The Prague Czech-English Dependency Treebank

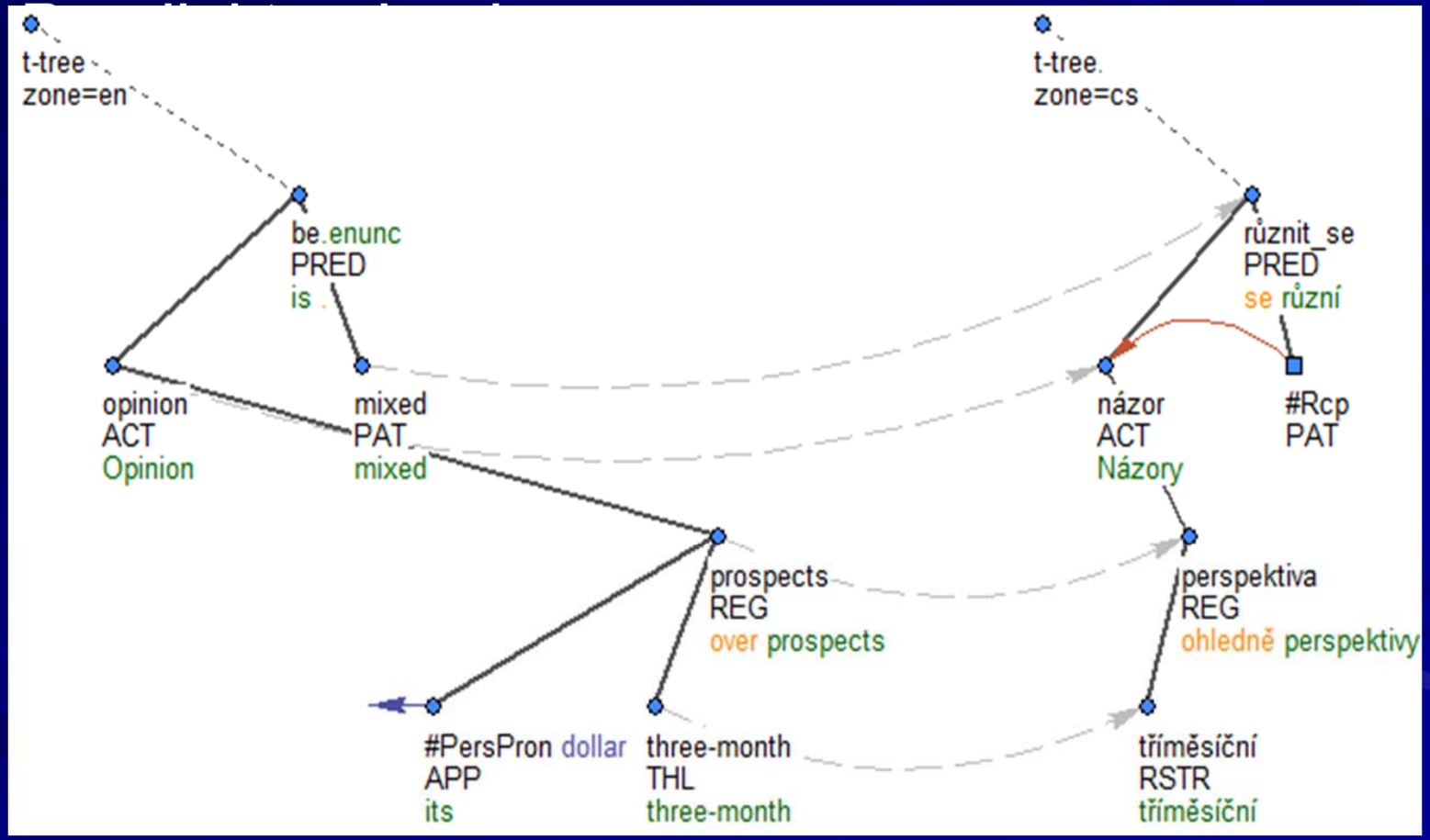
Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský

Charles University in Prague
Institute of Formal and Applied Linguistics
Czech Republic

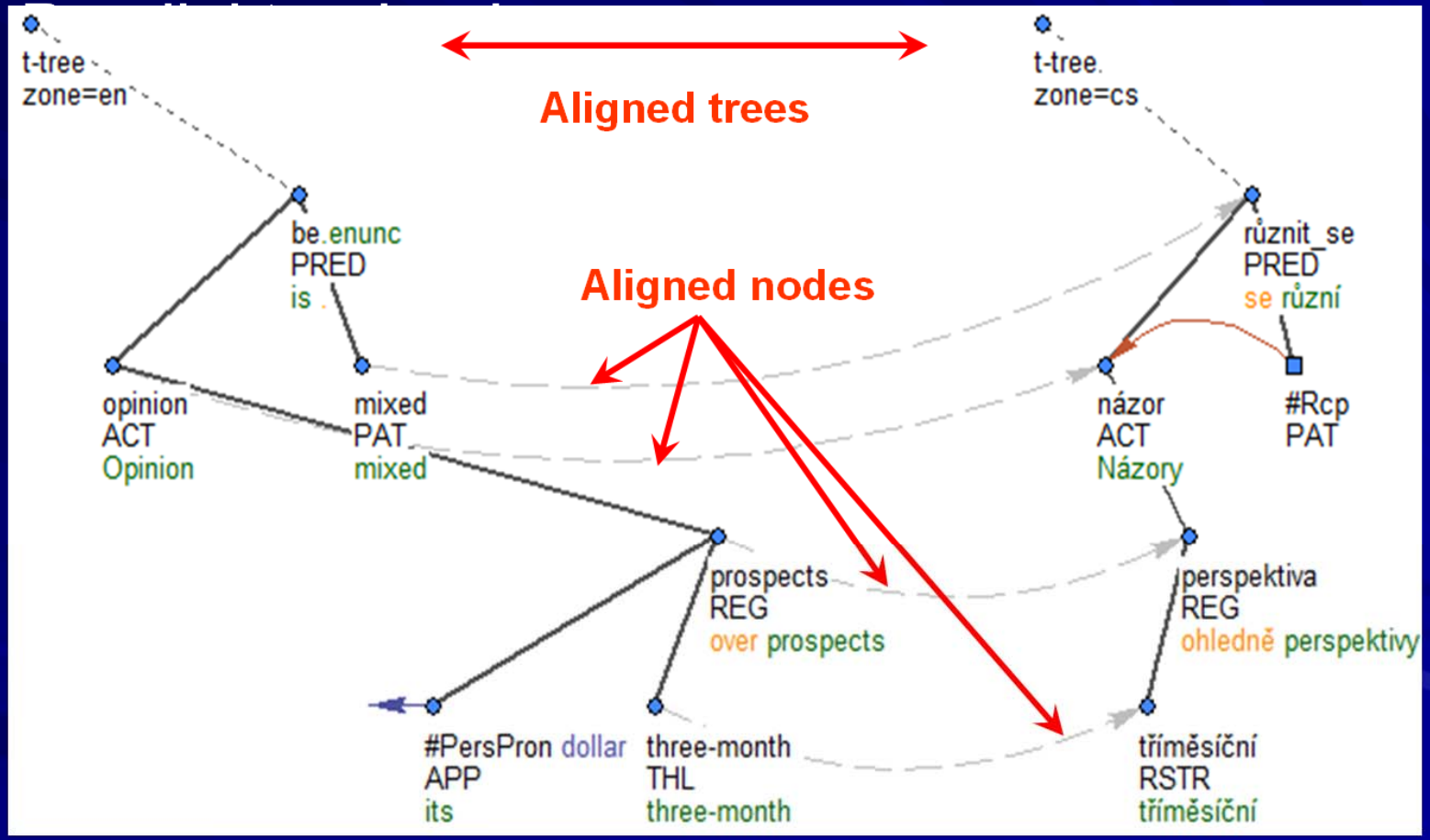
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank

