Dependency grammar and Universal Dependencies

an introduction and annotation exercise

Arianna Masciolini

LI2020 Syntax 2

4 March 2024

Who am I and why am I here?

- Arianna Masciolini
- background in **Computer Science**
- PhD student in **Natural Language Processing** at the Department of Swedish, Multilingualism, Language Technology
- interested in **Computational Syntax** and **Second Language Acquisition**
- currently working on
  - **UD treebank of L2 Swedish**
  - **automatic annotation of L2 texts**

Today's agenda

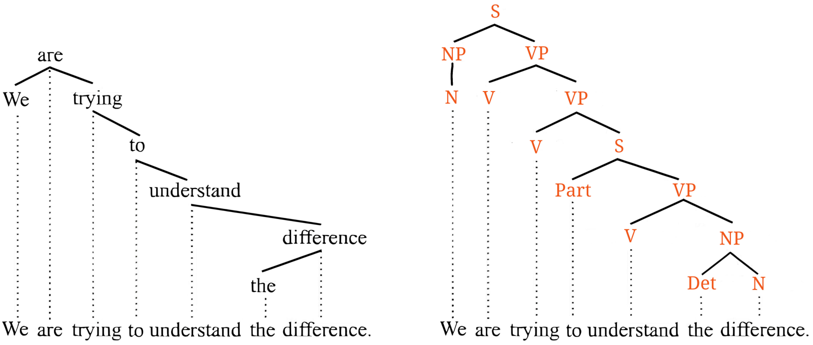
1. basics of **dependency grammar**
2. quick introduction to **Universal Dependencies**
3. **annotation exercise**

Dependency grammar

Dependency vs. phrase structure

dependency grammar	**phrase structure grammar**
- Lucien Tesnière (1959)	- Noam Chomsky (1956)
- descriptive	- generative
- (labelled) head-dependent links	- rewrite rules/transformations
- based on *dependency*	- based on *constituency*

Dependency vs. constituency


are
We
trying
to
understand
difference
the
We are trying to understand the difference.
S
NP
VP
N
V
VP
V
S
Part
VP
V
NP
Det
N
We are trying to understand the difference.

Dependency

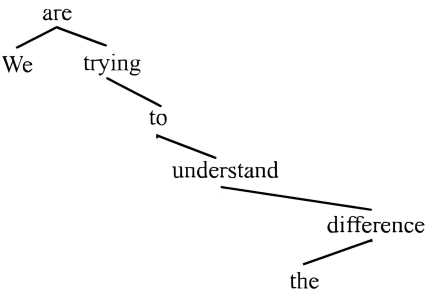
Constituency

original image: commons.wikimedia.org

Dependency

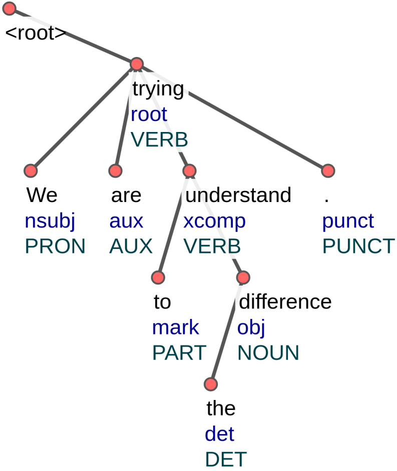
- **one-to-one correspondence** between two elements of a sentence
  - elements are typically words, but can also be subwords or larger semantic units
  - dependency trees typically have less nodes than phrase structure trees
- **directed link** between a *head* and a *dependent*
- links can be **labelled** to specify syntactic function

