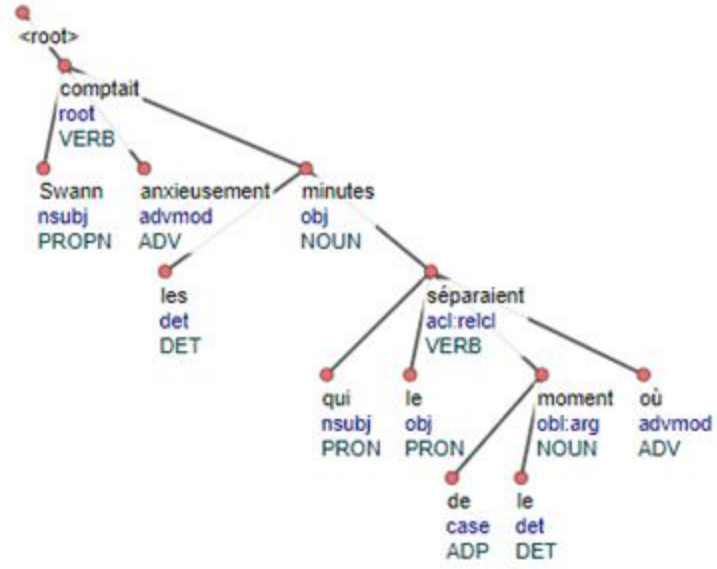
Vyhledávání v paralelním korpusu za použití anotace Universal Dependencies

Olga Nádvorníková[1] **Alexandr Rosen[2]**

[1] Ústav románských studií
[2] Ústav Českého národního korpusu

Bienále české lingvistiky
Filozofická fakulta Univerzity Karlovy
Praha, 17. září 2024



Osnova

1. Úvod 14:30-14:40
2. Paralelní korpus InterCorp 14:40-14:50
3. Anotace InterCorpu 14:50-15:05
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD 15:05-15:30
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity 15:30-15:40
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu 15:40-16:00
    3. Ukaž a hledej 16:00-16:20
6. Diskuse, otázky… 16:20-16:30

Tato prezentace:
https://shorturl.at/i1kkM

20. a 27. března 2024:
InterCorp a Universal Dependencies: nové možnosti výzkumu

Prezentace a video:
https://shorturl.at/SS1Ho

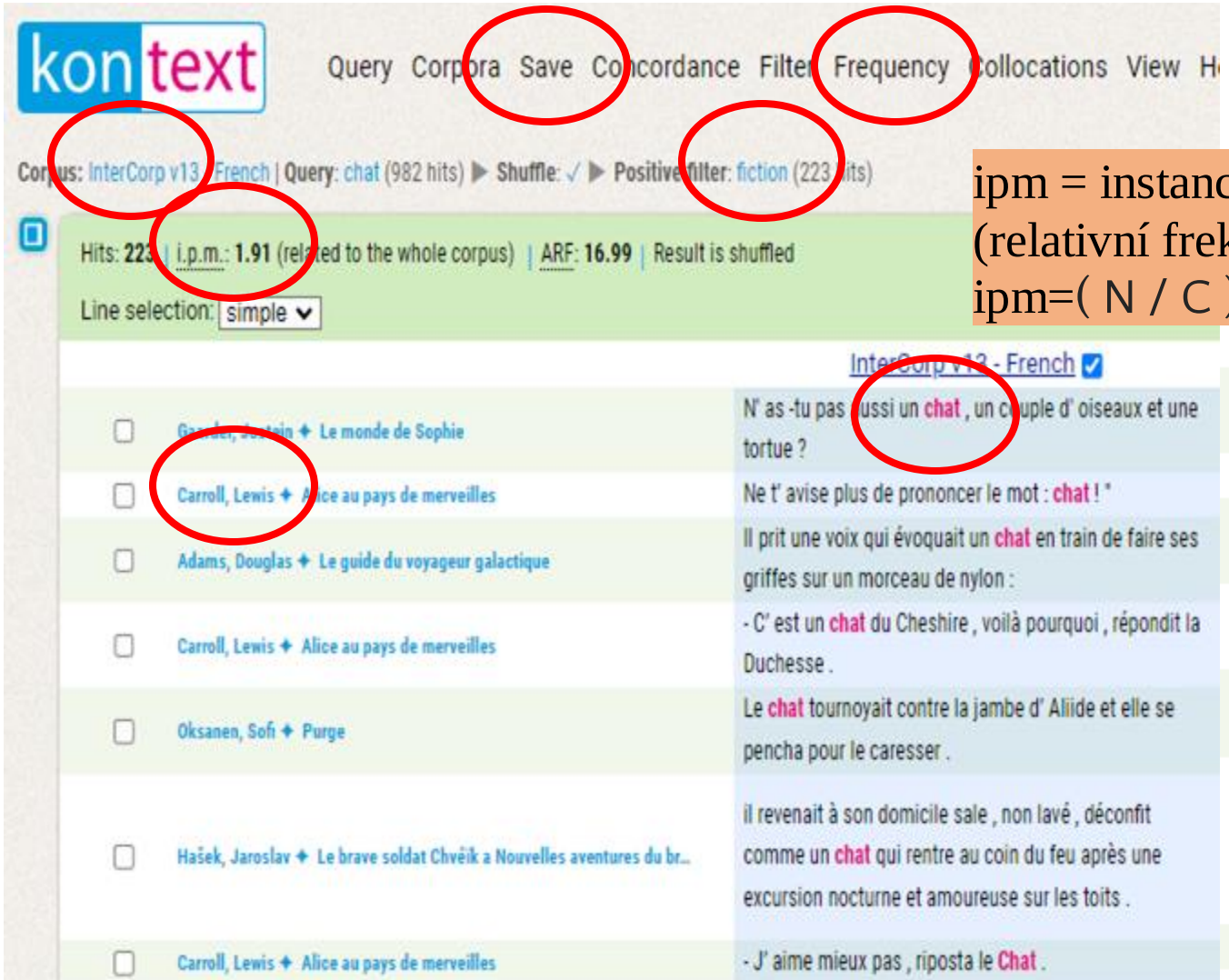
Osnova

1. Úvod
2. Paralelní korpus InterCorp
3. Anotace InterCorpu
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu
    3. Ukaž a hledej
6. Diskuse, otázky…

Co je to jazykový korpus a k čemu je dobrý

Jazykový korpus = soubor počítačově uložených textů (příp. přepisů mluvené řeči), který primárně slouží k jazykovému výzkumu (www.korpus.cz)

ipm = instances per million (relativní frekvence)
ipm=(N / C) * 10^6



Přehled korpusů dostupných přes rozhraní KonText Českého národního korpusu

- **Synchronní** korpusy psané češtiny (obecné: 5 mld slov; webové: 6 mld slov; akviziční, autorské, specializované)
- Korpusy **mluvené češtiny** (6 korpusů; celkem 13 mil. slov)
- **Diachronní** korpus češtiny (DiaKorp; 14.–19. století; 3,4 mil. slov)
- Jednojazyčné korpusy **jiných jazyků** (**de, en**, es, fi, **fr,** hu, **it,** nl, pl, pt, ru, sk, zh), dia/syn
- Mnohojazyčný **paralelní** korpus InterCorp (verze **16**: 62 jazyků; 5,3 mld slov)

Přístup do korpusu:

https://kontext.korpus.cz

Přihlášení:

a) **váš univerzitní login** (Shibboleth)
b) login **ud16test**, heslo **ud16test**

Hledat v korpusu:

syn2020 → InterCorp v16ud

Další informace a tutoriály:

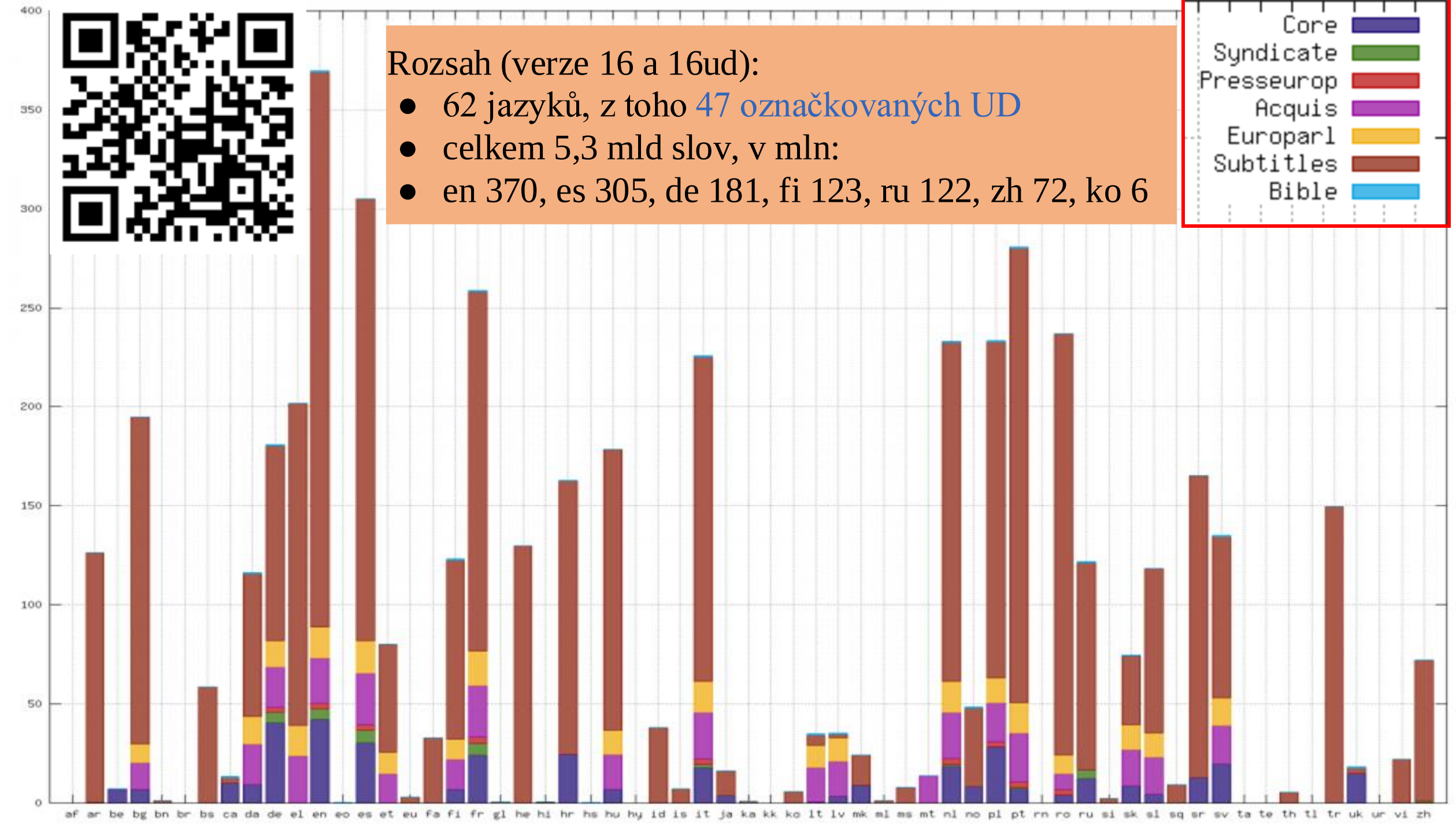
Tutoriál ke všem korpusům ČNK:
https://www.youtube.com/watch?v=EOuUdU-p8VQ&t=4112s
https://wiki.korpus.cz/doku.php/start

InterCorp: https://wiki.korpus.cz/doku.php/cnk:intercorp
Anotace podle UD:
https://wiki.korpus.cz/doku.php/pojmy:ud
Míry syntaktické komplexity a lexikální diverzity:
https://wiki.korpus.cz/doku.php/en:pojmy:syntakticka_komplexita

Universal Dependencies (oficiální popis a dokumentace):
https://universaldependencies.org

Rozsah (verze 16 a 16ud):

- 62 jazyků, z toho 47 označkovaných UD
- celkem 5,3 mld slov, v mln:
- en 370, es 305, de 181, fi 123, ru 122, zh 72, ko 6



V předchozích verzích:
nejednotná lingvistická anotace

Jazyky v korpusech InterCorp 16 a 16ud

S anotací v16ud

Bez anotace v 16ud

S beletrií

Víc než 15 knížek

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

Osnova

1. Úvod
2. Paralelní korpus InterCorp
3. Anotace InterCorpu
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu
    3. Ukaž a hledej
6. Diskuse, otázky…

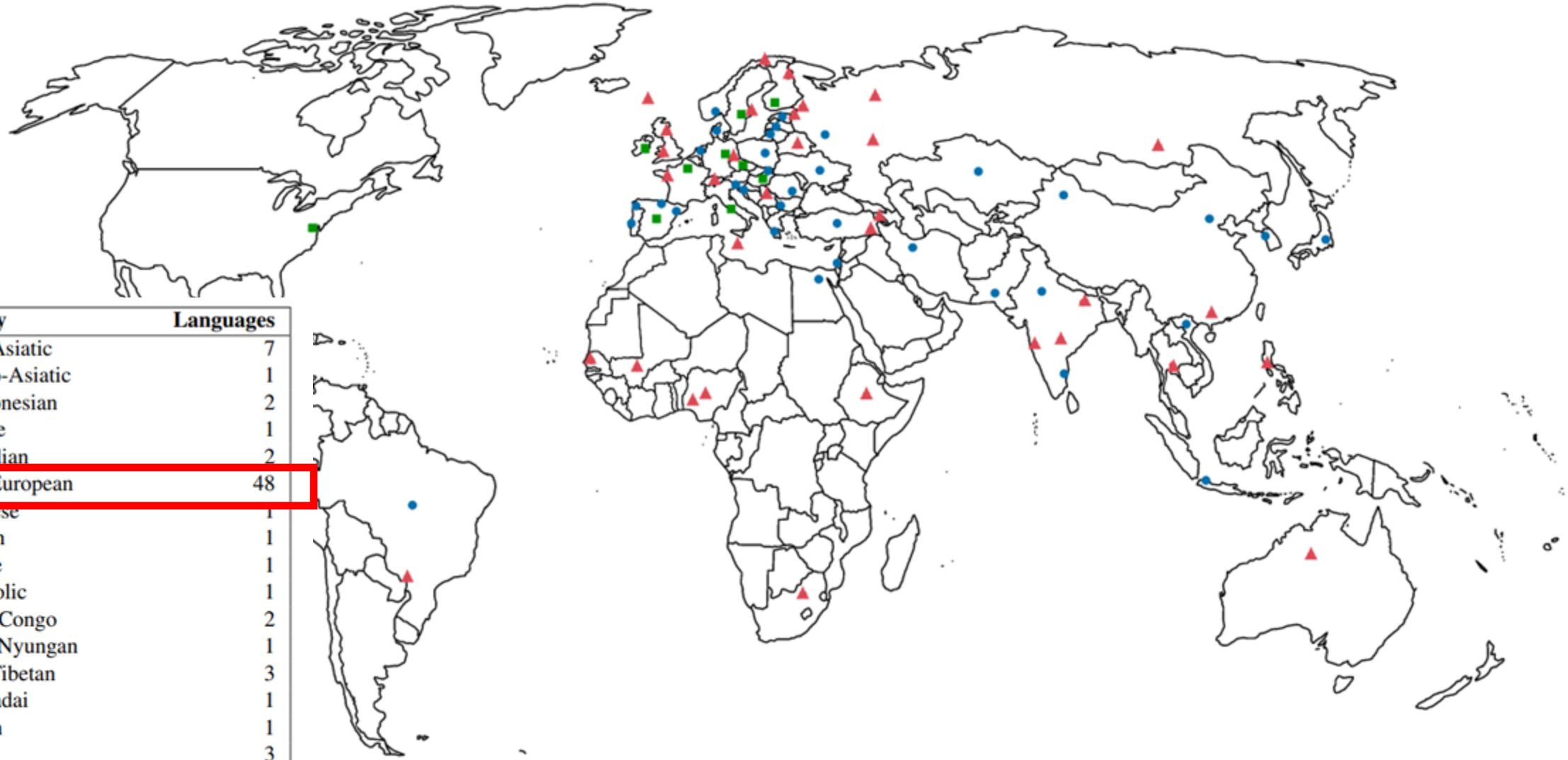
Proč zrovna *Universal Dependencies*?

- Faktický standard pro morfologickou a syntaktickou anotaci korpusů

- 161 jazyků, 283 *treebanků* – syntakticky anotovaných korpusů
  - verze 2.14, květen 2024
  - etalony pro trénování a testování (anotace by měla být správně)

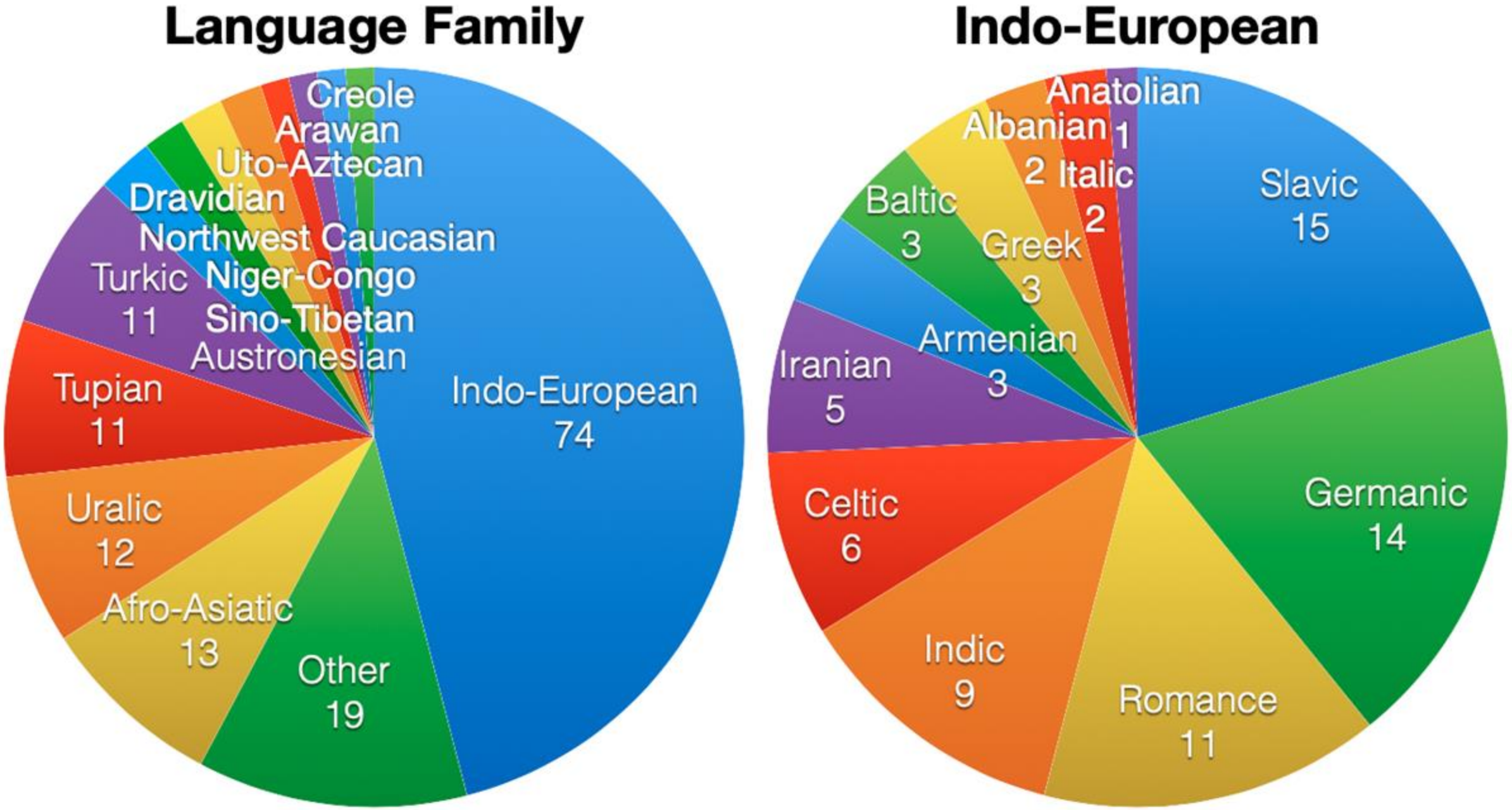
- *UDPipe* – nástroj pro 71 jazyků
  - v *InterCorp*u v16ud je 62 jazyků
  - *UDPipe* model 2.12 umí 47 z nich
  - anotace češtiny a “velkých jazyků” je relativně spolehlivá

- Aktivní komunita tvůrců a uživatelů



Family	Languages
Afro-Asiatic	7
Austro-Asiatic	1
Austronesian	2
Basque	1
Dravidian	2
Indo-European	48
Japanese	1
Korean	1
Mande	1
Mongolic	1
Niger-Congo	2
Pama-Nyungan	1
Sino-Tibetan	3
Tai-Kadai	1
Tupian	1
Turkic	3
Uralic	11
Code-Switching	1
Creole	1
Sign Language	1

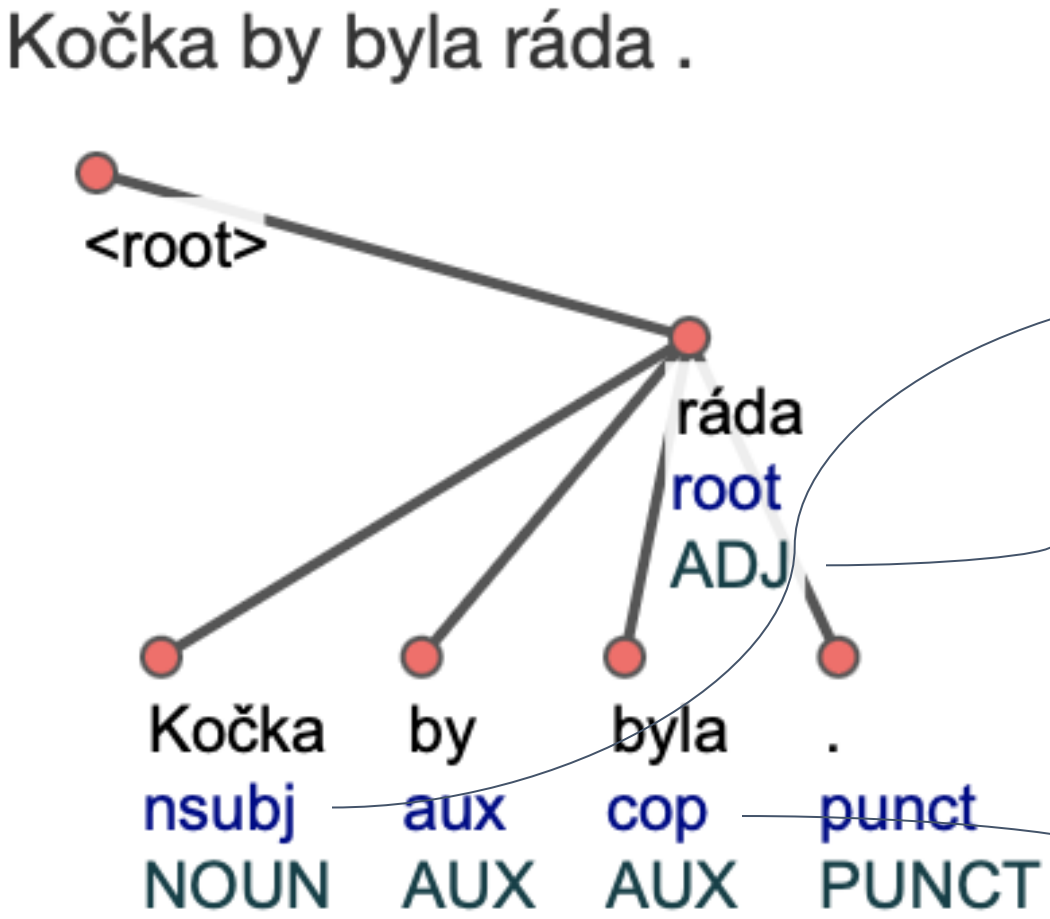
Table 4: Language families in UD v2.5.


