Using Treebanked Corpora: Universal Dependencies

Timo Korkiakangas, Marco Passarotti

Sunoikisis, Digital Classics, Spring 2019

What is a Treebank?

- Syntactically annotated corpus
- (usually) it features:
  - Lemmatization (disambiguated)
  - Morphological features (disambiguated)
  - Syntactic analysis

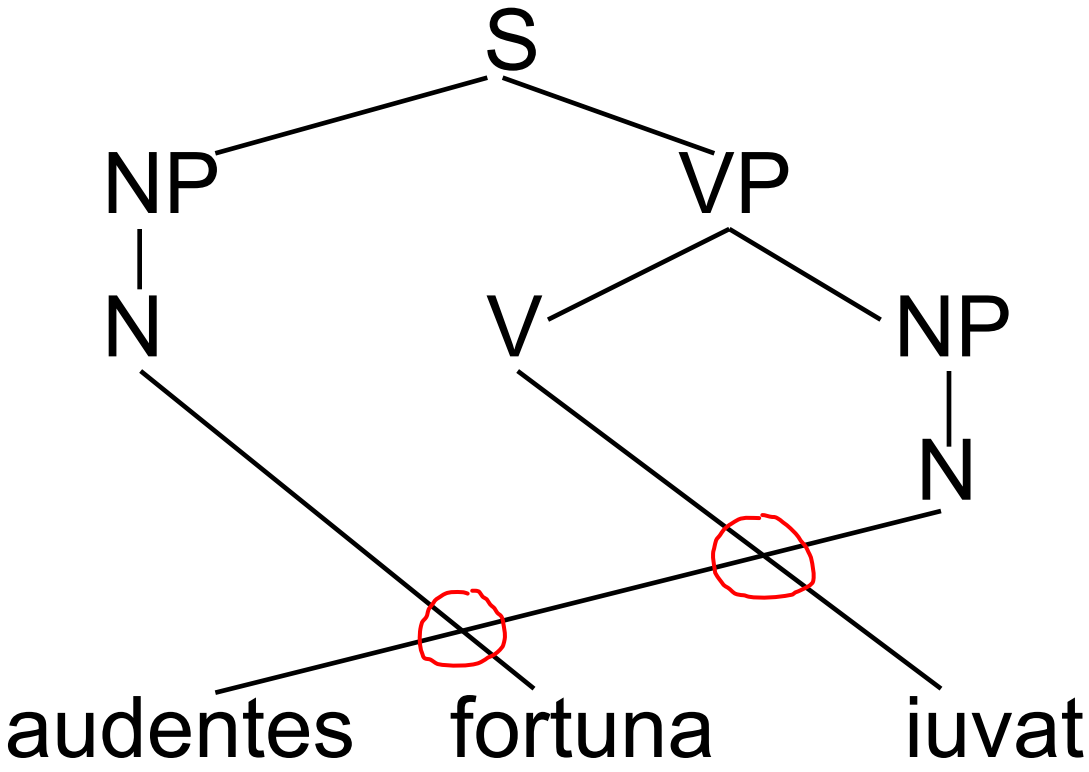
Which ‘Seeds’ for the Trees?

- Phrase Structure Grammars (PSG):
  - Words, PoS, Phrases, Start Symbols
  - Set inclusion: categorisation (e.g.: word, N, NP, S)

- Dependency Grammars (DG):
  - Only words (terminals in PSG)
  - Lexical nodes are connected via binary and hierarchical relations (“dependencies”)
  - Ho horizontal relation or cycle
  - No word order marking
  - Suitable for describing free-word-order languages: Dutch (Alpino), Italian (TUT), Czech (PDT), Latin (IT-TB; LDT; PROIEL)…

