Exploring InterCorp v16ud: the potential of a multilingual parallel treebank with complexity and diversity metrics

Alexandr Rosen
Ústav Českého národního korpusu
Filozofická fakulta Univerzity Karlovy, Praha

Instytut Slawistyki Zachodniej i Południowej
Uniwersytet Warszawski

10 July 2024

Outline

1. About InterCorp
2. Universal Dependencies (UD)
3. InterCorp with UD
4. Metrics of syntactic complexity and lexical diversity
5. Using the metrics
6. Perspectives, questions, discussion

Link to this presentation:

https://shorturl.at/fTJE3

Outline

1. About InterCorp
2. Universal Dependencies (UD)
3. InterCorp with UD
4. Metrics of syntactic complexity and lexical diversity
5. Using the metrics
6. Perspectives, questions, discussion

InterCorp – a multilingual parallel corpus

- Part of the *Czech National Corpus*
- Every text in Czech and at least one other language
- 2008: v0 (first online release)
- 2023: v16
- 62 languages
- 5.4 billion words
- March 2024: v16ud – pilot (Core only)
- June 2024: v16ud – full (all texts)

Access to:

➢ InterCorp **v16**

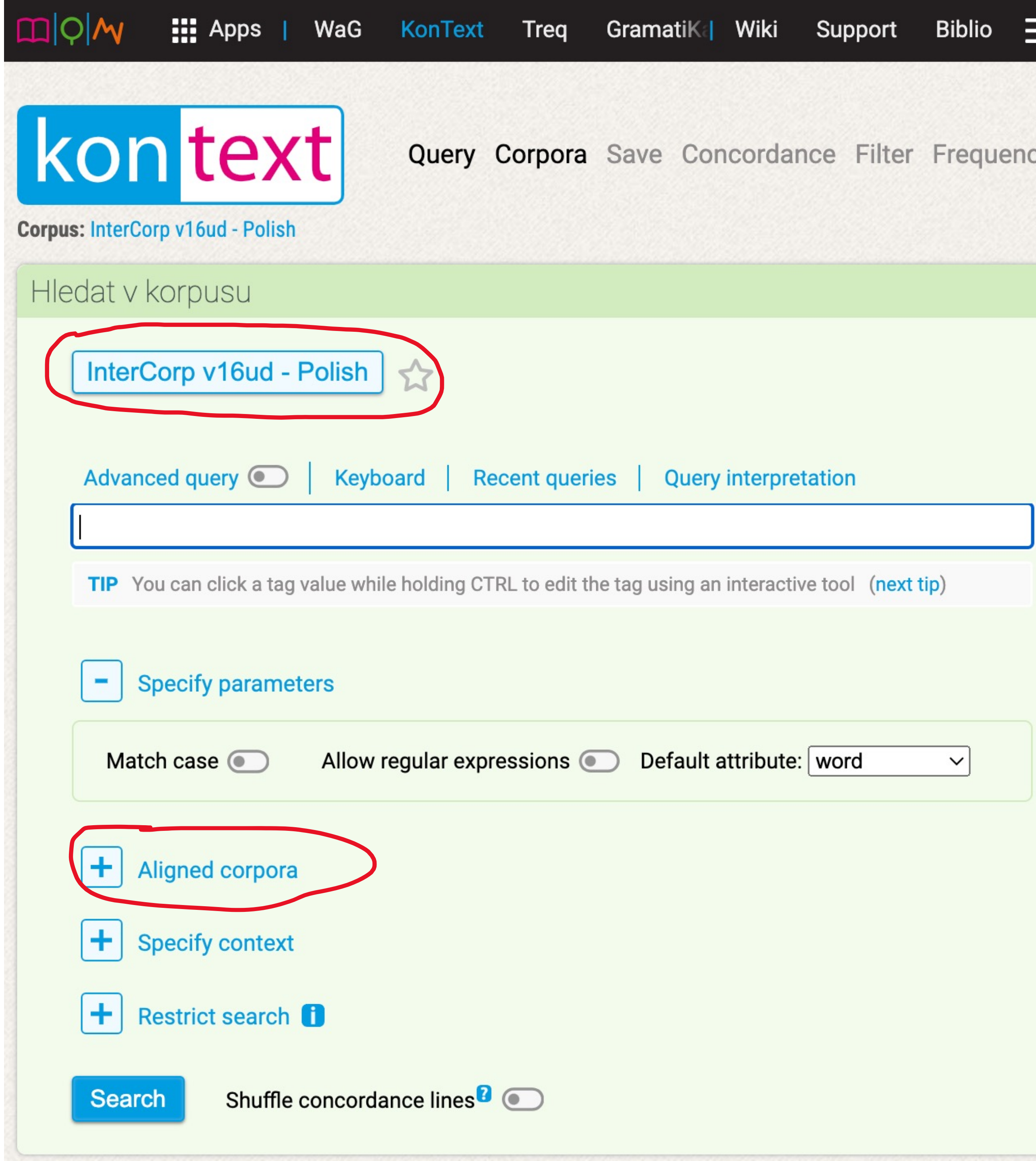
without login OR
with institutional login (Shibboleth)

1. Go to: [korpus.cz]()
2. Click on: **KonText**
3. Click on: **syn2020 > All corpora**
4. Select/Type in: **InterCorp v16 - Polish**

➢ InterCorp **v16ud pilot**

with temporary login

1. Go to: [korpus.cz]()
2. Login: **ud16test**
3. Password: **ud16test**
4. Click on: **KonText**


Apps | WaG KonText Treq GramatiKa | Wiki Support Biblio
kon text
Query Corpora Save Concordance Filter Frequenc
Hledat v korpusu
InterCorp v16ud - Polish
Advanced query | Keyboard | Recent queries | Query interpretation
- Specify parameters
+ Aligned corpora
+ Specify context
+ Restrict search
Search

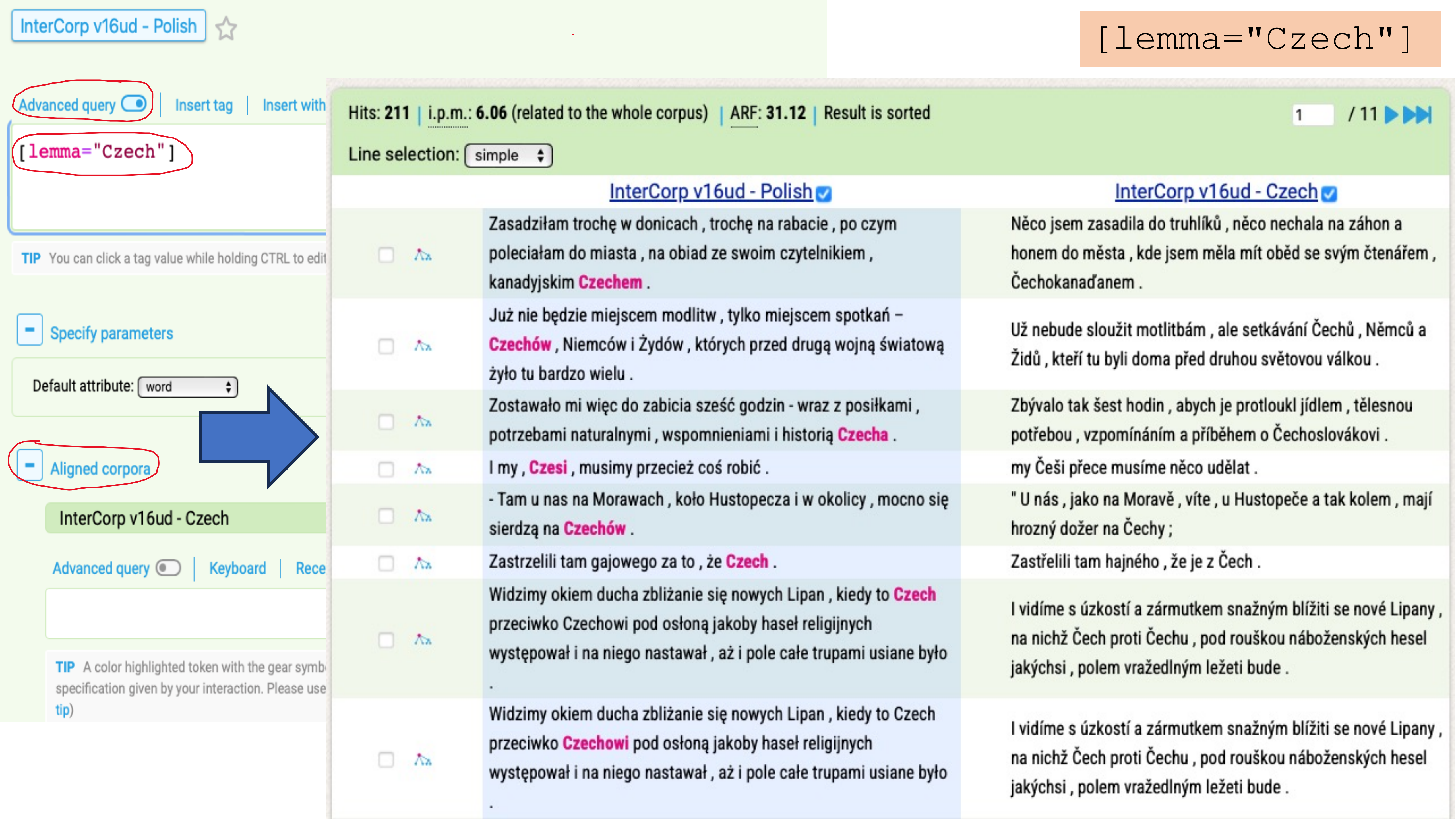
Corpus: InterCorp v16ud - Polish

TIP You can click a tag value while holding CTRL to edit the tag using an interactive tool (next tip)

Match case ○ Allow regular expressions ○ Default attribute: word

Shuffle concordance lines ○

[lemma="Czech"]



InterCorp v16ud - Polish ☆

Advanced query | Insert tag | Insert with

```
[lemma="Czech"]
```

TIP You can click a tag value while holding CTRL to edit

− Specify parameters

Default attribute: word

− Aligned corpora

InterCorp v16ud - Czech

Advanced query | Keyboard | Rece

TIP A color highlighted token with the gear symb specification given by your interaction. Please use tip)

Hits: 211 | i.p.m.: 6.06 (related to the whole corpus) | ARF: 31.12 | Result is sorted   1 / 11

Line selection: simple

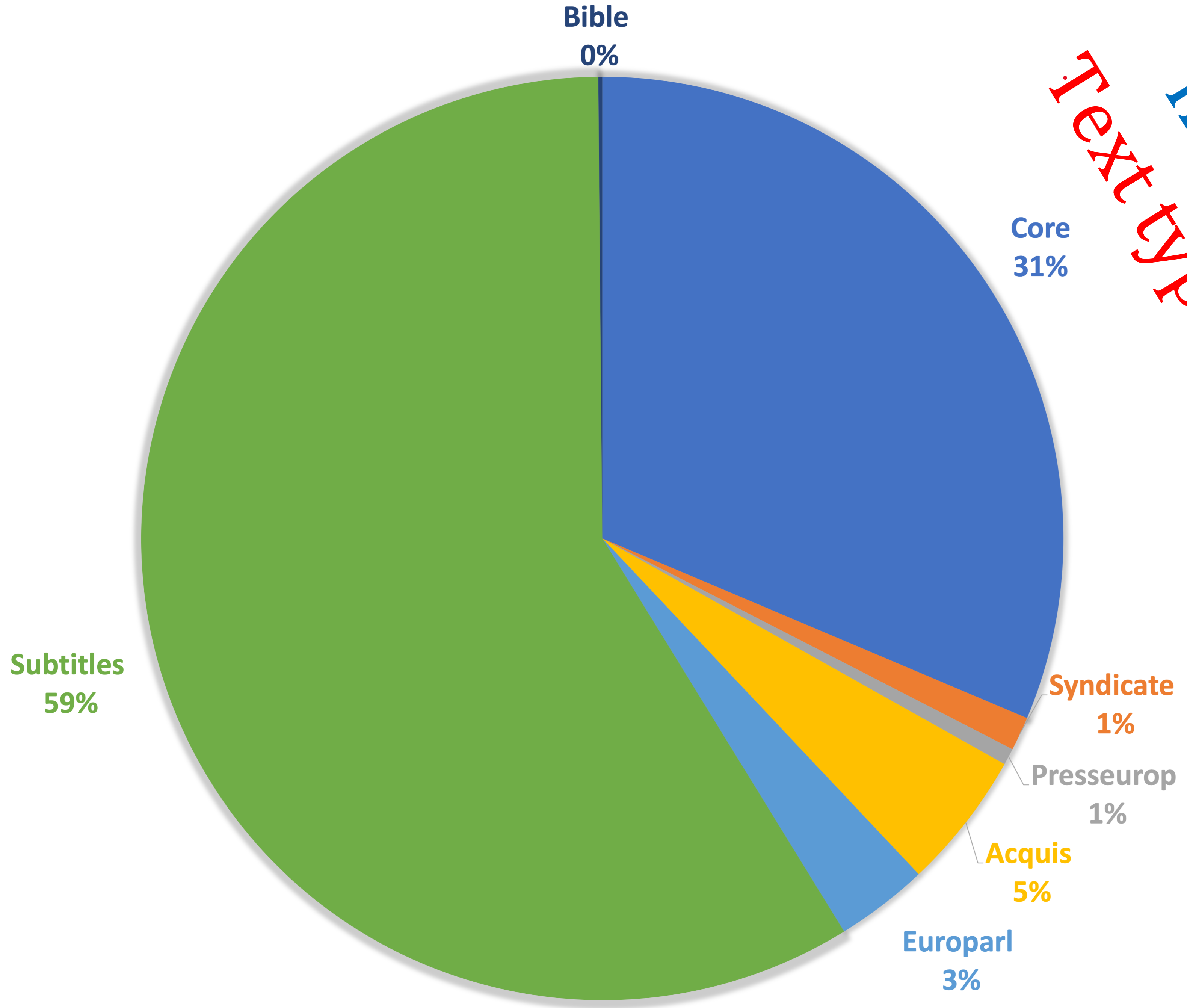
		InterCorp v16ud - Polish ☑	InterCorp v16ud - Czech ☑
☐		Zasadziłam trochę w donicach , trochę na rabacie , po czym poleciałam do miasta , na obiad ze swoim czytelnikiem , kanadyjskim **Czechem** .	Něco jsem zasadila do truhlíků , něco nechala na záhon a honem do města , kde jsem měla mít oběd se svým čtenářem , Čechokanaďanem .
☐		Już nie będzie miejscem modlitw , tylko miejscem spotkań – **Czechów** , Niemców i Żydów , których przed drugą wojną światową żyło tu bardzo wielu .	Už nebude sloužit motlitbám , ale setkávání Čechů , Němců a Židů , kteří tu byli doma před druhou světovou válkou .
☐		Zostawało mi więc do zabicia sześć godzin - wraz z posiłkami , potrzebami naturalnymi , wspomnieniami i historią **Czecha** .	Zbývalo tak šest hodin , abych je protloukl jídlem , tělesnou potřebou , vzpomínáním a příběhem o Čechoslovákovi .
☐		I my , **Czesi** , musimy przecież coś robić .	my Češi přece musíme něco udělat .
☐		- Tam u nas na Morawach , koło Hustopecza i w okolicy , mocno się sierdzą na **Czechów** .	" U nás , jako na Moravě , víte , u Hustopeče a tak kolem , mají hrozný dožer na Čechy ;
☐		Zastrzelili tam gajowego za to , że **Czech** .	Zastřelili tam hajného , že je z Čech .
☐		Widzimy okiem ducha zbliżanie się nowych Lipan , kiedy to **Czech** przeciwko Czechowi pod osłoną jakoby haseł religijnych występował i na niego nastawał , aż i pole całe trupami usiane było .	I vidíme s úzkostí a zármutkem snažným blížiti se nové Lipany , na nichž Čech proti Čechu , pod rouškou náboženských hesel jakýchsi , polem vražedlným ležeti bude .
☐		Widzimy okiem ducha zbliżanie się nowych Lipan , kiedy to Czech przeciwko **Czechowi** pod osłoną jakoby haseł religijnych występował i na niego nastawał , aż i pole całe trupami usiane było .	I vidíme s úzkostí a zármutkem snažným blížiti se nové Lipany , na nichž Čech proti Čechu , pod rouškou náboženských hesel jakýchsi , polem vražedlným ležeti bude .

