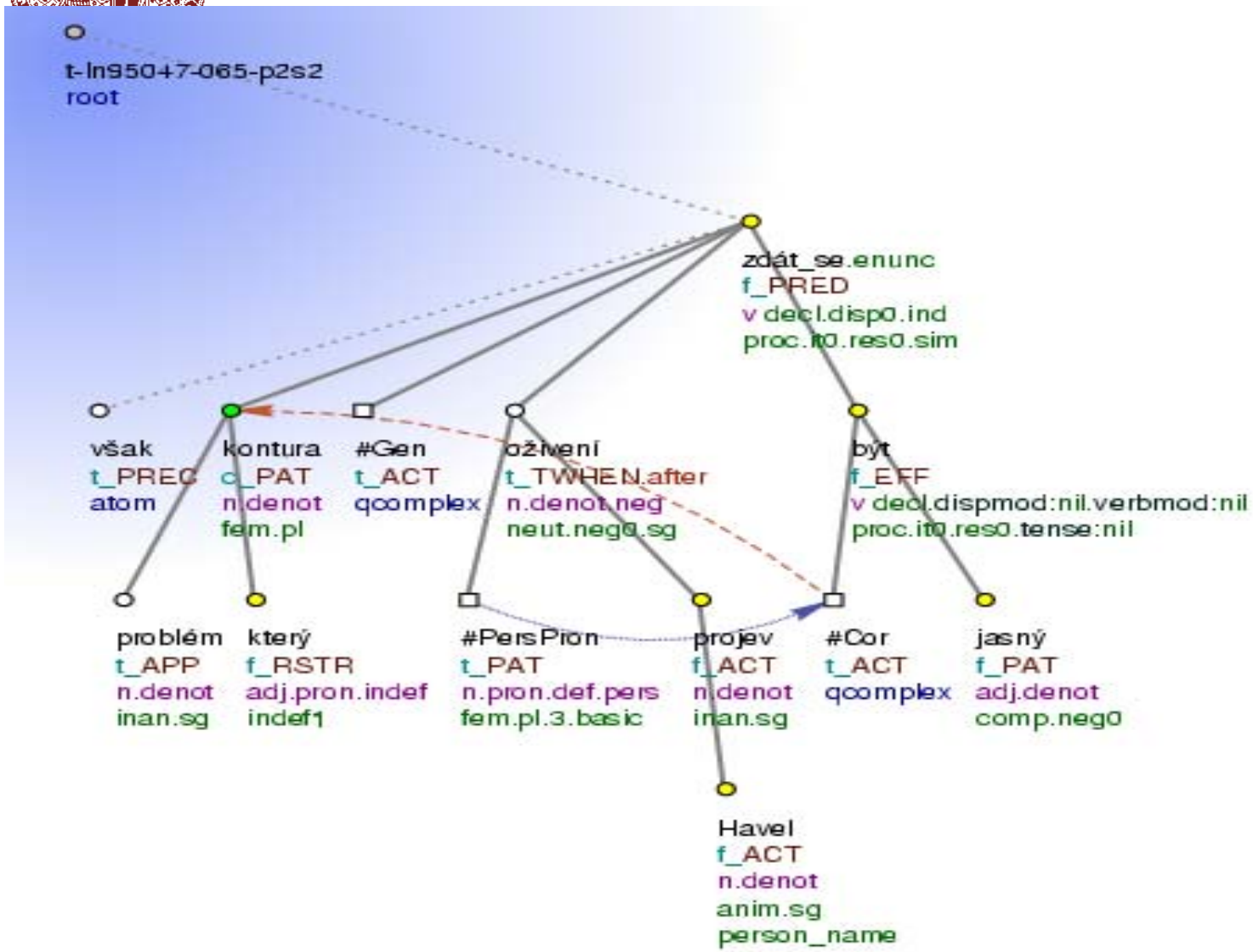
Layer 3: Tectogrammatical

- 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!)