The Prague Czech-English Dependency Treebank (PCEDT) 2.0



The Prague Czech-English Dependency Treebank (PCEDT) 2.0



The Prague Czech-English Dependency Treebank (PCEDT) 2.0

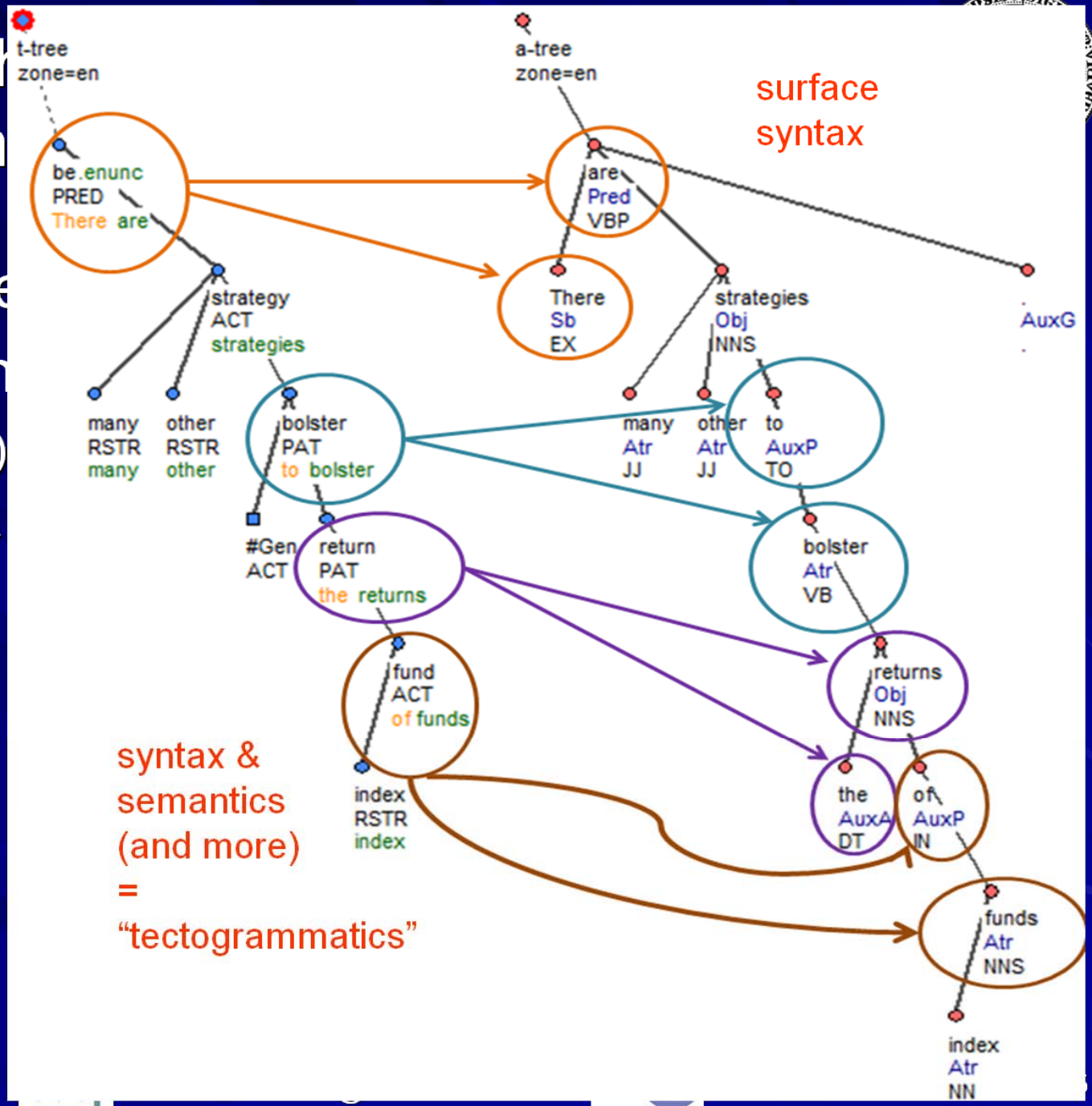
- Parallel treebank

The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")

Th
Depen

- Parallel tre
- Dependen
  - (surface)
  - syntax &



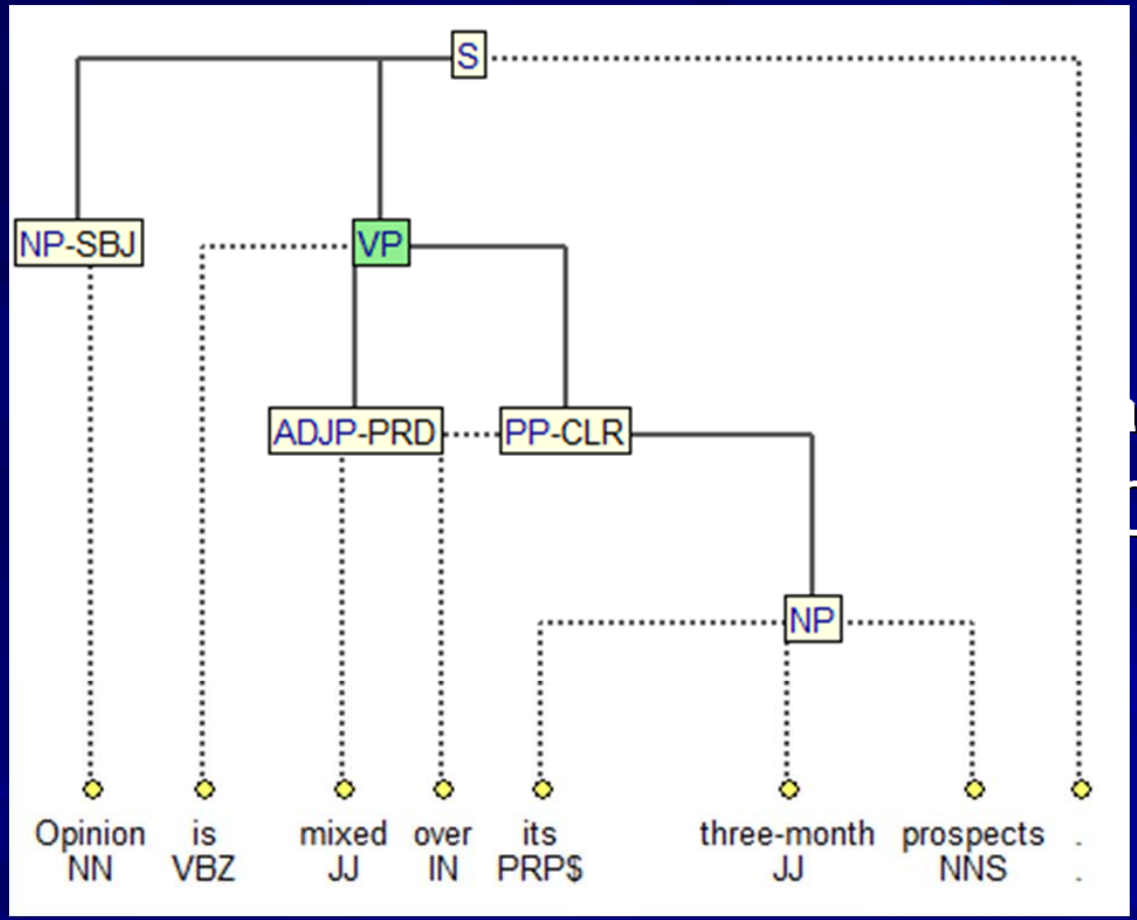
The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")

The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style (“Prague”)
  - (surface) syntax
  - syntax & semantics (“tectogrammatics”)
- Penn Treebank translation into Czech

