From Morphology to Semantics: the Prague Dependency Treebank Family

Jan Hajič

Charles University in Prague
Institute of Formal and Applied Linguistics
LINDAT-Clarin and META-NET (CZ)

Czech Republic

History

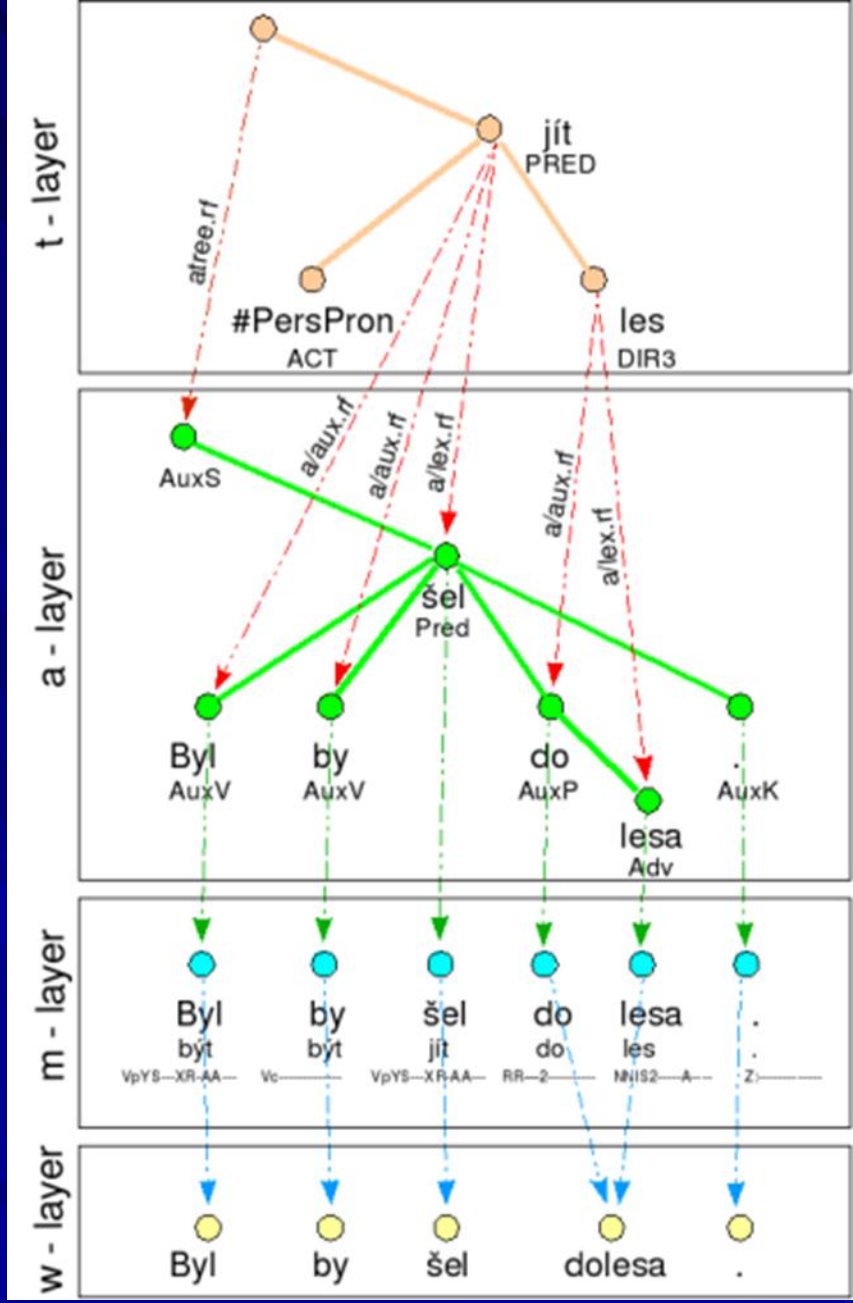
- LDC: Penn Treebank I (1993)
  - We want it too!
- But:
  - LDC's unlikely to do Czech (soon ☺)
  - Prague (old time structuralist) tradition: dependency
- 1995: decision to build our own treebank
  - Started 1996 with a specification grant
  - Tool development, annotation since 1997
  - First PDT (1.0) published in 2001 (LDC2001T10)
    - Morphology and syntax only, but > 1M words
  - PDT 2.0 2006 (LDC2006T01)
    - Full annotation & correction of 1.0
  - Other treebanks: 2004, 2012 (more to come, also by other groups)

