Tectogrammatical Representation of English

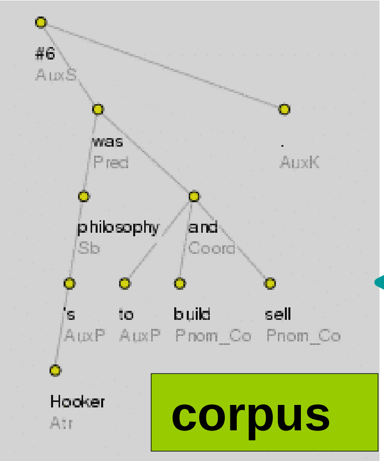
Silvie Cinková

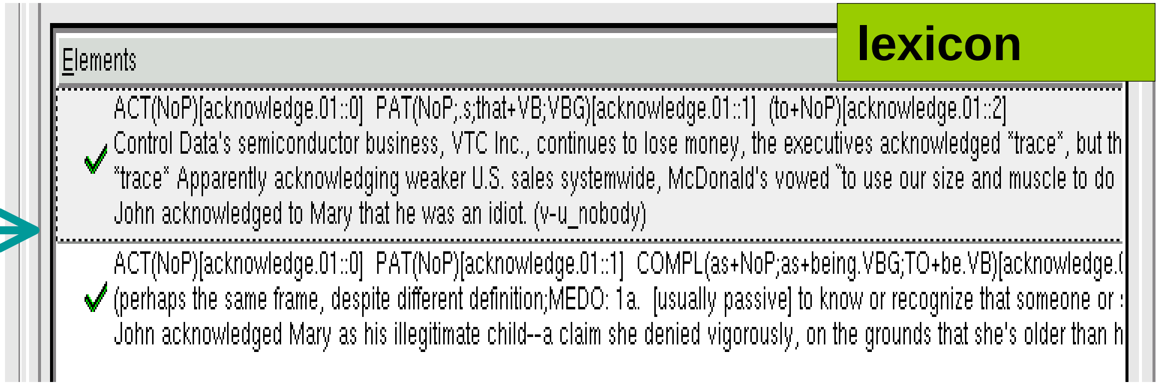
Lucie Mladová, Anja Nedoluzhko, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeněk Žabokrtský

Linguistic Annotation

- corpora + lexicons as gold-standard data for Machine Learning

- **Functional Generative Description (FGD):**
  - morphology (m-layer)
  - surface syntax (a-layer)
  - **deep syntax/semantics (t-layer)**