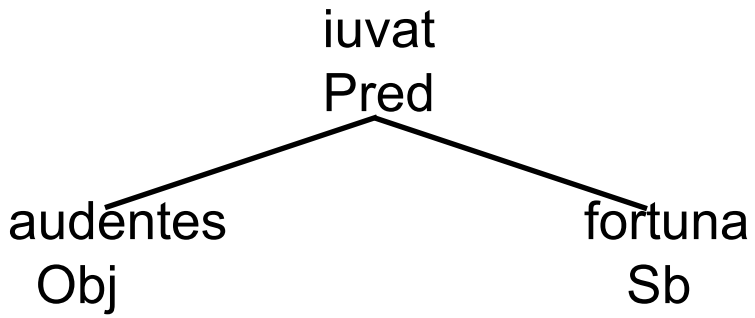
Example (PSG)

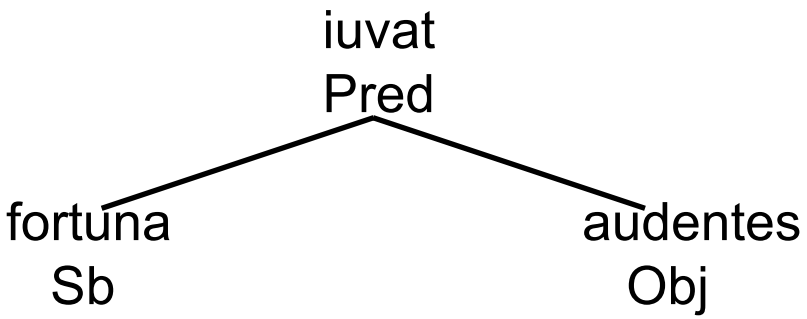
Audentes fortuna iuvat



Example (DG)

Audentes fortuna iuvat





Basics of Universal Dependencies

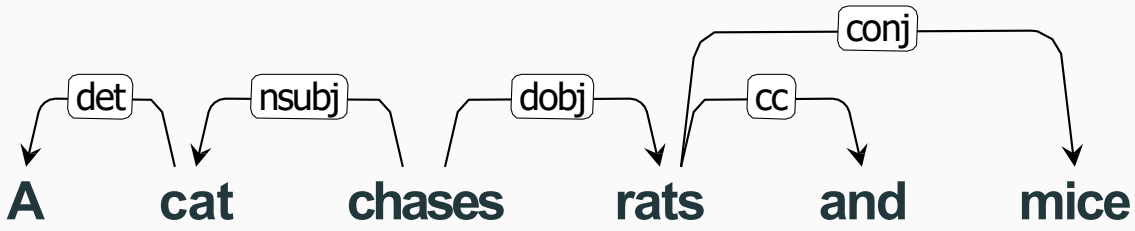
- http://universaldependencies.org/
- “cross-linguistically consistent treebank annotation for many languages” (129 treebanks for 76 languages in v. 2.3)
- available via LINDAT-CLARIN

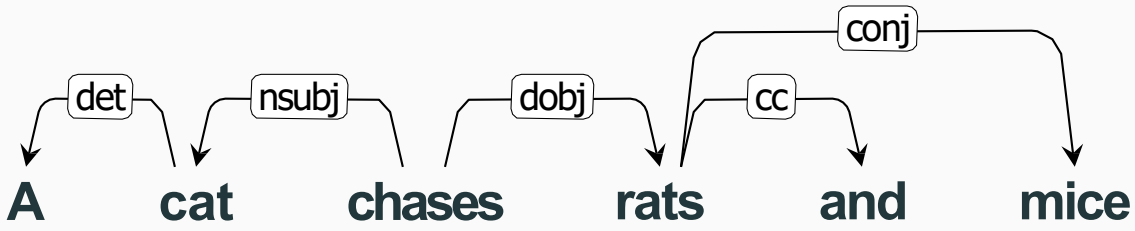
Slides based on those of the EACL Tutorial on Universal Dependencies
http://universaldependencies.org/eacl17tutorial/

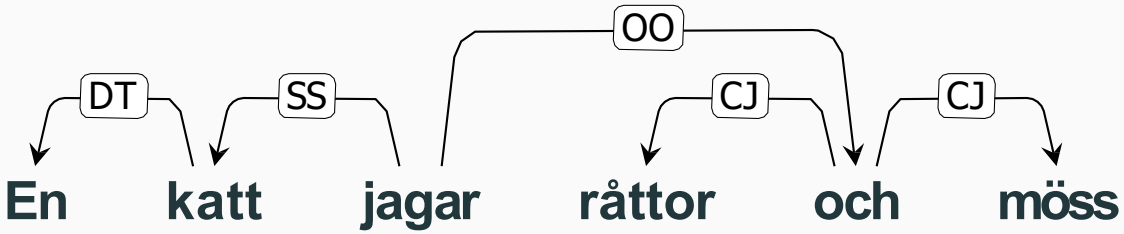
Why UDs?

