Tutorial on Universal Dependencies

Introduction

Joakim Nivre[1] Daniel Zeman[2] Filip Ginter[3] Francis M. Tyers[45]

[1]Department of Linguistics and Philology, Uppsala University, Sweden

[2]Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic

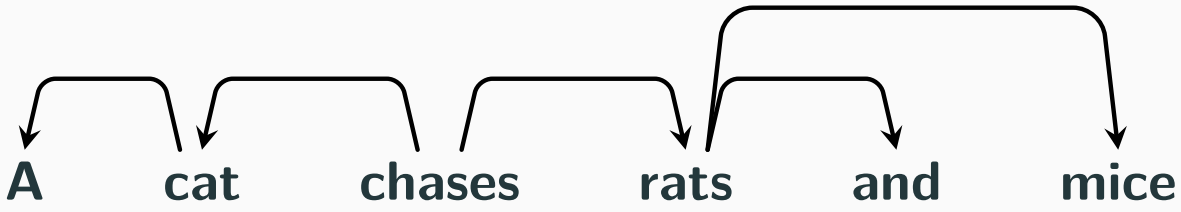
[3]Department of Information Technology, University of Turku, Finland

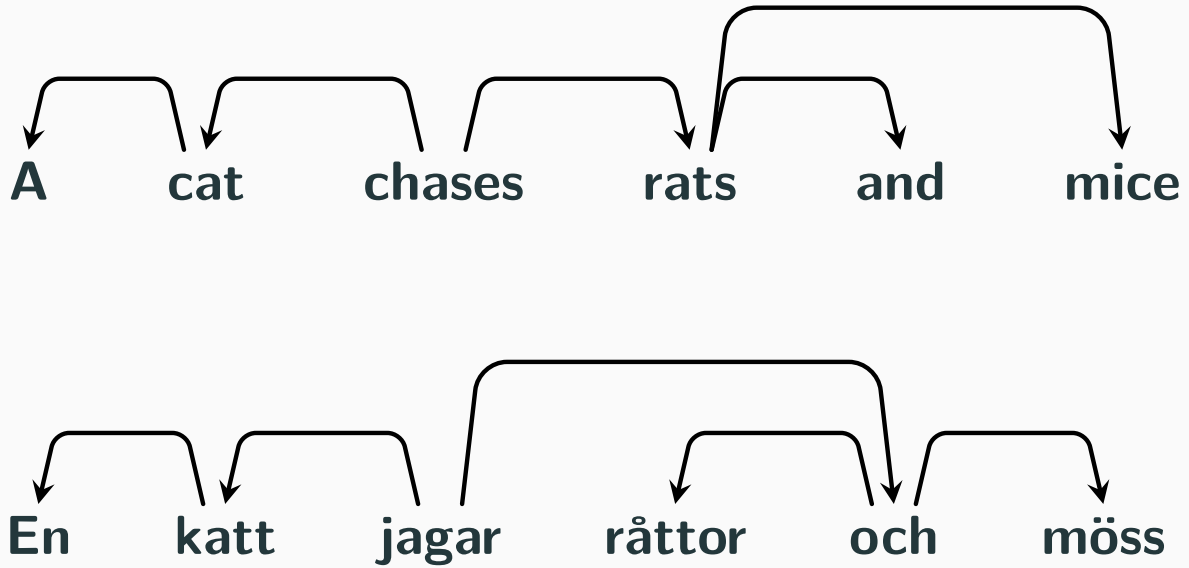
[4]Giela ja kultuvrra instituhtta, UiT Norgga árktalaš universitehta, Tromsø, Norway

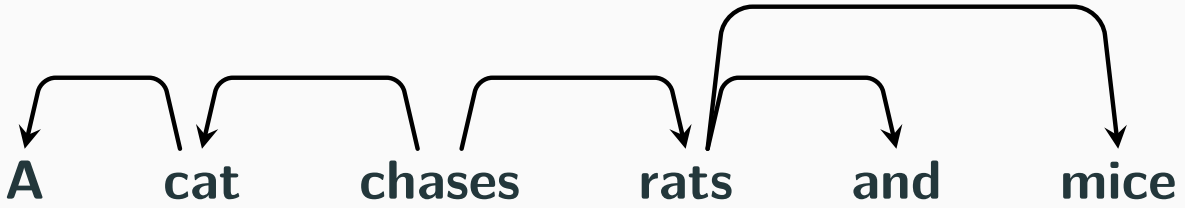
[5]Arvutiteaduse instituut, Tartu Ülikool, Estonia