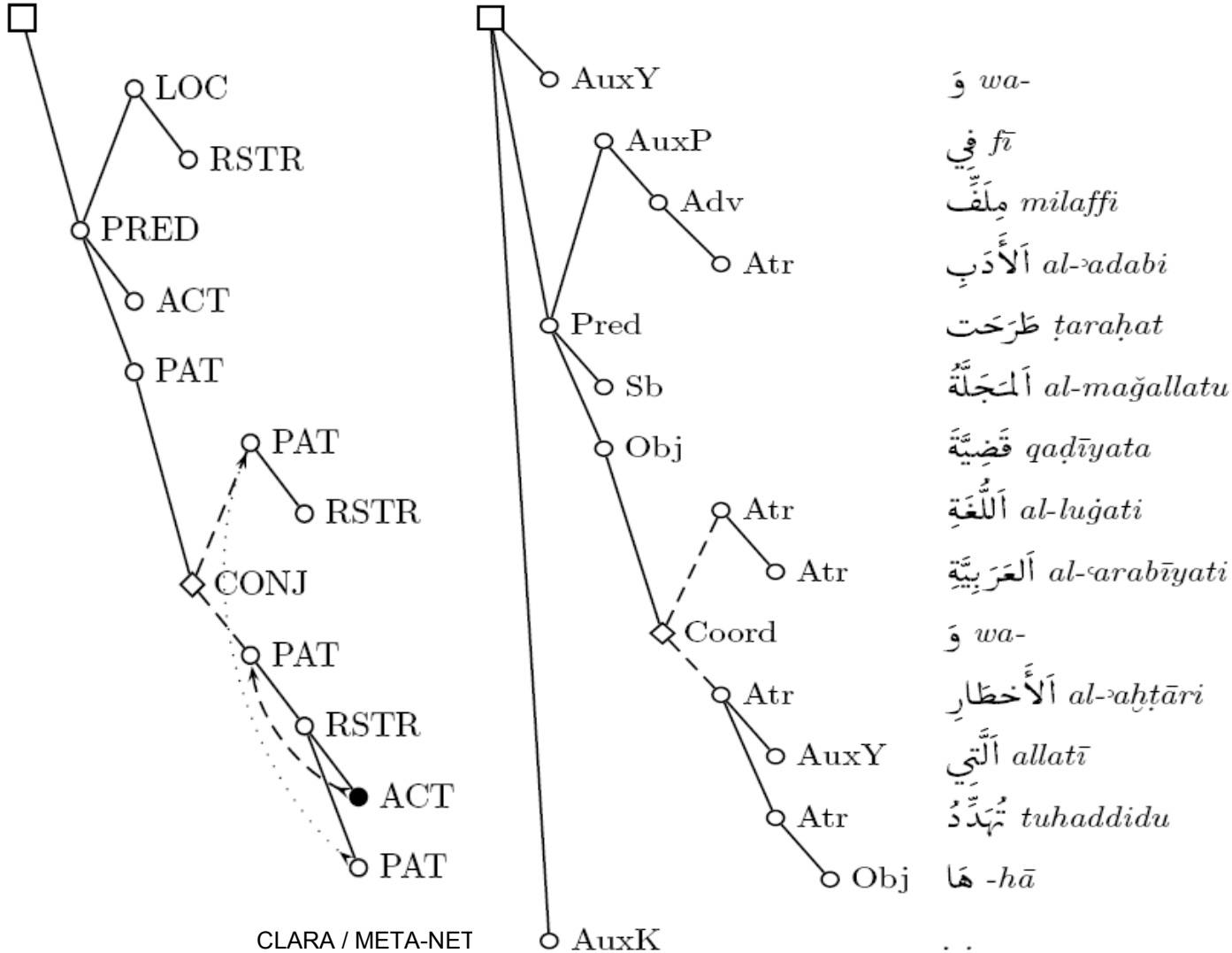
Fully Annotated Sentence



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.

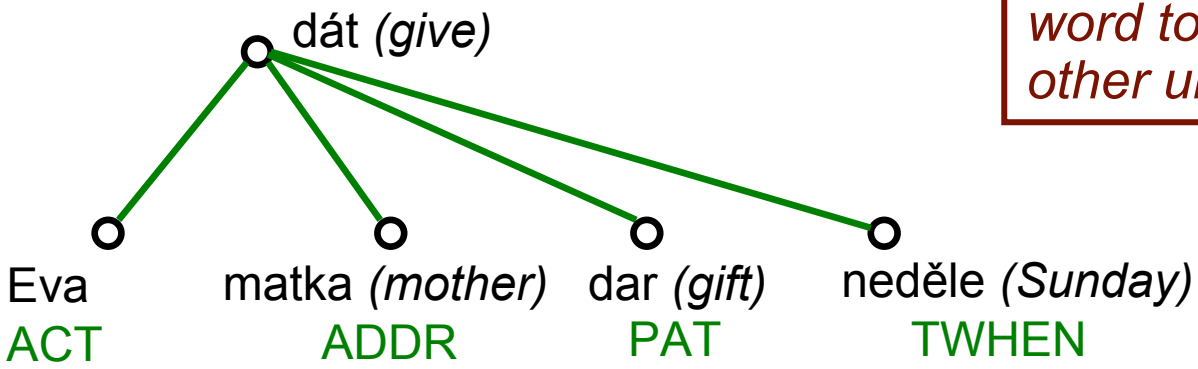


English PDT-style Annotation

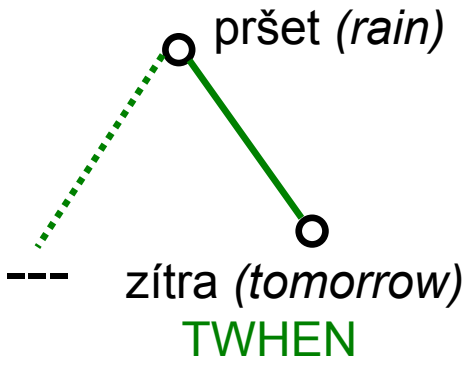
- Morphology and Syntax
  - By conversion
- Tectogrammatical annotation
  - Manual (English TR: by S. Cinková)
  - Pre-annotation
    - Transformation from Penn Treebank & Propbank (Palmer, Kingsbury) by Z. Žabokrtský et al.
  - Valency
    - From Propbank Frame Files (Cinková, Šindlerová, Nedolužko,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

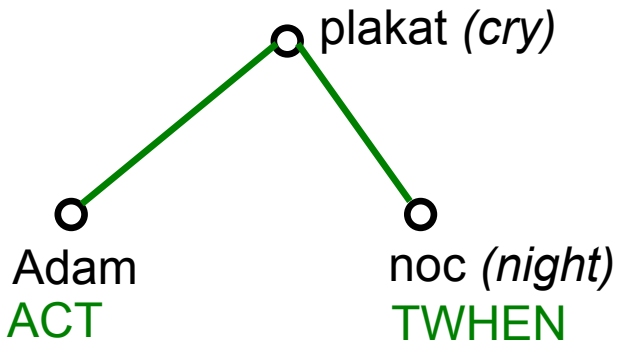
Valency in the PDT

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



Specifies actor





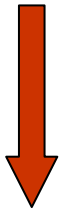
Valency - Basic Principles

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:

	obligatory	optional
argument		
adjunct		

Contents:

- functor
- obligatoriness
- surface form

one meaning of the word ➔ one valency frame

word: *leave*
meaning 1: *sb left sth* → frame1: ACT PAT
meaning 2: *sb left from somewhere* → frame2: ACT DIR1

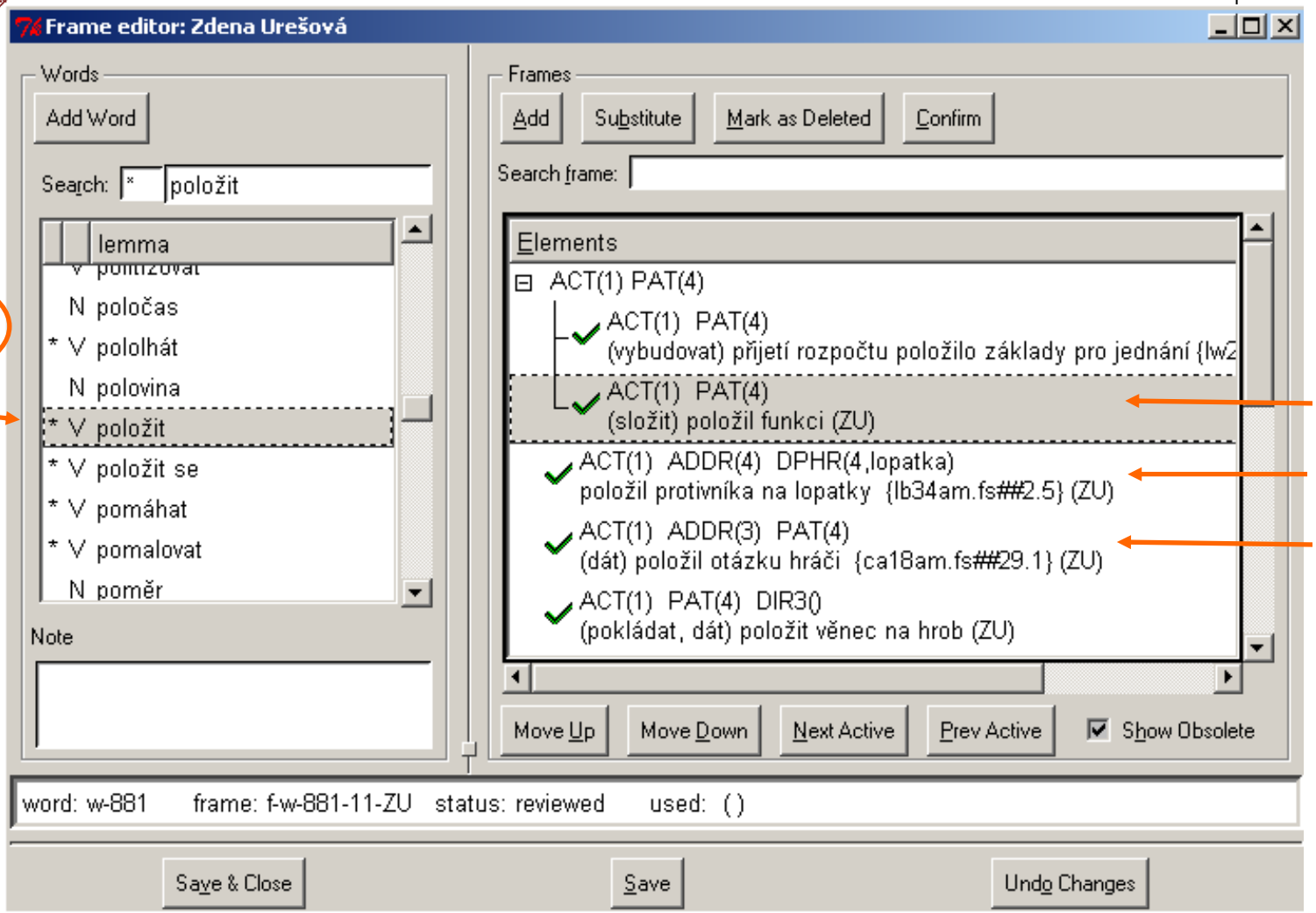
Valency lexicon: PDT-VALLEX

- 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:** 272×
dosáhnout určité úrovně
mzda d. v tomto oboru 80 tisíc
d. pokročilého věku
ACT(.1) PAT(.2,*aby*[.v]) ?ORIG(*na-1*[.6],*od-1*[.2]) v-w714f2 **Used:** 7×
dosáhl na něm slibu
dosáhli na sobě slibu
ACT(.1) DPHR(*svůj-1*.2) v-w714f3 **Used:** 2×
dosáhl svého
ACT(.1) DIR3(*) v-w714f4 **Used:** 2×
dosáhl na strop
rukou.MEANS

The PDT-VALLEX editor



'lay down'