Language Family
Creole
Arawan
Uto-Aztecan
Dravidian
Northwest Caucasian
Turkic 11
Niger-Congo
Sino-Tibetan
Austronesian
Tupian 11
Indo-European 74
Uralic 12
Afro-Asiatic 13
Other 19
Indo-European
Anatolian 1
Albanian 2
Italic 2
Baltic 3
Greek 3
Slavic 15
Armenian 3
Iranian 5
Germanic 14
Celtic 6
Indic 9
Romance 11

Zásady *Universal Dependencies*

- Jazykově nezávislé definice kategorií
- Kompromis mezi požadavky na anotaci (tzv. *Manningův zákon*):
  - **přijatelné** pro lingvistu
  - **stejné** jevy anotovat **stejně**
  - snadné pro **anotátory**
  - snadné pro **nelingvistu**
  - snadné pro **parser**
  - snadné **pokračování**: extrakce relací, porozumění textu, strojový překlad, ...
- https://en.wikipedia.org/wiki/Manning%27s_Law

Pomocná slova závisí na významových



- syntaktická funkce
  - deprel: dependency relation
- slovní druh
  - upos: universal part of speech
- funkce pro pomocná slova
  - ~~aux~~, cop, det, case, mark, clf

UD Guidelines verze 2 (verze 1: 2014)

- 17 slovních druhů – `upos`

  https://universaldependencies.org/u/pos/index.html
- 24 morfologických kategorií – `feats`

  https://universaldependencies.org/u/feat/index.html
- 37 syntaktických funkcí – `deprel`

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

17 upos (univerzálních slovních druhů) `[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	

24 morfologických kategorií

`[feats="Number=Sing"]`

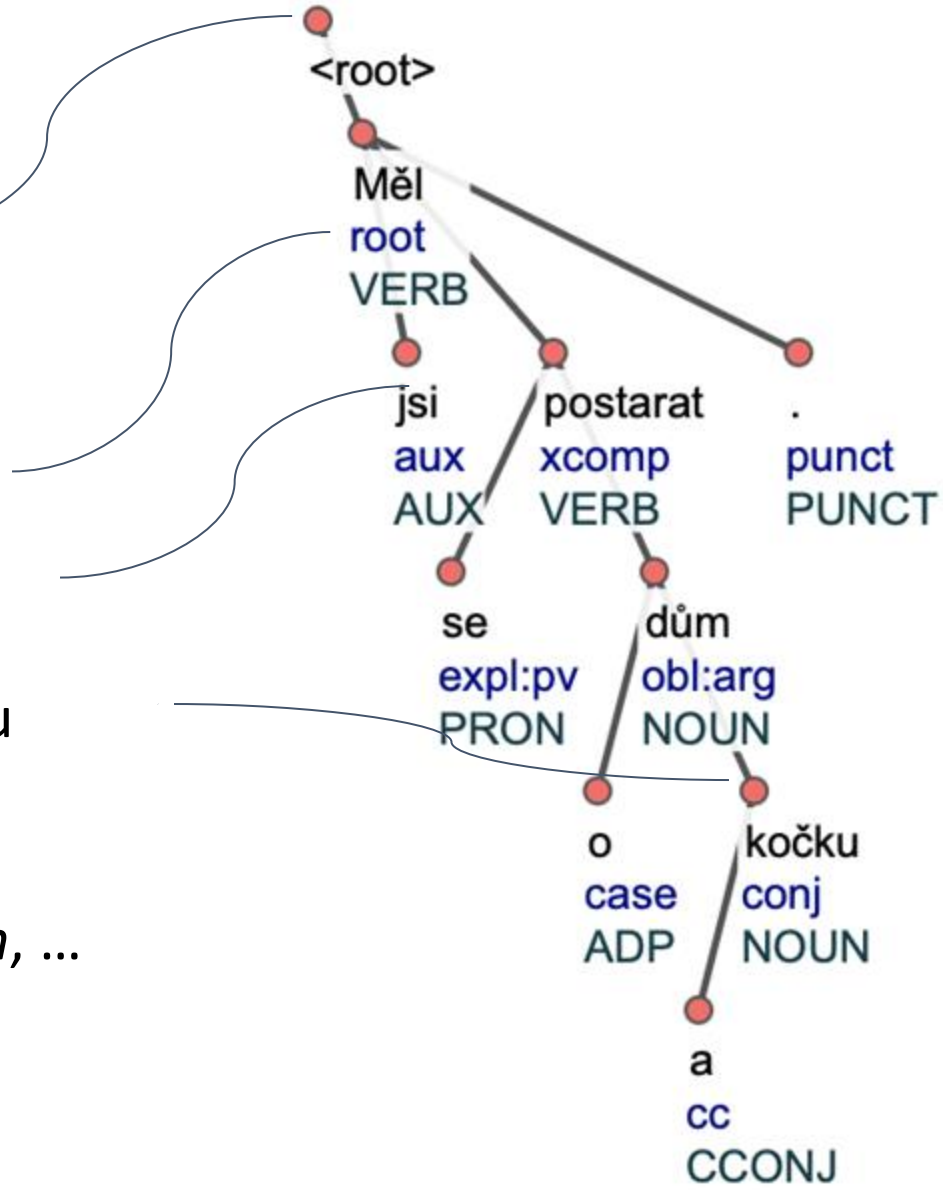
Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

Osnova

1. Úvod
2. Paralelní korpus InterCorp
3. Anotace InterCorpu
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu
    3. Ukaž a hledej
6. Diskuse, otázky…

Syntaktická anotace podle UD

- Každá věta jako jeden závislostní strom
- Jediná rovina (povrchová syntax)
- Každé slovo má svůj uzel a syntaktickou funkci
- Pomocná slova závisejí na významových slovech
- Druhý a další člen koordinace závisí na prvním členu
- Prázdné uzly neexistují (kromě uzlů technických)
- Víceslovné tokeny: *ses*, *sis*, *udělals*, *zač*, *proň*, *abych*, …



Měl ses postarat o dům a kočku.

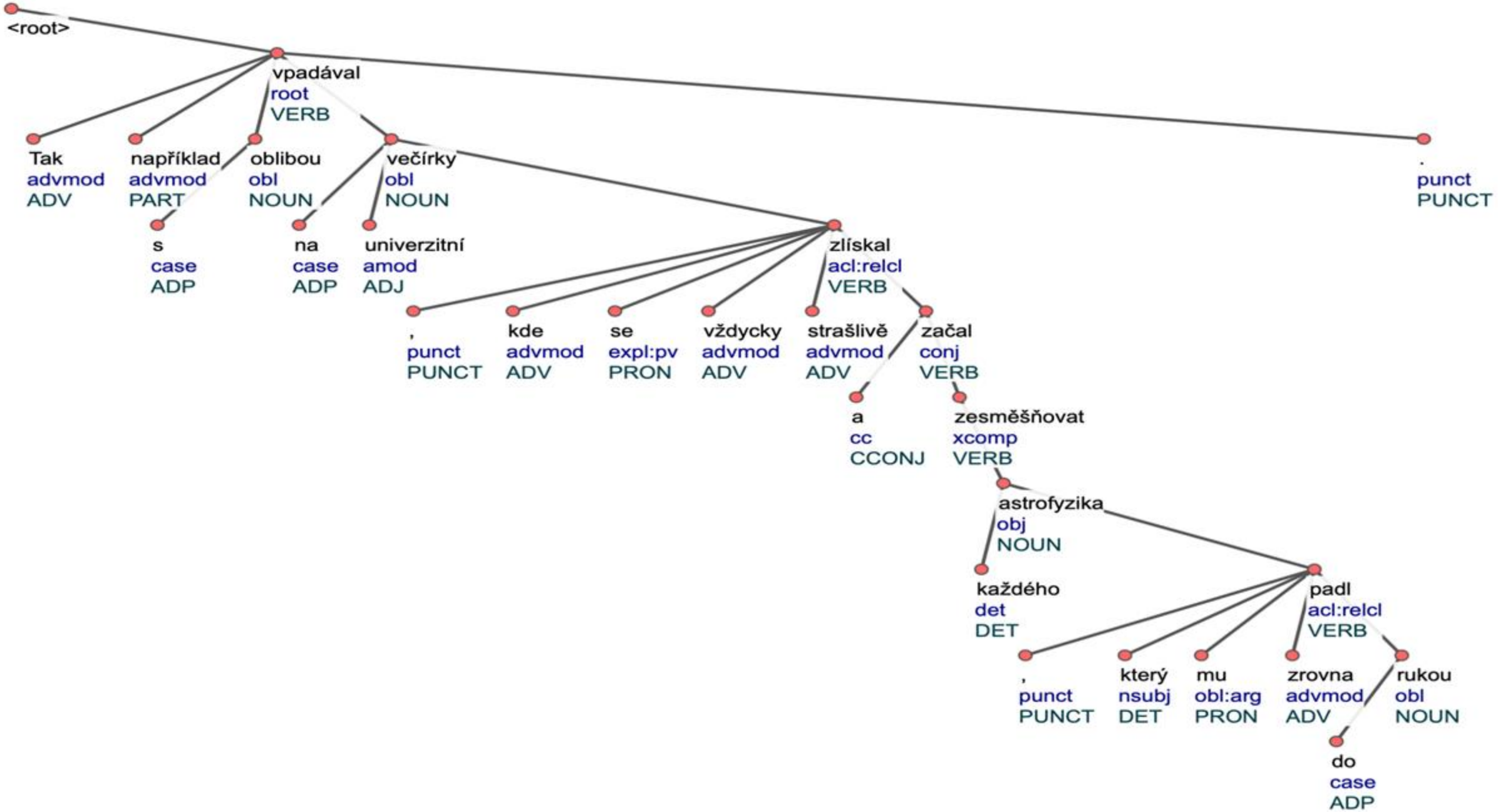
Syntaktické funkce (universal dependency relations) [deprel="acl"]

podle morfosyntaktických kategorií

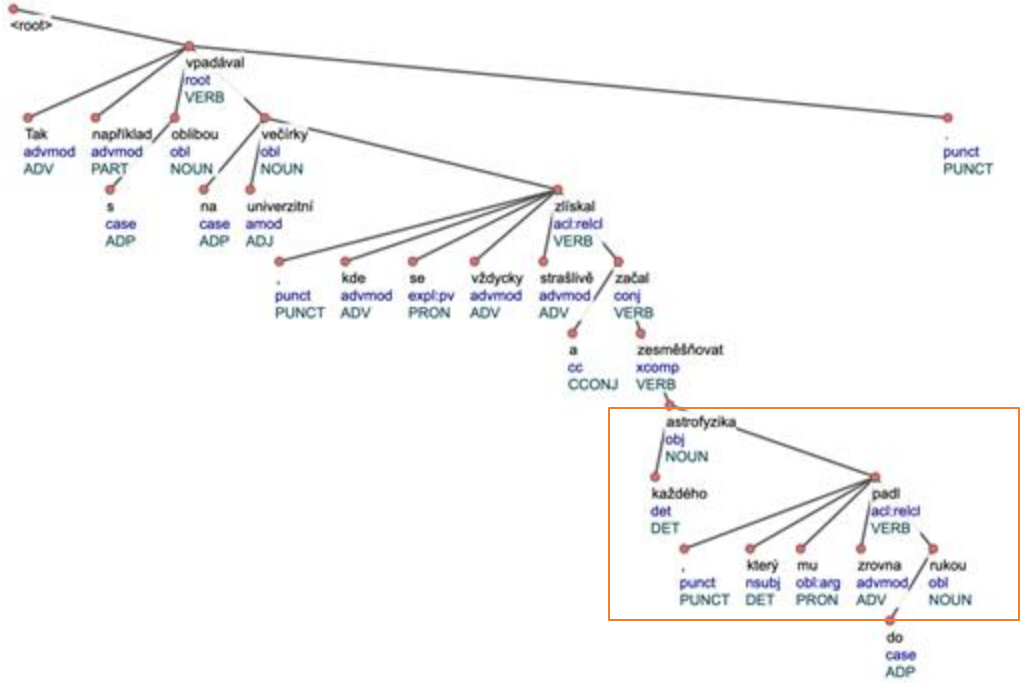
podle syntaktických funkcí

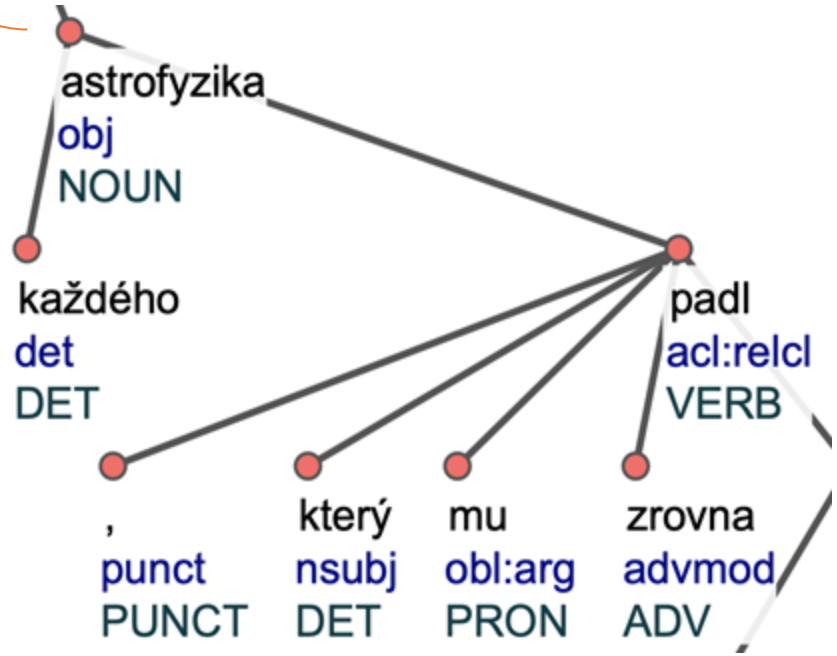
	Nominals	Clauses	Modifier words	Function words
Core arguments	**nsubj**	**csubj**		
	obj	**ccomp**		
	iobj	**xcomp**		
Non-core dependents	***obl***	***advcl***	***advmod***	***aux***
	vocative		*discourse*	***cop***
	expl			***mark***
	dislocated			
Nominal dependents	**nmod**	**acl**	**amod**	**det**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

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*

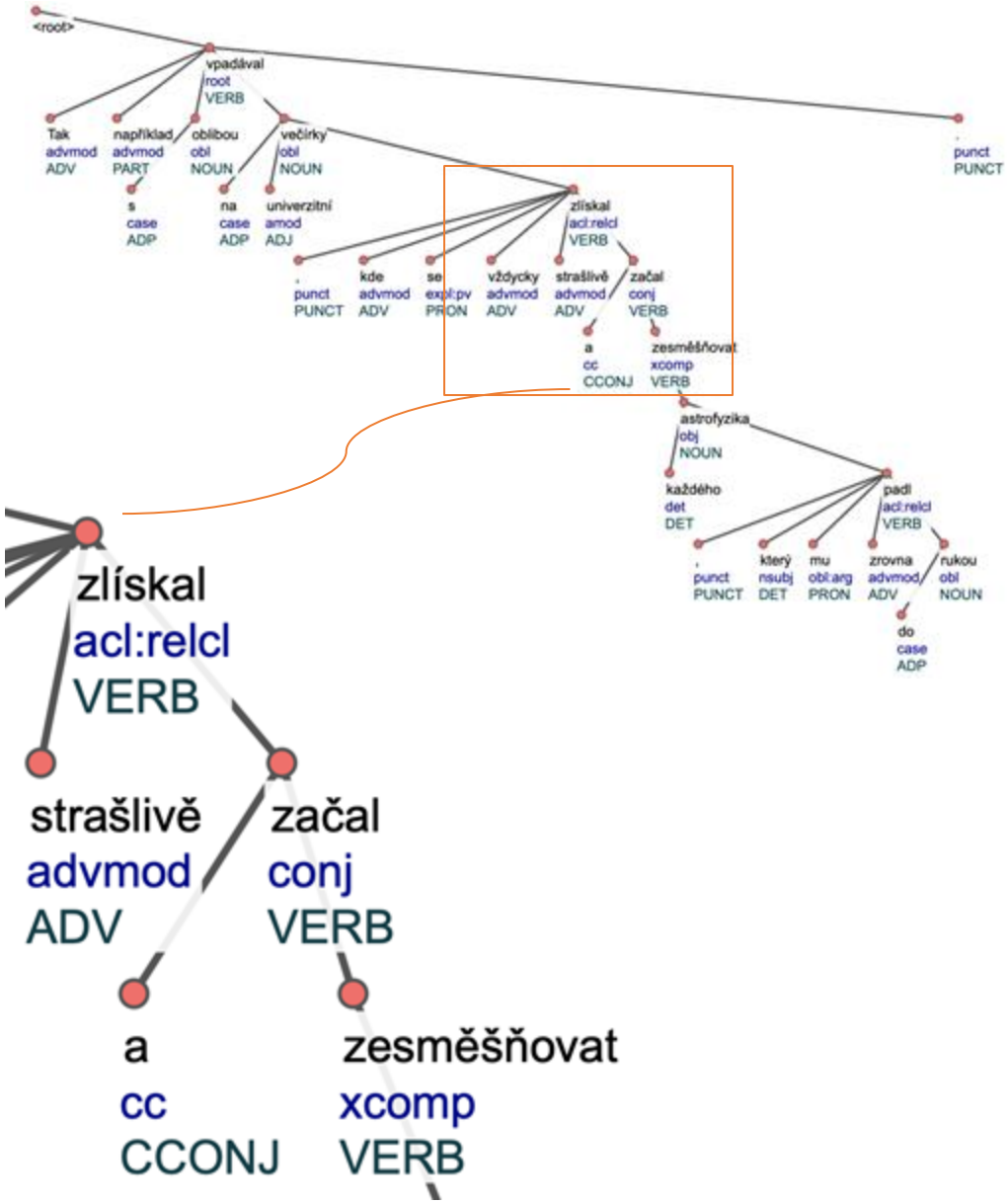


Tak například s oblibou vpadával na univerzitní večírky, kde se vždycky strašlivě zlískal a začal zesměšňovat každého astrofyzika, který mu zrovna padl do rukou.





