Deep Linguistic Information in Hybrid Machine Translation

Jan Hajič

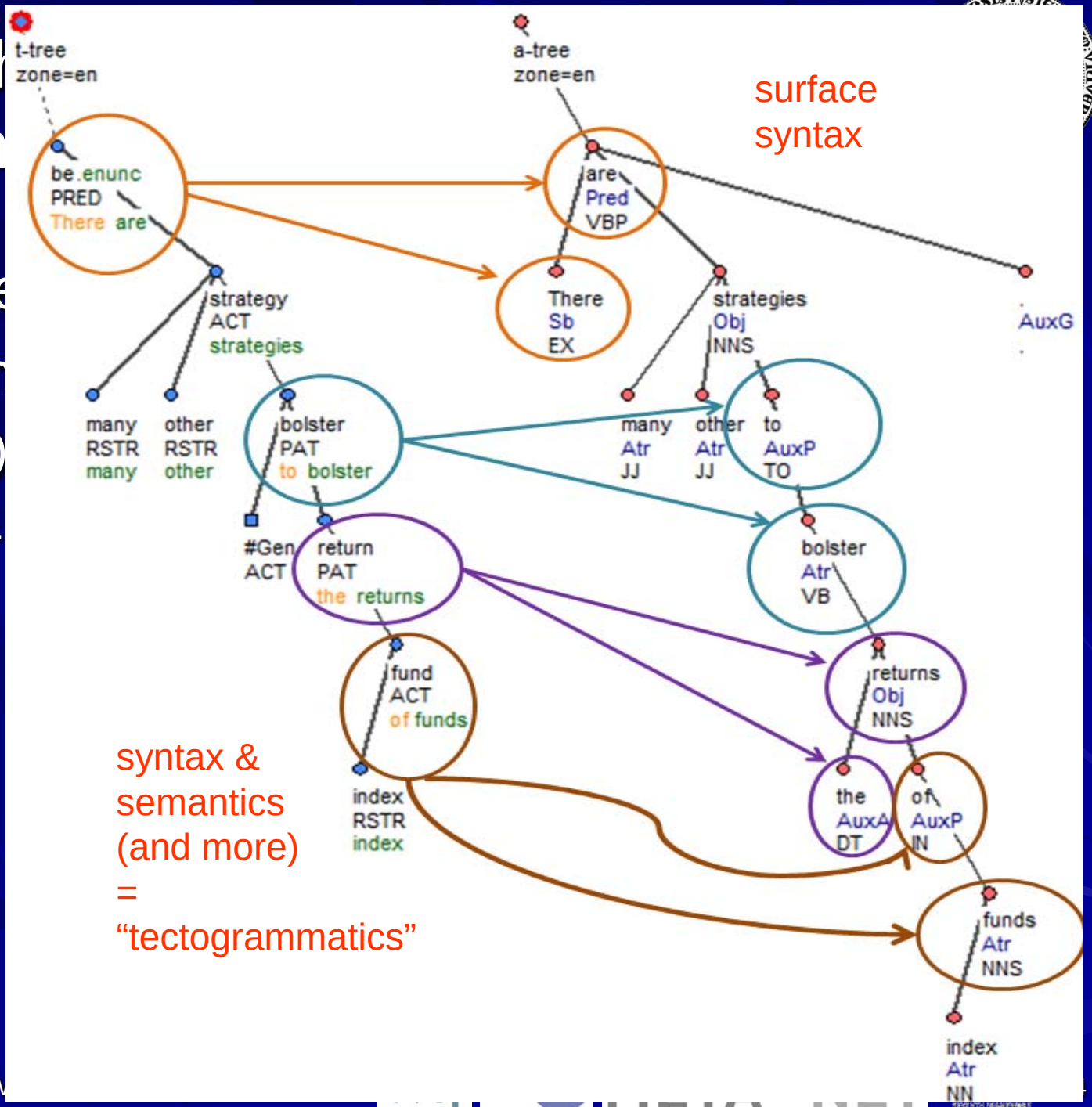
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic

Outline: From Data To an MT System

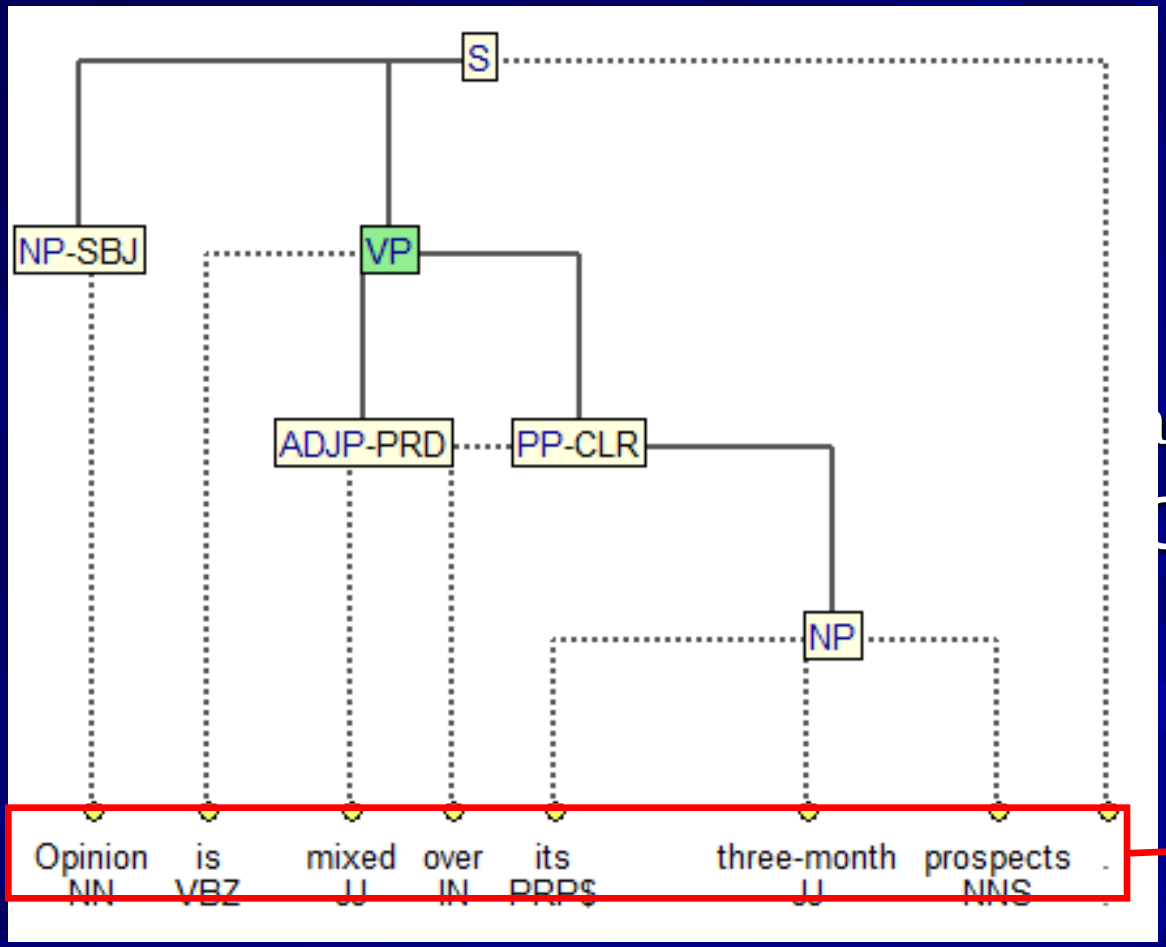
- "DeepBank:" The Prague Czech-English Dependency Treebank (2.0)
  - Texts, annotation style(s), alignment, tools
- The platform: Treex
- TectoMT: hybrid MT English → Czech
  - The (old) idea
  - Overall design
  - Core modules
- (A Speculation on) The Future

