ČESKÝ NÁRODNÍ
KORPUS

ČESKÝ NÁRODNÍ KORPUS

The *InterCorp* parallel corpus with a uniform annotation for all languages

Alexandr Rosen
Institute of the Czech National Corpus
Faculty of Arts, Charles University, Prague

Slovko 2023
12th International Conference
Natural Language Processing and Corpus Linguistics
Ľudovít Štúr Institute of Linguistics
Bratislava, 18–20th October 2023

OUTLINE

1. Linguistic categories and corpus annotation
2. InterCorp – a multilingual parallel corpus
3. InterCorp annotated by Universal Dependencies
4. Universal Dependencies in the KonText search interface
5. Using Universal Dependencies to query InterCorp
6. Other merits of a uniform linguistic annotation

OUTLINE

1. Linguistic categories and corpus annotation
2. InterCorp – a multilingual parallel corpus
3. InterCorp annotated by Universal Dependencies
4. Universal Dependencies in the KonText search interface
5. Using Universal Dependencies to query InterCorp
6. Other merits of a uniform linguistic annotation

1. Linguistic categories and corpus annotation

- Standard linguistic categories do not mean the same across languages
- Haspelmath (2010): comparative concepts, mapped to language-specific categories, **but:**
- A language-universal annotation scheme UD (*Universal Dependencies*) is gaining ground in corpus linguistics

de Marneffe et al. (2021)

https://universaldependencies.org

OUTLINE

1. Linguistic categories and corpus annotation
2. InterCorp – a multilingual parallel corpus
3. InterCorp annotated by Universal Dependencies
4. Universal Dependencies in the KonText search interface
5. Using Universal Dependencies to query InterCorp
6. Other merits of a uniform linguistic annotation

2. InterCorp – a multilingual parallel corpus

- Charles University, Institute of the Czech National Corpus

  https://www.korpus.cz

  https://intercorp.korpus.cz/?lang=en

- Since 2008, v13**ud** available, v16 just released, v16**ud** due soon

  https://kontext.korpus.cz

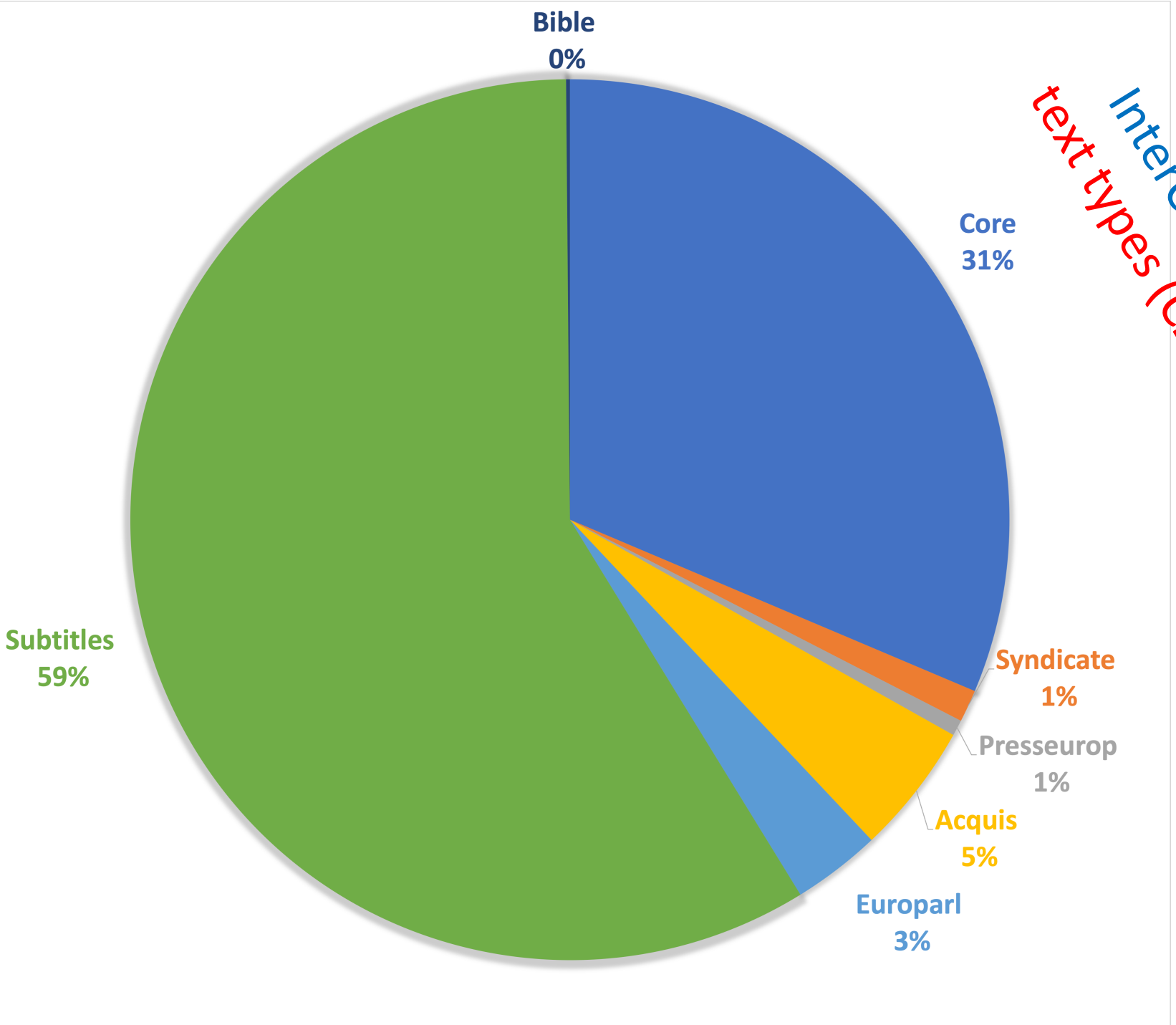
  https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze16

- 61 languages (4.9 bill. words) + Czech (0.4 bill. words)
- Each text in Czech and at least one foreign language
- Log in for all features: https://www.korpus.cz/login

