HamleDT 2.0

Harmonized **M**ulti-**l**anguag**e** **D**ependency **T**reebank

30 Dependency Treebanks Stanfordized

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

{rosa,masek,marecek,popel,zeman,zabokrtsky}@ufal.mff.cuni.cz

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

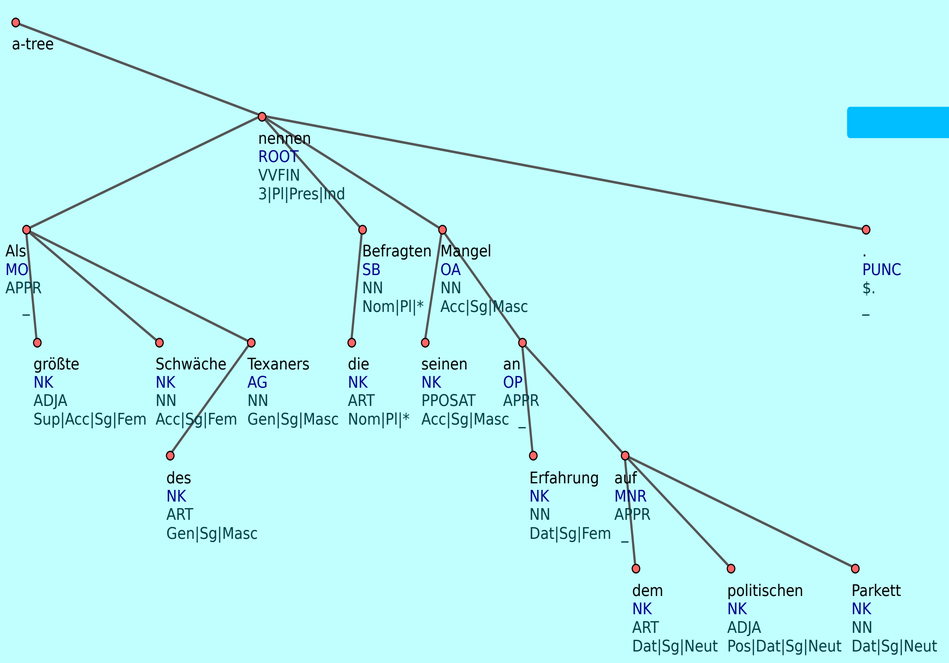
- 30 existing treebanks all converted to Prague Dependencies
- **No** need to learn 30 tagsets!
- **No** need to study 30 TB manuals!
- Now also in Stanford Dependencies!

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

- **free** download of 13 treebanks
  - .conll and .treex format
- **free** conversion pipeline for all 30 treebanks