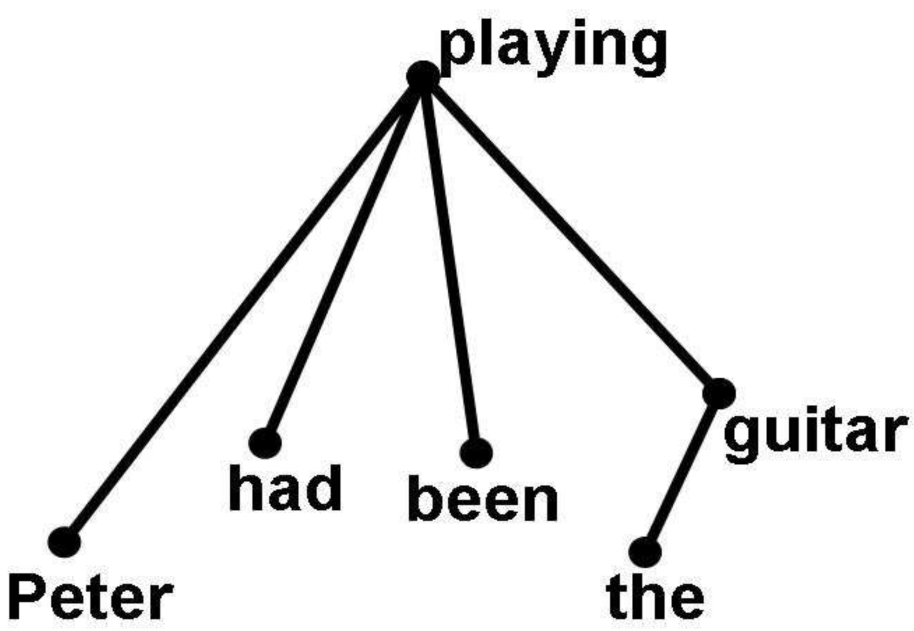


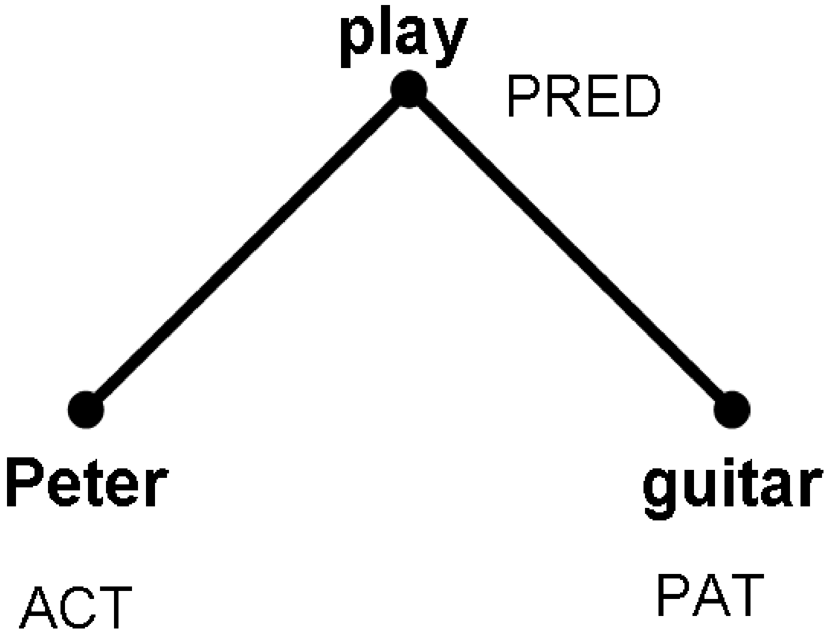
✓ ACT(NoP)[acknowledge.01::0] PAT(NoP;.s;that+VB;VBG)[acknowledge.01::1] (to+NoP)[acknowledge.01::2]
Control Data's semiconductor business, VTC Inc., continues to lose money, the executives acknowledged "trace", but th
"trace" Apparently acknowledging weaker U.S. sales systemwide, McDonald's vowed "to use our size and muscle to do
John acknowledged to Mary that he was an idiot. (v-u_nobody)

✓ ACT(NoP)[acknowledge.01::0] PAT(NoP)[acknowledge.01::1] COMPL(as+NoP;as+being.VBG;TO+be.VB)[acknowledge.(
(perhaps the same frame, despite different definition;MEDO: 1a. [usually passive] to know or recognize that someone or :
John acknowledged Mary as his illegitimate child--a claim she denied vigorously, on the grounds that she's older than h

a- vs. *t-layer* Representations

Peter had been playing the guitar.