Various standards and formats



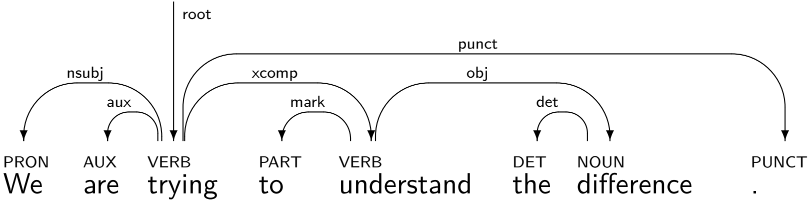
original image: commons.wikimedia.org

Various standards and formats



generated with UDPipe Online: lindat.mff.cuni.cz/services/udpipe

Various standards and formats



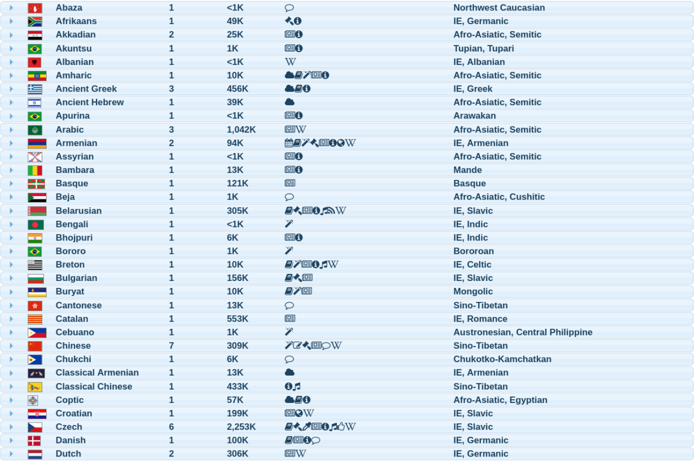
generated with gf-ud: github.com/GrammaticalFramework/gf-ud

Universal Dependencies 101

What is Universal Dependencies?

- a growing **collection of dependency treebanks** for many languages (over 140!)
- an **annotation scheme** for cross-lingually consistent grammatical annotation

Some UD languages



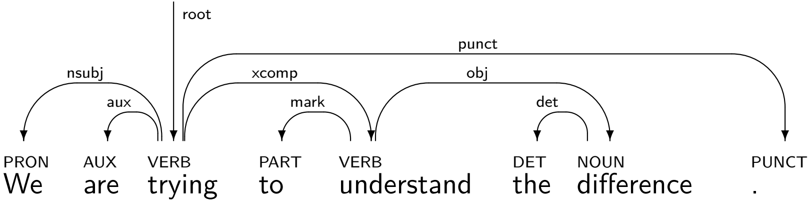
Abaza	1	<1K	💬	Northwest Caucasian
Afrikaans	1	49K	🔨ℹ	IE, Germanic
Akkadian	2	25K	📰ℹ	Afro-Asiatic, Semitic
Akuntsu	1	1K	📰ℹ	Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K	☁📖🪄📰ℹ	Afro-Asiatic, Semitic
Ancient Greek	3	456K	☁📖ℹ	IE, Greek
Ancient Hebrew	1	39K	☁	Afro-Asiatic, Semitic
Apurina	1	<1K	📰ℹ	Arawakan
Arabic	3	1,042K	📰W	Afro-Asiatic, Semitic
Armenian	2	94K	📅📖🪄🔨📰ℹ🌐W	IE, Armenian
Assyrian	1	<1K	📰ℹ	Afro-Asiatic, Semitic
Bambara	1	13K	📰ℹ	Mande
Basque	1	121K	📰	Basque
Beja	1	1K	💬	Afro-Asiatic, Cushitic
Belarusian	1	305K	📖🔨📰ℹ♫📶W	IE, Slavic
Bengali	1	<1K	🪄	IE, Indic
Bhojpuri	1	6K	📰ℹ	IE, Indic
Bororo	1	1K	🪄	Bororoan
Breton	1	10K	📖🪄📰ℹ♫W	IE, Celtic
Bulgarian	1	156K	📖🔨📰	IE, Slavic
Buryat	1	10K	📖🪄📰	Mongolic
Cantonese	1	13K	💬	Sino-Tibetan
Catalan	1	553K	📰	IE, Romance
Cebuano	1	1K	🪄	Austronesian, Central Philippine
Chinese	7	309K	🪄📝🔨📰💬W	Sino-Tibetan
Chukchi	1	6K	💬	Chukotko-Kamchatkan
Classical Armenian	1	13K	☁	IE, Armenian
Classical Chinese	1	433K	ℹ♫	Sino-Tibetan
Coptic	1	57K	☁📖ℹ	Afro-Asiatic, Egyptian
Croatian	1	199K	📰🌐W	IE, Slavic
Czech	6	2,253K	📖🔨💉📰ℹ♫👍W	IE, Slavic
Danish	1	100K	📖📰ℹ💬	IE, Germanic
Dutch	2	306K	📰W	IE, Germanic

source: universaldependencies.org

Design goals

- human *and* machine readability
  - ease of visualization and manual annotation
  - text-based format for straightforward computer processing
- suitability for both mono- and multilingual use cases
  - uniform morphosyntactic annotation layer complemented by language-specific guidelines
  - main fields of applications: typology and Natural Language Processing