Th
Depen

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



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

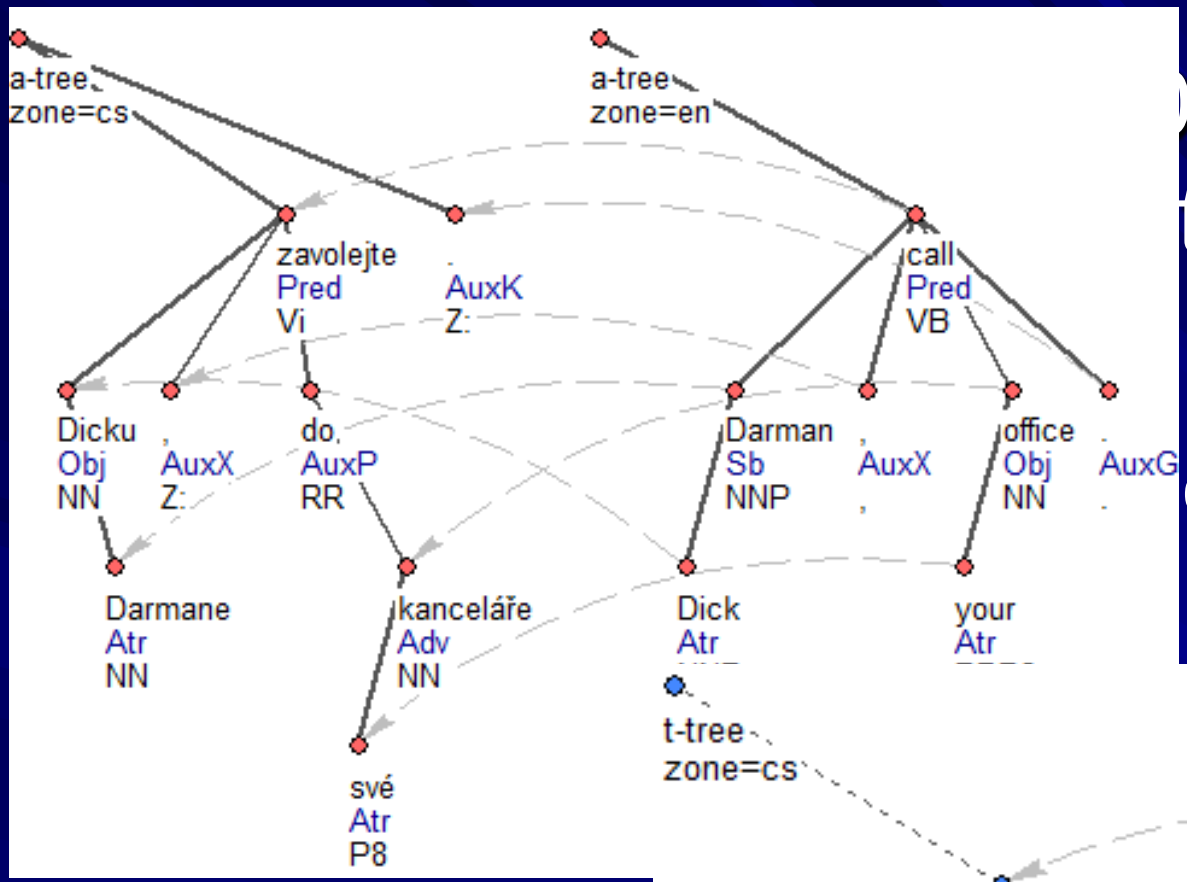
	Czech	English
Sentences	49,208	
a-nodes (automatic)	1,151,150	1,173,766
t-nodes (manual)	931,846	838,212

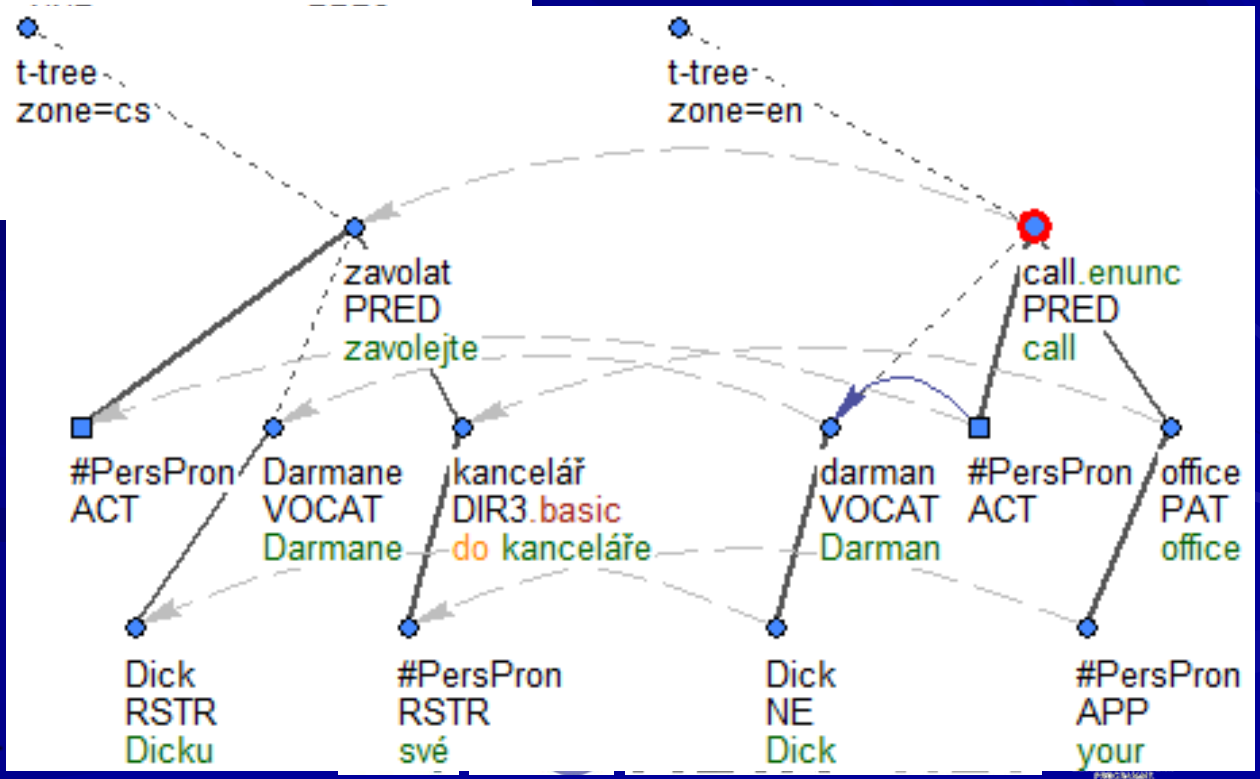
- Pub 8)
  - A TA-
    S

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

t(s)

due to translation)