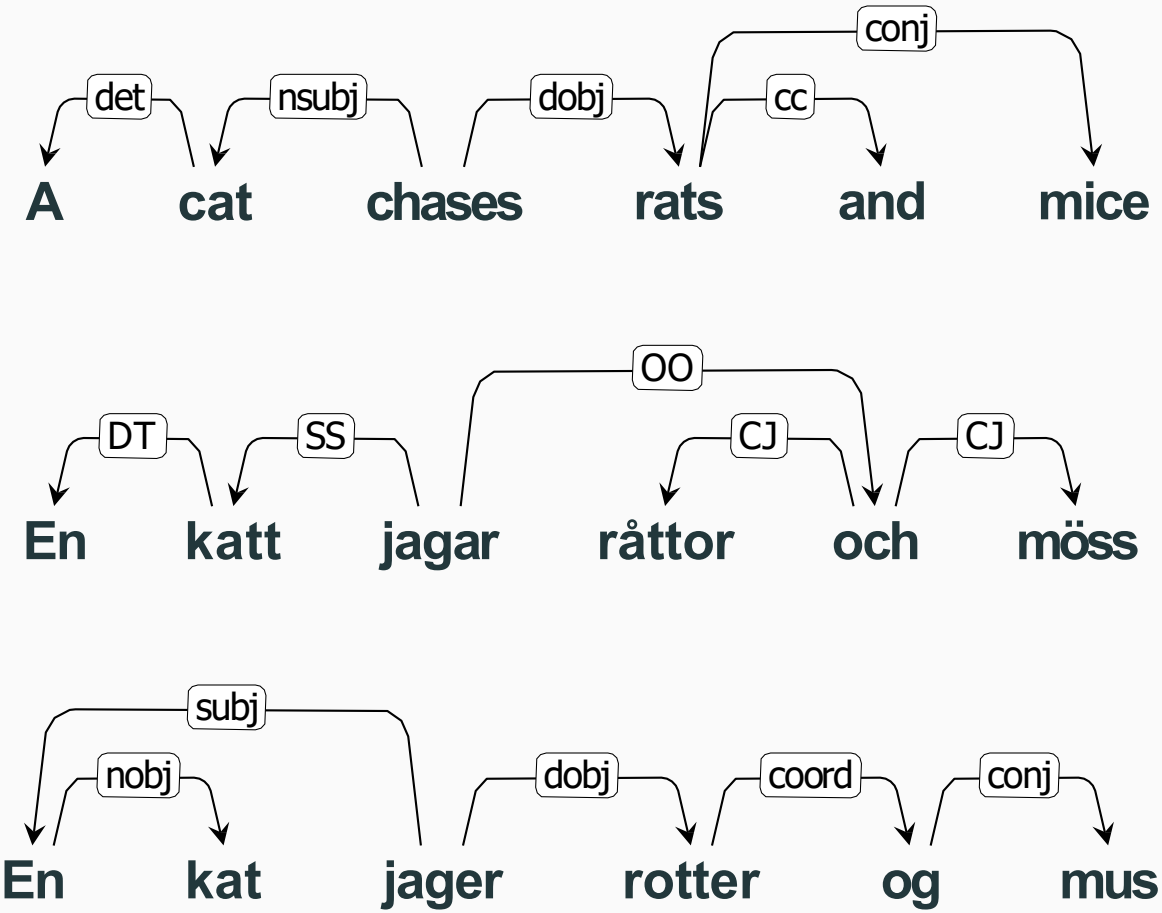
Increasing interest in multilingual NLP
- Multilingual evaluation campaigns to test generality
- Cross-lingual learning to support low-resource languages