UD sentences: tree format

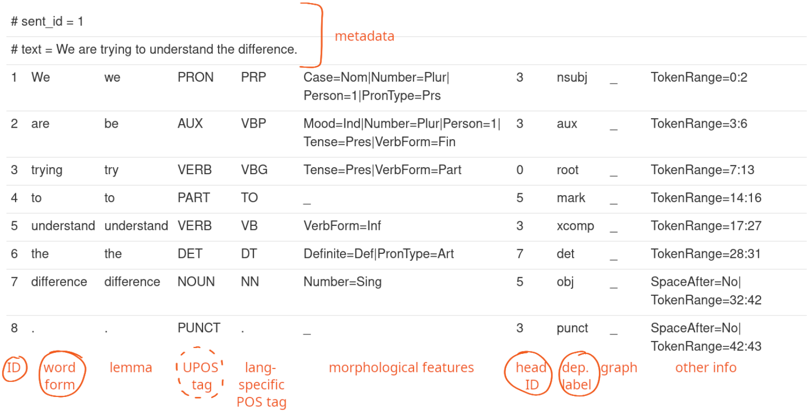


generated with gf-ud: github.com/GrammaticalFramework/gf-ud

UD sentences: CoNLL-U format

```
# sent_id = 1
# text = We are trying to understand the difference.
1   We  we  PRON    PRP Case=Nom|Number=Plur|Person=1|PronType=Prs  3   nsubj   _   TokenRange=0:2
2   are be  AUX VBP Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin   3   aux _   TokenRange=3:6
3   trying  try VERB    VBG Tense=Pres|VerbForm=Part    0   root    _   TokenRange=7:13
4   to  to  PART    TO  _   5   mark    _   TokenRange=14:16
5   understand  understand  VERB    VB  VerbForm=Inf    3   xcomp   _   TokenRange=17:27
6   the the DET DT  Definite=Def|PronType=Art   7   det _   TokenRange=28:31
7   difference  difference  NOUN    NN  Number=Sing 5   obj _   SpaceAfter=No|TokenRange=32:42
8   .   .   PUNCT   .   _   3   punct   _   SpaceAfter=No|TokenRange=42:43
```

UD sentences: table format



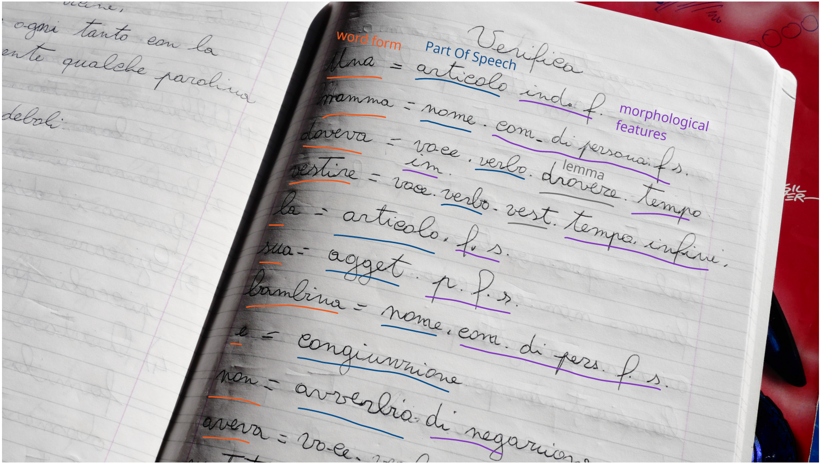
sent_id = 1

text = We are trying to understand the difference.