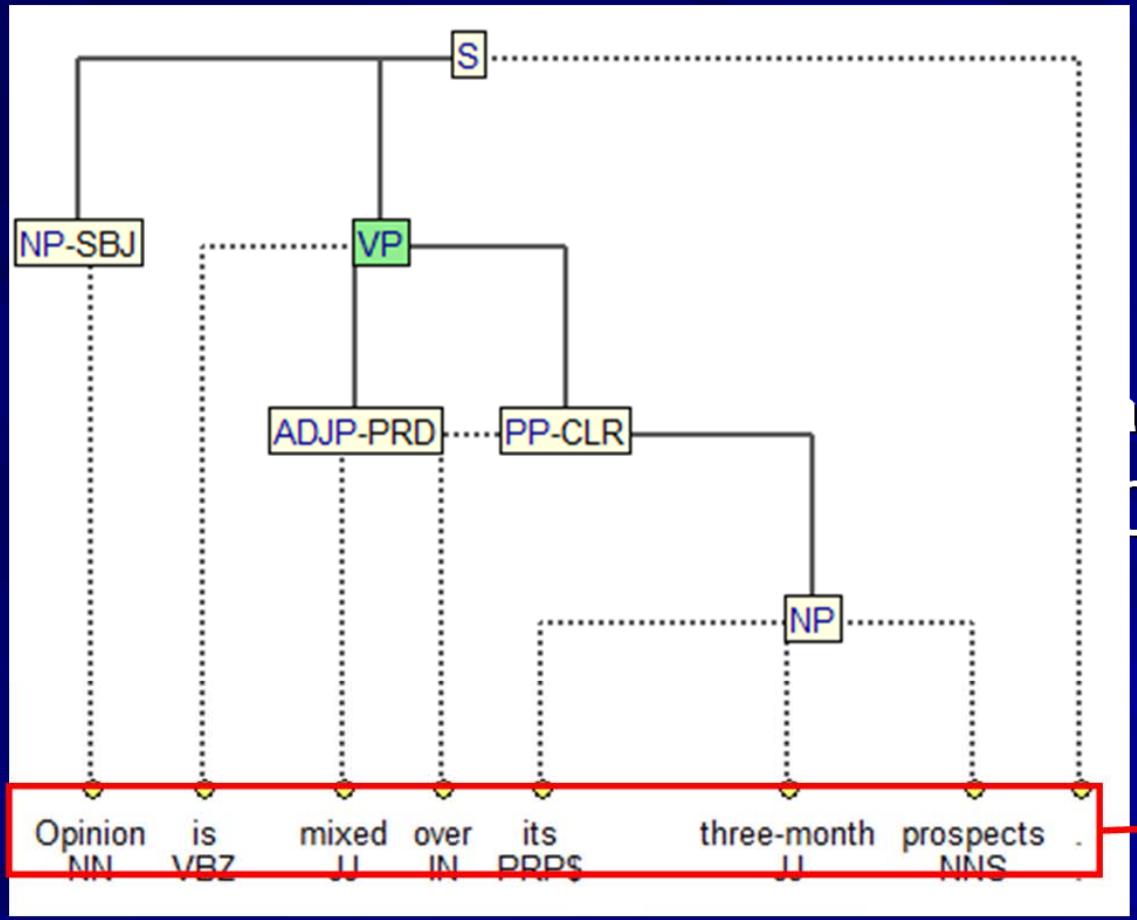
The Prague Czech-English Dependency Treebank (PCEDT) 2.0



tics”)

Czech

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
  - syntax & semantics (“tectogrammatics”)
- Penn Treebank translation into Czech

The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style (“Prague”)
  - (surface) syntax
  - syntax & semantics (“tectogrammatics”)
- Penn Treebank translation into Czech
- 1 million words

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

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

The Prague Czech-English Dependency Treebank (PCEDT) 2.0

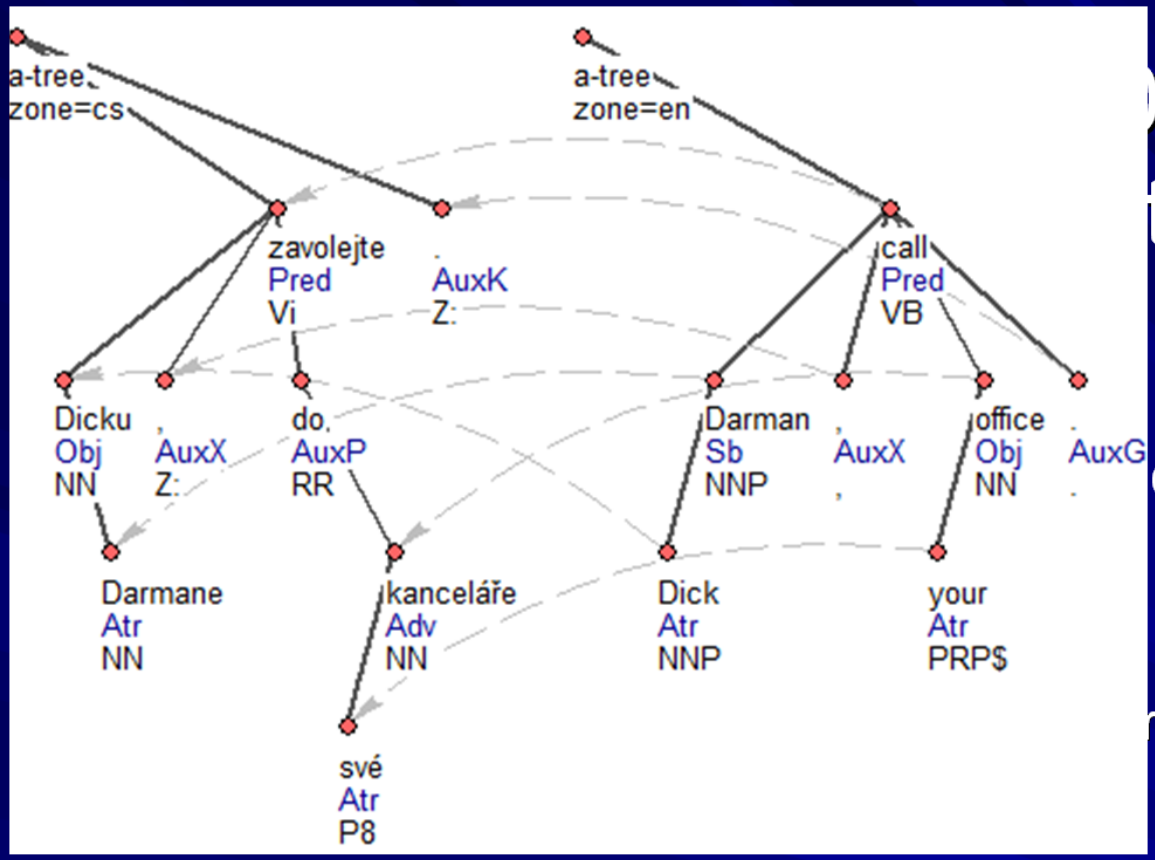
- Parallel treebank
- Dependency style (“Prague”)
  - (surface) syntax
  - syntax & semantics (“tectogrammatics”)
- Penn Treebank translation into Czech
- 1 million words

The Prague Czech-English Dependency Treebank (PCEDT) 2.0

- Parallel treebank
- Dependency style ("Prague")
  - (surface) syntax
  - syntax & semantics ("tectogrammatics")
- Penn Treebank translation into Czech
- 1 million words
- Getting ready at LDC
  - Also available through LINDAT-Clarin and META-SHARE

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)



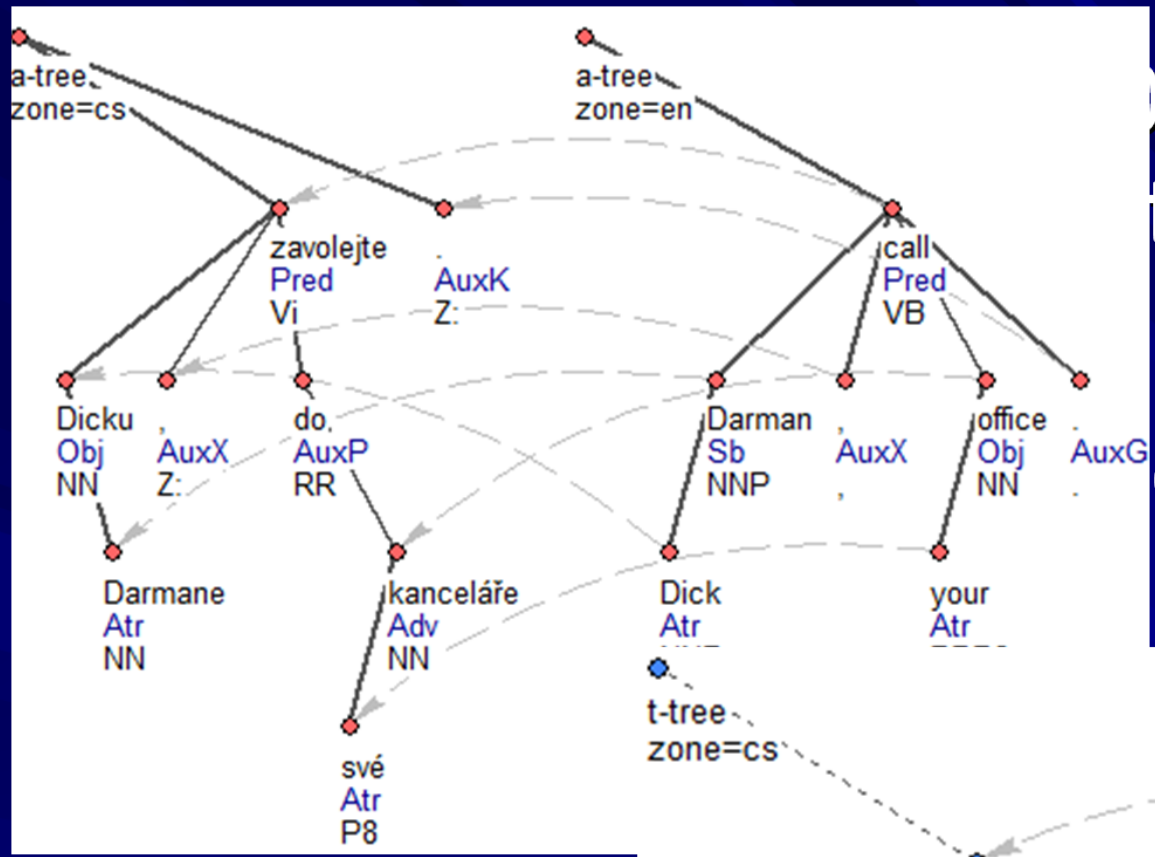
t(s)