Prague Dependency Treebanks the Basics

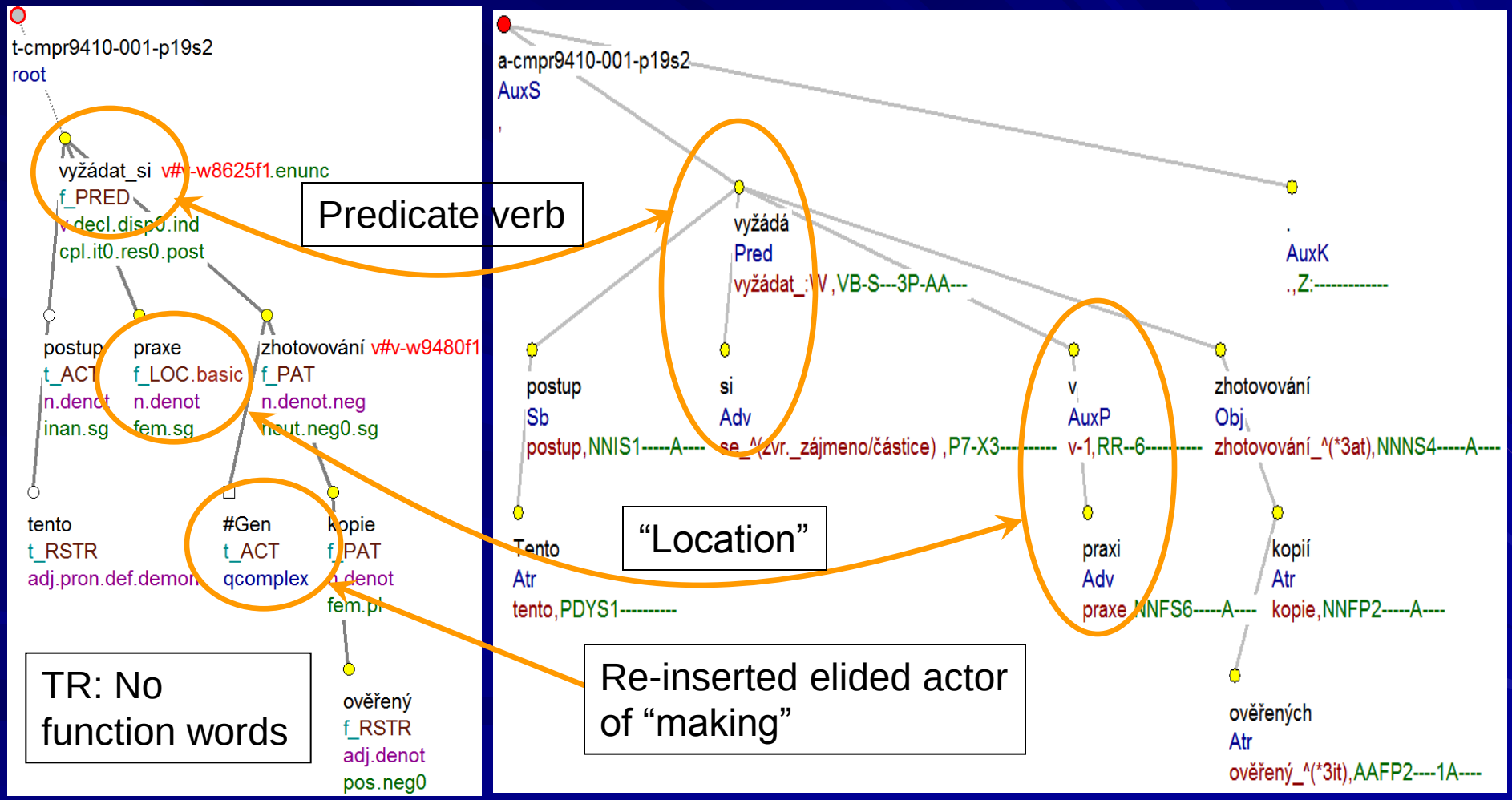
- General Features
  - Multilayered annotation, interlinked layers
  - Dependency-based syntax (both surface and deep)
    - Includes semantic functions, valency dictionary(-ies)
  - Information structure of the sentence (topic/focus)
  - Grammatical and textual co-reference, new: bridging
  - New: discourse relations (not published yet)
- Languages: Czech, English (also parallel), Arabic:
  - Indonesian, Urdu, Russian, … (Student work on samples)
  - (Auto) conversion from other treebanks (25 so far; experimental)
  - Spoken: Czech and English (non-parallel, dialogs)