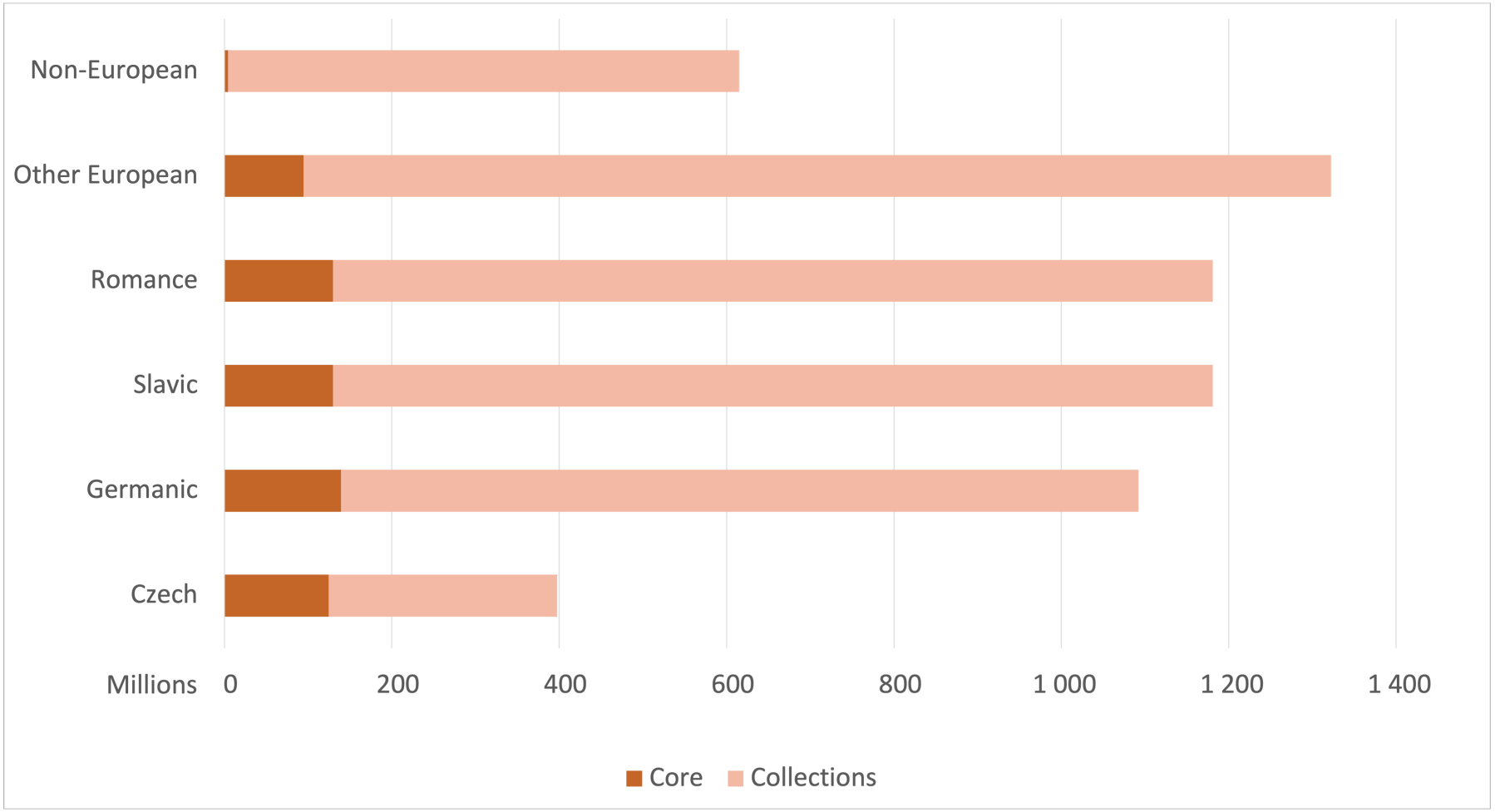
  user: **ic_ud** pw: **UnivDeps**

  or: institutional login (Shibboleth)

InterCorp v16 text types (Czech)



InterCorp v16
Language groups



InterCorp v16
62 languages

Afrikaans Albanian Arabic Armenian Basque **Belarusian** Bengali Bosnian Breton **Bulgarian** Catalan Chinese **Croatian Czech** Danish **Dutch English** Esperanto Estonian **Finnish French** Galician Georgian **German** Greek Hebrew Hindi Hungarian Icelandic Indonesian **Italian Japanese** Kazakh Korean **Latvian** Lithuanian Macedonian Malay Malayalam Maltese **Norwegian** Persian **Polish Portuguese** Romani Romanian **Russian** Serbian Sinhala **Slovak Slovene Spanish Swedish** Tagalog Tamil Telugu Thai Turkish Ukrainian Upper Sorbian Urdu Vietnamese

Linguistic annotation – language-specific (v16)

= lemmatization and tagging

Strategy:

use available tools (taggers), including:

- Tokenization bundled with the tool
- Existing tagsets
- Models trained elsewhere

Result:

tokenization, lemmatization and tagsets differ both conceptually and formally

Language-specific tools and tags (v16)

Lng	Tool	Preposition Determiner Adjective Noun
be	UD	ADP ADJ Case=Loc Degree=Pos Gender=Masc Number=Sing NOUN Animacy=Inan Case=Loc .
bg	TT	R Pde-os-n Ansi Ncnsi
ca	TT	ADP.Prep DET.Masc.Sing.Dem NOUN.Masc.Sing ADJ.Masc.Sing
cs	Morče	RR-6 PDXP6 AAFP6---3A NNFP6---A
de	RFT	APPR ART:Def:Dat:Pl:Masc ADJA:Pos:Dat:Pl:Masc N:Reg:Dat:Pl:Masc
en	TT	IN DT JJS NNS
es	TT	PREP ART NC ADJ
et	TT	P.sg.gen A.pos.sg.gen S.com.sg.kom
fi	OMorFi	A:Sg:Gen:Pos N:Sg:Gen Adp:Po
fr	TT	PRP DET:ART ADJ NOM
hr	ReIDI	Sl Pd-msl Agpmsly Ncmsl
hu	RFT	P:d:3:s:n T:f A:f:p:s: N:c:s:n
is	IceTagger	ao lhfove nhfog
it	TT	PRE PRO:demo NOM ADJ
lv	LVTagger	spsgy pd0msgn afmsgyp ncmsg1
nl	TT	prep det__demo adj nounpl
no	VISL	600 370 103 000 prep det adj subst
pl	TaKIPI	prep:loc:nwok adj:sg:loc:m3:pos adj:sg:loc:m3:pos subst:sg:loc:m3
pt	TT	SPS DA0 NCFS AQ0
ru	TT	Sp-l P--pl Afp-plf Ncmpln
sk	Morče	Eu6 PFfs6 AAfs6x SSfs6
sl	totale	Sl Pd-nsg Agpfsg Ncnsl
sr	ReLDI	Sa Pd-fsa Agpfsay Ncfsa
sv	Stagger	PP DT:NEU:SIN:DEF JJ:POS:UTR/NEU:SIN:DEF:NOM NN:NEU:SIN:IND:NOM
uk	UD	ADP Case=Loc PRON Animacy=Inan Case=Loc Gender=Neut Number=Sing PronType=Dem ADJ Case=Loc Degree=Pos Gender=Masc Number=Sing NOUN Animacy=Inan Case=Loc Gender=Masc Number=Sing

OUTLINE

1. Linguistic categories and corpus annotation
2. InterCorp – a multilingual parallel corpus
3. InterCorp annotated by Universal Dependencies
4. Universal Dependencies in the KonText search interface
5. Using Universal Dependencies to query InterCorp
6. Other merits of a uniform linguistic annotation

Linguistic annotation – language-universal **(v13ud)**

Strategy for all languages:

- Use the same annotation concepts and scheme
- Use a single tool

Why Universal Dependencies?

https://universaldependencies.org

- A de-facto standard for linguistic annotation
- Data and models for many languages
- Bonus: syntactic annotation
- Several parsers, including UDPipe

  https://lindat.mff.cuni.cz/services/udpipe/
- Active community of developers and users

UD Guidelines v.2 (2016, v.1: 2014)

- 17 parts of speech – **`upos`**

  https://universaldependencies.org/u/pos/index.html
- 24 morphological categories – **`feats`**

  https://universaldependencies.org/u/feat/index.html
- 37 syntactic functions – **`deprel`**

  https://universaldependencies.org/u/dep/index.html

Universal POS tags

`[upos="ADJ"]`

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Universal features [feats="VerbForm=Fin"]

Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

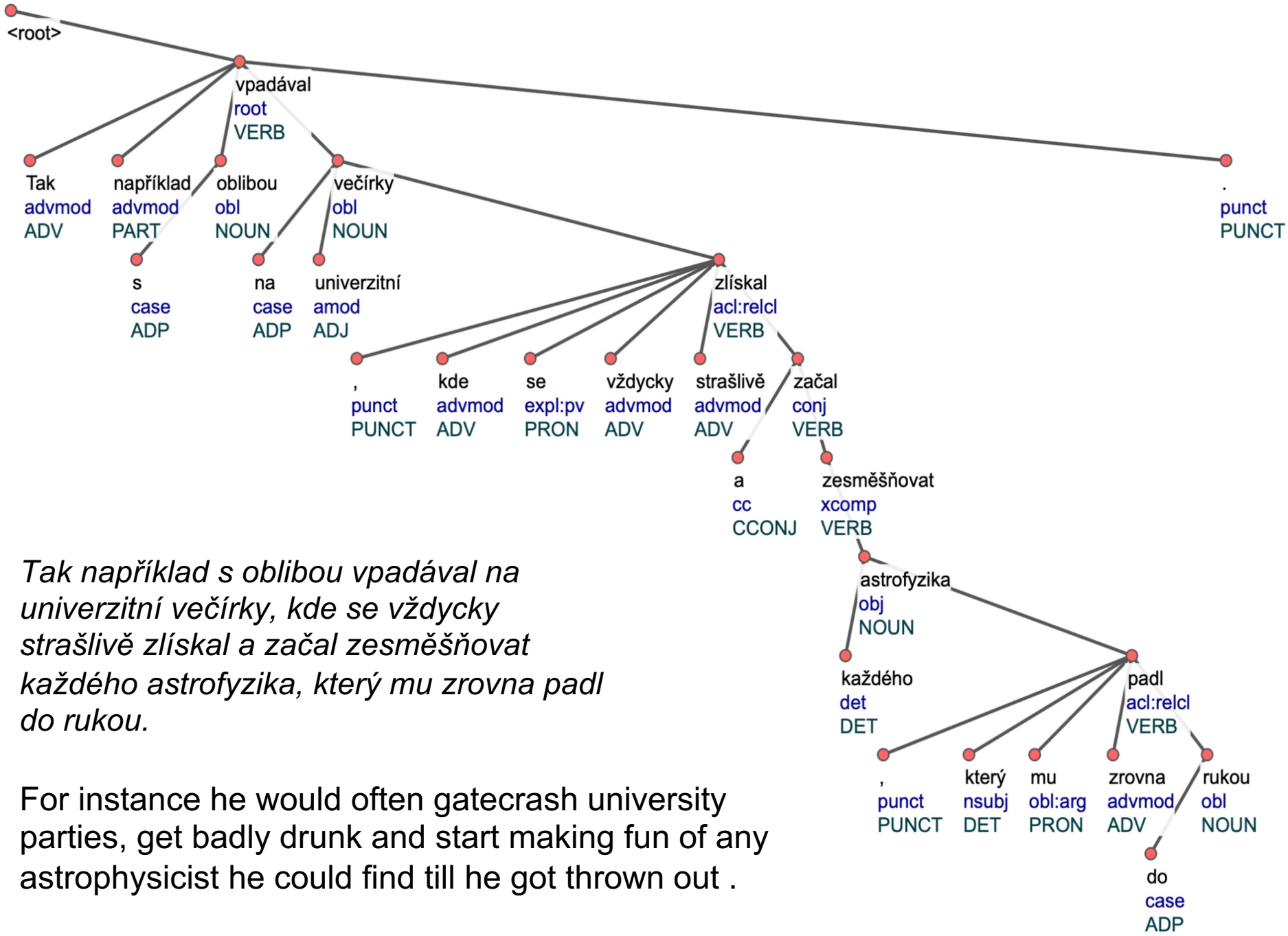
Universal dependency relations

[deprel="acl"]

MORPHOSYNTACTIC CATEGORIES

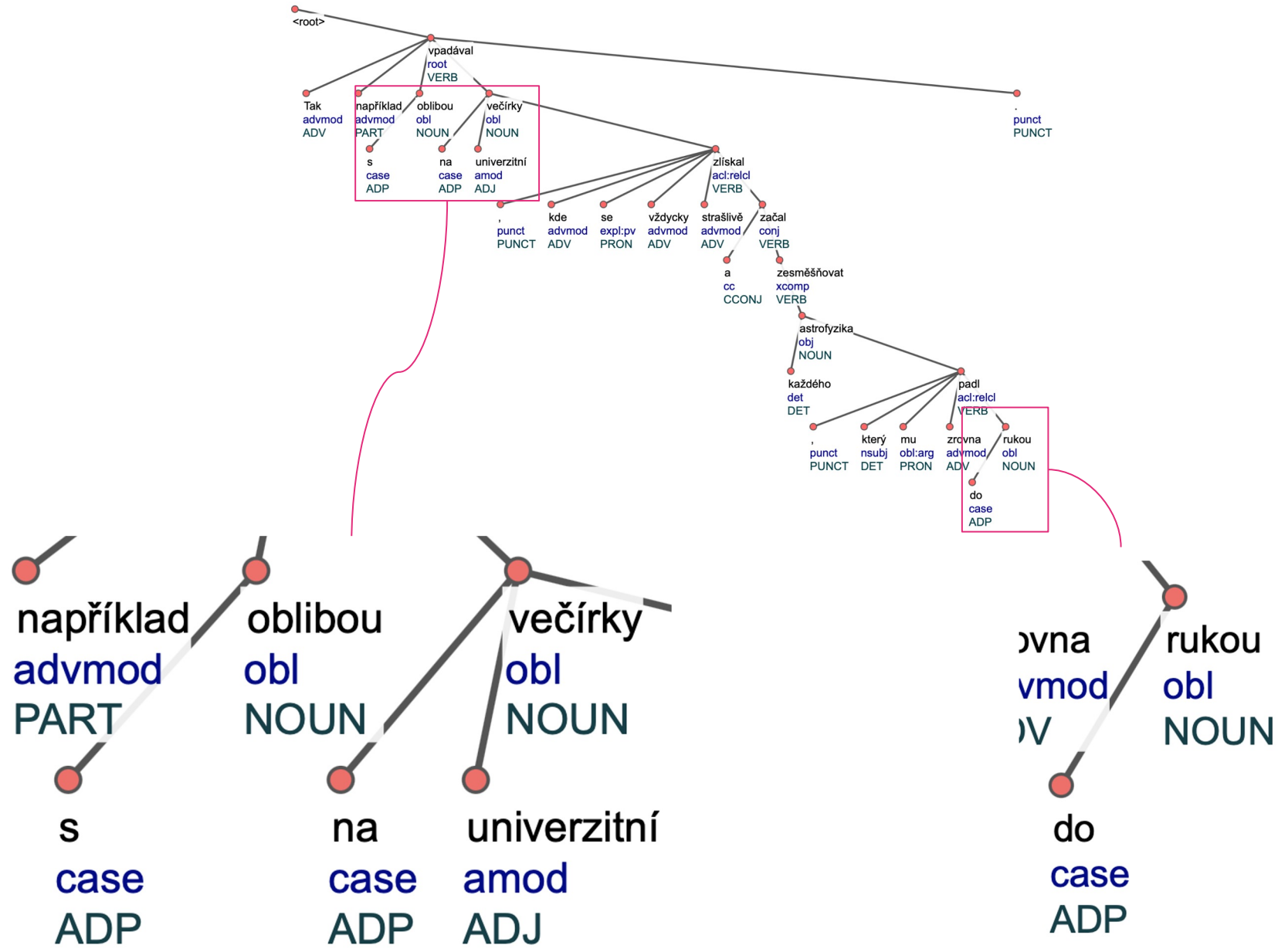
SYNTACTIC FUNCTIONS

	Nominals	Clauses	Modifier words	Function words
Core arguments	nsubj	csubj		
	obj	ccomp		
	iobj	xcomp		
Non-core dependents	obl	advcl	advmod	aux
	vocative		discourse	cop
	expl			mark
	dislocated			
Nominal dependents	nmod	acl	amod	det
	appos			clf
	nummod			case