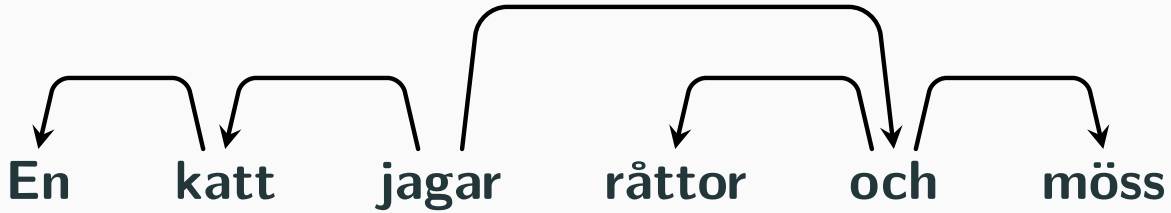
Introduction

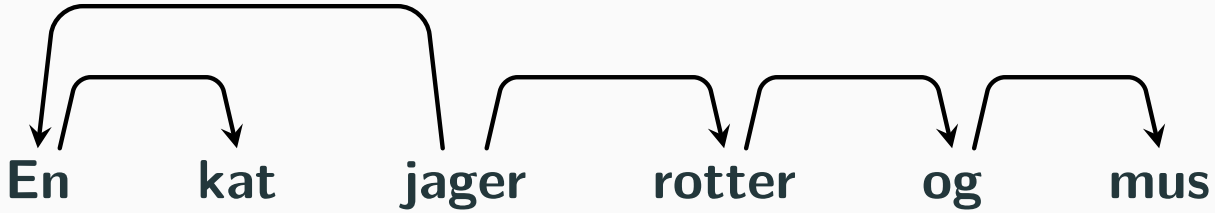
- Increasing interest in multilingual NLP
  - Multilingual evaluation campaigns to test generality
  - Cross-lingual learning to support low-resource languages
- Increasing awareness of methodological problems
  - Current NLP relies heavily on annotation
  - Annotation schemes vary across languages