More info:

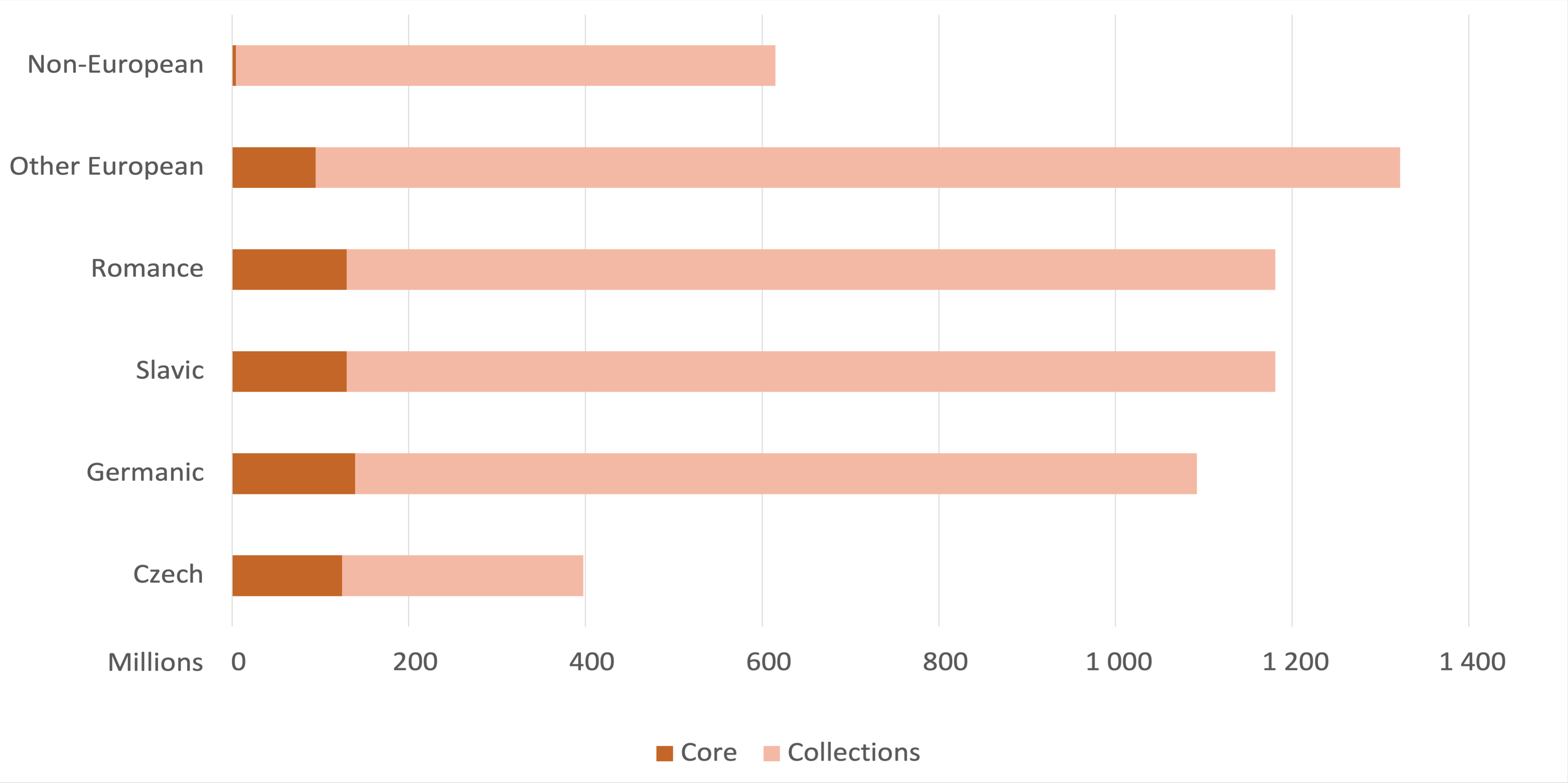
- All about InterCorp: https://wiki.korpus.cz/doku.php/en:cnk:intercorp
- On searching InterCorp: https://wiki.korpus.cz/doku.php/en:kurz:hledani_v_paralelnim_korpusu
- Tutorial for all CNC corpora (in Czech): https://wiki.korpus.cz/doku.php/start
- UD in InterCorp: https://wiki.korpus.cz/doku.php/en:pojmy:ud
- Complexity and diversity metrics in InterCorp v16ud: https://wiki.korpus.cz/doku.php/en:pojmy:syntakticka_komplexita

InterCorp v16

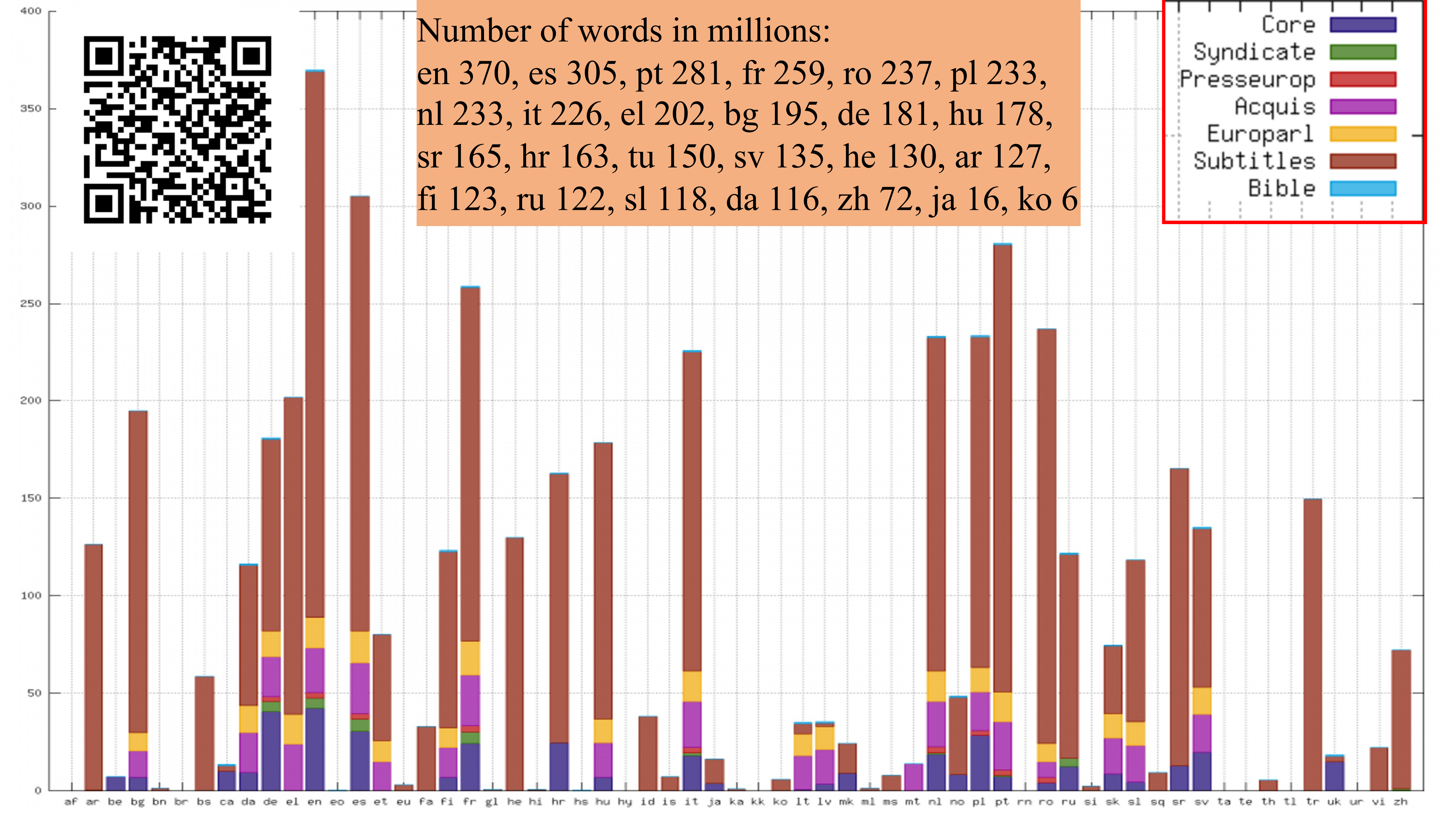
Text types (Czech)



InterCorp v16: language families



Number of words in millions:
en 370, es 305, pt 281, fr 259, ro 237, pl 233,
nl 233, it 226, el 202, bg 195, de 181, hu 178,
sr 165, hr 163, tu 150, sv 135, he 130, ar 127,
fi 123, ru 122, sl 118, da 116, zh 72, ja 16, ko 6



Red: *texts in Core*

Bold red: *10 and more texts*

InterCorp v16
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 UpperSorbian Urdu Vietnamese

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	RelDI	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

Language-specific morphosyntactic annotation:

– many tagsets

– various tokenization rules

Info on tagsets in InterCorp v16: https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze16#morphosyntactic_annotation

Outline

1. About InterCorp
2. Universal Dependencies (UD)
3. InterCorp with UD
4. Metrics of syntactic complexity and lexical diversity
5. Using the metrics
6. Perspectives, questions, discussion

Why *Universal Dependencies*?

- The *de facto* standard for morphological and syntactic annotation
- https://universaldependencies.org

- Version 2.14, May 2024
- 161 languages, 283 treebanks for training and testing linguistic models

- *UDPipe 2.12* – a tool with models for 71 languages
    - *InterCorp* v16ud: models for 47 out of 62 languages
    - If treebanks large enough, annotation fairly reliable

- Active community of UD experts and users

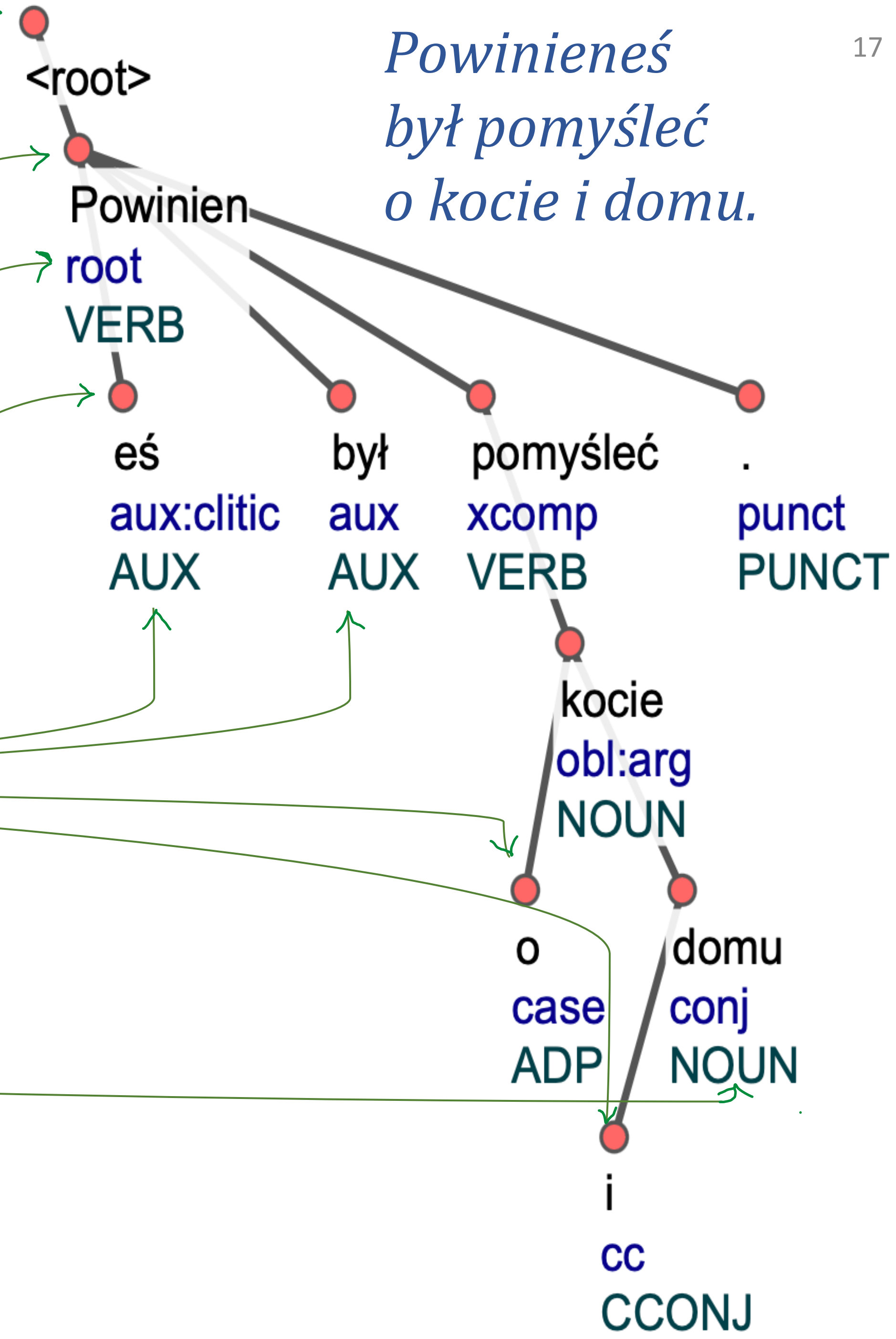
Universal Dependencies – principles

- Language-independent definition of linguistic categories
- Compromise between several requirements
- The annotation should be (https://en.wikipedia.org/wiki/Manning%27s_Law):
    - **satisfactory** for each language
    - **consistent** across languages
    - easy for **annotators**
    - easy for **non-lingusts**
    - easy for **parsers**
    - supportive to **downstream** tasks

UD – syntactic structure

- Single level (surface syntax)
- Every sentence as a dependency tree
- Every word has its node and dependency relation
- There are no empty nodes
- Multi-word tokens are split
- Function words depend on content words
- Non-initial conjuncts depend on the initial conjunct

Powinieneś
był pomyśleć
o kocie i domu.



Based on *Stanford Dependencies*

https://nlp.stanford.edu/software/stanford-dependencies.html

UD Guidelines version 2 (version 1: 2014)

- 37 syntactic functions – `deprel`

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

- 17 parts of speech – `upos`

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

- 24 morphological categories – `feats`

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

`[deprel="nsubj"]`

UD – syntactic functions (`deprel`)

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

Both finite and non-finite!

UD – other deprels

```
[deprel="conj"]
```

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*

acl:relcl for relative adnominal clauses
advcl:relcl for relative clauses whose antecedent is a clause
aux:pass for the passive auxiliary
csubj:outer for outer clausal subjects of predicates that are clauses
csubj:pass for clausal subjects of passive clauses
expl:impers for reflexive markers of impersonal clauses
expl:pass for reflexive markers of middle or passive clauses
expl:pv for reflexive clitics with inherently reflexive verbs
nsubj:outer for outer nominal subjects of predicates that are clauses
nsubj:pass for nominal subjects of passive clauses
obl:agent for demoted agents in passive clauses

Subtypes of deprels:

- Optional, language-specific

  Some semi-mandatory
- To search for any deprel no matter the subtype:

```
[deprel="nsubj.*"]
```

UD – *universal* parts of speech (upos)

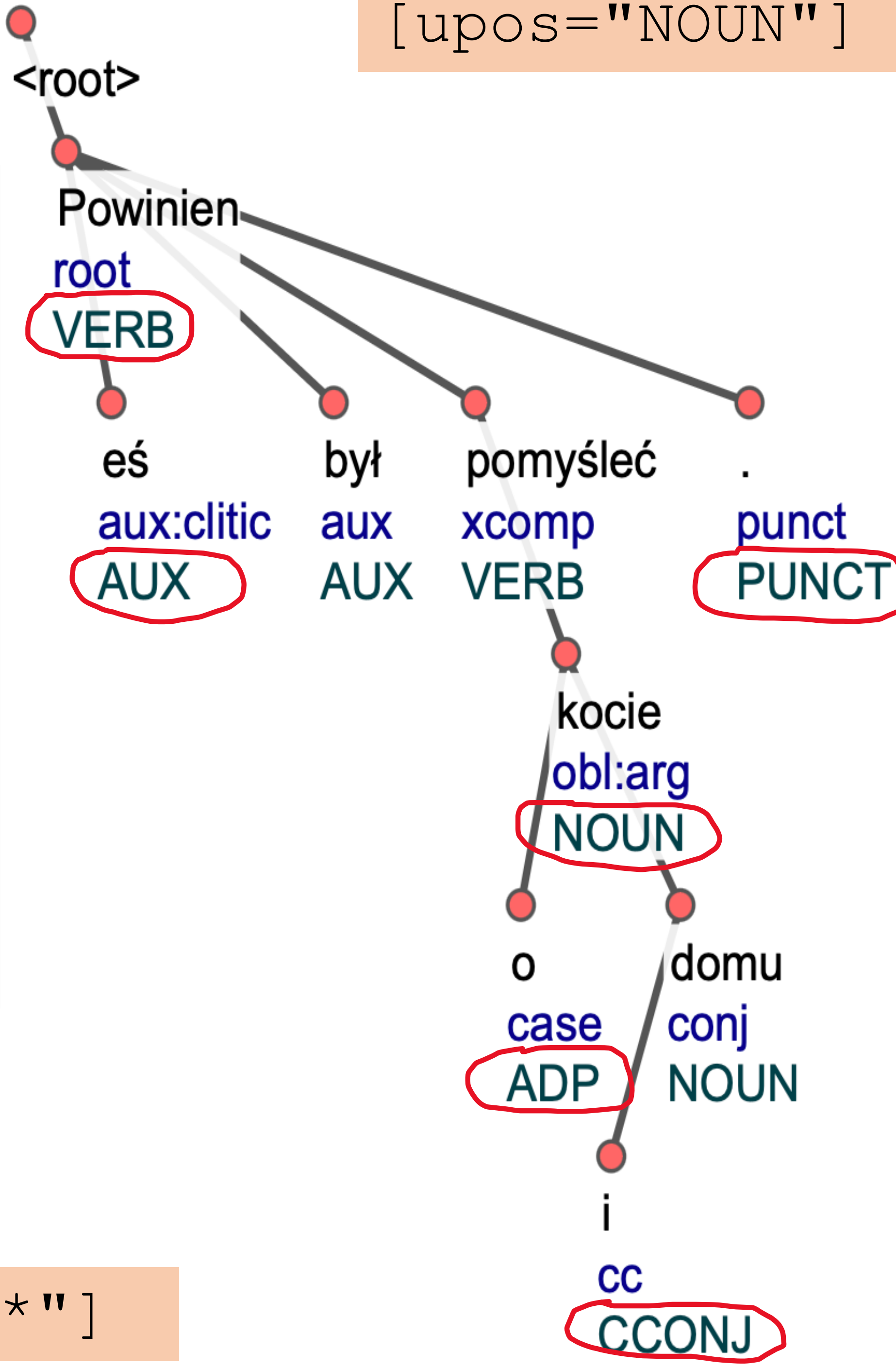
Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

= 17 word classes, based on 12 word classes in *Google Universal Tagset* 2007 https://github.com/slavpetrov/universal-pos-tags

In most treebanks also “legacy” language-specific tag (POS+categories):
xpos, e.g. subst:sg:loc:m3 for Polish NOUN

