# Universal Dependencies

Sara Stymne

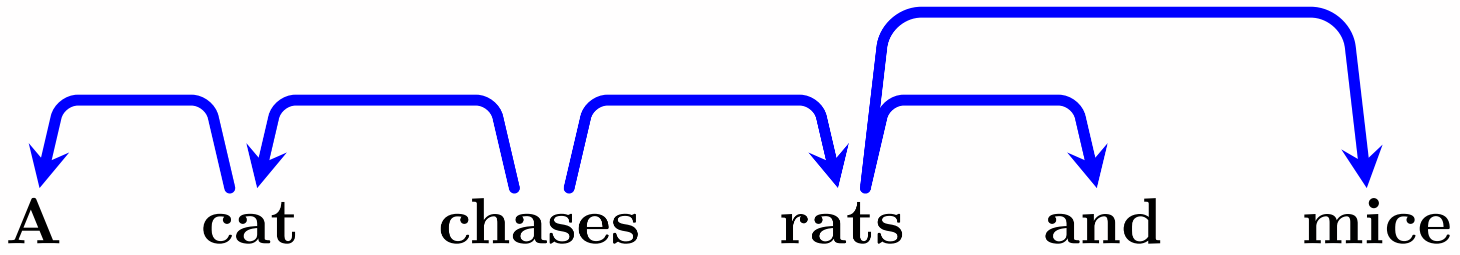
# Introduction

Growing interest in multilingual and cross-lingual NLP

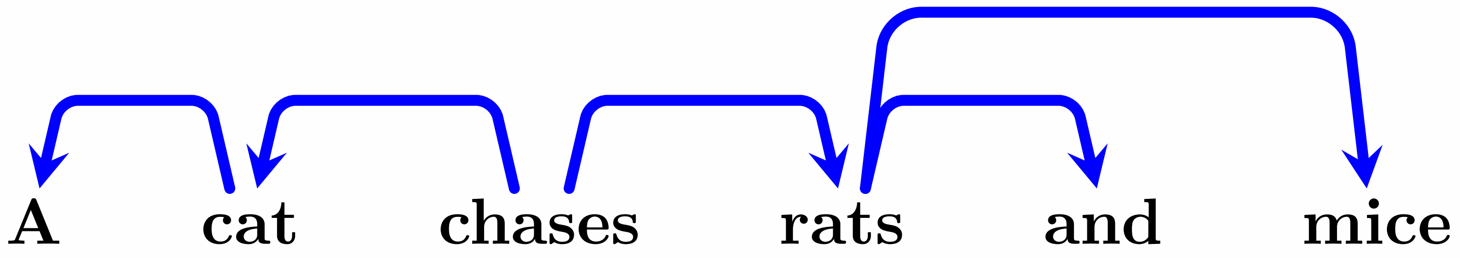
- Multilingual evaluation campaigns to test generality of approaches
- Cross-lingual learning to support low-resource languages

Growing awareness of methodological problems

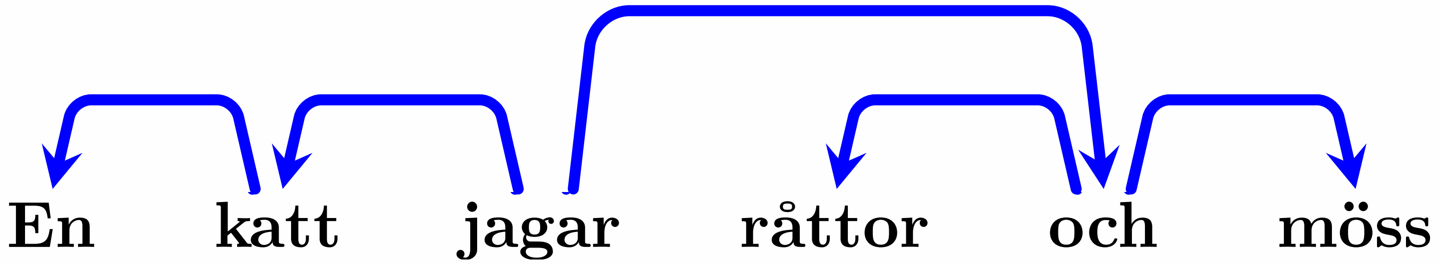
- Current NLP relies heavily on annotation
- Annotation guidelines vary across languages



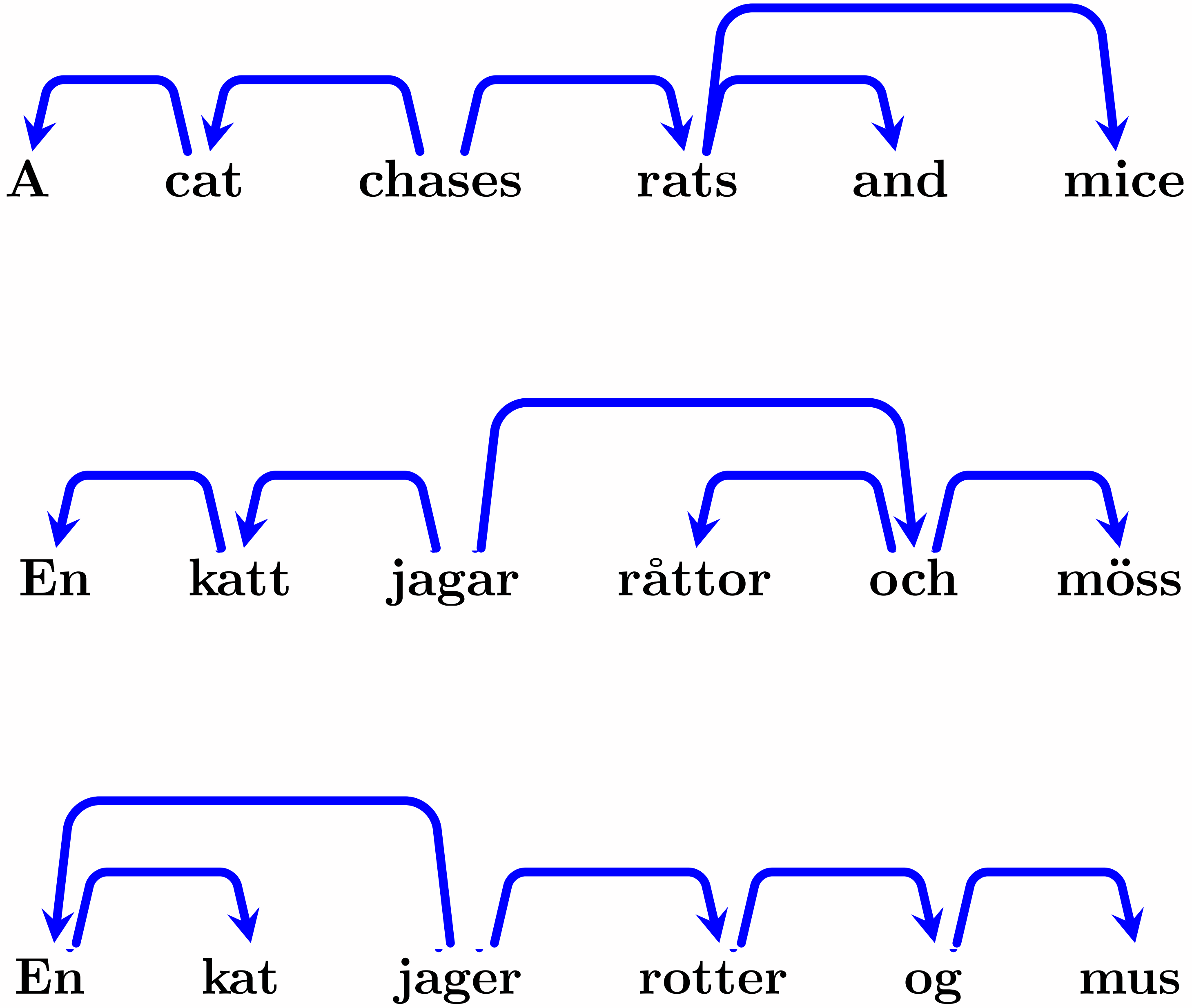
A
cat
chases
rats
and
mice



A cat chases rats and mice



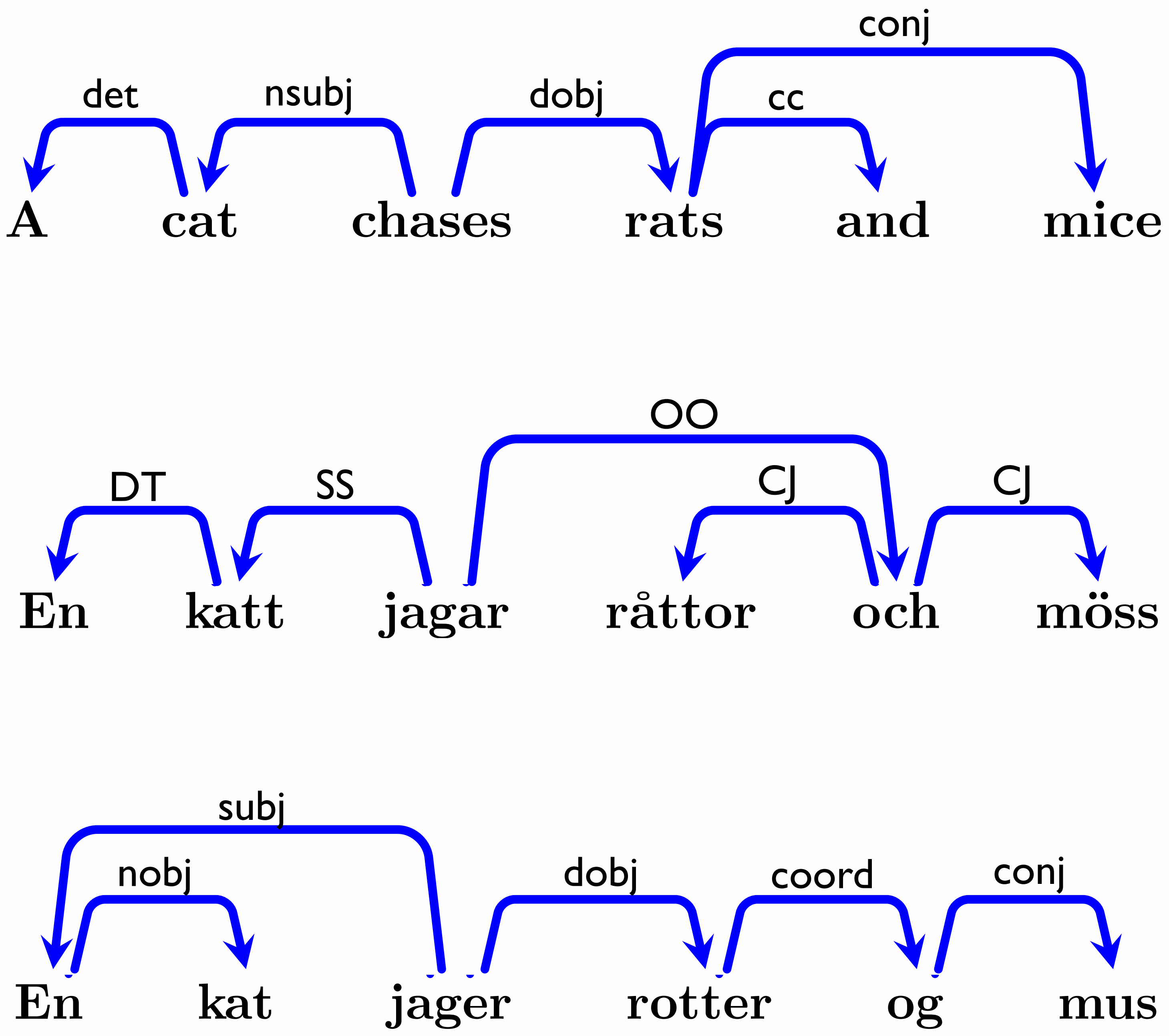
En katt jagar råttor och möss



A cat chases rats and mice

En katt jagar råttor och möss

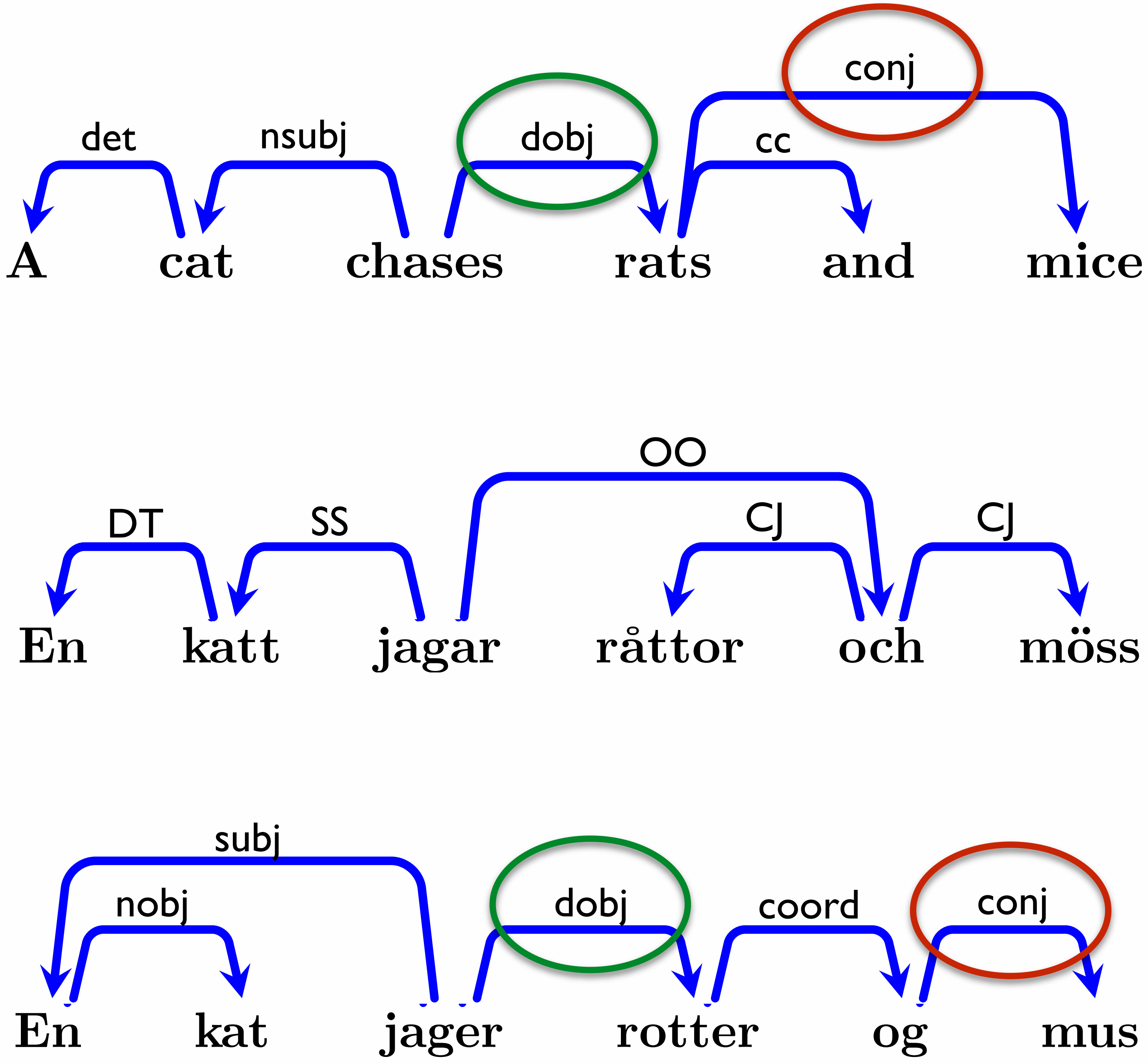
En kat jager rotter og mus



A cat chases rats and mice

En katt jagar råttor och möss

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?