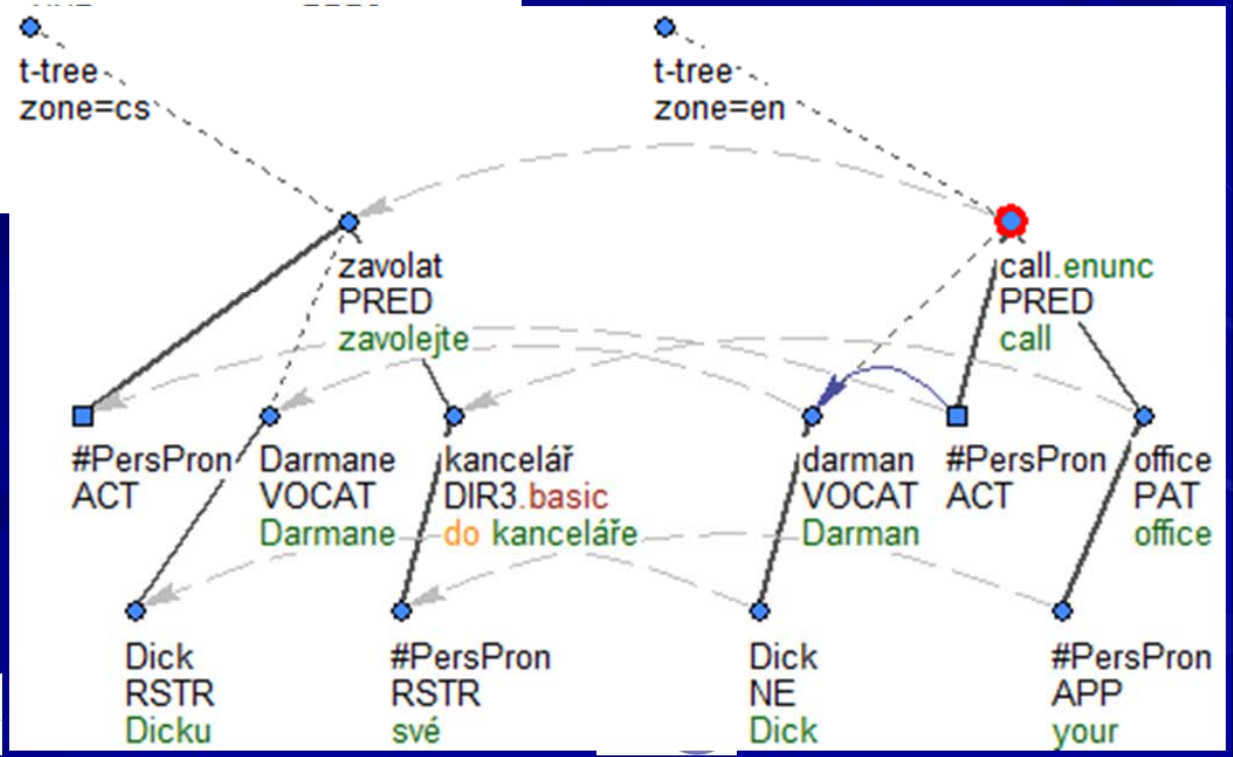
due to translation)

rected (in part)

t(s)

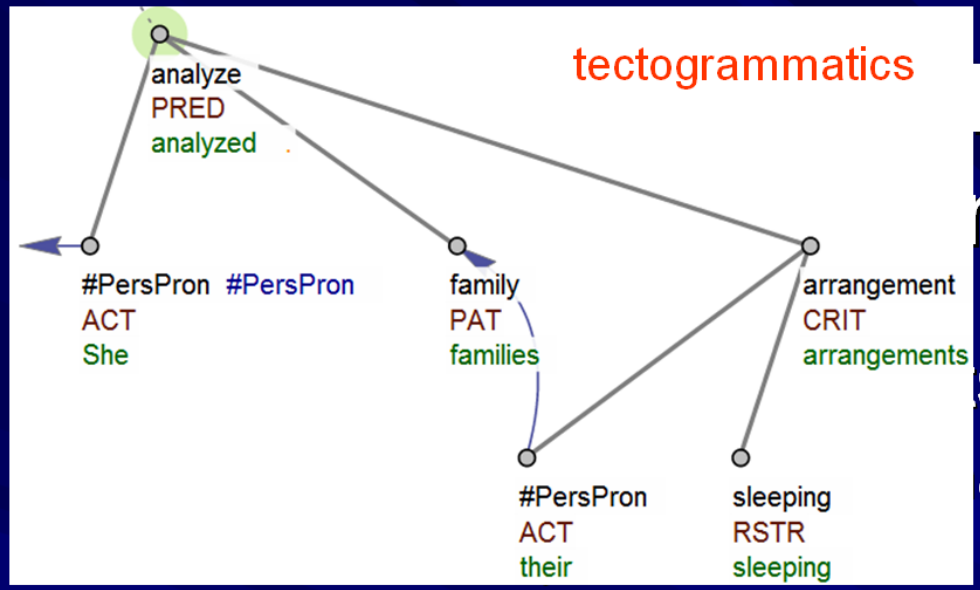
due to translation)

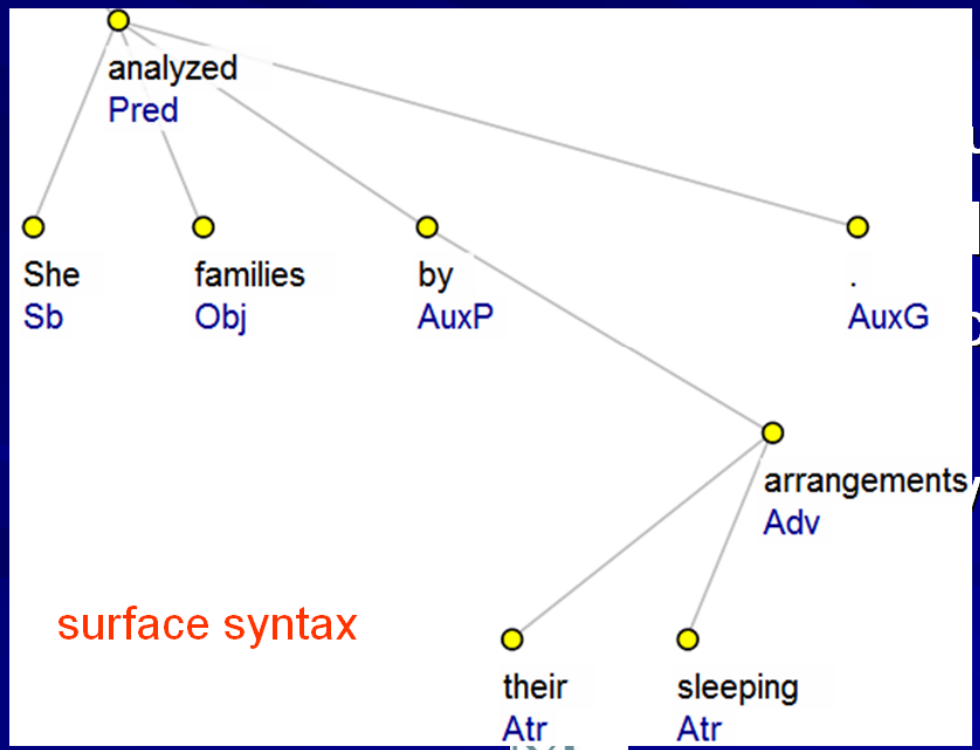




PCEDT 2.0
The Alignment(s)

