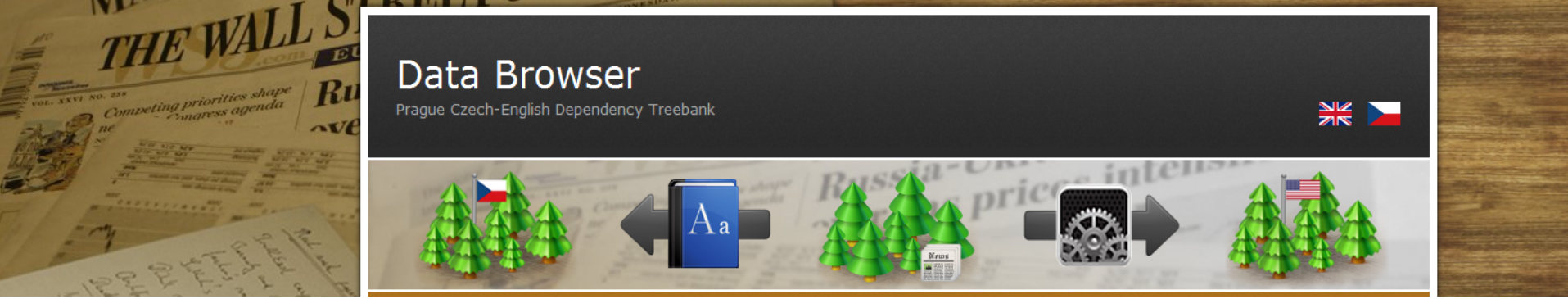
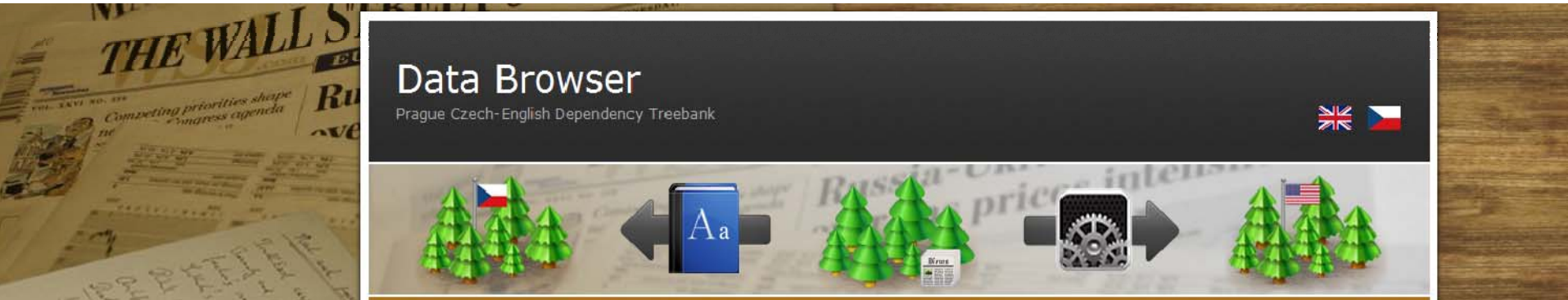


Prague Czech-English Dependency Treebank

Silvie Cinková

CLARA Joint Training Programme: Course on Treebank Annotation

Prague, December 16, 2010



PCEDT 2.0 at LDC

- Penn Treebank - Wall Street Journal
  - 49 208 sentences
  - 1 253 013 tokens
- Czech translations
- Manual tectogrammatical representation in both languages
- Valency lexicons for both languages

PCEDT **1.0** (Cuřín et al., 2004)

- 21 600 WSJ sentences + translations
- 515 manually annotated sentences on t-layer, both languages, retranslated from Czech into English by four different translation companies.

Prague English Dependency Treebank 1.0 (2009)