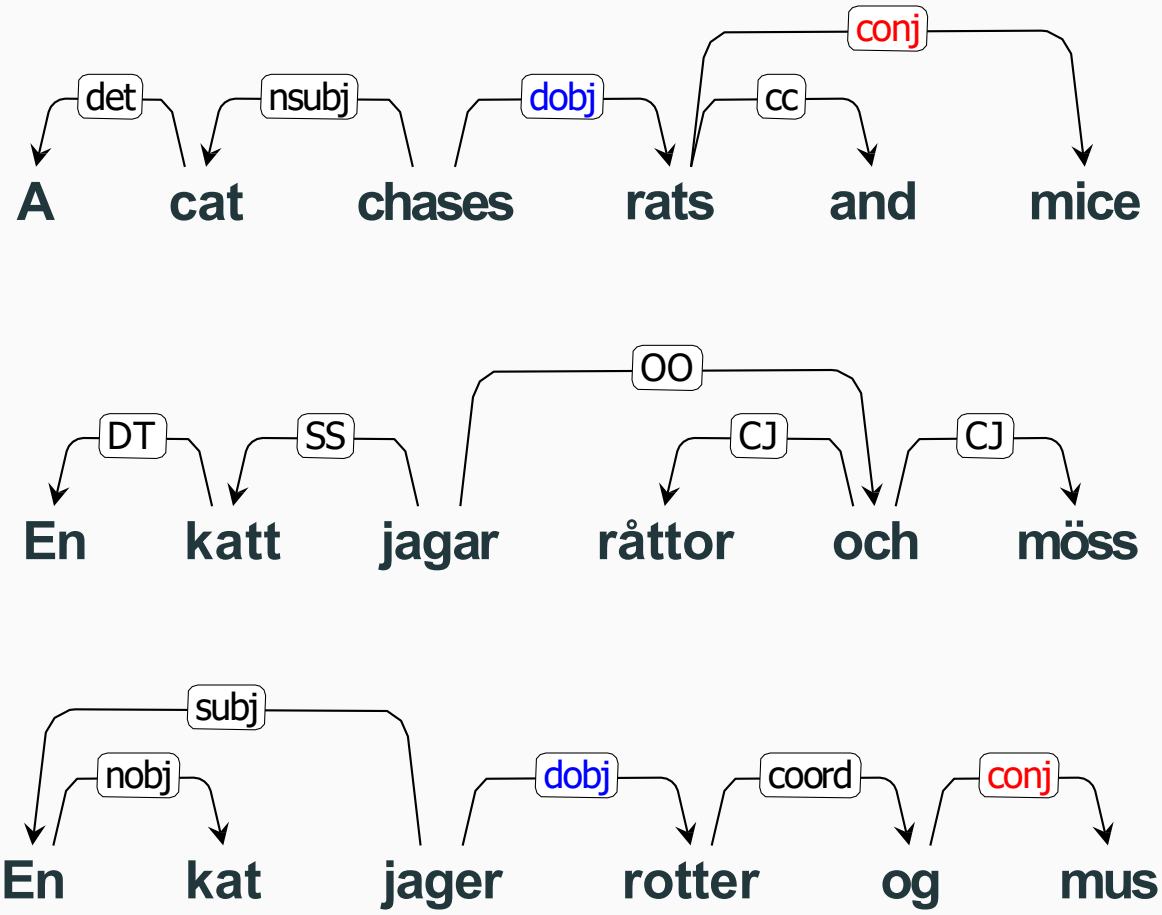
Increasing awareness of methodological problems
- Current NLP relies heavily on annotation
- Annotation schemes vary across languages


conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice


conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice


OO
DT
SS
CJ
CJ
En
katt
jagar
råttor
och
möss


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


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

Solution

- Cross-linguistically consistent grammatical annotation
- Build on common usage and existing de facto standards
- Complement – not replace – language-specific schemes
- Open community effort!