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

```
[upos="NOUN"]
```



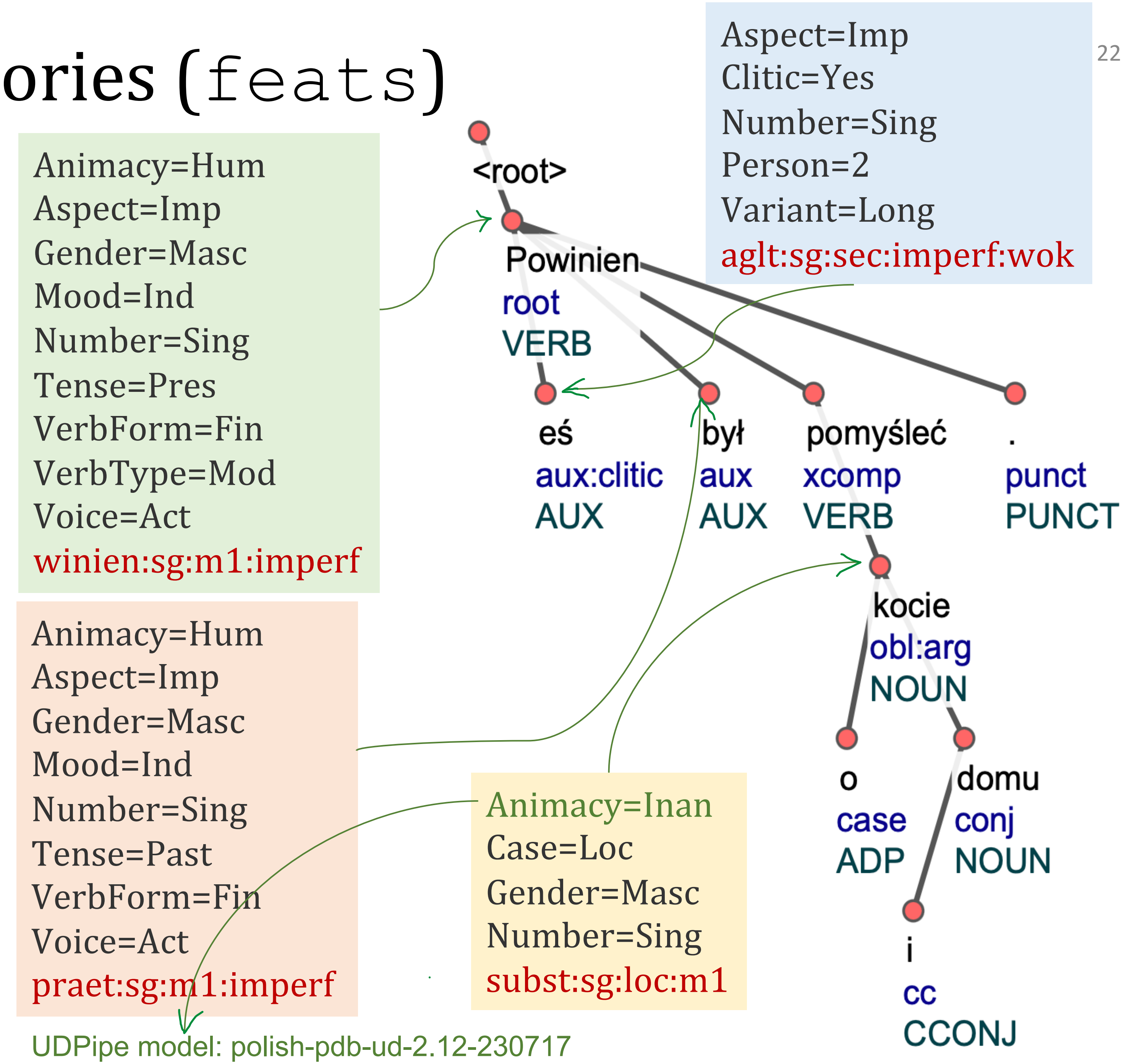
UD – morphological categories (feats)

[feats="Gender=Masc"]

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

= 24 categories, based on *Interset* 2006, used in CONLL shared tasks

https://github.com/dan-zeman/interset



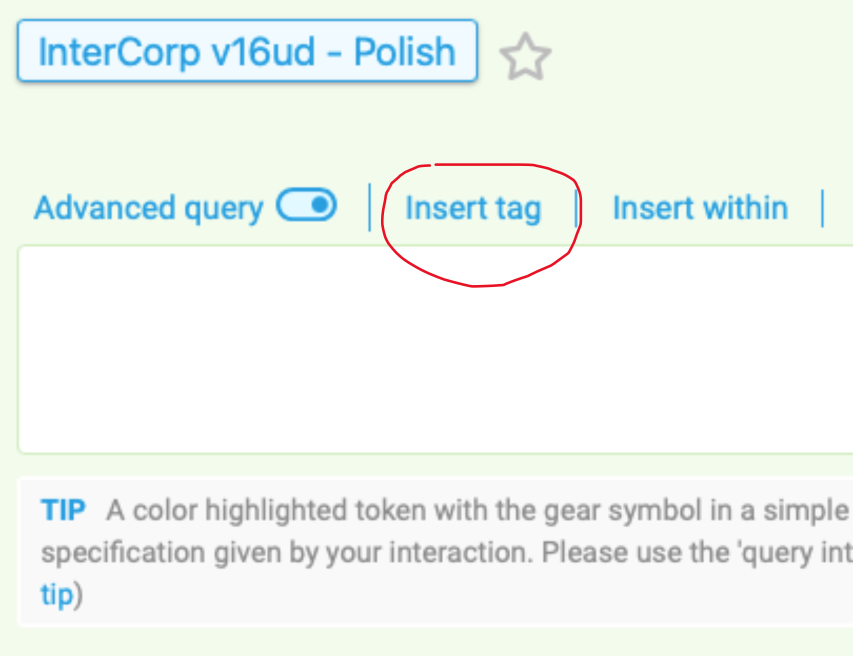
Animacy=Hum
Aspect=Imp
Gender=Masc
Mood=Ind
Number=Sing
Tense=Pres
VerbForm=Fin
VerbType=Mod
Voice=Act
winien:sg:m1:imperf

Aspect=Imp
Clitic=Yes
Number=Sing
Person=2
Variant=Long
aglt:sg:sec:imperf:wok

Animacy=Hum
Aspect=Imp
Gender=Masc
Mood=Ind
Number=Sing
Tense=Past
VerbForm=Fin
Voice=Act
praet:sg:m1:imperf

Animacy=Inan
Case=Loc
Gender=Masc
Number=Sing
subst:sg:loc:m1

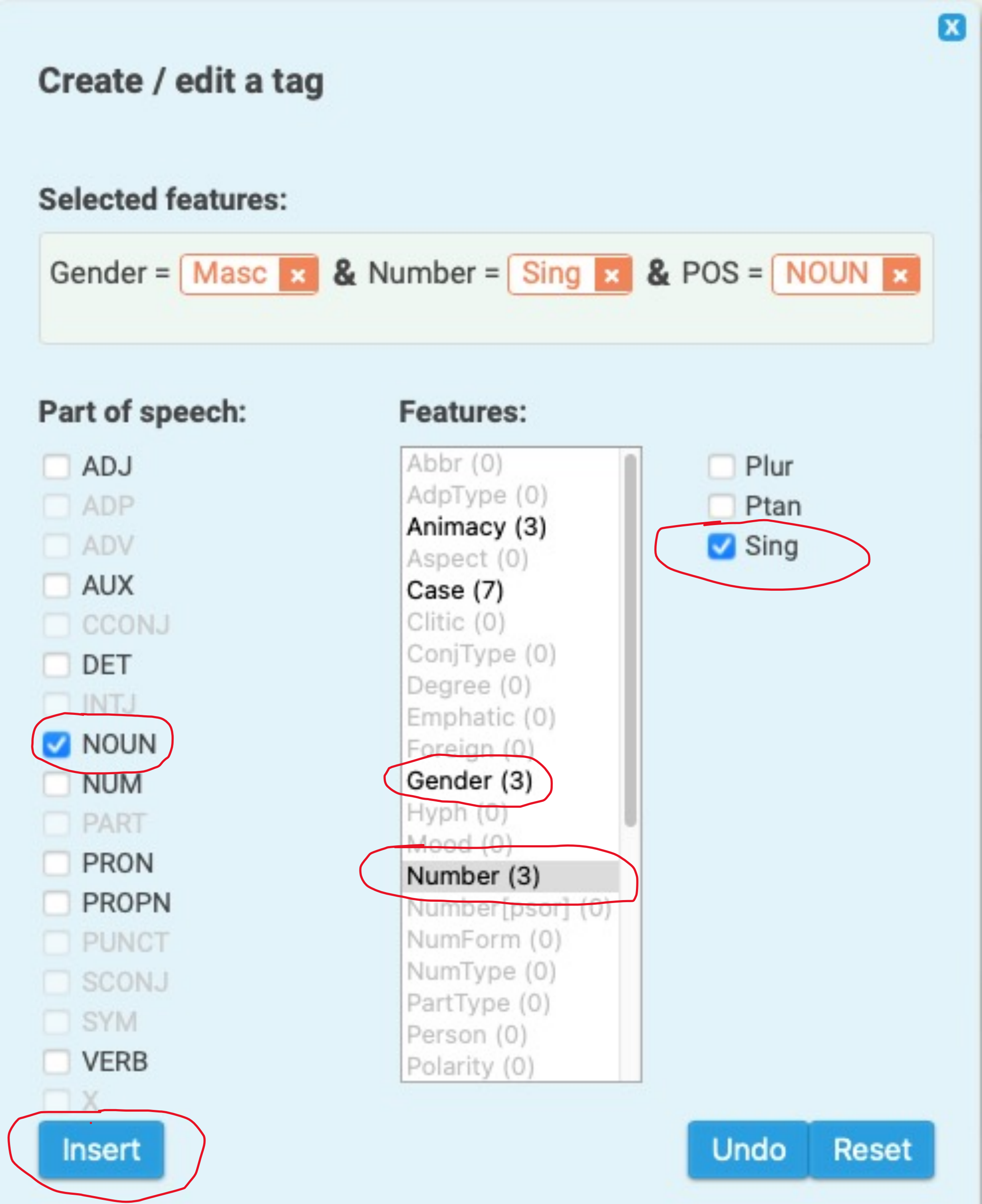
UDPipe model: polish-pdb-ud-2.12-230717
Polish-lfg-ud-2.12.230717: SubGender=Masc1



InterCorp v16ud - Polish ☆

Advanced query | Insert tag | Insert within |

TIP A color highlighted token with the gear symbol in a simple
specification given by your interaction. Please use the 'query int
tip)



Create / edit a tag

Selected features:

Gender = Masc × & Number = Sing × & POS = NOUN ×

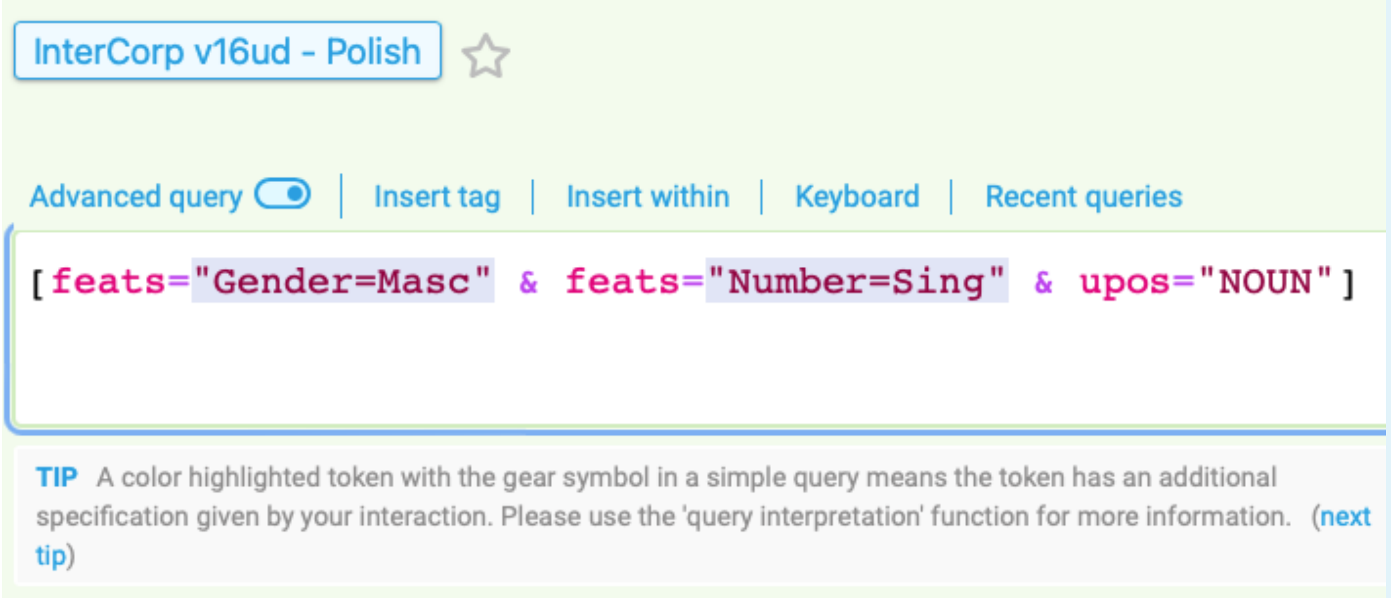
Part of speech:

- □ ADJ
- □ ADP
- □ ADV
- □ AUX
- □ CCONJ
- □ DET
- □ INTJ
- ☑ NOUN
- □ NUM
- □ PART
- □ PRON
- □ PROPN
- □ PUNCT
- □ SCONJ
- □ SYM
- □ VERB
- □ X

Features:

- Abbr (0)
- AdpType (0)
- Animacy (3)
- Aspect (0)
- Case (7)
- Clitic (0)
- ConjType (0)
- Degree (0)
- Emphatic (0)
- Foreign (0)
- Gender (3)
- Hyph (0)
- Mood (0)
- Number (3)
- Number[psor] (0)
- NumForm (0)
- NumType (0)
- PartType (0)
- Person (0)
- Polarity (0)

- □ Plur
- □ Ptan
- ☑ Sing



InterCorp v16ud - Polish ☆

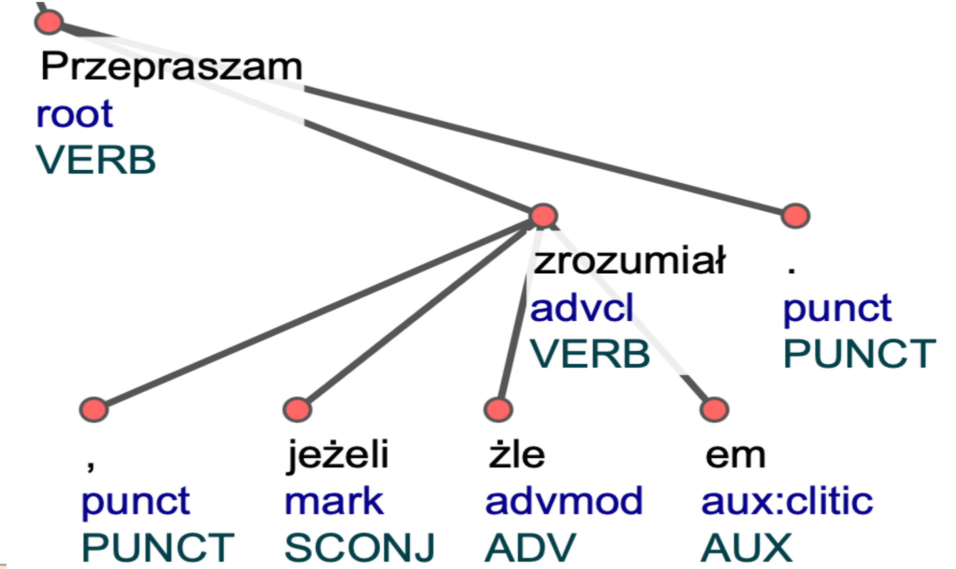
Advanced query | Insert tag | Insert within | Keyboard | Recent queries

```
[feats="Gender=Masc" & feats="Number=Sing" & upos="NOUN"]
```

TIP A color highlighted token with the gear symbol in a simple query means the token has an additional specification given by your interaction. Please use the 'query interpretation' function for more information. (next tip)

UD – tabular format (*CONLL-U*)

Based on *CONLL-X* 2007
https://web.archive.org/web/20160814191537/http://ilk.uvt.nl/conll/#dataformat



Przepraszam, jeżeli źle zrozumiałem.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL
1	*Przepraszam*	przepraszać	VERB	fin:sg:pri: imperf	Aspect=Imp|Mood=Ind|Number=Sing|Person=1| Tense=Pres|VerbForm=Fin|Voice=Act	0	root
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*	_	_	_	_	_	_
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=Period	1	punct

UD grows bottom up

- Concerns of typologists (Croft et al., 2017)
- Function words? (Osborne & Gerdes, 2019; Tuora et al., 2021)
- Core vs. non-core arguments (Przepiórkowski & Patejuk 2018)
- Coordination (Przepiórkowski & Patejuk 2019, Przepiórkowski et al. 2024)
- Some treebanks do not adhere to guidelines
- Treebanks differ in size and balance
- Success rate depends on language
  cs: 90% syntax, 97% morphology
  (Straka 2018, https://aclanthology.org/K18-2020.pdf)

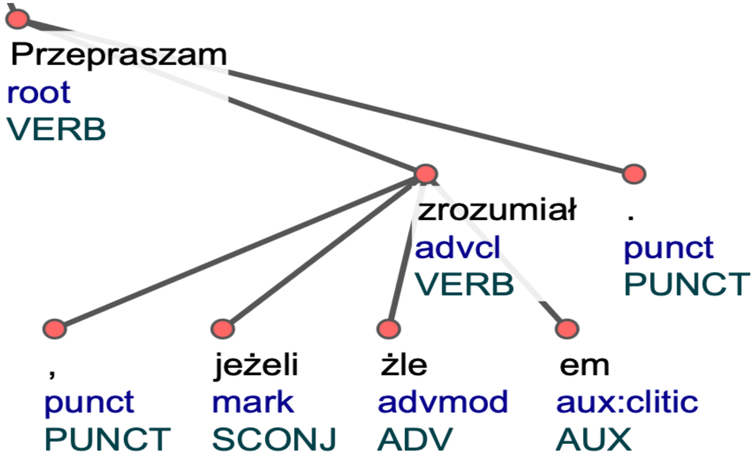
– Joakim Nivre (Uppsala) *chief cat herder*

– Dan Zeman (Praha) *every treebank must pass a test*

Outline

1. About InterCorp
2. Universal Dependencies (UD)
3. InterCorp with UD
4. Metrics of syntactic complexity and lexical diversity
5. Using the metrics
6. Perspectives, questions, discussion

UD in InterCorp – fused words



CONLL-U: two-level tokenization

InterCorp: graphical words as tokens, syntactic words as multivalues

Przepraszam, jeżeli źle zrozumiałem.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL
5-6	*zrozumiałem*	_	_	_	_	_	_
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

CONLL-U

InterCorp

id	word	sword	lemma	upos	xpos	feats	head	deprel
5|6	*zrozumiałem*	zrozumiał| em	zrozumieć| być	VERB| AUX	praet:sg:m1:perf| aglt:sg:pri:imperf:wok	Animacy=Hum|Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act|| Aspect=Imp|Clitic=Yes|Number=Sing|Person=1|Variant=Long	1|5	expl:pv| aux

UD in InterCorp – fused words

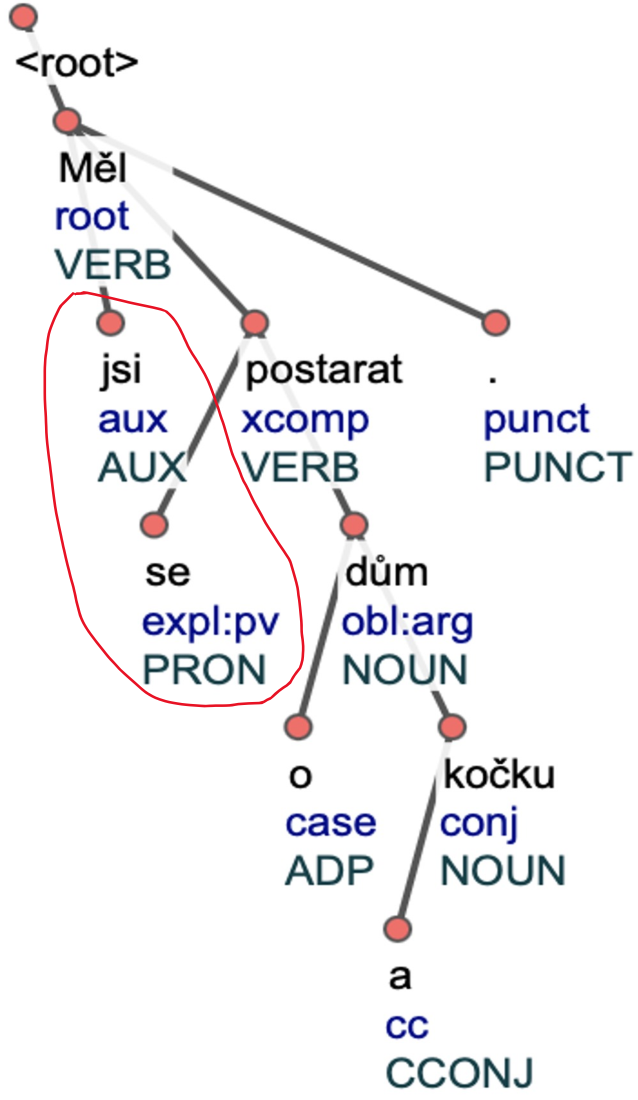
	word	sword	iword	lemma	upos
pl	*zrozumiałem*	zrozumiał|em	zrozumiał|em	zrozumieć|być	VERB|AUX
es	*hacerlo*	hacer|lo	hacer|lo	hacer|él	VERB|PRON
cs	*ses*	**se|jsi**	**se|s**	se|být	PRON|AUX
fr	*aux*	**à|les**	**au|x**	à|le	ADP|DET
de	*im*	**in|dem**	**i|m**	in|der	ADP|DET
it	*nel*	**in|il**	**ne|l**	in|il	ADP|DET
pt	*à*	**a|a**	**à|**	a|o	ADP|DET

```
[sword="em"]   [lemma="być"]   [upos="VERB"]

[word="ses"]   [sword="jsi"]   [lemma="být"]

[sword=".*\|.*"]

1:[sword=".*\|.*"] & 1.sword != 1.iword
```



Měl ses postarat o dům a kočku.