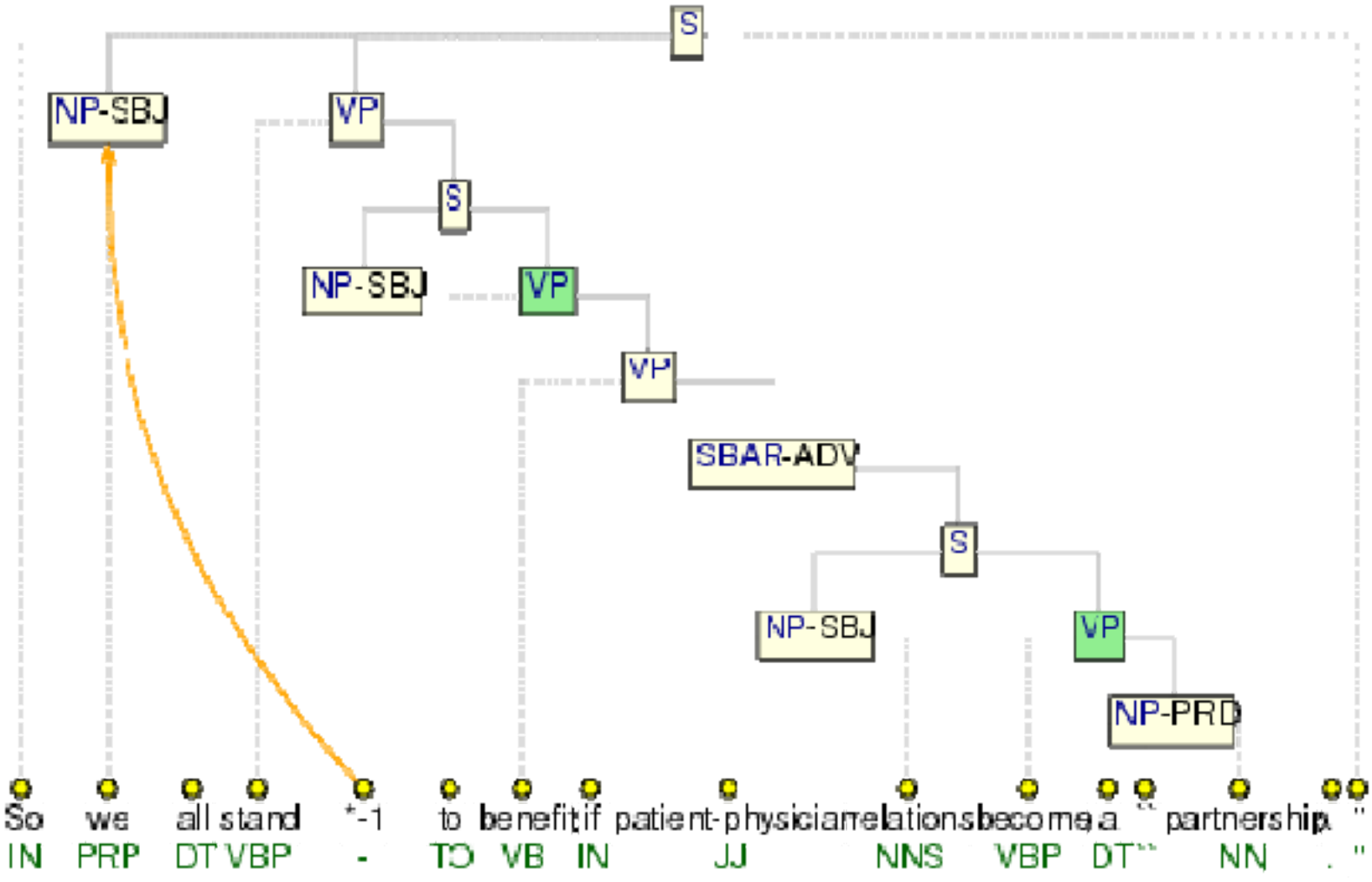
- http://ufal.mff.cuni.cz/pedt
- < 12,000 sentences (25%)