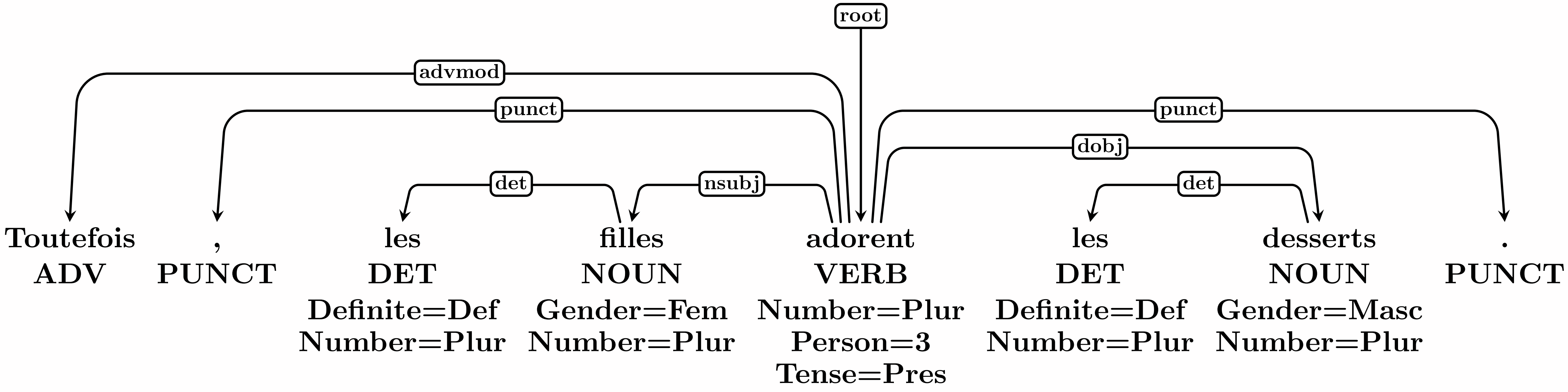
- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure transfer
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
- Hard to make progress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

# Universal Dependencies v2: A Multilingual Treebank Collection

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter
Yoav Goldberg, Jan Hajic, Christopher D. Manning
Ryan McDonald, Slav Petrov, Sampo Pyysalo
Natalia Silveira, Reut Tsarfaty, Daniel Zeman and many others

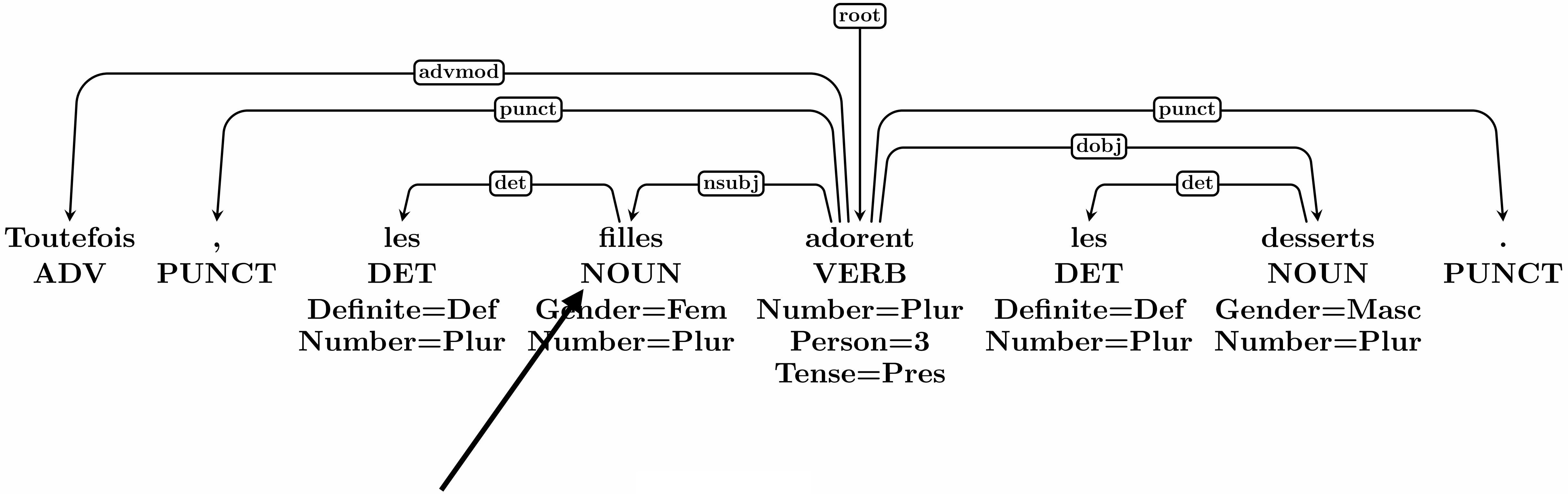
# Universal Dependencies

http://universaldependencies.org



# Universal Dependencies

http://universaldependencies.org

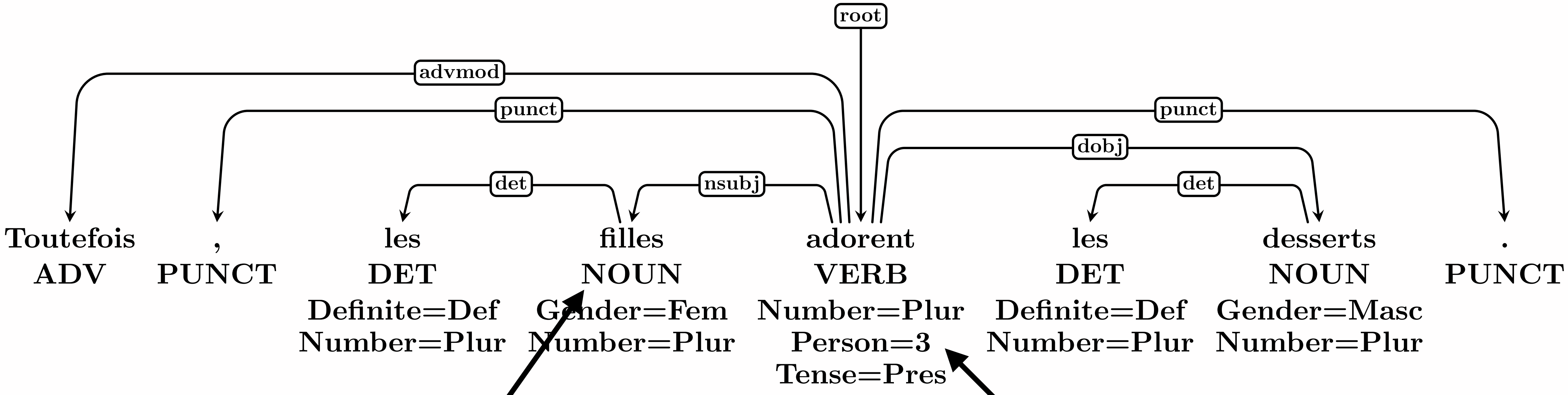


| Toutefois | , | les | filles | adorent | les | desserts | . |
|---|---|---|---|---|---|---|---|
| ADV | PUNCT | DET | NOUN | VERB | DET | NOUN | PUNCT |
| | | Definite=Def<br>Number=Plur | Gender=Fem<br>Number=Plur | Number=Plur<br>Person=3<br>Tense=Pres | Definite=Def<br>Number=Plur | Gender=Masc<br>Number=Plur | |

Part-of-speech tags Google

# Universal Dependencies

http://universaldependencies.org

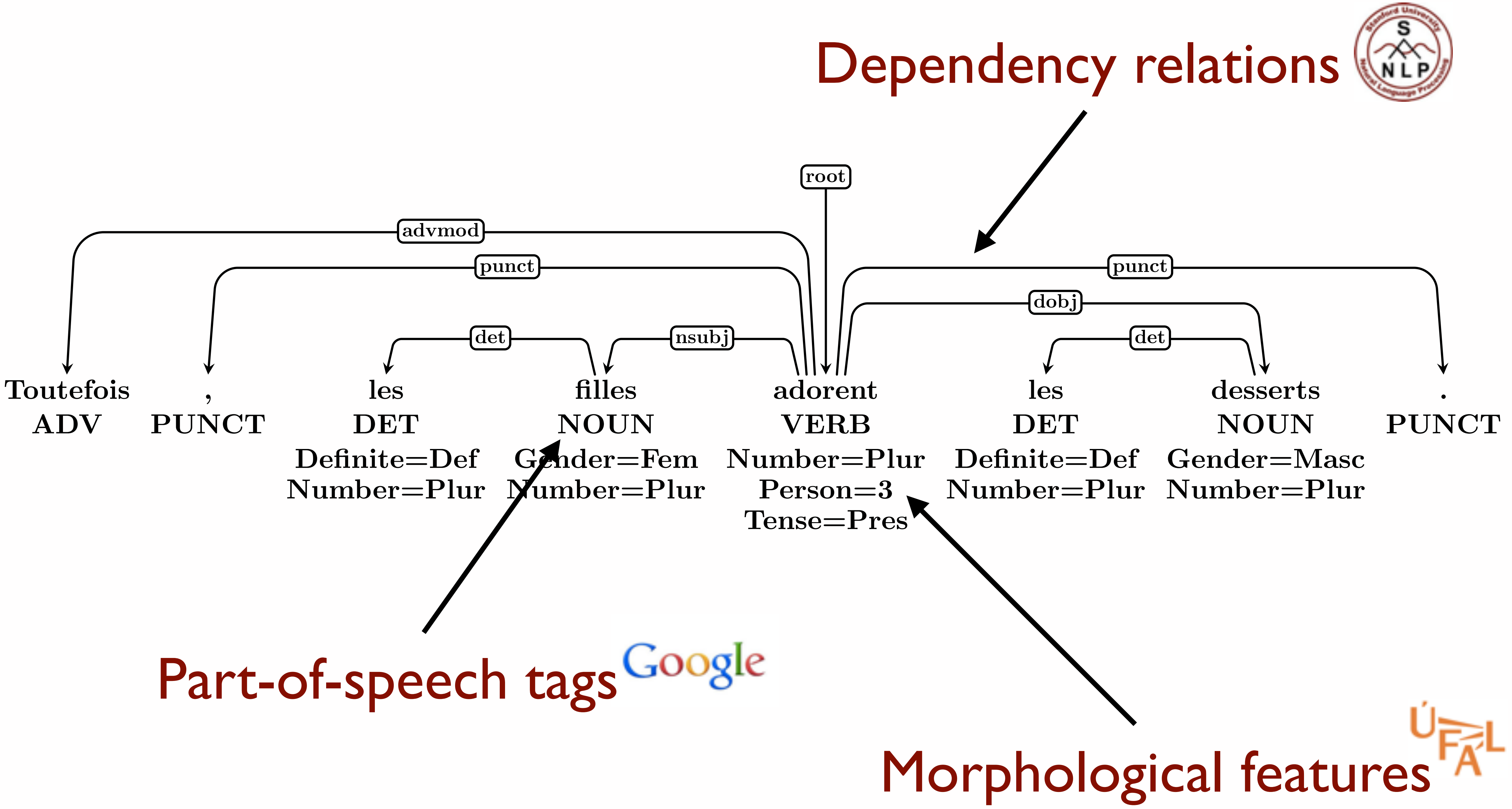

root
advmod
punct
punct
dobj
det
nsubj
det
Toutefois ADV
, PUNCT
les DET Definite=Def Number=Plur
filles NOUN Gender=Fem Number=Plur
adorent VERB Number=Plur Person=3 Tense=Pres
les DET Definite=Def Number=Plur
desserts NOUN Gender=Masc Number=Plur
. PUNCT

Part-of-speech tags

Morphological features

# Universal Dependencies

http://universaldependencies.org



Dependency relations

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Toutefois | , | les | filles | adorent | les | desserts | . |
| ADV | PUNCT | DET | NOUN | VERB | DET | NOUN | PUNCT |
| | | Definite=Def<br>Number=Plur | Gender=Fem<br>Number=Plur | Number=Plur<br>Person=3<br>Tense=Pres | Definite=Def<br>Number=Plur | Gender=Masc<br>Number=Plur | |

Part-of-speech tags

Morphological features

# 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

# Morphology

**Toutefois , les filles adorent les desserts .**

# Morphology

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Toutefois | , | les | filles | adorent | les | desserts | . |
| toutefois | , | le | fille | adorer | le | dessert | . |

- Lemma representing the semantic content of the word

# Morphology

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Toutefois | , | les | filles | adorent | les | desserts | . |
| toutefois | , | le | fille | adorer | le | dessert | . |
| ADV | PUNCT | DET | NOUN | VERB | DET | NOUN | PUNCT |

- Lemma representing the semantic content of the word
- Part-of-speech tag representing the abstract lexical category associated with the word

# Morphology

| | | | |
|---|---|---|---|
| Toutefois | , | desserts | . |
| toutefois | , | dessert | . |
| ADV | PUNCT | NOUN | PUNCT |

| Open | Closed | Other |
|---|---|---|
| ADJ | ADP | PUNCT |
| ADV | AUX | SYM |
| INTJ | CONJ | X |
| NOUN | DET | |
| PROPN | NUM | |
| VERB | PART | |
| | PRON | |
| | SCONJ | |

- Lemma rep of the word
- Part-of-speech tag representing the abstract lexical category associated with the word

# Morphology

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Toutefois | , | les | filles | adorent | les | desserts | . |
| toutefois | , | le | fille | adorer | le | dessert | . |
| ADV | PUNCT | DET | NOUN | VERB | DET | NOUN | PUNCT |
| | | Definite=Def<br>Number=Plur | Gender=Fem<br>Number=Plur | Number=Plur<br>Person=3<br>Tense=Pres | Definite=Def<br>Number=Plur | Gender=Masc<br>Number=Plur | |

- Lemma representing the semantic content of the word
- Part-of-speech tag representing the abstract lexical category associated with the word
- Features representing lexical and grammatical properties associated with the lemma or the particular word form

# Morphology

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Toutefois | , | les | filles | adorent | les | desserts | . |
| toutefois | , | le | fille | adorer | le | dessert | . |
| ADV | PUNCT | DET | NOUN | VERB | DET | NOUN | PUNCT |
| | | Definite=Def<br>Number=Plur | Gender=Fem<br>Number=Plur | Number=Plur<br>Person=3<br>Tense=Pres | Definite=Def<br>Number=Plur | Gender=Masc<br>Number=Plur | |

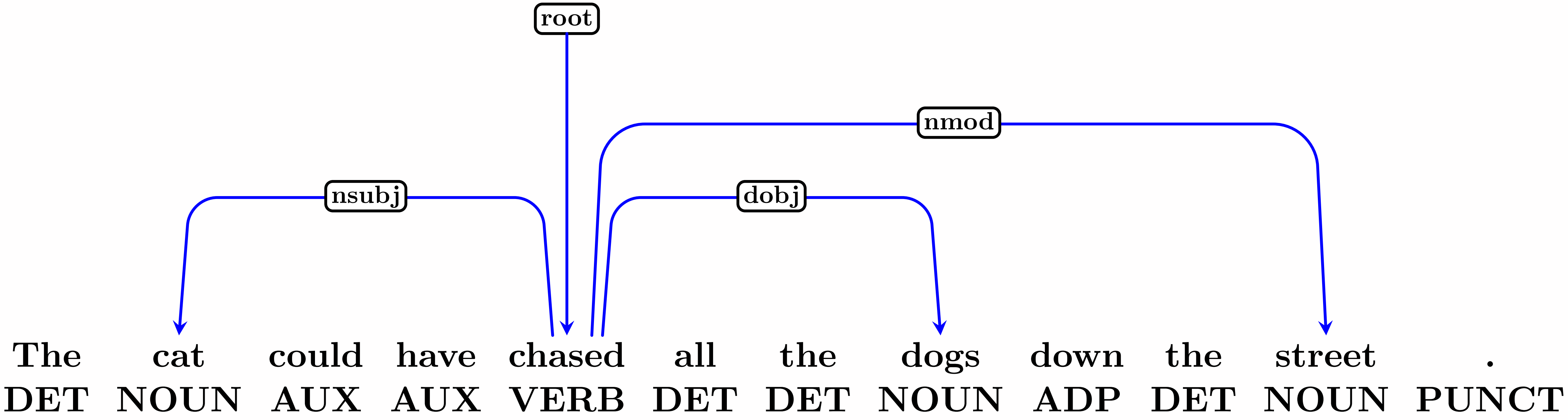
## Lexicalism

- Basic annotation units are words
- Morphological analysis through features, not segmentation
- Clitics and contractions are segmented if required by syntax