1	We	we	PRON	PRP	Case=Nom|Number=Plur|Person=1|PronType=Prs	3	nsubj	_	TokenRange=0:2
2	are	be	AUX	VBP	Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin	3	aux	_	TokenRange=3:6
3	trying	try	VERB	VBG	Tense=Pres|VerbForm=Part	0	root	_	TokenRange=7:13
4	to	to	PART	TO	_	5	mark	_	TokenRange=14:16
5	understand	understand	VERB	VB	VerbForm=Inf	3	xcomp	_	TokenRange=17:27
6	the	the	DET	DT	Definite=Def|PronType=Art	7	det	_	TokenRange=28:31
7	difference	difference	NOUN	NN	Number=Sing	5	obj	_	SpaceAfter=No|TokenRange=32:42
8	.	.	PUNCT	.	_	3	punct	_	SpaceAfter=No|TokenRange=42:43

original image generated with UDPipe Online: lindat.mff.cuni.cz/services/udpipe

UD sentences: table format



…iene,
ogni tanto con la
ente qualche parolina
deboli:

Verifica
Una = articolo ind. f.
mamma = nome. com. di persona. f. s.
doveva = voce. verbo. dovere. tempo
im.
vestire = voce. verbo. vest. tempo. infini.
la = articolo. f. s.
sua = agget. p. f. s.
bambina = nome. com. di pers. f. s.
e = congiunzione
non = avverbio. di negazione
aveva = voce. ve…

own image

Content vs. function words

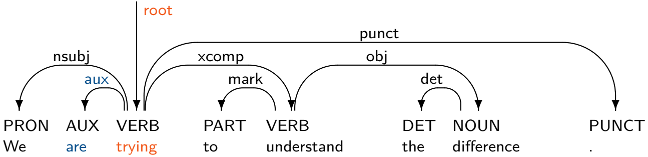
- *content words*: words with own lexical meaning
  - usually *open class*: nouns, lexical verbs, adjectives, adverbs. . .
- *function words*: words that primarily denote grammatical relationships between other words
  - usually *closed class*: prepositions, pronouns, auxiliaries. . .

Primacy of content words

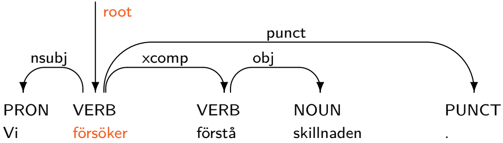
- syntactic heads tend to be content words
- as a rule of thumb, the root of a dependency tree is its main lexical verb or, in its absence, the complement of the copula

Example 1

The root is the present participle *trying*, not the finite auxiliary *are*:

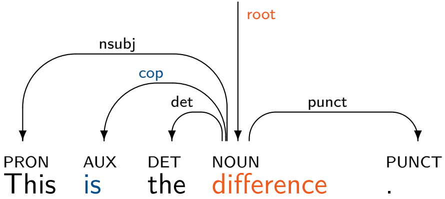


This facilitates comparisons with languages that don't use an auxiliary in this context:

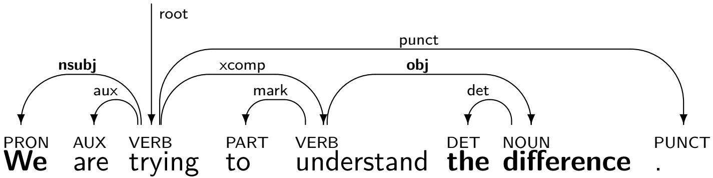


Example 2

The root is the noun *difference*, not the copula *is*:



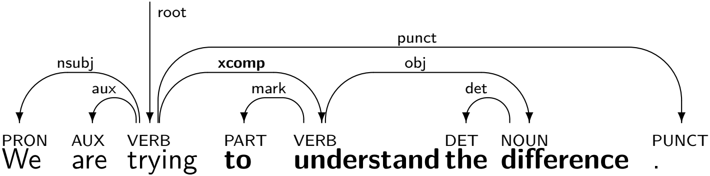
Some more dependency labels



Core nominal arguments of the verb

- **nsubj** (nominal subject)
- **obj** (direct object)

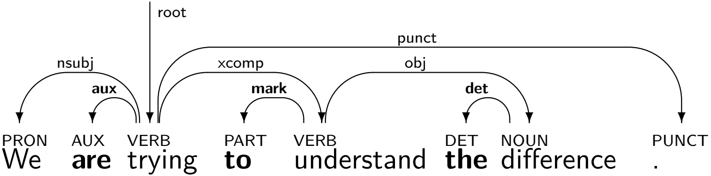
Some more dependency labels



Subordinate clauses

- **xcomp** (predicative complement whose subject is externally determined, as opposed to **ccomp** in sentences like *I think that **we understand the difference***)