2.0 English

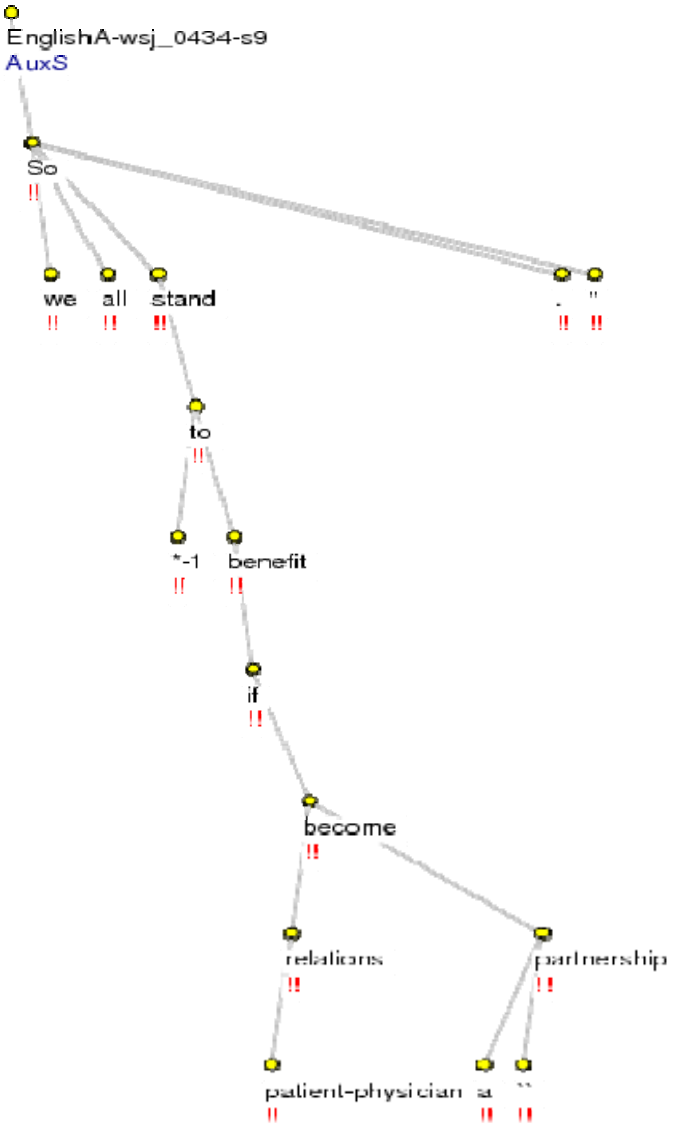
- Original phrase-structure annotation
  - Automatically converted into a-layer
  - A-layer automatically converted into t-layer
- Manual revision of pre-processed t-layer
  - Tree structure
  - Functor labels
  - Grammatical coreference
  - Verb valency (+ Engvallex lexicon)

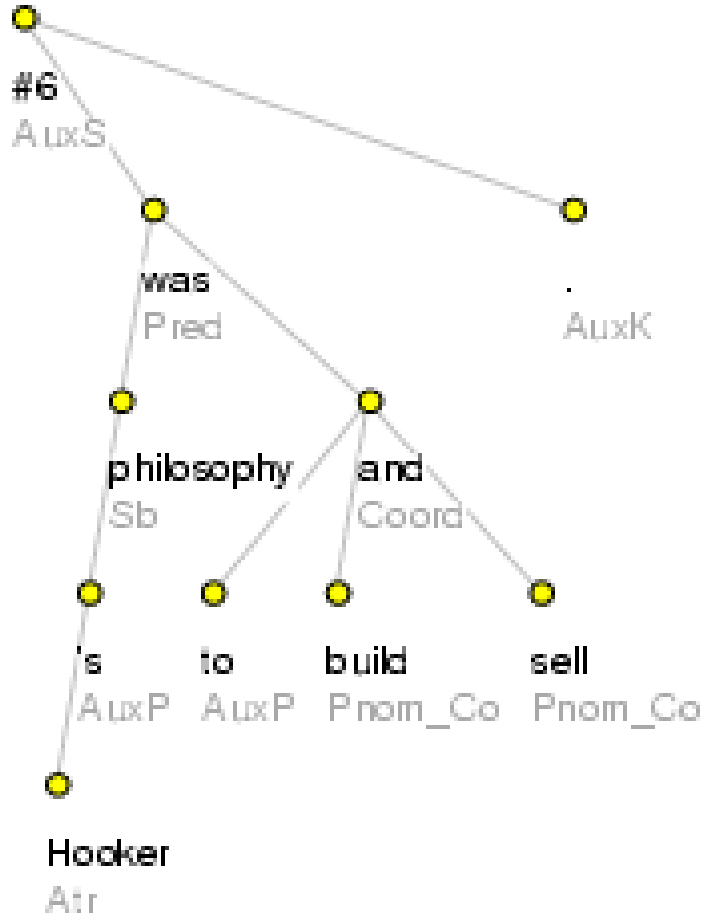
Original PTB-WSJ

So we all stand to benefit if patient-physician relations become a “partnership”.