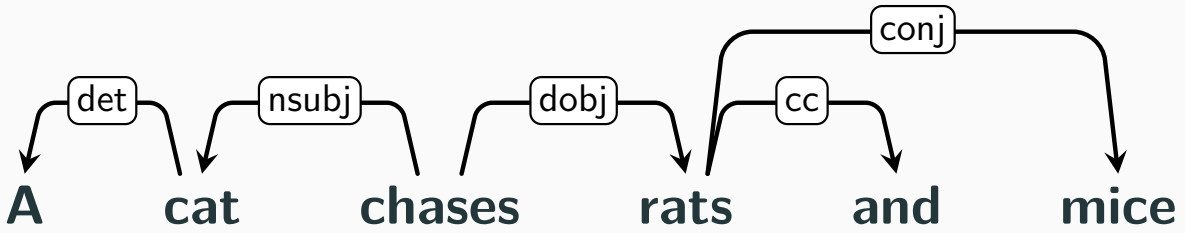

A
cat
chases
rats
and
mice

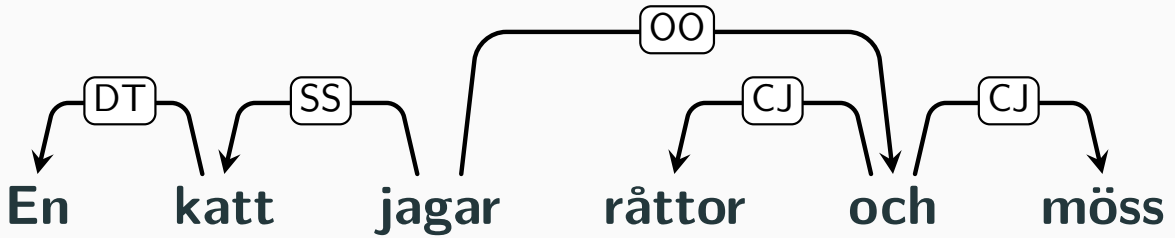

A
cat
chases
rats
and
mice
En
katt
jagar
råttor
och
möss

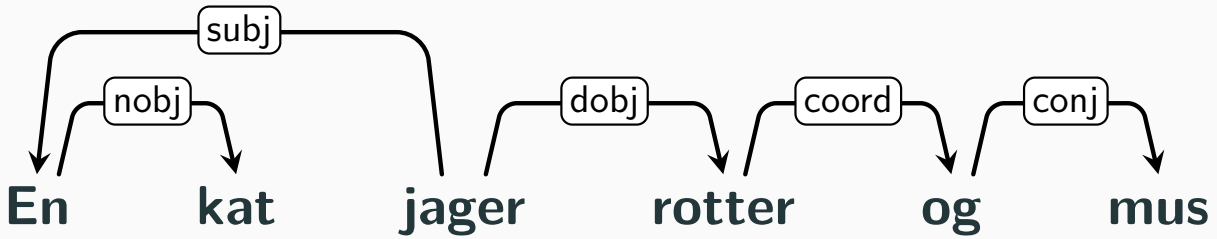

A
cat
chases
rats
and
mice

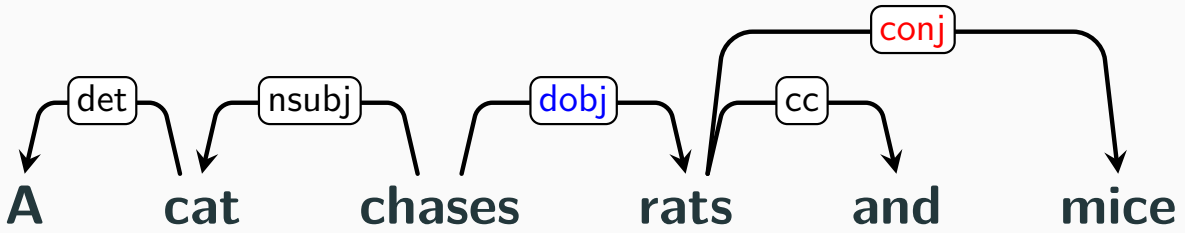

En
katt
jagar
råttor
och
möss

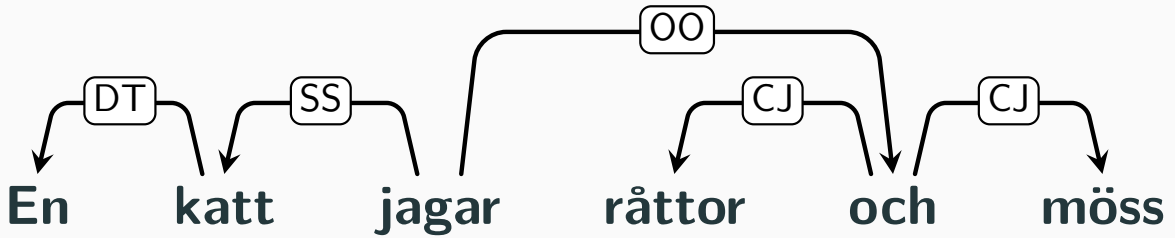

En
kat
jager
rotter
og
mus

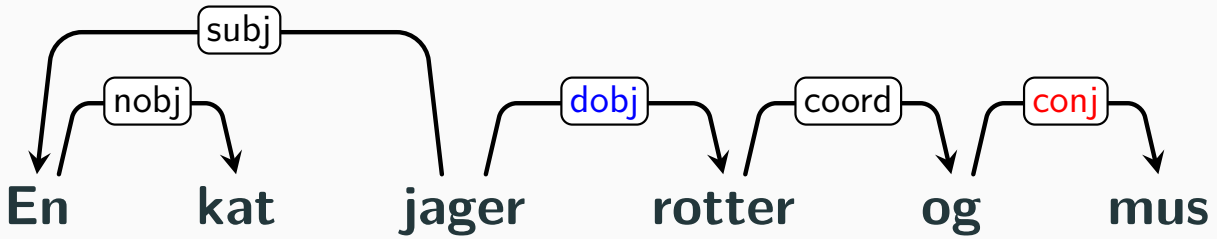

det
nsubj
dobj
cc
conj
A
cat
chases
rats
and
mice


DT
SS
OO
CJ
CJ
En
katt
jagar
råttor
och
möss


subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus


conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice


OO
DT
SS
CJ
CJ
En
katt
jagar