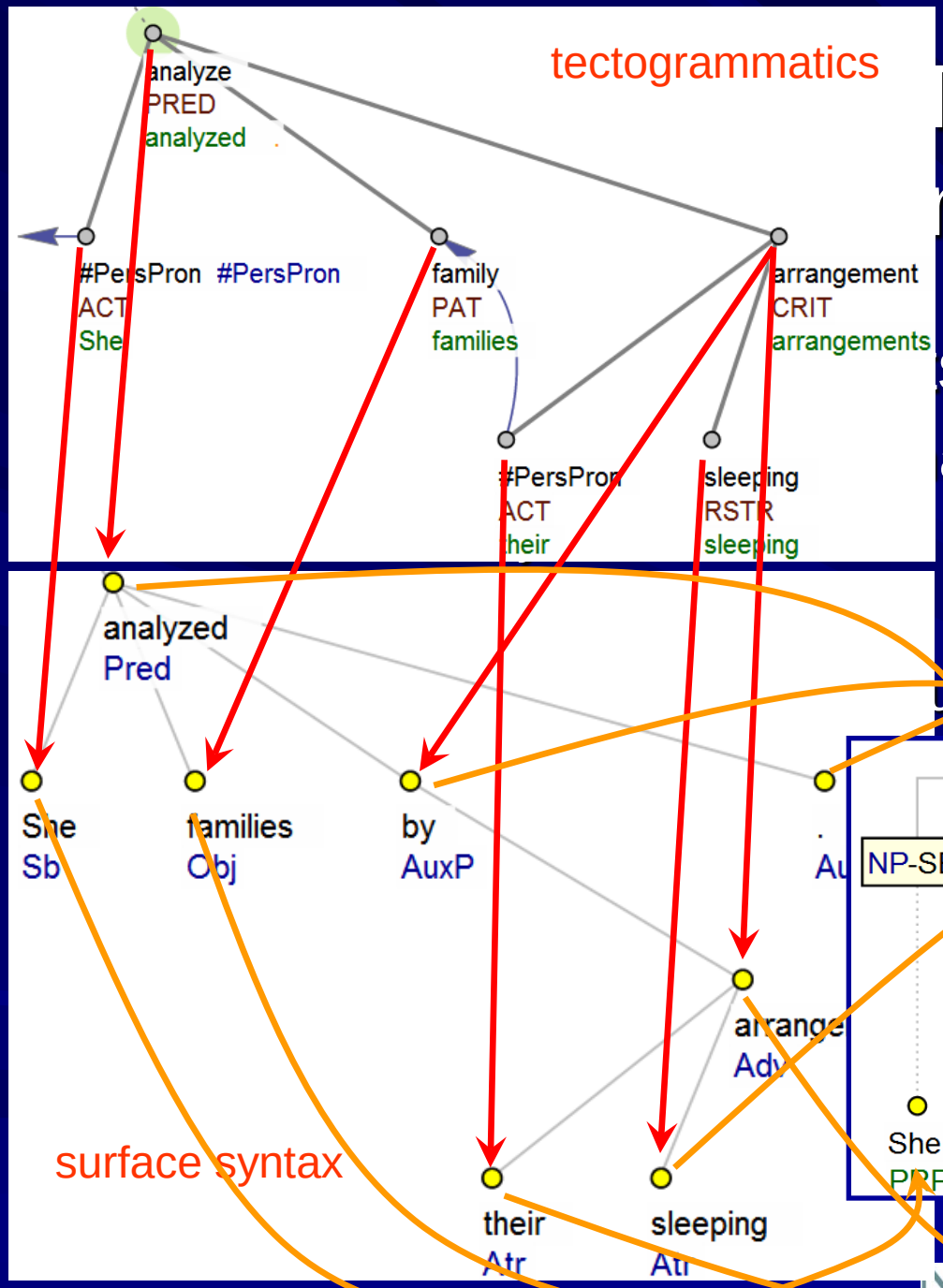
T 2.0
ment(s)

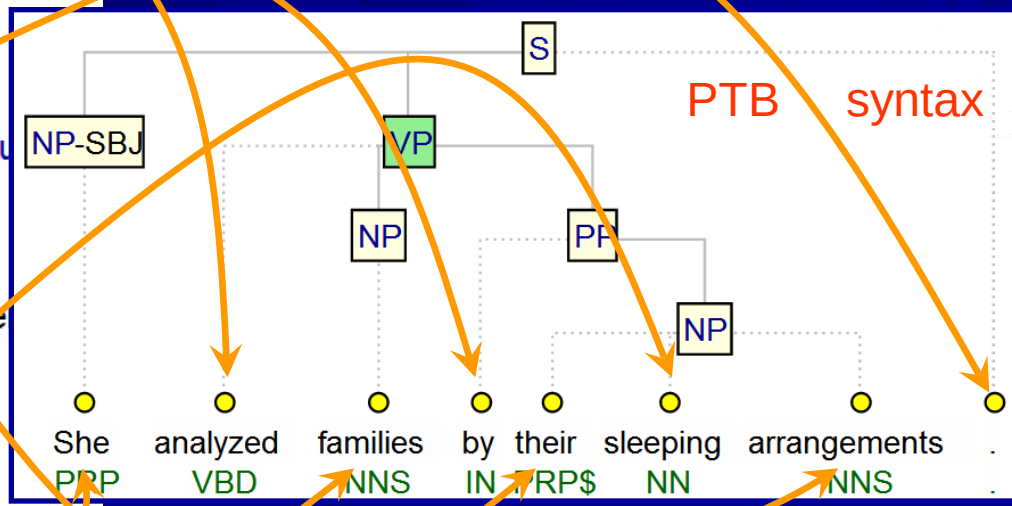
s

atural due to translation)

1

ally corrected (in part), m → n

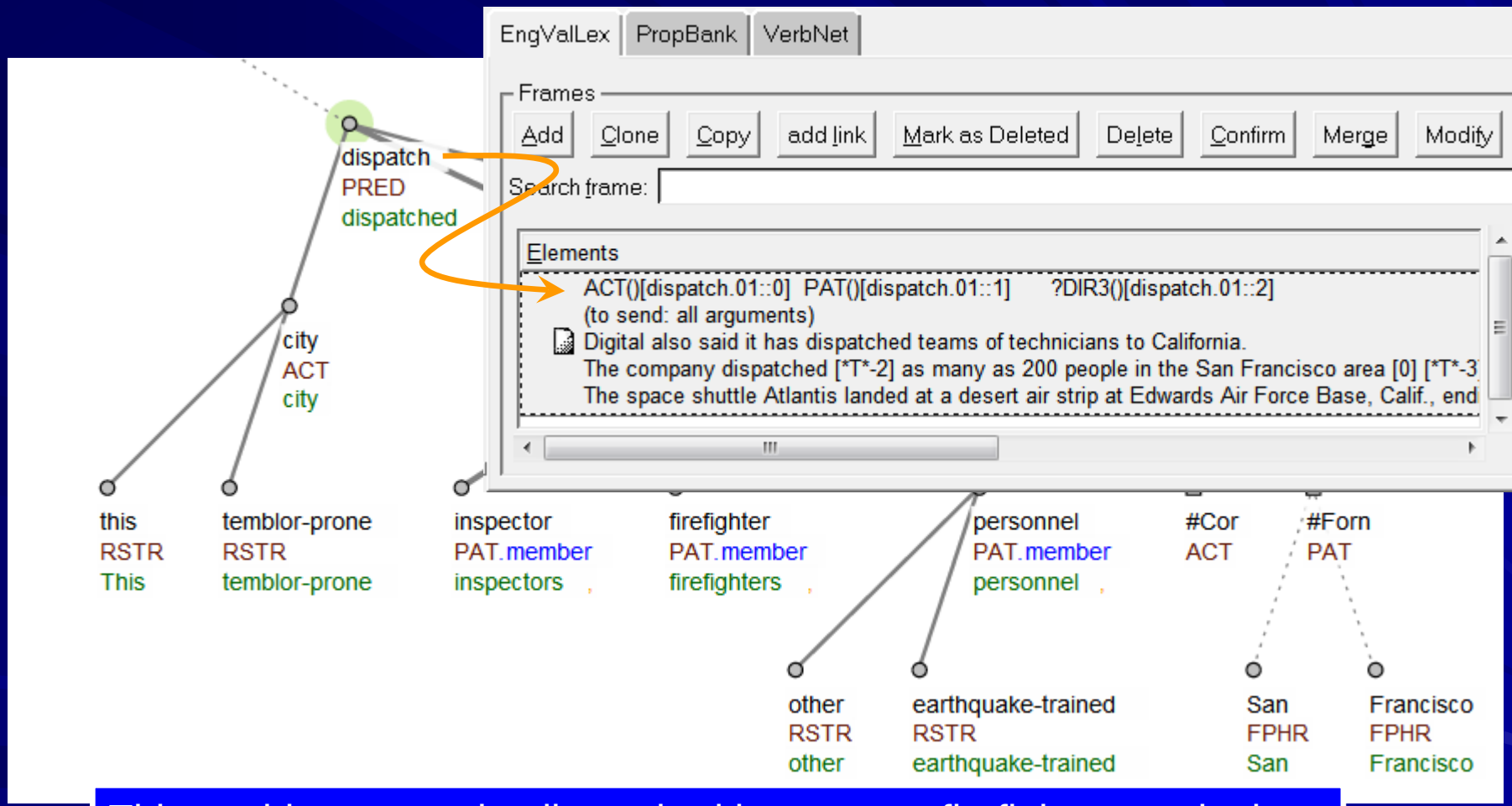




Surface syntax annotation

- 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 (MST); 2000 sent. 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 ☺)

