Coordination Structures in Dependency Treebanks

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, Zdeněk Žabokrtský

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

Introduction

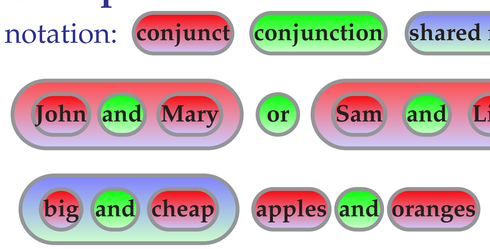
Coordination structures (CS)

are difficult to represent in dependency treebanks:

- coordination vs. dependency are fundamentally different relations
- nested coordination
- shared vs. private modifiers
- multiconjunct CS, punctuation, etc.

Examples:

notation: conjunct conjunction shared modifier



Problem

- large inter-treebank variation
- obstacle for multi-lingual parsing

Our Goal

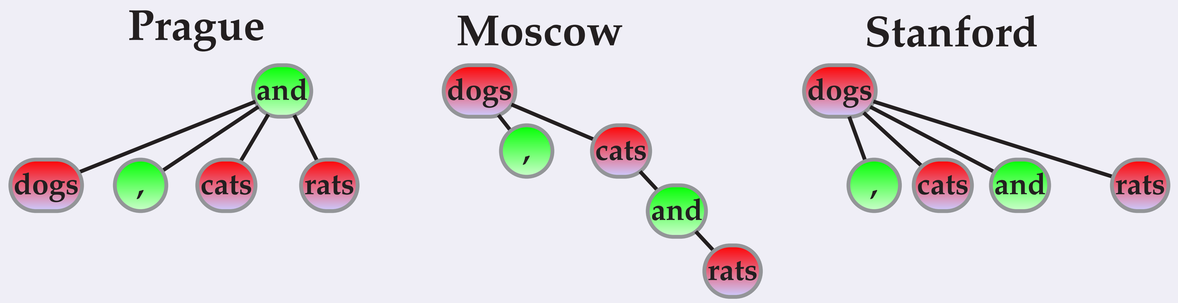
- explore the CS variations in a systematic way
- convert the treebanks into a common CS style

Novel taxonomy of CS styles

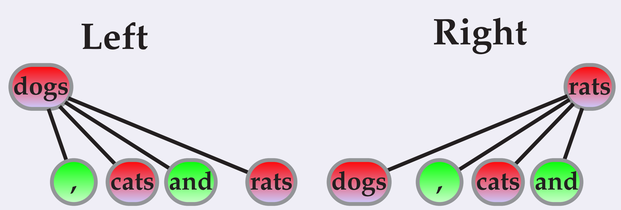
We identified

- 5 dimensions in CS tree shape variations
- 3 dimensions in CS labeling
- a few additional subtle variations
- in theory over one thousand possible styles
- 16 styles found in the real treebanks

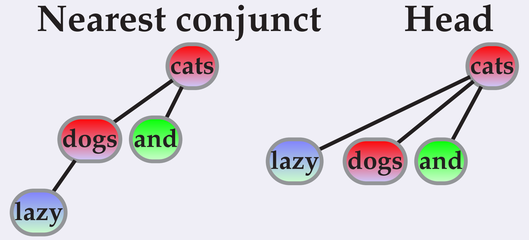
Family – configuration of conjuncts



Choice of head



Shared modifiers below



Convertibility

Different CS styles do not have equivalent expressive power ⇒ no chance for a lossless conversion.

- We developed an algorithm that decomposes a CS in one style and assembles it in another style.
- Empirical roundtrip accuracy: usually > 99%

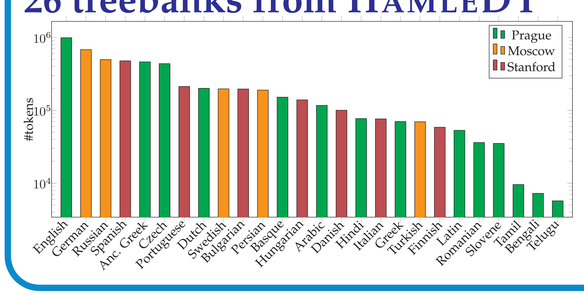
Roundtrip means e.g.
Prague → Moscow → Prague
evaluated by unlabeled attachment score.

Conclusions

- a survey of coordination styles in 26 treebanks
- a general taxonomy which covers most of the variations
- 26 treebanks converted into a common style available at http://ufal.ms.mff.cuni.cz/hamledt/
- relatively high convertibility accuracy should allow future experiments with learnability of CS by parsers

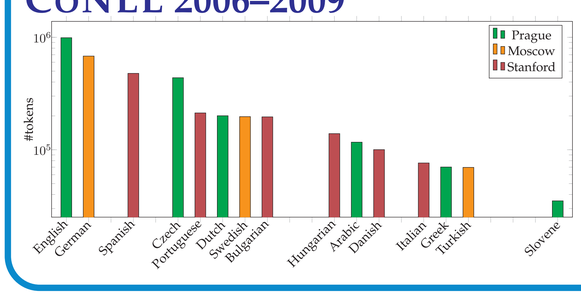
Analyzed Treebanks

26 treebanks from HamleDT



Similar collections of treebanks

CoNLL 2006–2009



Google Universal Treebanks v1.0