råttor
och
möss

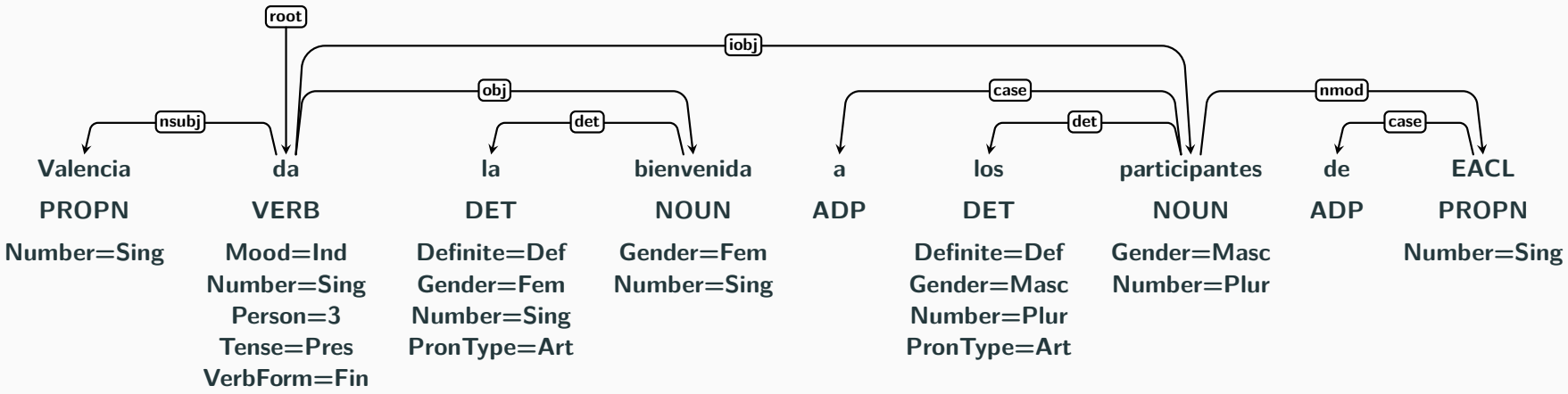

subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus

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

Universal Dependencies

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

Le	chat	chasse	les	chiens	.
le	chat	chasser	le	chien	.
DET	NOUN	VERB	DET	NOUN	PUNCT
Definite=Def Gender=Masc Number=Sing	Gender=Masc Number=Sing	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	Definite=Def Gender=Masc Number=Plur	Gender=Masc Number=Plur	

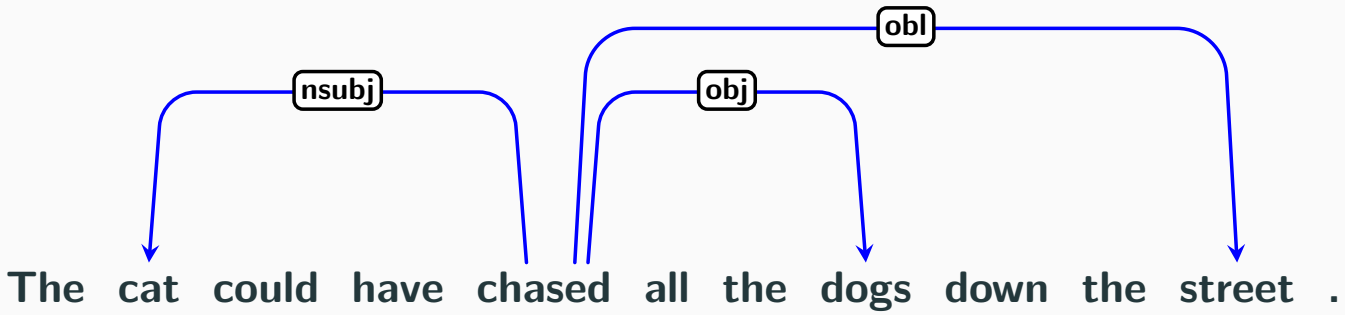
- Lemma representing the semantic content of a word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Syntactic Annotation