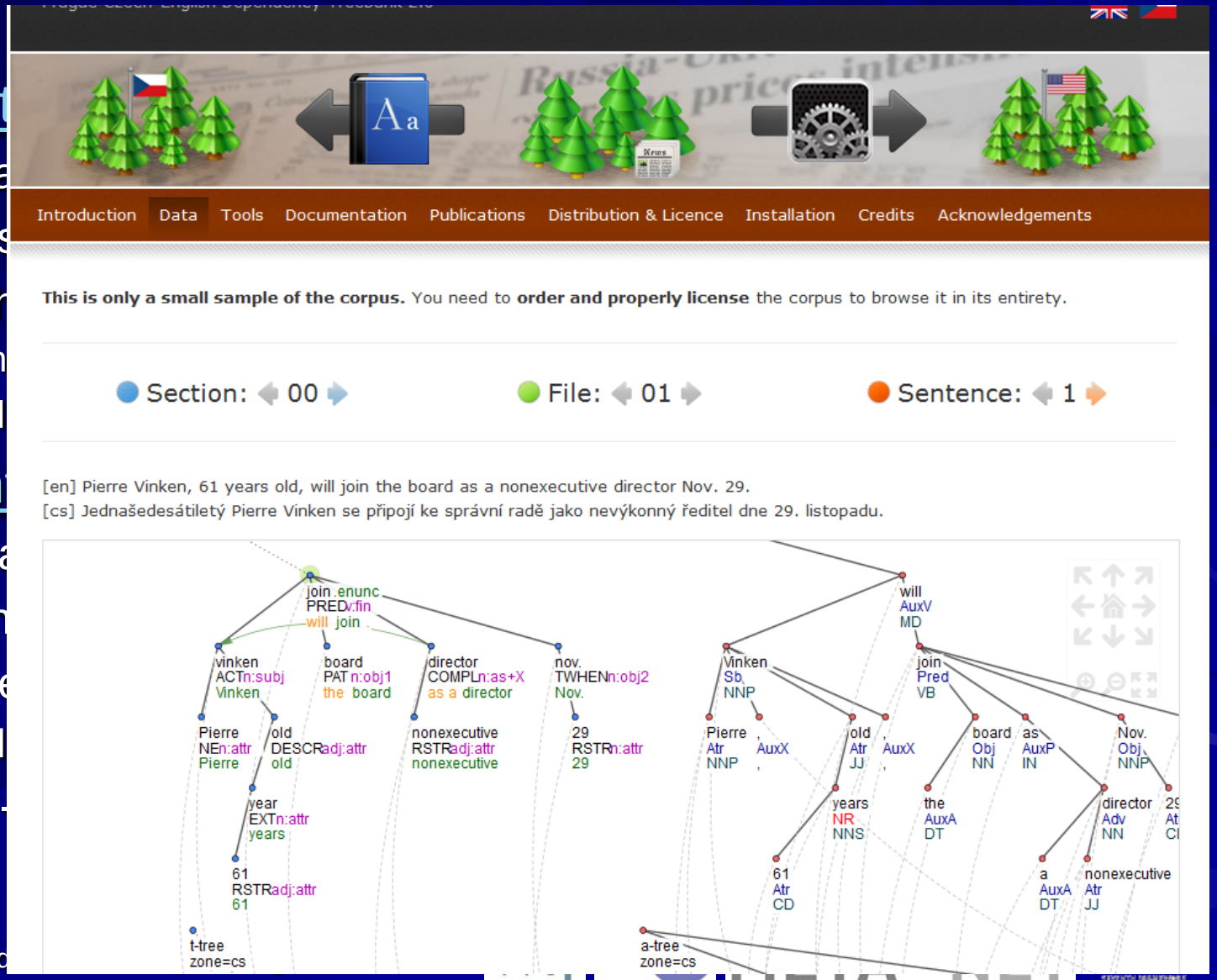
Tectogrammatical annotation



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

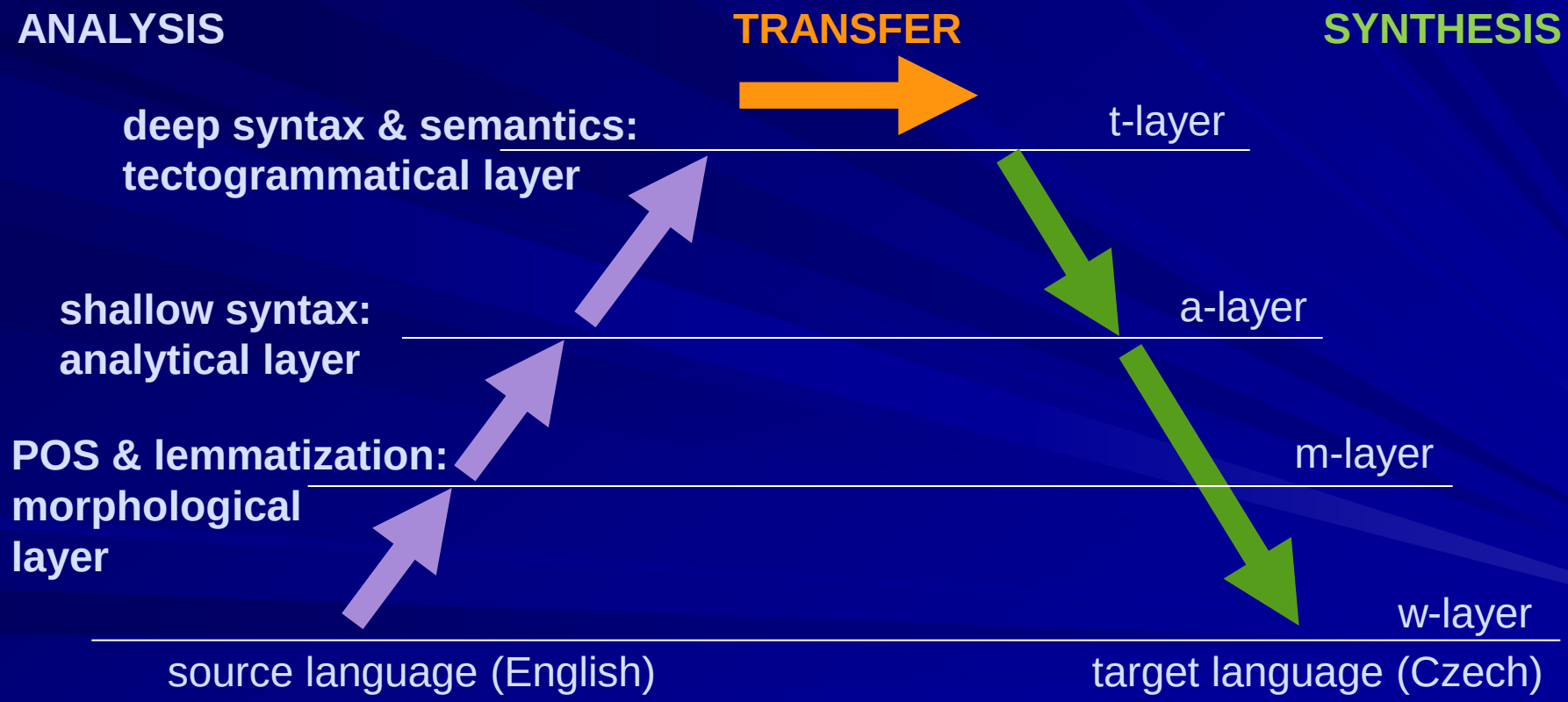
Accompanying Tools

- TrEd (htt
  - Annota
  - Open s
  - Search
    - Sim
    - PM
- Treex (h
  - Modul
  - Easy h
  - Module
    - incl
  - CPAN-