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



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

Strategie úpravy CONLL-U pro Manatee

Kromě atributů z CONLL-U nové atributy pro snadnější …

- … orientaci v syntaktické struktuře (`p_lemma, e_id`):
  - `lemma`, `upos`, `feats`, `deprel` a relativní pozice rodiče
  - ID a relativní pozici efektivního rodiče
- … přístup k údajům u pomocných slov (`aux_feats, case_lemma`):
  - `lemma`, `upos`, `feats` a podtyp `deprel`
- … hledání a statistiky podle některých kategorií
  - některé atributy z `feats`

→ Minimalizace počtu nových atributů

- jen ty, které pro daný jazyk mají smysl
- celkem 20 až 44

Atributy z CONLL-U

atribut	popis
word	slovní tvar: *abys*
sword	slovní tvar rozdělený na interpretovaná syntaktická slova: *aby*|*bys*
iword	rozdělený slovní tvar bez interpretace: *aby*|*s*
lc	slovní tvar malými písmeny: *abys*
lemma	lemma (základní tvar): *aby*|*být*
lc_lemma	lemma malými písmeny: *aby*|*být*
upos	slovní druh podle UD: SCONJ|AUX
xpos	jazykově specifický slovní druh: J,----------|Vc-S---2-------
feats	kategorie podle UD: ||Mood=Cnd|Number=Sing|Person=2|VerbForm=Fin
id	pořadí slova ve větě
head	odkaz na řídící uzel
deprel	syntaktická funkce

Přidané atributy simulující syntaktickou strukturu

atribut	popis
parent	relativní pozice řídícího uzlu, např. -1, +2
p_lemma	lemma řídícího uzlu
p_upos	slovní druh řídícího uzlu
p_feats	morfologické kategorie řídícího uzlu
p_deprel	syntaktická funkce řídícího uzlu
e_id	pořadové číslo skutečného řídícího uzlu (tj. u 2. a dalšího členu koordinace pořadové číslo 1. členu)
e_deprel	syntaktická funkce skutečného řídícího uzlu
eparent	relativní pozice skutečného řídícího uzlu

Atributy pro závislá pomocná slova

- Uvádějí se u příslušných významových
- Víc pomocných slov? Atribut má multihodnotu. Ve `feats` je oddělovač "||".
- Názvy atributů: **typ pom. slova** _ **atribut pom. slova**, např. **aux**_**feats**
- **Typ pom. slova**:
  - **aux**: pomocné sloveso
  - **case**: předložka, postpozice
  - **clf**: klasifikátory (čínština, japonština)
  - **cop**: spona
  - **det**: determinátor (člen, ukazovací zájmeno, číslovky)
  - **mark**: podřadicí spojka
- **Atribut pom. slova**:
  - **lemma**: lemma pom. slova
  - **upos**: slovní druh pom. slova
  - **feats**: morfologické kategorie pom. slova
  - **type**: podtyp syntaktické funkce, pokud je uveden
    několik deprel=det:numgov, upos=DET
    koček det_type=numgov

Osnova

1. Úvod
2. Paralelní korpus InterCorp
3. Anotace InterCorpu
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu
    3. Ukaž a hledej
6. Diskuse, otázky…

Než začneme… V čem je to nové?

Klasické verze InterCorpu

[word="kočka"]

[lemma="kočka"]

UD verze

[word="kočka"]

[lemma="kočka"]

[tag="N.*, NOUN atd."]

[upos="NOUN"]

[xpos="N.*"]

[feats="Case=Nom|
Gender=Fem|
Number=Sing"]

[deprel="nsubj"]

UD verze – parent

[p_word=""]

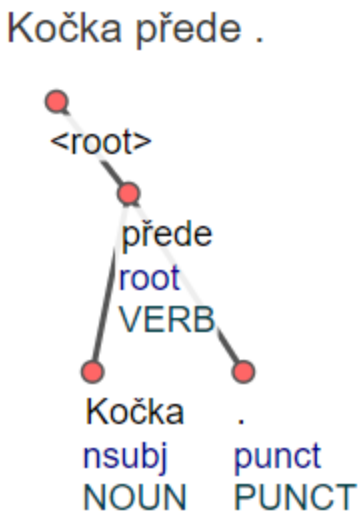
[p_lemma=""]

[p_upos=""]

[p_xpos=""]

[p_feats=""]

[p_deprel=""]

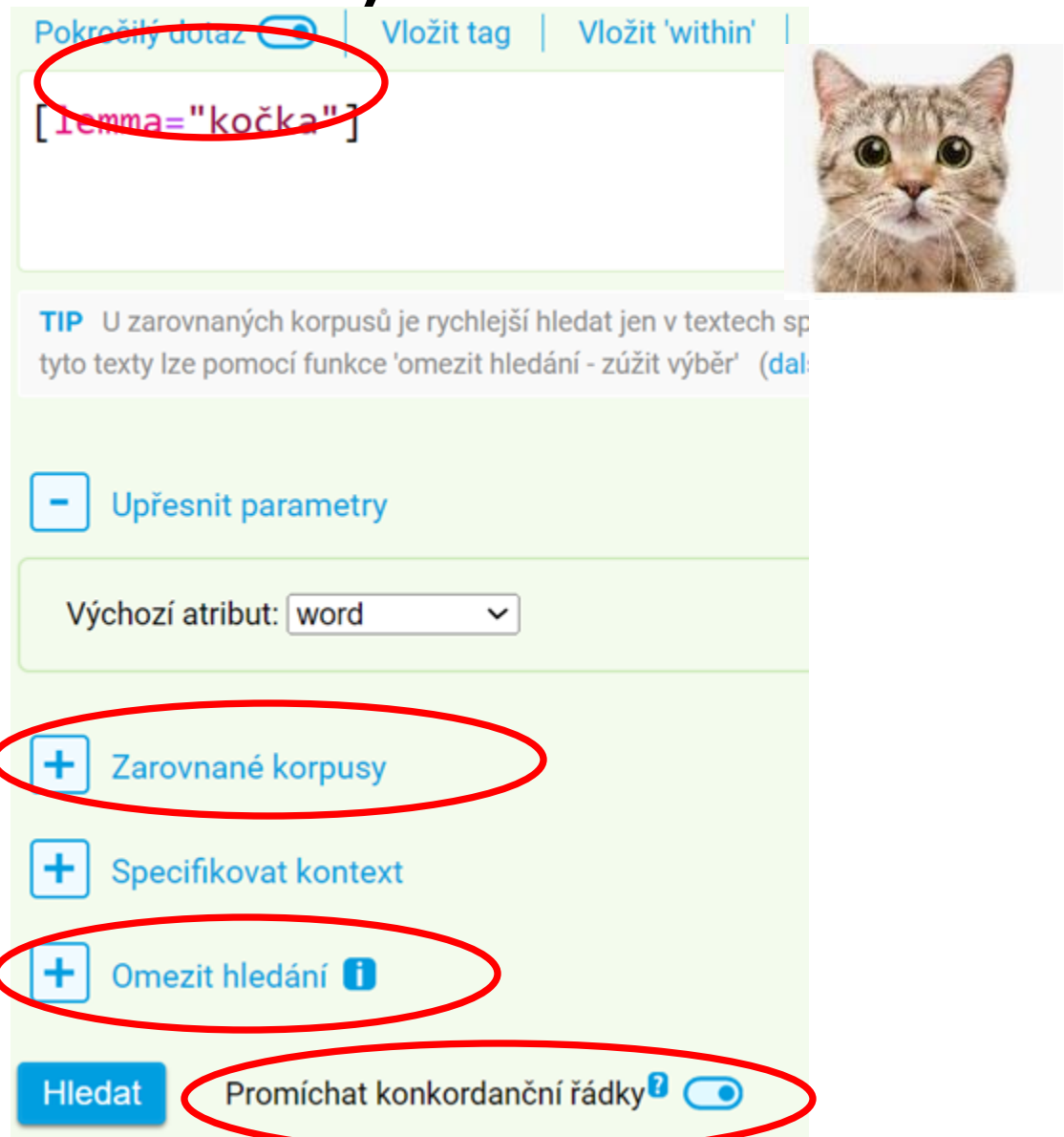


a) v dotazu [lemma="kočka"
& p_upos="V.*"]

b) ve frekvenčních seznamech

PŘÍSTUP KE KORPUSU: http://kontext.korpus.cz (univerzitní login NEBO ud16test, heslo ud16test)

EX = příklad/cvičení
TIP = tip pro vyhledávání
POZOR = pozor past



4.1 Základní vyhledávání v IC 16ud

EX 1 - lemma *kočka* (v různých jazycích) - jaké má nejčastější syntaktické funkce?

Řešení: **`[lemma="kočka"]  Frekvence > Vlastní > p_deprel`**

TIP: lze zobrazit syntaktickou strukturu (ikona vlevo od výskytu)

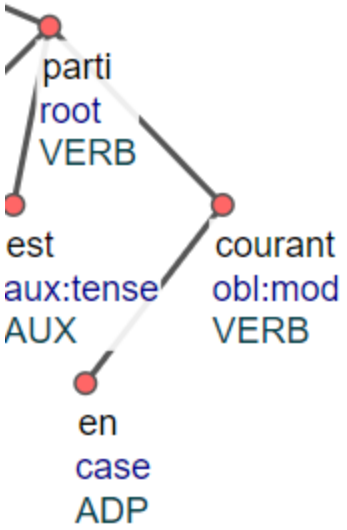
EX 2 - Opačně: Jaké větné členy nejčastěji rozvíjejí lemma *kočka*?

Řešení: **`[p_lemma="kočka"] - Frekvence > Vlastní > deprel`**

TIP: Seznam `deprel` https://universaldependencies.org/u/dep/index.html (projít si je, rozkliknout, podívat se na příklady)

Frekvenční seznamy podle vlastností řídícího členu: Co můžete dělat *corriendo / en courant (v běhu / běhaje / běhajíc.e)?*

PATH/DRÁHA + MANNER/ZPŮSOB



Verbálně rámcující jazyky

fr. ***il est arrivé en courant***

es. ***llegó corriendo***

Satelitně rámcující jazyky

cs. ***přiběhl***

DOTAZ: `en courant / corriendo > Frekvence > p_lemma` *(fr: slova KWIC nejvíce vpravo)*

- *ekvivalenty v češtině?*

(Talmy 2000, Martinková 2014 a v přípravě aj.)

TIP: Vizualizace syntaktického stromu v nástroji UDPipe (https://lindat.mff.cuni.cz/services/udpipe/) > vybrat treebank > zadat větu > process > show trees

4437 výskytů

p_lemma	Freq	i.p.m.
salir	1 011	7,27
seguir	240	1,73
venir	151	1,09
entrar	120	0,86
bajar	117	0,84
ver	109	0,78
llegar	107	0,77
pasar	97	0,7
volver	94	0,68
subir	82	0,59
tener	71	0,51
acercar	52	0,37
cruzar	46	0,33
ir	41	0,3
andar	38	0,27

perifrastická konstrukce

Vi al muchacho **corriendo**

A jak jsou tato spojení přeložena do češtiny/angličtiny?

zdrhaly, dal se na útěk, odběhl do, utíkal, spasit se úprkem, dal se do klusu, vyběhla ven, uprchli, zmizel jako stín, běžel/pádil pryč, vzali nohy na ramena…

Johansson (2007): Seeing through multilingual corpora

918 výskytů

p_lemma	Freq	i.p.m.
partir	107	0,91
sortir	90	0,77
arriver	63	0,54
traverser	50	0,43
descendre	40	0,34
revenir	32	0,27
venir	25	0,21
monter	24	0,21
passer	23	0,2
enfuir	21	0,18
aller	17	0,15
rentrer	15	0,13
alimenter	14	0,12
repartir	14	0,12
sauver	13	0,11
retourner	10	0,09

Bible: ca cs da de en fi fr hr it lt lv nl no pl pt ru sk sv uk

Konkrétní slovo v konkrétní syntaktické funkci

TIP: operátor & (AND) umožňuje v dotazu kombinovat atributy:

EX 3: lemma kočka v pozici subjektu (`nsubj`)

Řešení: **`[lemma="kočka" & deprel="nsubj.*"]`**

EX 4: Jaké jsou nejčastější predikáty subjektu "kočka"?

Řešení: **`[lemma="kočka" & deprel="nsubj.*"] Frekvence > Vlastní > p_lemma`**

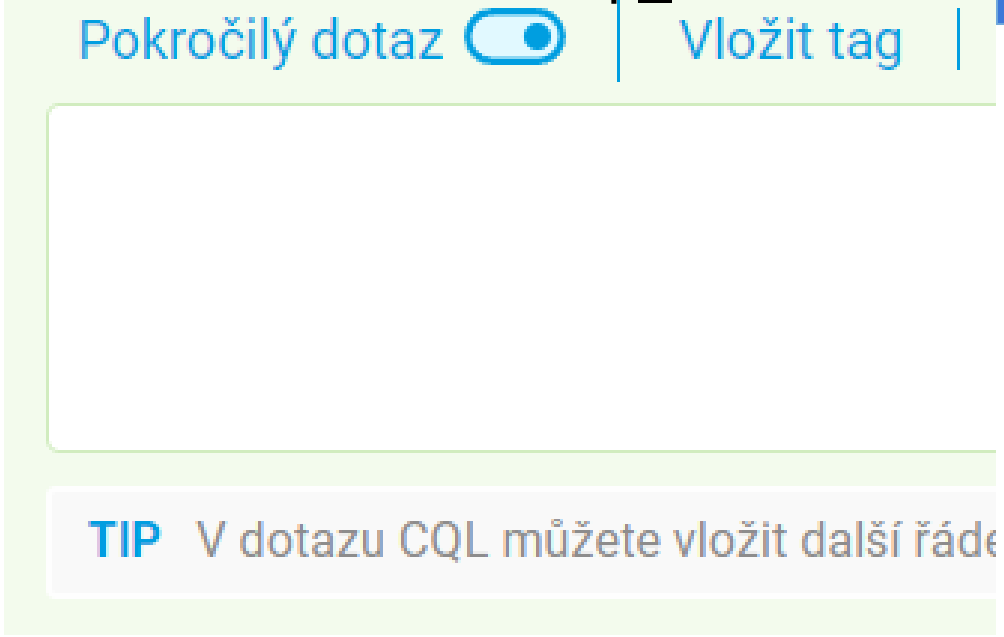
Doplňkově: Jaké jsou nejčastější nominální subjekty (`nsubj`) a přímé předměty (`obj`) v Bibli?
Řešení: Omezit hledání > text.group: Bible > [Zúžit výběr]
[deprel="nsubj.*"] nebo [deprel="nsubj.*" & upos="NOUN|PROPN"], Frekvence > Lemmata
Výsledek (cs): *Hospodin, Bůh, král, muž, Ježíš, lid, syn, David, otec, kněz, slovo, …* 17. *žena*
Nebo přímý předmět: [deprel="obj.*" & upos="NOUN|PROPN"] nebo přímý i nepřímý předmět: [deprel="i?obj.*" & upos="NOUN|PROPN"]

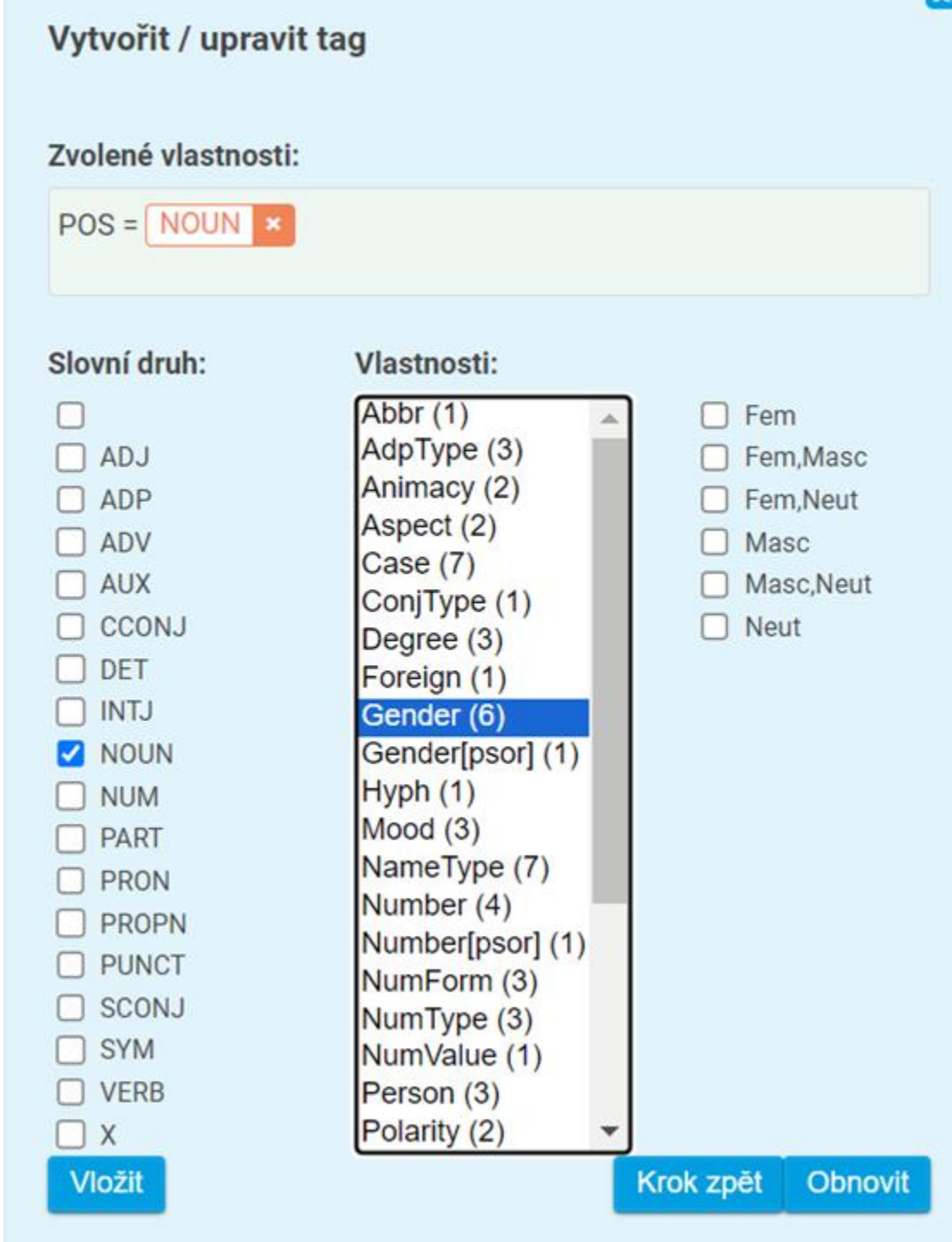
TIP: Kombinace s `feats`

Např. rod/Gender):

```
[deprel="nsubj.*" & upos="NOUN" &
feats="Gender=Fem"]
```

Frekvence > Vlastní > p_lemma





Bude také deprel!