How to make easier …

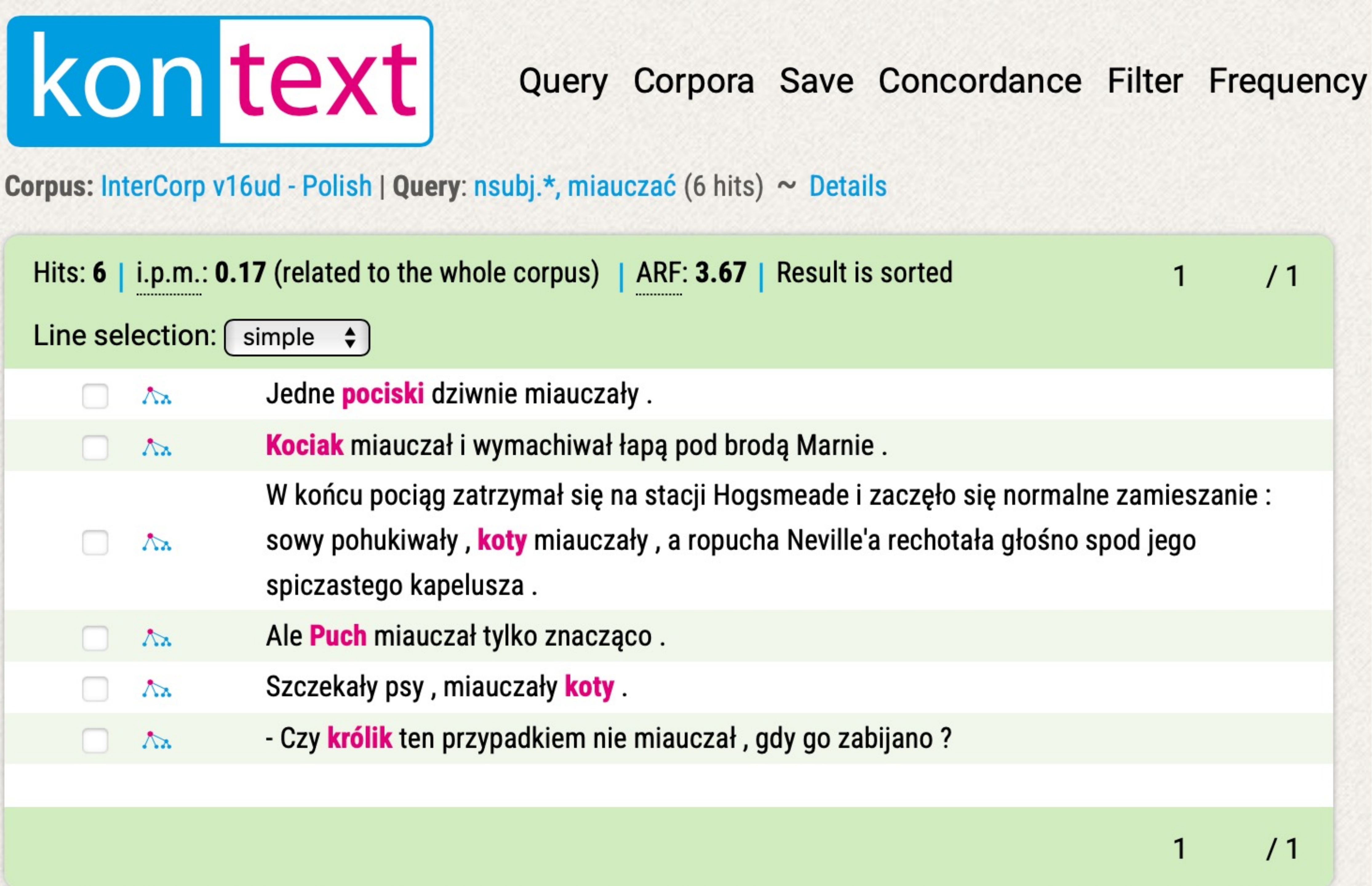
```
[deprel="nsubj.*" & p_lemma="miauczać"]
```

```
(1:[lemma="miauczać"]
2:[deprel="nsubj.*"]) |
(2:[deprel="nsubj.*"]
1:[lemma="miauczać"])
 & 1.id = 2.head within <s/>
```

🙀

- … navigating syntactic structure (`p_lemma, e_deprel`):
  - ○ lemma, upos, feats, deprel and relative position of the head
  - ○ ID, relative position and deprel of the effective head (for coordination)
- … access to info about function words (`aux_feats, case_lemma`):
  - ○ lemma, upos, feats and deprel subtype
- … queries and statistics using some common morphological categories
  - ○ some attributes from the feats list
  - ○ language-specific (20–44)

➡ new attributes in addition to those from CONLL-U



Corpus: InterCorp v16ud - Polish | **Query:** nsubj.*, miauczać (6 hits) ~ Details

Hits: **6** | i.p.m.: **0.17** (related to the whole corpus) | ARF: **3.67** | Result is sorted   1 / 1

Line selection: simple

☐ Jedne **pociski** dziwnie miauczały .

☐ **Kociak** miauczał i wymachiwał łapą pod brodą Marnie .

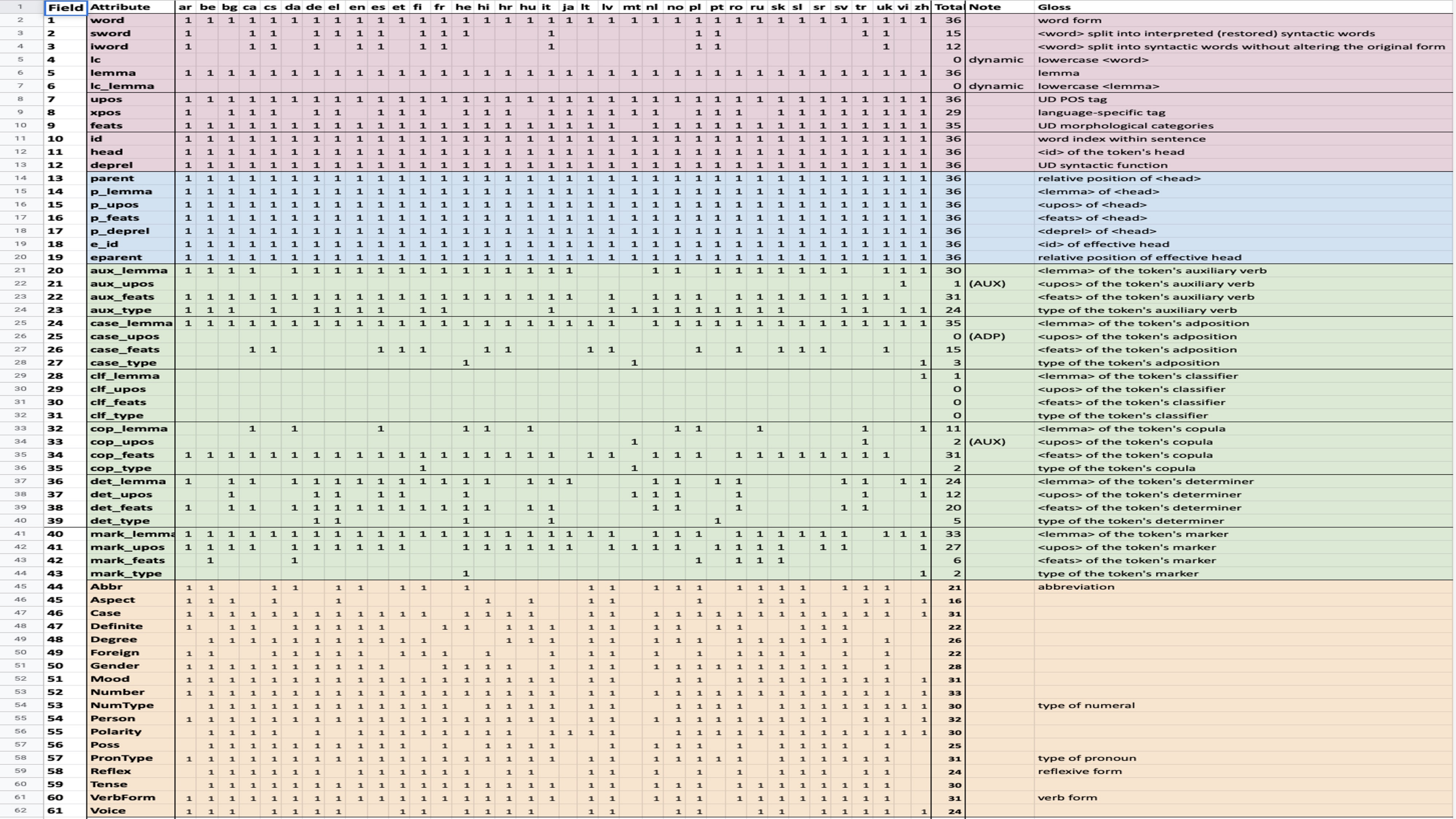
☐ W końcu pociąg zatrzymał się na stacji Hogsmeade i zaczęło się normalne zamieszanie : sowy pohukiwały , **koty** miauczały , a ropucha Neville'a rechotała głośno spod jego spiczastego kapelusza .

☐ Ale **Puch** miauczał tylko znacząco .

☐ Szczekały psy , miauczały **koty** .

☐ - Czy **królik** ten przypadkiem nie miauczał , gdy go zabijano ?

1 / 1



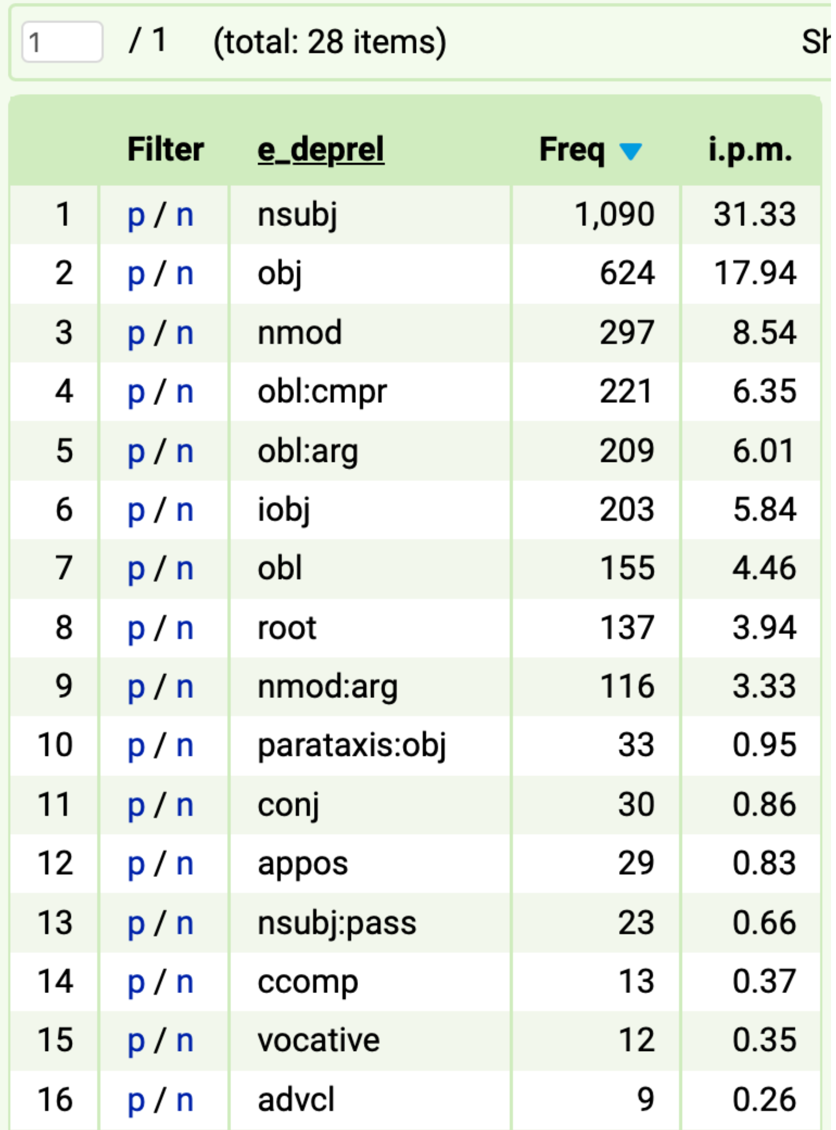
Field	Attribute	ar	be	bg	ca	cs	da	de	el	en	es	et	fi	fr	he	hi	hr	hu	it	ja	lt	lv	mt	nl	no	pl	pt	ro	ru	sk	sl	sr	sv	tr	uk	vi	zh	Total	Note	Gloss
1	word	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	36		word form
2	sword	1			1	1		1	1	1	1		1	1	1				1							1	1							1	1			15		<word> 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	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	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	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	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	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	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	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	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	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	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	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	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	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	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			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	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	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	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		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			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		

Attributes to navigate syntactic structure

Attribute	Description
parent	relative position of the head, e.g. -1, +2
p_lemma	lemma of the head
p_upos	upos of the head
p_feats	feats of the head
p_deprel	deprel of the head
e_id	ID of the effective head (for non-initial conjuncts: ID of the head of the initial conjunct)
e_deprel	deprel of the effective head
eparent	relative position of the effective head

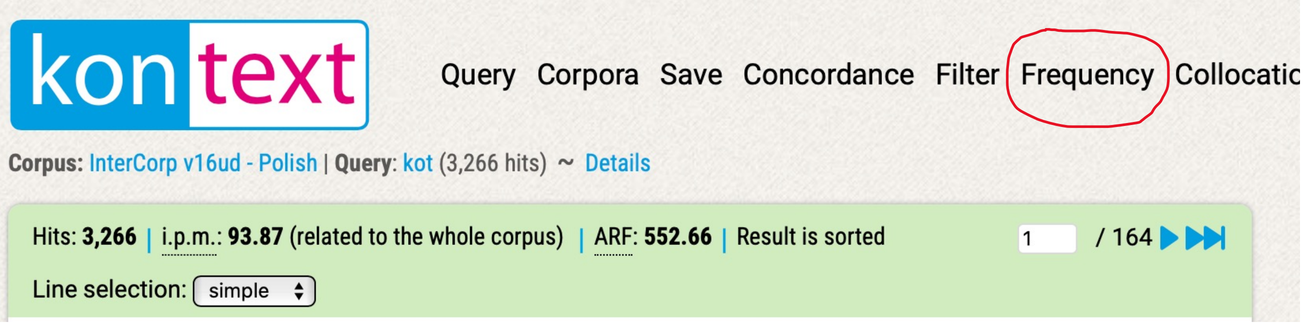
To list typical deprels of a lemma:

[lemma="kot"]
Frequency > Custom > e_deprel



1 / 1 (total: 28 items) Sh

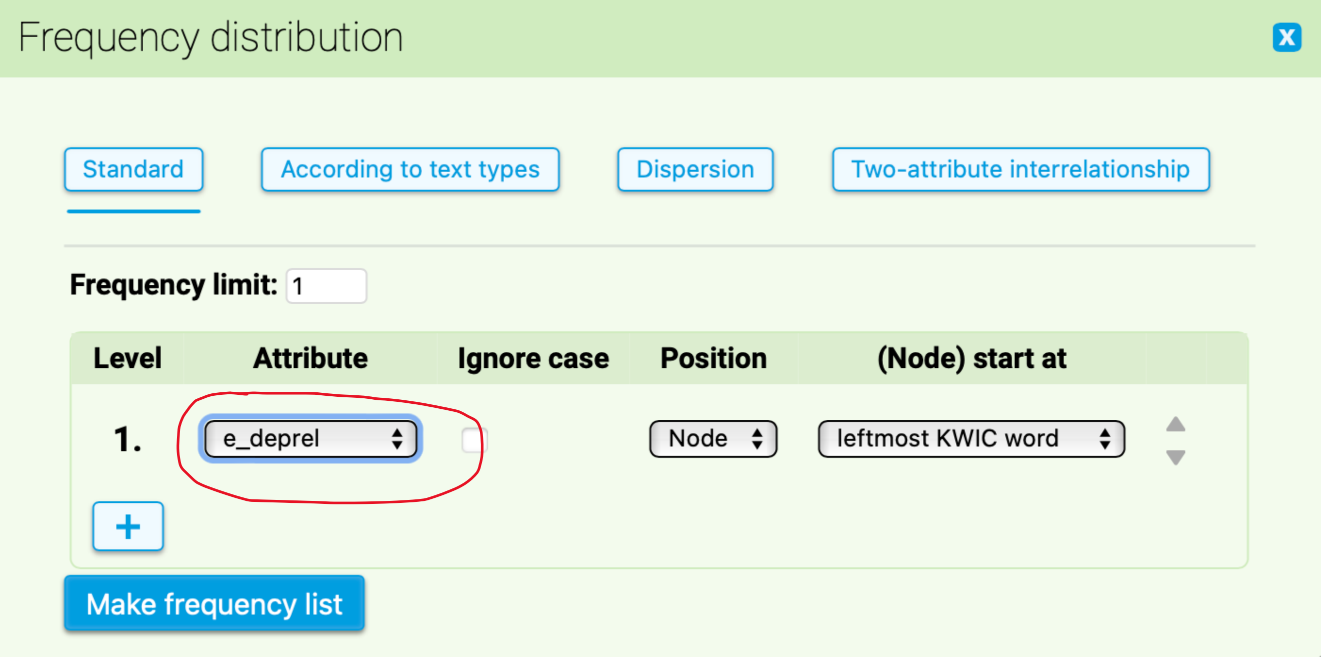
	Filter	e_deprel	Freq ▼	i.p.m.
1	p / n	nsubj	1,090	31.33
2	p / n	obj	624	17.94
3	p / n	nmod	297	8.54
4	p / n	obl:cmpr	221	6.35
5	p / n	obl:arg	209	6.01
6	p / n	iobj	203	5.84
7	p / n	obl	155	4.46
8	p / n	root	137	3.94
9	p / n	nmod:arg	116	3.33
10	p / n	parataxis:obj	33	0.95
11	p / n	conj	30	0.86
12	p / n	appos	29	0.83
13	p / n	nsubj:pass	23	0.66
14	p / n	ccomp	13	0.37
15	p / n	vocative	12	0.35
16	p / n	advcl	9	0.26



Corpus: InterCorp v16ud - Polish | Query: kot (3,266 hits) ~ Details

Hits: 3,266 | i.p.m.: 93.87 (related to the whole corpus) | ARF: 552.66 | Result is sorted 1 / 164

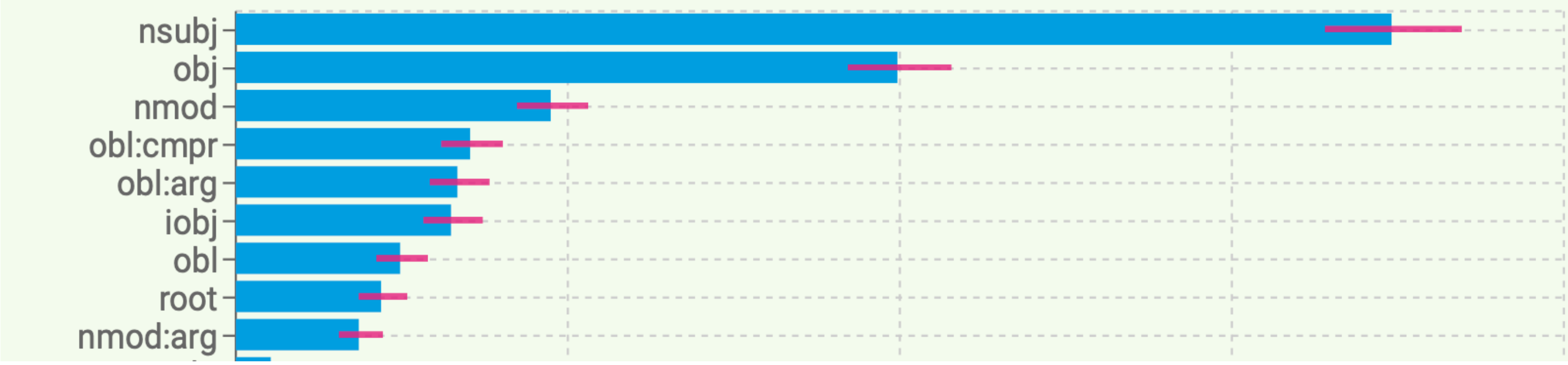
Line selection: simple



Frequency distribution

Frequency limit: 1

Level	Attribute	Ignore case	Position	(Node) start at
1.	e_deprel	□	Node	leftmost KWIC word

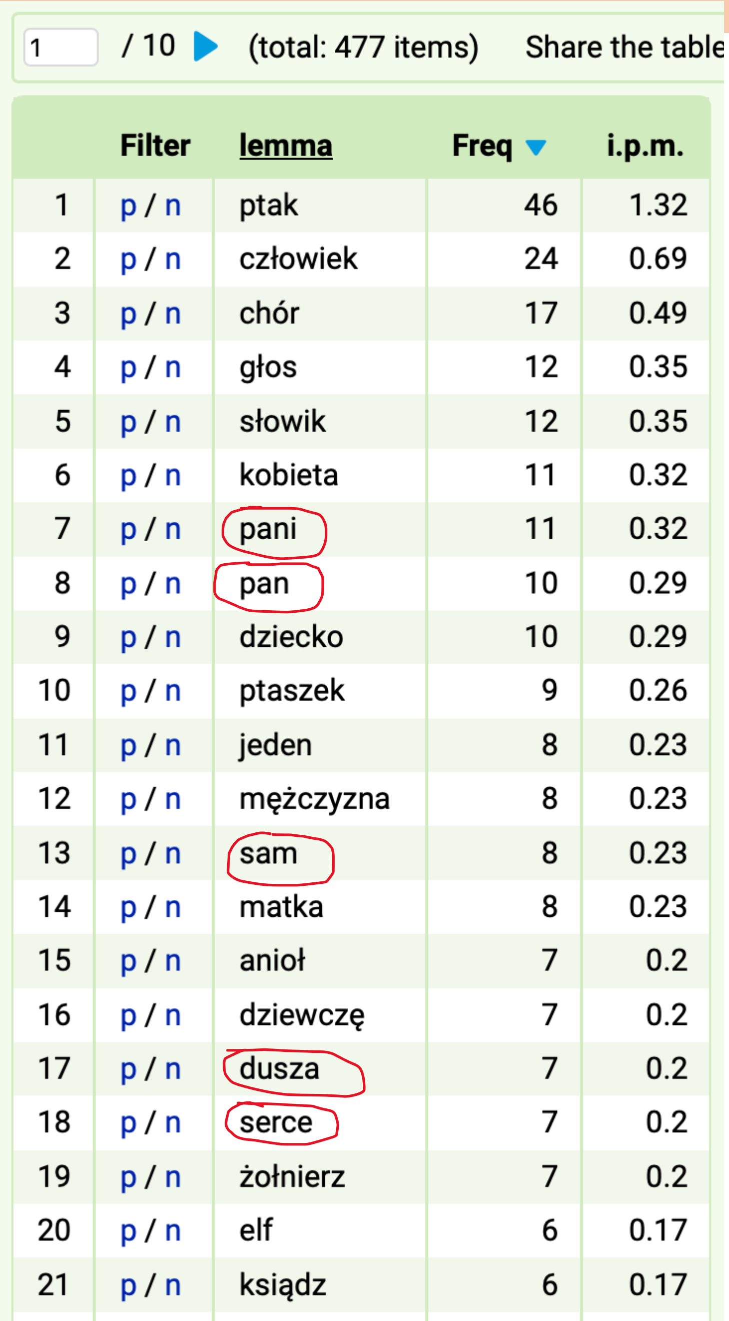


[deprel="nsubj.*" & upos!="PRON|DET" & p_lemma="śpiewać"]
Frequency > Lemmas

to list typical subjects of a predicate

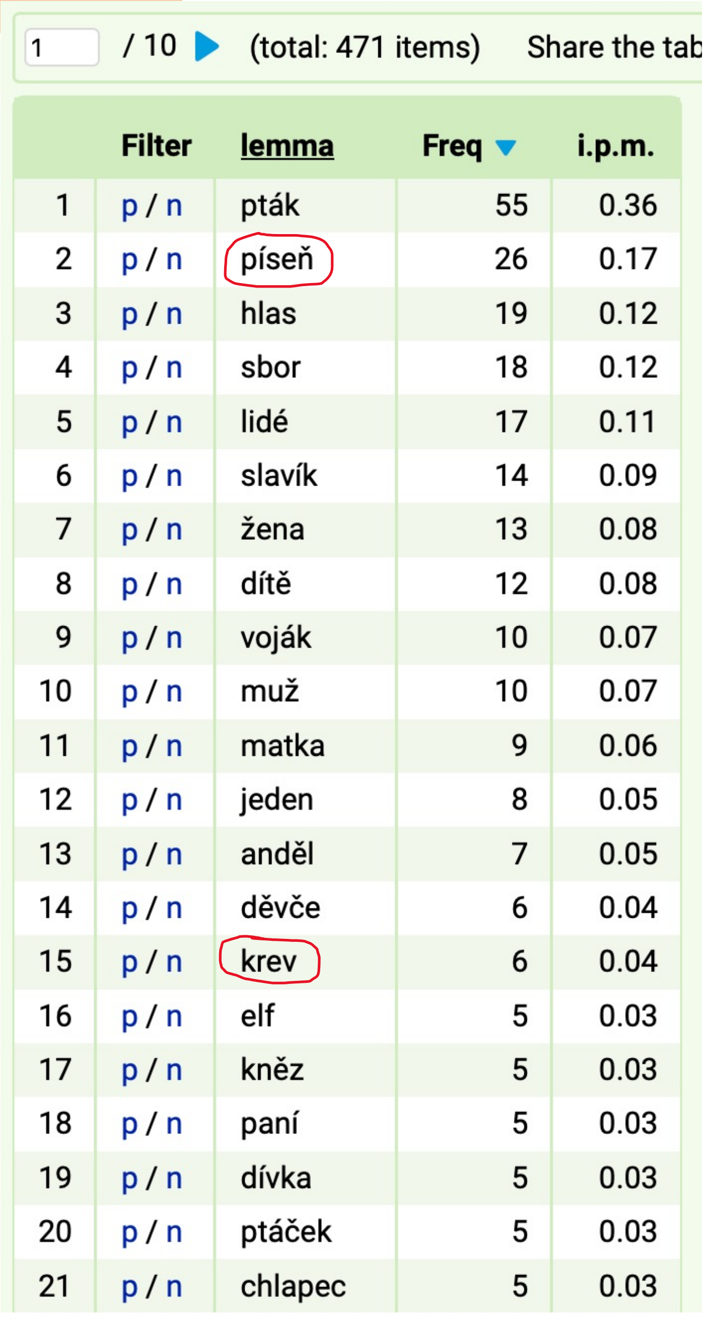
Who sings…

…in Polish?



1 / 10 (total: 477 items) Share the table

	Filter	lemma	Freq	i.p.m.
1	p / n	ptak	46	1.32
2	p / n	człowiek	24	0.69
3	p / n	chór	17	0.49
4	p / n	głos	12	0.35
5	p / n	słowik	12	0.35
6	p / n	kobieta	11	0.32
7	p / n	pani	11	0.32
8	p / n	pan	10	0.29
9	p / n	dziecko	10	0.29
10	p / n	ptaszek	9	0.26
11	p / n	jeden	8	0.23
12	p / n	mężczyzna	8	0.23
13	p / n	sam	8	0.23
14	p / n	matka	8	0.23
15	p / n	anioł	7	0.2
16	p / n	dziewczę	7	0.2
17	p / n	dusza	7	0.2
18	p / n	serce	7	0.2
19	p / n	żołnierz	7	0.2
20	p / n	elf	6	0.17
21	p / n	ksiądz	6	0.17



1 / 10 (total: 471 items) Share the tab

	Filter	lemma	Freq	i.p.m.
1	p / n	pták	55	0.36
2	p / n	píseň	26	0.17
3	p / n	hlas	19	0.12
4	p / n	sbor	18	0.12
5	p / n	lidé	17	0.11
6	p / n	slavík	14	0.09
7	p / n	žena	13	0.08
8	p / n	dítě	12	0.08
9	p / n	voják	10	0.07
10	p / n	muž	10	0.07
11	p / n	matka	9	0.06
12	p / n	jeden	8	0.05
13	p / n	anděl	7	0.05
14	p / n	děvče	6	0.04
15	p / n	krev	6	0.04
16	p / n	elf	5	0.03
17	p / n	kněz	5	0.03
18	p / n	paní	5	0.03
19	p / n	dívka	5	0.03
20	p / n	ptáček	5	0.03
21	p / n	chlapec	5	0.03

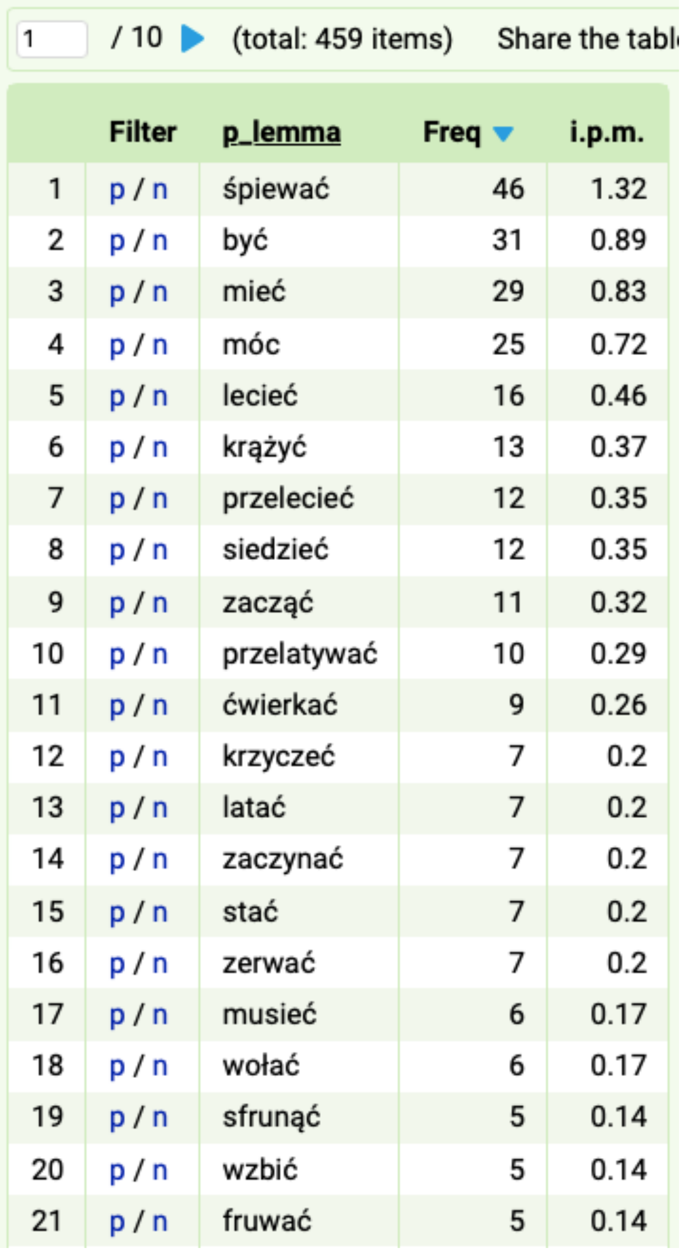
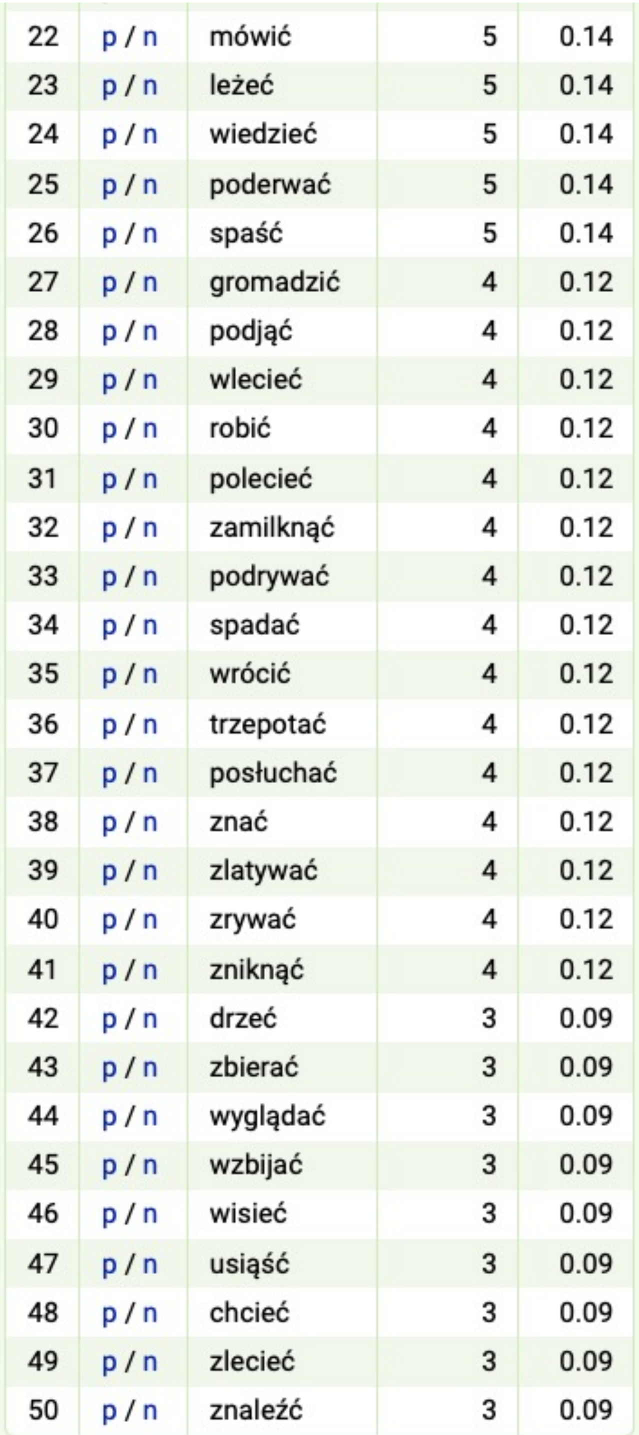
InterCorp v16ud Czech
+ InterCorp v16ud Polish
… & p_lemma="zpívat"]

…in Czech?