The UDs Philosophy

Merging existing initiatives to build consistent LRs:

- Interset Interlingua for morphosyntactic tagsets ← a tool for conversion between tagsets of multiple languages (2006 → HamleDT, 2013)
- Google Universal PoS tags ← cross-linguistic error analysis based on the CoNLL-X shared task data (2007)
- Stanford Dependencies: de facto standard for dependency analysis of English (2005)
- Google Dependency Scheme: first attempt to combine Stanford Dependencies & Google Universal PoS tags towards a universal annotation scheme (UDT, 2013) → Universal Stanford Dependencies (USD, 2014)

UD wants to replace all these with a single coherent standard

UDs Design Principles

- *Dependency*
  - –Widely used in practical NLP systems
  - –Available in treebanks for many languages
- *Lexicalism*
  - –Basic annotation units are words – syntactic words: clitics are split off (Spanish: *dámelo* = *dá me lo*) and contractions are undone (French *au* = *à le*)
  - –Words have morphological properties
  - –Words enter into syntactic relations
- *Recoverability*
  - –Syntactic wordhood does not always coincide with wehitespace-sperated orthographic units
  - –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	

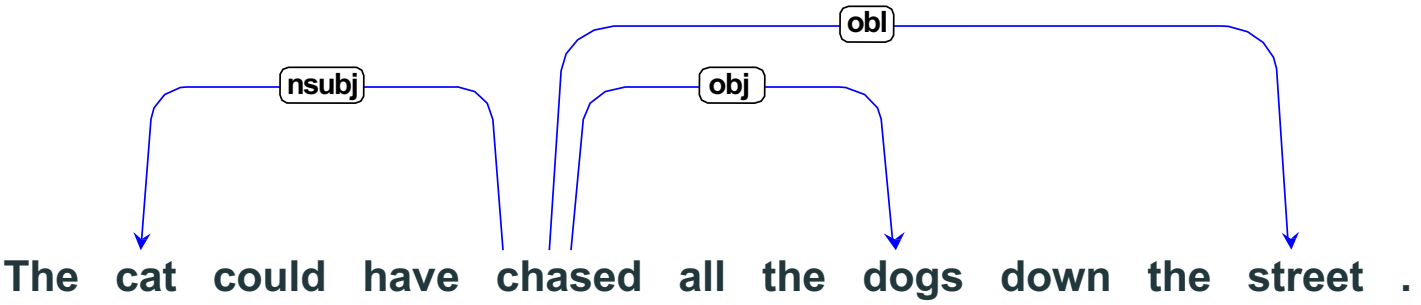
- Lemma
- Part-of-speech tag
- Features

Syntactic Annotation

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

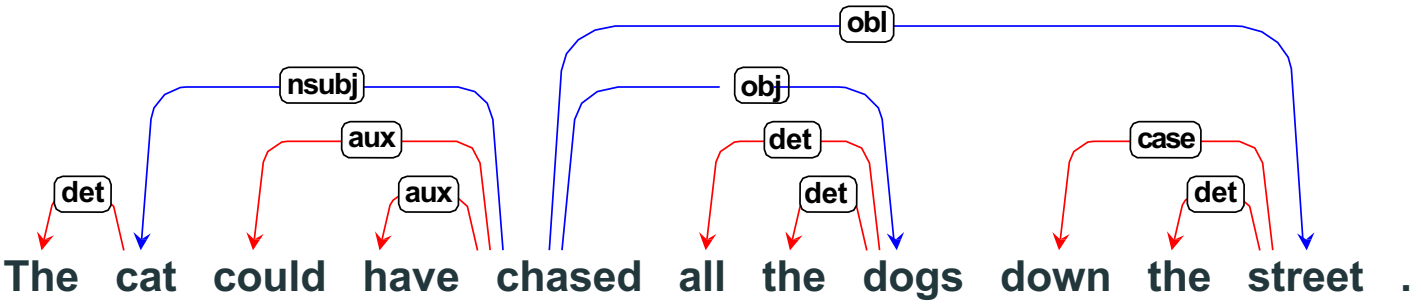
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



