Prague Dependency Treebank
PDT (since 1998, now PDT 2.0)

Eva Hajičová
Institute of Formal and Applied Linguistics
Charles university in Prague

(http://ufal.mff.cuni.cz/pdt2.0)

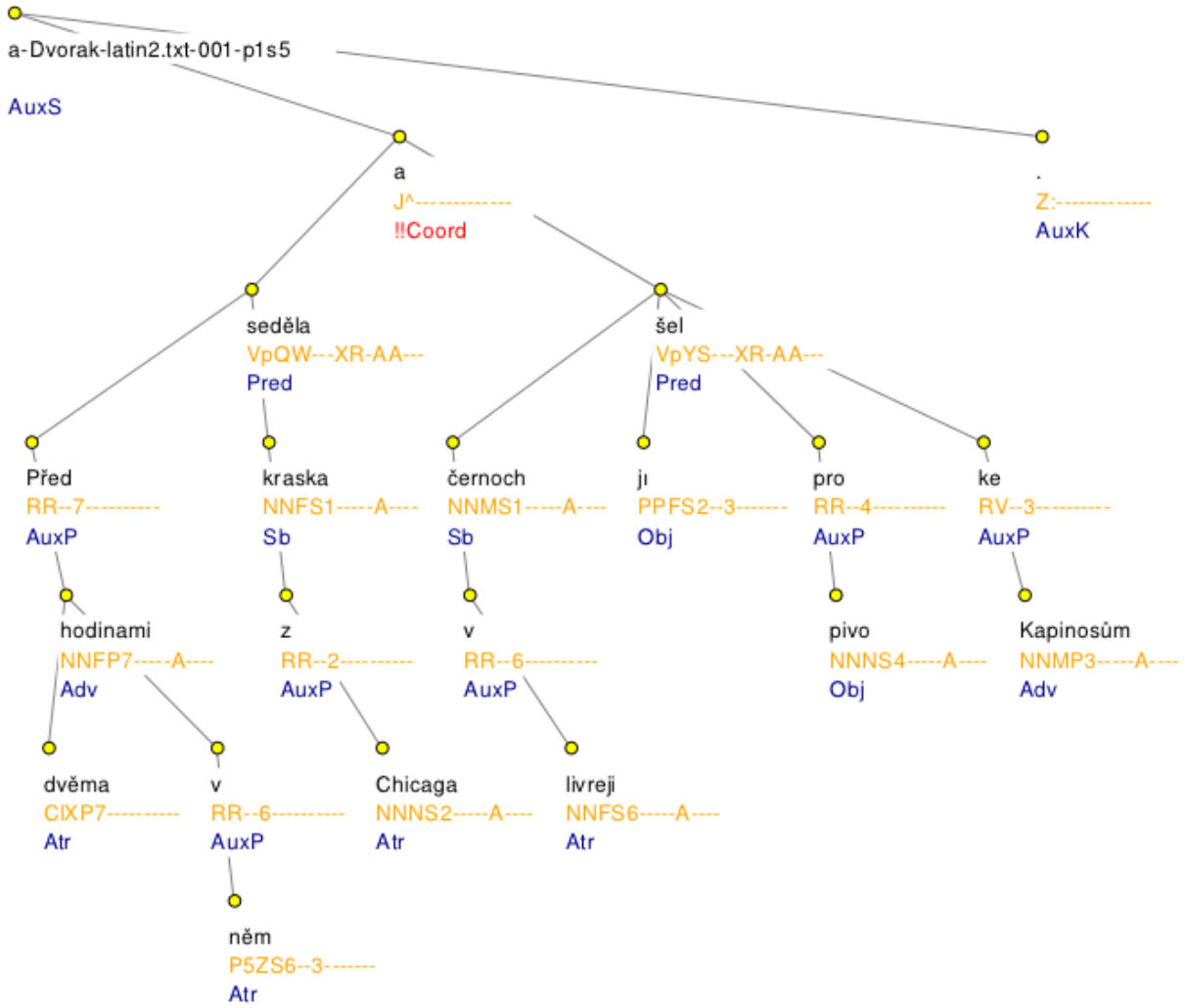
Annotation scheme of PDT

- Three layers:
  - Morphological (>100.000 sentences)
  - Surface syntax – dependency (80.000 sentences)
  - Tectogrammatical (underlying, dependency)
    - incl. information structure (topic/focus) – 50.000 sentences
- Additional annotation in PDT 2.5 (2011)
  - multiword expressions, clause boundaries
  - some additional morphological and syntactic features
- Partially completed (for PDT 2.6, 2012)
  - Coreference relations (nominal, incl. bridging)
  - Basic discourse relations