To list typical predicates of a subject

[deprel="nsubj.*" & lemma="ptak|ptasiek"]
Frequency > Custom > p_lemma

What do the Polish birds do?

1 / 10 ▶ (total: 459 items) Share the tabl

	Filter	p_lemma	Freq ▼	i.p.m.
1	p / n	śpiewać	46	1.32
2	p / n	być	31	0.89
3	p / n	mieć	29	0.83
4	p / n	móc	25	0.72
5	p / n	lecieć	16	0.46
6	p / n	krążyć	13	0.37
7	p / n	przelecieć	12	0.35
8	p / n	siedzieć	12	0.35
9	p / n	zacząć	11	0.32
10	p / n	przelatywać	10	0.29
11	p / n	ćwierkać	9	0.26
12	p / n	krzyczeć	7	0.2
13	p / n	latać	7	0.2
14	p / n	zaczynać	7	0.2
15	p / n	stać	7	0.2
16	p / n	zerwać	7	0.2
17	p / n	musieć	6	0.17
18	p / n	wołać	6	0.17
19	p / n	sfrunąć	5	0.14
20	p / n	wzbić	5	0.14
21	p / n	fruwać	5	0.14
22	p / n	mówić	5	0.14
23	p / n	leżeć	5	0.14
24	p / n	wiedzieć	5	0.14
25	p / n	poderwać	5	0.14
26	p / n	spaść	5	0.14
27	p / n	gromadzić	4	0.12
28	p / n	podjąć	4	0.12
29	p / n	wlecieć	4	0.12
30	p / n	robić	4	0.12
31	p / n	polecieć	4	0.12
32	p / n	zamilknąć	4	0.12
33	p / n	podrywać	4	0.12
34	p / n	spadać	4	0.12
35	p / n	wrócić	4	0.12
36	p / n	trzepotać	4	0.12
37	p / n	posłuchać	4	0.12
38	p / n	znać	4	0.12
39	p / n	zlatywać	4	0.12
40	p / n	zrywać	4	0.12
41	p / n	zniknąć	4	0.12
42	p / n	drzeć	3	0.09
43	p / n	zbierać	3	0.09
44	p / n	wyglądać	3	0.09
45	p / n	wzbijać	3	0.09
46	p / n	wisieć	3	0.09
47	p / n	usiąść	3	0.09
48	p / n	chcieć	3	0.09
49	p / n	zlecieć	3	0.09
50	p / n	znaleźć	3	0.09

Attributes for function word dependents

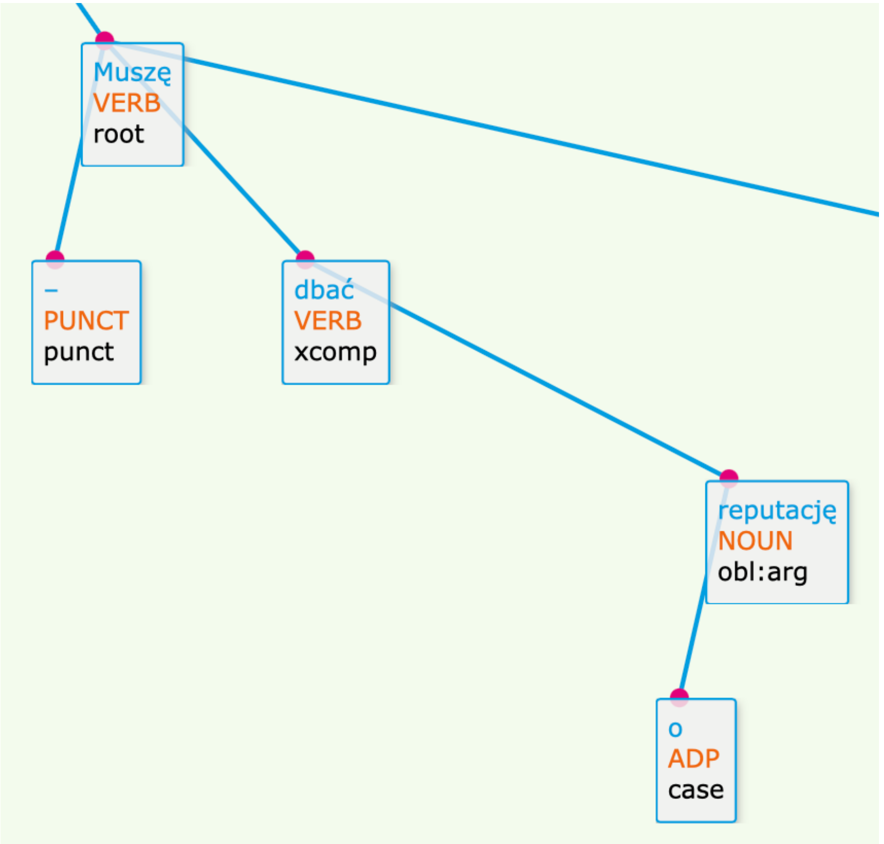
- Specified for content words
- More function words? The attribute is multivalued. In feats with "||" as the separator.
- Attribute names: **function word type_function word attribute**, e.g. `aux_feats`
- **Function word type**:
  - `aux`: auxiliary
  - `case`: preposition, postposition
  - `clf`: classifier (Chinese, Japanese)
  - `cop`: copula
  - `det`: determiner
  - `mark`: subordinating conjunction
- **Function word attribute**:
  - `lemma`
  - `upos`
  - `feats`
  - `type`: subtype of deprel, if any

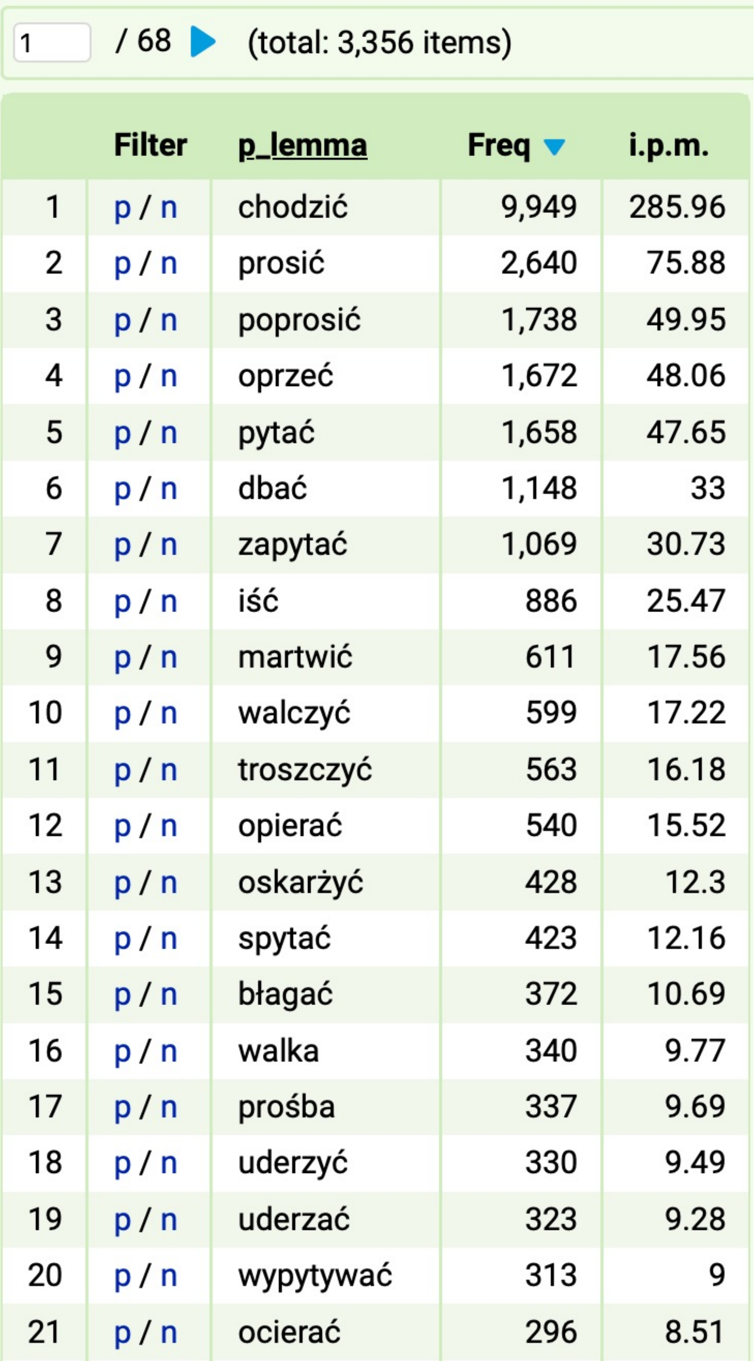
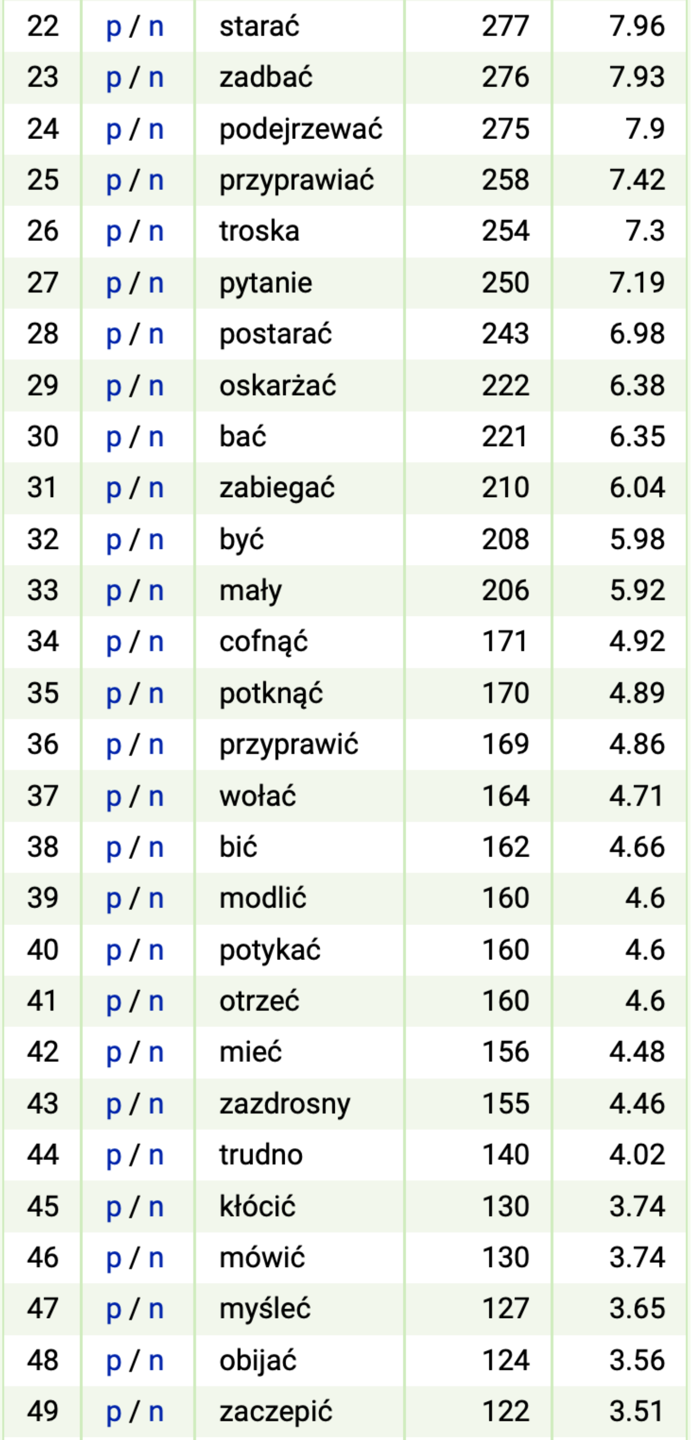
    dużo `deprel=det:numgov upos=DET`

    czasu `det_type=numgov`

To list words heading nouns or pronouns in a specific case, with a specific preposition

[case_lemma="o" & case="Acc"]
Frequency > Custom > p_lemma



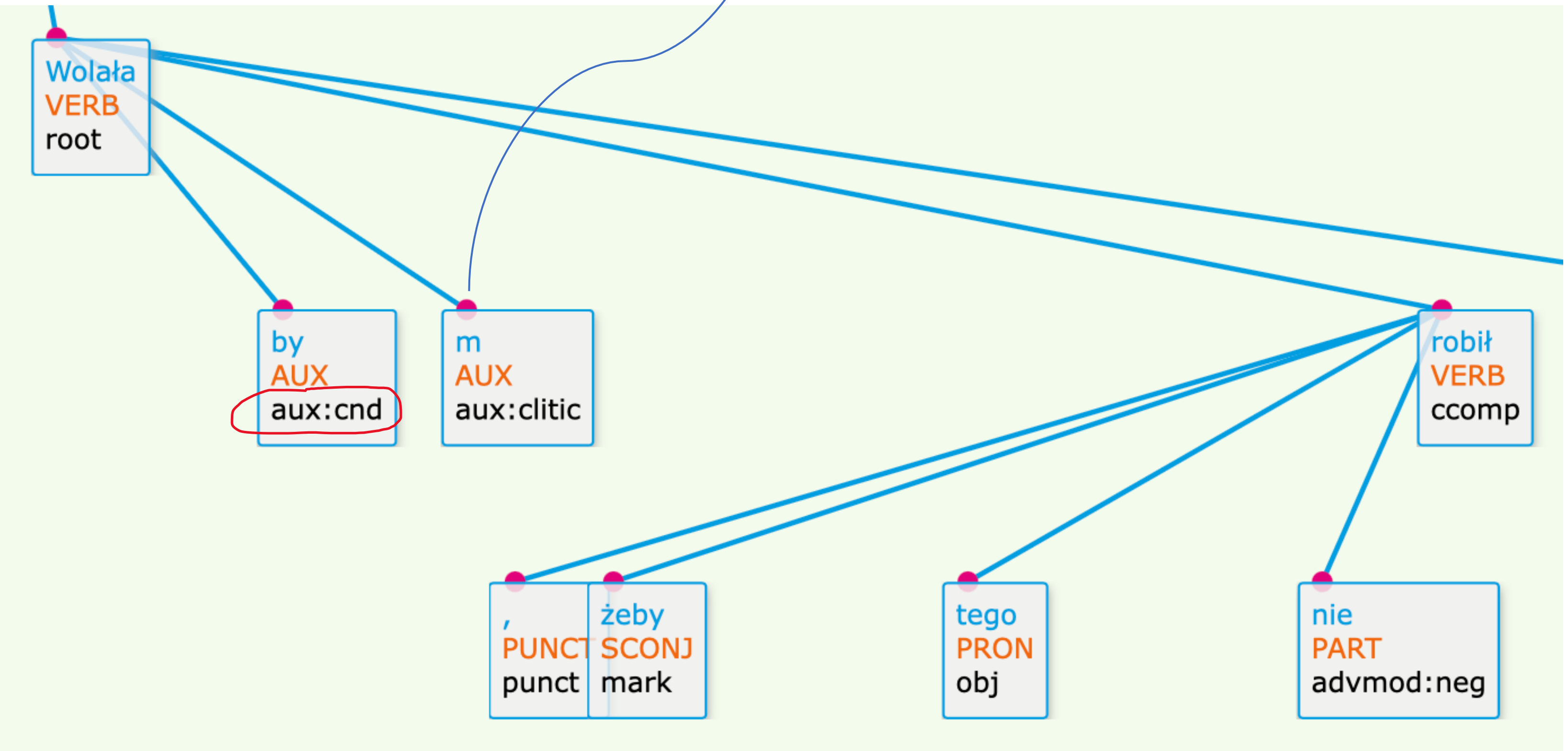
1 / 68 ▶ (total: 3,356 items)

	Filter	p_lemma	Freq ▼	i.p.m.
1	p / n	chodzić	9,949	285.96
2	p / n	prosić	2,640	75.88
3	p / n	poprosić	1,738	49.95
4	p / n	oprzeć	1,672	48.06
5	p / n	pytać	1,658	47.65
6	p / n	dbać	1,148	33
7	p / n	zapytać	1,069	30.73
8	p / n	iść	886	25.47
9	p / n	martwić	611	17.56
10	p / n	walczyć	599	17.22
11	p / n	troszczyć	563	16.18
12	p / n	opierać	540	15.52
13	p / n	oskarżyć	428	12.3
14	p / n	spytać	423	12.16
15	p / n	błagać	372	10.69
16	p / n	walka	340	9.77
17	p / n	prośba	337	9.69
18	p / n	uderzyć	330	9.49
19	p / n	uderzać	323	9.28
20	p / n	wypytywać	313	9
21	p / n	ocierać	296	8.51
22	p / n	starać	277	7.96
23	p / n	zadbać	276	7.93
24	p / n	podejrzewać	275	7.9
25	p / n	przyprawiać	258	7.42
26	p / n	troska	254	7.3
27	p / n	pytanie	250	7.19
28	p / n	postarać	243	6.98
29	p / n	oskarżać	222	6.38
30	p / n	bać	221	6.35
31	p / n	zabiegać	210	6.04
32	p / n	być	208	5.98
33	p / n	mały	206	5.92
34	p / n	cofnąć	171	4.92
35	p / n	potknąć	170	4.89
36	p / n	przyprawić	169	4.86
37	p / n	wołać	164	4.71
38	p / n	bić	162	4.66
39	p / n	modlić	160	4.6
40	p / n	potykać	160	4.6
41	p / n	otrzeć	160	4.6
42	p / n	mieć	156	4.48
43	p / n	zazdrosny	155	4.46
44	p / n	trudno	140	4.02
45	p / n	kłócić	130	3.74
46	p / n	mówić	130	3.74
47	p / n	myśleć	127	3.65
48	p / n	obijać	124	3.56
49	p / n	zaczepić	122	3.51