Procvičení (doma :-)

1. Kdo nejčastěji zpívá? (nejčastější lemmata subjektu slovesa *zpívat*)
2. Co dělají ptáci nejčastěji? (nejčastější lemmata přísudku lemmatu *pták* ve funkci subjektu)

Řešení: https://wiki.korpus.cz/doku.php/pojmy:ud

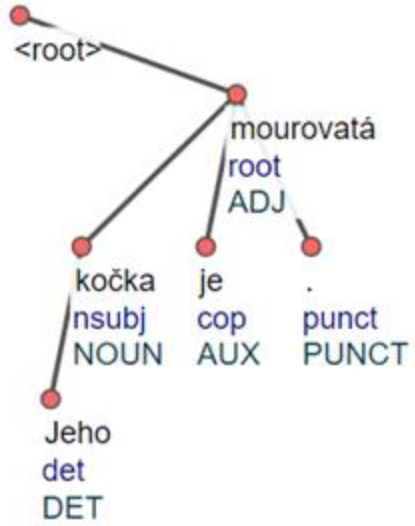
Konkrétní slovo v konkrétní syntaktické funkci: PAST 1

```
[lemma="kočka" & deprel="nsubj"]
```

Frekvence > Vlastní > p_lemma

Trvalý odkaz: https://www.korpus.cz/kontext/view?q=~gGmg44SIO0Is

POZOR 1: Spona visí jako pomocné slovo na lexikálním slově

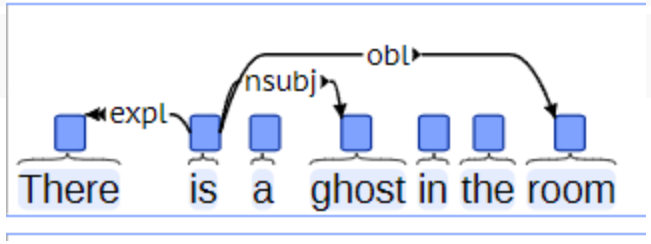


Konkrétní slovo v konkrétní syntaktické funkci: PAST 2

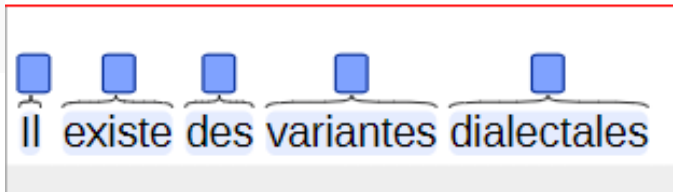
```
[lemma="chat" & deprel="nsubj"]
Frekvence > Vlastní > p_lemma
```

POZOR 2: Dotaz **nenajde** *Kočka* `[nsubj:pass]` *byla nakrmena.*

[deprel="**expl**"]



`expl:subj`



TIP: Při zadávání `deprel` uvádět operátory `.*`

např. `[nsubj.*]`, aby se pokryly i podtypy

+ vždy se podívat na definici `deprel` a možné podtypy:

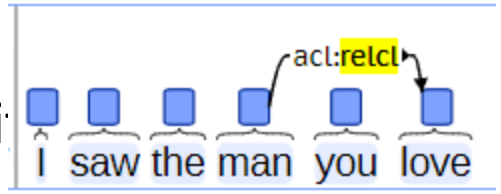
(https://universaldependencies.org/u/dep/index.html)

Podtyp `acl` : vztažné věty (`acl:relcl`)

deprel `acl` = adnominal clause

`[deprel="acl:relcl"] = vztažné věty`

pro jazyky, v nichž existuje vztažná věta (cs, fr, en, es, de, el, fi, i



naopak v japonštině nikoli:

Tomio	**ga**	**tabeta**	*o-bentō*	wa	watași	no	mono	dešita
Tomio	SUBJ	snědl	svačina	TOP	já	GEN	věc	byla

Svačina, **kterou Tomio snědl**, byla moje.

https://universaldependencies.org/u/dep/acl-relcl.html

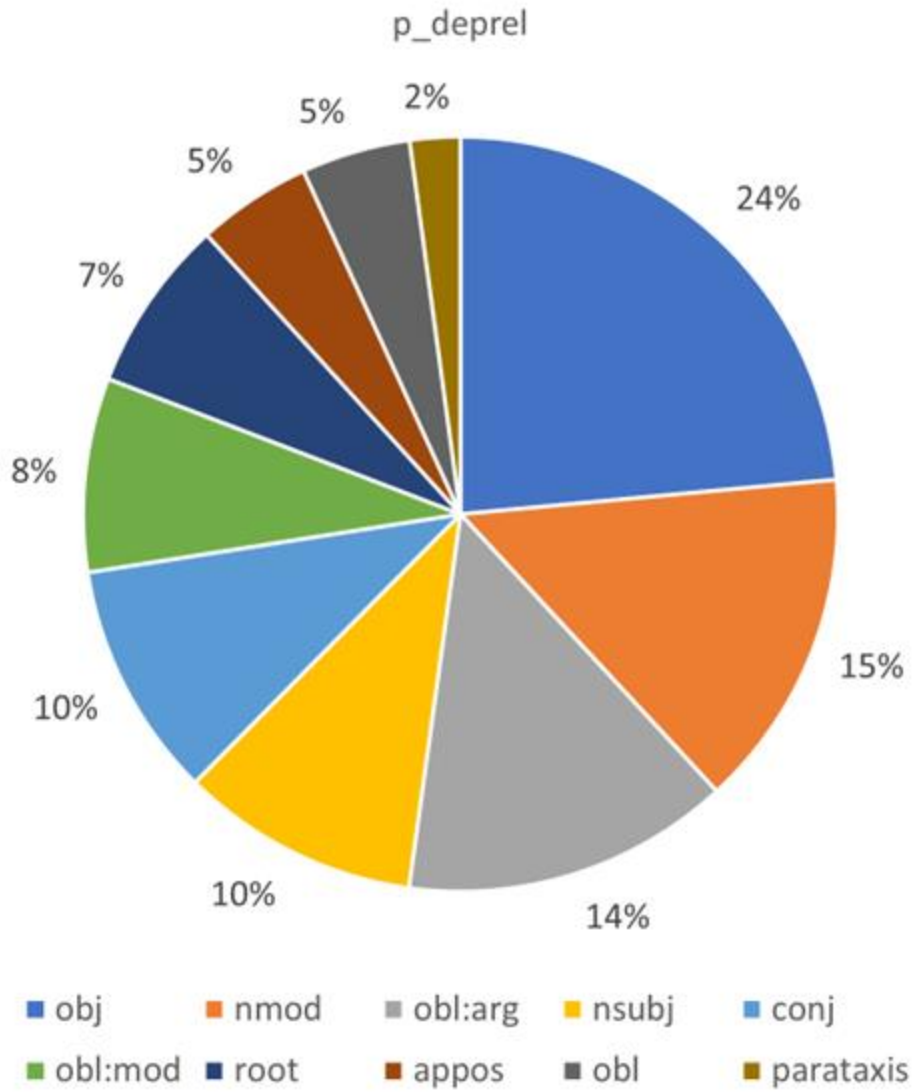
Úplně nové možnosti výzkumu – i kontrastivního, translatologického aj.

EX 5 – vlastnosti řídícího členu vztažné věty (např. deprel)

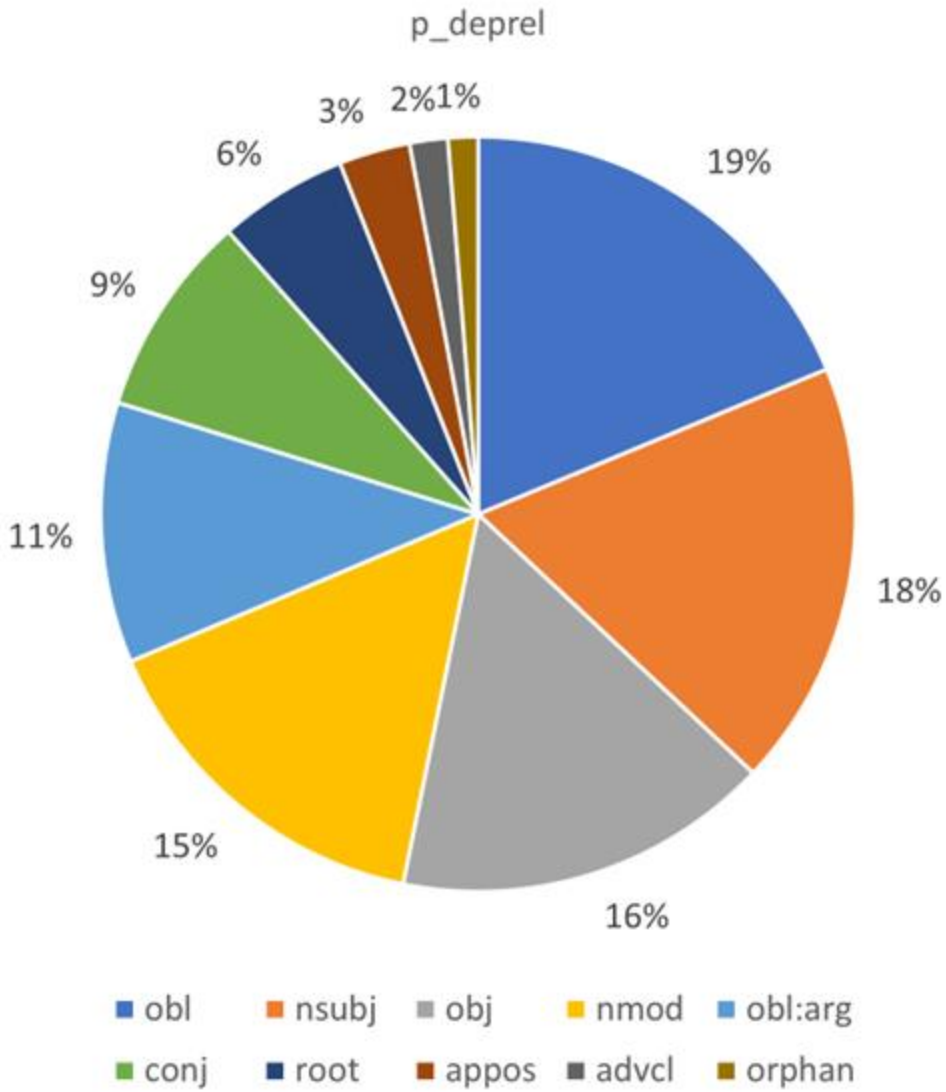
Frekvence > Vlastní > p_deprel

Karolína Pavlíková, IC v13ud, fiction

1) French text.original ∈ {Yes}



2) Czech text.original ∈ {Yes}

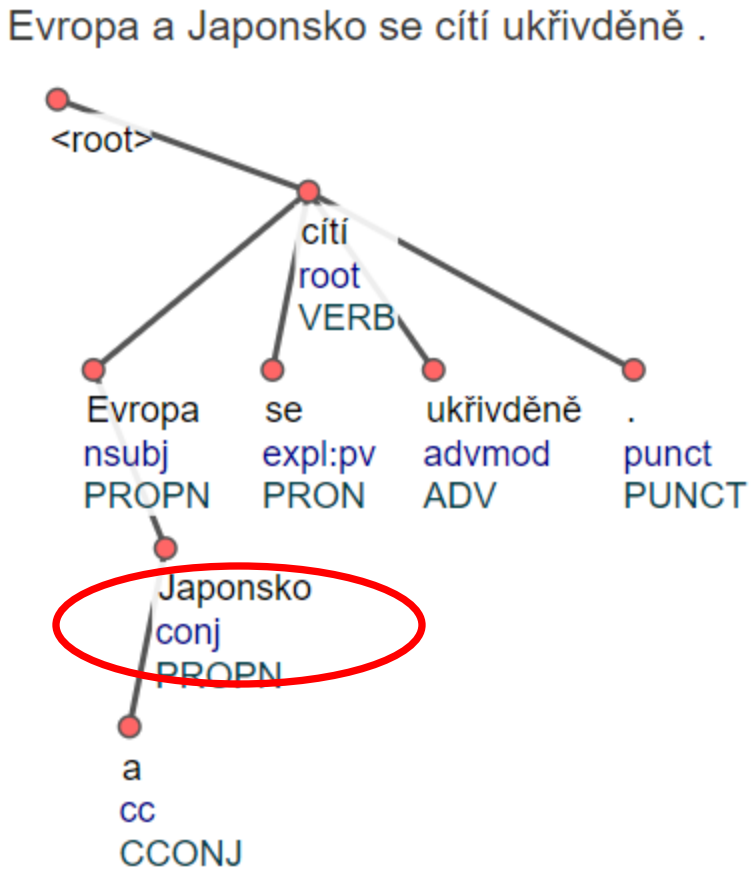


PAST 3: Koordinace – `[deprel="conj"]`

KOORDINACE conj = conjunct

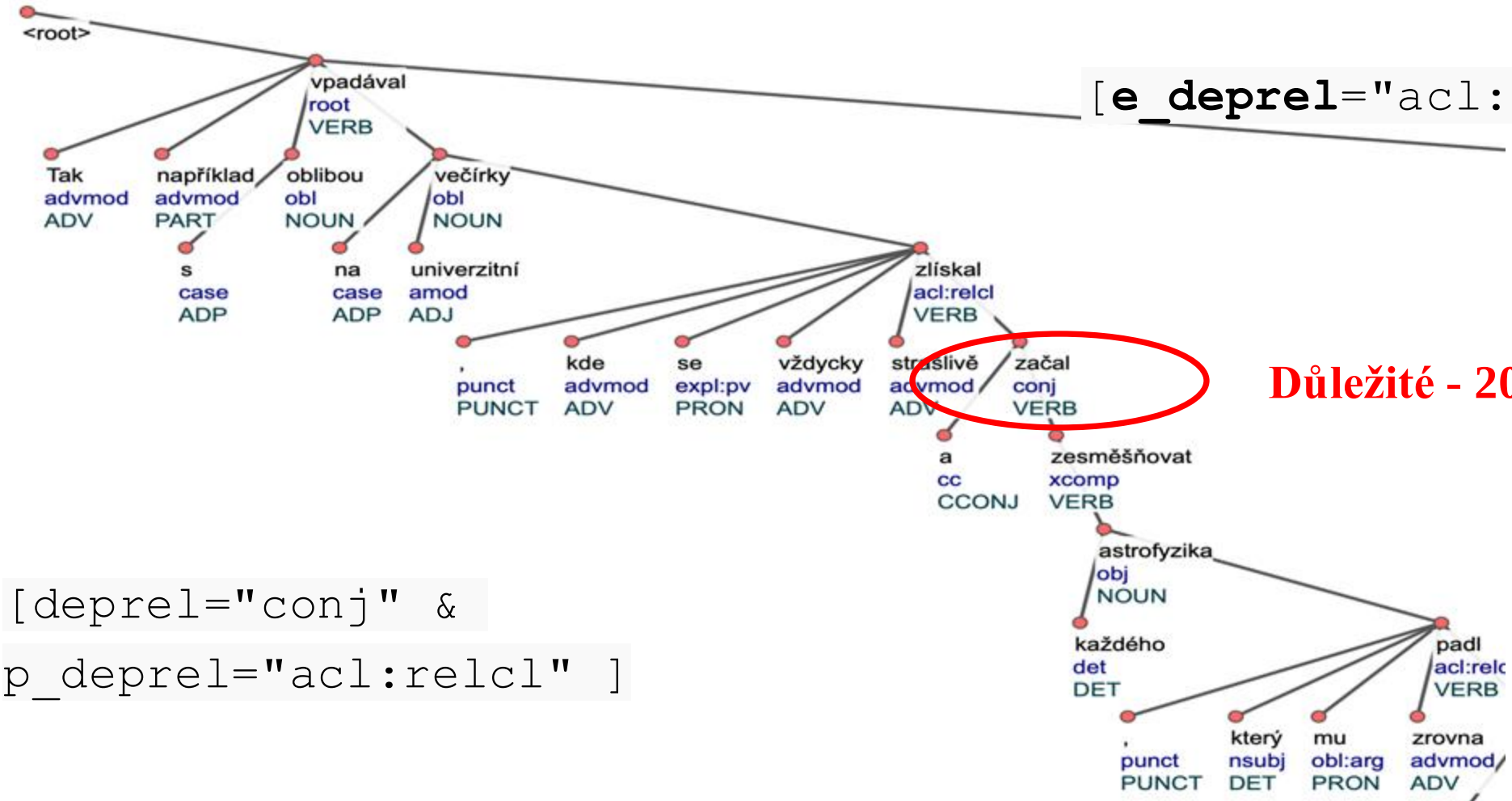
= druhý a každý další člen koordinačního řetězce

```
[deprel="nsubj.*" ] PROBLÉM?
Řešení: [e_deprel="nsubj.*" ]
nebo [deprel="nsubj.*" |
deprel="conj" & p_deprel="nsubj.*"]
```



Koordinace vztažných vět – deprel `conj`

the first conjunction is by convention treated as the parent (or "technical head") of all subsequent coordinated clauses via the `conj`



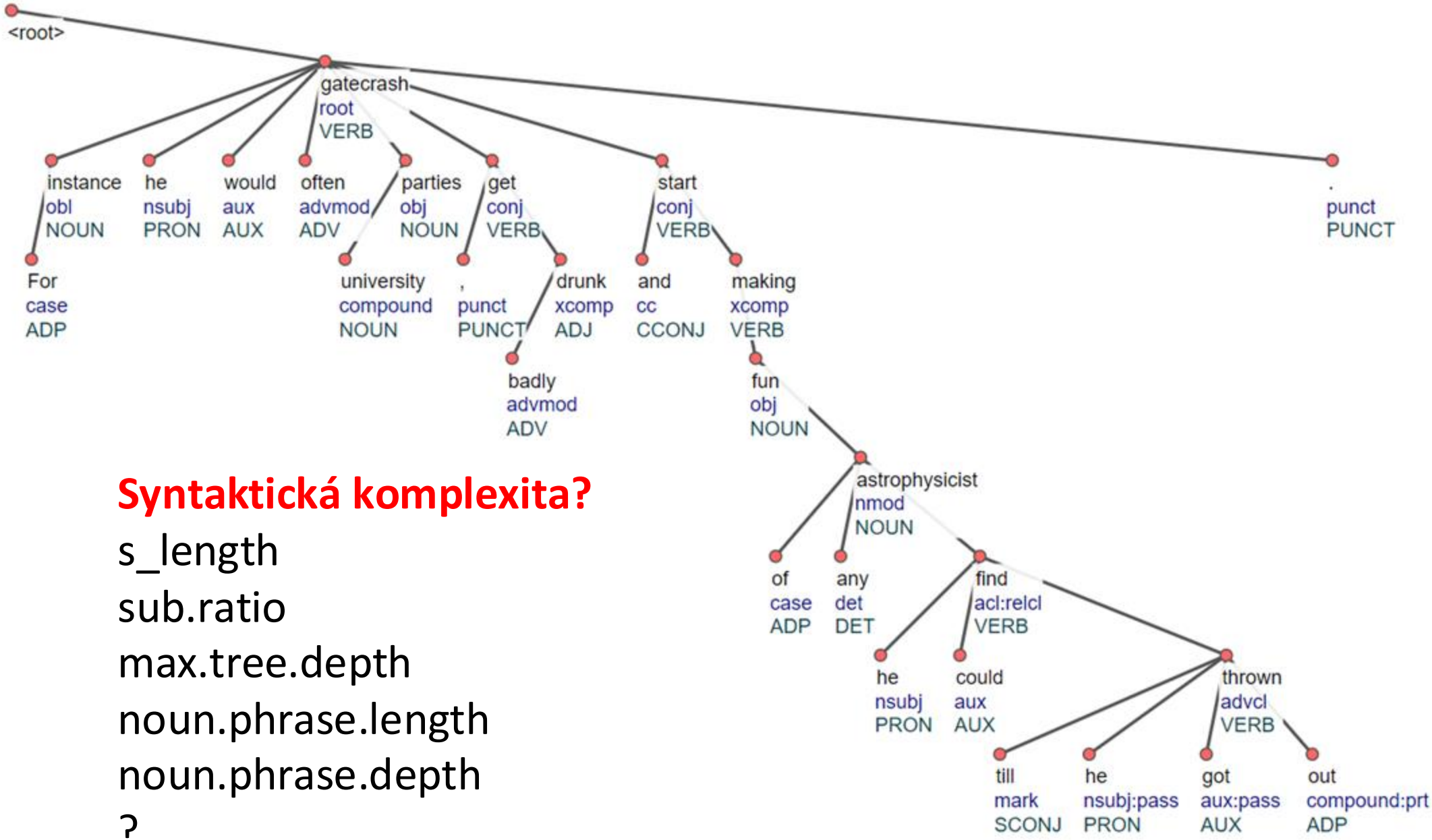

```
[e_deprel="acl:relcl"]
```

Důležité - 20%!

```
[deprel="conj" &
p_deprel="acl:relcl" ]
```

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 .



Syntaktická komplexita?

s_length

sub.ratio

max.tree.depth

noun.phrase.length

noun.phrase.depth

?