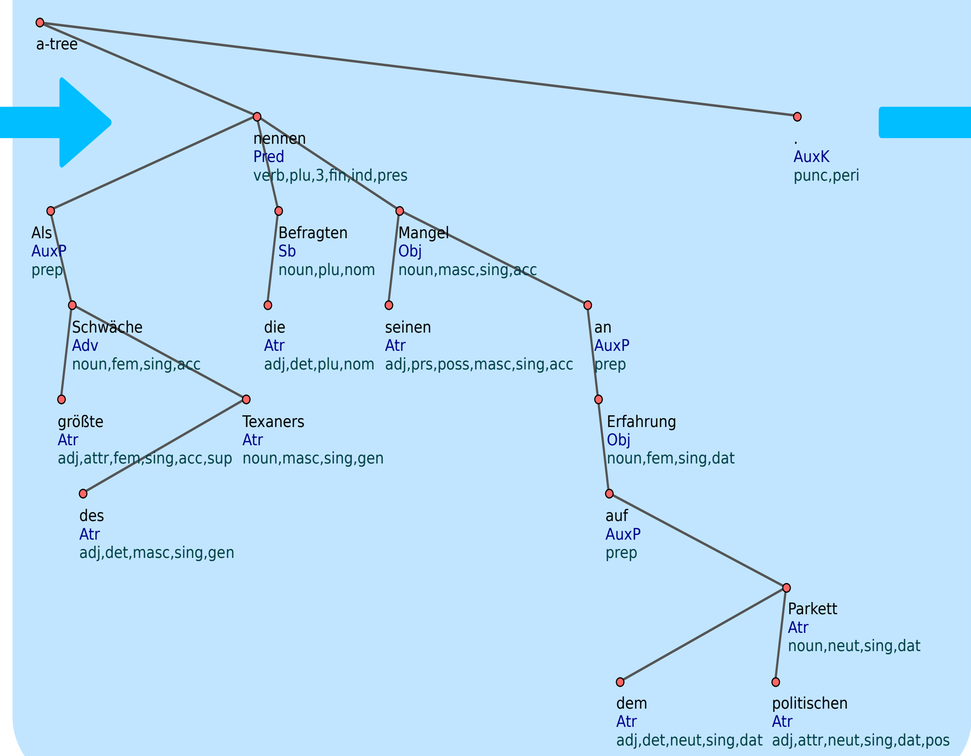
Existing Treebank

treebank-specific annotation of labelled dependecies, POS tags and morphological features



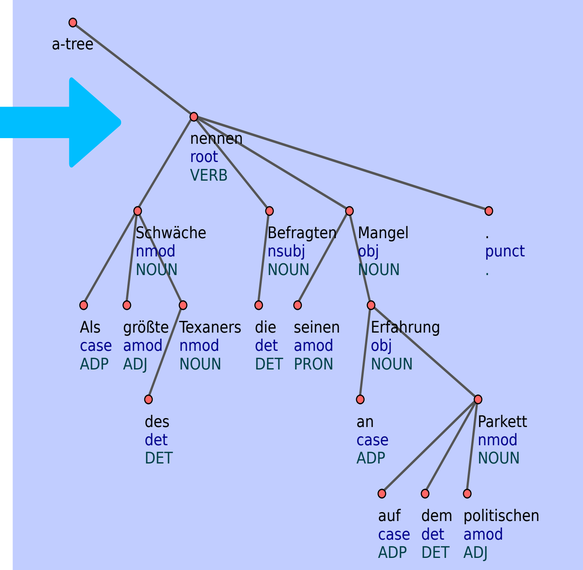
Harmonization

treebank-specific conversion to Prague Dependencies and Interset morphological features



Stanfordization

conversion to Universal Stanford Dependencies and Universal POS Tagset



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
- German
- Hungarian
- Japanese
- Slovene
- Spanish
- Turkish

< 100 000 tokens
< 500 000 tokens
> 500 000 tokens

9 TBs harder to get

(have to ask/pay the owners)

- Basque
- Bengali
- English
- Greek
- Hindi
- Italian
- Russian
- Slovak
- Telugu

Prague Dependencies

- adapted from Prague Dependency Treebank
  - used in 9 other treebanks
- adpositions and conjunctions are heads
- coordinated nodes are children of the conjunction
  - private/shared modifier distinction
- 23 dependency relation labels:
  Adv Apposition Atr Atv AtvV AuxC AuxG AuxK AuxO AuxP AuxR AuxT AuxV AuxX AuxY AuxZ Coord ExD Neg Obj Pnom Pred Sb
- tagsets converted to Interset
  - treebank tagsets very different: different granularities, labels...
  - Interset tries to capture all important morphological features
  - part of speech, gender, number, case, finitness, voice, tense...

Universal Stanford Dependencies

- new language-universal version of basic Stanford Dependencies
- based on lexicalist approach
  - edges connect lexical nodes
  - auxiliary nodes are leaves: conjunctions, copulas, adpositions...
- used in Universal Dependency Treebanks (Google)
- suggests a two-layer taxonomy
  - universal grammatical relation labels
  - fine-grained language-specific labels
- HamleDT uses 30 universal labels:
  advcl advmod amod appos aux auxpass case cc compound conj ccomp cop csubj csubjpass dep det mark mwe neg nfincl nmod nsubj nsubjpass nummod obj punct relcl remnant root xcomp

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap), SVV 260 104, Czech Science Foundationgrant no. P406/14/06548P, and GAUK 1572314. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).