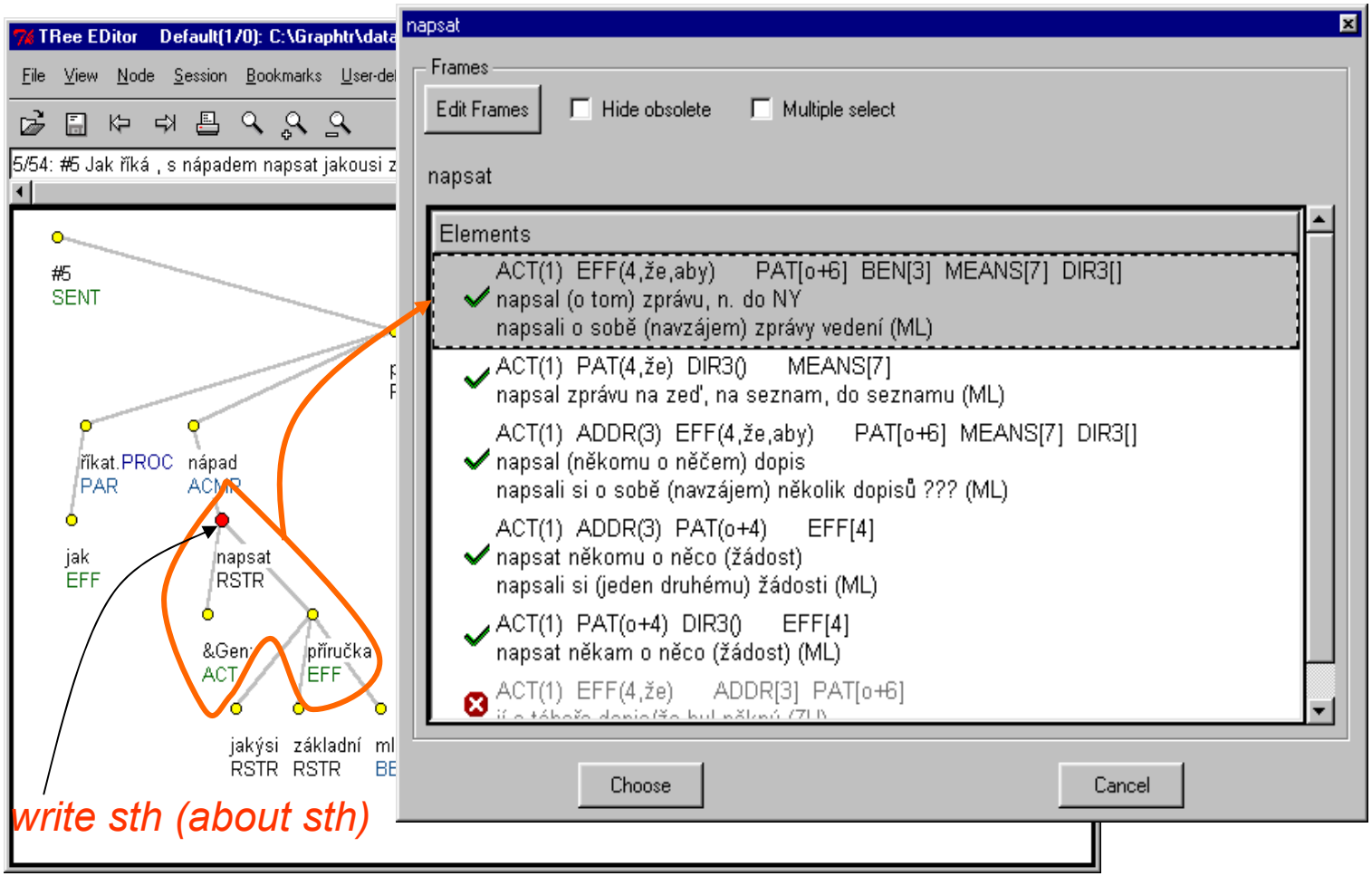
senses:

resign

win

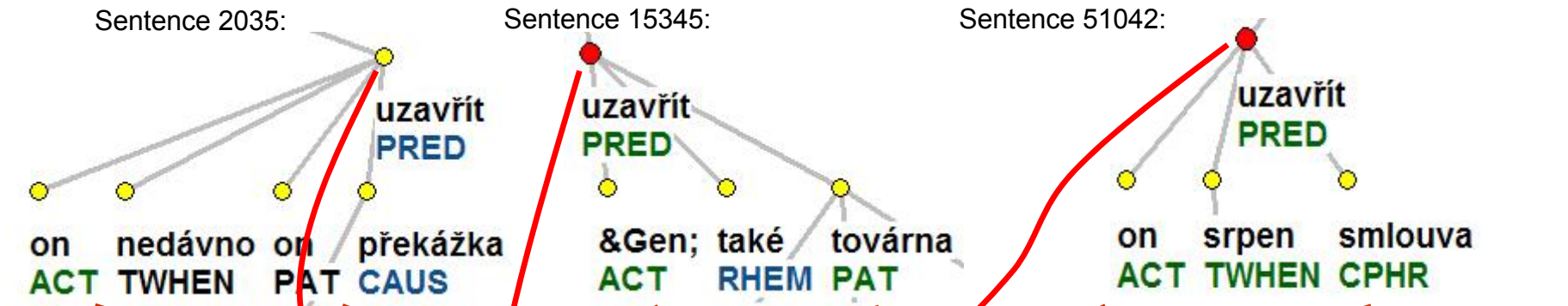
ask

Valency Lexicon and TrEd



Corpus <-> Valency Lexicon

- **Corpus** – occurrences of „uzavřít“ (*to close*) :



- **Lexicon**:

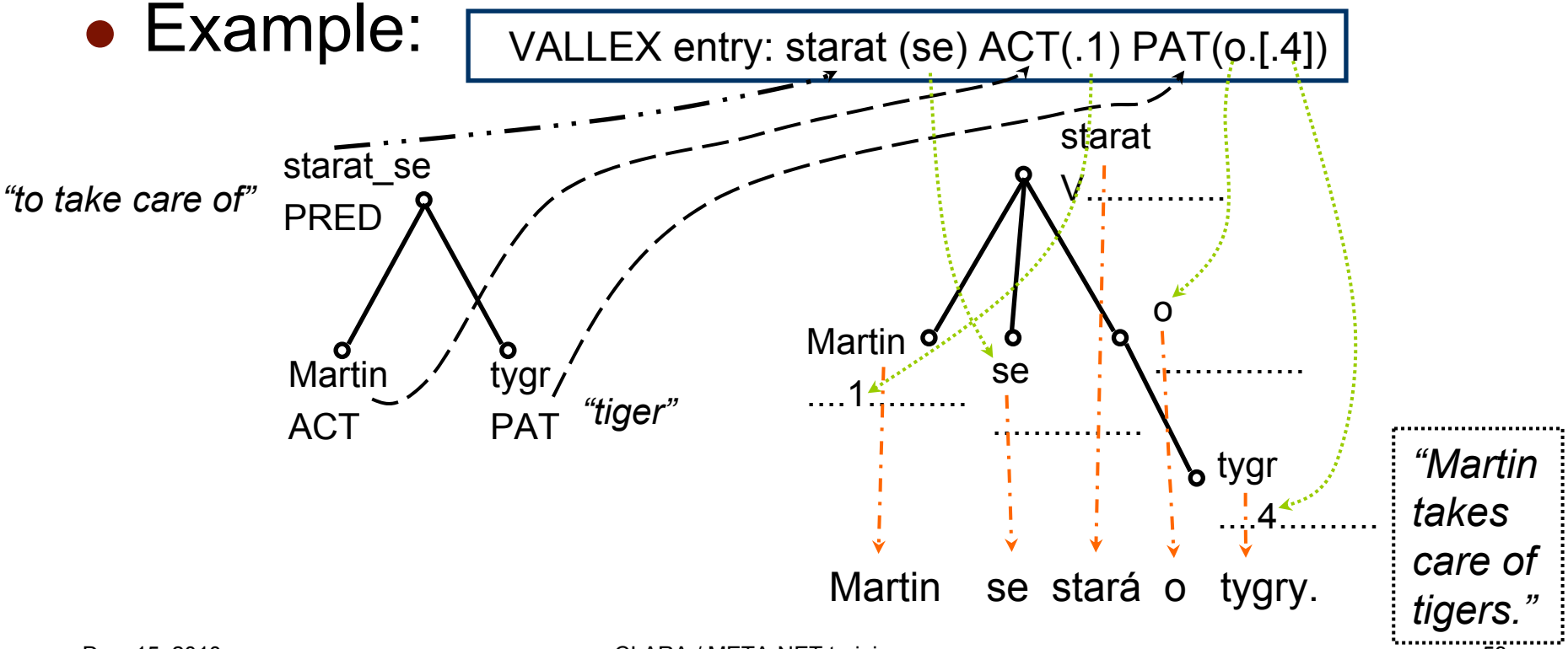
ENTRY: uzavřít

vf$_1$: ACT(.1) CPHR({smlouva}.4)
ex: u. dohodu *(close a contract)*

vf$_2$: 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:



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

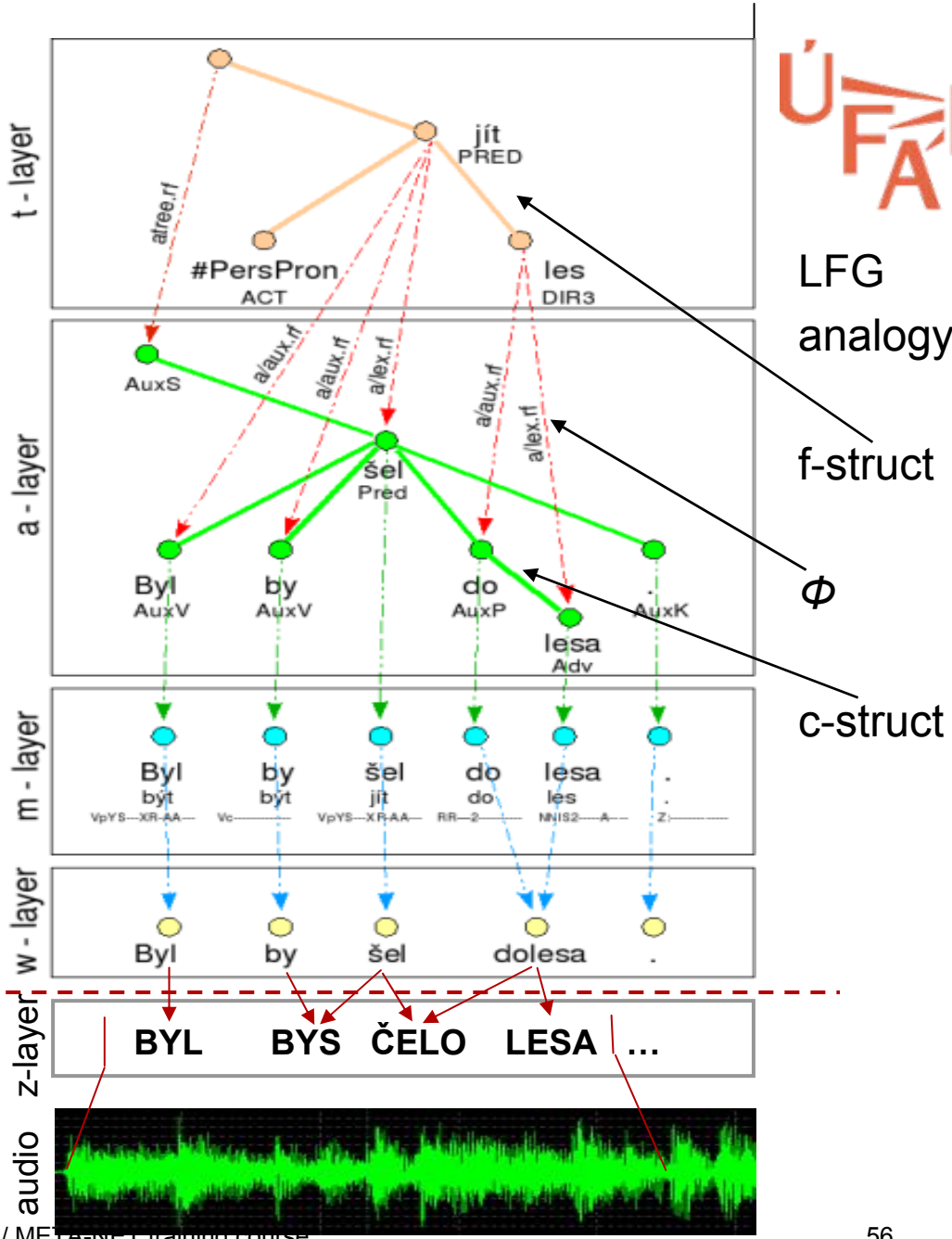
$\Rightarrow$ **PML**

(Prague Markup Language)

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

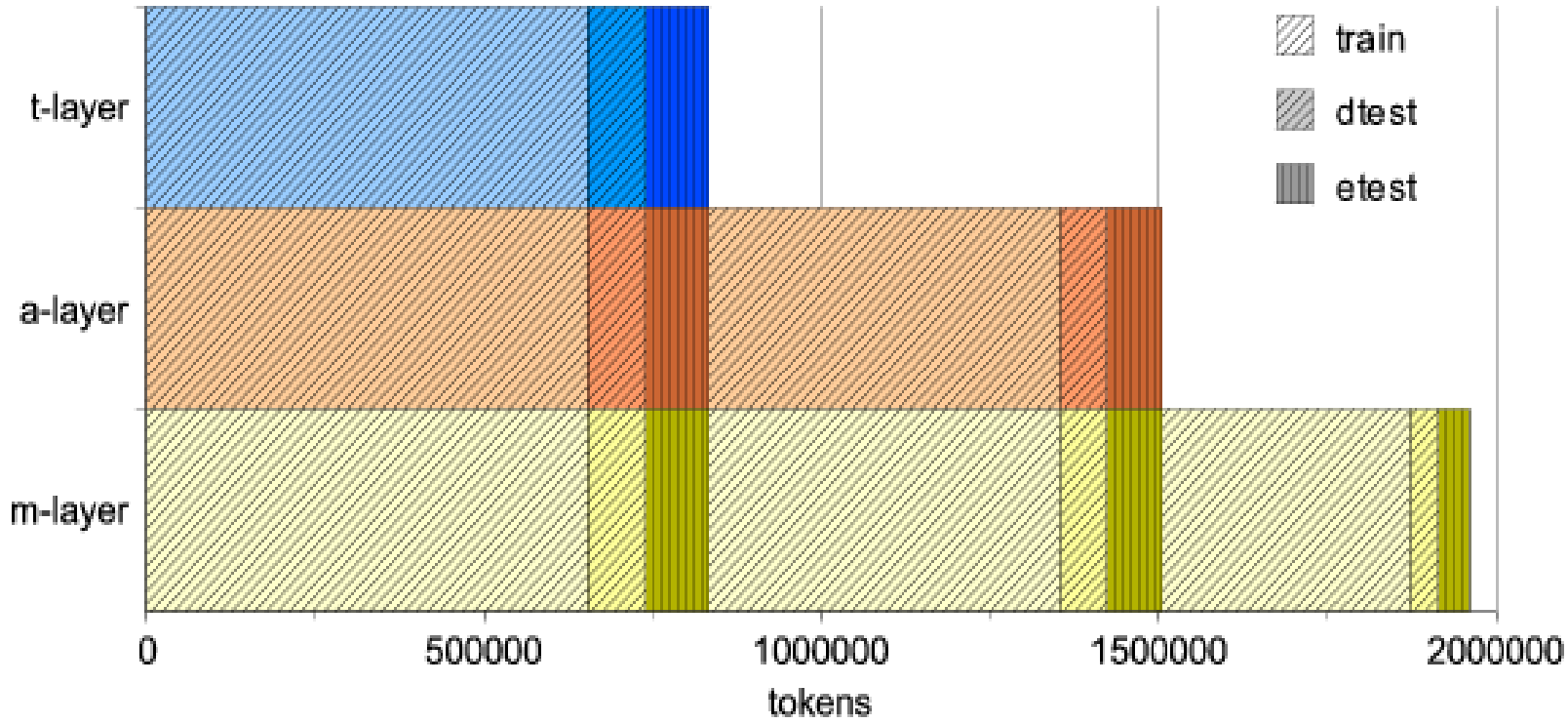
PML/XML Annotation Layers

- 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")]