Osnova

1. Úvod
2. Paralelní korpus InterCorp
3. Anotace InterCorpu
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu
    3. Ukaž a hledej
6. Diskuse, otázky…

5.1 Co to je a **proč** to měřit?

FR

Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts.
(A. Camus, *La Peste*)

[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs, and floundered on to its side, its body writhing in long convulsions.
(transl. S. Gilbert)

en

V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami, svalil se na zem a dodělal v škubavých křečích.
(transl. M. Tomášková)

Míry syntaktické komplexity

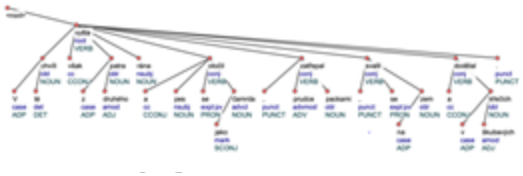
$$\frac{\text{No T-units + No Sub}}{\text{No T-units}}$$

Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts.
(A. Camus, *La Peste*)

Sub.ratio = 2.5 ((2+3)/2)
Max.Tree.Depth = 3

[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs, and floundered on to its side, its body writhing in long convulsions.
(transl. S. Gilbert)

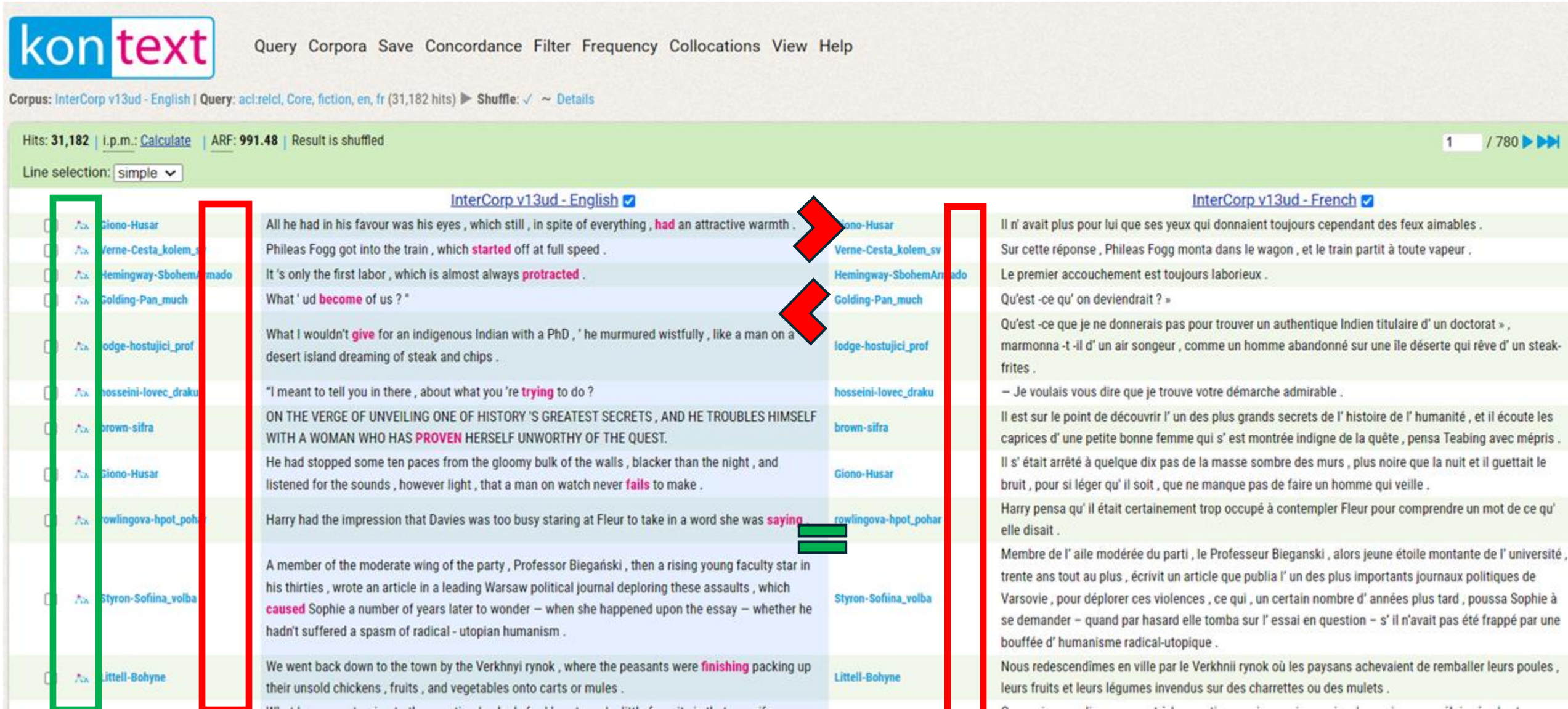
(SPLIT) Sub.ratio = 1.33 (3+1)/3)
Max.Tree.Depth = 1



V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami, svalil se na zem a dodělal v škubavých křečích.
(transl. M. Tomášková)

Sub.ratio = 1 (5/5)
Max.Tree.Depth = 0

PLÁN:
implementace měr syntaktické komplexity
na základě anotace UD do paralelního korpusu InterCorp



Pokus o definici syntaktické komplexity

- Obecná definice komplexity systémů:

“the number and variety of elements and the elaborateness of their interrelational structure“ (Rescher 1998:1, Hübler 2007:10; cited by Álvarez González et al. 2019)

- **Beaman** (1984: 45; *Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse*):

“**syntactic complexity in language is related to the number, type, and depth of embedding in a text**. Syntactically simple authors use short, single clause sentences and rely more heavily on coordinated structures [...]. Syntactically complex authors [...] use longer sentences and more subordinate clauses that reveal more complex structural relationships.” (viz podobně také De Clerq 2016)

Syntaktická komplexita věty tak může být definována počtem a variabilitou klauzí, z nichž se skládá, a mírou hierarchičnosti vztahů mezi nimi.

Simplifikace komplexnosti

- syntaktická komplexita musí brát v úvahu **nejen počet a variabilitu entit, ale také míru hierarchičnosti jejich uspořádání** (viz Beaman 1984: 45)
- (syntaktická) komplexita je **multidimenzionální jev** (Biber, Larsson & Hancock 2023). Pouze jedna míra komplexity je tak vždy nutně určitým zkreslením, proto je třeba kombinovat více měr. Navíc každá míra se bude chovat odlišně podle žánru (registru) a podle jazyka, příp. podle dostupnosti a spolehlivosti značkování v InterCorpu.
- cílem je analyzovat syntaktickou **komplexitu *absolutní*** („objektivní“, tj. *formal properties*, viz Brunato and Venturi 2022: 1) a nikoli *relativní* („subjektivní“ – orientovaná na uživatele jazyka a posuzující míru náročnosti zpracování dané věty/textu, „readability“), viz Szmrecsanyi and Kortmann 2012: 10.

Míry syntaktické komplexity (a lexikální diverzity) v InterCorpu v16ud

Míry (na úrovni věty):

1. sLength (délka věty v počtu slov)
2. subRatio (subordination ratio) $\frac{(\text{Nº T-units} + \text{Nº Sub})}{\text{Nº T-units}}$
3. maxTreeDepth
4. maxNP**Depth**
5. maxNP**Length**
6. mdd = *mean dependency distance*
7. LEXIKÁLNÍ: lexDivWord a lexDivLemma

“Manning’s law” pro SCMs:

Míry syntaktické komplexity implementované do mnohojazyčného korpusu musí být zároveň:

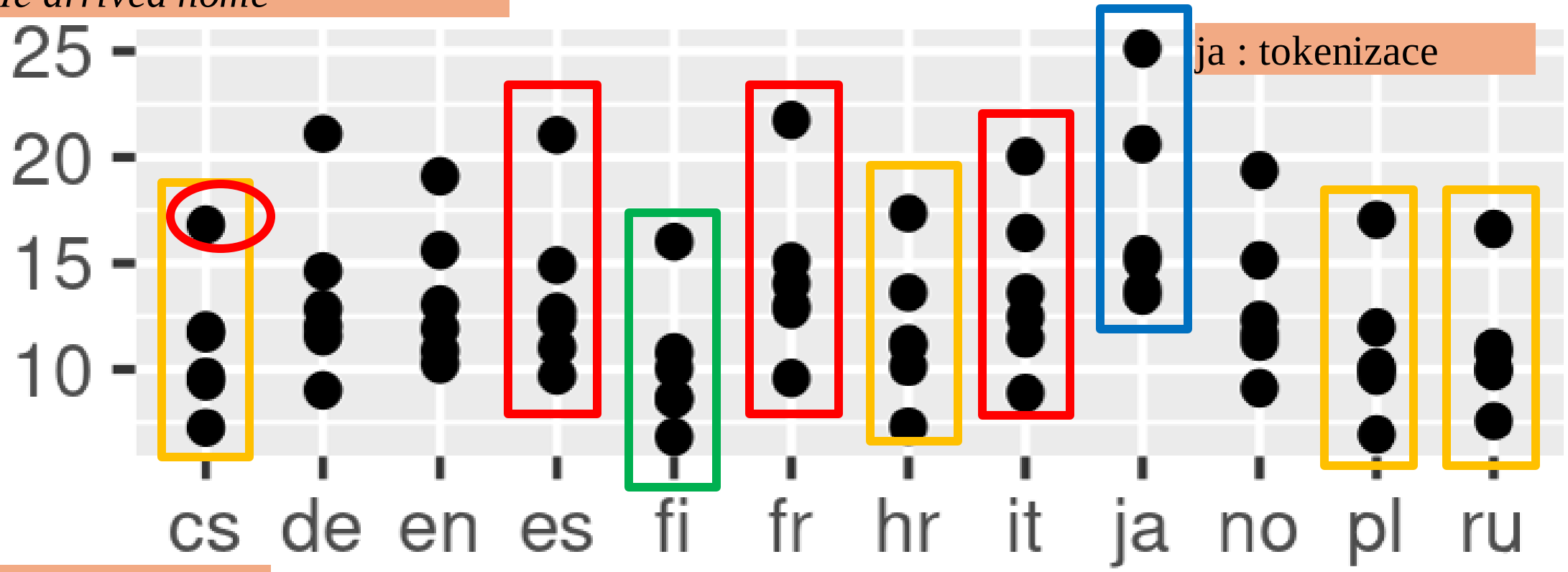
1) implementovatelné pomocí UD
2) spolehlivé a srovnatelné napříč jazyky
3) spolehlivé a srovnatelné napříč textovými typy
4) známé a zavedené (replikovatelnost předchozích v srovnatelnost)

fr *Il est arrivé à la maison.(6)*
cs: *Přišel domů (2 slova)*
(Nádvorníková 2020)

s_length

Scatterplot – average sentence length (6 textů / 12 jazyků, *fiction*)

fr *Il est arrivé à la maison.(6 w.)*
cs: *Přišel domů (2 words)*
‚He arrived home‘



ja : tokenizace

faktor ve *fiction* :
styl textu

Osnova

1. Úvod
2. Paralelní korpus InterCorp
3. Anotace InterCorpu
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu
    3. Ukaž a hledej
6. Diskuse, otázky…

Míry komplexity a diverzity jako metadata

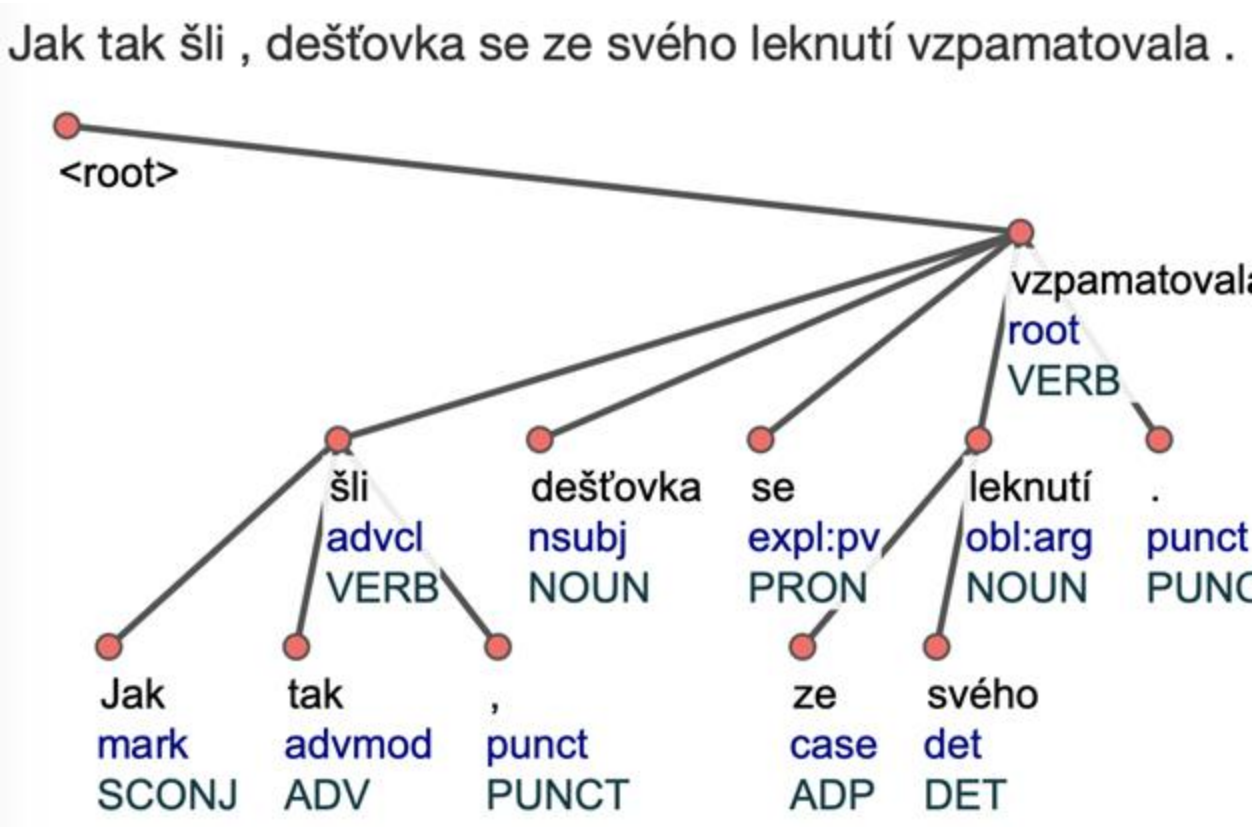
Zobrazení → Korpusová nastavení → Struktury: **<text>, <s>** Zobrazení → KWIC/**Věta**

Atributy **<text>**u:

<text
author=Čapek, Josef
title=Povídání o pejskovi a kočičce
lexDivWord=463.83
lexDivLemma=304.68
subRatioAvg=1.72
maxTreeDepthAvg=0.89
sLengthAvg=14.08
mdd=2.69
maxNPLengthAvg=2.65
maxNPDepthAvg=1.02
…
>

Atributy **<s>** (věty):

<s
id=cs:Capek-O_pejskovi_a_koc:0:28:1
maxNPDepth=1
subRatio=2.0
sLength=9
maxNPLength=3
mdd=2.75
maxTreeDepth=1
>



Míry syntaktické komplexity na úrovni **věty**

<table>
<tr><th></th><th>jmenná fráze</th><th>věta (klauze)</th></tr>
<tr><td rowspan="2">horizontální dimenze</td><td rowspan="2">maxNPLength
maximální délka</td><td>sLength
délka v počtu slov</td></tr>
<tr><td rowspan="2">subRatio
subordinační poměr</td></tr>
<tr><td>vertikální dimenze</td><td>maxNPDepth
maximální hloubka</td><td>maxTreeDepth
maximální hloubka stromu</td></tr>
<tr><td>kognitivní náročnost</td><td></td><td>mdd
průměrná délka závislostí</td></tr>
</table>

Míry syntaktické komplexity na úrovni **věty**

	jmenná fráze	**věta (klauze)**
horizontální dimenze	maxNPLength *maximální délka*	sLength *délka v počtu slov*
		subRatio *subordinační poměr*
vertikální dimenze	maxNPDepth *maximální hloubka*	maxTreeDepth *maximální hloubka stromu*
kognitivní náročnost		mdd *průměrná délka závislostí*

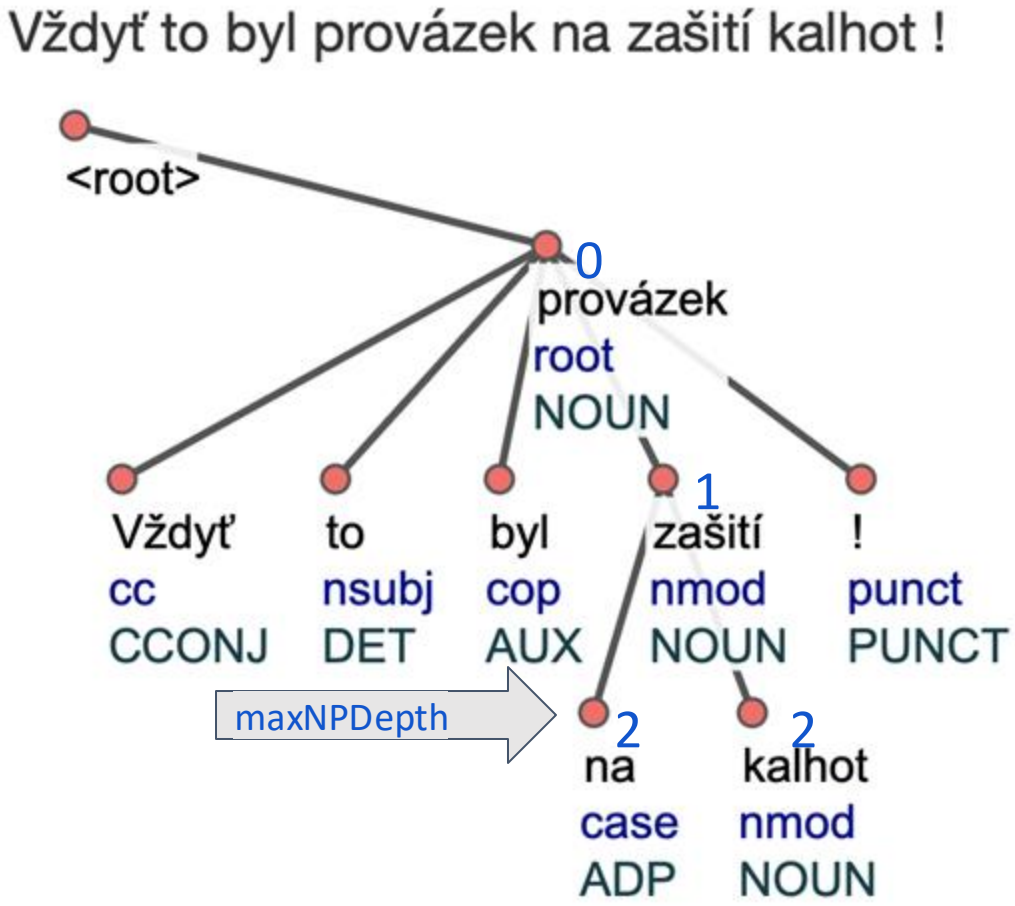
Jmenná fráze – míry syntaktické komplexity

MaxNPLength:

- počet slov v nejdelší jmenné frázi
- *provázek na zašití kalhot*
- = 4

MaxNPDepth:

- maximální počet zanoření ve jmenné frázi
- *provázek* … 0
- *zašití* … 1
- *na* … 2
- *kalhot* … 2
- = 2



Míry syntaktické komplexity na úrovni **věty**

	jmenná fráze	**věta (klauze)**
horizontální dimenze	maxNPLength *maximální délka*	sLength *délka v počtu slov*
		subRatio *subordinační poměr*
vertikální dimenze	maxNPDepth *maximální hloubka*	maxTreeDepth *maximální hloubka stromu*
kognitivní náročnost		mdd *průměrná délka závislostí*

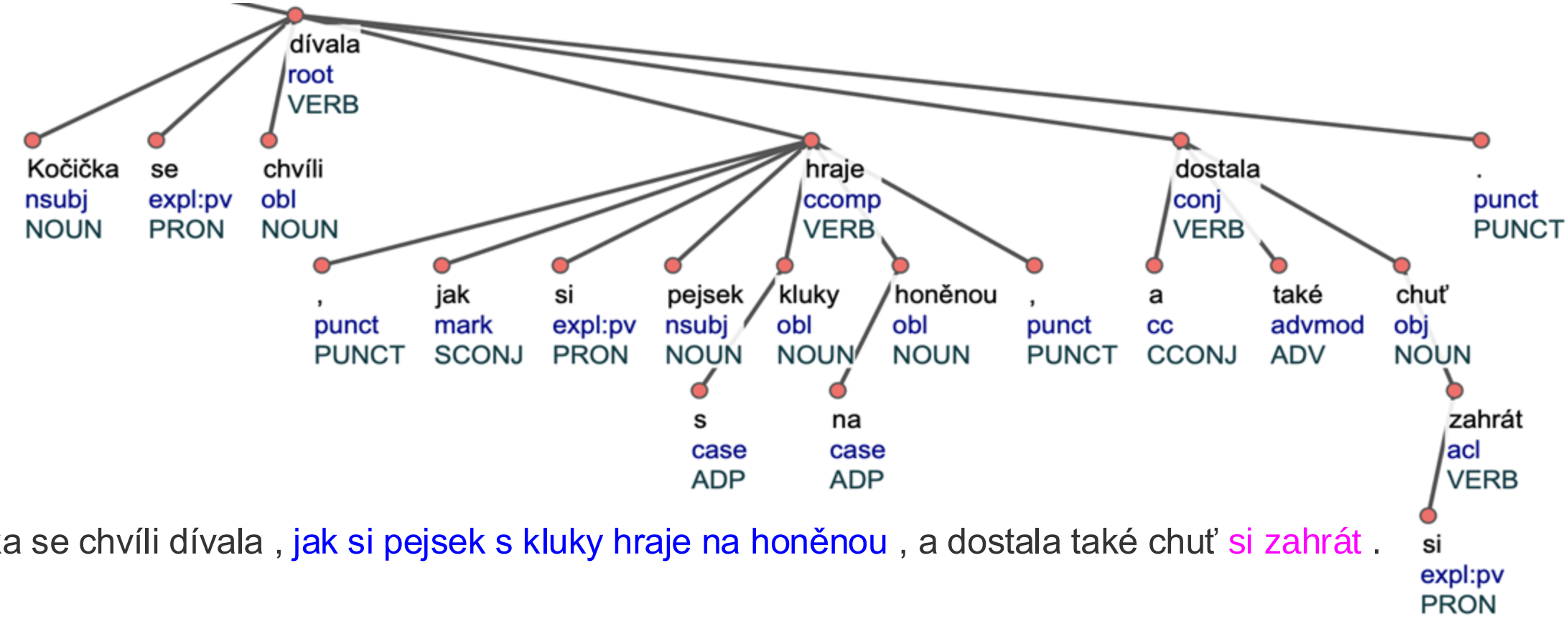
Co je to věta

T-unit:

- hlavní věta včetně všech na ní závislých (Hunt 1965)
- každá koordinovaná hlavní věta, včetně všech závislých, je jeden T-unit

Vedlejší věta (klauze), i nefinitní:

- csubj – podmětová
- ccomp – předmětová
- xcomp – “otevřený” predikát (doplněk, …)
- advcl – příslovečná
- acl – přívlastková



Kočička se chvíli dívala , jak si pejsek s kluky hraje na honěnou , a dostala také chuť si zahrát .