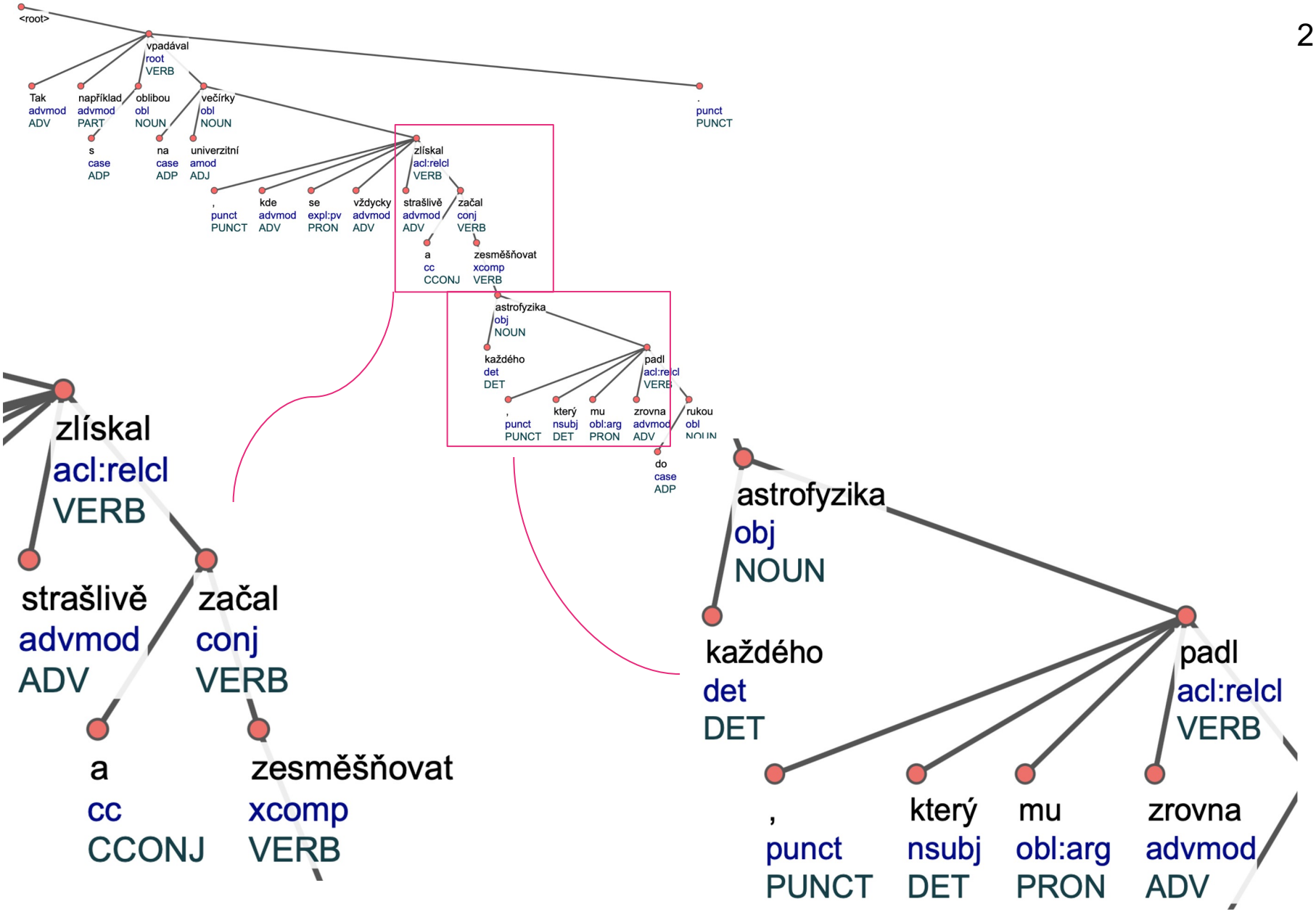


Tak například s oblibou vpadával na univerzitní večírky, kde se vždycky strašlivě zlískal a začal zesměšňovat každého astrofyzika, který mu zrovna padl do rukou.

For instance he would often gatecrash university parties, get badly drunk and start making fun of any astrophysicist he could find till he got thrown out .



Tak například s oblibou vpadával na univerzitní večírky, kde se vždycky strašlivě zlískal a začal zesměšňovat každého astrofyzika, který mu zrovna padl do rukou.



Tak například s oblibou vpadával na univerzitní večírky, kde se vždycky strašlivě zlískal a začal zesměšňovat každého astrofyzika, který mu zrovna padl do rukou.

UDPipe output: the CONLL-U format

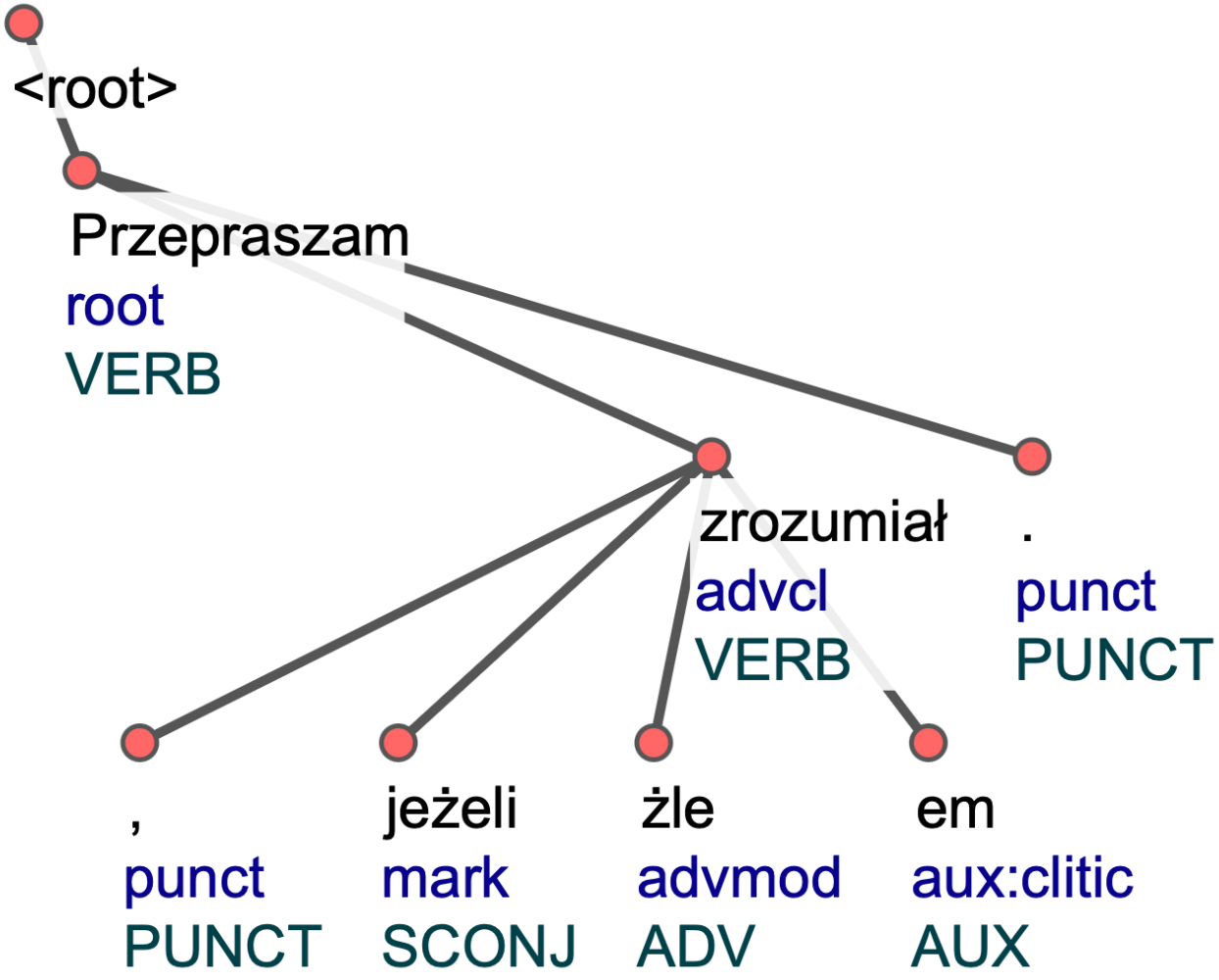
A table with each token on a new line, 10 columns:

1. ID – order no. for each token, interval for fused words
2. FORM
3. LEMMA
4. UPOS – UD part of speech
5. XPOS – language-specific (legacy) tag
6. FEATS – list of morphological categories
7. HEAD – ID of the token's governor, the root = 0
8. DEPREL – syntactic function
9. DEPS – reserved for *Enhanced Dependencies*
10. MISC – varia (e.g. no space in between tokens)

1	*Tak*	CCONJ	_	5	cc
2	*například*	ADV	_	5	advmod
3	*s*	ADP	AdpType=Prep|Case=Ins	4	case
4	*oblibou*	NOUN	Case=Ins|Gender=Fem|Number=Sing|Polarity=Pos	5	obl
5	*vpadával*	VERB	Aspect=Imp|Gender=Masc|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act	0	root
6	*na*	ADP	AdpType=Prep|Case=Acc	8	case
7	*univerzitní*	ADJ	Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing|Polarity=Pos	8	amod
8	*večírky*	NOUN	Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur	5	obl:arg
9	*,*	PUNCT	_	14	punct
10	*kde*	ADV	PronType=Int,Rel	14	advmod
11	*se*	PRON	Case=Acc|PronType=Prs|Reflex=Yes|Variant=Short	14	expl:pv
12	*vždycky*	ADV	_	14	advmod
13	*strašlivě*	ADV	Degree=Pos|Polarity=Pos	14	advmod
14	*zlískal*	VERB	Aspect=Perf|Gender=Masc|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act	8	**acl**
15	*a*	CCONJ	_	16	cc
16	*začal*	VERB	Gender=Masc|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act	**14**	**conj**
17	*zesměšňovat*	VERB	Aspect=Imp|Polarity=Pos|VerbForm=Inf	16	xcomp
18	*každého*	DET	Animacy=Anim|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing|Polarity=Pos|PronType=Tot	19	det
19	*astrofyzika*	NOUN	Animacy=Anim|Case=Acc|Gender=Masc|Number=Sing	17	obj
20	*,*	PUNCT	_	24	punct
21	*který*	DET	Case=Nom|Gender=Masc|Number=Sing|PronType=Int,Rel	24	nsubj
22	*mu*	ADV		24	advmod

Przepraszam, jeżeli źle zrozumiałem.

'I apologize if I didn't understand well.'



Przepraszam, jeżeli źle zrozumiałem.
'I apologize if I didn't understand well.'

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	*Przepraszam*	przepraszać	VERB	fin:sg:pri:imperf	Aspect=Imp|Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act	0	root	_	SpaceAfter=No
2	*,*	,	PUNCT	interp	PunctType=Comm	5	punct	_	
3	*jeżeli*	jeżeli	SCONJ	comp	_	5	mark	_	
4	*źle*	źle	ADV	adv:pos	Degree=Pos	5	advmod	_	
5-6	*zrozumiałem*	_	_	_	_	_	_	_	SpaceAfter=No
5	*zrozumiał*	zrozumieć	VERB	praet:sg:m1:perf	Animacy=Hum|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act	1	advcl	_	_
6	*em*	być	AUX	aglt:sg:pri:imperf:wok	Aspect=Imp|Clitic=Yes|Number=Sing|Person=1|Variant=Long	5	aux:clitic	_	_
7	.	.	PUNCT	interp	PunctType=Peri	1	punct	_	SpaceAfter=No

OUTLINE

1. Linguistic categories and corpus annotation
2. InterCorp – a multilingual parallel corpus
3. InterCorp annotated by Universal Dependencies
4. UD in the KonText search interface
5. Using Universal Dependencies to query InterCorp
6. Other merits of a uniform linguistic annotation

Why CONLL-U is not good enough? (1/2)

- Double tokenization: orthography vs. syntax
- Solution: orthographical words as tokens, syntactic words as multivalues

word	*abys*	*zrozumiałem*
sword	aby|bys	zrozumiał|em
iword	a|bys	zrozumiał|em
lemma	aby|být	zrozumieć|być
upos	SCONJ|AUX	VERB|AUX
xpos	J,----------|Vc-S---2-------	praet:sg:m1:perf|aglt:sg:pri:imperf:wok
feats	||Mood=Cnd|Number=Sing|Person=2|VerbForm=Fin	..|Tense=Past|VerbForm=Fin|Voice=Act||Clitic=Yes|Number=Sing|Person=1|..
deprel	mark|aux	root|aux:clitic

Why CONLL-U is not good enough? (2/2)

- Searching syntactic structure
  - The CQL **meet** command with a global condition

```
(meet 1:[upos="VERB"] 2:[deprel="nsubj" & lemma="dog"])
& 1.ID = 2.HEAD within <s/>
```

  - Instead: add attributes about the head's properties

    lemma, **deprel**, **xpos**, **feats** (cf. syn2020)

- Searching for properties of function words
  - Instead: raise function verb attributes to content words

- Morphological categories as a list of attribute=value pairs
  - Instead: add selected categories as standard attributes

Strategy of modifying CONLL-U for Manatee

To facilitate:

- … navigation within syntactic structure (**`p_lemma`**), **add**:
  - the head's **`lemma`**, **`upos`**, **`feats`**, **`deprel`** and relative position

- … access to function words' properties (**`aux_feats`**, **`case_lemma`**), **add its**:
  - **`lemma`**, **`upos`**, **`feats`** and **`deprel`**'s subtype

- … search and statistics based on some categories, **add:**
  - some categories from the **`feats`** list