- Czech-English alignments
  - Sentence-level (manual, natural due to translation)
    - At both syntactic levels
  - Word (node) level
    - automatic, test section manually corrected (in part)
- Between annotation levels
  - Tectogrammatics to surface syntax
    - m → n, incl. 1 → 0
  - Surface syntax to word level (1 → 1)





T 2.0

ment(s)

s

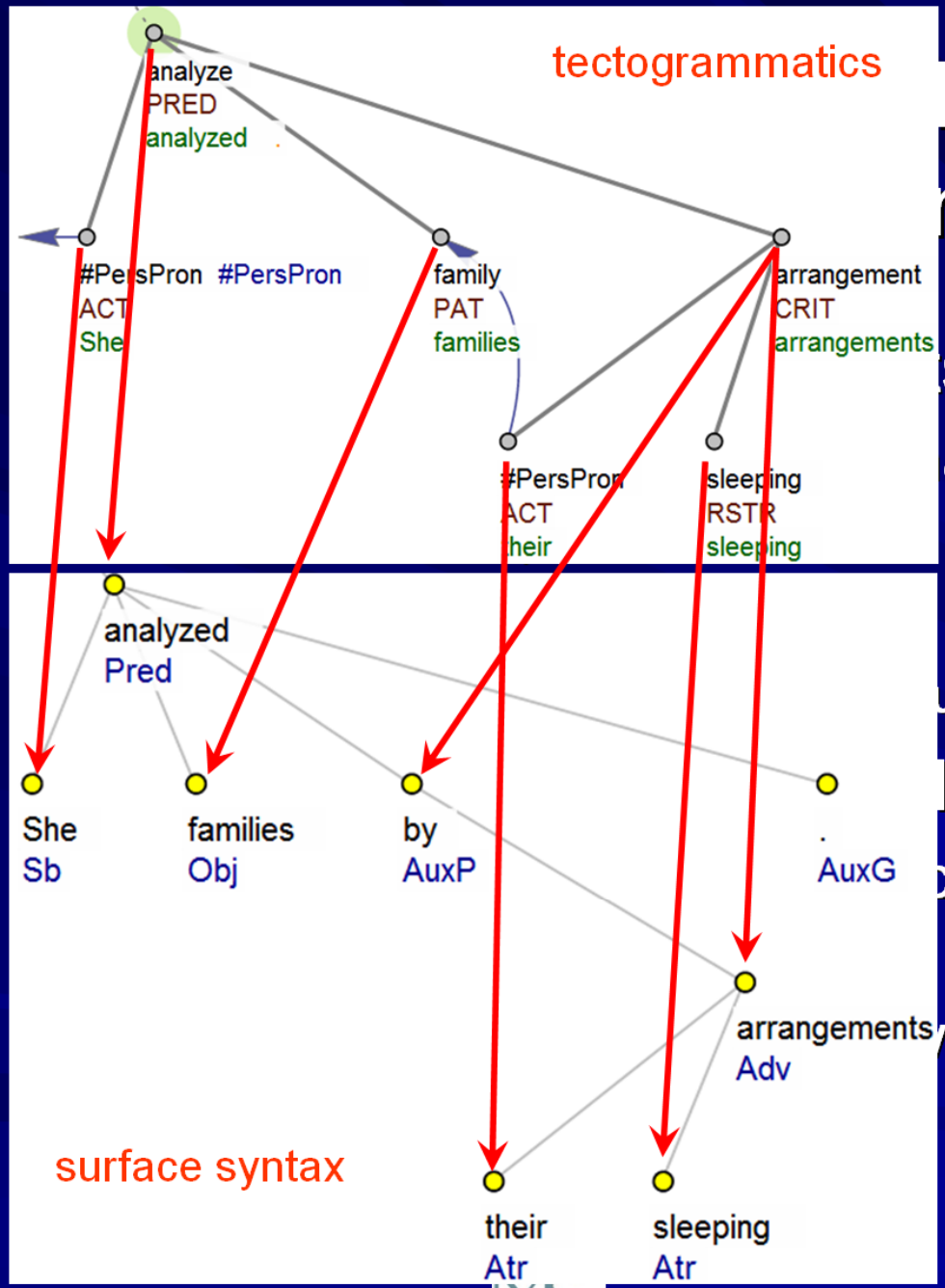
atural due to translation)

ually corrected (in part)

s

ce syntax

vel (1 → 1)



T 2.0
ment(s)

s

atural due to translation)

ually corrected (in part)

ls

ce syntax

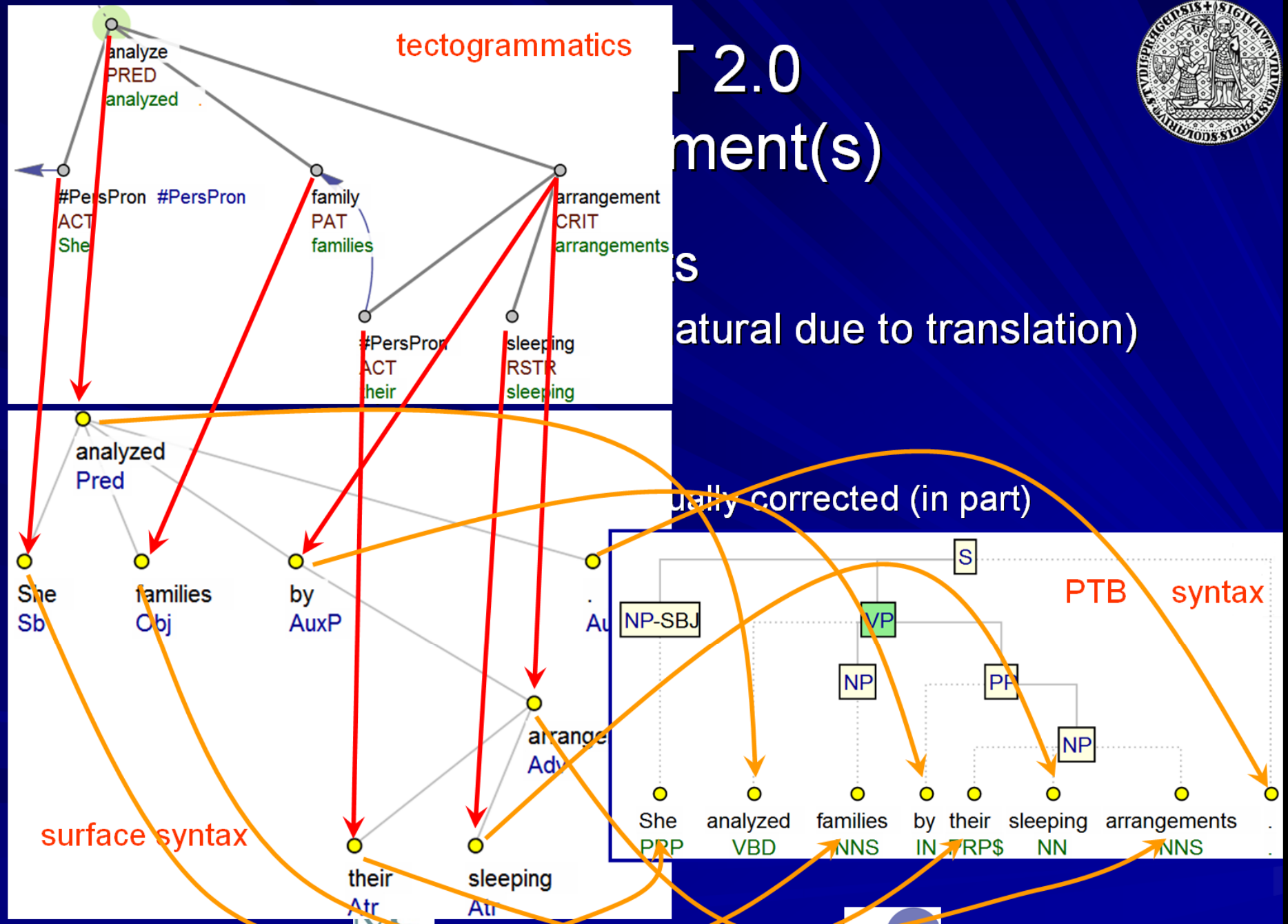
vel (1 → 1)