The Layers

- Three basic layers
  - Morphological layer
  - Surface syntax (“a”) layer
  - “Tectogrammatical” layer: underlying syntax, semantic roles (valency), inf. structure, co-reference (anaphora)
- Format
  - Prague Markup Language (XML + Schema)
- (Speech:
  - Additional layers:
    - audio, transcript)



Tectogrammatical vs. Analytical (Surface) Syntax

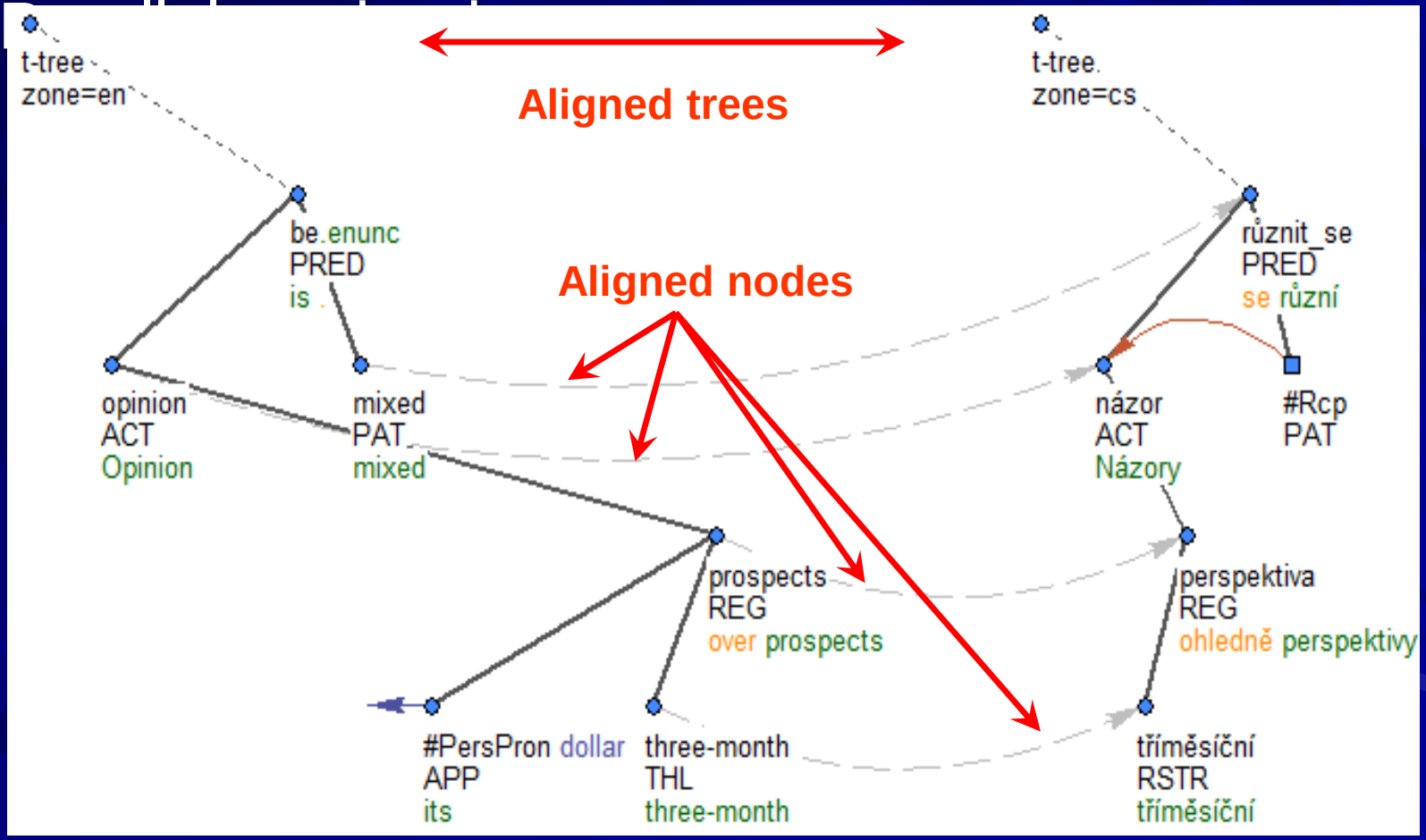


In practice, that procedure will require making of certified copies.

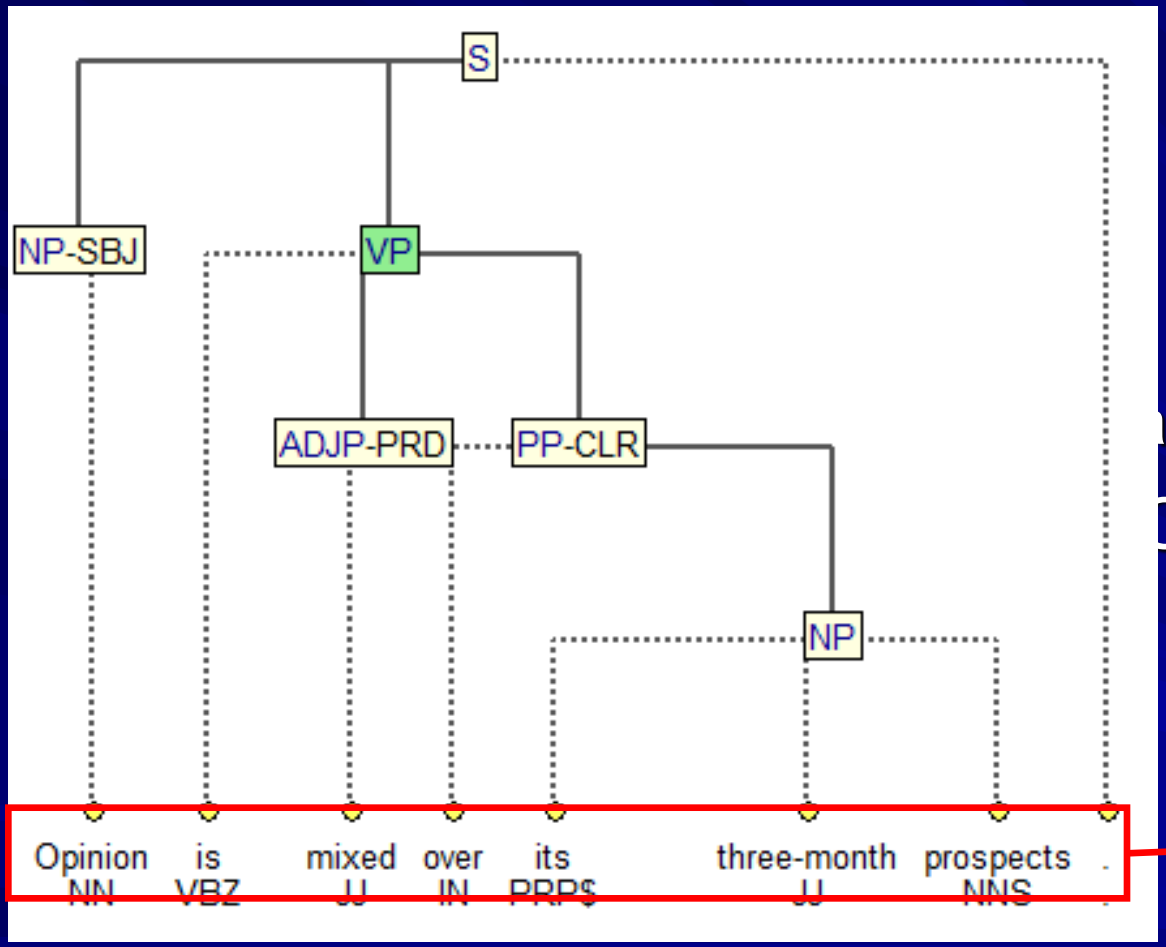
PDT-style Treebanks (written language)