Example

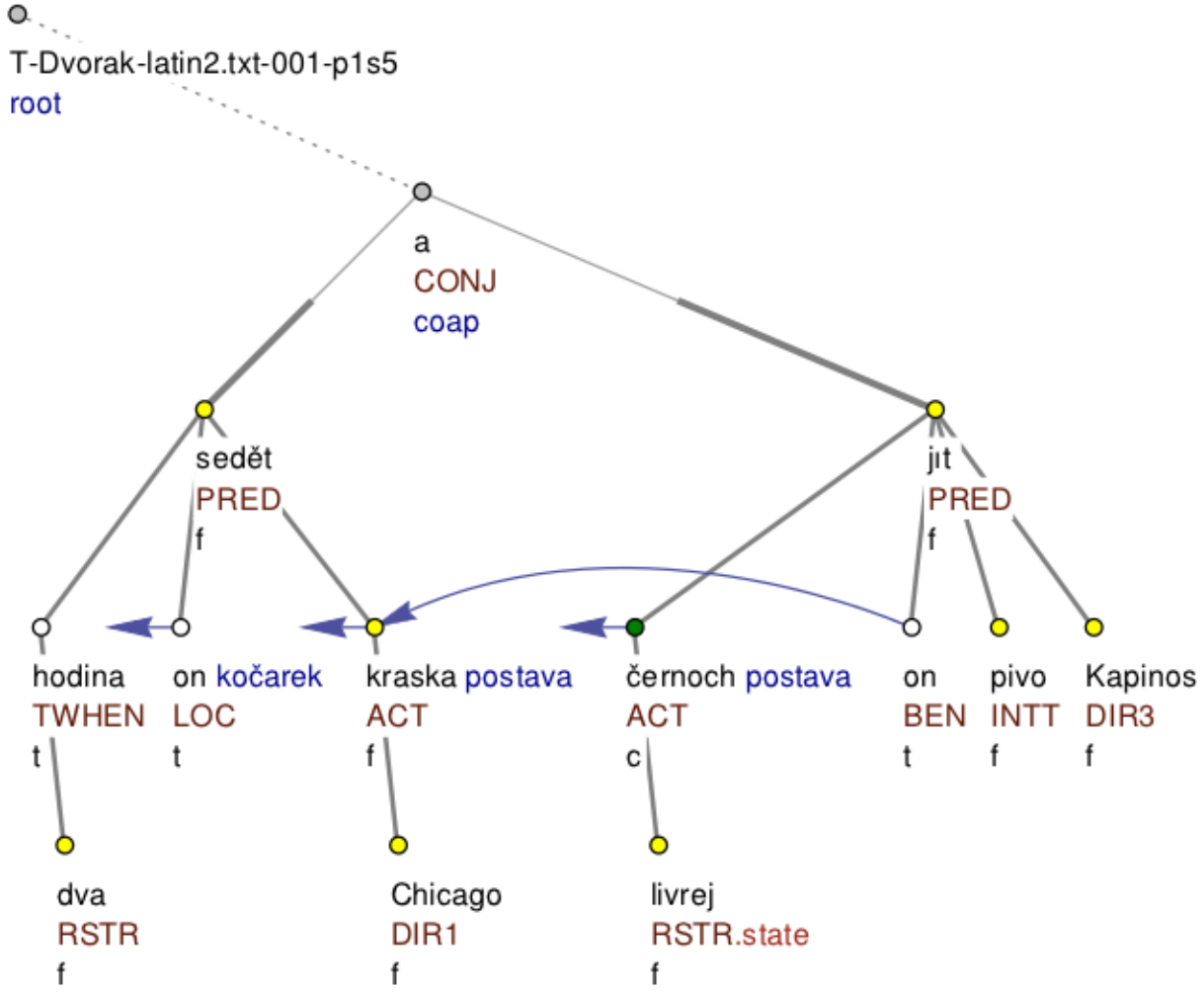
- Před dvěma hodinami v něm seděla kráska z Chicaga a černoch v livreji ji šel pro pivo ke Kapinosům.
- *Before two hours in it was-sitting beauty from Chicago and black-man in suit for-her went for beer to Capinose*
- Two hours ago a beauty from Chicago was sitting in it and a black man in a suit went for beer for her to Capinose.

Morphological and syntactic layers



Před dvěma hodinami v něm seděla kraska z Chicaga a černoch v livreji jı šel pro pivo ke Kapinosům.

Tectogrammatical layer



Před dvěma hodinami v něm seděla kraska z Chicaga a černoch v livreji jı šel pro pivo ke Kapinosům.

Prague Dependency Treebank, PDT

(i) Successful application of syntactic and semantic features:

- Dependency syntax of the tectogrammatical layer: (underlying) syntactic relations (functors): useful

BUT:

- Not fully exploited
- sparse data →
  - lower results in evaluation of tools (semantic role labeling, deep parsing, coreference)

Prague Dependency Treebank, PDT

(ii) What applications held back?