# Syntax

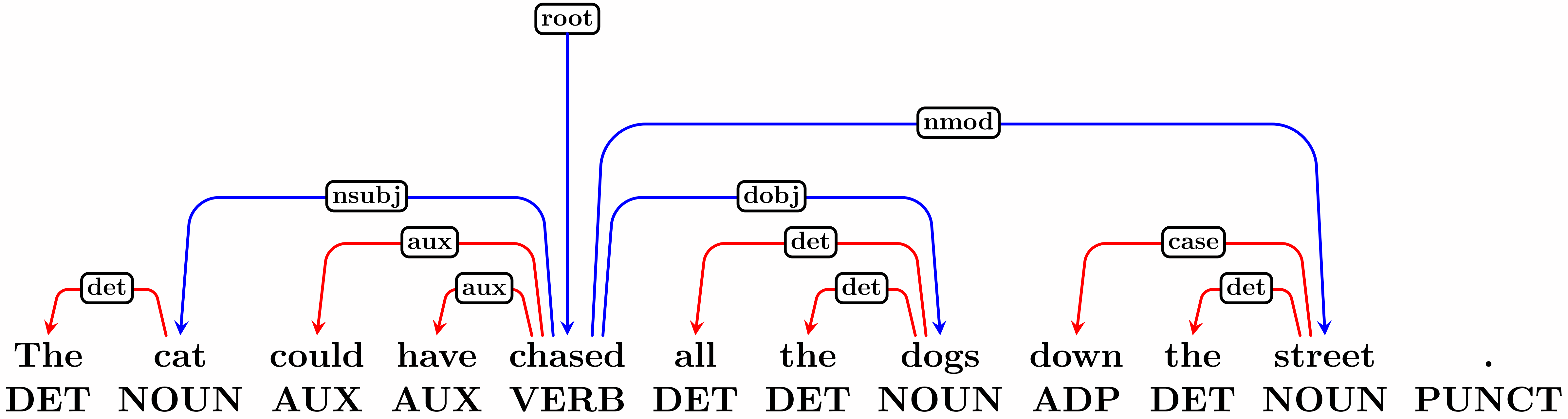
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **The** | **cat** | **could** | **have** | **chased** | **all** | **the** | **dogs** | **down** | **the** | **street** | **.** |
| **DET** | **NOUN** | **AUX** | **AUX** | **VERB** | **DET** | **DET** | **NOUN** | **ADP** | **DET** | **NOUN** | **PUNCT** |

# Syntax


root
nmod
nsubj
dobj
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT

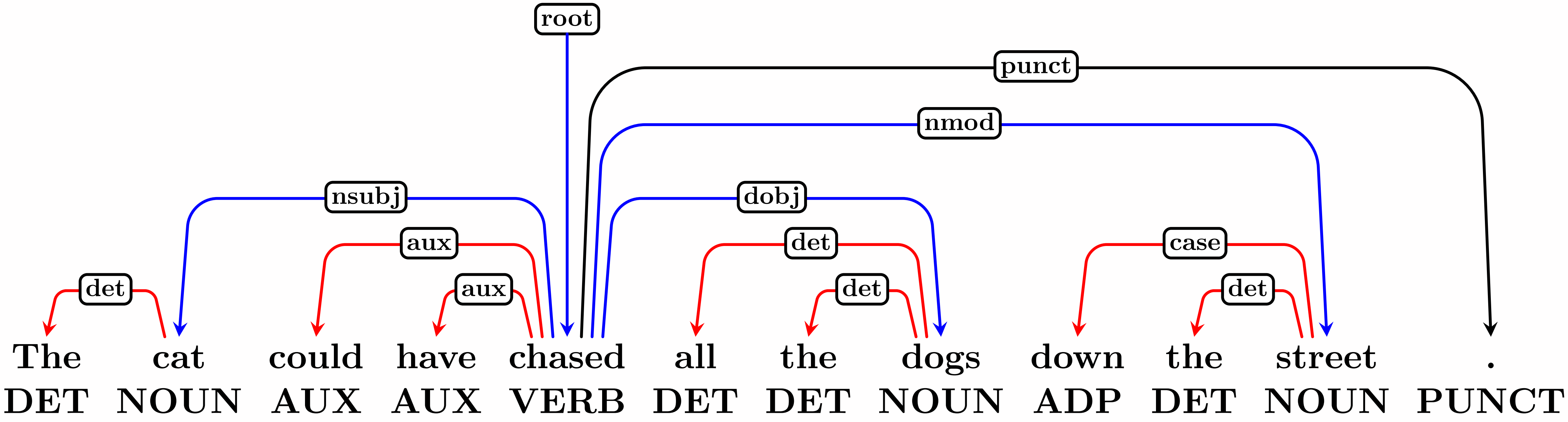
- Content words are related by dependency relations

# Syntax

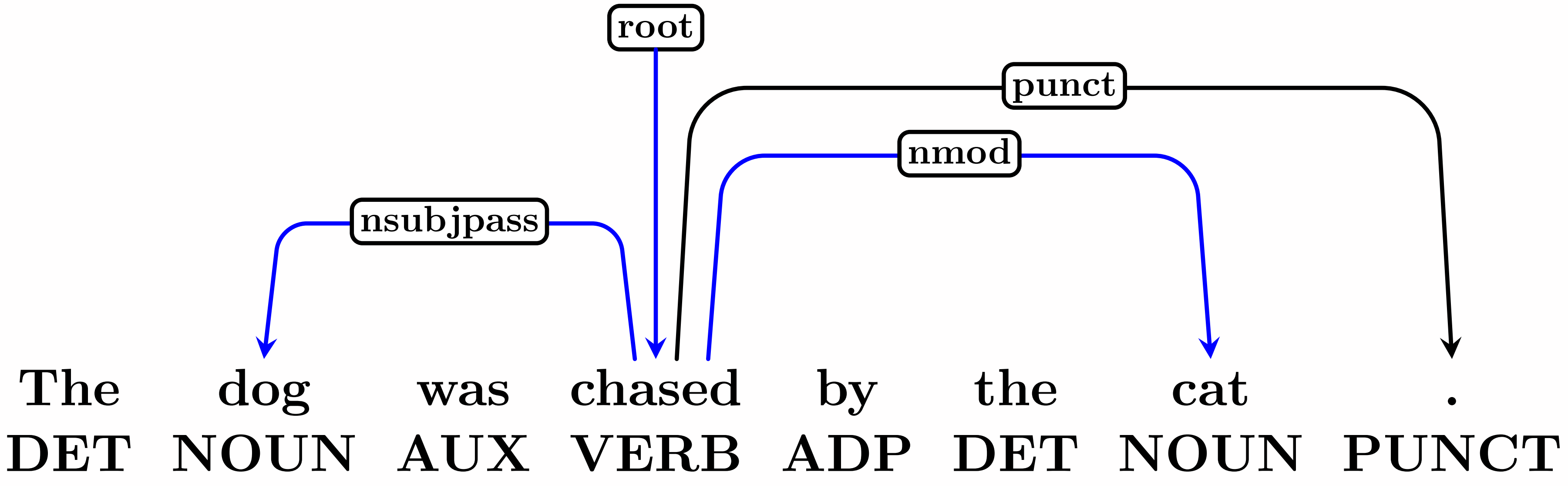


- Content words are related by dependency relations
- Function words attach to the content word they modify

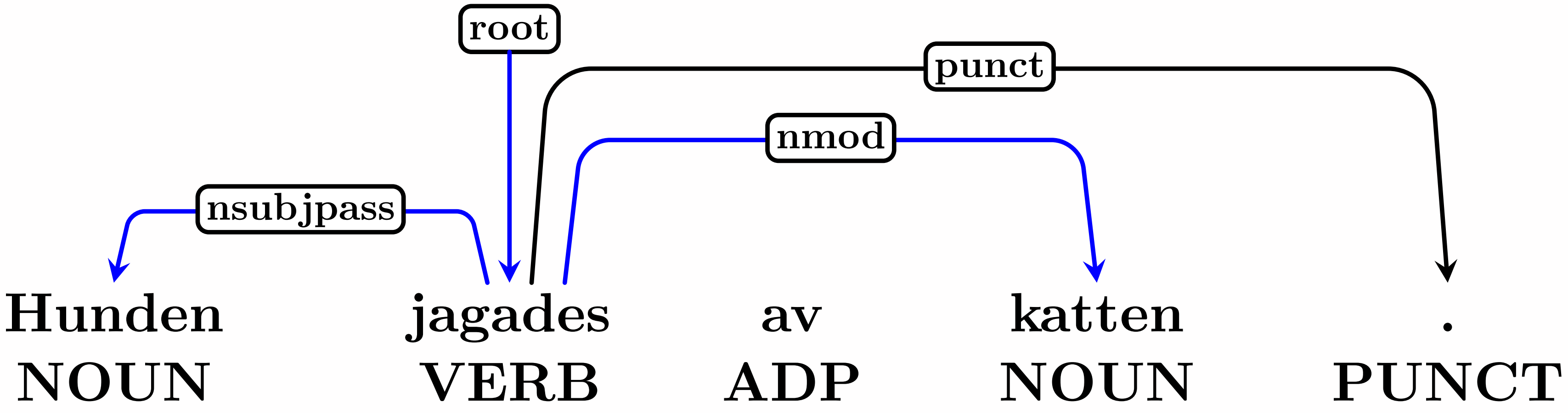
# Syntax


root
punct
nmod
nsubj
dobj
aux
det
case
aux
det
det
det
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT

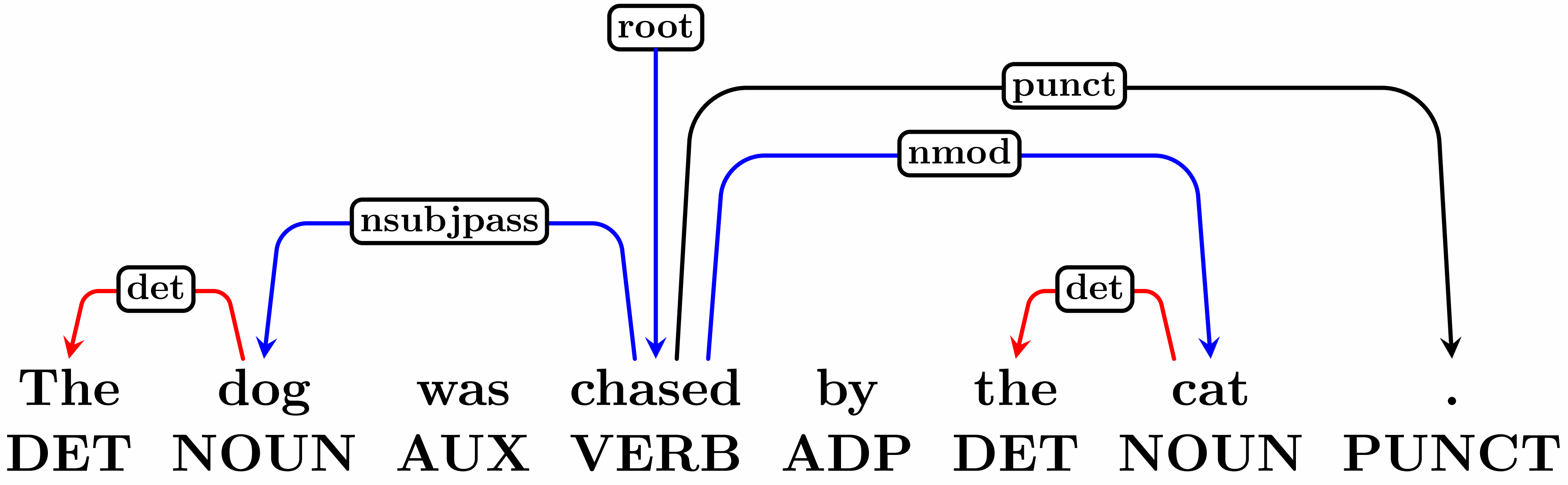
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause



The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT



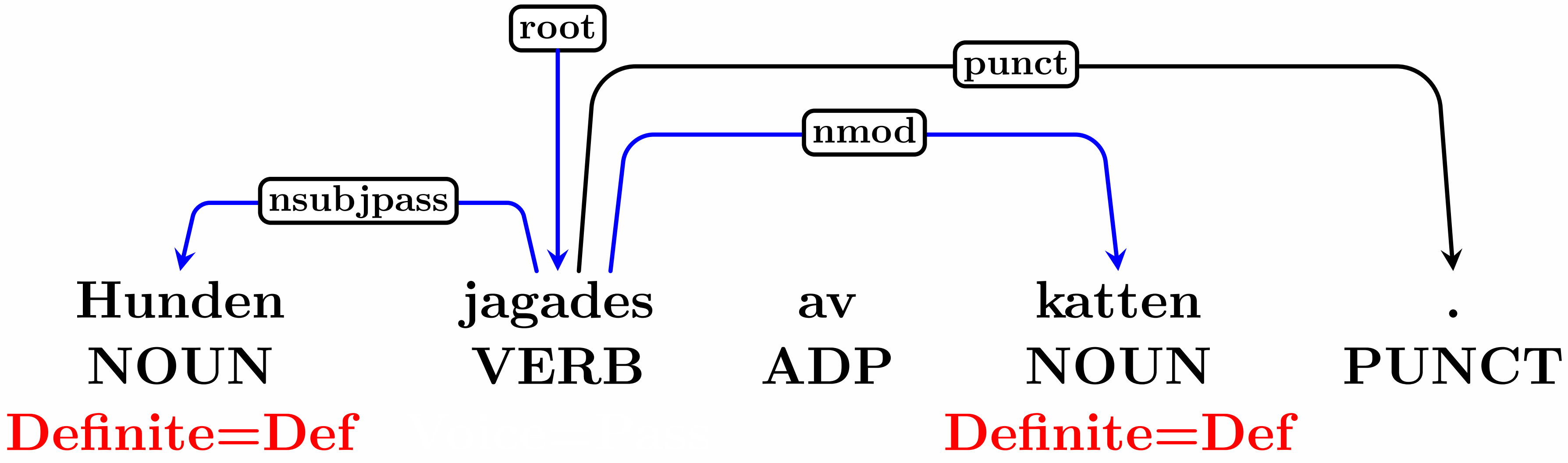
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT



root
punct
nmod
nsubjpass
det
det

The dog was chased by the cat .

DET NOUN AUX VERB ADP DET NOUN PUNCT

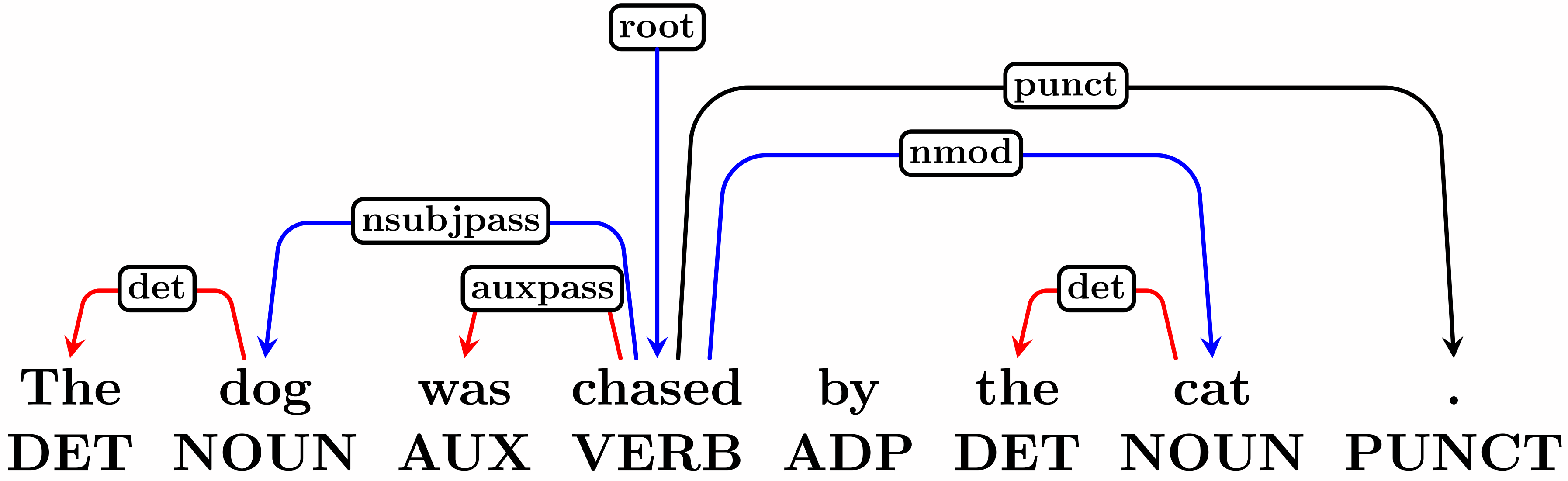


root
punct
nmod
nsubjpass

Hunden jagades av katten .