- Czech
  - Prague Dependency Treebank
    - Complex annotation, all levels, additional annotation
  - Translation of Penn Treebank, aligned
    - Tectogrammatical layer only, no information structure
      - Analytical, morphology: automatic tools
        - Will be manually revised later
- English
  - Re-annotation of Penn Treebank, TR only so far
- Arabic
  - New morphology, analytical syntax, sample TR only

The Prague Czech-English Dependency Treebank (PCEDT) 2.0



The Prague Czech-English Dependency Treebank (PCEDT) 2.0



tics")

Czech

Názory na její tříměsíční perspektivu se různí.

The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax

	Czech	English
Sentences	49,208	
a-nodes (automatic)	1,151,150	1,173,766
t-nodes (manual)	931,846	838,212

- Pub
  - A wsing
    ar

	Alignment links
a-layer	1,214,441
t-layer	727,415

PCEDT 2.0
The Alignment(s)

- Czech-English alignments
  - Sentence-level (manual, natural due to translation)
    - At both syntactic levels
  - Word (node) level
    - automatic, test section manually corrected (in part)

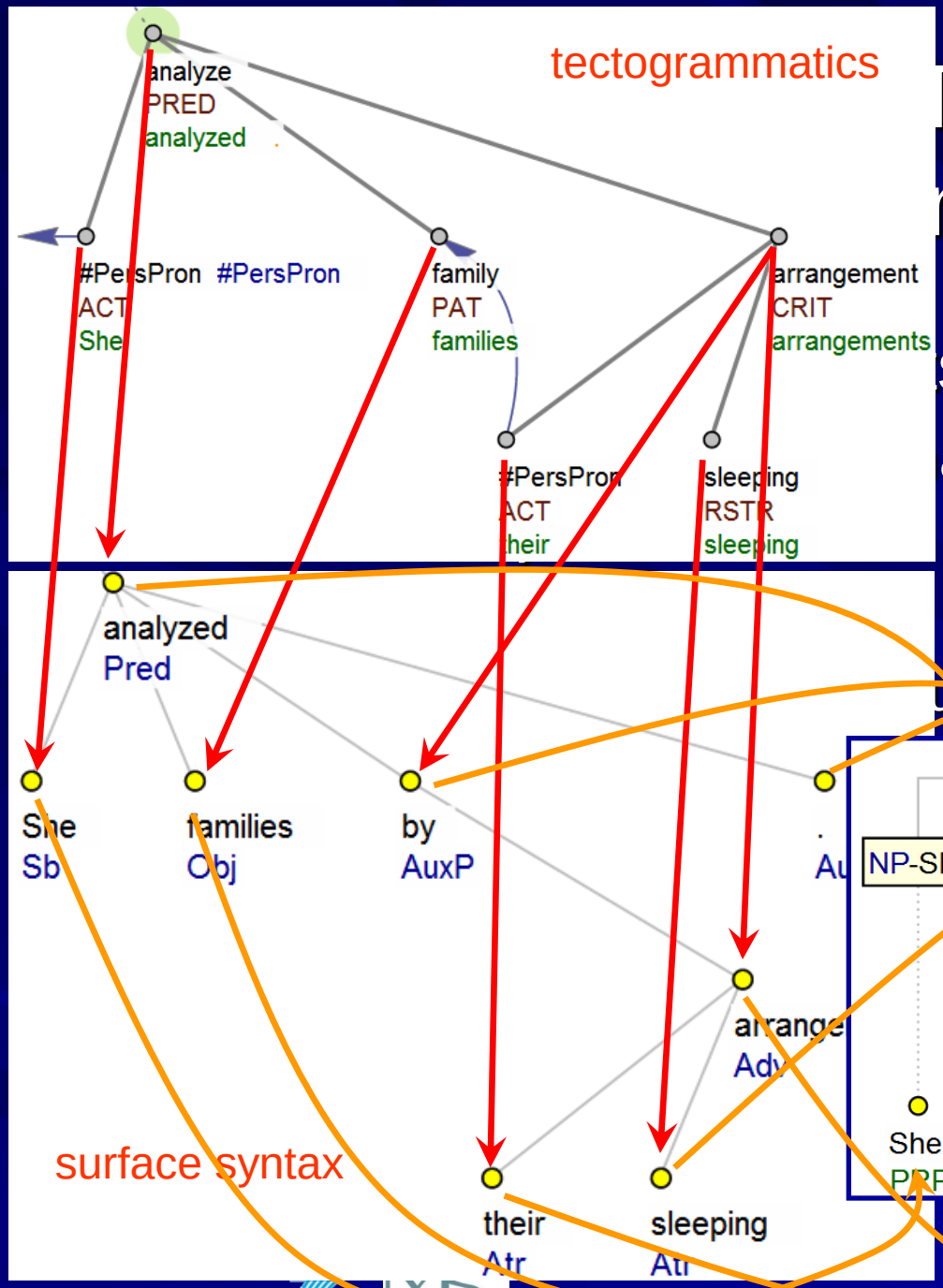
T 2.0
ment(s)