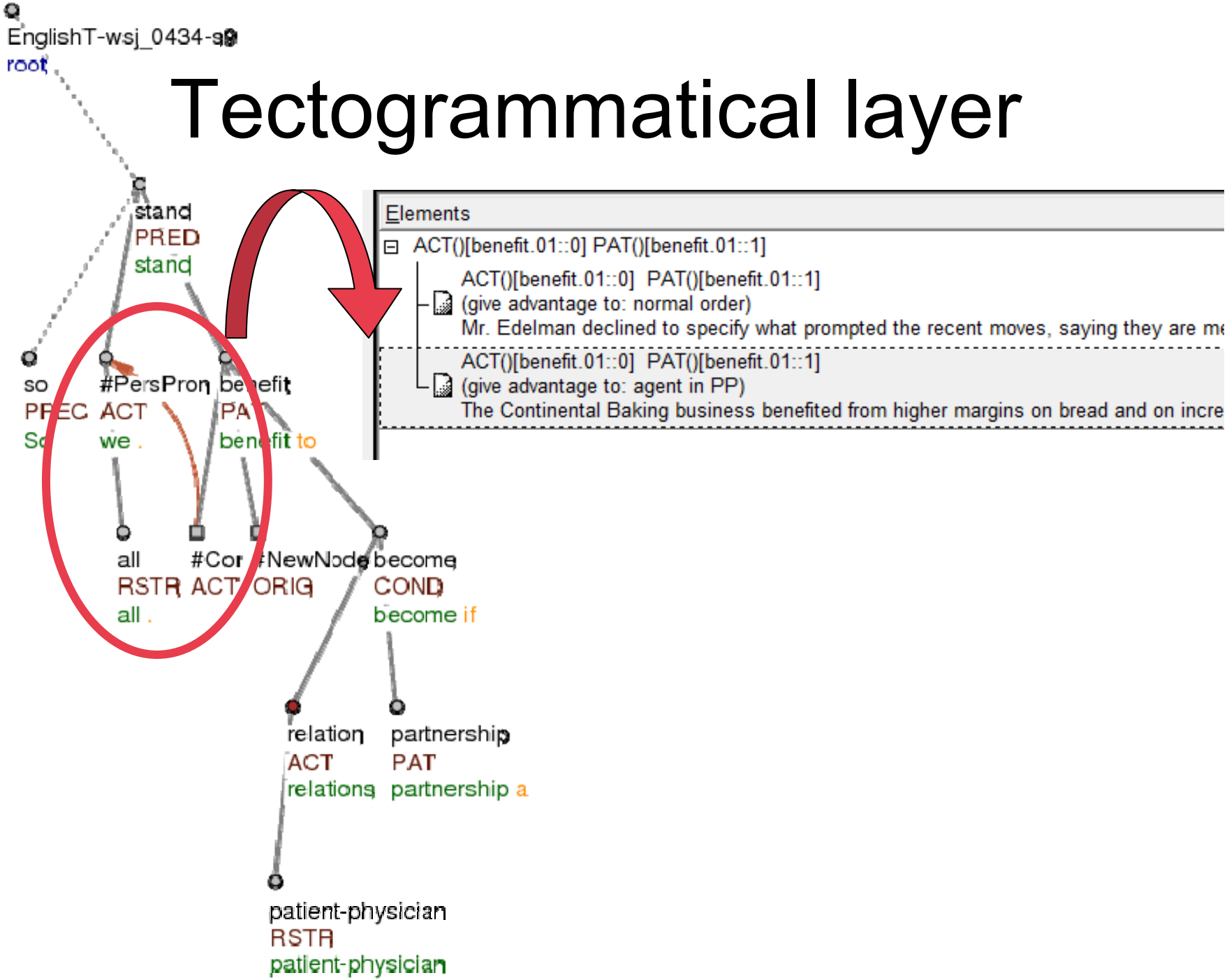
Analytical layer

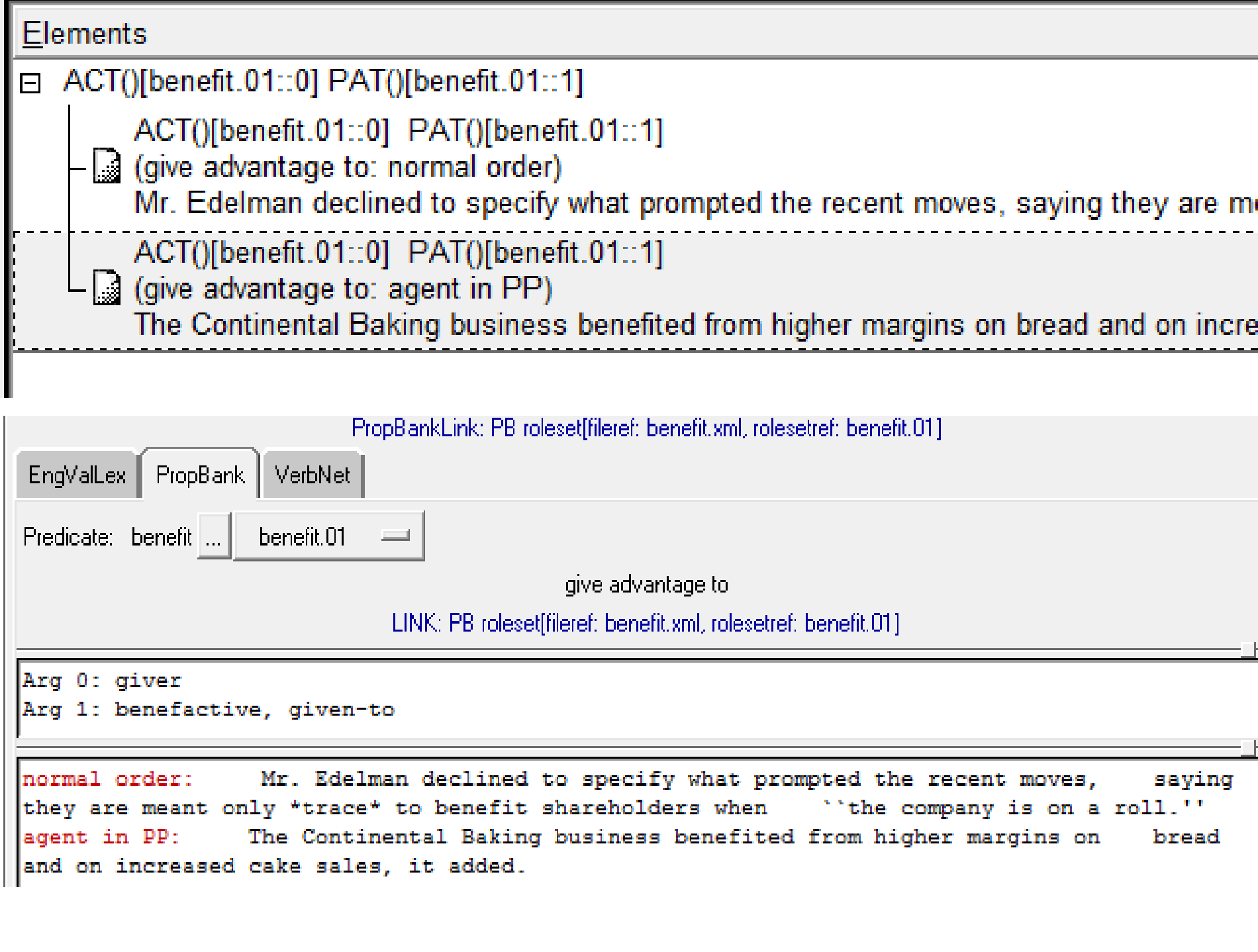




Hooker 's philosophy was to build and sell .

Tectogrammatical layer




Elements
ACT()[benefit.01::0] PAT()[benefit.01::1]
ACT()[benefit.01::0] PAT()[benefit.01::1]
(give advantage to: normal order)
Mr. Edelman declined to specify what prompted the recent moves, saying they are m
ACT()[benefit.01::0] PAT()[benefit.01::1]
(give advantage to: agent in PP)
The Continental Baking business benefited from higher margins on bread and on incre
PropBankLink: PB roleset[fileref: benefit.xml, rolesetref: benefit.01]
EngValLex
PropBank
VerbNet
Predicate: benefit ...
benefit.01
give advantage to
LINK: PB roleset[fileref: benefit.xml, rolesetref: benefit.01]
Arg 0: giver
Arg 1: benefactive, given-to
normal order: Mr. Edelman declined to specify what prompted the recent moves, saying they are meant only *trace* to benefit shareholders when ``the company is on a roll.''
agent in PP: The Continental Baking business benefited from higher margins on bread and on increased cake sales, it added.