Keep the new attributes at a minimum

- Only those that make sense for a given language
- Between 20 and 44

Field	Attribute	ar	be	bg	ca	cs	da	de	el	en	es	et	fi	fr	he	hi	hr	hu	it	ja	lt	lv	mt	nl	no	pl	pt	ro	ru	sk	sl	sr	sv	tr	uk	vi	zh	Tota	Note	Gloss
1	word	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		word form
2	sword	1			1	1		1	1	1	1		1	1	1				1							1	1							1	1		15		<word> split into interpreted (restored) syntactic words
3	iword	1			1	1		1		1	1		1	1					1							1	1								1		12		<word> split into syntactic words without altering the original form
4	lc																																				0	dynamic	lowercase <word>
5	lemma	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		lemma
6	lc_lemma																																				0	dynamic	lowercase <lemma>
7	upos	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		UD POS tag
8	xpos	1		1	1	1		1	1		1	1	1		1	1	1		1	1	1	1	1	1		1	1	1		1	1	1	1	1	1	1	29		language-specific tag
9	feats	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	35		UD morphological categories
10	id	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		word index within sentence
11	head	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		<id> of the token's head
12	deprel	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		UD syntactic function
13	parent	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		relative position of <head>
14	p_lemma	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		<lemma> of <head>
15	p_upos	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		<upos> of <head>
16	p_feats	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		<feats> of <head>
17	p_deprel	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		<deprel> of <head>
18	e_id	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		<id> of effective head
19	eparent	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		relative position of effective head
20	aux_lemma	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1				1	1		1	1	1	1	1	1	1		1	1	1	30		<lemma> of the token's auxiliary verb
21	aux_upos																																			1	1	(AUX)	<upos> of the token's auxiliary verb
22	aux_feats	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1		1	1	1		1	1	1	1	1	1	1	1	31		<feats> of the token's auxiliary verb
23	aux_type	1	1	1		1		1	1	1	1		1	1					1			1	1	1	1	1	1	1	1	1			1	1		1	1	24		type of the token's auxiliary verb
24	case_lemma	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	35		<lemma> of the token's adposition
25	case_upos																																				0	(ADP)	<upos> of the token's adposition
26	case_feats				1	1					1	1	1			1	1				1	1				1		1		1	1	1			1		15		<feats> of the token's adposition
27	case_type														1								1														1	3		type of the token's adposition
28	clf_lemma																																				1		<lemma> of the token's classifier
29	clf_upos																																				0		<upos> of the token's classifier
30	clf_feats																																				0		<feats> of the token's classifier
31	clf_type																																				0		type of the token's classifier
32	cop_lemma				1		1				1				1	1		1							1	1			1				1			1	11		<lemma> of the token's copula
33	cop_upos																						1											1				2	(AUX)	<upos> of the token's copula
34	cop_feats	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1		1	1	1	1	1	1	1	1		31		<feats> of the token's copula
35	cop_type												1										1														2		type of the token's copula
36	det_lemma	1		1	1		1	1	1	1	1	1	1	1	1	1		1	1	1				1	1		1	1					1	1		1	1	24		<lemma> of the token's determiner
37	det_upos			1				1	1		1	1			1								1	1	1			1						1			1	12		<upos> of the token's determiner
38	det_feats	1		1	1		1	1	1	1	1	1	1	1	1	1		1	1					1	1			1					1	1			20		<feats> of the token's determiner
39	det_type							1	1						1				1							1											5		type of the token's determiner
40	mark_lemma	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1		1	1	1	1	1	1		1	1	1	33		<lemma> of the token's marker
41	mark_upos	1	1	1	1		1	1	1	1	1	1			1	1	1	1	1	1		1	1	1	1		1	1	1	1		1	1				1	27		<upos> of the token's marker
42	mark_feats		1				1																			1		1	1	1							6		<feats> of the token's marker
43	mark_type														1																						1	2		type of the token's marker
44	Abbr	1	1			1	1		1	1		1	1		1						1	1		1	1	1		1	1	1	1		1	1	1			21		abbreviation
45	Aspect	1	1	1		1			1							1		1			1	1				1			1	1	1			1	1		1	16		
46	Case	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1			1	1		1	1	1	1	1	1	1	1	1	1	1	1		1	31		
47	Definite	1		1	1		1	1	1	1	1			1	1		1	1	1		1	1		1	1		1	1			1	1	1					22		
48	Degree		1	1	1	1	1	1	1	1	1	1					1	1	1		1	1		1	1	1	1	1	1	1				1			26		
49	Foreign	1	1			1	1	1	1	1		1	1	1		1			1		1	1		1		1		1	1	1	1		1		1			22		
50	Gender	1	1	1	1	1	1	1	1	1	1			1	1	1	1		1		1	1		1	1	1	1	1	1	1	1	1	1		1			28		
51	Mood	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1			1	1		1	1	1	1	1	1	1	1		1	31		
52	Number	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1	1	1	1	1		1	33		
53	NumType		1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1		1	1		1	1	1	1		1	1	1	1	1	1	1	1	1	30		type of numeral
54	Person	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1		1	1		1	32		
55	Polarity		1	1	1	1		1		1	1	1	1	1	1	1	1		1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	30		
56	Poss		1	1	1	1	1	1	1	1	1	1		1		1	1	1	1			1		1	1	1		1		1	1	1	1		1			25		
57	PronType	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1	1	1		1	1	1	1	1	1			31		type of pronoun
58	Reflex		1	1	1	1	1	1		1	1	1	1	1	1		1	1			1	1		1		1		1		1	1	1		1	1			24		reflexive form
59	Tense		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1	1	1	1	1			30		
60	VerbForm	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1		1	1	1	1	1	1	1	1			31		verb form
61	Voice	1	1	1		1	1	1	1			1	1		1	1	1	1			1	1			1	1			1	1		1	1	1	1		1	24		

The CONLL-U attributes

	Attribute	Total	Gloss
1	**word**	36	word form
2	**sword**	15	<word> split into interpreted (restored) syntactic words
3	**iword**	12	<word> split into syntactic words without altering the original form
4	**lc**	0	lowercase <word>
5	**lemma**	36	lemma
6	**lc_lemma**	0	lowercase <lemma>
7	**upos**	36	UD POS tag
8	**xpos**	29	language-specific tag
9	**feats**	35	UD morphological categories
10	**id**	36	word index within sentence
11	**head**	36	<id> of the token's head
12	**deprel**	36	UD syntactic function

Attributes concerning syntactic structure

ID	Attribute	Total	Gloss
13	**parent**	36	relative position of <head>
14	**p_lemma**	36	<lemma> of <head>
15	**p_upos**	36	<upos> of <head>
16	**p_feats**	36	<feats> of <head>
17	**p_deprel**	36	<deprel> of <head>
18	**e_id**	36	<id> of effective head
19	**eparent**	36	relative position of effective head

Attributes concerning function words

20	**aux_lemma**	30	<lemma> of the token's auxiliary verb
21	**aux_upos**	1	<upos> of the token's auxiliary verb
22	**aux_feats**	31	<feats> of the token's auxiliary verb
23	**aux_type**	24	type of the token's auxiliary verb
24	**case_lemma**	35	<lemma> of the token's adposition
25	**case_upos**	0	<upos> of the token's adposition
26	**case_feats**	15	<feats> of the token's adposition
27	**case_type**	3	type of the token's adposition
28	**clf_lemma**	1	<lemma> of the token's classifier
29	**clf_upos**	0	<upos> of the token's classifier
30	**clf_feats**	0	<feats> of the token's classifier
31	**clf_type**	0	type of the token's classifier
32	**cop_lemma**	11	<lemma> of the token's copula
33	**cop_upos**	2	<upos> of the token's copula
34	**cop_feats**	31	<feats> of the token's copula
35	**cop_type**	2	type of the token's copula
36	**det_lemma**	24	<lemma> of the token's determiner
37	**det_upos**	12	<upos> of the token's determiner
38	**det_feats**	20	<feats> of the token's determiner
39	**det_type**	5	type of the token's determiner
40	**mark_lemma**	33	<lemma> of the token's marker
41	**mark_upos**	27	<upos> of the token's marker
42	**mark_feats**	6	<feats> of the token's marker
43	**mark_type**	2	type of the token's marker

Attributes concerning some morphological categories

44	**Abbr**	21	abbreviation
45	**Aspect**	16	
46	**Case**	31	
47	**Definite**	22	
48	**Degree**	26	
49	**Foreign**	22	
50	**Gender**	28	
51	**Mood**	31	
52	**Number**	33	
53	**NumType**	30	type of numeral
54	**Person**	32	
55	**Polarity**	30	
56	**Poss**	25	possessive
57	**PronType**	31	type of pronoun
58	**Reflex**	24	reflexive form
59	**Tense**	30	
60	**VerbForm**	31	verb form
61	**Voice**	24	