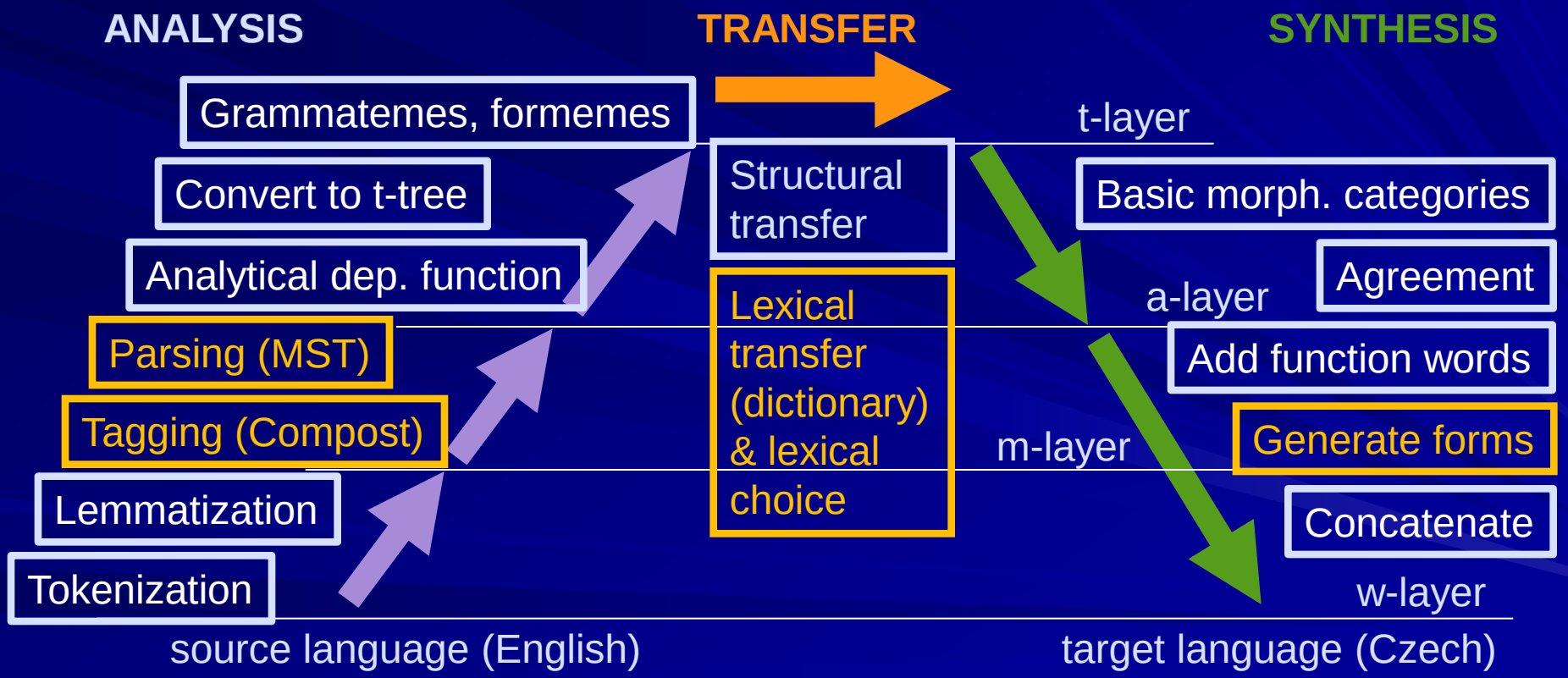
PCEDT and Tectogrammatics in (hybrid) MT

- The famous, (almost) “Vauquois” triangle:



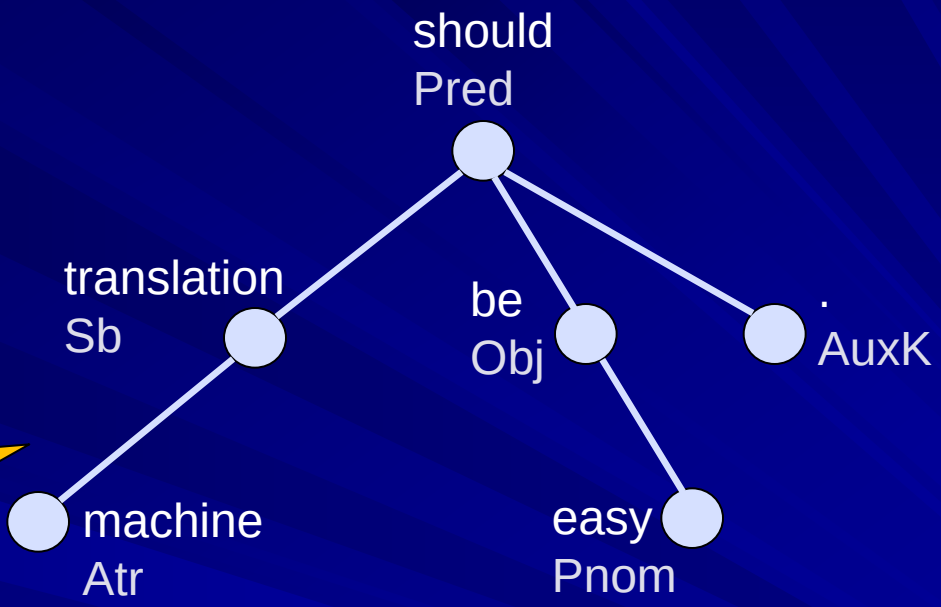
Analysis-Transfer-Synthesis Hybrid System

- Over 90 steps: both rule-based and statistical



Example Translation

a-layer (parse) + functions



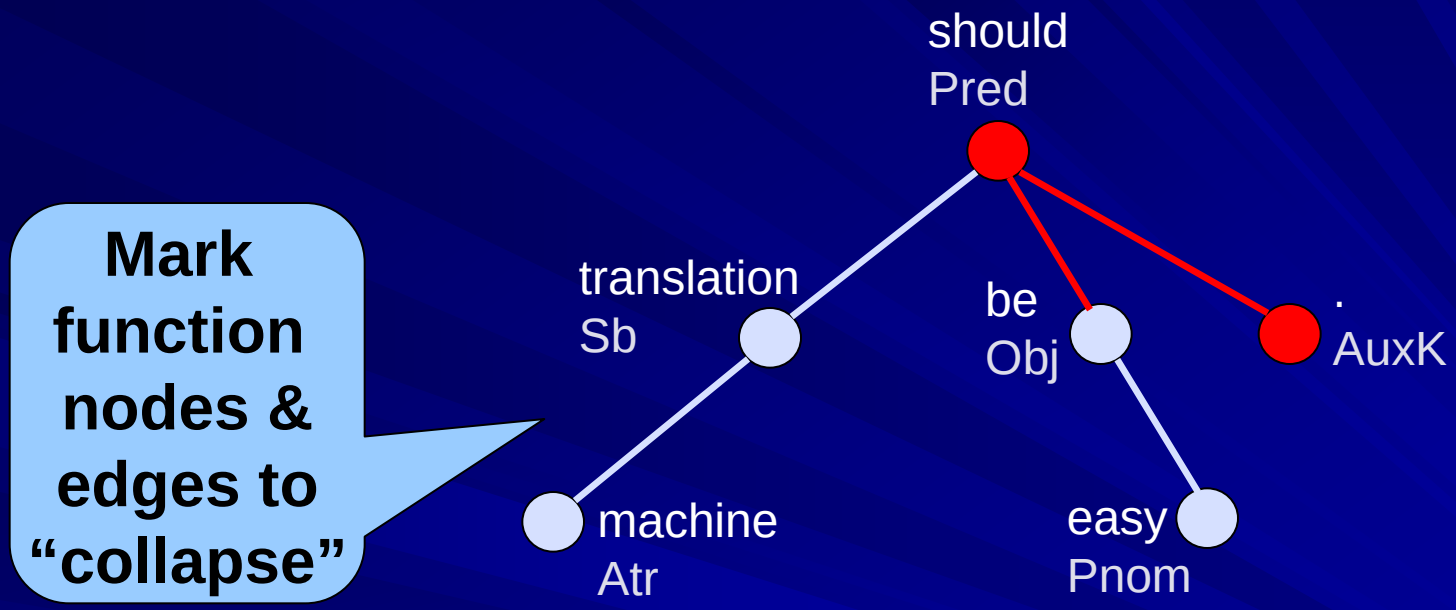
Lemmatized & POS tagged

machine translation should be easy .
NN NN MD VB JJ .