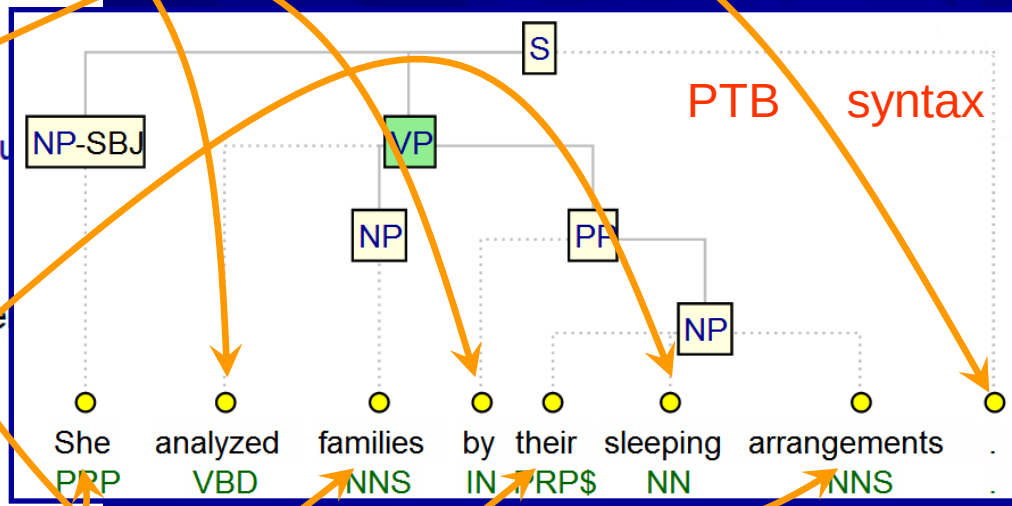
s

atural due to translation)