Other English Annotations We Used

- PropBank
- Flat noun phrases
  - `(NP (NNP Air) (NNP Force) (NN contract))` → `(NP(NML (NNP Air) (NNP Force)) (NN contract))`
- BBN Pronoun Coreference and Entity Type Corpus

Czech

- Automatic parsing of texts (a-, t-layers)
- Manual revision of t-layer
- Valency (PDT-Vallex lexicon)
- Grammatemes
- Grammatical & **textual** coreference

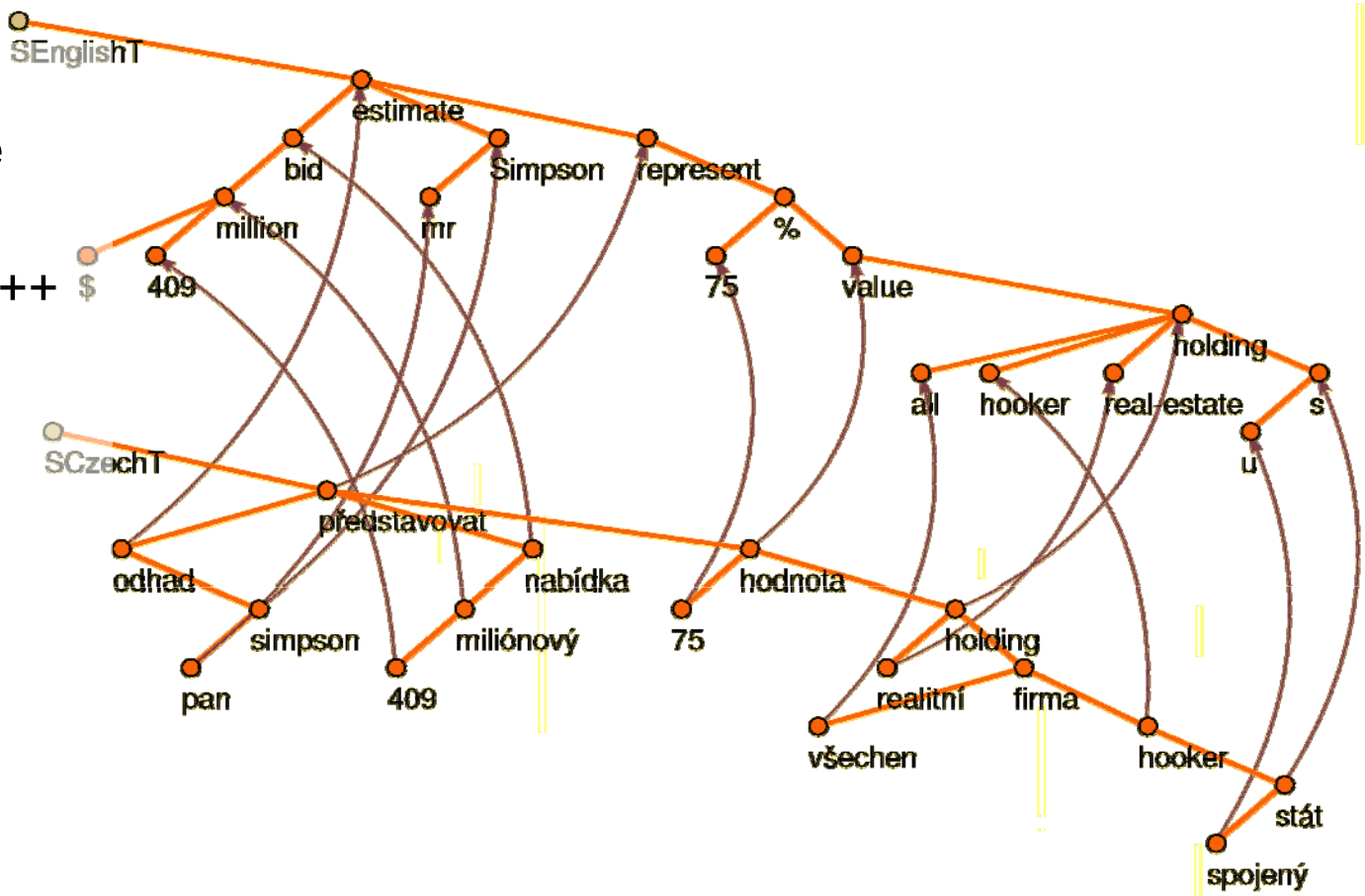
Background: PDT 2.0

Parallel Features

- Sentence alignment (implicit, sentence-to-sentence translation)
- T-node alignment (automatic)

t-aligner
(Mareček, Žabokrtský, Novák, 2008)