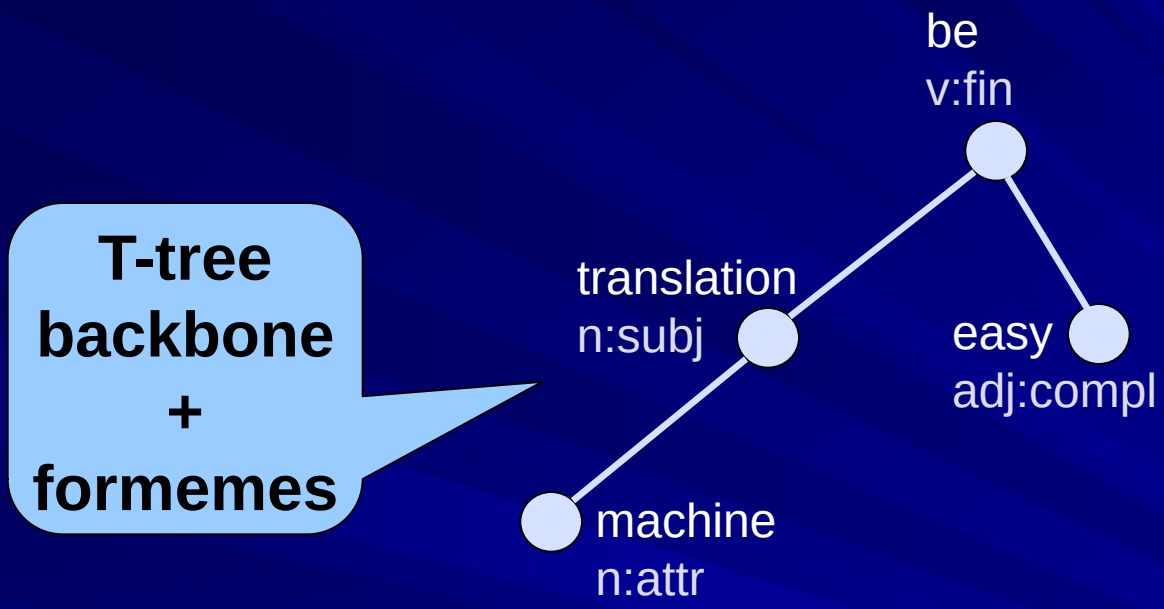
Tokenized

Machine translation should be easy .