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

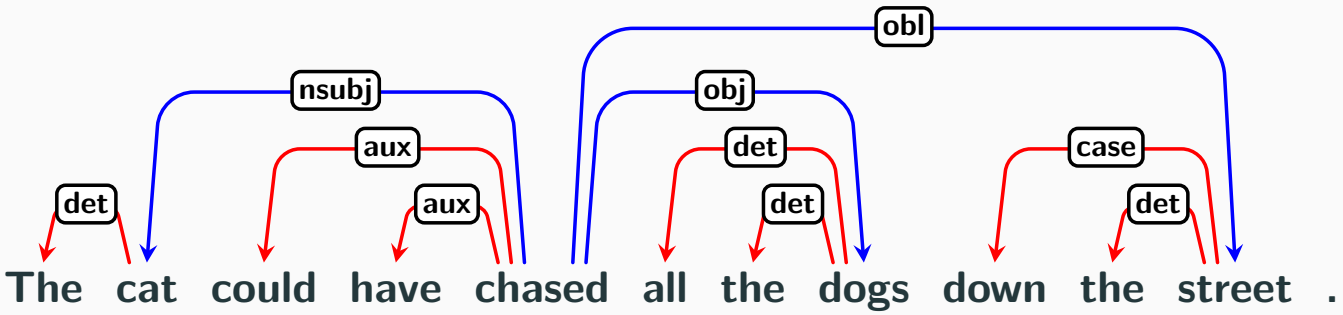
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



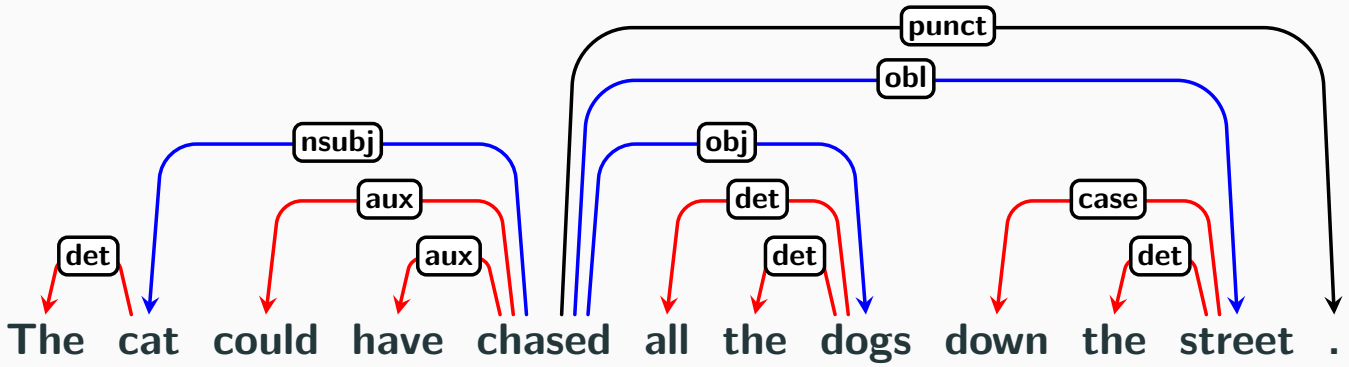
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

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

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

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	_	_	2	det	_	_
2	chat	chat	NOUN	_	_	3	nsubj	_	_
3	boit	boire	VERB	_	_	0	root	_	_
4-5	du	_	_	_	_	_	_	_	_
4	de	de	ADP	_	_	6	case	_	_
5	le	le	DET	_	_	6	det	_	_
6	lait	lait	NOUN	_	_	3	obj	_	SpaceAfter=No
7	.	.	PUNCT	_	_	3	punct	_	_

- 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?