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)
So we all stand to benefit if patient-physician relations become a “partnership”.
Hooker's philosophy was to build and sell.
Tectogrammatical layer

Elements
- 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 more...
- 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...
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.'"

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
  \[- (NP (NNP Air) (NNP Force) (NN contract)) \rightarrow (NP(NML (NNP Air) (NNP Force)) (NN contract))\]
• 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 alignment
- 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
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.
References