Věta – míry syntaktické komplexity

sLength:

- počet slov ve větě
- bez interpunkce

MaxTreeDepth:

- maximální počet zanoření vedlejších vět
- koordinace se nepočítá

subRatio:

- subordinanční poměr
- (počet T-units + počet vedlejších vět) / počet T-units

Věta

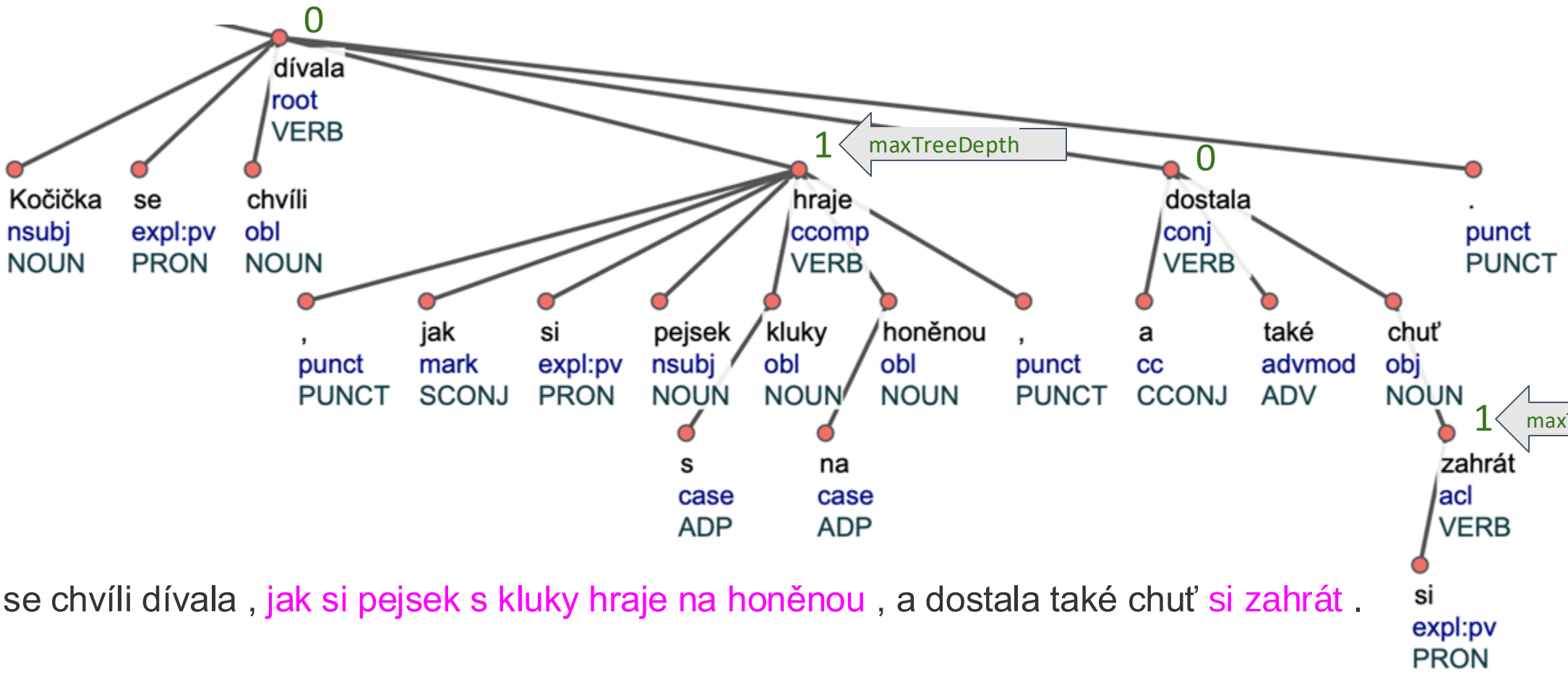
maxNPDepth=2 **subRatio=2.0** sLength=18

maxNPLength=3 mdd=2.71 **maxTreeDepth=1**

počet T-units = 2

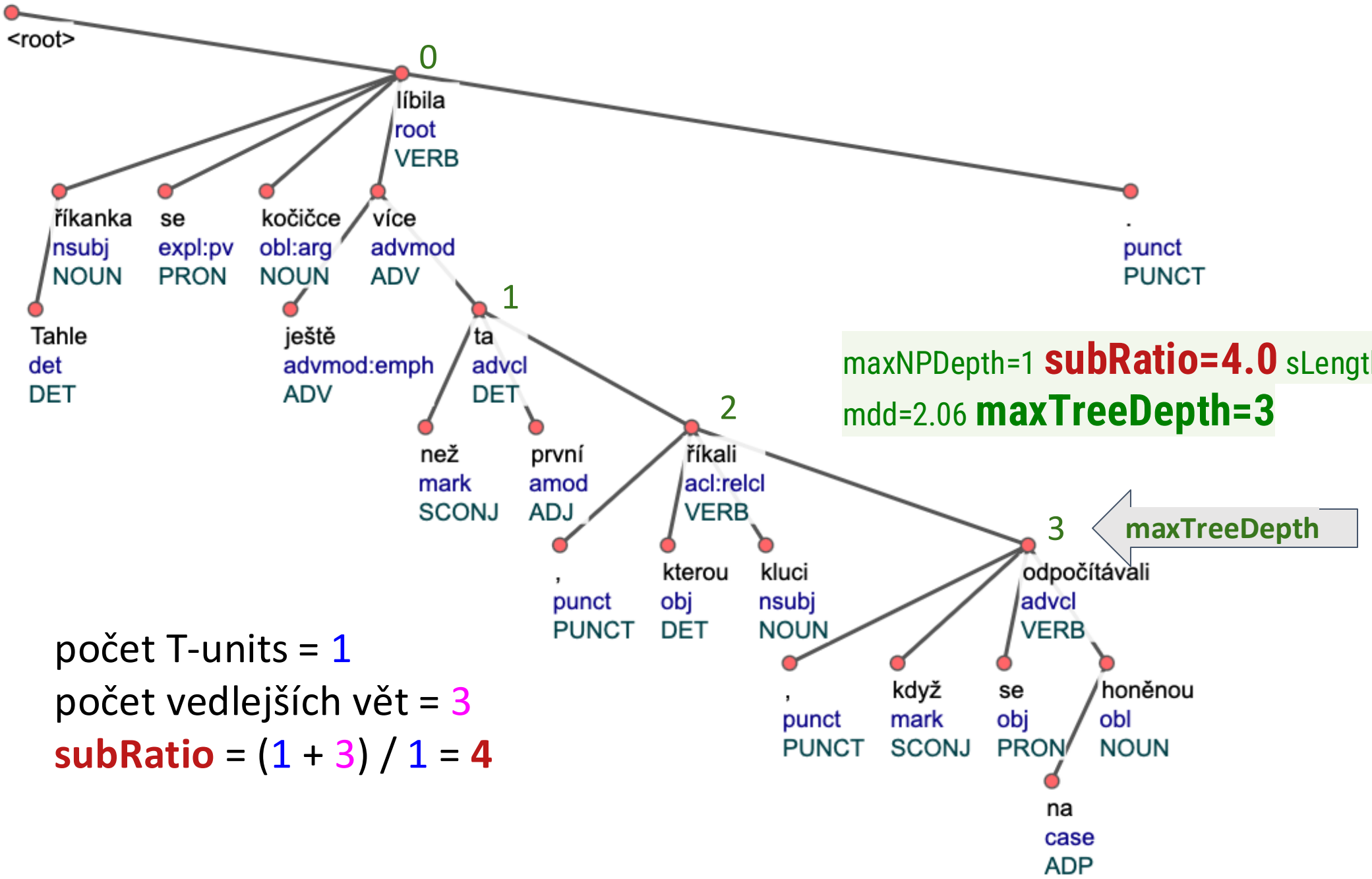
počet vedlejších vět = 2

subRatio = (2 + 2) / 2 = **2**



Kočička se chvíli dívala , jak si pejsek s kluky hraje na honěnou , a dostala také chuť si zahrát .

Tahle říkanka se kočičce ještě více líbila než ta první , kterou říkali kluci , když se odpočítávali na honěnou .



maxNPDepth=1 **subRatio=4.0** sLength=18 maxNPLength=2
mdd=2.06 **maxTreeDepth=3**

počet T-units = 1
počet vedlejších vět = 3
subRatio = (1 + 3) / 1 = **4**

Míry syntaktické komplexity na úrovni **věty**

<table>
<tr><th></th><th>jmenná fráze</th><th>věta (klauze)</th></tr>
<tr><td rowspan="2">horizontální dimenze</td><td rowspan="2">maxNPLength
maximální délka</td><td>sLength
délka v počtu slov</td></tr>
<tr><td rowspan="2">subRatio
subordinační poměr</td></tr>
<tr><td>vertikální dimenze</td><td>maxNPDepth
maximální hloubka</td><td>maxTreeDepth
maximální hloubka stromu</td></tr>
<tr><td>kognitivní náročnost</td><td></td><td>mdd
průměrná délka závislostí</td></tr>
</table>

Věta – kognitivní náročnost

MDD:

ze všech měr nejmíň závislá na typu jazyka

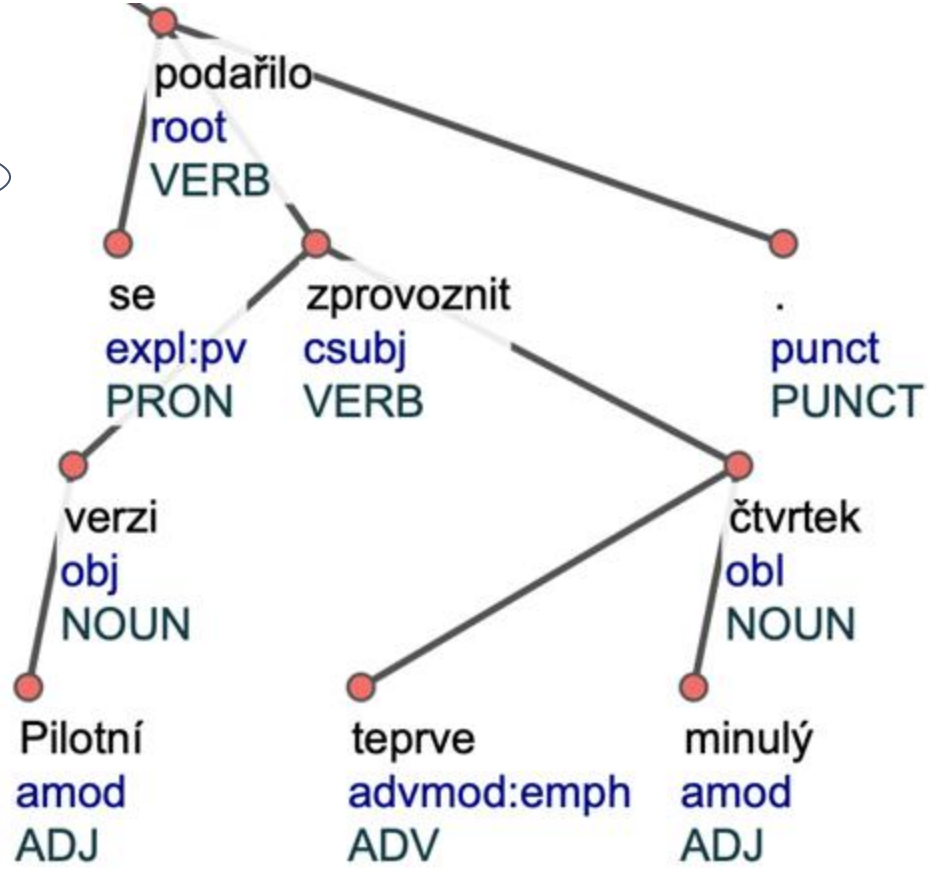
- Mean Dependency Distance
  (Yan & Li, 2019; Mačutek et al., 2021)
- průměrná vzdálenost závislého členu od řídícího v textu
- bez interpunkce
- výpočet (*n* … počet slov ve větě)

  $DD_i = | ID_i - head_i |$

  $DD = \sum_{i = 0 \text{ až } n} DD_i$

  $mdd = DD / (n - 1)$
- DD = 12

  MDD = 12 / 7 ≅ 1,71



	Pilotní	verzi	se	podařilo	zprovoznit	teprve	minulý	čtvrtek
ID (= *i*)	1	2	3	4	5	6	7	8
$head_i$	2	5	4	0	4	8	8	5
DD_i	1	3	1	0	1	2	1	3

Míry syntaktické komplexity na úrovni **textu**

	jmenné fráze	**věty**
horizontální dimenze	maxNPLength**Avg** ***průměrná*** *maximální délka*	sLength**Avg** ***průměrná*** *délka v počtu slov*
		subRatio**Avg** ***průměrný*** *subordinační poměr*
vertikální dimenze	maxNPDepth**Avg** ***průměrná*** *maximální hloubka*	maxTreeDepth**Avg** ***průměrná*** *maximální hloubka stromu*
kognitivní náročnost		mdd *průměrná délka závislostí*

Míry lexikální diverzity na úrovni **textu**

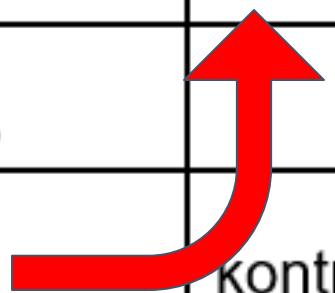
- varianta míry *type-token ratio*
- počet různých *typů* v pohyblivém okně o 1000 tokenech
- nedefinováno, je-li text kratší
- počet různých *slovních tvarů*: lexDivWord
  - cs: 421–732, en: 350–563
- počet různých *lexémů*: lexDivLemma
  - cs: 279–629, en: 281–494

Osnova

1. Úvod
2. Paralelní korpus InterCorp
3. Anotace InterCorpu
    1. Universal Dependencies
    2. Syntaktická anotace a její implementace v InterCorpu
4. Praktické ukázky vyhledávání pomocí UD
5. InterCorp: Míry syntaktické komplexity a lexikální diverzity
    1. Co to je a proč to měřit?
    2. Anotace komplexity a diverzity v InterCorpu
    3. Ukaž a hledej
6. Diskuse, otázky…

Možné aplikace

FICTION	cs	fr	en	fi	další…
CS		kontr/transl	kontr/transl	kontr/transl	
FR	kontr/transl		kontr/transl	kontr/transl	
EN	kontr/transl	kontr/transl		kontr/transl	
FI	kontr/transl	kontr/transl	kontr/transl		
další…					

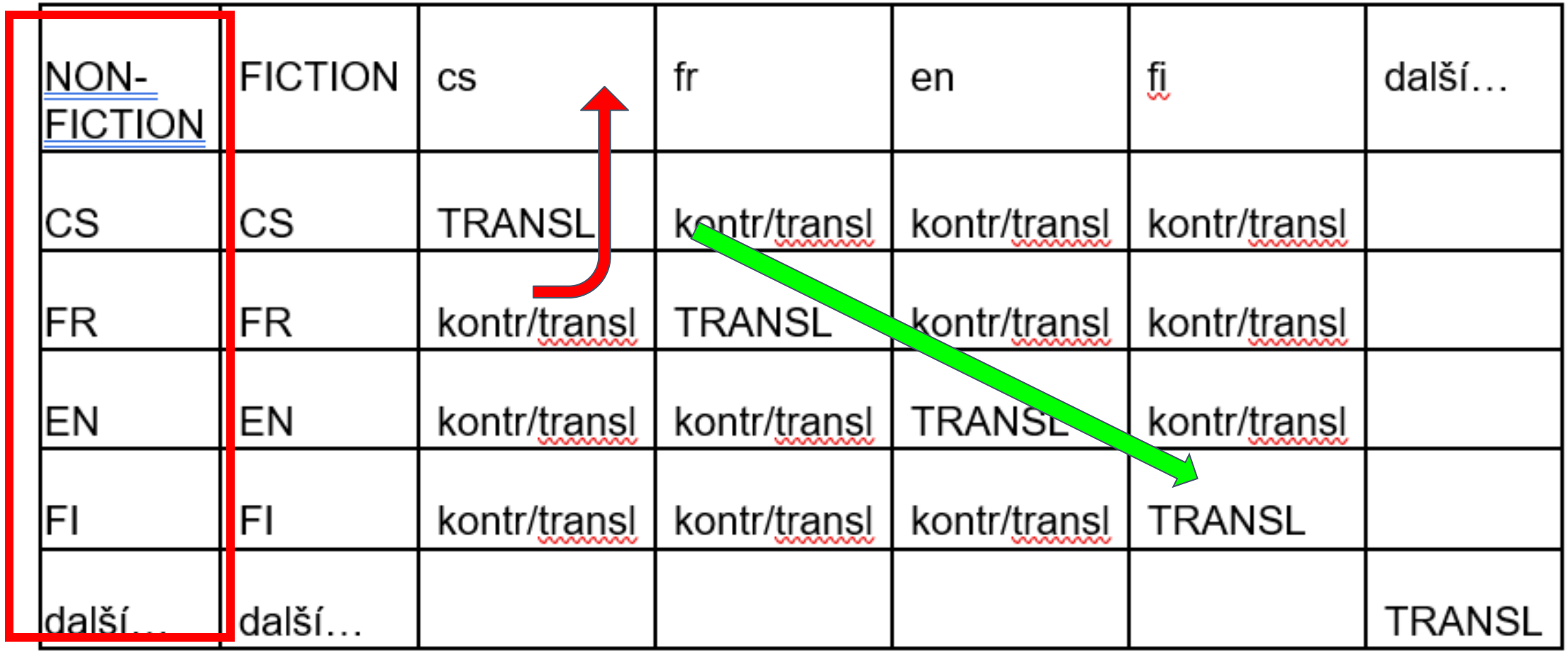


Možné aplikace

FICTION	cs	fr	en	fi	další…
CS	TRANSL	kontr/transl	kontr/transl	kontr/transl	
FR	kontr/transl	TRANSL	kontr/transl	kontr/transl	
EN	kontr/transl	kontr/transl	TRANSL	kontr/transl	
FI	kontr/transl	kontr/transl	kontr/transl	TRANSL	
další…					TRANSL

Možné aplikace

NON-FICTION	FICTION	cs	fr	en	fi	další…
CS	CS	TRANSL	kontr/transl	kontr/transl	kontr/transl	
FR	FR	kontr/transl	TRANSL	kontr/transl	kontr/transl	
EN	EN	kontr/transl	kontr/transl	TRANSL	kontr/transl	
FI	FI	kontr/transl	kontr/transl	kontr/transl	TRANSL	
další…	další…					TRANSL



Vyhledávání měr syntaktické komplexity (a lexikální diverzity)

1. Běžný dotaz → Zobrazení měr (na úrovni vět)
    a) Metadata
    b) Struktury
2. Specifikace měr přímo v dotazu
3. Zobrazení měr u jednotlivých textů, stažení výsledku a vyhodnocování (**NEW!** a porovnání s údaji pro subkorpusy kolekcí a textových typů: https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze16ud#detailed_statistics)

Míry lexikální diverzity a syntaktické komplexity pro textové typy/jazyky (**NEW**)

Jazyk	Kolekce	Počet		Tisíce			Lexikální diverzita		Syntaktická komplexita (průměr)					
		dokumentů	textů	vět	slov	tokenů	lexDivWord	lexDivLemma	sLength	subRatio	maxTreeDepth	maxNPLength	maxNPDepth	mdd
af	Subtitles	1	24	23,0	134,6	161,7	406,4	347,2	5,887	1,093	0,095	2,377	0,811	2,251
ar	Core-fiction	2	2	2,1	28,8	35,6	620,3	576,6	13,830	2,712	1,310	5,293	2,016	2,817
	Core-misc	1	1	1,3	5,5	7,4	451,4	421,4	4,150	1,330	0,290	1,870	0,840	2,010
	Subtitles	1	34 193	28 726,4	126 195,5	157 188,9	592,8	557,3	4,421	1,338	0,336	2,216	0,986	1,678
	Syndicate	3	433	19,0	384,5	439,0	622,7	560,3	20,513	2,485	1,312	11,036	3,940	2,405
be	Core-fiction	104	104	625,1	7 068,7	8 978,9	615,4	492,7	11,583	1,865	0,804	4,122	1,436	2,316
	Core-misc	4	4	7,6	57,7	76,0	556,2	425,6	7,608	1,672	0,605	2,870	1,002	2,254
bg	Core-fiction	87	87	559,6	7 067,3	8 597,7	548,3	439,5	13,125	1,728	0,732	4,255	1,532	2,497
	Acquis	1	10 846	862,3	13 582,3	16 991,2	392,4	306,3	18,073	1,801	0,514	9,389	2,805	3,265
	Europarl	1	45 271	408,3	9 082,0	10 379,8	498,4	386,3	23,014	2,538	1,263	10,961	3,402	2,581
	Subtitles	1	40 986	32 591,1	164 644,1	214 988,4	518,2	384,6	5,089	1,336	0,322	1,861	0,706	1,931

1. Běžný dotaz → Zobrazení měr

- InterCorp v16ud – Czech
- zarovnaný korpus English/French…
- pokročilý dotaz
- koordinované hlavní věty: [deprel="conj" & p_deprel="root"]

Zobrazení → Korpusová nastavení → Metainformace → <s> →

subRatio + sLength + maxTreeDepth (příp. další)

→ dole: POUŽÍT VOLBY ZOBRAZENÍ

Korpusová nastavení pro InterCorp v16ud - Czech

Poziční atributy | Struktury | Metainformace | Rozšiřující funkce

- [ ] <#>
  - [ ] Počet tokenů
- [ ] <doc>
  - [ ] Pořadí dokumentu
  - [ ] doc.id
  - [ ] doc.tag_model
- [ ] <text>
  - [ ] text.lang
  - [ ] text.pubyear
  - [ ] text.version
  - [ ] text.pubmonth
  - [ ] text.pubDateYear
  - [ ] text.pubDateMonth
  - [ ] text.id
  - [ ] text.author
  - [ ] text.title
  - [ ] text.group
  - [ ] text.publisher
- [ ] <p>
  - [ ] p.id
- [x] <s>
  - [ ] s.id
  - [ ] s.maxNPDepth
  - [x] s.subRatio
  - [x] s.sLength
  - [ ] s.maxNPLength
  - [ ] s.mdd
  - [x] s.maxTreeDepth

CS

Vaculik-Cesky_snar ✦ 2.5 ✦ 18 ✦ 2

Dal jsem mu povinnou výstrahu , aby to nedělal , a **doporučil** / mu dotáhnout pointu pro případ , že to udělá .