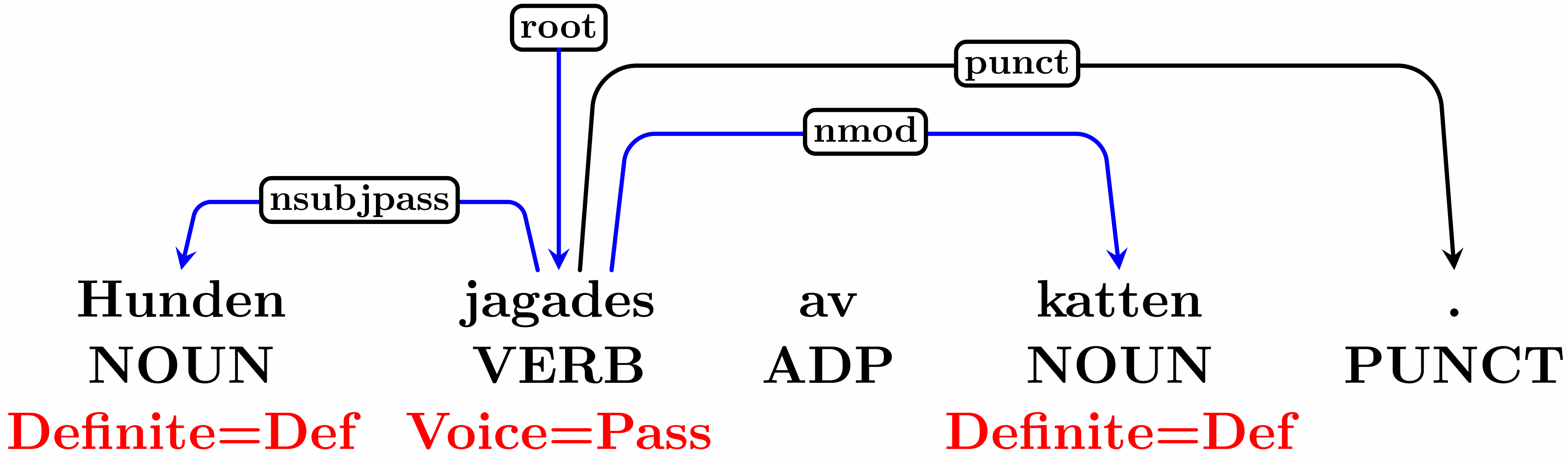
NOUN VERB ADP NOUN PUNCT

Definite=Def Definite=Def



The dog was chased by the cat .

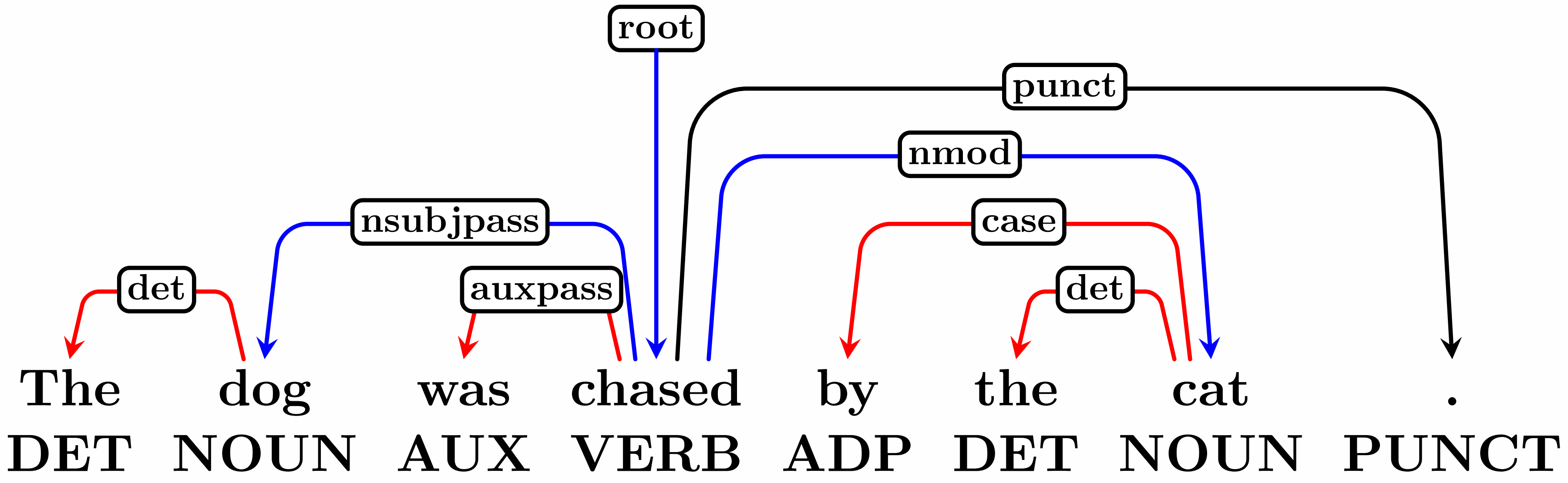
DET NOUN AUX VERB ADP DET NOUN PUNCT



Hunden jagades av katten .

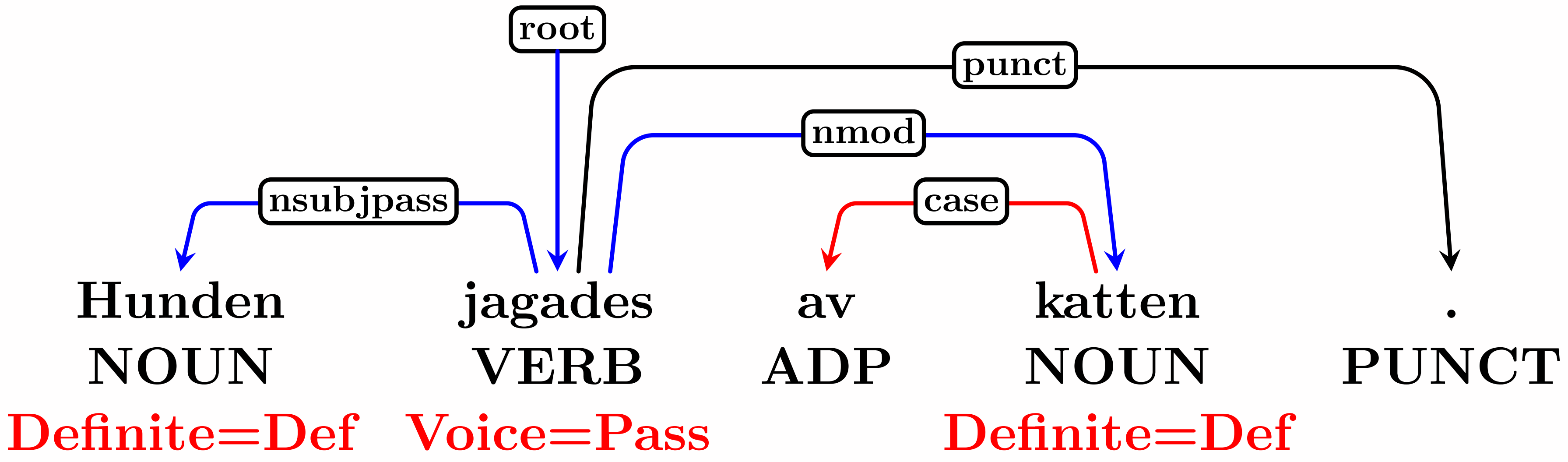
NOUN VERB ADP NOUN PUNCT

Definite=Def Voice=Pass Definite=Def



root
punct
nmod
nsubjpass
case
det
auxpass
det

The DET
dog NOUN
was AUX
chased VERB
by ADP
the DET
cat NOUN
. PUNCT



root
punct
nmod
nsubjpass
case

Hunden NOUN Definite=Def
jagades VERB Voice=Pass
av ADP
katten NOUN Definite=Def
. PUNCT

# Dependency Relations

## Taxonomy of 40 universal grammatical relations

- Three types of structures: nominals, clauses, modifiers
- Core arguments vs. other dependents (not complements vs. adjuncts)

## A two-level architecture

- Universal: broad categories to allow cross-linguistic comparison
- Language-specific: subtypes to capture language-specific phenomena

| **Universal** | **Subtype** |
|---|---|
| acl | acl:relcl |
| compound | compound:prt |
| nmod | nmod:poss |

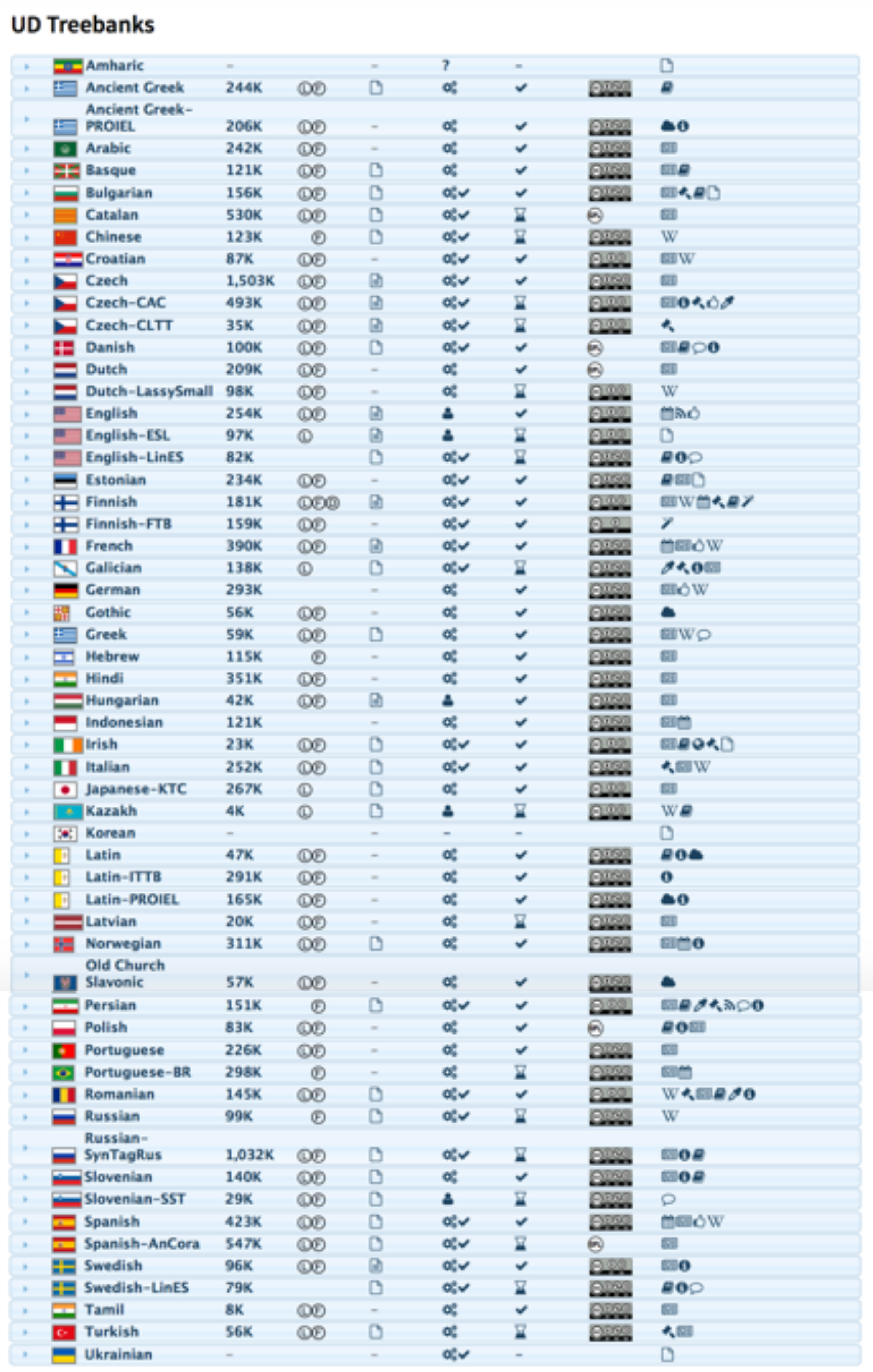
# Release v2

102 treebanks

61 languages

145 contributors

http://universaldependencies.org



## UD Treebanks

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Amharic | - | | - | ? | - | | 🗋 |
| Ancient Greek | 244K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | 📖 |
| Ancient Greek-PROIEL | 206K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁❶ |
| Arabic | 242K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▭ |
| Basque | 121K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▭📖 |
| Bulgarian | 156K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | ▭🔨📖🗋 |
| Catalan | 530K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | ○ | ▭ |
| Chinese | 123K | Ⓕ | 🗋 | ⚙✓ | ⌛ | 🅭 | W |
| Croatian | 87K | ⓁⒻ | - | ⚙✓ | ✓ | 🅭 | ▭W |
| Czech | 1,503K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | 🅭 | ▭ |
| Czech-CAC | 493K | ⓁⒻ | 🗎 | ⚙✓ | ⌛ | 🅭 | ▭❶🔨👍✎ |
| Czech-CLTT | 35K | ⓁⒻ | 🗎 | ⚙✓ | ⌛ | 🅭 | 🔨 |
| Danish | 100K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | ○ | ▭📖💬❶ |
| Dutch | 209K | ⓁⒻ | - | ⚙ | ✓ | ○ | ▭ |
| Dutch-LassySmall | 98K | ⓁⒻ | - | ⚙ | ⌛ | 🅭 | W |
| English | 254K | ⓁⒻ | 🗎 | 👤 | ✓ | 🅭 | 📅≋👍 |
| English-ESL | 97K | Ⓛ | 🗎 | 👤 | ⌛ | 🅭 | 🗋 |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⌛ | 🅭 | 📖❶💬 |
| Estonian | 234K | ⓁⒻ | - | ⚙✓ | ✓ | 🅭 | 📖▭🗋 |
| Finnish | 181K | ⓁⒻⒹ | 🗎 | ⚙✓ | ✓ | 🅭 | ▭W📅🔨📖✎ |
| Finnish-FTB | 159K | ⓁⒻ | - | ⚙✓ | ✓ | 🅭 | ✎ |
| French | 390K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | 🅭 | 📅▭👍W |
| Galician | 138K | Ⓛ | 🗋 | ⚙✓ | ⌛ | 🅭 | ✎🔨❶▭ |
| German | 293K | | - | ⚙ | ✓ | 🅭 | ▭👍W |
| Gothic | 56K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁ |
| Greek | 59K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▭W💬 |
| Hebrew | 115K | Ⓕ | - | ⚙ | ✓ | 🅭 | ▭ |
| Hindi | 351K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▭ |
| Hungarian | 42K | ⓁⒻ | 🗎 | 👤 | ✓ | 🅭 | ▭ |
| Indonesian | 121K | | - | ⚙ | ✓ | 🅭 | ▭📅 |
| Irish | 23K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | ▭📖♁🔨🗋 |
| Italian | 252K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | 🔨▭W |
| Japanese-KTC | 267K | Ⓛ | 🗋 | ⚙ | ✓ | 🅭 | ▭ |
| Kazakh | 4K | Ⓛ | 🗋 | 👤 | ⌛ | 🅭 | W📖 |
| Korean | - | | - | - | - | | 🗋 |
| Latin | 47K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | 📖❶☁ |
| Latin-ITTB | 291K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ❶ |
| Latin-PROIEL | 165K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁❶ |
| Latvian | 20K | ⓁⒻ | - | ⚙ | ⌛ | 🅭 | ▭ |
| Norwegian | 311K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▭📅❶ |
| Old Church Slavonic | 57K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁ |
| Persian | 151K | Ⓕ | 🗋 | ⚙✓ | ✓ | 🅭 | ▭📖✎🔨≋💬❶ |
| Polish | 83K | ⓁⒻ | - | ⚙ | ✓ | ○ | 📖❶▭ |
| Portuguese | 226K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▭ |
| Portuguese-BR | 298K | Ⓕ | - | ⚙ | ⌛ | 🅭 | ▭📅 |
| Romanian | 145K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | W🔨▭📖✎❶ |
| Russian | 99K | Ⓕ | 🗋 | ⚙✓ | ⌛ | 🅭 | W |
| Russian-SynTagRus | 1,032K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | 🅭 | ▭❶📖 |
| Slovenian | 140K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▭❶📖 |
| Slovenian-SST | 29K | ⓁⒻ | 🗋 | 👤 | ⌛ | 🅭 | 💬 |
| Spanish | 423K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | 📅▭👍W |
| Spanish-AnCora | 547K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | ○ | ▭ |
| Swedish | 96K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | 🅭 | ▭❶ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⌛ | 🅭 | 📖❶💬 |
| Tamil | 8K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▭ |
| Turkish | 56K | ⓁⒻ | 🗋 | ⚙ | ⌛ | 🅭 | 🔨▭ |
| Ukrainian | - | | - | ⚙✓ | - | | 🗋 |

