HamleDT 2.0: Thirty Dependency Treebanks Stanfordized

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský

Charles University in Prague,
Faculty of Mathematics and Physics,
ÚFAL (Institute of Formal and Applied Linguistics)

September 14th 2014, Sedlec
HamleDT 2.0

- 30 existing treebanks all converted to Prague Dependencies
- No need to learn 30 tagsets!
- No need to study 30 TB manuals!
- Now also in Universal Stanford Dependencies!

13 TBs free to download
- Ancient Greek
- Arabic
- Czech
- Danish
- Dutch
- Estonian
- Finnish
- Latin
- Persian
- Portuguese
- Romanian
- Swedish
- Tamil

8 TBs easy to get
(free download from owners)
- Bulgarian
- Catalan
- Dutch
- English
- Estonian
- Finnish
- Spanish
- Turkish
- Hungarian
- Japanese
- Slovene

9 TBs harder to get
(have to ask/pay the owners)
- Basque
- Bengali
- Greek
- Hindi
- Italian
- Russian
- Slovak
- Telugu

- < 100,000 tokens
- < 500,000 tokens
- > 500,000 tokens
Step 1: Existing TB

treebank-specific annotation (dependency labels, POS tags, morphological features)
Step 2: Harmonization

treebank-specific conversion to Prague dependencies and Interset
Step 3: Stanfordization

conversion to Universal Stanford Dependencies (USD) and Universal POS Tagset
USD: all auxiliary nodes are leaves (conjunctions, copulas, adpositions, articles,...)
Coordination

PRG:

John brought apples and Mary brought oranges

USD:
Coordination with ellipsis

PRG:

John brought apples and Mary oranges

USD:
Visit us at...

http://ufal.mff.cuni.cz/hamledt
http://ufal.mff.cuni.cz/hamledt

- 13 treebanks for download from LINDAT
  - 3 formats: *conll, *treex.gz, *.stanford
  - 3 styles: Prague, Stanford, original
  - Train & test split

- Conversion pipeline for all 30 treebanks available in Treex

- Queryable via PML-TQ from LINDAT
Future work

- HamleDT 3.0?
  - more languages, more treebanks per language
    (We know about 110 dependency treebanks.)
  - more “Free” treebanks
  - English translations and alignments (Google Translate)
  - Interset vs. new Universal Features
  - new CoNLL-U format
- Closer cooperation with Stanford and Google teams
- PML-TQ (update, more public treebanks)
- Experiments with parsers and learnability
Thank you

Questions?