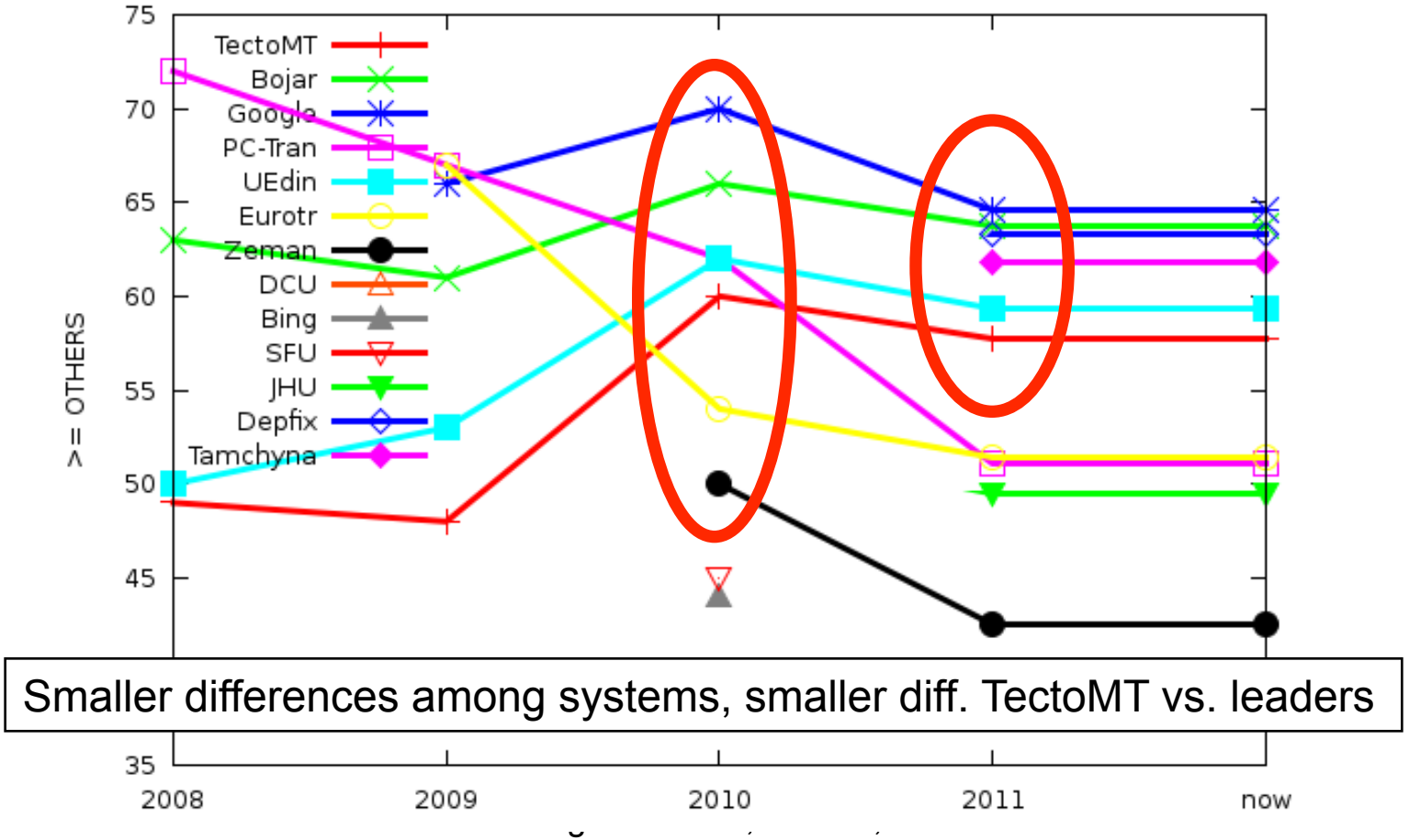
- Machine translation via the tectogrammatical layer

BUT

- recent results optimistic (Zabokrtsky, Popel, et al., WMT at EMNLP, Edinburgh, July 2011) – gradual (substantial) improvement since 2008

WMT 2011 results (preliminary)

Subjective evaluation, English -> Czech



Smaller differences among systems, smaller diff. TectoMT vs. leaders

Prague Dependency Treebank, PDT

(iii) What other kinds of tasks if richer data sources available:

- Prague Dependency Treebank, Parallel Prague Czech-English Dep. Treebank:

  „rich“ data already available

- BUT:
  - not large enough (and partly not rich enough) for tasks such as
    - Information Extraction, IR, summarization, sentiment detection, machine translation etc.