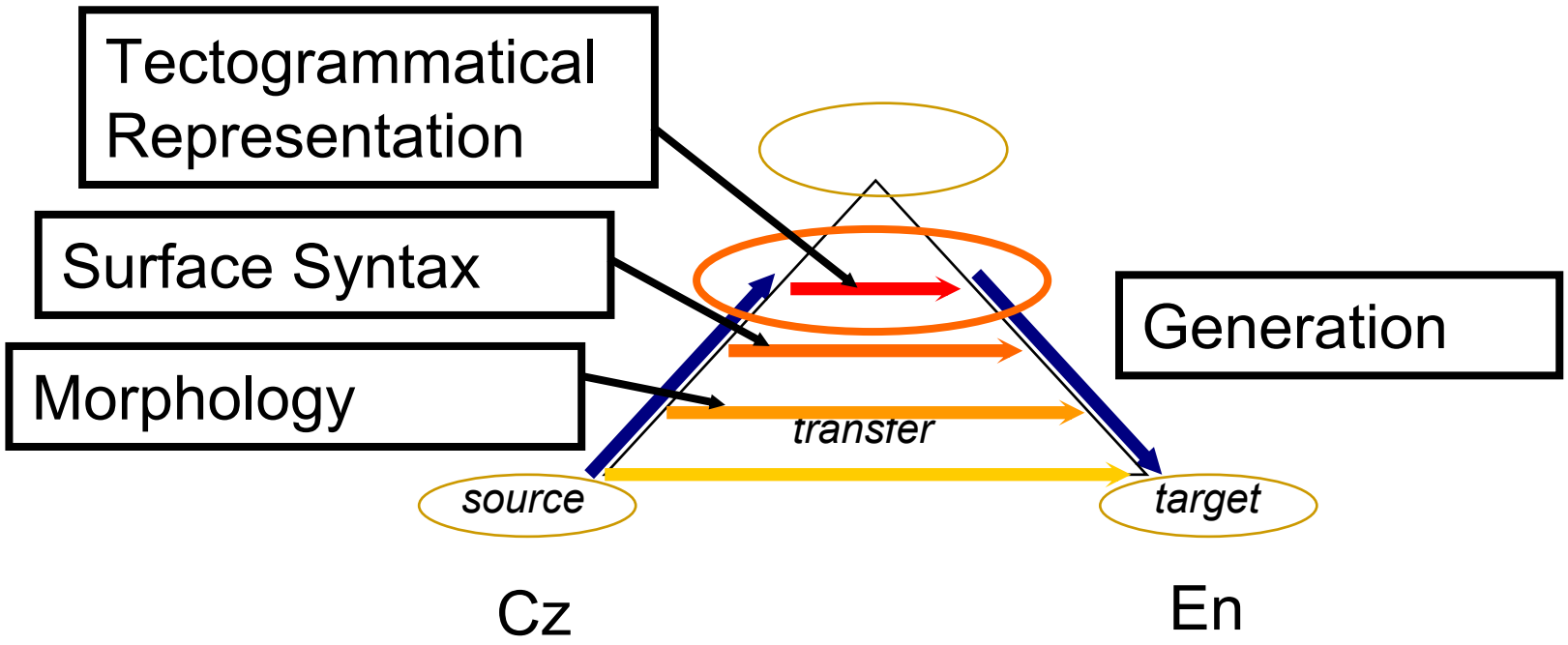
PDT 2.0: The Data

- Data sizes

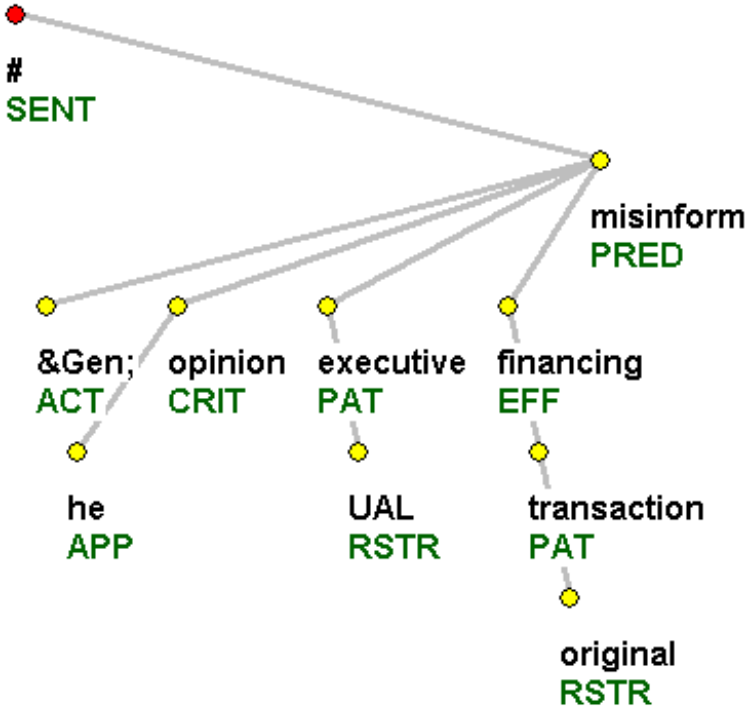


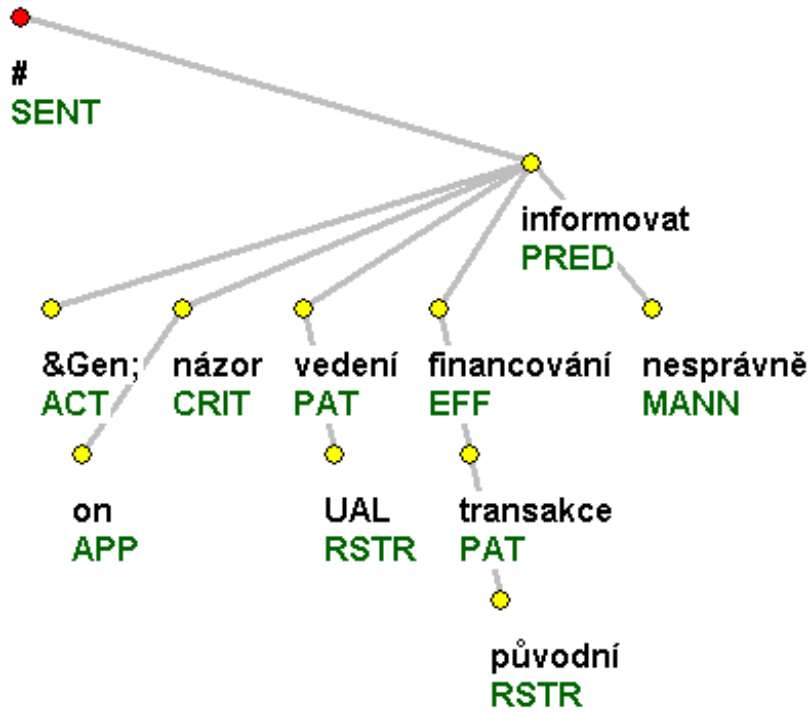
Tectogrammatical Layer in Machine Translation

- The Translation ("Vauquois") triangle



Dependency trees in MT



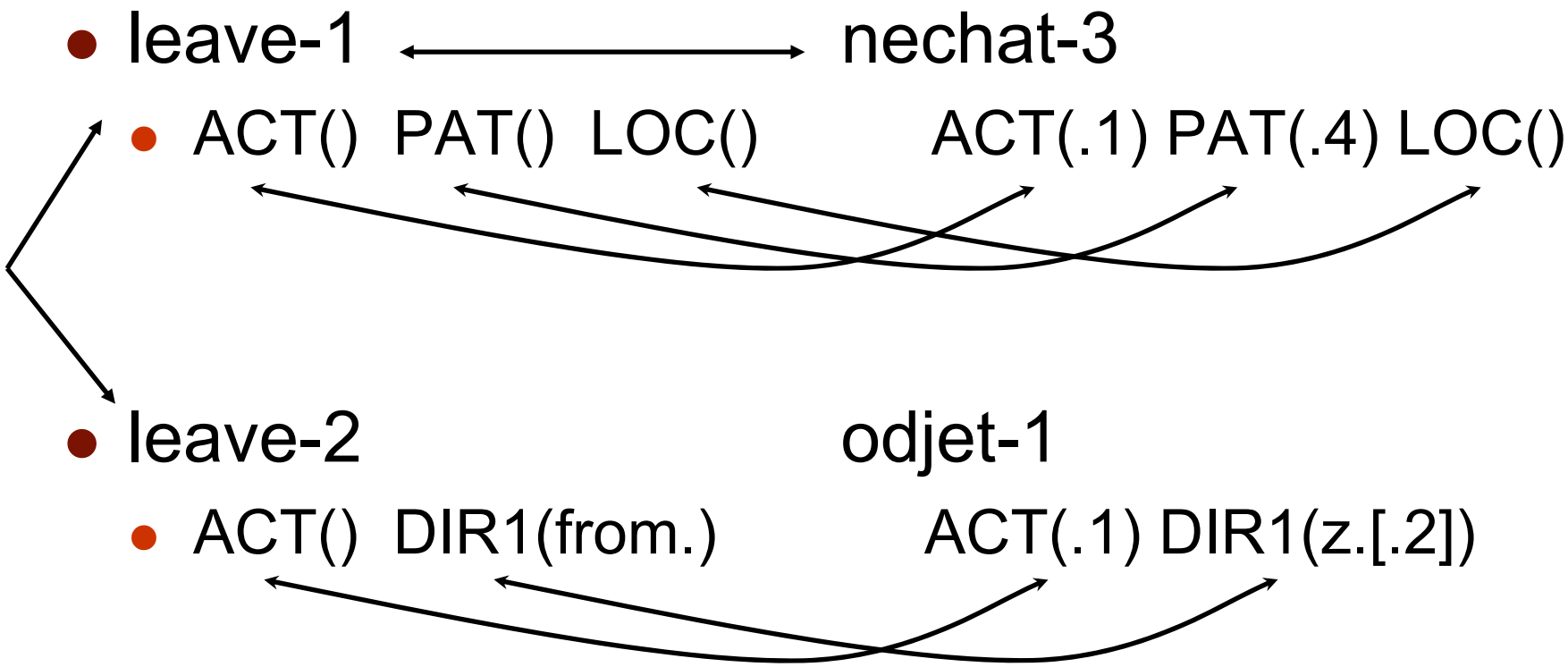


According to his opinion UAL's executives were misinformed about the financing of the original transaction.

Podle jeho názoru bylo vedení UAL o financování původní transakce nesprávně informováno.

Transfer:
- structure (~0)
- lexical
- functions
- grammatical

Valency and Translation



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)
- http://www.clsp.jhu.edu: Workshop 2002
  - Using TL for MT Generation