Example Translation

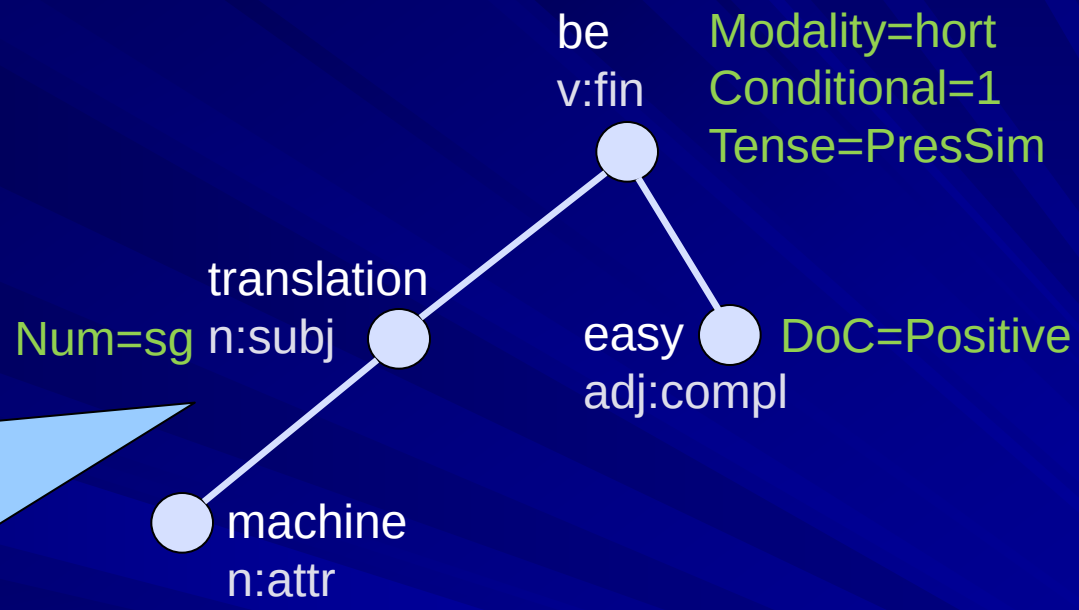


Example Translation

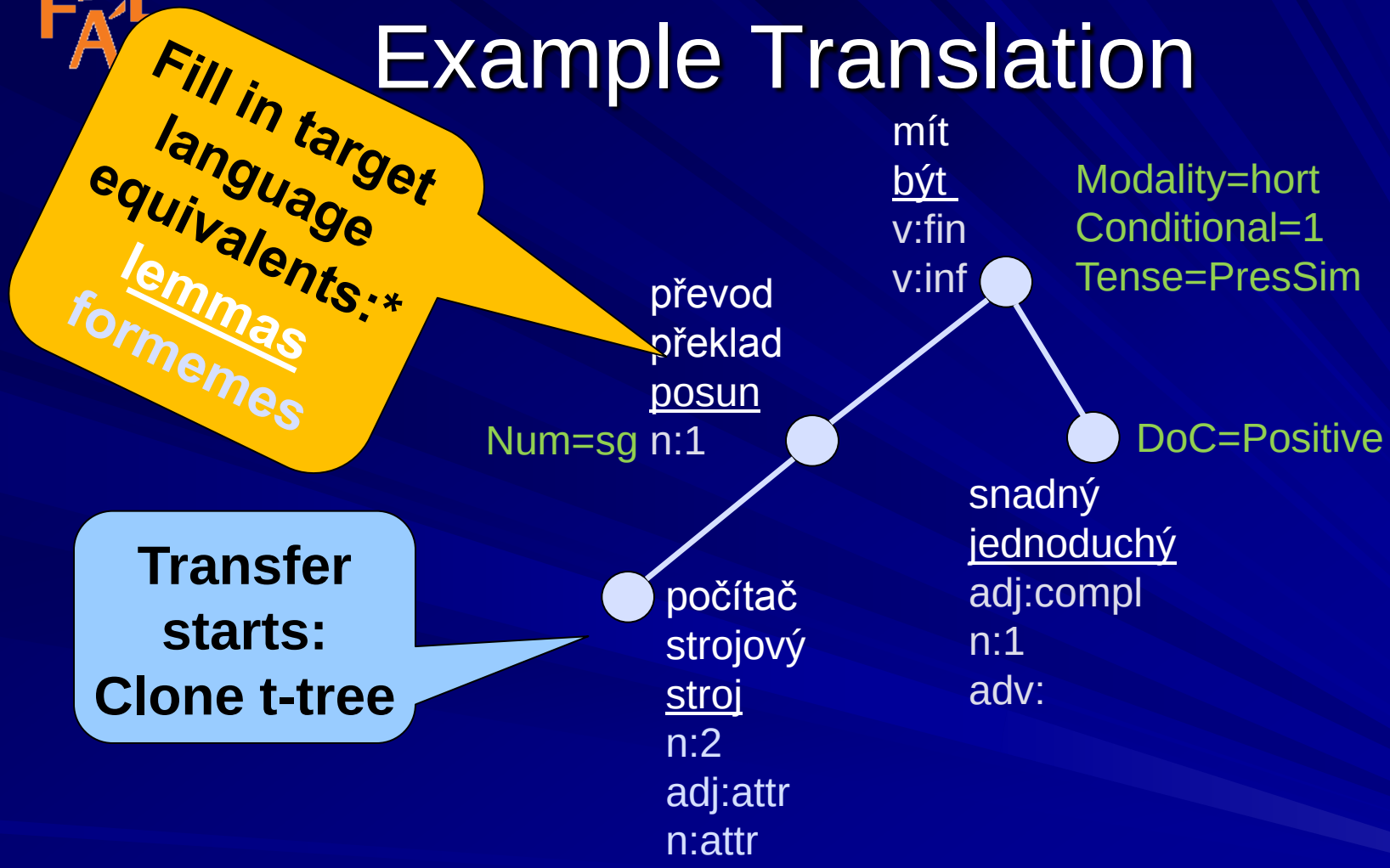


Example Translation

T-tree
backbone
+
formemes
+
grammatemes

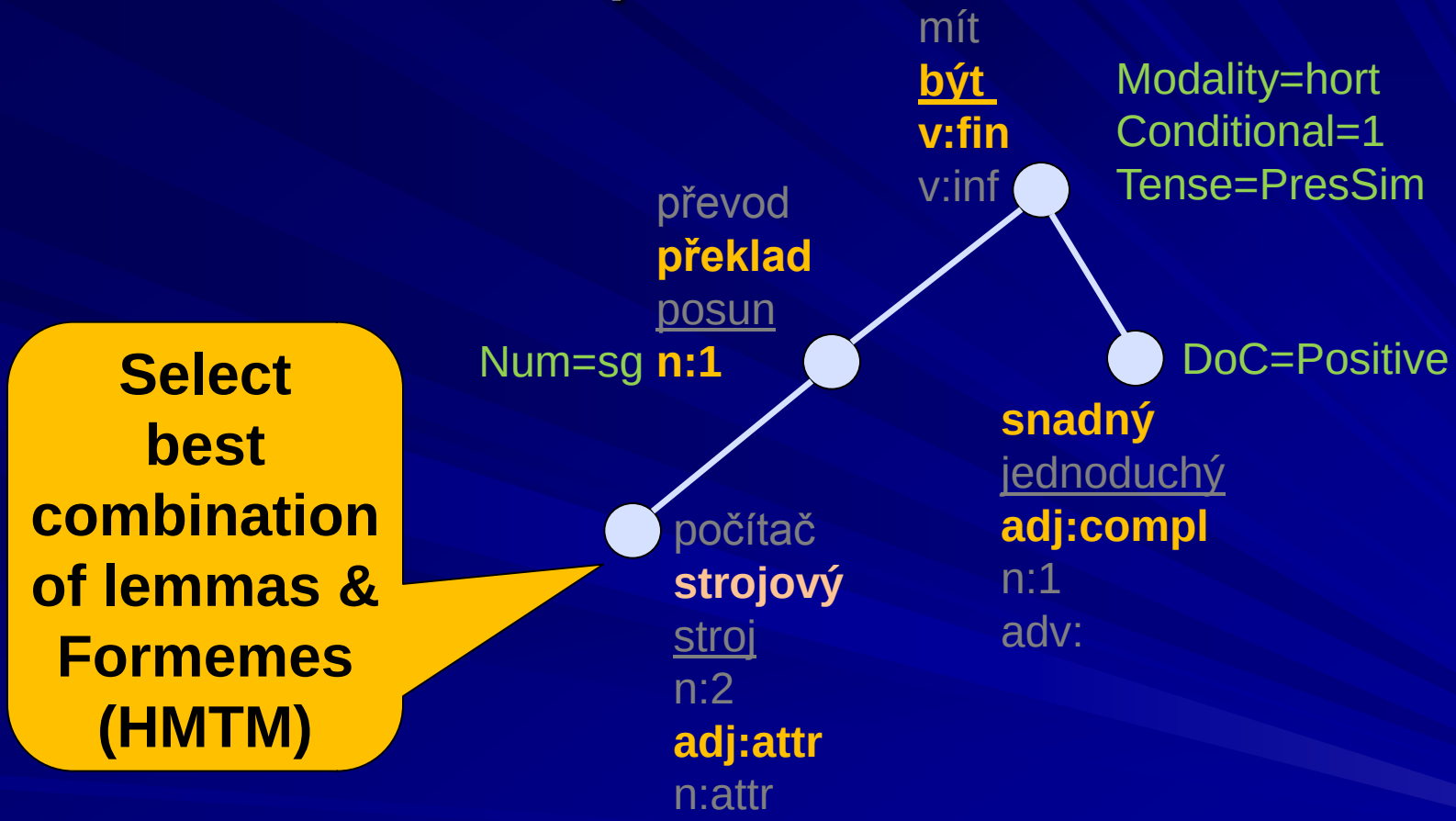


Example Translation



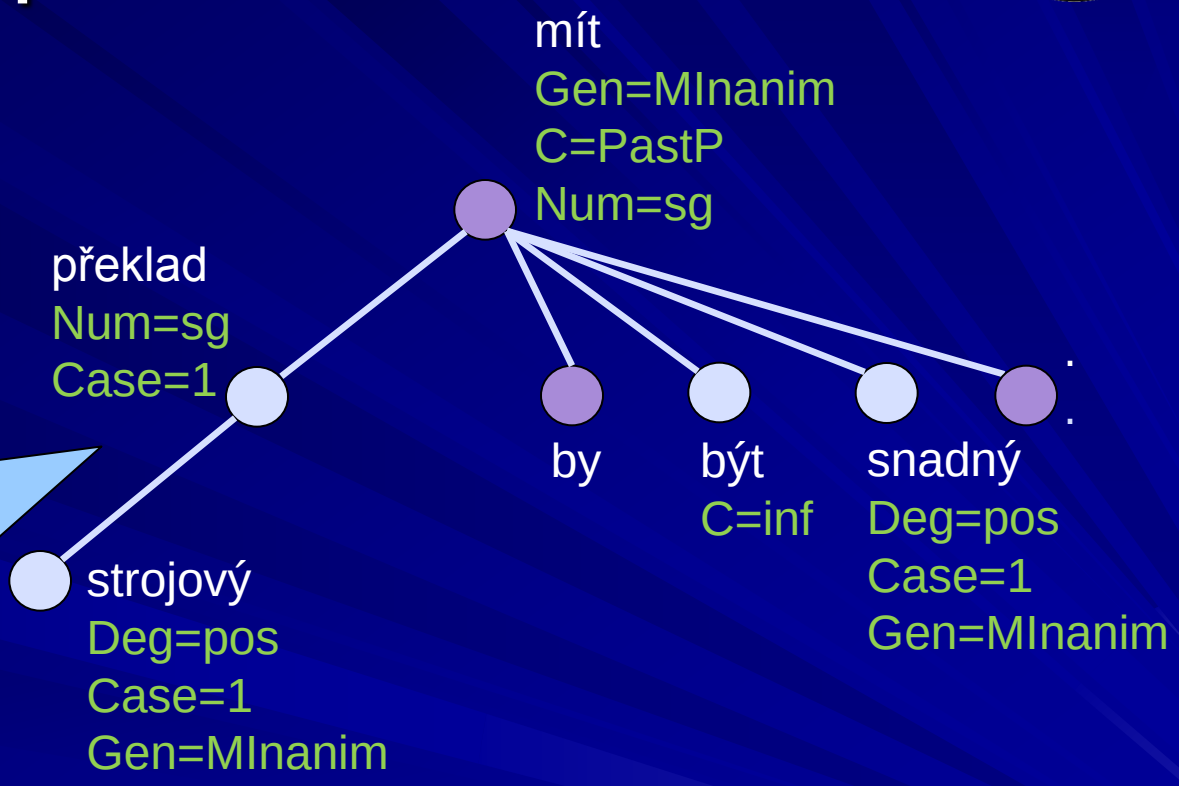
* Dictionary translation: MaxEnt classifier, ~10^6 features