## Release v2

102 treebanks

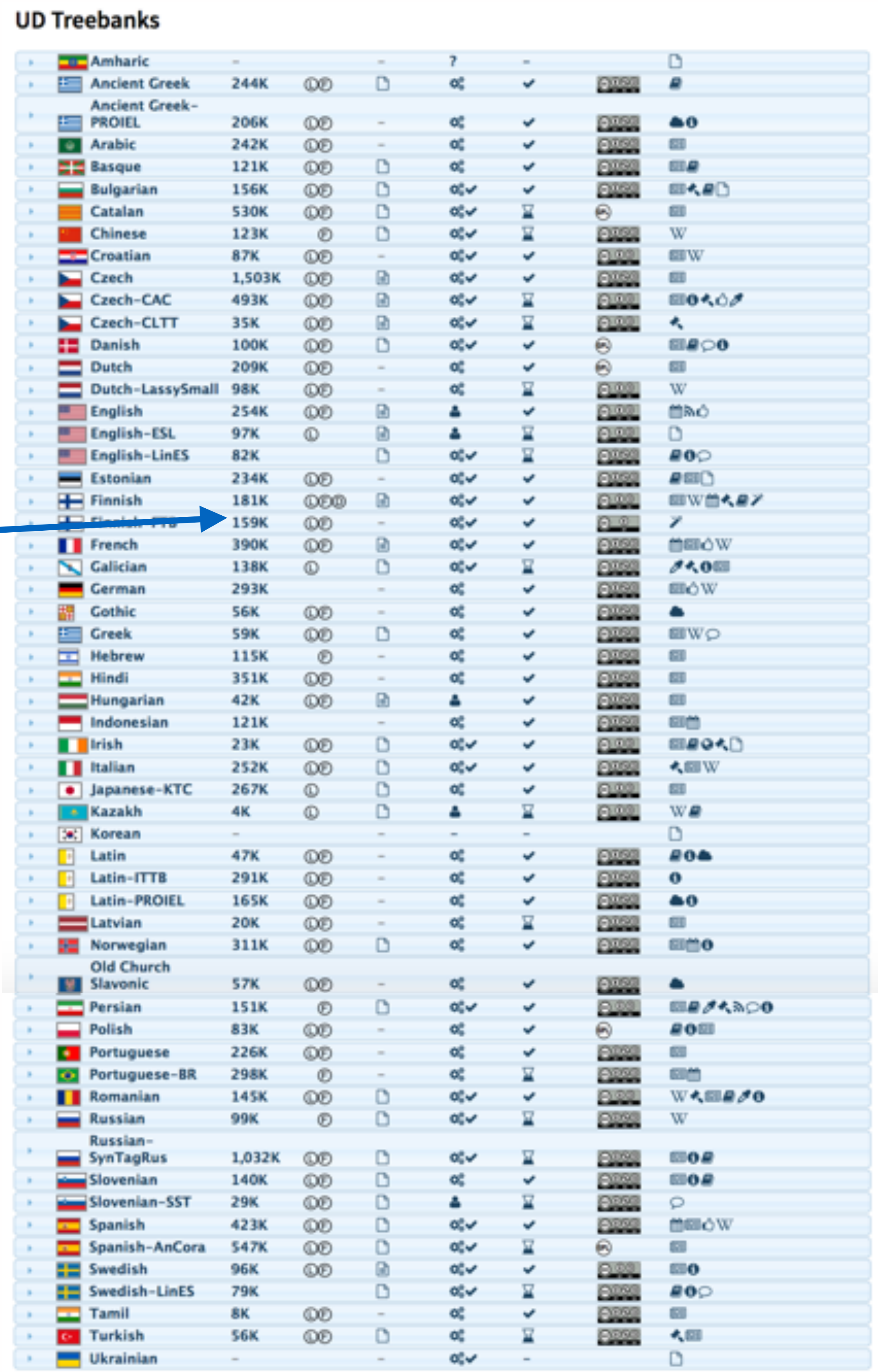
61 languages

145 contributors

http://universaldependencies.org

## Size:

<1K to 1,5M tokens



### UD Treebanks

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Amharic | – | | – | ? | – | | 🗋 |
| Ancient Greek | 244K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | 📘 |
| Ancient Greek-PROIEL | 206K | ⓁⒻ | – | ⚙ | ✓ | CC | ☁ⓘ |
| Arabic | 242K | ⓁⒻ | – | ⚙ | ✓ | CC | ▤ |
| Basque | 121K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤📘 |
| Bulgarian | 156K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤⚒📘🗋 |
| Catalan | 530K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ | ○ | ▤ |
| Chinese | 123K | Ⓕ | 🗋 | ⚙✓ | ⧗ | CC | W |
| Croatian | 87K | ⓁⒻ | – | ⚙✓ | ✓ | CC | ▤W |
| Czech | 1,503K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | CC | ▤ |
| Czech-CAC | 493K | ⓁⒻ | 🗎 | ⚙✓ | ⧗ | CC | ▤ⓘ⚒👍✐ |
| Czech-CLTT | 35K | ⓁⒻ | 🗎 | ⚙✓ | ⧗ | CC | ⚒ |
| Danish | 100K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | ○ | ▤📘💬ⓘ |
| Dutch | 209K | ⓁⒻ | – | ⚙ | ✓ | ○ | ▤ |
| Dutch-LassySmall | 98K | ⓁⒻ | – | ⚙ | ⧗ | CC | W |
| English | 254K | ⓁⒻ | 🗎 | 👤 | ✓ | CC | 📅📡👍 |
| English-ESL | 97K | Ⓛ | 🗎 | 👤 | ⧗ | CC | 🗋 |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⧗ | CC | 📘ⓘ💬 |
| Estonian | 234K | ⓁⒻ | – | ⚙✓ | ✓ | CC | 📘▤🗋 |
| Finnish | 181K | ⓁⒻⒹ | 🗎 | ⚙✓ | ✓ | CC | ▤W📅⚒📘✐ |
| [illegible] | 159K | ⓁⒻ | – | ⚙✓ | ✓ | CC | ✐ |
| French | 390K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | CC | 📅▤👍W |
| Galician | 138K | Ⓛ | 🗋 | ⚙✓ | ⧗ | CC | ✐⚒ⓘ▤ |
| German | 293K | | – | ⚙ | ✓ | CC | ▤👍W |
| Gothic | 56K | ⓁⒻ | – | ⚙ | ✓ | CC | ☁ |
| Greek | 59K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤W💬 |
| Hebrew | 115K | Ⓕ | – | ⚙ | ✓ | CC | ▤ |
| Hindi | 351K | ⓁⒻ | – | ⚙ | ✓ | CC | ▤ |
| Hungarian | 42K | ⓁⒻ | 🗎 | 👤 | ✓ | CC | ▤ |
| Indonesian | 121K | | – | ⚙ | ✓ | CC | ▤📅 |
| Irish | 23K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤📘🌐⚒🗋 |
| Italian | 252K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ⚒▤W |
| Japanese-KTC | 267K | Ⓛ | 🗋 | ⚙ | ✓ | CC | ▤ |
| Kazakh | 4K | Ⓛ | 🗋 | 👤 | ⧗ | CC | W📘 |
| Korean | – | | – | – | – | | 🗋 |
| Latin | 47K | ⓁⒻ | – | ⚙ | ✓ | CC | 📘ⓘ☁ |
| Latin-ITTB | 291K | ⓁⒻ | – | ⚙ | ✓ | CC | ⓘ |
| Latin-PROIEL | 165K | ⓁⒻ | – | ⚙ | ✓ | CC | ☁ⓘ |
| Latvian | 20K | ⓁⒻ | – | ⚙ | ⧗ | CC | ▤ |
| Norwegian | 311K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤📅ⓘ |
| Old Church Slavonic | 57K | ⓁⒻ | – | ⚙ | ✓ | CC | ☁ |
| Persian | 151K | Ⓕ | 🗋 | ⚙✓ | ✓ | CC | ▤📘✐⚒📡💬ⓘ |
| Polish | 83K | ⓁⒻ | – | ⚙ | ✓ | ○ | 📘ⓘ▤ |
| Portuguese | 226K | ⓁⒻ | – | ⚙ | ✓ | CC | ▤ |
| Portuguese-BR | 298K | Ⓕ | – | ⚙ | ⧗ | CC | ▤📅 |
| Romanian | 145K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | W⚒▤📘✐ⓘ |
| Russian | 99K | Ⓕ | 🗋 | ⚙✓ | ⧗ | CC | W |
| Russian-SynTagRus | 1,032K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ | CC | ▤ⓘ📘 |
| Slovenian | 140K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤ⓘ📘 |
| Slovenian-SST | 29K | ⓁⒻ | 🗋 | 👤 | ⧗ | CC | 💬 |
| Spanish | 423K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | 📅▤👍W |
| Spanish-AnCora | 547K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ | ○ | ▤ |
| Swedish | 96K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | CC | ▤ⓘ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⧗ | CC | 📘ⓘ💬 |
| Tamil | 8K | ⓁⒻ | – | ⚙ | ✓ | CC | ▤ |
| Turkish | 56K | ⓁⒻ | 🗋 | ⚙ | ⧗ | CC | ⚒▤ |
| Ukrainian | – | | – | ⚙✓ | – | | 🗋 |

# Release v2

102 treebanks

61 languages

145 contributors

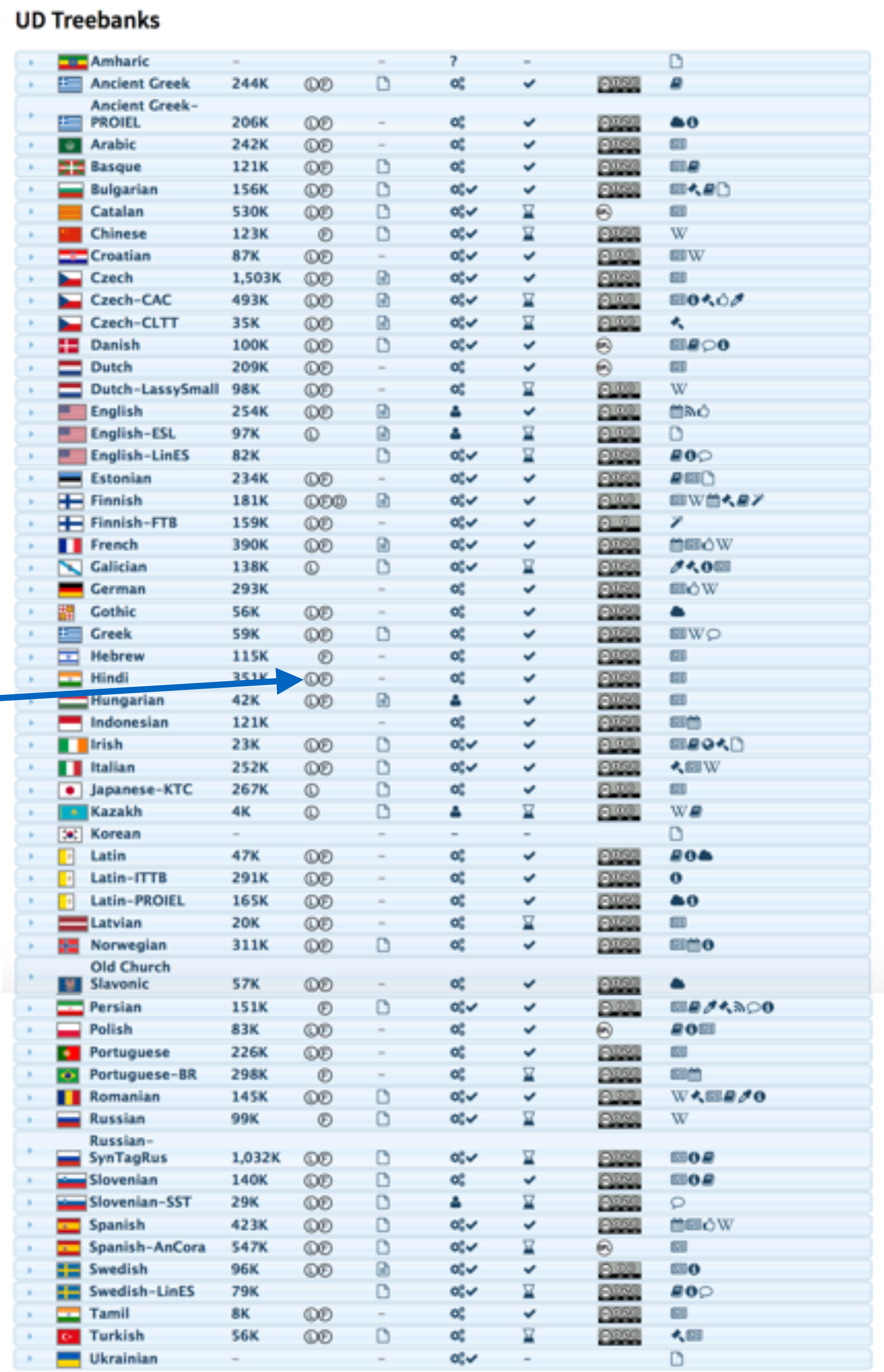
http://universaldependencies.org

## Size:

<1K to 1,5M tokens

## Annotation:

lemmas (83), features (94)