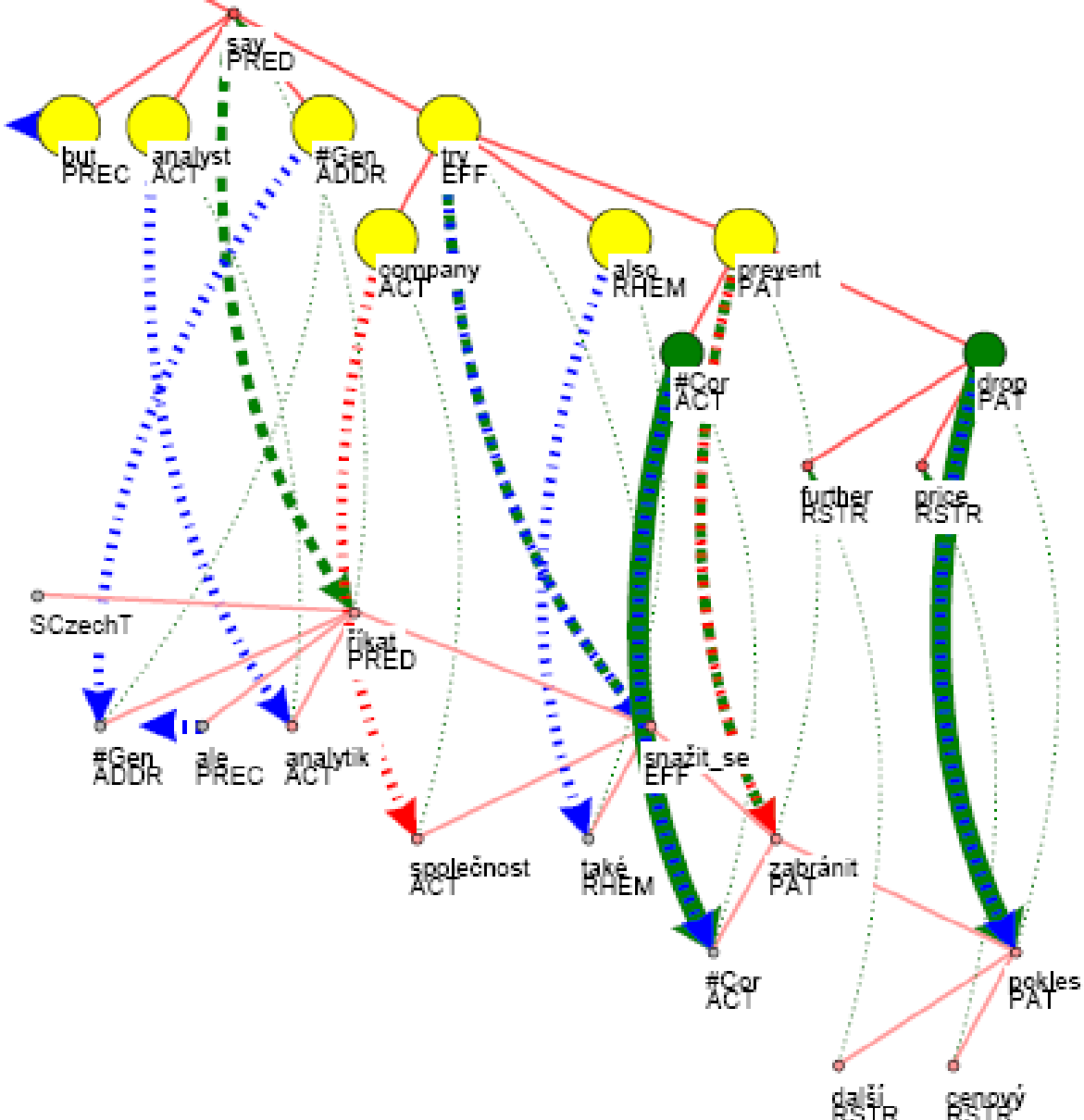
- Only t-nodes (i.e. content words) are aligned
- Outperforms GIZA++ word aligment
- Evaluated on 515 double-annotated sentences
- F-measure=90,4% (IAA was 94,7%)



Other Interesting Work

- Alignment of valency slots across the languages in the valency lexicons
- Work in progress, less than 100 paired verbs