T 2.0
ment(s)

s

atural due to translation)

ually corrected (in part)



Surface syntax annotation

- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + „pure“ POS tags from PTB
  - Automatic (from PTB) + a few manual corrections

Surface syntax annotation

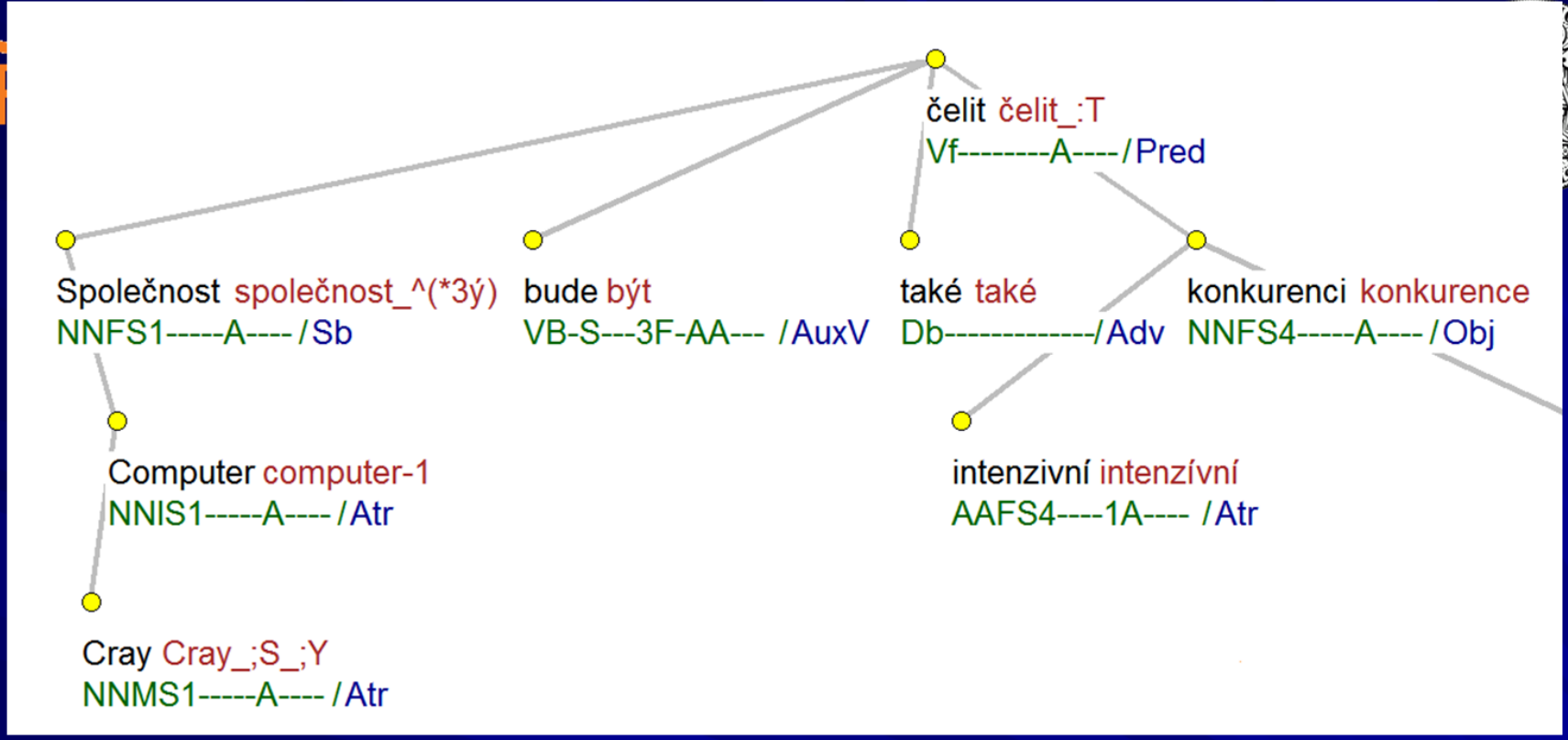
- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + „pure“ POS tags from PTB
  - Automatic (from PTB) + a few manual corrections

Surface syntax annotation

- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + „pure“ POS tags from PTB
  - Automatic (from PTB) + a few manual corrections
- Czech
  - PDT style, no change
  - Syntax: automatic, 2000 sentences fully manual for testing
  - Lemmatization and tagging: auto
    - 99%/96%, Spoustová et al. EACL 2009 (COMPOST tagger)
    - http://ufal.mff.cuni.cz/compost (Czech, English & other)
  - No p-level (of course ☺)

Surface syntax annotation

- English
  - Dependency (head rules + additions, manual corrections)
  - Function label (PDT-style) at all nodes (from PTB + rules)
  - Lemmatization + „pure“ POS tags from PTB
  - Automatic (from PTB) + a few manual corrections
- Czech
  - PDT style, no change
  - Syntax: automatic, 2000 sentences fully manual for testing
  - Lemmatization and tagging: auto
    - 99%/96%, Spoustová et al. EACL 2009 (COMPOST tagger)
    - http://ufal.mff.cuni.cz/compost (Czech, English & other)
  - No p-level (of course ☺)

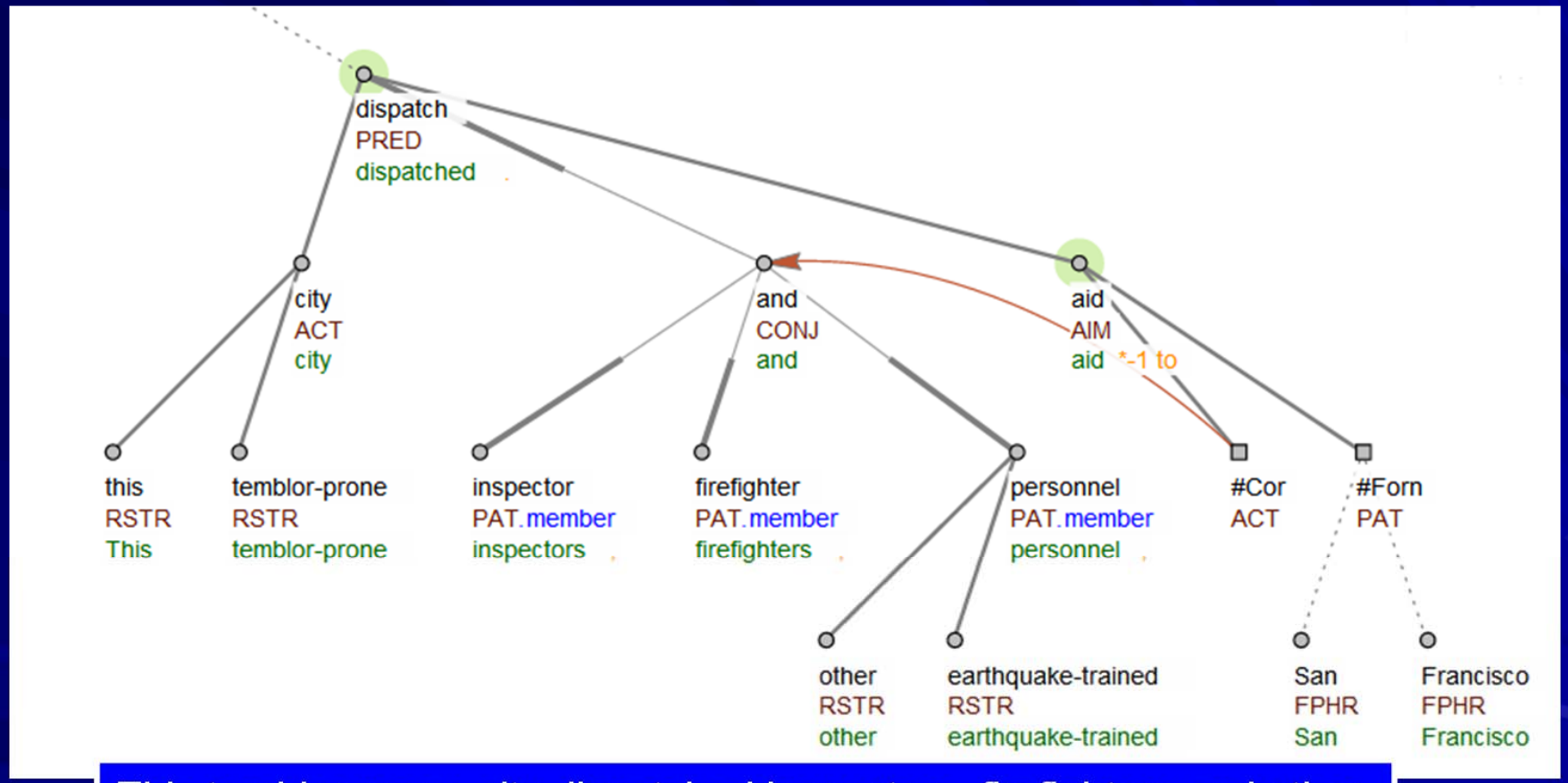


- PDT style, no change
- Syntax: automatic, 2000 sentences fully manual for testing
- Lemmatization and tagging: auto
  - 99%/96%, Spoustová et al. EACL 2009 (COMPOST tagger)
  - http://ufal.mff.cuni.cz/compost (Czech, English & other)
- No p-level (of course ☺)