UD Treebanks

| | | | | | |
|---|---|---|---|---|---|
| Amharic | - | | - | ? | - |
| Ancient Greek | 244K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Ancient Greek-PROIEL | 206K | ⓁⒻ | - | ⚙ | ✓ |
| Arabic | 242K | ⓁⒻ | - | ⚙ | ✓ |
| Basque | 121K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Bulgarian | 156K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Catalan | 530K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ |
| Chinese | 123K | Ⓕ | 🗋 | ⚙✓ | ⧗ |
| Croatian | 87K | ⓁⒻ | - | ⚙✓ | ✓ |
| Czech | 1,503K | ⓁⒻ | 🗎 | ⚙✓ | ✓ |
| Czech-CAC | 493K | ⓁⒻ | 🗎 | ⚙✓ | ⧗ |
| Czech-CLTT | 35K | ⓁⒻ | 🗎 | ⚙✓ | ⧗ |
| Danish | 100K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Dutch | 209K | ⓁⒻ | - | ⚙ | ✓ |
| Dutch-LassySmall | 98K | ⓁⒻ | - | ⚙ | ⧗ |
| English | 254K | ⓁⒻ | 🗎 | 👤 | ✓ |
| English-ESL | 97K | Ⓛ | 🗎 | 👤 | ⧗ |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⧗ |
| Estonian | 234K | ⓁⒻ | - | ⚙✓ | ✓ |
| Finnish | 181K | ⓁⒻⒹ | 🗎 | ⚙✓ | ✓ |
| Finnish-FTB | 159K | ⓁⒻ | - | ⚙✓ | ✓ |
| French | 390K | ⓁⒻ | 🗎 | ⚙✓ | ✓ |
| Galician | 138K | Ⓛ | 🗋 | ⚙✓ | ⧗ |
| German | 293K | | - | ⚙ | ✓ |
| Gothic | 56K | ⓁⒻ | - | ⚙ | ✓ |
| Greek | 59K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Hebrew | 115K | Ⓕ | - | ⚙ | ✓ |
| Hindi | 351K | ⓁⒻ | - | ⚙ | ✓ |
| Hungarian | 42K | ⓁⒻ | 🗎 | 👤 | ✓ |
| Indonesian | 121K | | - | ⚙ | ✓ |
| Irish | 23K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Italian | 252K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Japanese-KTC | 267K | Ⓛ | 🗋 | ⚙ | ✓ |
| Kazakh | 4K | Ⓛ | 🗋 | 👤 | ⧗ |
| Korean | - | | - | - | - |
| Latin | 47K | ⓁⒻ | - | ⚙ | ✓ |
| Latin-ITTB | 291K | ⓁⒻ | - | ⚙ | ✓ |
| Latin-PROIEL | 165K | ⓁⒻ | - | ⚙ | ✓ |
| Latvian | 20K | ⓁⒻ | - | ⚙ | ⧗ |
| Norwegian | 311K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Old Church Slavonic | 57K | ⓁⒻ | - | ⚙ | ✓ |
| Persian | 151K | Ⓕ | 🗋 | ⚙✓ | ✓ |
| Polish | 83K | ⓁⒻ | - | ⚙ | ✓ |
| Portuguese | 226K | ⓁⒻ | - | ⚙ | ✓ |
| Portuguese-BR | 298K | Ⓕ | - | ⚙ | ⧗ |
| Romanian | 145K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Russian | 99K | Ⓕ | 🗋 | ⚙✓ | ⧗ |
| Russian-SynTagRus | 1,032K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ |
| Slovenian | 140K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Slovenian-SST | 29K | ⓁⒻ | 🗋 | 👤 | ⧗ |
| Spanish | 423K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Spanish-AnCora | 547K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ |
| Swedish | 96K | ⓁⒻ | 🗎 | ⚙✓ | ✓ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⧗ |
| Tamil | 8K | ⓁⒻ | - | ⚙ | ✓ |
| Turkish | 56K | ⓁⒻ | 🗋 | ⚙ | ⧗ |
| Ukrainian | - | | - | ⚙✓ | - |

## Release v2

102 treebanks

61 languages

145 contributors

http://universaldependencies.org

## Size:

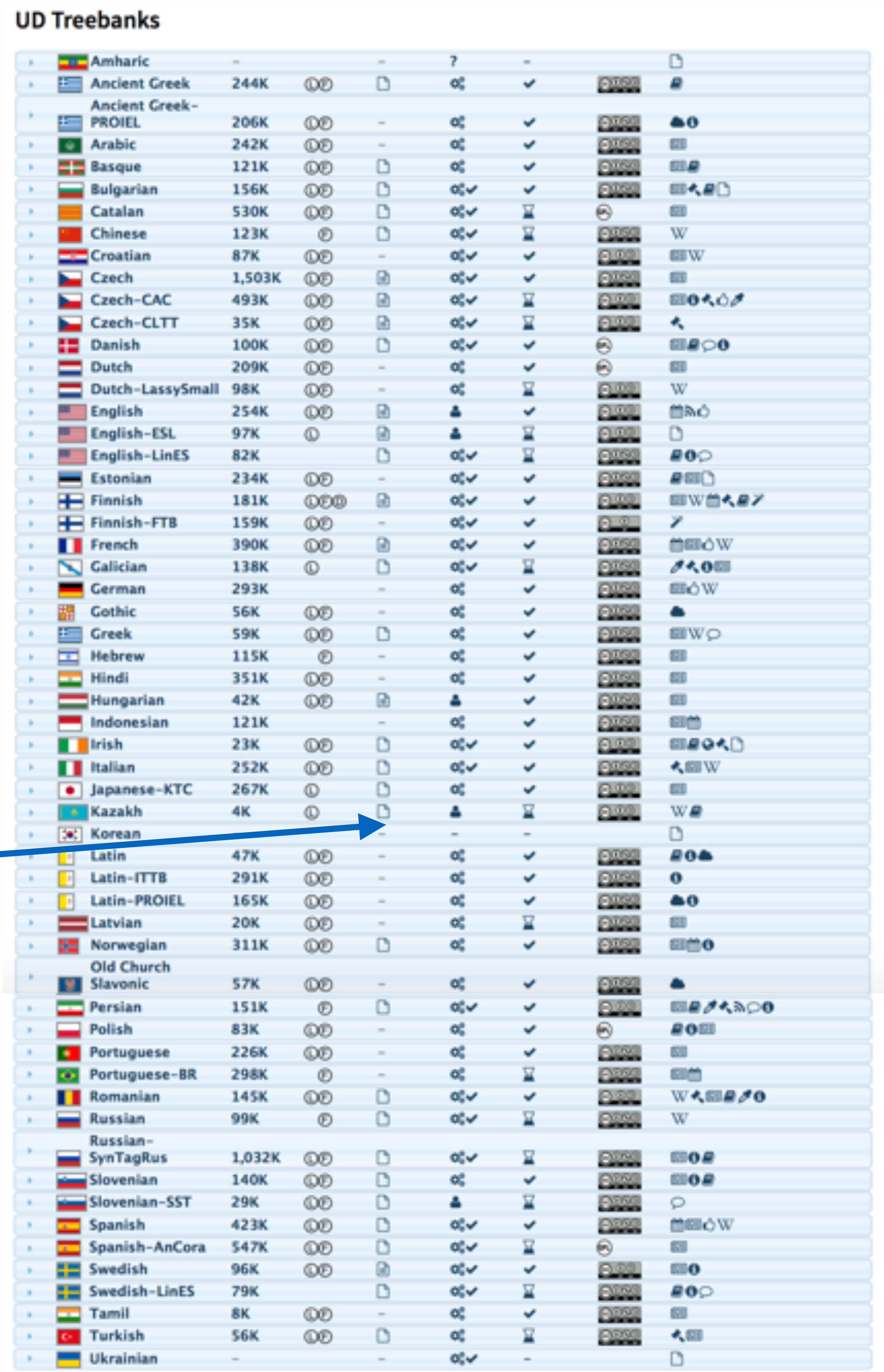
<1K to 1,5M tokens

## Annotation:

lemmas (83), features (94)

## Language-specific guidelines:

complete (14), partial (34)



### UD Treebanks

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Amharic | - | | - | ? | - | | 🗋 |
| Ancient Greek | 244K | LF | 🗋 | ⚙ | ✓ | cc | ■ |
| Ancient Greek-PROIEL | 206K | LF | - | ⚙ | ✓ | cc | ☁ⓘ |
| Arabic | 242K | LF | - | ⚙ | ✓ | cc | ▤ |
| Basque | 121K | LF | 🗋 | ⚙ | ✓ | cc | ▤■ |
| Bulgarian | 156K | LF | 🗋 | ⚙✓ | ✓ | cc | ▤⚒■🗋 |
| Catalan | 530K | LF | 🗋 | ⚙✓ | ⧗ | ⊜ | ▤ |
| Chinese | 123K | F | 🗋 | ⚙✓ | ⧗ | cc | W |
| Croatian | 87K | LF | - | ⚙✓ | ✓ | cc | ▤W |
| Czech | 1,503K | LF | 🗎 | ⚙✓ | ✓ | cc | ▤ |
| Czech-CAC | 493K | LF | 🗎 | ⚙✓ | ⧗ | cc | ▤ⓘ⚒♁✒ |
| Czech-CLTT | 35K | LF | 🗎 | ⚙✓ | ⧗ | cc | ⚒ |
| Danish | 100K | LF | 🗋 | ⚙✓ | ✓ | ⊜ | ▤■💬ⓘ |
| Dutch | 209K | LF | - | ⚙ | ✓ | ⊜ | ▤ |
| Dutch-LassySmall | 98K | LF | - | ⚙ | ⧗ | cc | W |
| English | 254K | LF | 🗎 | 👤 | ✓ | cc | 📅⊃♁ |
| English-ESL | 97K | L | 🗎 | 👤 | ⧗ | cc | 🗋 |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⧗ | cc | ■ⓘ💬 |
| Estonian | 234K | LF | - | ⚙✓ | ✓ | cc | ■▤🗋 |
| Finnish | 181K | LFD | 🗎 | ⚙✓ | ✓ | cc | ▤W📅⚒■✗ |
| Finnish-FTB | 159K | LF | - | ⚙✓ | ✓ | cc | ✗ |
| French | 390K | LF | 🗎 | ⚙✓ | ✓ | cc | 📅▤♁W |
| Galician | 138K | L | 🗋 | ⚙✓ | ⧗ | cc | ✒⚒ⓘ▤ |
| German | 293K | | - | ⚙ | ✓ | cc | ▤♁W |
| Gothic | 56K | LF | - | ⚙ | ✓ | cc | ☁ |
| Greek | 59K | LF | 🗋 | ⚙ | ✓ | cc | ▤W💬 |
| Hebrew | 115K | F | - | ⚙ | ✓ | cc | ▤ |
| Hindi | 351K | LF | - | ⚙ | ✓ | cc | ▤ |
| Hungarian | 42K | LF | 🗎 | 👤 | ✓ | cc | ▤ |
| Indonesian | 121K | | - | ⚙ | ✓ | cc | ▤📅 |
| Irish | 23K | LF | 🗋 | ⚙✓ | ✓ | cc | ▤■♁⚒🗋 |
| Italian | 252K | LF | 🗋 | ⚙✓ | ✓ | cc | ⚒▤W |
| Japanese-KTC | 267K | L | 🗋 | ⚙ | ✓ | cc | ▤ |
| Kazakh | 4K | L | 🗋 | 👤 | ⧗ | cc | W■ |
| Korean | | | - | - | - | | 🗋 |
| Latin | 47K | LF | - | ⚙ | ✓ | cc | ■ⓘ☁ |
| Latin-ITTB | 291K | LF | - | ⚙ | ✓ | cc | ⓘ |
| Latin-PROIEL | 165K | LF | - | ⚙ | ✓ | cc | ☁ⓘ |
| Latvian | 20K | LF | - | ⚙ | ⧗ | cc | ▤ |
| Norwegian | 311K | LF | 🗋 | ⚙ | ✓ | cc | ▤📅ⓘ |
| Old Church Slavonic | 57K | LF | - | ⚙ | ✓ | cc | ☁ |
| Persian | 151K | F | 🗋 | ⚙✓ | ✓ | cc | ▤■✒⚒⊃💬ⓘ |
| Polish | 83K | LF | - | ⚙ | ✓ | ⊜ | ■ⓘ▤ |
| Portuguese | 226K | LF | - | ⚙ | ✓ | cc | ▤ |
| Portuguese-BR | 298K | F | - | ⚙ | ⧗ | cc | ▤📅 |
| Romanian | 145K | LF | 🗋 | ⚙✓ | ✓ | cc | W⚒▤■✒ⓘ |
| Russian | 99K | F | 🗋 | ⚙✓ | ⧗ | cc | W |
| Russian-SynTagRus | 1,032K | LF | 🗋 | ⚙✓ | ⧗ | cc | ▤ⓘ■ |
| Slovenian | 140K | LF | 🗋 | ⚙ | ✓ | cc | ▤ⓘ■ |
| Slovenian-SST | 29K | LF | 🗋 | 👤 | ⧗ | cc | 💬 |
| Spanish | 423K | LF | 🗋 | ⚙✓ | ✓ | cc | 📅▤♁W |
| Spanish-AnCora | 547K | LF | 🗋 | ⚙✓ | ⧗ | ⊜ | ▤ |
| Swedish | 96K | LF | 🗎 | ⚙✓ | ✓ | cc | ▤ⓘ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⧗ | cc | ■ⓘ💬 |
| Tamil | 8K | LF | - | ⚙ | ✓ | cc | ▤ |
| Turkish | 56K | LF | 🗋 | ⚙ | ⧗ | cc | ⚒▤ |
| Ukrainian | - | | - | ⚙✓ | - | | 🗋 |

# Release v2

102 treebanks

61 languages

145 contributors

http://universaldependencies.org

## Size:

<1K to 1,5M tokens

## Annotation:

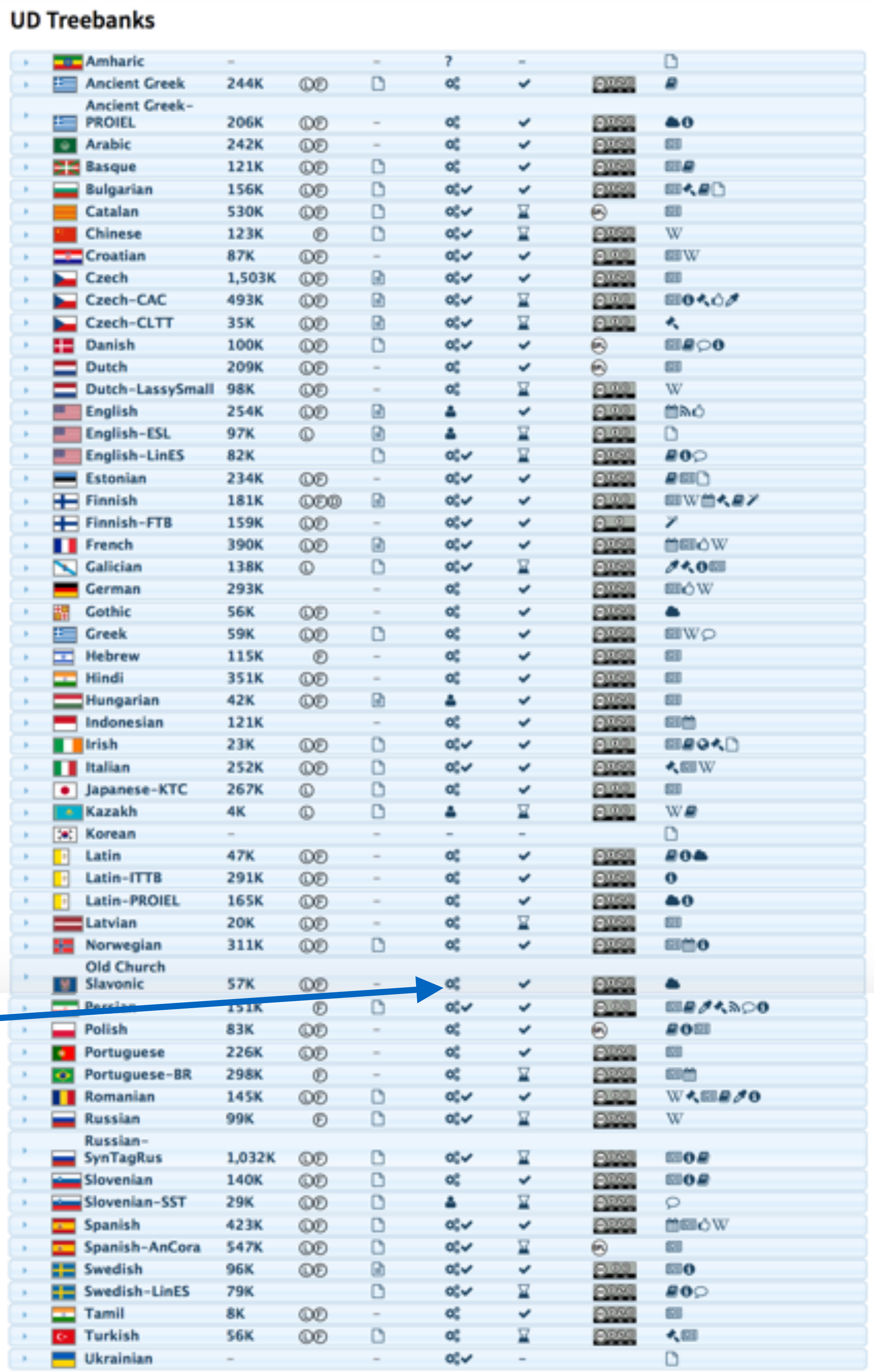
lemmas (83), features (94)