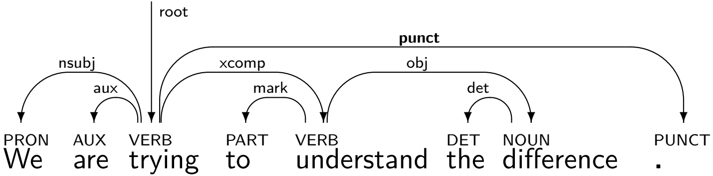
Some more dependency labels



Function words

- **aux** (auxiliary)
- **mark** (word marking a subordinate clause)
- **det** (determiner of a nominal)

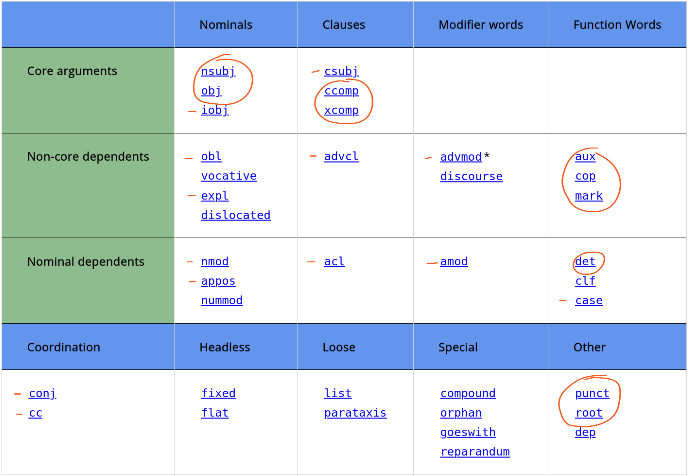
Some more dependency labels



Others

- **punct** (punctuation mark)

Dependency labels: overview



	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj – iobj	– csubj ccomp xcomp		
Non-core dependents	– obl vocative – expl dislocated	– advcl	– advmod* discourse	aux cop mark
Nominal dependents	– nmod – appos nummod	– acl	– amod	det clf – case
Coordination	Headless	Loose	Special	Other
– conj – cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

source: universaldependencies.org

Annotation exercise

Annotation exercise

- 10 hand-picked sentences from the ESL (English as a Second Language) treebank
- 2 different methods:
  1. manual annotation
  2. automatic parsing + manual validation

Sentence 1

I do not want to spend much time on computers.

Sentence 1

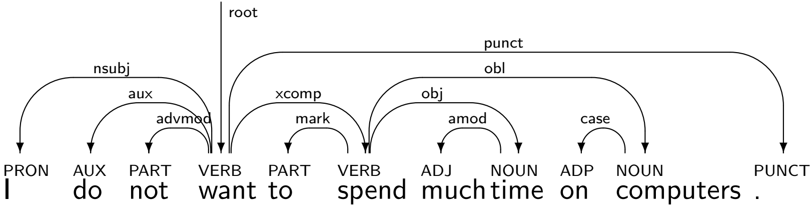
I do not want to spend much time on computers.

- what clause is the subject of the subordinate clause controlled by?

Sentence 1

I do not want to spend much time on computers.

- what clause is the subject of the subordinate clause controlled by?



Sentence 2

All your tasks will be performed by computers.

Sentence 2

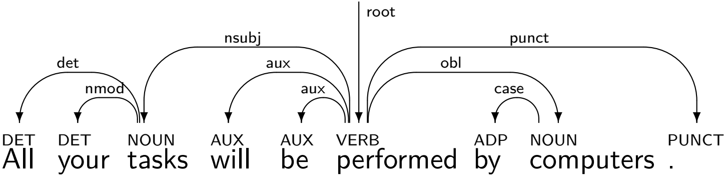
All your tasks will be performed by computers.

- what are the **logical** and **syntactic** subjects of this sentence?

Sentence 2

All your tasks will be performed by computers.

- what are the **logical** and **syntactic** subjects of this sentence?



Sentence 3

Can you imagine life before computers?

Sentence 3

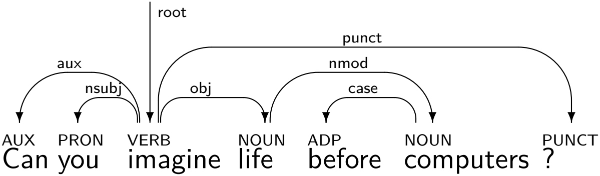
Can you imagine life before computers?

- question
- what does “before computers” modify?

Sentence 3

Can you imagine life before computers?