Tectogrammatical annotation

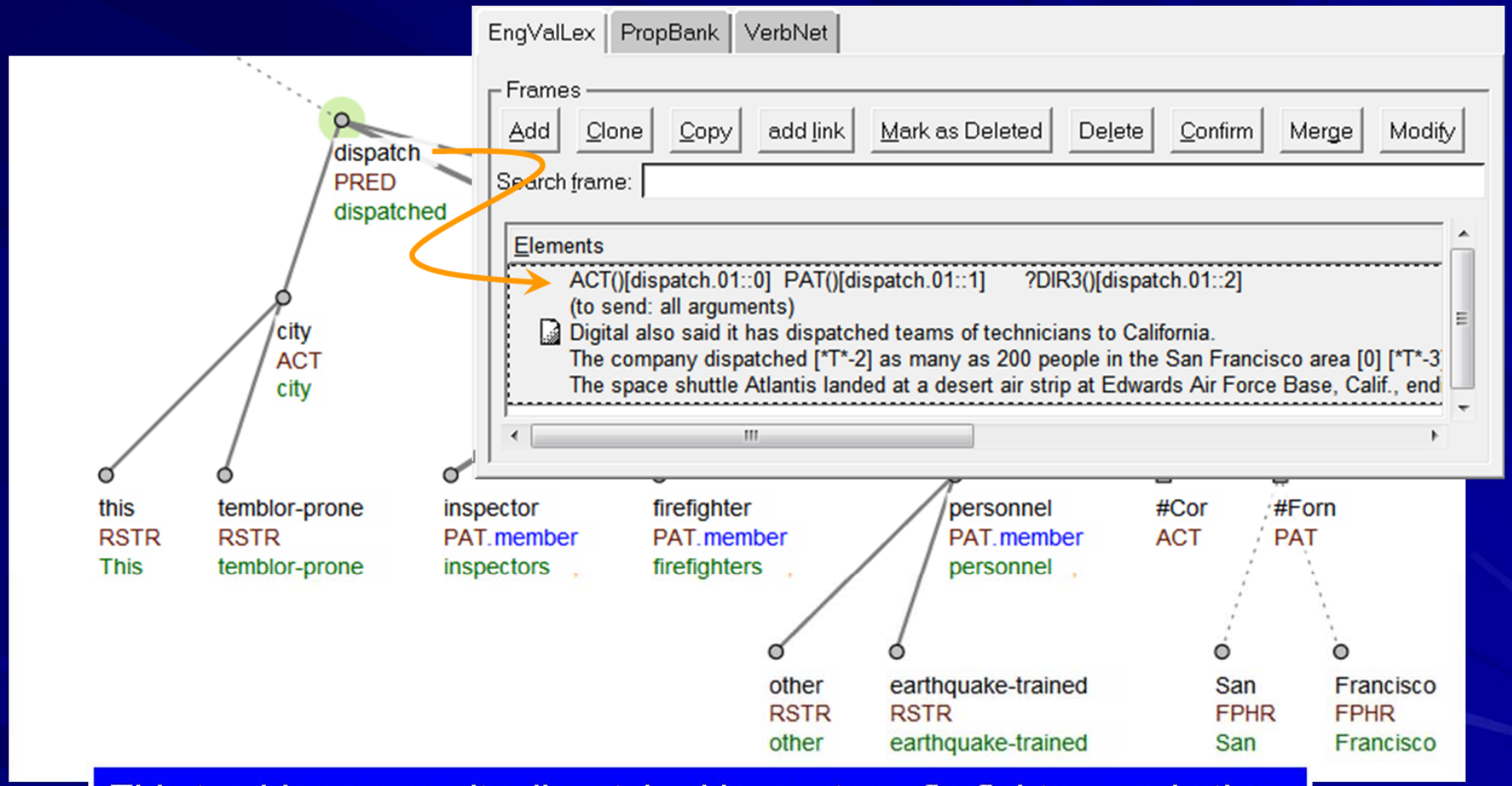
- Manual (both languages)
- Major features
  - Nodes with „autosemantic“ words only (no function words)
    - Ellipsis „restored“ (new node for verbal arguments)
  - (Semantic) function (dependent → head relation)
    - Verb arguments + ca 50 functions for other relations
  - Valency lexicons attached (Eng: links to PropBank)
  - “Formemes”: prep+case style label (useful in MT and search)
  - Co-reference integrated (Eng: BBN + more), Czech: manually
- Alignment
  - To surface syntax & between Czech and English

Tectogrammatical annotation



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

Tectogrammatical annotation



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

Accompanying Tools

- TrEd (http://ufal.mff.cuni.cz/tred)
  - Annotation, View/Browse and Search environment
  - Open source, perl
  - Search and visualization: PML-TQ
    - Powerfull query language for complex NLP annotations, 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

FAQ

(with 'take-home' answers)

Acknowledgements:

- What is different from PCEDT 1.0 (2004)?
  - Size: full PTB & translation (1 mil.) vs. ~400k
  - Full tecto-manual annotation both Czech and English
  - Integrated co-reference, named entities
  - Valency lexicons on both sides, links to PropBank, VerbNet
  - "Formemes": prep+case label
  - English lemmatization (m-level, t-level)
- Next?
  - Manual "native" annotation of surface dependencies
  - Extended semantic annotation
    - modalities, tense, number, etc.
  - Aligned valency lexicon entries (Cze ↔ Eng)
  - Remote access and „EZ" search for data and lexicons
    - http://ufal.mff.cuni.cz/lindat/PDT-Vallex.html
    - http://ufal.mff.cuni.cz/lindat/EngVallex.html (now in beta)

FAQ

(with 'take-home' answers)

Acknowledgements:
Charles University research funds
("PRVOUK")

- What is different from PCEDT 1.0 (2004)?
  - Size: full PTB & translation (1 mil.) vs. ~400k
  - Full tecto-manual annotation both Czech and English
  - Integrated co-reference, named entities
  - Valency lexicons on both sides, links to PropBank, VerbNet
  - "Formemes": prep+case label
  - English lemmatization (m-level, t-level)
- Next?
  - Manual "native" annotation of surface dependencies
  - Extended semantic annotation
    - modalities, tense, number, etc.
  - Aligned valency lexicon entries (Cze ↔ Eng)
  - Remote access and „EZ" search for data and lexicons
    - http://ufal.mff.cuni.cz/lindat/PDT-Vallex.html
    - http://ufal.mff.cuni.cz/lindat/EngVallex.html (now in beta)