OUTLINE

1. Linguistic categories and corpus annotation
2. InterCorp – a multilingual parallel corpus
3. InterCorp annotated by Universal Dependencies
4. Universal Dependencies in the KonText search interface
5. Using Universal Dependencies to query InterCorp
6. Other merits of a uniform linguistic annotation

Create / edit a tag

Selected features:

Case = Ins × & Gender = Fem × & Number = Plur × & POS = NOUN ×

Part of speech:

- [ ]
- [ ] ADJ
- [ ] ADP
- [ ] ADV
- [ ] AUX
- [ ] CCONJ
- [ ] DET
- [ ] INTJ
- [x] NOUN
- [ ] NUM
- [ ] PART
- [ ] PRON
- [ ] PROPN
- [ ] PUNCT
- [ ] SCONJ
- [ ] SYM
- [ ] VERB
- [ ] X

Features:

Abbr (0)
AdpType (0)
Animacy (0)
Aspect (0)
Case (6)
Clitic (0)
ConjType (0)
Degree (1)
Emphatic (0)
Foreign (0)
Gender (3)
Hyph (0)
Mood (0)
Number (2)
Number[psor] (0)
NumForm (1)
NumType (0)
PartType (0)
Person (0)
Polarity (0)

- [ ] Acc
- [ ] Dat
- [ ] Gen
- [x] Ins
- [ ] Loc
- [ ] Nom
- [ ] Voc

Insert

Undo Reset

Searching for morphological categories

```
[upos="NOUN"
& feats="Gender=Fem"
& feats="Number=Plur"
& feats="Case=Ins"]

[upos="NOUN"
& feats=".*Case=Ins.*Gender=Fem.*Number=Plur.*"]

[upos="NOUN"
& gender="Fem"
& case="Ins"
& number="Plur"]

[xpos="NNFP7.*"]
```

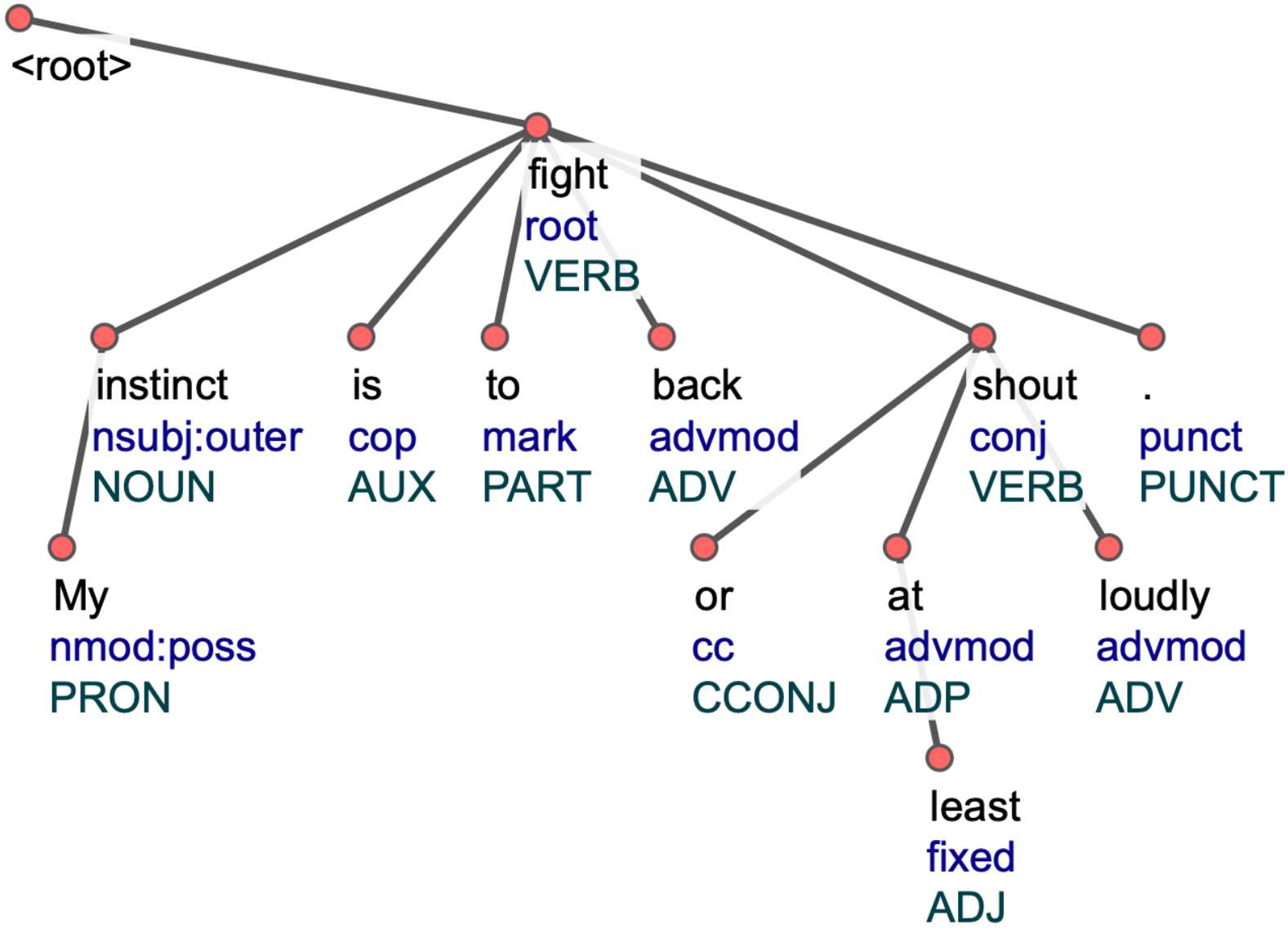
Searching for syntactic functions

run as the head of an adnominal clause:

```
[lemma="run" & deprel="acl"]
```

Everyone of the rabbits was seized by the instinct ***to run away****.*

Some people have the idea that rabbits spend a good deal of their time ***running away from foxes.***


<root>
fight
root
VERB
instinct
nsubj:outer
NOUN
is
cop
AUX
to
mark
PART
back
advmod
ADV
shout
conj
VERB
.
punct
PUNCT
My
nmod:poss
PRON
or
cc
CCONJ
at
advmod
ADP
loudly
advmod
ADV
least
fixed
ADJ

