Tutorial on Universal Dependencies

Introduction

Joakim Nivre\textsuperscript{1}  Daniel Zeman\textsuperscript{2}  Filip Ginter\textsuperscript{3}  Francis M. Tyers\textsuperscript{4,5}

\textsuperscript{1}Department of Linguistics and Philology, Uppsala University, Sweden

\textsuperscript{2}Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic

\textsuperscript{3}Department of Information Technology, University of Turku, Finland

\textsuperscript{4}Giela ja kultuvra instituhtta, UiT Norgga árktalaš universitehta, Tromsø, Norway

\textsuperscript{5}Arvutiteaduse instituut, Tartu Ülikool, Estonia
Introduction

- Increasing interest in multilingual NLP
  - Multilingual evaluation campaigns to test generality
  - Cross-lingual learning to support low-resource languages

- Increasing awareness of methodological problems
  - Current NLP relies heavily on annotation
  - Annotation schemes vary across languages
A cat chases rats and mice
A cat chases rats and mice

En katt jagar rättor och möss
A cat chases rats and mice

En katt jagar råttor och möss

En kat jager rotter og mus
A cat chases rats and mice.
A cat chases rats and mice

En katt jagar råttor och möss

En kat jager rotter og mus
Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure transfer
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
- Hard to make progress towards a universal parser
http://universaldependencies.org

- Part-of-speech tags
- Morphological features
- Syntactic dependencies
Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual NLP and linguistic research
- Build on common usage and existing de facto standards
- Complement – not replace – language-specific schemes
- Open community effort – anyone can contribute!
The UD Philosophy

- Maximize parallelism – but don’t overdo it
  - Don’t annotate the same thing in different ways
  - Don’t make different things look the same
  - Don’t annotate things that are not there

- Universal taxonomy with language-specific elaboration
  - Languages select from a universal pool of categories
  - Allow language-specific extensions
Design Principles

- Dependency
  - Widely used in practical NLP systems
  - Available in treebanks for many languages

- Lexicalism
  - Basic annotation units are words – syntactic words
  - Words have morphological properties
  - Words enter into syntactic relations

- Recoverability
  - Transparent mapping from input text to word segmentation
Morphological Annotation

- Lemma representing the semantic content of a word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form
The cat could have chased all the dogs down the street.

- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause
Syntactic Annotation

The cat could have chased all the dogs down the street.

- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause
The cat could have chased all the dogs down the street.

- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause
Syntactic Annotation

- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause
CoNLL-U Format

- Revised and extended version of CoNLL-X format
- Two-level segmentation and enhanced dependencies
Where are we today?

- Brief history of UD:
  - First guidelines launched in October 2014
  - Treebank releases (roughly) every six months
  - Version 2 in December 2016 (guidelines) and March 2017 (treebanks)

- UD in numbers:
  - 50 languages
  - 70 treebanks
  - 164 contributors
  - 8000+ downloads

- Current UD events:
  - CoNLL shared task on UD parsing
  - First UD workshop (Gothenburg, May 22)
  - Next release in November (v2.1)
Outline

1. Introduction [Joakim]
2. Cross-linguistically consistent syntactic annotation [Fran]
3. Word segmentation and morphological annotation [Dan]

BREAK

4. Infrastructure, resources and tools for UD [Filip]
5. Making use of UD in NLP and linguistics [Joakim]
6. Adding a new language to UD [Fran]
7. CoNLL shared task on UD parsing [Dan]
Questions?