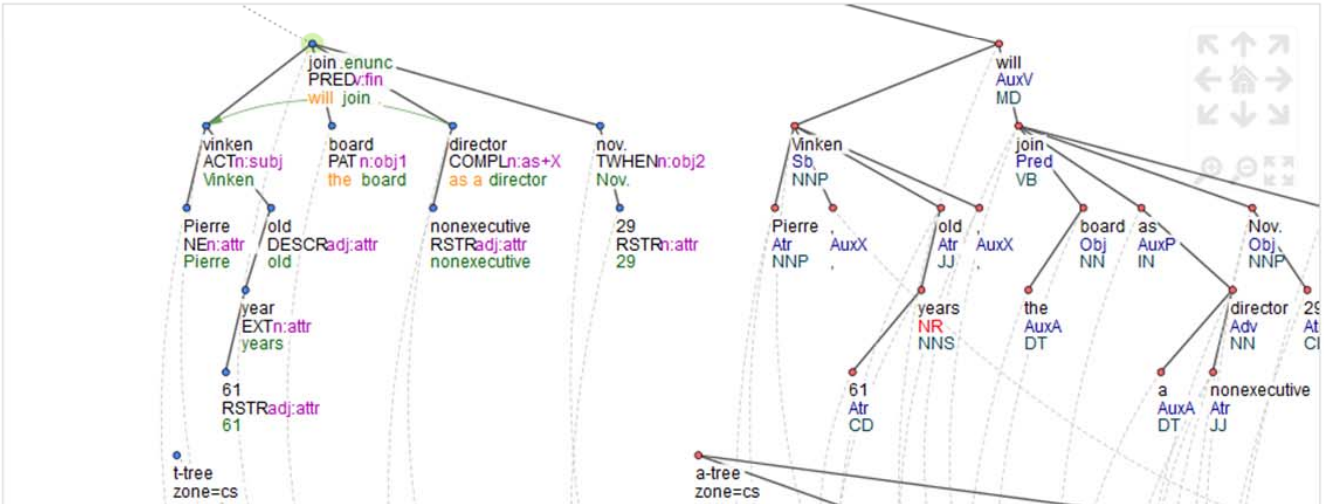
Explore & enjoy!

Introduction | Data | Tools | Documentation | Publications | Distribution & Licence | Installation | Credits | Acknowledgements

This is only a small sample of the corpus. You need to **order and properly license** the corpus to browse it in its entirety.

Section: 00 File: 01 Sentence: 1

[en] Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
[cs] Jednašedesátiletý Pierre Vinken se připojí ke správní radě jako nevýkonný ředitel dne 29. listopadu.



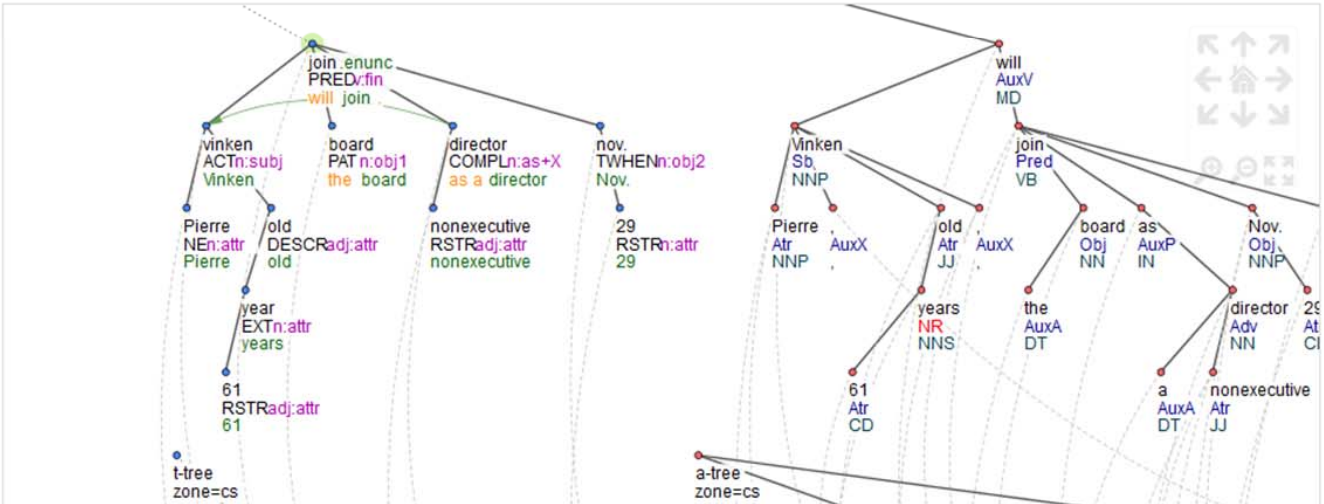
Explore & enjoy!

Introduction | Data | Tools | Documentation | Publications | Distribution & Licence | Installation | Credits | Acknowledgements

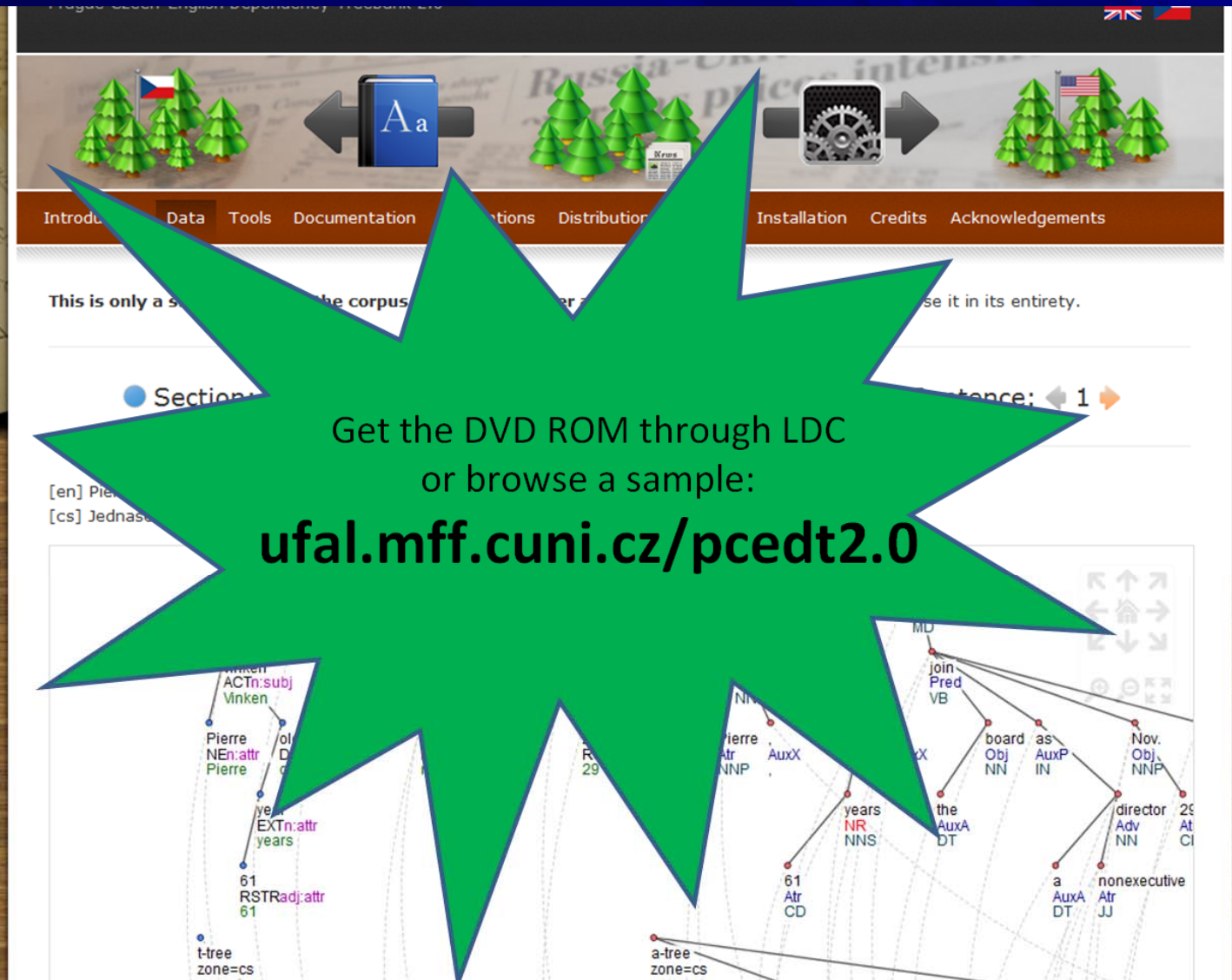
This is only a small sample of the corpus. You need to **order and properly license** the corpus to browse it in its entirety.

Section: 00 File: 01 Sentence: 1

[en] Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
[cs] Jednašedesátiletý Pierre Vinken se připojí ke správní radě jako nevýkonný ředitel dne 29. listopadu.



Explore & enjoy!



Get the DVD ROM through LDC
or browse a sample:

ufal.mff.cuni.cz/pcedt2.0