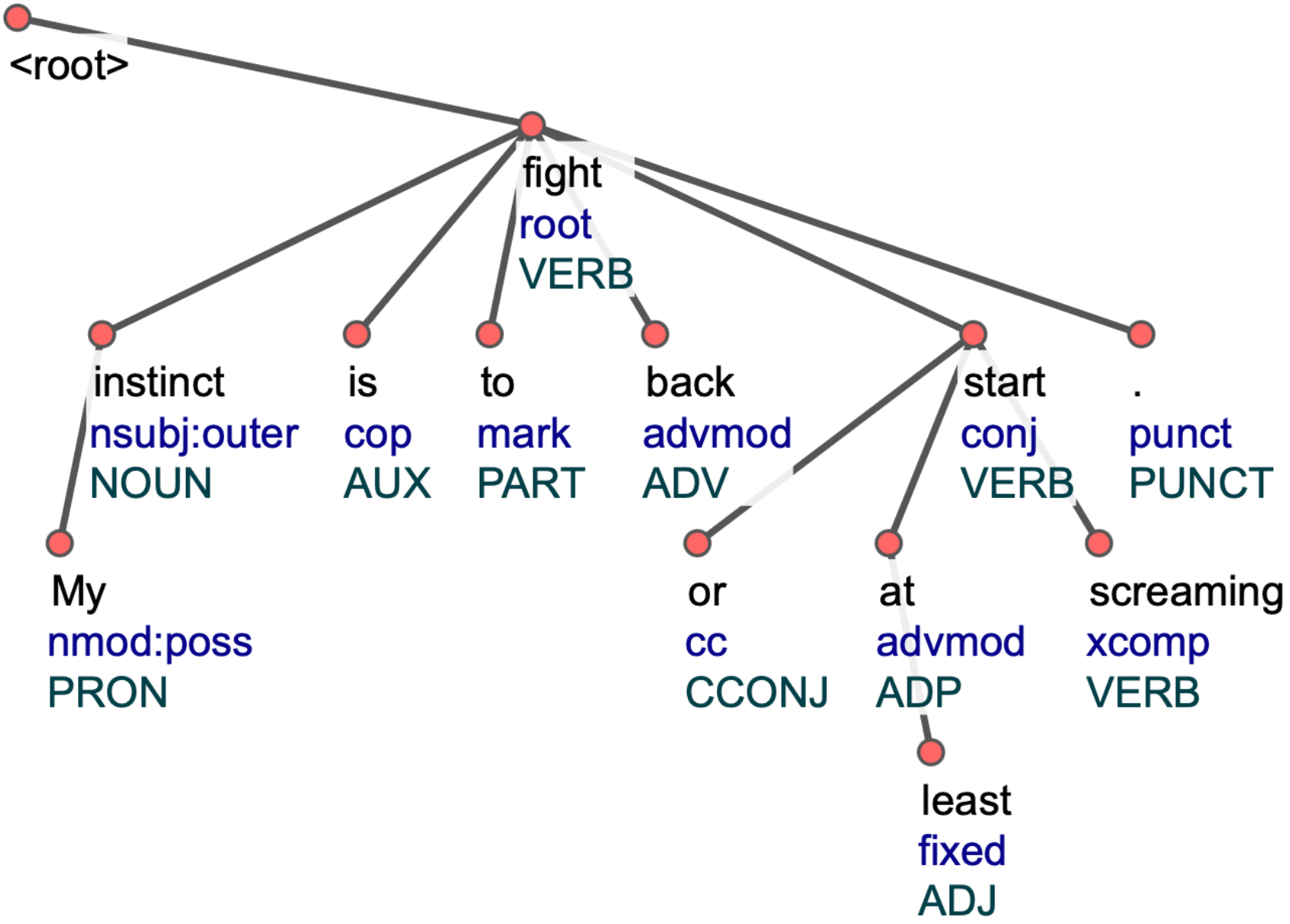
Coordination

- Non-initial conjuncts are marked as `deprel="conj"`
- To query all conjuncts:

```
[deprel="obj" | deprel="conj" & p_deprel="obj"]
```

- To avoid this in InterCorp 16ud:

An attribute in non-initial conjuncts copies the initial conjunct's deprel, otherwise `= deprel`.


<root>
fight
root
VERB
instinct
nsubj:outer
NOUN
is
cop
AUX
to
mark
PART
back
advmod
ADV
start
conj
VERB
.
punct
PUNCT
My
nmod:poss
PRON
or
cc
CCONJ
at
advmod
ADP
screaming
xcomp
VERB
least
fixed
ADJ

OUTLINE

1. Linguistic categories and corpus annotation
2. InterCorp – a multilingual parallel corpus
3. InterCorp annotated by Universal Dependencies
4. Universal Dependencies in the KonText search interface
5. Using Universal Dependencies to query InterCorp
6. Other merits of a uniform linguistic annotation

Complexity measures

To appear in InterCorp v16ud as metadata for:

- Sentences
- Texts
- Text types

Useful for:

- L1 or L2 learning/teaching
- Contrastive studies

Complexity measures

1. Lexical diversity

2. Syntactic complexity

According to syntactic category:

- Clausal complexity
- Noun phrase complexity

According to dimension

- Vertical (levels of embedding)
- *Horizontal (number of subtree nodes)*

Complexity measures

- Lexical diversity
  - lexical types within a moving window of 1000 tokens
- Syntactic complexity
  - maximum tree depth (for clauses)
  - *subordination ratio: T-units + subord.clauses / T-units*
  - average NP tree depth
  - *average NP complexity: nominal dependents / nouns*

References

- Croft, W., Nordquist, D., Looney, K., and Regan, M. **2017**. Linguistic typology meets Universal Dependencies. In Dickinson, M., Hajič, J., Kübler, S., and Przepiórkowski, A., editors, *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75. Indiana University, Bloomington, Bloomington, IN, USA.
- Křen, M., Rosen, A., Štourač, M., Vavřín, M., and Vondřička, P. **2011**. Paralelní korpus InterCorp po sedmi letech. In Čermák, F., editor, Korpusová lingvistika Praha 2011: 2 – Výzkum a výstavba korpusů, volume 15 of Studie z korpusové lingvistiky, pages 105–115, Praha. Ústav Českého národního korpusu.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, Daniel Zeman. **2020**. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4034-4043, European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4.
- Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, Daniel Zeman **2021**. Universal Dependencies. In: *Computational Linguistics*, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.
- Osborne, T. and Gerdes, K. **2019**. The status of function words in dependency grammar: A critique of universal dependencies (UD). *Glossa: a journal of general linguistics*, 4(1):17.
- Przepiórkowski, A. and Patejuk, A. **2018**. Arguments and adjuncts in Universal Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3837–3852, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Tuora, R., Przepiórkowski, A., and Leczkowski, A. **2021**. Comparing learnability of two dependency schemes: 'semantic' (UD) and 'syntactic' (SUD). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2987–2996, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thank you for your kind attention!

Origins of Universal Dependencies

- Stanford Dependencies 2005: content words as heads
  https://nlp.stanford.edu/software/stanford-dependencies.html
- Google Universal Tagset 2007: 12 parts-of-speech
  https://github.com/slavpetrov/universal-pos-tags
- Interset 2006: a single set of morphological categories for shared tasks – Conference on Computational Natural Language Learning
  https://github.com/dan-zeman/interset
- CONLL-X 2007: format for shared tasks
  https://web.archive.org/web/20160814191537/http://ilk.uvt.nl/conll/#dataformat

	Nominals	Clauses	Modifier words	Function words
Core arguments	nsubj *nominal subject*	csubj *clausal subject*		
	obj *object*	**ccomp** *clausal complement*		
	iobj *indirect object*	**xcomp** *open clausal complement*		
Non-core dependents	obl *oblique nominal*	**advcl** *adverbial clause modifier*	advmod *adverbial modifier*	aux *auxiliary verb*
	vocative		discourse *~element*	cop *copula*
	expl *expletive*			mark *marker* (*sconj*)
	dislocated *~element*			
Nominal dependents	nmod *nominal modifier*	**acl** *adnominal clause*	**amod** *adjectival modifier*	det *determiner*
	appos *appositional modifier*			clf *classifier*
	nummod *numeric modifier*			case *case marking* (*e.g. prepositions*)

Coordination	MWE	Loose	Special	Other
conj *conjunct*	fixed *multiword expression*	list	orphan (*when head is elided*)	punct *punctuation*
cc *coordinating conjunction*	flat *multiword expression*	**parataxis** (*direct speech*)	goeswith (*split words*)	**root**
	compound		reparandum *overridden disfluency*	dep *unspecified dependency*