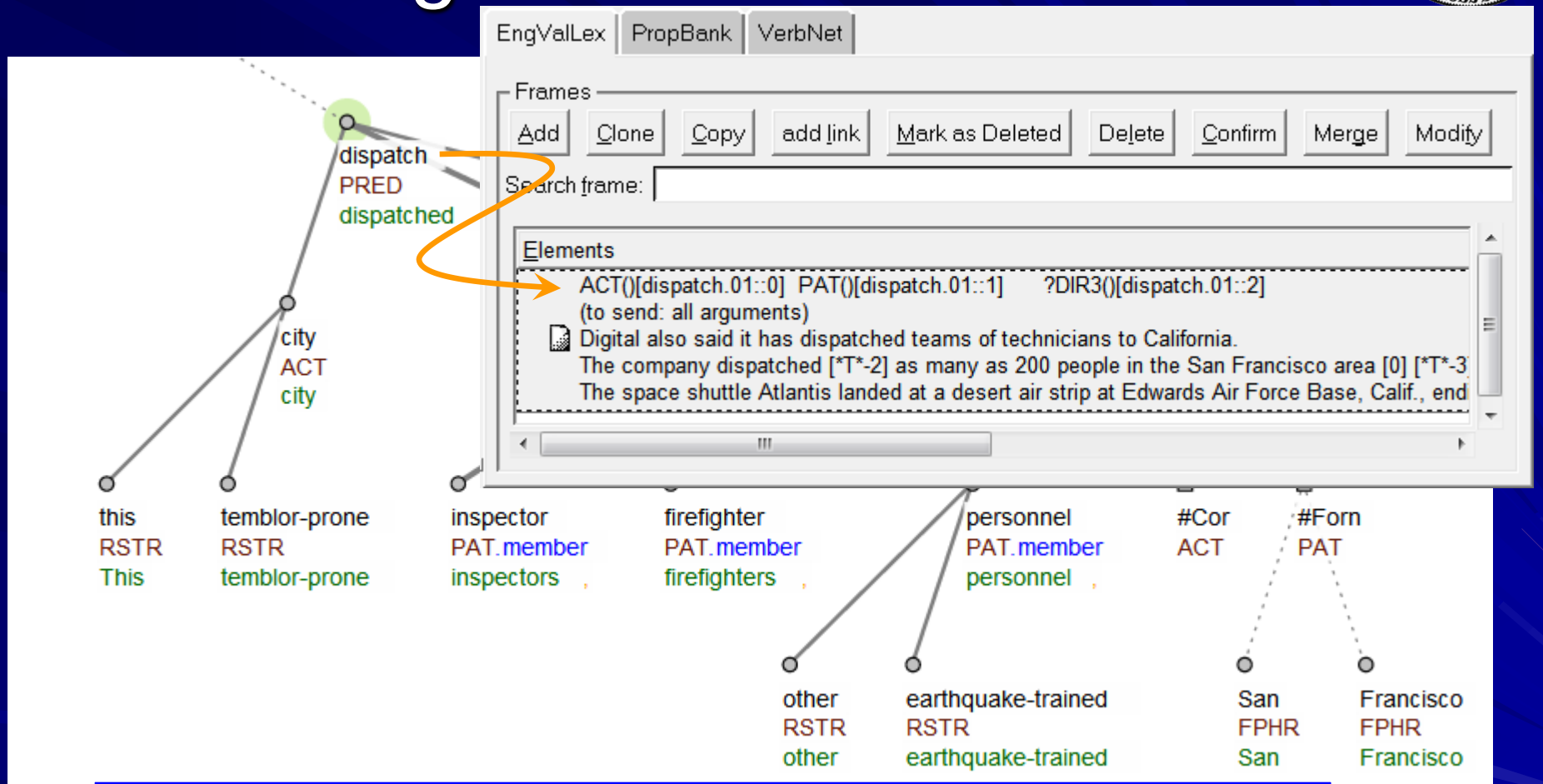
1

ally corrected, m → n





Tectogrammatical annotation



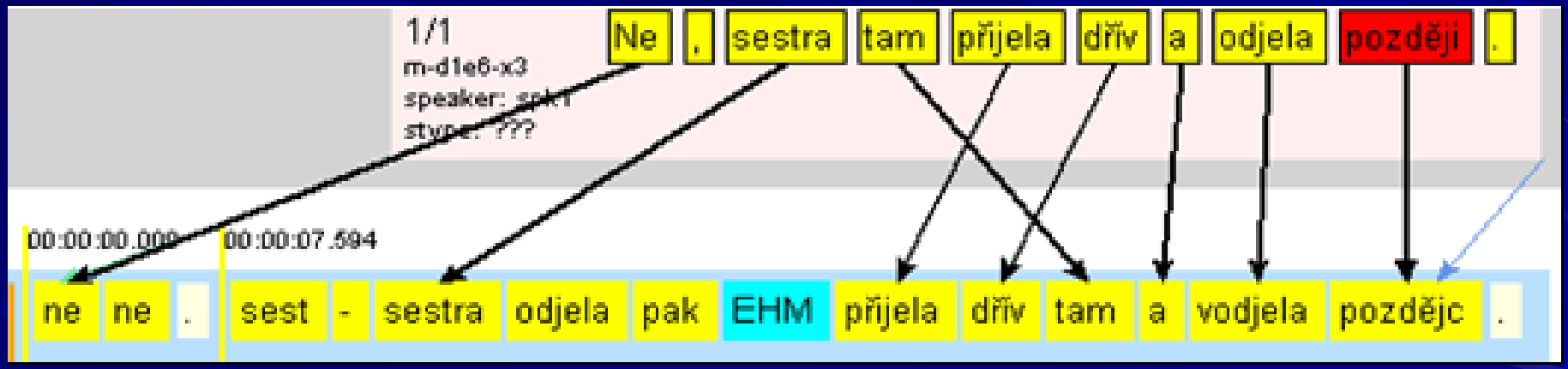
This temblor-prone city dispatched inspectors, firefighters and other earthquake-trained personnel *-1 to aid San Francisco.

PDT-style Treebanks (spoken language)

- Specifics of spoken language
  - Short sentences but unclear segmentation
    - Sentence breaks must be (re)annotated
  - Ungrammatical (esp. for Czech – coll.)
    - Annotation based on written-language rules difficult if not impossible
- …additional decisions:
  - Change annotation?
  - Change the input? (but original must be kept)

Spoken corpora

- Solution: “Speech reconstruction”
  - Keep audio, word-for-word transcript
    - Adds two layers to the annotation scheme: audio, transcript
  - Add edited text: ***LINKS*** to original transcript / audio



  - Annotate edited text (using usual guidelines)

Accompanying Tools

- TrEd (http://ufal.mff.cuni.cz/tred)
  - Annotation, View/Browse and Search environment
  - Open source, perl
  - Search and visualization: **PML-TQ**
    - Powerful query language for complex NLP annotation, esp. tree-based
- Treex (http://ufal.mff.cuni.cz/treex)
  - Modular NLP processing environment
  - Easy handling of complex NLP-annotated data
  - Modules exists for Czech, English data processing
    - incl. 3rd-party tools integrated into Treex