Alignment of Lexicon Frames



Data Browser

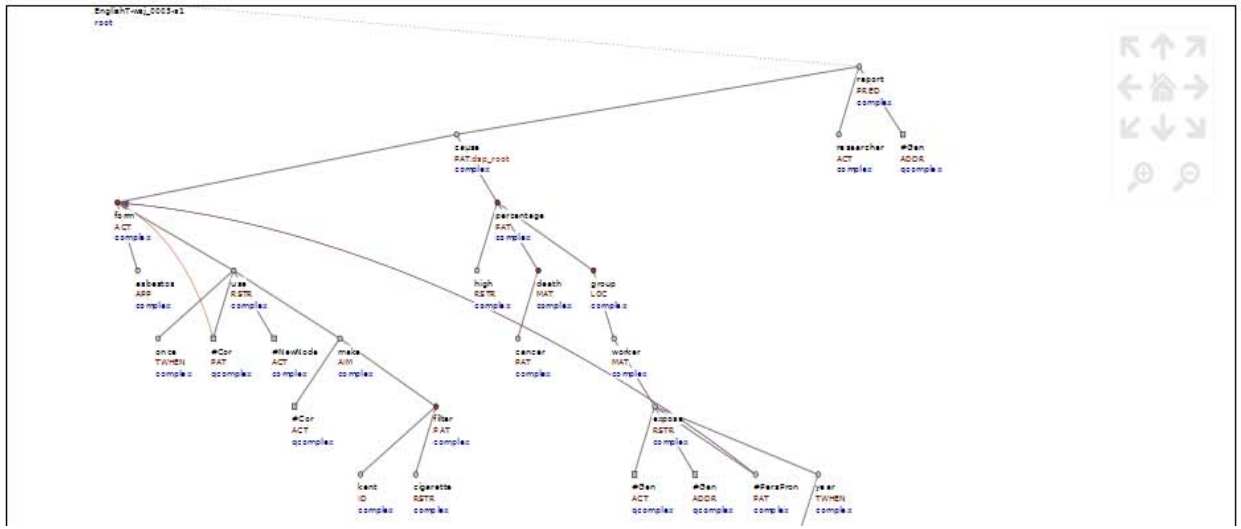
Prague Czech-English Dependency Treebank

Section 000

File no.3

Sentence no.1

A form of asbestos once used * * to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed * to it more than 30 years ago , researchers reported *T*-1 .



References

- Bojar O., Prokopová M.: **Czech-English Manual Word Alignments,** Software or data, Charles University in Prague, UFAL, Oct 2009
- Bojar O., Prokopová M.: **Czech-English Machine Translation Dictionary,** Tech. report no. 2007/-, ÚFAL MFF UK, Prague, Czech Republic, 12 pp., Apr 2007
- Bojar O., Šindlerová J.: **Building a Bilingual ValLex Using Treebank Token Alignment: First Observations,** in Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, ISBN 2-9517408-6-7, pp. 304-309, 2010
- Cinková S., Toman J., Hajič J., Čermáková K., Klimeš V., Mladová L., Šindlerová J., Tomšů K., Žabokrtský Z.: **Tectogrammatical Annotation of the Wall Street Journal,** in The Prague Bulletin of Mathematical Linguistics, No. 92, Univerzita Karlova, Prague, Czech Republic, ISSN 0032-6585, 2009
- Cuřín J., Čmejrek M., Havelka J., Hajič J., Kuboň V. and Žabokrtský Z. 2004. **Prague Czech-English Dependency Treebank, Version 1.0.** Linguistic Data Consortium, LDC2004T25.
- Hajič J., Cinková S., Čermáková K., Mladová L., Nedolužko A., Pajas P., Semecký J., Šindlerová J., Toman J., Tomšů K., Korvas M., Rysová M., Veselovská K., Žabokrtský Z.: **Prague English Dependency Treebank 1.0,** Software or data, Institute of Formal and Applied Linguistics, Charles University in Prague, Malostranské nám. 25, 118 00 Praha 1, ISBN 978-80-904175-0-2 , Jan 2009
- Mareček D.: **Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus,** Master thesis, Charles University, MFF UK, 90 pp., Jul 2008
- Mareček D., Žabokrtský Z., Novák V.: **Automatic Alignment of Czech and English Deep Syntactic Dependency Trees,** in Proceedings of the Twelfth EAMT Conference, Copyright © HITEC e.V., Hamburg, Germany, ISBN 978-3-00-025770-4, pp. 102-111, 2008
- Mikulová M., Štěpánek J.: **Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank,** in Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, ISBN 2-9517408-6-7, pp. 1836-1839, 2010
- Palmer M., Gildea D., Kingsbury P., **The Proposition Bank: A Corpus Annotated with Semantic Roles**, Computational Linguistics Journal, 31:1, 2005.
- Vadas, D. and Curran, J. R.: **Adding Noun Phrase Structure to the Penn Treebank.** Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 240–247, Prague, Czech Republic, June 2007.
- Weischedel R. and Brunstein A. **BBN Pronoun Coreference and Entity Type Corpus**. 2005. Linguistic Data Consortium, LDC2005T33.