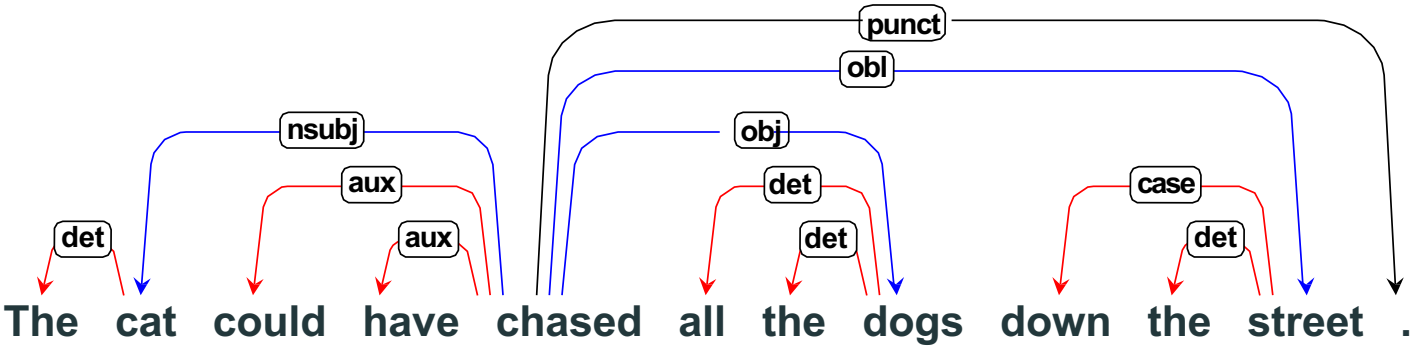
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

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

Different in UDs v2

```
# sent_id 1
# ...
1    They     they     PRON    PRP    Case=Nom|Number=Plur                2   nsubj    4:nsubj    _
2    buy      buy      VERB    VBP     Number=Plur|Person=3|Tense=Pres    0    root     _          _
3    and      and      CONJ    CC     _                                   2   cc       _          _
4    sell     sell     VERB    VBP     Number=Plur|Person=3|Tense=Pres    2    conj     0:root     _
5    books    book     NOUN    NNS    Number=Plur                         2   dobj     4:dobj     SpaceAfter=No
6    .        .        PUNCT   .      _                                   2   punct    _          _

# sent_id 2
# ...
1     I         I       PRON    PRP    Case=Nom|Number=Sing|Person=1       2   nsubj    _   _
2-3   haven't   _       _       _      _                                   _   _        _   _
2     have      have    VERB    VBP     Number=Sing|Person=1|Tense=Pres    0    root     _   _
3     not       not     PART    RB     Negative=Neg                        2   neg      _   _
4     a         a       DET     DT     Definite=Ind|PronType=Art           4   det      _   _
5     clue      clue    NOUN    NN     Number=Sing                         2   dobj     _   SpaceAfter=No
6     .         .       PUNCT   .      _                                   2   punct    _   _
```

CYCLE!: 4 instead of 5

Querying UDs
Introducing PML-TQ

PML-TQ

Query Elements

- Node-types:
  - `a-node / a-root`
  - `t-node / t-root`
- Attributes: `deprel, tag ...`
- Relations: `child, parent ...`
- Operators: `=, !=, > ...`
- Naming the nodes: `$a :=`
- Lists in output: `>> for…give…sort by…desc`

PML-TQ

http://lindat.mff.cuni.cz/services/pmltq/#!/home

Query examples. Treebank: UDLA-PROIEL 2.3

```
# Nominal Subjects
a-node [ deprel = 'nsubj' ]

# Word Order: Verb - Nominal Subject
a-node $a := [ deprel='nsubj', tag='NOUN',
parent a-node $b := [ tag='VERB', id < $a.id] ]

# Conjunct Nominal Subjects (list of couples)
a-node $a := [ deprel='nsubj', tag='NOUN',
child a-node $b := [ deprel='conj' ] ]
>> for $a.lemma, $b.lemma give $1, $2, count()
sort by $3 desc,$1,$2
```