- question
- what does “before computers” modify?



Sentence 4

There are only ten computers in the school.

Sentence 4

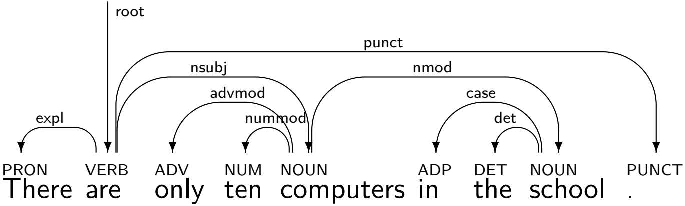
There are only ten computers in the school.

- is the use of the verb “to be” the same as in sentence 2?

Sentence 4

There are only ten computers in the school.

- is the use of the verb “to be” the same as in sentence 2?



Sentence 5

But the most important innovation in technological development is the computer.

Sentence 5

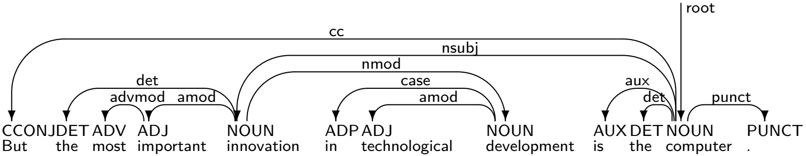
But the most important innovation in technological development is the computer.

- what is the subject here and how many dependents does it have?

Sentence 5

But the most important innovation in technological development is the computer.

- what is the subject here and how many dependents does it have?



Sentence 6

In particular, the computer has changed my daily life dramatically.

Sentence 6

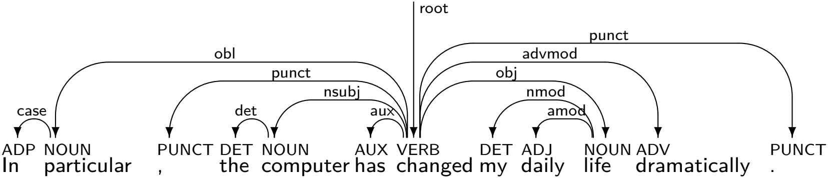
In particular, the computer has changed my daily life dramatically.

- what is “in particular”?

Sentence 6

In particular, the computer has changed my daily life dramatically.

- what is “in particular”?


root
punct
obl
advmod
punct
obj
nsubj
nmod
case
det
aux
amod
ADP NOUN PUNCT DET NOUN AUX VERB DET ADJ NOUN ADV PUNCT
In particular , the computer has changed my daily life dramatically .

Sentence 7

Maybe, technology will never stop advancing and our life will never work without computers.

Sentence 7

Maybe, technology will never stop advancing and our life will never work without computers.

- what are the two conjuncts in this sentence?

Sentence 8

I work with children and the computer helps me in my job but affects it too.

Sentence 8

I work with children and the computer helps me in my job but affects it too.

- two coordinating conjunctions here: what is conjunted to what?

Sentence 9

When I was a child I didn't use the computer because I didn't know what it was.

Sentence 9

When I was a child I didn't use the computer because I didn't know what it was.

- how many clauses are there?
- what is the relationship between them?

Sentence 10

With the introduction of the computer in our civilization we can access the Internet to communicate with our relatives and friends living abroad or far from us.

Sentence 10

With the introduction of the computer in our civilization we can access the Internet to communicate with our relatives and friends living abroad or far from us.

- what is “living” referred to?

Readings & useful links

Learn more

- a more in-depth introduction to UD by its creators and treebank maintainers: **amupod.univ-amu.fr** (video)
- official UD documentation, at **universaldependencies.org**
- a (relatively) up-to-date scientific publication: **Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. *Universal Dependencies.* Computational Linguistics, 47(2):255–308, 2021** (available through the GU library)
- Computational Syntax course, part of the Master in Language Technology, usually in the Spring semester (detailed course notes are available at **cse.chalmers.se/∼aarne/grammarbook.pdf**)

Other useful links

- UDPipe online, a user-friendly online parser with models for many languages: **lindat.mff.cuni.cz/services/udpipe**
- official online viewer for CoNNL-U files: **universaldependencies.org/conllu_viewer.html**
- latest version (2.13) of the UD treebanks: **lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5287**
- to contact me after this lecture: **arianna.masciolini@gu.se**

Thank you for today!