Example Translation

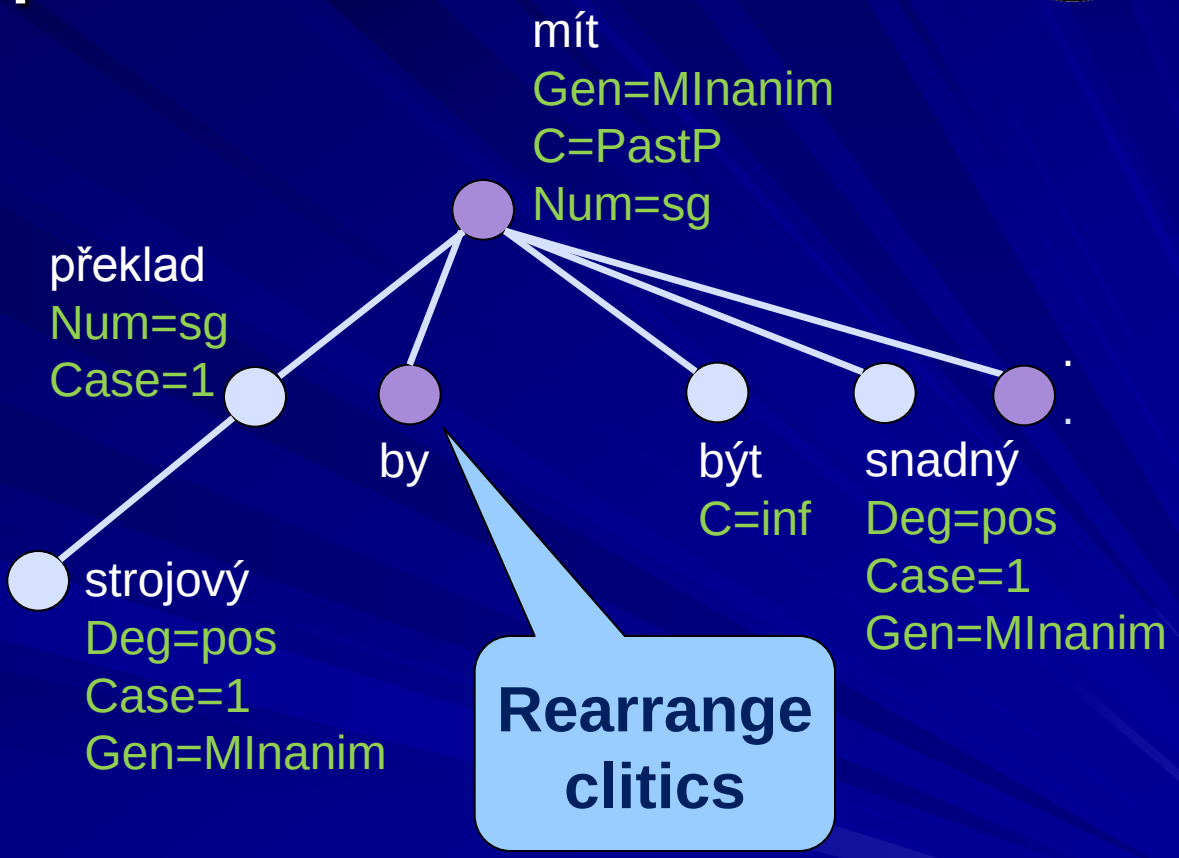


Example Translation

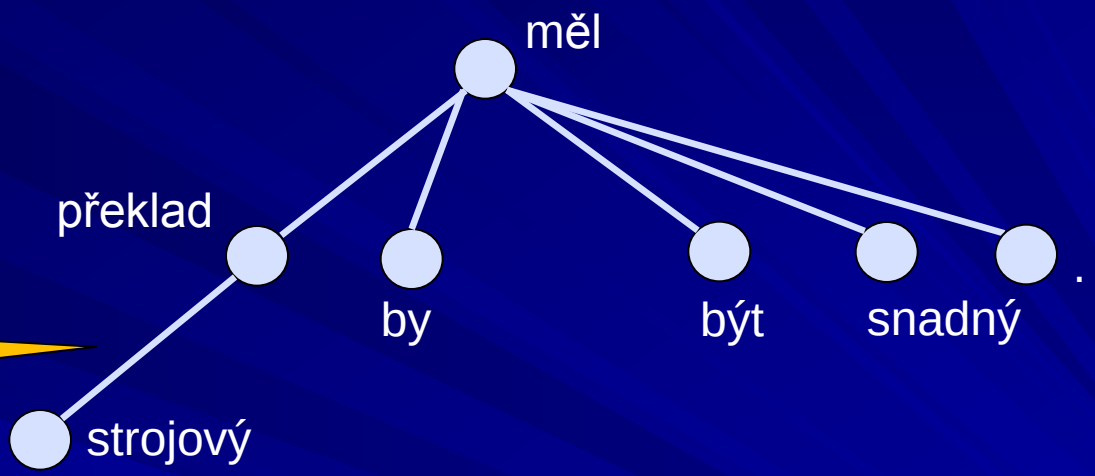
Clone to a-tree, add core morphological & POS tags + agreement + function words



Example Translation



Example Translation



Synthesize word forms

... and flatten the tree: (capitalize, space)

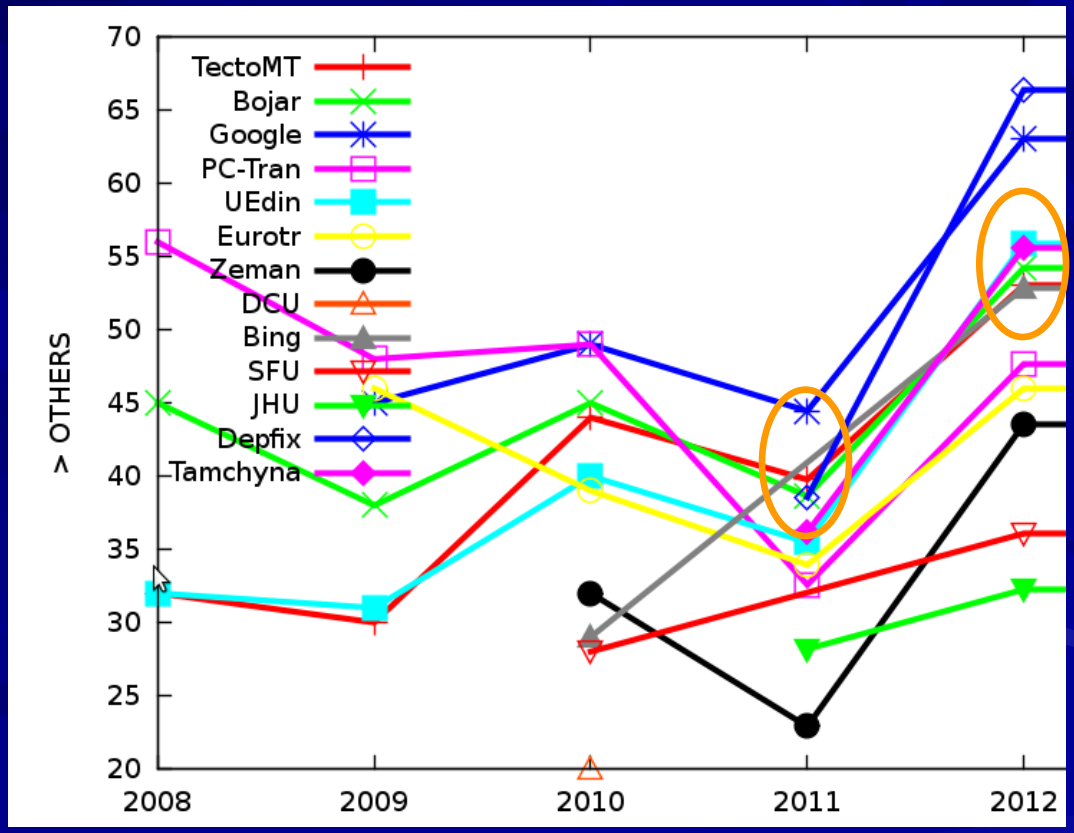
Strojový překlad by měl být snadný.

Results

- WMT Constrained task en → cs:
  - TectoMT, Moses (Prague), Moses (Edinburgh) tied 1st
- Unconstrained:

  (subj. eval.)
- BLEU

  All < 0.17



The Future

Acknowledgements:
Charles University research funds
("PRVOUK")

- Non-isomorphic trees
  - Better breakdown to treelets and/or parameter training (than in STSG)
- Multiple paths / n-best lists
  - At least until statistical components
- Combine with Moses (using input lattices)
  - Two „languages": original & Czech by TectoMT
- Moses with syntactic and semantic factors
- Still more generalized syntax and semantics (AMR/MRS and beyond?)

References

Zdeněk Žabokrtský, Martin Popel: Hidden Markov Tree Model in Dependency-based Machine Translation. In *ACL 2009*, pp. 145-148

David Mareček, Martin Popel, Zdeněk Žabokrtský: Maximum Entropy Translation Model in Dependency-Based MT Framework. *Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, ACL 2010, Uppsala, Sweden, pp. 201-206.

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák and David Mareček: Formemes in English-Czech Deep Syntactic MT. In W*MT'12*, Montréal, Canada, pp. 267-274.

Martin Popel, Zdeněk Žabokrtský: TectoMT: Modular NLP Framework. *IceTAL 2010*, 7th International Conference on Natural Language Processing, Reykjavík, Iceland, pp. 293-304.

TectoMT at WMT 12: http://www.statmt.org/wmt12/pdf/WMT02.pdf

Thank you!