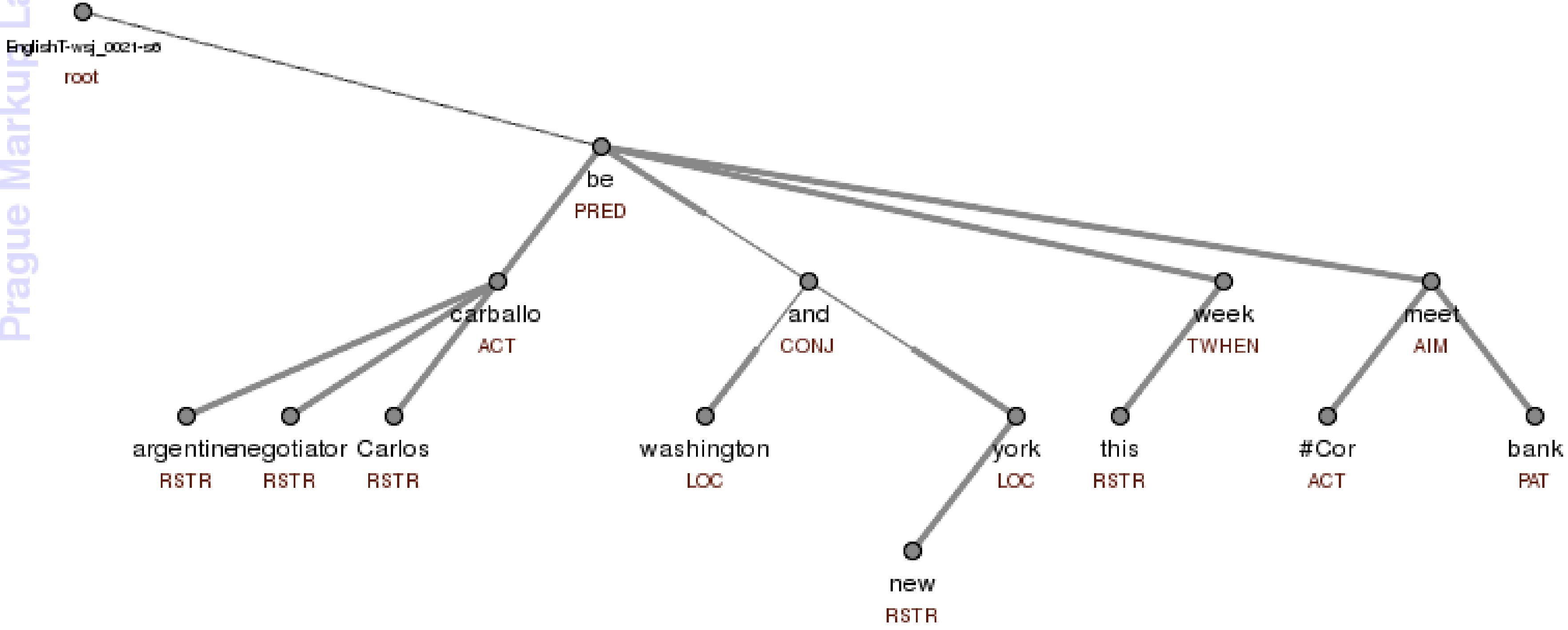




What TR captures

- syntactic dependency (autosemantic words)
- semantic relations
- links to the lower layers (a-, m-)
- valency
- coreference (grammatical and textual)
- topic-focus articulation

“Real” TR - Example



Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks.

FGD-Compliant Language Resources

- **Prague Dependency Treebank (PDT)**
  - Czech texts
    - m-layer: 2 million words
    - a-layer: 1.5 million words
    - t-layer: 0.8 million words
- **PDT-Vallex**
  - valency lexicon interlinked with the data (verbs, nouns, adjectives)
- **Prague Czech-English Dependency Treebank 1.0 (PCEDT 1.0)**
  - English-Czech parallel corpus of approx. 22k sentences, automatic a- and t-layer annotation

Our Project Goal: Prague Czech-English Dependency Treebank 2.0

An approx. 50 000-sentence parallel corpus

- **English:** PennTreebank - Wall Street Journal
  - automatic conversion into the PDT-like shape (*a* layer)
  - automatic *t*-layer procession, manual annotation
- **Czech:**
  - PTB-WSJ translated into Czech
  - automatic *a*- and *t*- layer procession
  - manual *t*-layer annotation

English TR Annotation Stage I

ANNOTATING

- dependencies
- links to a-layer
- t-lemmas
- functors
- valency frames for verbs