## Language-specific guidelines:

complete (14), partial (34)

## Manual annotation/ validation:

complete (34), partial (31)



UD Treebanks

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Amharic | - | | - | ? | - | | 🗋 |
| Ancient Greek | 244K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | 📓 |
| Ancient Greek-PROIEL | 206K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁ⓘ |
| Arabic | 242K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▤ |
| Basque | 121K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▤📓 |
| Bulgarian | 156K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | ▤🔨📓🗋 |
| Catalan | 530K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | Ⓒ | ▤ |
| Chinese | 123K | Ⓕ | 🗋 | ⚙✓ | ⌛ | 🅭 | W |
| Croatian | 87K | ⓁⒻ | - | ⚙✓ | ✓ | 🅭 | ▤W |
| Czech | 1,503K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | 🅭 | ▤ |
| Czech-CAC | 493K | ⓁⒻ | 🗎 | ⚙✓ | ⌛ | 🅭 | ▤ⓘ🔨👍✎ |
| Czech-CLTT | 35K | ⓁⒻ | 🗎 | ⚙✓ | ⌛ | 🅭 | 🔨 |
| Danish | 100K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | Ⓒ | ▤📓💬ⓘ |
| Dutch | 209K | ⓁⒻ | - | ⚙ | ✓ | Ⓒ | ▤ |
| Dutch-LassySmall | 98K | ⓁⒻ | - | ⚙ | ⌛ | 🅭 | W |
| English | 254K | ⓁⒻ | 🗎 | 👤 | ✓ | 🅭 | 📅📡👍 |
| English-ESL | 97K | Ⓛ | 🗎 | 👤 | ⌛ | 🅭 | 🗋 |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⌛ | 🅭 | 📓ⓘ💬 |
| Estonian | 234K | ⓁⒻ | - | ⚙✓ | ✓ | 🅭 | 📓▤🗋 |
| Finnish | 181K | ⓁⒻⒹ | 🗎 | ⚙✓ | ✓ | 🅭 | ▤W📅🔨📓✎ |
| Finnish-FTB | 159K | ⓁⒻ | - | ⚙✓ | ✓ | 🅭 | ✎ |
| French | 390K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | 🅭 | 📅▤👍W |
| Galician | 138K | Ⓛ | 🗋 | ⚙✓ | ⌛ | 🅭 | ✎🔨ⓘ▤ |
| German | 293K | | - | ⚙ | ✓ | 🅭 | ▤👍W |
| Gothic | 56K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁ |
| Greek | 59K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▤W💬 |
| Hebrew | 115K | Ⓕ | - | ⚙ | ✓ | 🅭 | ▤ |
| Hindi | 351K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▤ |
| Hungarian | 42K | ⓁⒻ | 🗎 | 👤 | ✓ | 🅭 | ▤ |
| Indonesian | 121K | | - | ⚙ | ✓ | 🅭 | ▤📅 |
| Irish | 23K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | ▤📓🌐🔨🗋 |
| Italian | 252K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | 🔨▤W |
| Japanese-KTC | 267K | Ⓛ | 🗋 | ⚙ | ✓ | 🅭 | ▤ |
| Kazakh | 4K | Ⓛ | 🗋 | 👤 | ⌛ | 🅭 | W📓 |
| Korean | - | | - | - | - | | 🗋 |
| Latin | 47K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | 📓ⓘ☁ |
| Latin-ITTB | 291K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ⓘ |
| Latin-PROIEL | 165K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁ⓘ |
| Latvian | 20K | ⓁⒻ | - | ⚙ | ⌛ | 🅭 | ▤ |
| Norwegian | 311K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▤📅ⓘ |
| Old Church Slavonic | 57K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ☁ |
| Persian | 151K | Ⓕ | 🗋 | ⚙✓ | ✓ | 🅭 | ▤📓✎🔨📡💬ⓘ |
| Polish | 83K | ⓁⒻ | - | ⚙ | ✓ | Ⓒ | 📓ⓘ▤ |
| Portuguese | 226K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▤ |
| Portuguese-BR | 298K | Ⓕ | - | ⚙ | ⌛ | 🅭 | ▤📅 |
| Romanian | 145K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | W🔨▤📓✎ⓘ |
| Russian | 99K | Ⓕ | 🗋 | ⚙✓ | ⌛ | 🅭 | W |
| Russian-SynTagRus | 1,032K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | 🅭 | ▤ⓘ📓 |
| Slovenian | 140K | ⓁⒻ | 🗋 | ⚙ | ✓ | 🅭 | ▤ⓘ📓 |
| Slovenian-SST | 29K | ⓁⒻ | 🗋 | 👤 | ⌛ | 🅭 | 💬 |
| Spanish | 423K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | 🅭 | 📅▤👍W |
| Spanish-AnCora | 547K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | Ⓒ | ▤ |
| Swedish | 96K | ⓁⒻ | 🗎 | ⚙✓ | ✓ | 🅭 | ▤ⓘ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⌛ | 🅭 | 📓ⓘ💬 |
| Tamil | 8K | ⓁⒻ | - | ⚙ | ✓ | 🅭 | ▤ |
| Turkish | 56K | ⓁⒻ | 🗋 | ⚙ | ⌛ | 🅭 | 🔨▤ |
| Ukrainian | - | | - | ⚙✓ | - | | 🗋 |

# Release v2

102 treebanks

61 languages

145 contributors

http://universaldependencies.org

## Size:

<1K to 1,5M tokens

## Annotation:

lemmas (83), features (94)

## Language-specific guidelines:

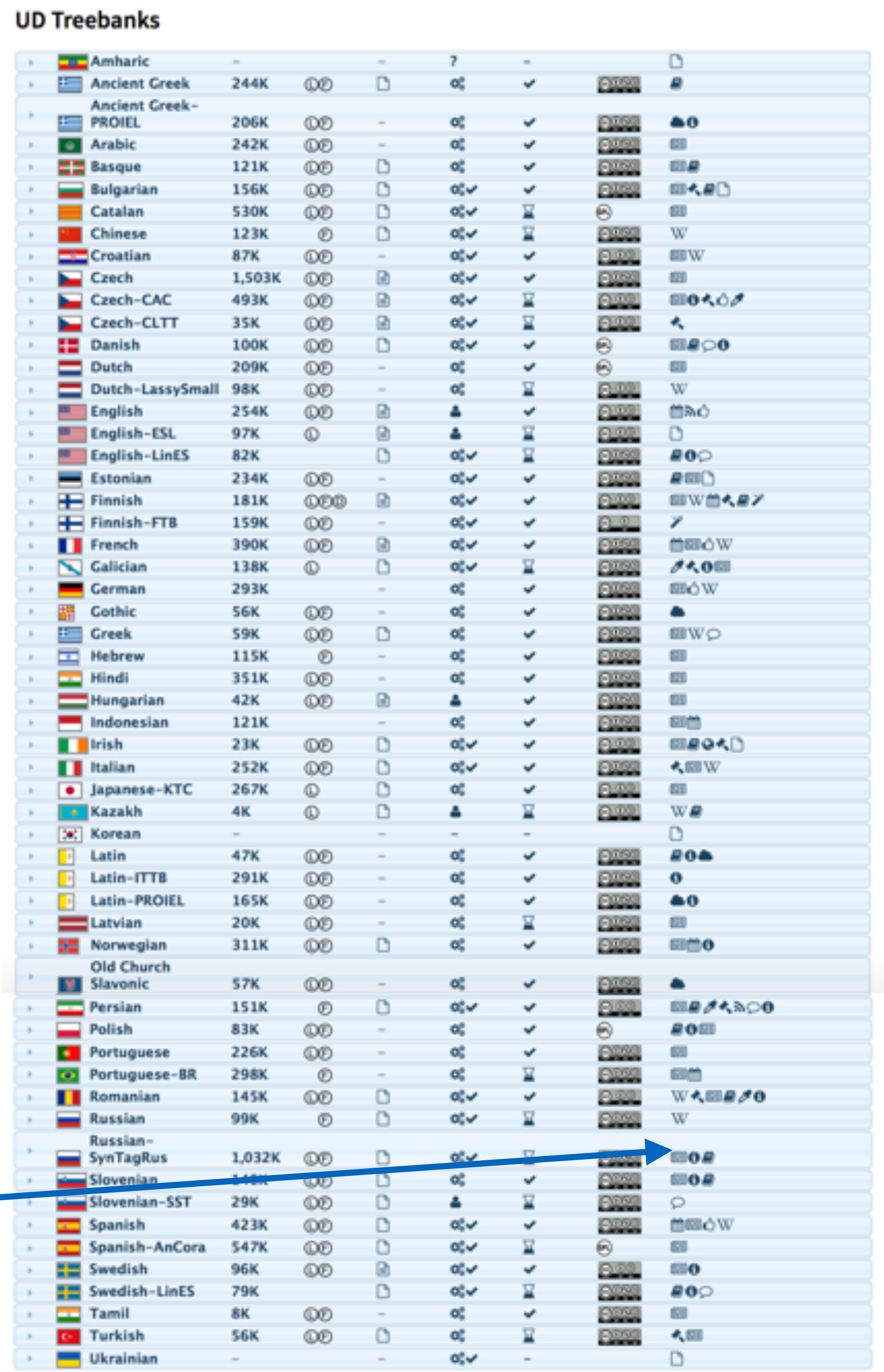
complete (14), partial (34)

## Manual annotation/ validation:

complete (34), partial (31)

## Genres:

bible, blog, fiction, grammar examples, legal text, medical text, news, non-fiction, reviews, spoken, social, web, wikipedia



### UD Treebanks

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Amharic | - | | - | ? | - | | 🗋 |
| Ancient Greek | 244K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | 📖 |
| Ancient Greek-PROIEL | 206K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ⓘ |
| Arabic | 242K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Basque | 121K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤📖 |
| Bulgarian | 156K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤🔨📖🗋 |
| Catalan | 530K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ | ◉ | ▤ |
| Chinese | 123K | Ⓕ | 🗋 | ⚙✓ | ⧗ | CC | W |
| Croatian | 87K | ⓁⒻ | - | ⚙✓ | ✓ | CC | ▤W |
| Czech | 1,503K | ⓁⒻ | 🖹 | ⚙✓ | ✓ | CC | ▤ |
| Czech-CAC | 493K | ⓁⒻ | 🖹 | ⚙✓ | ⧗ | CC | ▤ⓘ🔨👍💉 |
| Czech-CLTT | 35K | ⓁⒻ | 🖹 | ⚙✓ | ⧗ | CC | 🔨 |
| Danish | 100K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | ◉ | ▤📖💬ⓘ |
| Dutch | 209K | ⓁⒻ | - | ⚙ | ✓ | ◉ | ▤ |
| Dutch-LassySmall | 98K | ⓁⒻ | - | ⚙ | ⧗ | CC | W |
| English | 254K | ⓁⒻ | 🖹 | 👤 | ✓ | CC | 📅📶👍 |
| English-ESL | 97K | Ⓛ | 🖹 | 👤 | ⧗ | CC | 🗋 |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⧗ | CC | 📖ⓘ💬 |
| Estonian | 234K | ⓁⒻ | - | ⚙✓ | ✓ | CC | 📖▤🗋 |
| Finnish | 181K | ⓁⒻⒺ | 🖹 | ⚙✓ | ✓ | CC | ▤W📅🔨📖✎ |
| Finnish-FTB | 159K | ⓁⒻ | - | ⚙✓ | ✓ | CC | ✎ |
| French | 390K | ⓁⒻ | 🖹 | ⚙✓ | ✓ | CC | 📅▤👍W |
| Galician | 138K | Ⓛ | 🗋 | ⚙✓ | ⧗ | CC | 💉🔨ⓘ▤ |
| German | 293K | | - | ⚙ | ✓ | CC | ▤👍W |
| Gothic | 56K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ |
| Greek | 59K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤W💬 |
| Hebrew | 115K | Ⓕ | - | ⚙ | ✓ | CC | ▤ |
| Hindi | 351K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Hungarian | 42K | ⓁⒻ | 🖹 | 👤 | ✓ | CC | ▤ |
| Indonesian | 121K | | - | ⚙ | ✓ | CC | ▤📅 |
| Irish | 23K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤📖🌐🔨🗋 |
| Italian | 252K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | 🔨▤W |
| Japanese-KTC | 267K | Ⓛ | 🗋 | ⚙ | ✓ | CC | ▤ |
| Kazakh | 4K | Ⓛ | 🗋 | 👤 | ⧗ | CC | W📖 |
| Korean | - | | - | - | - | | 🗋 |
| Latin | 47K | ⓁⒻ | - | ⚙ | ✓ | CC | 📖ⓘ☁ |
| Latin-ITTB | 291K | ⓁⒻ | - | ⚙ | ✓ | CC | ⓘ |
| Latin-PROIEL | 165K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ⓘ |
| Latvian | 20K | ⓁⒻ | - | ⚙ | ⧗ | CC | ▤ |
| Norwegian | 311K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤📅ⓘ |
| Old Church Slavonic | 57K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ |
| Persian | 151K | Ⓕ | 🗋 | ⚙✓ | ✓ | CC | ▤📖💉🔨📶💬ⓘ |
| Polish | 83K | ⓁⒻ | - | ⚙ | ✓ | ◉ | 📖ⓘ▤ |
| Portuguese | 226K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Portuguese-BR | 298K | Ⓕ | - | ⚙ | ⧗ | CC | ▤📅 |
| Romanian | 145K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | W🔨▤📖💉ⓘ |
| Russian | 99K | Ⓕ | 🗋 | ⚙✓ | ⧗ | CC | W |
| Russian-SynTagRus | 1,032K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ | CC | ▤ⓘ📖 |
| Slovenian | 140K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤ⓘ📖 |
| Slovenian-SST | 29K | ⓁⒻ | 🗋 | 👤 | ⧗ | CC | 💬 |
| Spanish | 423K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | 📅▤👍W |
| Spanish-AnCora | 547K | ⓁⒻ | 🗋 | ⚙✓ | ⧗ | ◉ | ▤ |
| Swedish | 96K | ⓁⒻ | 🖹 | ⚙✓ | ✓ | CC | ▤ⓘ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⧗ | CC | 📖ⓘ💬 |
| Tamil | 8K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Turkish | 56K | ⓁⒻ | 🗋 | ⚙ | ⧗ | CC | 🔨▤ |
| Ukrainian | - | | - | ⚙✓ | - | | 🗋 |

## Release v2

102 treebanks
61 languages
145 contributors

http://universaldependencies.org

## Size:

<1K to 1,5M toke s

## Annotation:

lemmas (83), fe (94)