  - CPAN-distributed

Lessons Learned (1)

- Positive experience
  - Dependency style
  - Separate layers of annotation
    - Most importantly: separate surface syntax vs. deep syntax
  - Specific format and specific graphical tools (TrEd et al.)
    - Stand-off annotation
  - Spoken annotation “trick” with speech reconstruction
    - Still, additional guidelines needed
- Negative experience
  - Lots of time spent on consistency checking
    - Annotator training: guidelines too detailed
    - Prevents crowdsourcing
  - Lots of time goes to final quality checking and corrections
    - min. 3 PY for PDT, PCEDT

Lessons Learned (2)

Acknowledgements:
Charles University research funds ("PRVOUK")

- For future projects
  - Annotation in small teams
    - "Phenomenon-by-phenomenon"
  - Ongoing quality checking, time allotted for final QC
    - Error discovered at annotation time *much* cheaper to correct
    - Consequences for tool selection ("intelligent" annotation SW)
  - Need for excellent software and annotator's support
    - Programmers' efforts always underestimated
    - "helpdesk" for annotators important (usually former annotator)
    - Organization, statistics, watchdog
      - Single repository for annotated data
  - Payment
    - Annotator's incentives work (for speed of annotation)
  - Speed of annotation vs. quality
    - Almost no correlation

Happy [redacted] Birthday!