POSTPONED

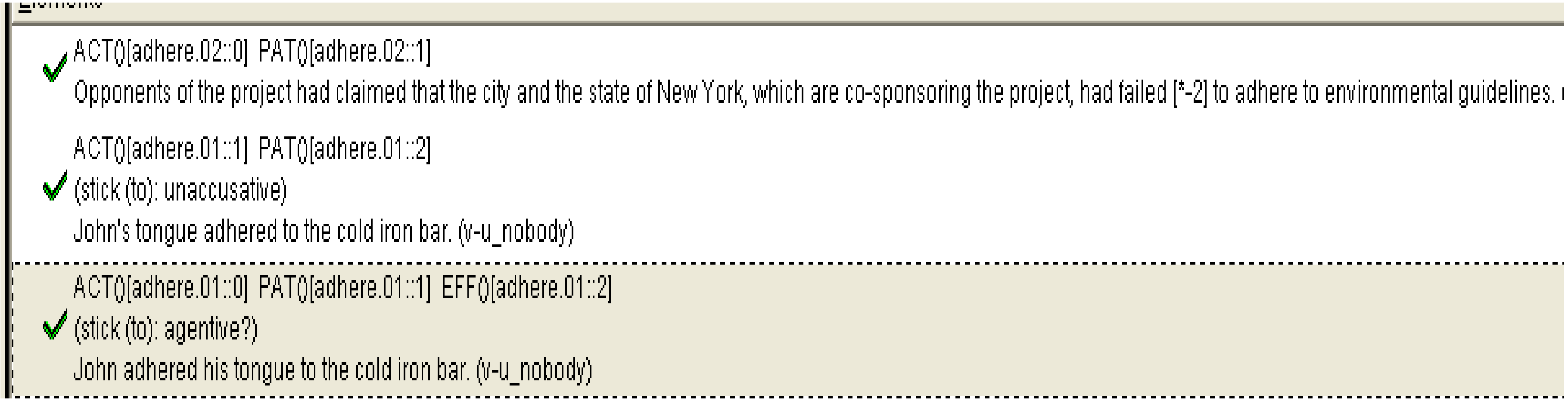
- valency of other POS
- coreference
- TFA
- subfunctors
- grammatemes

Goals for 2006 (PIRE, January 06)

- Conversion of the PropBank-Lexicon (EngValLex)
- Preparation of PTB – WSJ for extensive annotation (building on PCEDT).
- Annotation of the PDT-compatible version of PTB-WSJ (PEDT)

EngValLex

- Automatic conversion ✓ ✓ ✓
- Manual corrections ✓ ✓
  - continuous proofreading during the annotation
- Morphosyntactic representations ✓
  - formal description designed, test annotation of 400 rolesets



✓ ACT()[adhere.02::0] PAT()[adhere.02::1]
Opponents of the project had claimed that the city and the state of New York, which are co-sponsoring the project, had failed [*-2] to adhere to environmental guidelines.

ACT()[adhere.01::1] PAT()[adhere.01::2]
✓ (stick (to): unaccusative)
John's tongue adhered to the cold iron bar. (v-u_nobody)

ACT()[adhere.01::0] PAT()[adhere.01::1] EFF()[adhere.01::2]
✓ (stick (to): agentive?)
John adhered his tongue to the cold iron bar. (v-u_nobody)

Corpus Preparation

- TrEd adjustment
  - new macros, scheme alterations
  - original phrase-structure tree view
  - compliance with TrEd settings for Czech annotation
- adjustment of t-layer conversion
  - node hiding (auxiliaries, particles...)
  - functor assignment
- subversion configuration
  - reliable data storage
  - no uncontrolled annotation overlap

Annotation

- **approx. 1000 trees annotated** (the first 3 months of “routine” annotation)
- **interannotator-agreement watching**
  - monthly, approx. 30 trees
  - November 2006:
    - functor agreement: 77,2% - 79,1%
    - parent node agreement: 83,7% - 87,6%
- **identifying recurrent structures, unifying the annotation**
  - “established” phenomena (e.g. tree structure with raised objects, nominal *-ing* clauses, existential *there, dummy-do* etc.)
  - text phenomena to be resolved arbitrarily but consistently (e.g. tree structure with *10:50 a.m. EST*)

Next To-Do’s

- Annotation manual in English (draft): 2006
- EngValLex correction (without morphosyntactic representation description)
- Increasing the annotation speed to 500 trees/annotator/month
- Increasing (or at least keeping) the interannotator agreement

Future Work (end 2008)

- Converting NomBank into EngValLex (valency of nouns)
- Manual annotation of the PCEDT 1.0 data (22 000 sentences)
- Annotation manual updated, refined
- Morphosyntactic form description in EngValLex