## Language-s uidelines:

co

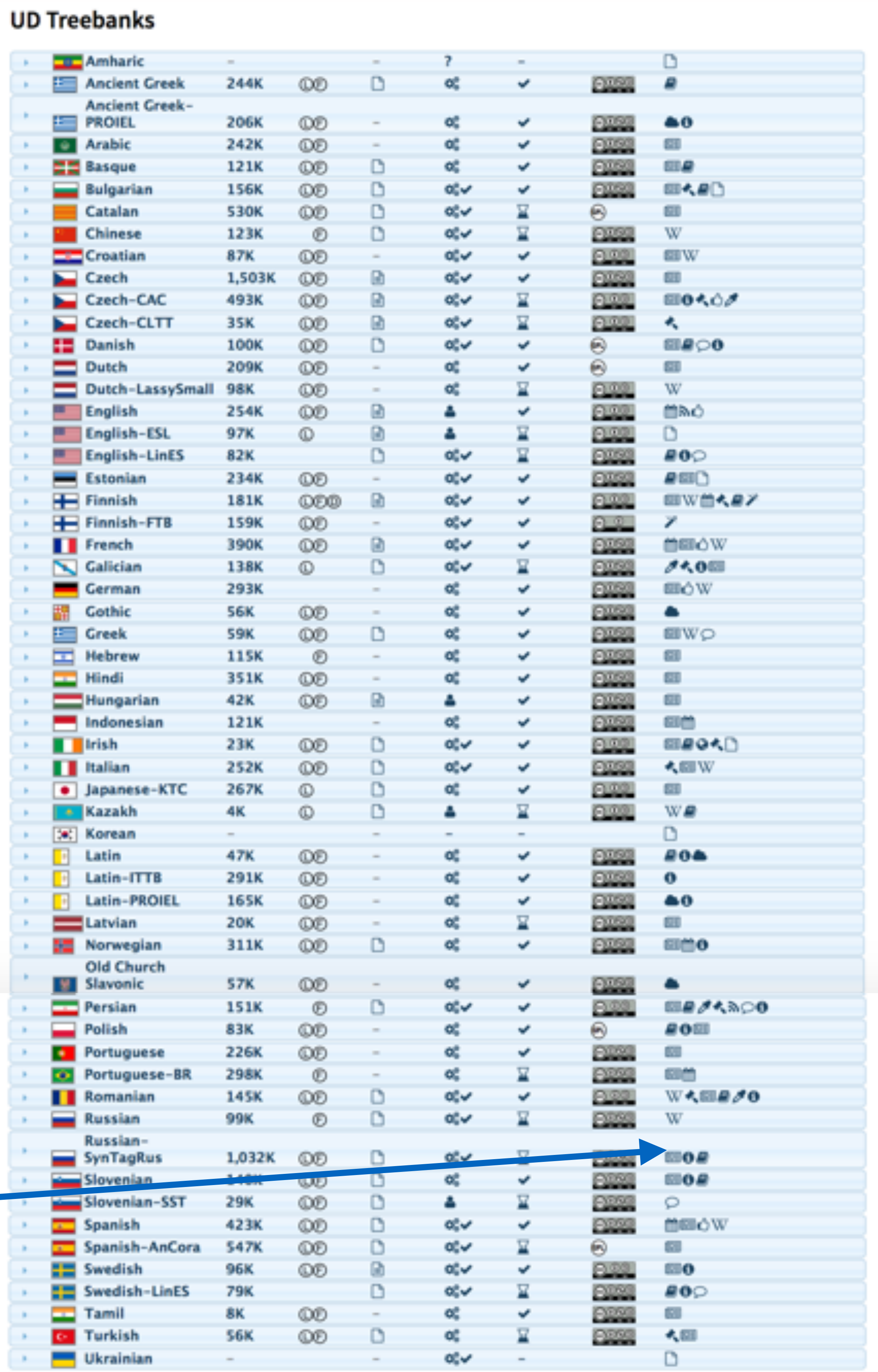
## Manual validation

complete

## Genres

bible, bl ction, grammar exa es, legal te t, medical text, news, non-fi tion, reviews, spoken, social, web, wikipedia

Over 50% of the world's native speakers



UD Treebanks

| | | | | | |
|---|---|---|---|---|---|
| Amharic | - | | - | ? | - |
| Ancient Greek | 244K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Ancient Greek-PROIEL | 206K | ⓁⒻ | - | ⚙ | ✓ |
| Arabic | 242K | ⓁⒻ | - | ⚙ | ✓ |
| Basque | 121K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Bulgarian | 156K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Catalan | 530K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ |
| Chinese | 123K | Ⓕ | 🗋 | ⚙✓ | ⌛ |
| Croatian | 87K | ⓁⒻ | - | ⚙✓ | ✓ |
| Czech | 1,503K | ⓁⒻ | 🗎 | ⚙✓ | ✓ |
| Czech-CAC | 493K | ⓁⒻ | 🗎 | ⚙✓ | ⌛ |
| Czech-CLTT | 35K | ⓁⒻ | 🗎 | ⚙✓ | ⌛ |
| Danish | 100K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Dutch | 209K | ⓁⒻ | - | ⚙ | ✓ |
| Dutch-LassySmall | 98K | ⓁⒻ | - | ⚙ | ⌛ |
| English | 254K | ⓁⒻ | 🗎 | 👤 | ✓ |
| English-ESL | 97K | Ⓛ | 🗎 | 👤 | ⌛ |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⌛ |
| Estonian | 234K | ⓁⒻ | - | ⚙✓ | ✓ |
| Finnish | 181K | ⓁⒻⒹ | 🗎 | ⚙✓ | ✓ |
| Finnish-FTB | 159K | ⓁⒻ | - | ⚙✓ | ✓ |
| French | 390K | ⓁⒻ | 🗎 | ⚙✓ | ✓ |
| Galician | 138K | Ⓛ | 🗋 | ⚙✓ | ⌛ |
| German | 293K | | - | ⚙ | ✓ |
| Gothic | 56K | ⓁⒻ | - | ⚙ | ✓ |
| Greek | 59K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Hebrew | 115K | Ⓕ | - | ⚙ | ✓ |
| Hindi | 351K | ⓁⒻ | - | ⚙ | ✓ |
| Hungarian | 42K | ⓁⒻ | 🗎 | 👤 | ✓ |
| Indonesian | 121K | | - | ⚙ | ✓ |
| Irish | 23K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Italian | 252K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Japanese-KTC | 267K | Ⓛ | 🗋 | ⚙ | ✓ |
| Kazakh | 4K | Ⓛ | 🗋 | 👤 | ⌛ |
| Korean | - | | - | - | - |
| Latin | 47K | ⓁⒻ | - | ⚙ | ✓ |
| Latin-ITTB | 291K | ⓁⒻ | - | ⚙ | ✓ |
| Latin-PROIEL | 165K | ⓁⒻ | - | ⚙ | ✓ |
| Latvian | 20K | ⓁⒻ | - | ⚙ | ⌛ |
| Norwegian | 311K | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Old Church Slavonic | 57K | ⓁⒻ | - | ⚙ | ✓ |
| Persian | 151K | Ⓕ | 🗋 | ⚙✓ | ✓ |
| Polish | 83K | ⓁⒻ | - | ⚙ | ✓ |
| Portuguese | 226K | ⓁⒻ | - | ⚙ | ✓ |
| Portuguese-BR | 298K | Ⓕ | - | ⚙ | ⌛ |
| Romanian | 145K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Russian | 99K | Ⓕ | 🗋 | ⚙✓ | ⌛ |
| Russian-SynTagRus | 1,032K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ |
| Slovenian | [illegible] | ⓁⒻ | 🗋 | ⚙ | ✓ |
| Slovenian-SST | 29K | ⓁⒻ | 🗋 | 👤 | ⌛ |
| Spanish | 423K | ⓁⒻ | 🗋 | ⚙✓ | ✓ |
| Spanish-AnCora | 547K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ |
| Swedish | 96K | ⓁⒻ | 🗎 | ⚙✓ | ✓ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⌛ |
| Tamil | 8K | ⓁⒻ | - | ⚙ | ✓ |
| Turkish | 56K | ⓁⒻ | 🗋 | ⚙ | ⌛ |
| Ukrainian | - | | - | ⚙✓ | - |

## Release v2

102 treebanks
61 languages
145 contributors

http://universaldependencies.org

## Size:

<1K to 1,5M tokens

## Annotation:

lemmas (83), features (94)

## Language-specific guidelines:

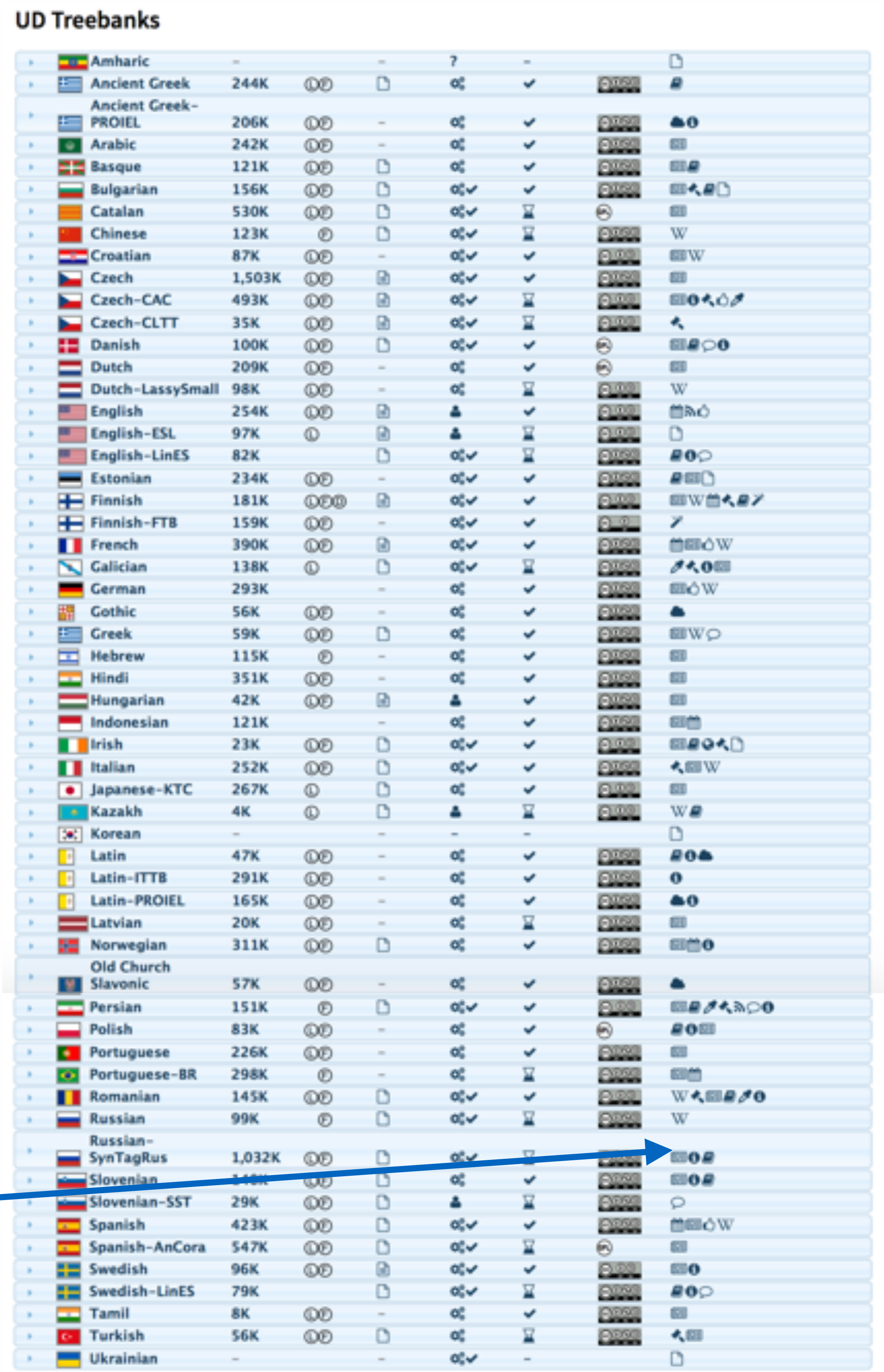
complete (14), partial (34)

## Manual annotation/ validation:

complete (34), partial (31)

## Genres:

bible, blog, fiction, grammar examples, legal text, medical text, news, non-fiction, reviews, spoken, social, web, wikipedia



UD Treebanks

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Amharic | - | | - | ? | - | | 🗋 |
| Ancient Greek | 244K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | 📘 |
| Ancient Greek-PROIEL | 206K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ⓘ |
| Arabic | 242K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Basque | 121K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤📘 |
| Bulgarian | 156K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤🔨📘🗋 |
| Catalan | 530K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | ◉ | ▤ |
| Chinese | 123K | Ⓕ | 🗋 | ⚙✓ | ⌛ | CC | W |
| Croatian | 87K | ⓁⒻ | - | ⚙✓ | ✓ | CC | ▤W |
| Czech | 1,503K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤ |
| Czech-CAC | 493K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | CC | ▤ⓘ🔨👍💉 |
| Czech-CLTT | 35K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | CC | 🔨 |
| Danish | 100K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | ◉ | ▤📘💬ⓘ |
| Dutch | 209K | ⓁⒻ | - | ⚙ | ✓ | ◉ | ▤ |
| Dutch-LassySmall | 98K | ⓁⒻ | - | ⚙ | ⌛ | CC | W |
| English | 254K | ⓁⒻ | 🗋 | 👤 | ✓ | CC | 🗓📶👍 |
| English-ESL | 97K | Ⓛ | 🗋 | 👤 | ⌛ | CC | 🗋 |
| English-LinES | 82K | | 🗋 | ⚙✓ | ⌛ | CC | 📘ⓘ💬 |
| Estonian | 234K | ⓁⒻ | - | ⚙✓ | ✓ | CC | 📘▤🗋 |
| Finnish | 181K | ⓁⒻⓂ | 🗋 | ⚙✓ | ✓ | CC | ▤W🗓🔨📘✐ |
| Finnish-FTB | 159K | ⓁⒻ | - | ⚙✓ | ✓ | CC | ✐ |
| French | 390K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | 🗓▤👍W |
| Galician | 138K | Ⓛ | 🗋 | ⚙✓ | ⌛ | CC | 💉🔨ⓘ▤ |
| German | 293K | | - | ⚙ | ✓ | CC | ▤👍W |
| Gothic | 56K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ |
| Greek | 59K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤W💬 |
| Hebrew | 115K | Ⓕ | - | ⚙ | ✓ | CC | ▤ |
| Hindi | 351K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Hungarian | 42K | ⓁⒻ | 🗋 | 👤 | ✓ | CC | ▤ |
| Indonesian | 121K | | - | ⚙ | ✓ | CC | ▤🗓 |
| Irish | 23K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤📘🌐🔨🗋 |
| Italian | 252K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | 🔨▤W |
| Japanese-KTC | 267K | Ⓛ | 🗋 | ⚙ | ✓ | CC | ▤ |
| Kazakh | 4K | Ⓛ | 🗋 | 👤 | ⌛ | CC | W📘 |
| Korean | - | | - | - | - | | 🗋 |
| Latin | 47K | ⓁⒻ | - | ⚙ | ✓ | CC | 📘ⓘ☁ |
| Latin-ITTB | 291K | ⓁⒻ | - | ⚙ | ✓ | CC | ⓘ |
| Latin-PROIEL | 165K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ⓘ |
| Latvian | 20K | ⓁⒻ | - | ⚙ | ⌛ | CC | ▤ |
| Norwegian | 311K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤🗓ⓘ |
| Old Church Slavonic | 57K | ⓁⒻ | - | ⚙ | ✓ | CC | ☁ |
| Persian | 151K | Ⓕ | 🗋 | ⚙✓ | ✓ | CC | ▤📘💉🔨📶💬ⓘ |
| Polish | 83K | ⓁⒻ | - | ⚙ | ✓ | ◉ | 📘ⓘ▤ |
| Portuguese | 226K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Portuguese-BR | 298K | Ⓕ | - | ⚙ | ⌛ | CC | ▤🗓 |
| Romanian | 145K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | W🔨▤📘💉ⓘ |
| Russian | 99K | Ⓕ | 🗋 | ⚙✓ | ⌛ | CC | W |
| Russian-SynTagRus | 1,032K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | CC | ▤ⓘ📘 |
| Slovenian | 140K | ⓁⒻ | 🗋 | ⚙ | ✓ | CC | ▤ⓘ📘 |
| Slovenian-SST | 29K | ⓁⒻ | 🗋 | 👤 | ⌛ | CC | 💬 |
| Spanish | 423K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | 🗓▤👍W |
| Spanish-AnCora | 547K | ⓁⒻ | 🗋 | ⚙✓ | ⌛ | ◉ | ▤ |
| Swedish | 96K | ⓁⒻ | 🗋 | ⚙✓ | ✓ | CC | ▤ⓘ |
| Swedish-LinES | 79K | | 🗋 | ⚙✓ | ⌛ | CC | 📘ⓘ💬 |
| Tamil | 8K | ⓁⒻ | - | ⚙ | ✓ | CC | ▤ |
| Turkish | 56K | ⓁⒻ | 🗋 | ⚙ | ⌛ | CC | 🔨▤ |
| Ukrainian | - | | - | ⚙✓ | - | | 🗋 |

# Research Problems (examples)

- Annotation
  - Produce guidelines and/or annotation for language X
  - Study the annotation of construction Y across languages
- Parsing
  - Develop and/or evaluate a parser for language X
  - Study cross-lingual transfer learning and/or annotation projection
- Typology
  - Study word order patterns in language X
  - Compare the realisation of construction Y across languages
- Applications
  - Investigate UD-based features for your favorite NLP problem

# Questions?