Houellebecq-Moznost_os ✦ 1.0 ✦ 20 ✦ 0

Vyhýbal jsem se revolucím bolestným a zbytečným - neboť původ všeho zla je **biologický** / a nezávislý na jakékoli představitelné společenské transformaci ;

FR

Vaculik-Cesky_snar ✦ 3.5 ✦ 38 ✦ 2

Comme il se doit , je lui ai donné l' avertissement obligatoire en lui demandant de ne pas le faire et je lui ai recommandé de travailler la péroraison dans le cas où il le ferait tout de même .

Houellebecq-Moznost_os ✦ 3.0 ✦ 22 ✦ 1

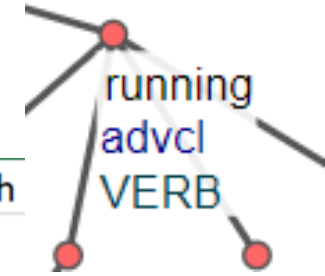
J' évitais au monde des révolutions douloureuses et inutiles – puisque la racine de tout mal était biologique , et indépendante d' aucune transformationsocialeimaginable ;

TIP:https://lindat.mff.cuni.cz//services/udpipe/

TIP: stáhnout konkordanci (nebo vzorek)

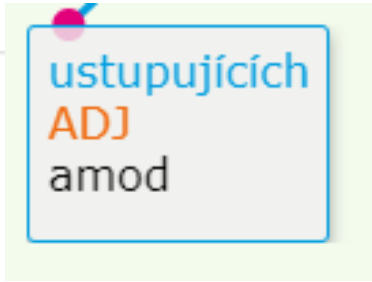
Konkordance → Vzorek
Uložit → csv, xlsx…

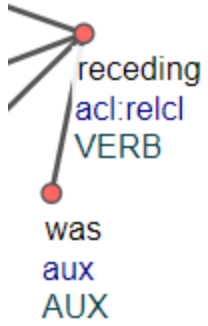
doc_id	sub.ra	s_leng	max.tr	cs1	KWIC	cs2	sub.ra	s_leng	max.tree.depth
adams-daleka	1.0	19	0	Krátce vyštěkl a vzápětí nato	byli/	venku z houští na otevřené stráni a pes uháněl bez hlesu za nimi .	2.0	19	2
adams-preva	1.0	19	0	Laseroví odečítači kmitali sem a tam , snímali mu otisky prstů ,	prosvítili/	sítnici a vlasový míšek v místech ustupujících vlasů .	4.0	25	2



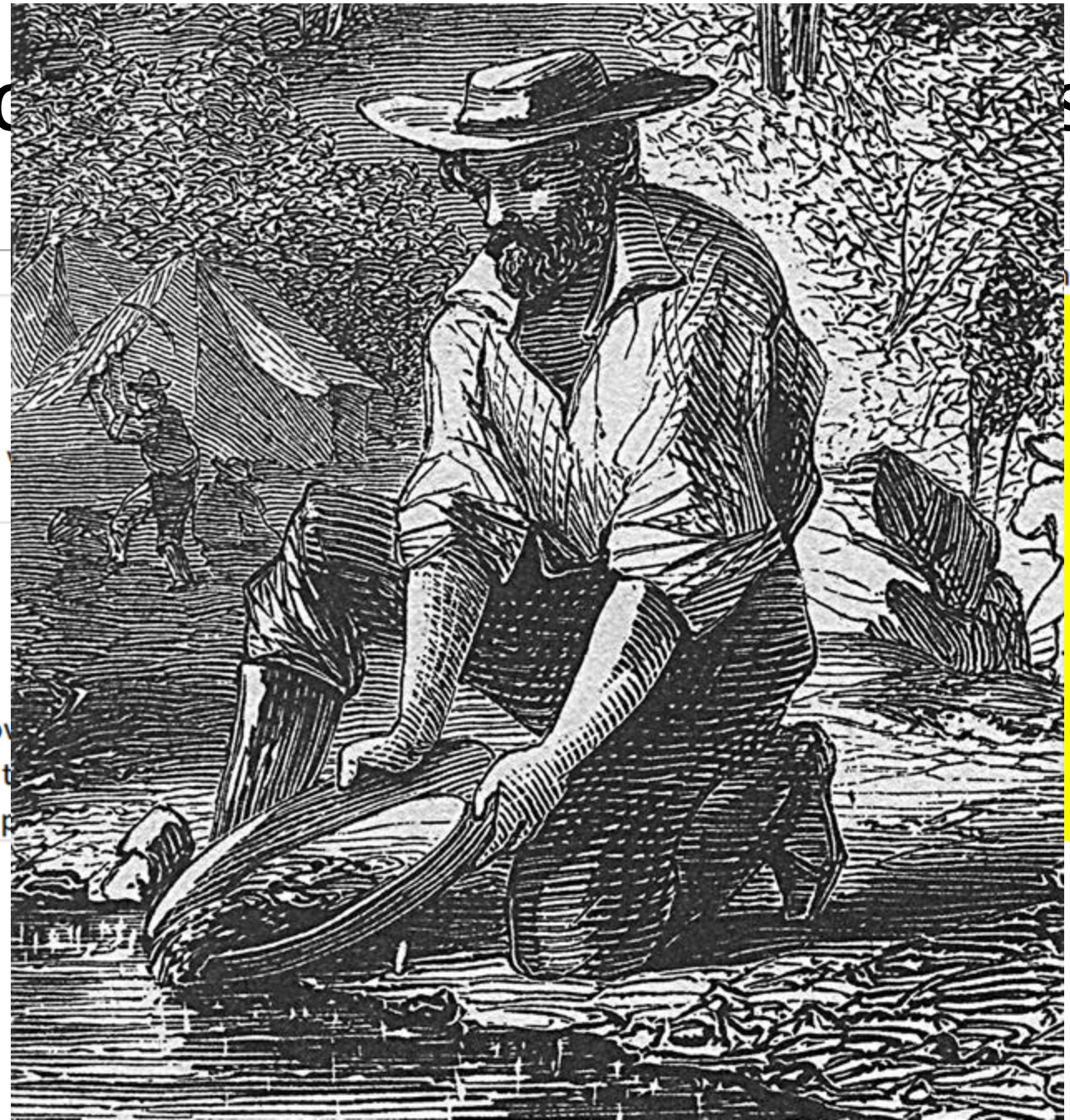
It barked once and then they were out on the open slope with the dog running mute behind them .

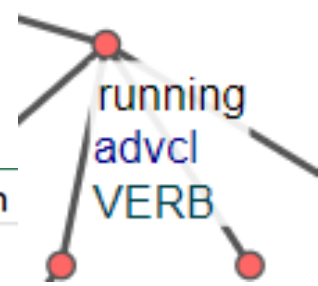
The laser readers were becoming very agitated as they flickered over his fingerprints , his retina and the follicle pattern where his hair line was receding .





Konkordanc sx aj.





doc_id	sub.ra	s_leng	max.tr	cs1	ax.tree.depth	
adams-daleka	1.0	19	0	Krátce v nato		It barked once and then they were out on the open slope with the dog running mute behind them .
adams-preva	1.0	19	0	Laserov sem a t otisky p		The laser readers were becoming very agitated as they flickered over his fingerprints , his retina and the follicle pattern where his hair line was receding .

2. Specifikace měr přímo v dotazu

PŘÍKLAD 1 – co najde tento výraz?

<s maxTreeDepth="0" & sLength <= "10" />

Chci v takovém typu vět najít koordinované podměty

(typ *Pejsek a kočička vařili dort*):

[deprel="conj" & p_deprel="nsubj.*"]

within <s maxTreeDepth="0" & sLength <= "10" />

Specifikace měr přímo v dotazu – kontrastivní vyhledávání

PŘÍKLAD 2 – koordinované vztažné věty kontrastivně

French – zarovnaný korpus Czech

CO NAJDE TENTO VÝRAZ?

[deprel="conj" & p_deprel="acl:relcl"] within <s maxTreeDepth >= "3" />

UPŘESNĚNÍ V ZAROVNANÉM KORPUSU – CZECH:

“pokročilý dotaz”, “Překlad **obsahuje** odpovídající výsledky”

[deprel="conj" & p_deprel="acl:relcl"] **+ NEW - lze podmínku within**

Specifikace měr přímo v dotazu – maxNPDepth a maxNPLength

```
[deprel="nsubj:pass*"]
within <s maxNPDepth >="10" & maxNPLength >="40"/>
```

Zarovnané korpusy CS-EN

https://www.korpus.cz/kontext/view?q=~GiOiMyS0uU6c

(Konkordance ➞ Trvalý odkaz)

3. Zobrazení měr u jednotlivých textů, stažení výsledku a vyhodnocování

InterCorp v16ud – Czech

Dotaz: <text>

→ Zobrazení

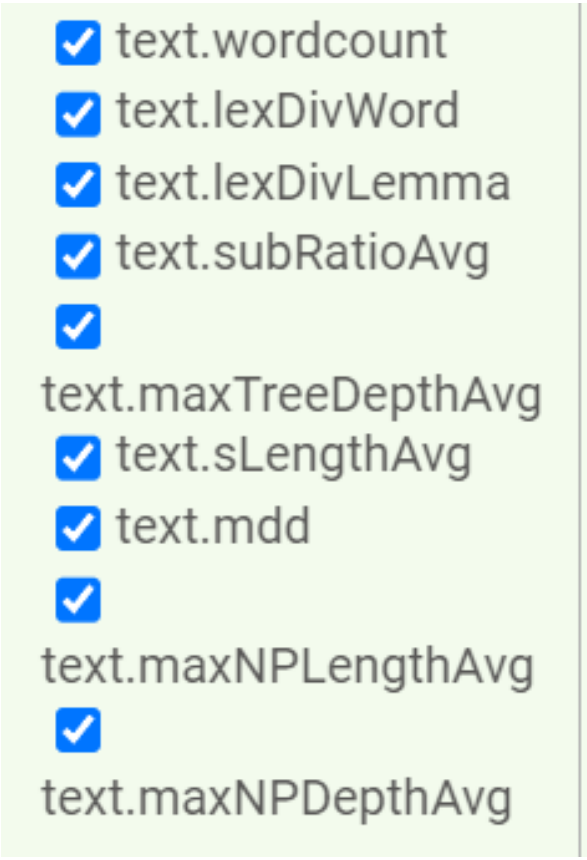
→ Korpusová nastavení

→ Metainformace

→ Použít volby zobrazení

→ Uložit

→ .csv/.xlsx aj. (1811 textů)



Vlastnosti věty – údaje za celé texty (NP)

text.maxNPDepthAvg

text.maxNPLengthAvg

MAX: 1. text.maxNPDepthAvg, 2. text.maxNPLengthAvg

author	title	srclang	wordcount	subRatioAvg	maxTreeDepth	sLengthAvg	mdd	maxNPLengt	maxNPDepthAvg
García Márquez, Gabri	Podzim patriarchy	es	70478	4,32	4,29	310,53	7,36	68,23	7,72
Hrabal, Bohumil	Taneční hodiny pro s	cs	17460	2,37	2,70	873,05	10,37	72,20	5,20
Bourdieu, Pierre	Teorie jednání	fr	49271	3,83	2,07	35,57	3,12	17,35	4,30
Antunes, António Lobo	Jidášova díra	pt	47151	2,56	1,74	41,59	3,49	16,10	4,08
Meyer, Thomas	Transformace sociá	de	47109	2,74	1,39	29,75	2,95	14,43	3,85
Patočka, Jan	Kacířske esejé o filo	cs	42207	2,96	1,59	27,88	2,81	13,27	3,61
	NATO v 21. století	en	4667	1,76	0,65	22,54	2,49	11,30	3,54
Agamben, Giorgio	Prostředky bez účel	it	23433	3,11	1,69	26,45	2,88	12,51	3,52
Hayek, Friedrich A.	Cesta do otroctví	en	59790	3,49	1,89	25,55	2,89	11,30	3,47
Mandiargues, André Pi	Vlčí slunce	fr	36051	2,98	1,66	28,89	2,97	12,08	3,46
	Transformované NA	en	16272	1,78	0,72	21,13	2,52	11,33	3,46
Patočka, Jan	Úvod do Husserlovy	cs	54680	2,83	1,44	25,08	2,78	11,92	3,43
Lévi-Strauss, Claude	Rasa a dějiny	fr	13159	3,48	1,87	26,72	2,81	11,32	3,41
Procacci, Giuliano	Dějiny Itálie	it	134343	2,29	1,18	25,82	2,76	11,83	3,40
Havel, Václav	Moc bezmocných	cs	24098	3,40	1,71	33,78	3,33	14,33	3,39
Souček, Ludvík	Tušení stínů	cs	98280	2,09	1,04	23,30	2,77	11,88	3,35

Vlastnosti věty – údaje za celé texty (NP)

text.maxNPDepthAvg
text.maxNPLengthAvg

Souvislosti např.
s frekvencí koordinovaných vztažných vět?

[deprel="conj"
& p_deprel="acl:relcl"]

	Filtr	doc.id	Freq	i.p.m. ▼
1	p / n	Garcia_Marquez-podzim	513	6 267,64
2	p / n	Foucault-Slova_a_veci	700	4 803,93
3	p / n	Andric-Most_na_Drine	608	4 653,05
4	p / n	Obama-Inauguracni_rec	11	4 539,83
5	p / n	Andric-Travnicka_kron	752	4 478,35
6	p / n	Faulkner-Mesto	575	3 867,18
7	p / n	Ajvaz-Zlaty_vek	366	3 865,12
8	p / n	Proust-Swann	611	3 722,8
9	p / n	Hrabal-Obsluhoval_pov	262	3 325,97
10	p / n	Bruckner-Pokuseni	246	3 304,1
11	p / n	Ajvaz-Druhe_mesto	155	3 255,55
12	p / n	Ourednik-Europeana	94	3 193,59
13	p / n	allende-dum_duchu	524	3 113,77

Vlastnosti věty – údaje za celé texty

text.subRatioAvg

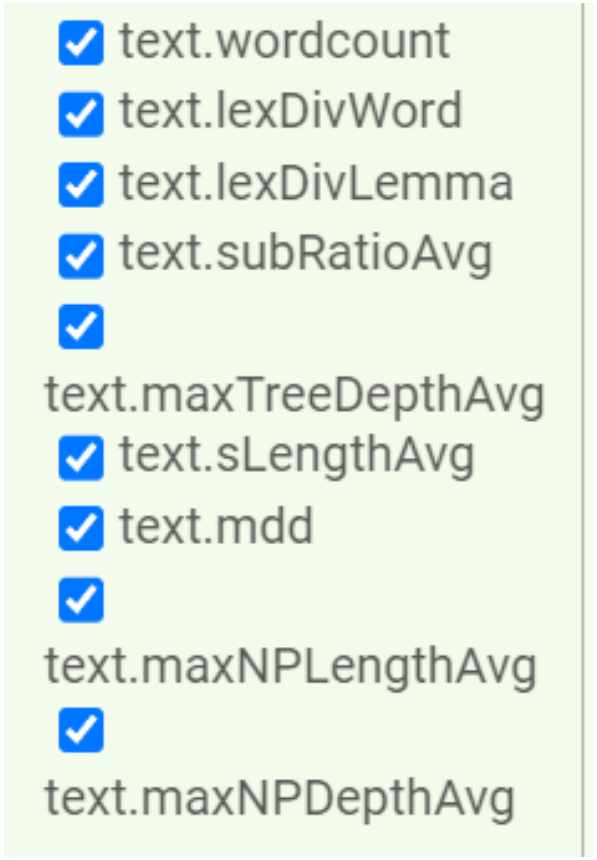
text.maxTreeDepthAvg

Nejvyšší v češtině? (CS, cs, core)

- src.lang?
- autor
- typ textu? (fiction, non-fiction, poetry, drama…)

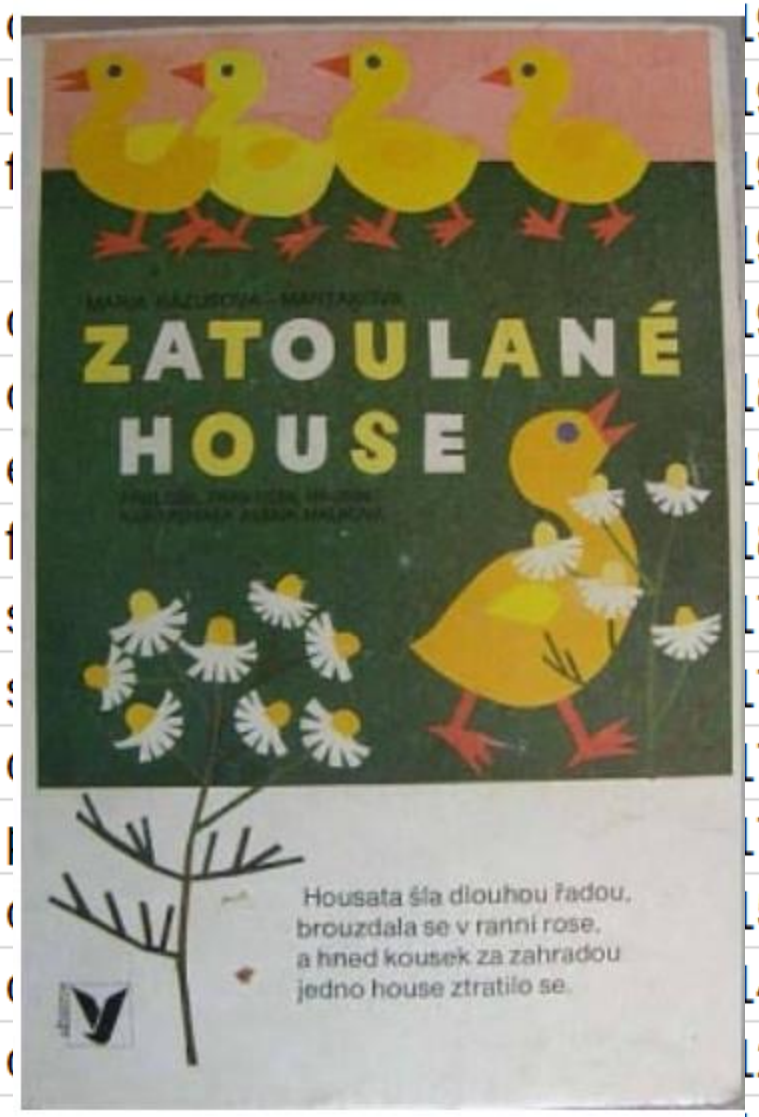
Viz tabulka ve wiki.korpus.cz:

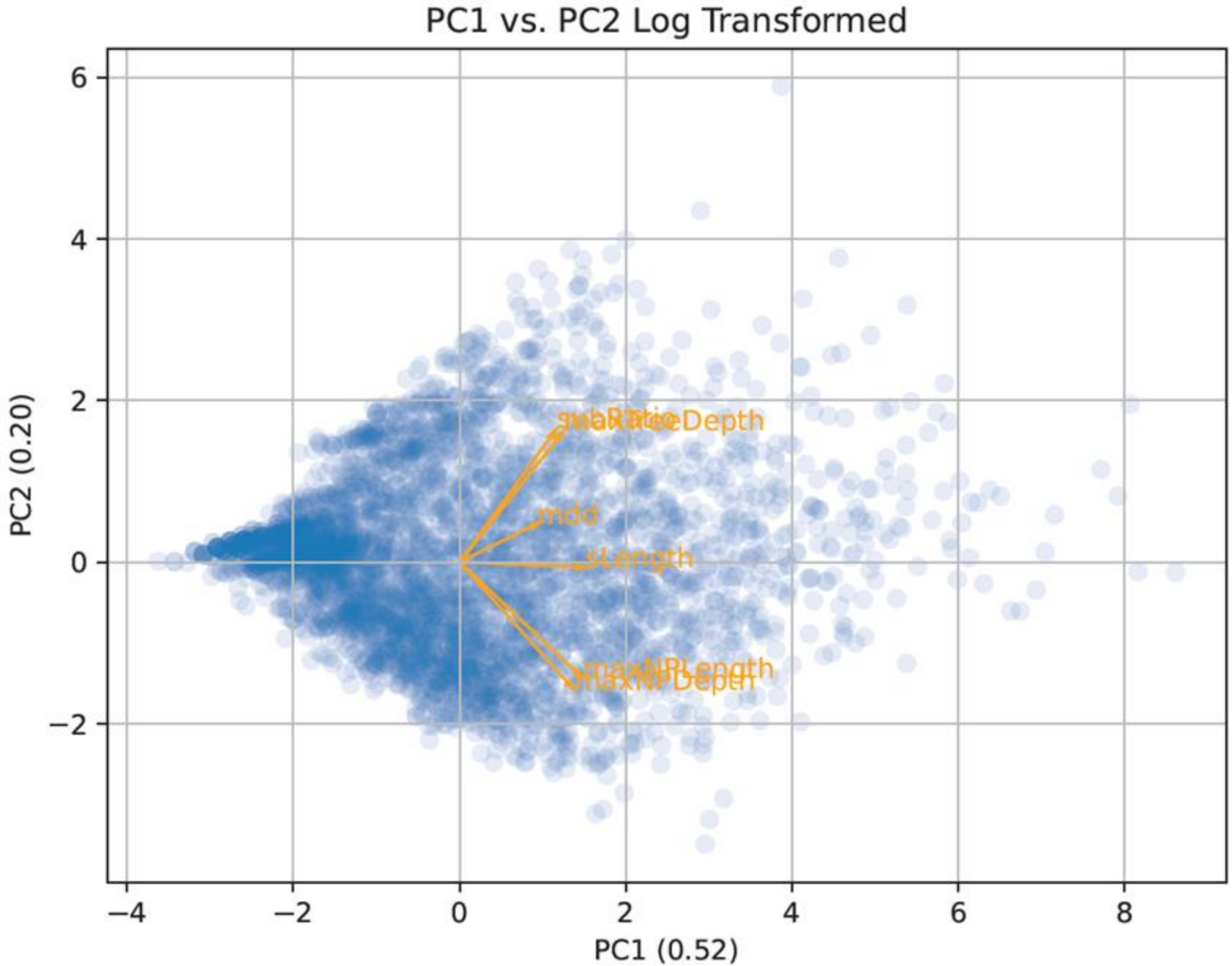
en:cnk:intercorp:verze16ud



author	title	srclang	wordcount	subRatioAvg	maxTreeDepth	sLengthAvg
Melchor, Fernanda	Období hurikánů	es	60085	4,53	2,18	63,19
García Márquez, Gabri	Podzim patriarchy	es	70478	4,32	4,29	310,53
Böll, Heinrich	Konec jedné služebi	de	48109	3,99	1,95	40,79
Bourdieu, Pierre	Teorie jednání	fr	49271	3,83	2,07	35,57
Hayek, Friedrich A.	Cesta do otroctví	en	59790	3,49	1,89	25,55
Lévi-Strauss, Claude	Rasa a dějiny	fr	13159	3,48	1,87	26,72
Proust, Marcel	Hledání ztraceného	fr	135949	3,43	1,79	27,39
Havel, Václav	Moc bezmocných	cs	24098	3,40	1,71	33,78
Čapek, Karel	O věcech obecných	cs	30381	3,33	1,58	20,94
Leiris, Michael	Věk dospělosti	fr	42802	3,26	1,70	29,89
Carpentier, Alejo	Harfa a stín	es	43193	3,16	1,62	26,76
Pamuk, Orhan	Istanbul: vzpomínky	tr	94327	3,13	1,63	29,41
Agamben, Giorgio	Prostředky bez účel	it	23433	3,11	1,69	26,45
Čapek, Karel	Výlet do Španěl	cs	18663	3,04	1,32	24,66
Böll, Heinrich	Biliár o půl desáté	de	76616	2,99	1,05	24,14
Čep, Jan	Proměny	cs	1891	2,98	1,77	30,50
Mandiargues, André Pi	Vlčí slunce	fr	36051	2,98	1,66	28,89
Bernhard, Thomas	Wittgensteinův sync	de	27487	2,97	1,68	29,85
Patočka, Jan	Kacířske esejé o filo	cs	42207	2,96	1,59	27,88

author	title	srclang	wordcount	subRatioAvg	maxTreeDepth	sLengthAvg
Goscinny, René; Uderz	Asterix z Galie	fr	4380	1,20	0,21	4,20
Venclova, Tomas	Čas rozpůlil se... / Jp	lt	6467	1,20	0,22	6,22
Topol, Josef	Kočka na kolejích			19	0,20	4,33
Ābele, Inga	Ostřice			19	0,19	4,13
Goscinny, René; Uderz	Asterix a cesta kolen			19	0,20	4,10
Sofokles	Antigoné			19	0,21	4,75
Šotola, Jiří	Podzim v zahradní re			19	0,19	6,32
Čapek, Karel	Věc Makropulos			18	0,18	3,69
Arriaga, Guillermo	Psí lásky			18	0,20	4,81
Jarry, Alfred	Ubu			18	0,18	4,68
Karvaš, Peter	Antigona a ti druzí			17	0,16	3,46
Karvaš, Peter	Půlnoční mše			17	0,24	5,80
Biebl, Konstantín	Nový Ikaros			17	0,16	5,06
Krynicki, Ryszard	Kámen, jinovatka			17	0,15	4,44
	Historie města Brna			15	0,19	10,83
Pešková, Vlastimila	Biologie člověka			14	0,13	10,81
Fischerová, Daniela	Hodina mezi psem a			12	0,13	4,29
Rázusová-Martáková, I	Zatoulané house	sk	170	1,11	0,19	6,54



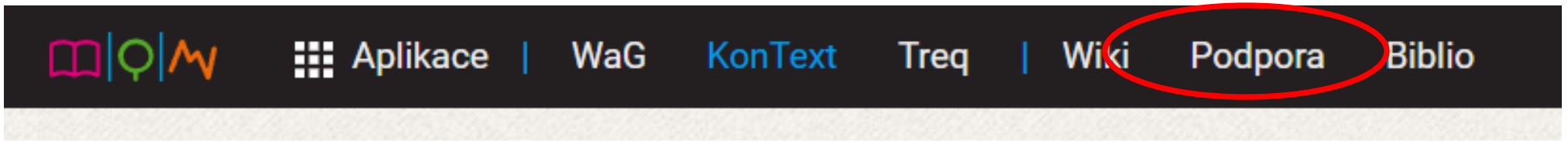

PC1 vs. PC2 Log Transformed
PC2 (0.20)
PC1 (0.52)
6
4
2
0
−2
−4
−2
0
2
4
6
8
mdd
sLength

Závěry a výhledy

Nové možnosti ve výzkumu

- kontrastivním
- typologickém
- translatologickém
- lingvoliterárním
- didaktika jazyka
- registry/textové typy
- simplifikace (dětská literatura?)
- autorské styly
- …. ?


remercie
root
VERB
je
nsubj
PRON
vous
iobj
PRON
de
case
ADP
votre
det
DET
attention
oblig:arg
NOUN
.
punct
PUNCT



https://podpora.korpus.cz/projects/poradna

Shrnutí a doporučení

Před výzkumem v IC-UD

1. definice deprel v popisu UD (https://universaldependencies.org/u/dep/index.html) + ověření pod-typů (acl:relcl atd.)
2. projít data – chybovost, false negatives/positives
3. don’t panic :-)

Co si přečíst?

a) základní přehled:

dokumentace (wiki) k UD v InterCorpu (https://wiki.korpus.cz/doku.php/pojmy:ud)

a) popis Universal Dependencies (hlavně `deprel` – https://universaldependencies.org/u/dep/index.html)

Zdroje

Tutoriál ke všem korpusům ČNK:

- https://www.youtube.com/watch?v=EOuUdU-p8VQ&t=4112s
- https://wiki.korpus.cz/doku.php/start

InterCorp:

- https://wiki.korpus.cz/doku.php/cnk:intercorp
- Anotace podle UD: https://wiki.korpus.cz/doku.php/pojmy:ud
- Míry syntaktické komplexity a lexikální diverzity: https://wiki.korpus.cz/doku.php/en:pojmy:syntakticka_komplexita
- Olga Nádvorníková, Alexandr Rosen, Martin Vavřín: InterCorp s jednotnou morfologickou a syntaktickou anotací podle Universal Dependencies: zážitky tvůrců a uživatelů. Praha, 16/11/2021. Video, pdf: zážitky tvůrců, zážitky uživatelů.
- 20. a 27. března 2024: *InterCorp a Universal Dependencies: nové možnosti výzkumu* Prezentace a video: https://shorturl.at/SS1Ho

Zdroje

Universal Dependencies :

- Oficiální popis a dokumentace: https://universaldependencies.org
- Daniel Zeman: Universal Dependencies and the Slavic Languages. Warszawa, 19.11.2018.
- Lindat UD Corpora (online search): https://lindat.mff.cuni.cz/services/kontext/corpora/corplist
- Lindat UDPipe: https://lindat.mff.cuni.cz/services/udpipe/

Álvarez González, A., Zarina Estrada Fernández and a Claudine Chamoreau (2019). *Diverse scenarios of syntactic complexity*. Amsterdam: John Benjamins Publishing Company.

Arnold J., Wasow T., Losongco A. and Ginstrom R. (2000). Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. Language, vol. (17/1): 28-55.

Beaman K. (1984). Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse. In Tannen D. and Freedle R. (Eds), Coherence in Spoken and Written Discourse: 45-80.

BERREWAERTS Joëlle, DEMANET Laurence, SCHELSTRAETE Marie-Anne *et al.*, « Chapitre IV. Les explications des changements liés au vieillissement dans l'utilisation du langage », dans : Pierre Feyereisen éd., *Parler et communiquer chez la personne âgée. Psychologie du vieillissement cognitif.* Paris cedex 14, Presses Universitaires de France, « Psychologie et sciences de la pensée », 2002, p. 169-218. DOI : 10.3917/puf.feyer.2002.01.0169. URL : https://www.cairn.info/--9782130527558-page-169.htm

Biber, D. and Bethany Gray. Grammatical complexity in academic English. Linguistic change in writing. *ICAME Journal*. 41(1), 215-219. ISSN 1502-5462. Dostupné z: doi:10.1515/icame-2017-0009

Canavese, P. and L. Mori (2021). Testing the hypothesis of "translation as a catalyst for plain legislation" on the syntactic level: A comparison of different varieties of legislative Italian. In: Castagnoli, S., S. Bernardini, A. Ferraresi, M. Miličević Petrović (eds) 2021. Using Corpora in Contrastive and Translation Studies Conference (6th Edition). Bertinoro (Italy), 9-11 September 2021.

Čermák, Petr et al. (2020). Complex Words, Causatives, Verbal Periphrases and the Gerund: Romance Languages Versus Czech (A Parallel Corpus-Based Study). Praha: Karolinum.

Chunxiao Yan. Complexité syntaxique et flux de dépendance : études quantitatives dans les treebanks universal dependencies. Linguistique. Université de Nanterre - Paris X, 2021. Français. ffNNT : 2021PA100127ff. fftel-03649621f

Cosme, Ch. (2006). Clause combining across languages. A corpus-based study of English-French translation shifts. *Languages in Contrast* 6(1), 71-108.

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.

Cvrček, V. et al. (2020). *Registry v češtině*. Praha: NLN, 2020. De Clercq, B. (2016) Le développement de la complexité syntaxique en français langue seconde : complexité structurelle et diversité. SHS Web of Conferences (27) 07006 (2016). DOI: 10.1051/shsconf/20162707006

Dell'Orletta F., Montemagni S., Venturi G. "*READ-IT: assessing readability of Italian texts with a view to text simplification".* In: SLPAT '11 – SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (Edimburgo, UK, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

Ebeling Oksefjell, S., Ebeling, J. (2020). Dialogue vs. narrative in fiction: A cross-linguistic comparison. *Languages in Contrast* 20(2), 2020, pp. 288-313.

Fabricius-Hansen, Cathrine. 1996. “Informational Density: A Problem for Translation and Translation Theory.” Linguistics 34: 521–65.

Fabricius-Hansen, C. (1999). Information packaging and translation: aspects of translational sentence splitting (German–English/Norwegian). In Monika Doherty (ed.), *Sprachspezifische Aspekte der Informationsverteilung*. 175–214. Berlin: Akademie Verlag.

Ferreira F. (1991). Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances. Journal of Memory and Language, vol. (30/2): 2110-2233.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Givón T. (1991). Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. Studies in Language, vol. (15/2): 335-370.

Bruno Guillaume, Marie-Catherine de Marneffe, Guy Perrier. Conversion et améliorations de corpus du français annotés en Universal Dependencies. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), 2019, 60 (2), pp.71-95. ffhal-02267418f Hunt, K. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3. Champaign, IL, USA: NCTE.

Chlumská, L. (2017). *Překladová čeština a její charakteristiky*. Praha: Nakladatelství Lidové noviny.

Jagaiah, T., Olinghouse, N.G. & Kearns, D.M. (2020). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. *Read Writ* 33, 2577–2638 (2020). https://doi.org/10.1007/s11145-020-10057-x Johansson, S. 2007. Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies. Amsterdam: John Benjamins.

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.

Kuboň, V. (2001). A Method for Analyzing Clause Complexity. Prague Bulletin of Mathematical Linguistics, vol. (75): 5-28

Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies, *Linguistic Typology*, vol. 23, no. 3, 2019, pp. 533-572. https://doi.org/10.1515/lingty-2019-0025

Johansson, S. 2007. Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies. Amsterdam: John Benjamins.

Marneffe, M.-C. de ; Christopher Manning, Joakim Nivre, Daniel Zeman (2021). Universal Dependencies. In: *Computational Linguistics*, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.

Mačutek, J., Čech, R., and Courtin, M. (2021). The Menzerath-Altmann law in syntactic structure revisited. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 65–73, Sofia, Bulgaria. Association for Computational Linguistics.

Mačutek, J.,, R. Čech, and J. Milička. 2019. Length of non-projective sentences: A pilot study using a Czech UD treebank. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 110–117, Paris, France. Association for Computational Linguistics.

Mondorf, B. (2003). Support for More-Support. In Rohdenburg G. and Mondorf B. (Eds), Determinants of Grammatical Variation in English: 251-304.

NÁDVORNÍKOVÁ, Olga a Jovanka ŠOTOLOVÁ, 2016. Za hranice věty: analýza změn v segmentaci na věty v překladových textech na základě francouzsko-českého paralelního korpusu. In: Jazykové paralely. Praha: NLN, s. 188–235.

Nádvorníková, O. (2017). Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus. In: *Language Use and Linguistic Structure*. Olomouc: Palacký University Olomouc, s. 445–461. http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016-proceedings.pdf

Nádvorníková, O. (2020). The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: a question of methodology. Linguistica Pragensia. 30(2), 30-50. ISSN 1805-9635. Dostupné z: doi:10.14712/18059635.2020.1.2

Nádvorníková, O. (2021). Contexts and Consequences of Sentence Splitting in Translation (English-French-Czech). *Research in Language 19(3),* pp. 229-250. https://czasopisma.uni.lodz.pl/research/issue/view/1045

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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.

Rescher, N. (1998). Complexity: A Philosophical Overview. New Brunswick NJ: Transaction Rohdenburg G. (1996). Cognitive complexity and increased grammatical explicitness in English. Cognitive Linguistics, vol. (7): 149-182.

Schleppegrell M. (1992). Subordination and Linguistic Complexity. Discourse Processes: A Multidisciplinary Journal, vol. (15/1): 117-131.

Solfjeld, Kåre. (1996). Sententiality and translation strategies German-Norwegian. *Linguistics* 34. 567–590. Straka, Milan (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics. https://aclanthology.org/K18-2020.pdf

Szmrecsanyi, B. (2004). On operationalizing syntactic complexity. In Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis Louvain-la-Neuve, March 10–12, 2004, Vol. 2, Gérard Purnelle, Cédrick Fairon & Anne Dister (eds), 1032–1039. Louvain-la-Neuve: Presses Universitaires de Louvain.

Yan, H. and Li, Y. (2019). Beyond length: Investigating dependency distance across L2 modalities and proficiency levels. *Open Linguistics*, 5(1):601–614.

Wasow T. (1997). Remarks on grammatical weight. Language Variation and Change, vol. (9): 81-105.

Zeman, Daniel (2018): The World of Tokens, Tags and Trees. Praha: ÚFAL. ISBN 978-80-88132-09-7.

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, et al. (2020). Universal Dependencies 2.6, LINDAT/ CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available at: http://hdl.handle.net/11234/1- 3226. See also http://universaldependencies.org. 0

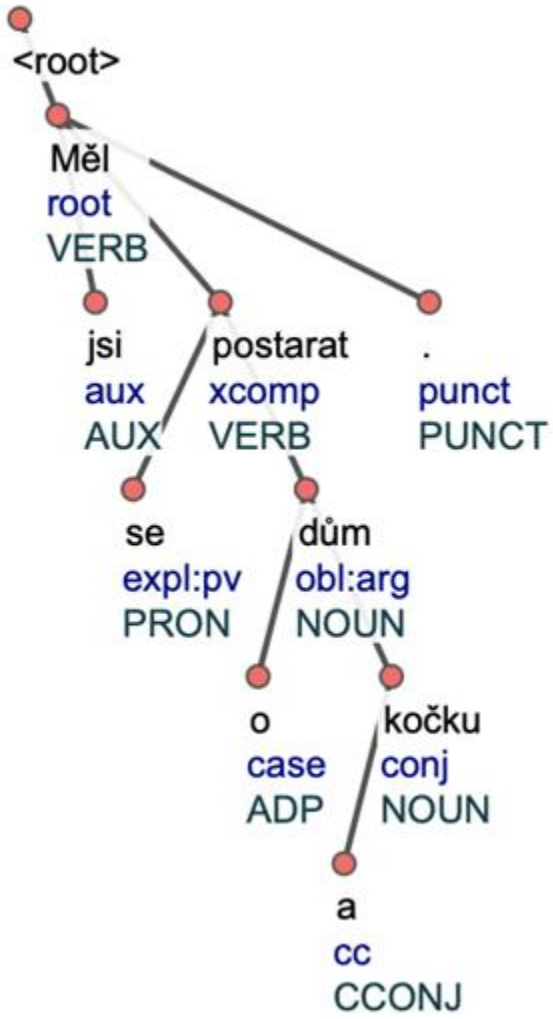
Otázky?
Diskuse!

Proč nepoužít rovnou CONLL-U? (1/2)

- Dvojí tokenizace u agregátů jako grafických a syntaktických slov
  - řešení: grafická slova jako tokeny, syntaktická jako multihodnoty

Měl ses postarat o dům a kočku.

word	*ses*
sword	se|jsi
iword	se|s
lemma	se|být
upos	PRON|AUX
xpos	P7--4----------|VB-S---2P-AA--
feats	Case=Acc|Number=Sing|PronType=Prs|Reflex=Yes|Variant=Short||Mood=Ind|Number=Sing|Person=2|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act
deprel	expl:pv|aux



Dostupnost jazykových verzí textů v jádru (*core*) korpusu InterCorp

32 jazykových verzí

ID	jazyky
SaintExupery-Malyprinc	cs be bg ca da de ds el en es fi fr hi hr hs hu it ja la lv mk nl no pl pt ro ru sk sl sv sy uk
rowlingova-hpot_kamen	cs be bg ca da de el en es fi fr hr hu it ja lv mk nl no pl pt ro ru sk sl sv sy uk
Orwell-1984	cs be bg ca da de en es fi fr hr hu it ja lt lv mk nl no pl pt ro ru sk sl sr uk
adams-stoparuv_pruvodc	cs be bg ca da de en es fi fr hr hs hu it ja lv mk nl no pl pt ru sk sl sr sv uk
carroll-alenka_v_kraji	cs be bg da de el en es fi fr hr hu it la lv mk nl no pl pt ro ru sk sl sv sy uk
Kundera-Nesnesit_lehko	cs bg ca da de en es fi fr hr hu it lt lv mk nl no pl pt ru sl sv sy uk
Kafka-Proces	cs be bg ca da de en es fr hr hu it lv mk nl no pl pt ro ru sl sv sy uk
Eco-Jmeno_ruze	cs bg ca da de en es fi fr hr hu it lv mk nl no pl pt ru sl sr sv sy uk
Frankova-DenikA_Franko	cs bg ca da de el en es fi fr hr hu it mk nl no pl pt ru sk sl sr sv uk
brown-sifra	cs be bg ca da de en es fi fr hr it ja mk nl no pl pt ru sk sl sr sv uk
Tolkien-Hobit	cs be bg ca da de en es fi hr it lv mk nl no pl pt ru sk sl sr sv uk
Bulgakov-MistrAMarketk	cs be bg da de en es fr hr hu it lv mk nl pl pt ru sk sl sr sv sy uk