To find verbs in conditional mood 1st person singular

```
[aux_feats="Number=Sing" &
aux_feats="Person=1" &
aux_type="cnd"]
```

feats= Aspect=Imp
Clitic=Yes
Number=Sing
Person=1
Variant=Short



Attributes for some morphological categories

Abbr	abbreviation: Yes
Aspect	Perf, Imp, Prog, Hab, Prosp, Iter
Case	**Nom, Gen, Dat, Acc, Voc, Loc, Ins,** Erg, Abs, …
Definite	Def, Ind, Cons, Spec, Com
Degree	Pos, Cmp, Sup, Abs, Aug, Dim, Equ
Foreign	Yes
Gender	**Fem, Masc, Neut, Com**
Mood	Ind, Cnd, Imp, Int, Sub, Adm, Des, Irr, Jus, Nec, Opt, Pot, Prp, Qot
Number	**Sing, Plur, Dual**, Tri, Coll, Count, Grpa, Grpl, Inv, Pauc, Ptan
NumType	type of numeral: Card, Ord, Dist, Frac, Mult, Range, Sets
Person	**1, 2, 3**, 0, 4
Polarity	Neg, Pos
Poss	possessive: Yes
PronType	type of pronoun: Art, Dem, Emp, Exc,. Ind, Int, Neg, Prs, Rcp, Rel, Tot
Reflex	reflexive: Yes
Tense	Pres, Past, Fut, Imp, Pqp
VerbForm	Fin, Inf, Part, Noun, Conv, Ger, Gdv
Voice	Act, Pass, Mid, Cau, Antip, Bfoc, Dir, Inv, Lfoc, Rcp

Equivalent queries

```
[upos="NOUN" & feats="Gender=Fem" & feats="Case=Gen"]
```

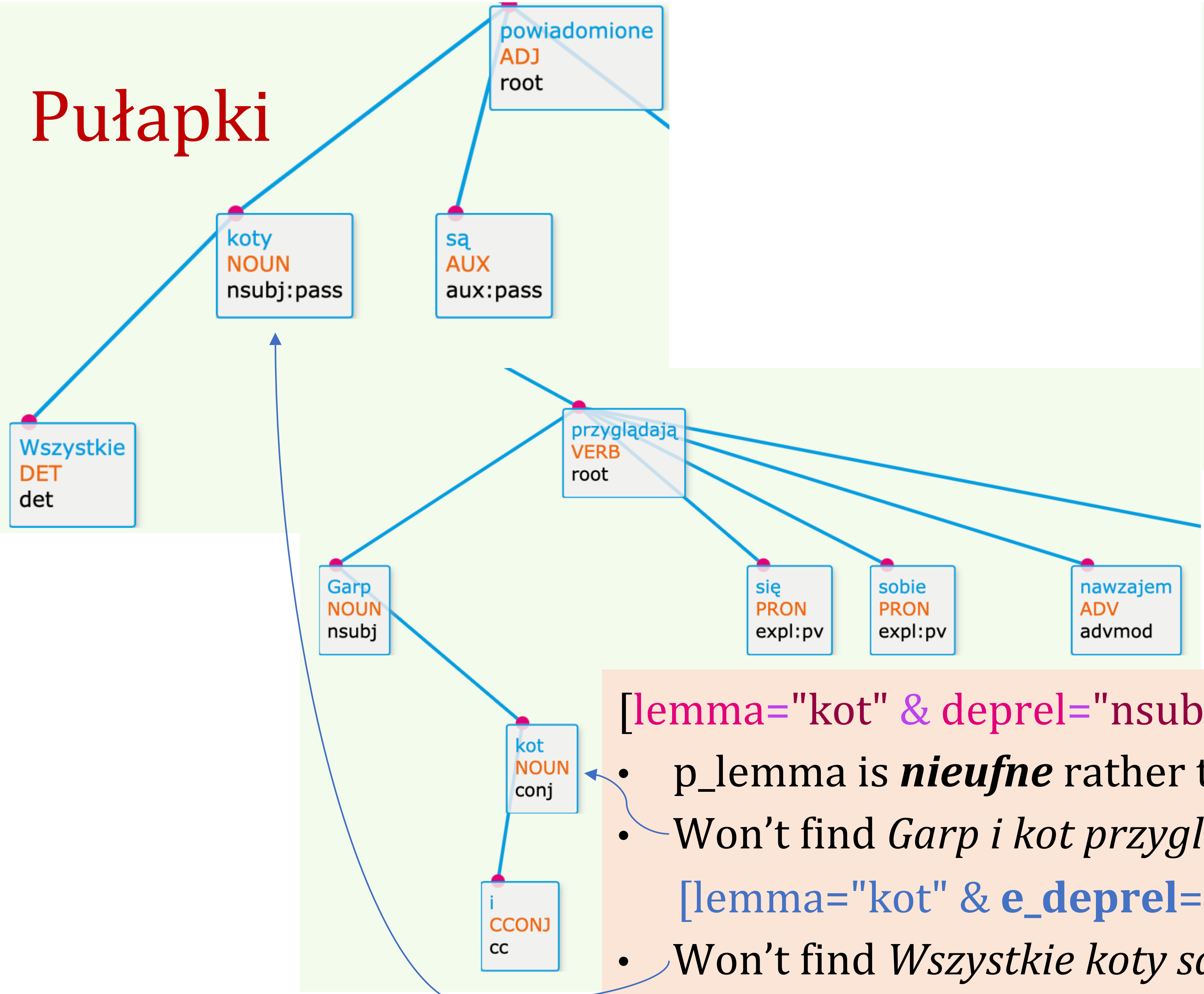
```
[upos="NOUN" & feats=".*Case=Gen.*Gender=Fem.*"]
```

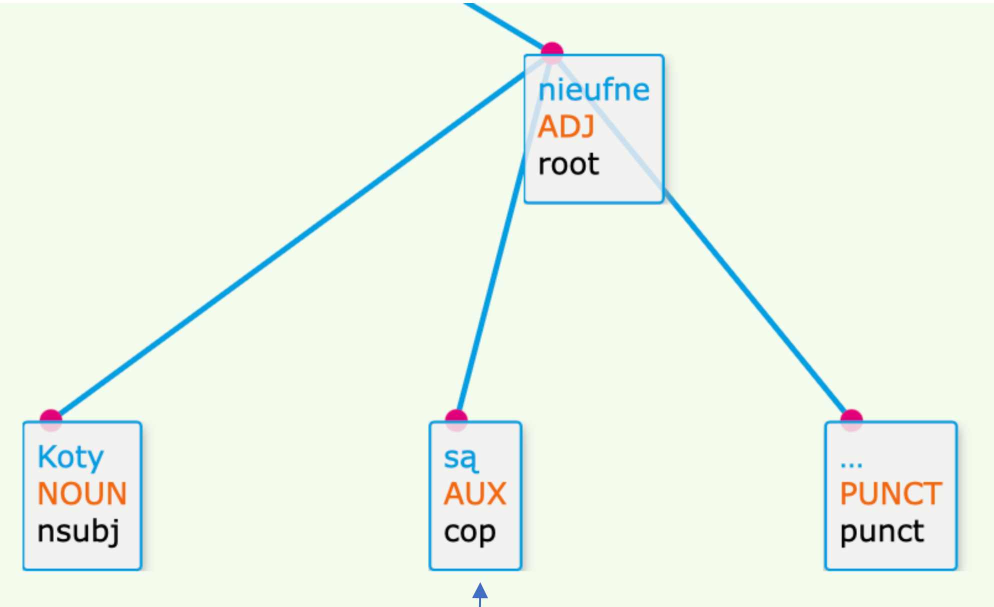
```
[upos="NOUN" & gender= "Fem" & case="Gen"]
```

```
[xpos="subst:...:gen:f"]
```

More results – includes names (upos="PROPN")

Pułapki





[lemma="kot" & deprel="nsubj"]

- p_lemma is ***nieufne*** rather than ***są***
- Won’t find *Garp i kot przyglądają się* [deprel="**conj**"]

  [lemma="kot" & **e_deprel**="nsubj"]
- Won’t find *Wszystkie koty są powiadomione* [deprel="nsubj**:pass**"]

  [lemma=“kot" & deprel="**nsubj.***"]

Outline

1. About InterCorp
2. Universal Dependencies (UD)
3. InterCorp with UD
4. Metrics of syntactic complexity and lexical diversity
5. Using the metrics
6. Perspectives, questions, discussion

Thanks to many are due for the idea, design and implementation of the metrics:

- *Olga Nádvorníková (2021–) **
- *Martin Vavřín (2021–2022) ***
- *Bohumil Šimčík (2023–) ***
- *Michal Křen and Michal Škrabal ***
- *Jiří Milička (metrics of lexical diversity) ***

** Institute of Romance Studies*

*** Institute of the Czech National Corpus*

What is syntactic complexity?

… syntactic complexity in language is related to the number, type, and depth of embedding in a text … (Beaman 1984: 45)

… can be determined by:

- number and variability of clauses
- their hierarchy within the sentence

Simplifying complexity

- Complexity is **multi-dimensional,** thus more metrics should be combined (Biber, Larsson & Hancock 2023).
- Metrics are **specific** to genre and language.
- We aim at **absolute** (*objective*) complexity; rather than relative (*subjective*, reader-oriented, measuring processing load, *readability*) (Brunato and Venturi 2022: 1, Szmrecsanyi and Kortmann 2012: 10).

Research of complexity in a wider context

- **syntactic** complexity (e.g. Ferreira: 1991; Givón: 1991; Szmrecsanyi: 2004, *complexité syntaxique* De Clercq 2016 aj.)
- **cognitive** complexity (e.g. Mondorf: 2003; Givón: 1991; Rohdenburg: 1996)
- **clause** complexity (e.g. Kuboň: 2001)
- **linguistic** complexity (e.g. Schleppegrell: 1992)
- **structural** complexity (e.g. Givón: 1991; Arnold et al.: 2000)
- **grammatical / syntactic weight** (e.g. Wasow: 1997; Wasow and Arnold: 2003)
- **information density** (Fabricius-Hansen 1999 aj.)

What can be done with syntactic complexity

a) **Language development** (Givón 2009:4)

d) **Monolingual studies** (Mačutek, Čech & Milička 2019), propositions relatives (Hudelot 1980), Biber, Larsson & Hancock 2023 (English), etc.

c) **Contrastive studies**: clause-linking (Lehmann 1988), clause-combining (Cosme 2006, etc.), information packaging (Solfjeld 1996, Fabricius-Hansen 1999), *shared task UD* (Berdicevskis et al. 2018, etc.),

f) **Translation studies**: Izquierdo & Marco 2000, Canavese & Mori 2021; comparable or parallel corpora (*translation universals* – simplification, normalisation, etc.)

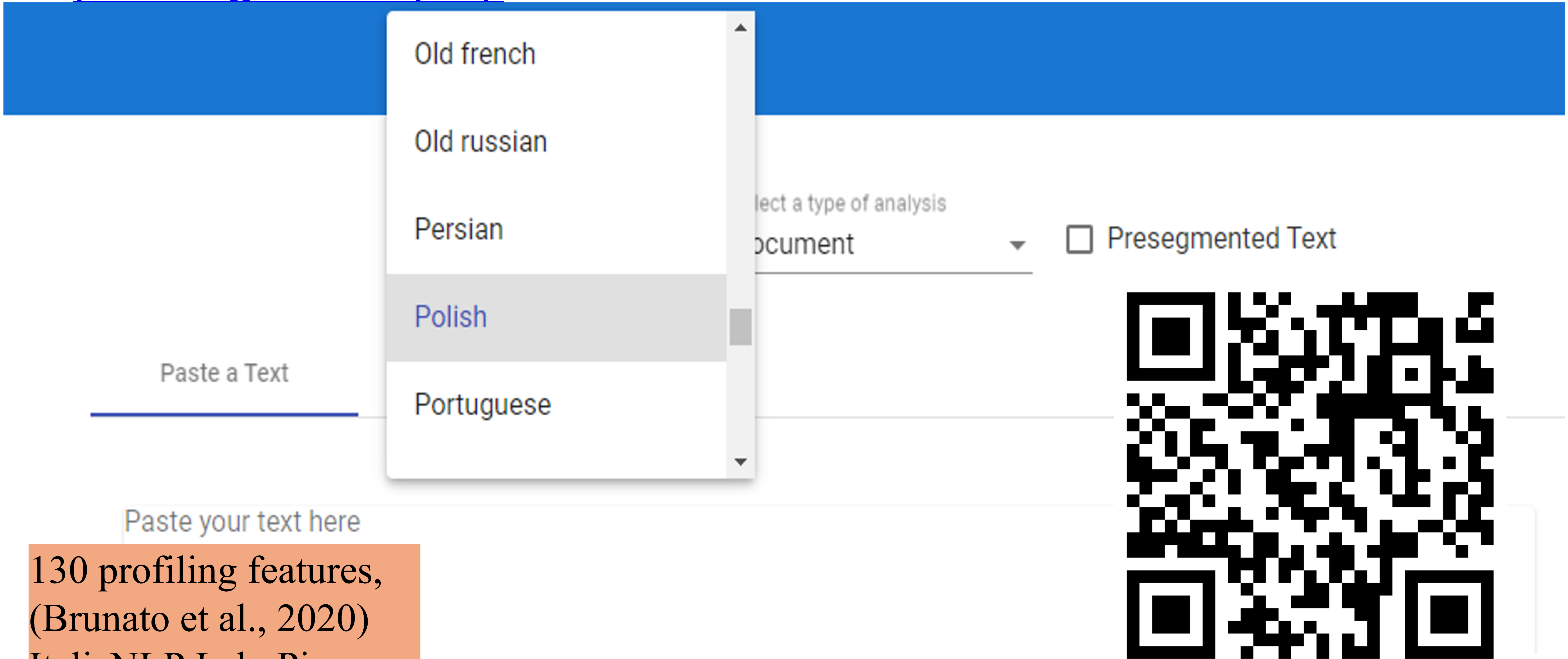
b) ***Register variation*****:** spoken/written (Beaman 1984 etc.), academic: Biber & Gray 2017, etc..

e) ***Typology*****:** Levshina 2019, 2021 – Leipzig Corpora Collection (comparable, UD)

g) **Readability:** Kincaid et al. 1975, Dell'Orletta et al. 2011, Gruszczyński & Ogrodniczuk 2015 ***Jasnopis***.

h) **Language acquisition, proficiency assessment** (L1 et L2), Lu 2010, etc.

Linguistic Profiling Tool (UD) http://linguistic-profiling.italianlp.it/



130 profiling features, (Brunato et al., 2020) ItaliaNLP Lab, Pisa

Metrics of syntactic complexity and lexical diversity

Syntactic complexity

by syntactic category:

- clauses
- noun phrases

by tree dimension:

- vertical (no. of embeddings)
- horizontal (no. of words)

Lexical diversity

no. of lexical types in a moving window 1000 tokens wide:

- word forms
- lexemes

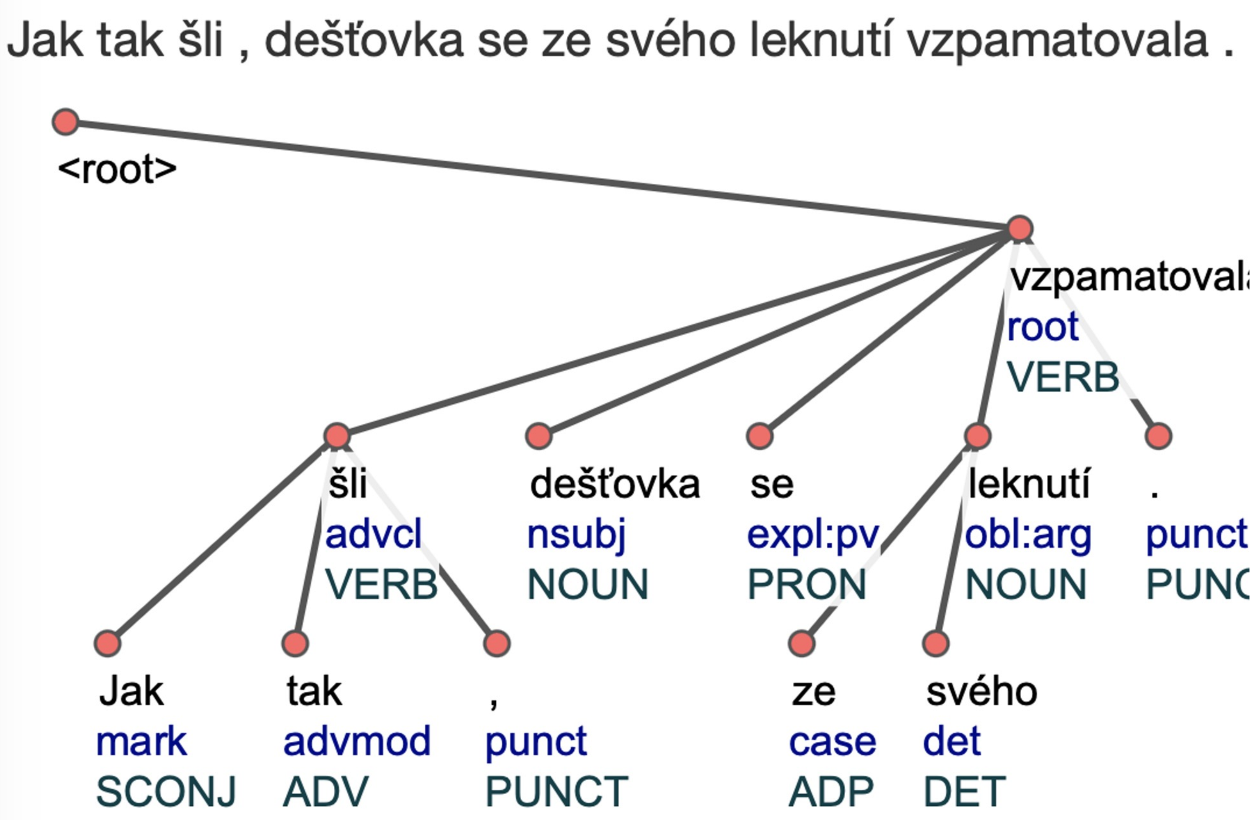
Metrics as metadata

Attributes of **<text>**:

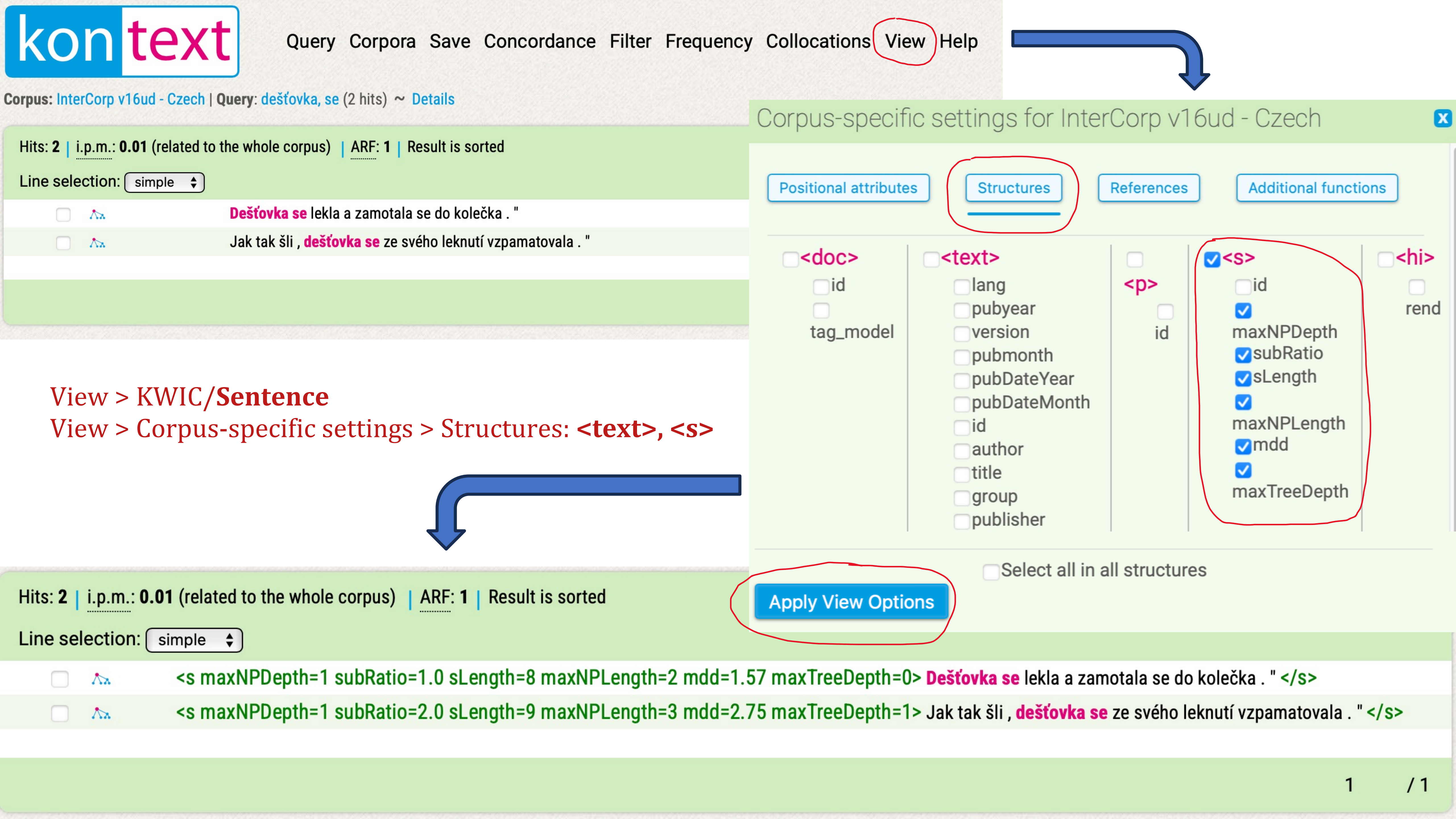
<text
author=Čapek, Josef
title=Povídání o pejskovi a kočičce
maxNPLengthAvg=2.65
maxNPDepthAvg=1.02
subRatioAvg=1.72
maxTreeDepthAvg=0.89
sLengthAvg=14.08
mdd=2.69
lexDivWord=463.83
lexDivLemma=304.68 **… >**

Attributes of **<s>** (sentence):

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


Jak tak šli , dešťovka se ze svého leknutí vzpamatovala .

Jak tak szli, dżdżownica otrząsnęła się ze swojego przerażenia.


kontext
Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: InterCorp v16ud - Czech | **Query:** dešťovka, se (2 hits) ~ Details

Hits: **2** | i.p.m.: **0.01** (related to the whole corpus) | ARF: **1** | Result is sorted

Line selection: simple

□ **Dešťovka se** lekla a zamotala se do kolečka . "

□ Jak tak šli , **dešťovka se** ze svého leknutí vzpamatovala . "

Corpus-specific settings for InterCorp v16ud - Czech

Positional attributes
Structures
References
Additional functions

□ <doc>
- □ id
- □ tag_model

□ <text>
- □ lang
- □ pubyear
- □ version
- □ pubmonth
- □ pubDateYear
- □ pubDateMonth
- □ id
- □ author
- □ title
- □ group
- □ publisher

□ <p>
- □ id

☑ <s>
- □ id
- ☑ maxNPDepth
- ☑ subRatio
- ☑ sLength
- ☑ maxNPLength
- ☑ mdd
- ☑ maxTreeDepth

□ <hi>
- □ rend

□ Select all in all structures

View > KWIC/**Sentence**
View > Corpus-specific settings > Structures: **<text>, <s>**

Hits: **2** | i.p.m.: **0.01** (related to the whole corpus) | ARF: **1** | Result is sorted

Line selection: simple

□ <s maxNPDepth=1 subRatio=1.0 sLength=8 maxNPLength=2 mdd=1.57 maxTreeDepth=0> **Dešťovka se** lekla a zamotala se do kolečka . " </s>

□ <s maxNPDepth=1 subRatio=2.0 sLength=9 maxNPLength=3 mdd=2.75 maxTreeDepth=1> Jak tak šli , **dešťovka se** ze svého leknutí vzpamatovala . " </s>

1 / 1

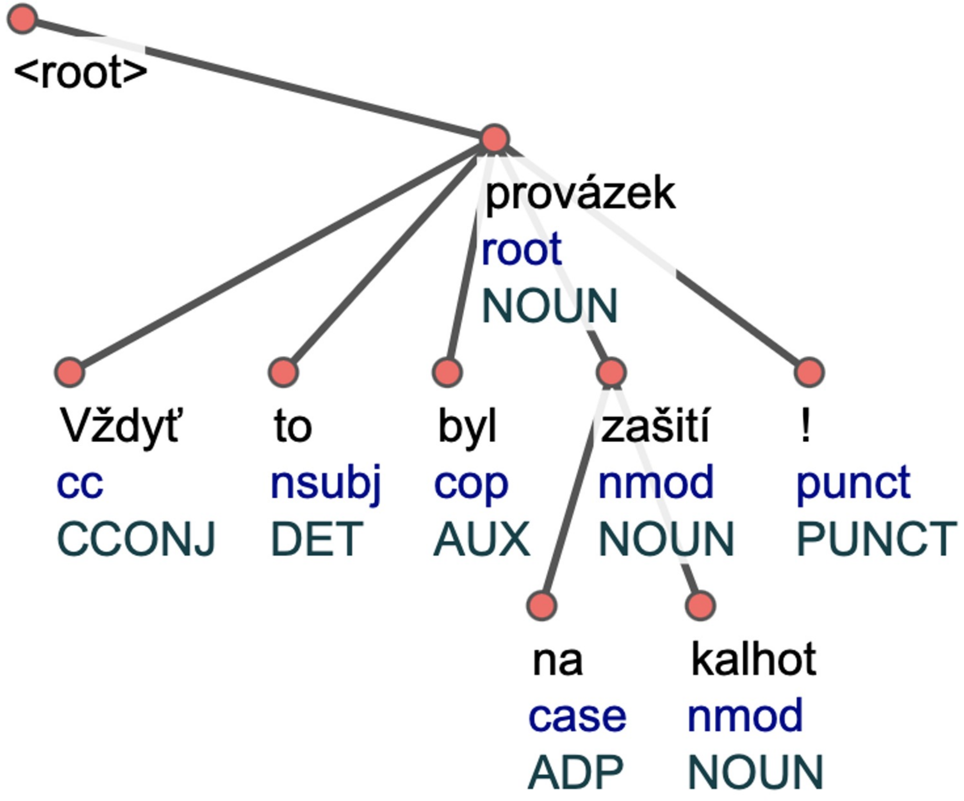
Sentence-level complexity metrics

	Noun phrase	**Sentence**
horizontal dimension	maxNPLength *maximum length*	sLength *sentence length in words*
		subRatio *subordination ratio*
vertical dimension	maxNPDepth *maximum depth*	maxTreeDepth *maximum tree depth*
cognitive load		mdd *mean dependency distance*

What is a noun phrase?

- Subtree with NOUN, PNOM, PRON as the head
- Every conjunct separately
- Ignoring: punctuation, conjunction
- Nominal predicate? Part of the NP (nmod: *provázek na zašití kalhot*), not of the whole predicate (nsubj, cop: *Vždyť to byl …*)

Vždyť to byl provázek na zašití kalhot !



<root>

provázek
root
NOUN

Vždyť
cc
CCONJ

to
nsubj
DET

byl
cop
AUX

zašití
nmod
NOUN

!
punct
PUNCT

na
case
ADP

kalhot
nmod
NOUN

Przecież to był sznurek do zszycia spodni!

Noun phrase – complexity metrics

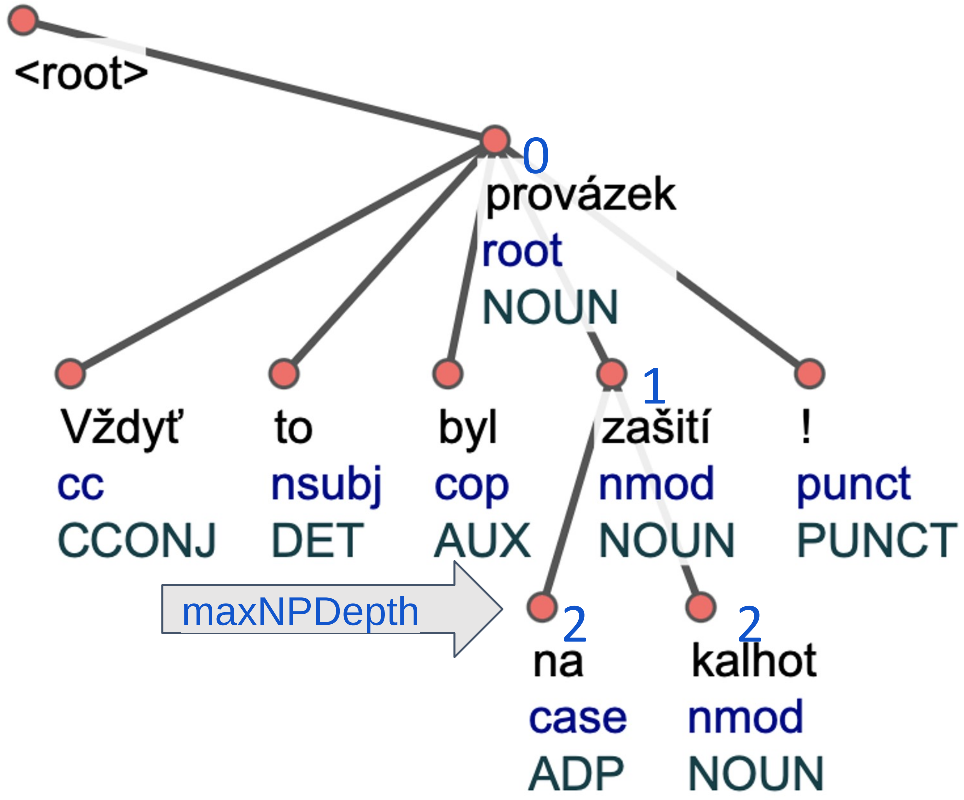
MaxNPLength:

- no. of words in the longest NP
- *provázek na zašití kalhot*
- = 4

MaxNPDepth:

- maximum no. of embeddings in any NP
- *provázek* … 0
- *zašití* … 1
- *na* … 2
- *kalhot* … 2
- = 2

Vždyť to byl provázek na zašití kalhot !



Sentence-level complexity metrics

	Noun phrase	Sentence
horizontal dimension	maxNPLength *maximum length*	sLength *sentence length in words*
		subRatio *subordination ratio*
vertical dimension	maxNPDepth *maximum depth*	maxTreeDepth *maximum tree depth*
cognitive load		mdd *mean dependency distance*

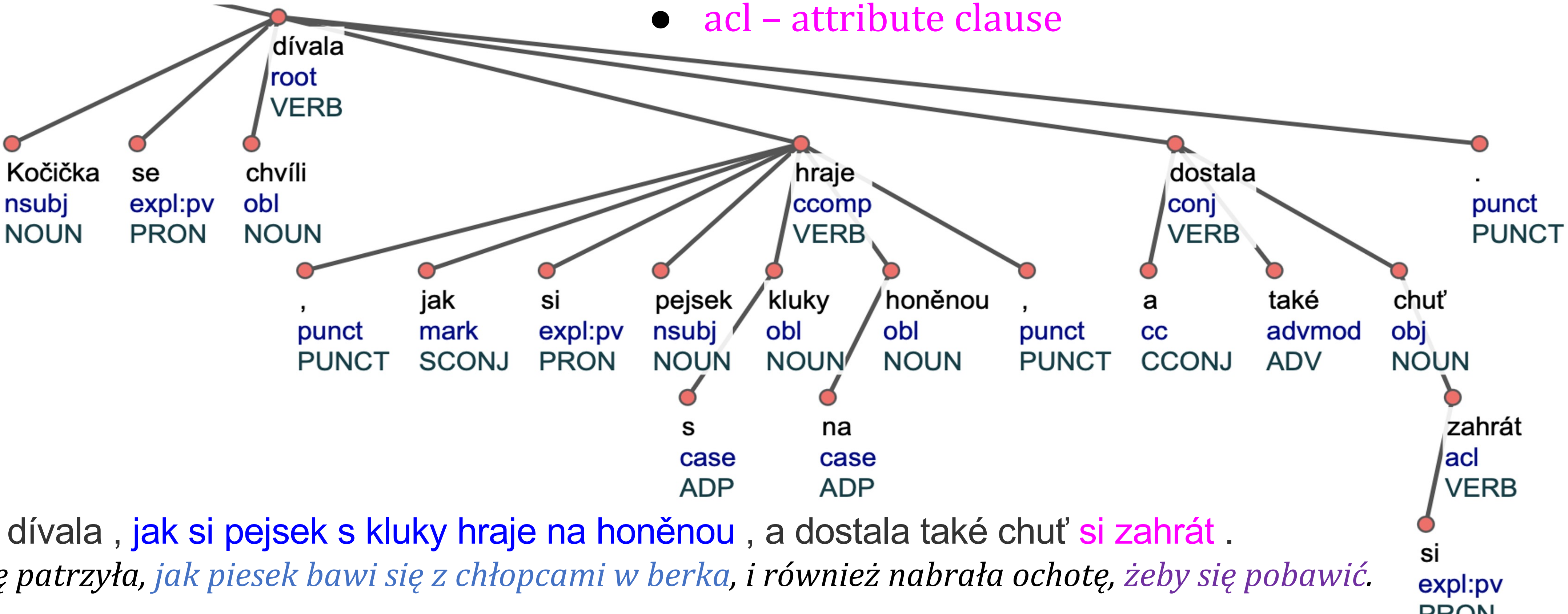
What is a sentence?

T-unit:

- main clause including all dependent clauses (Hunt 1965)
- each main conjunct clause, including all dependent clauses, is one T-unit

(Subordinate) clause, even non-finite:

- csubj – subject clause
- ccomp – complement clause
- xcomp – open predicate (predicative complement)
- advcl – adverbial clause
- acl – attribute clause



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

Kotka przez chwilę patrzyła, jak piesek bawi się z chłopcami w berka, i również nabrała ochotę, żeby się pobawić.

Sentence – complexity metrics

sLength:

- no. of words in the sentence
- punctuation is ignored

MaxTreeDepth:

- maximum number of embedded clauses in the sentence
- coordination is skipped

subRatio:

- subordination ratio
- (no. of T-units + no. of clauses) / no. of T-units

Sentence

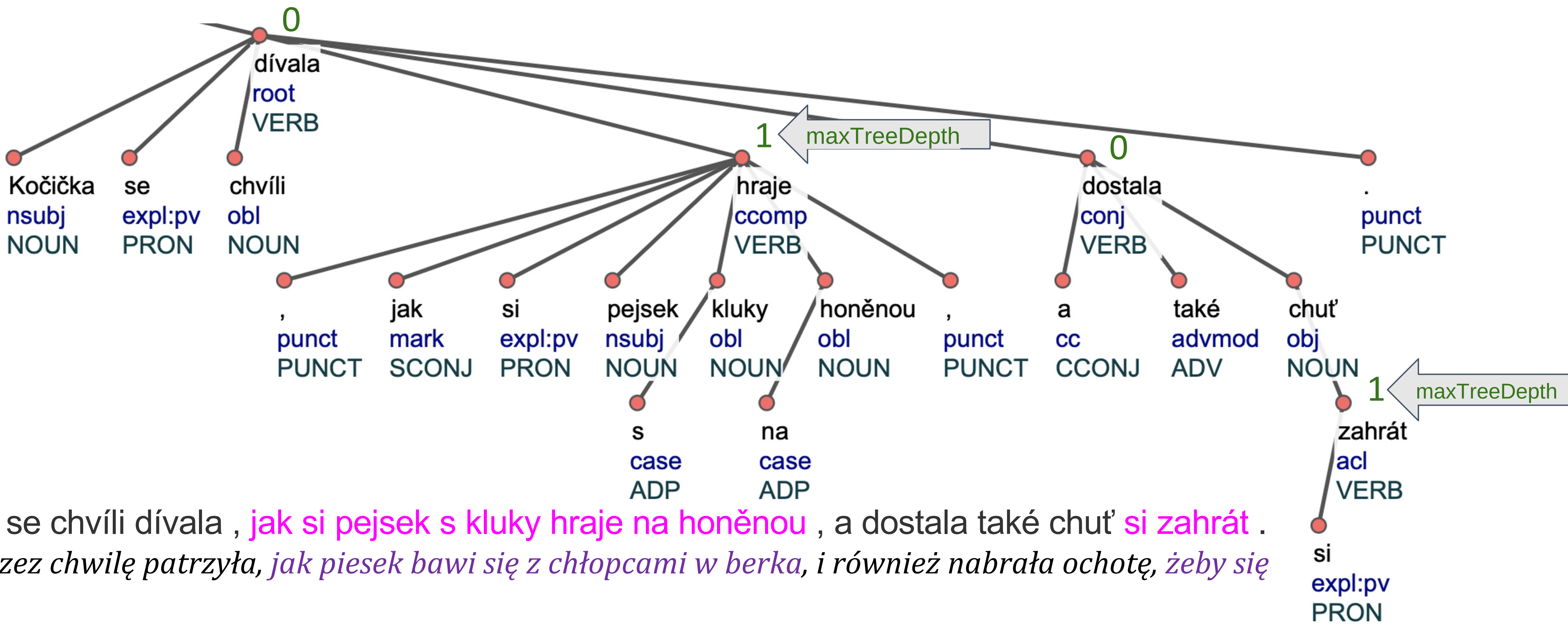
maxNPDepth=2 **subRatio=2.0** sLength=18

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

No. of T-units = 2

No. of clauses = 2

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

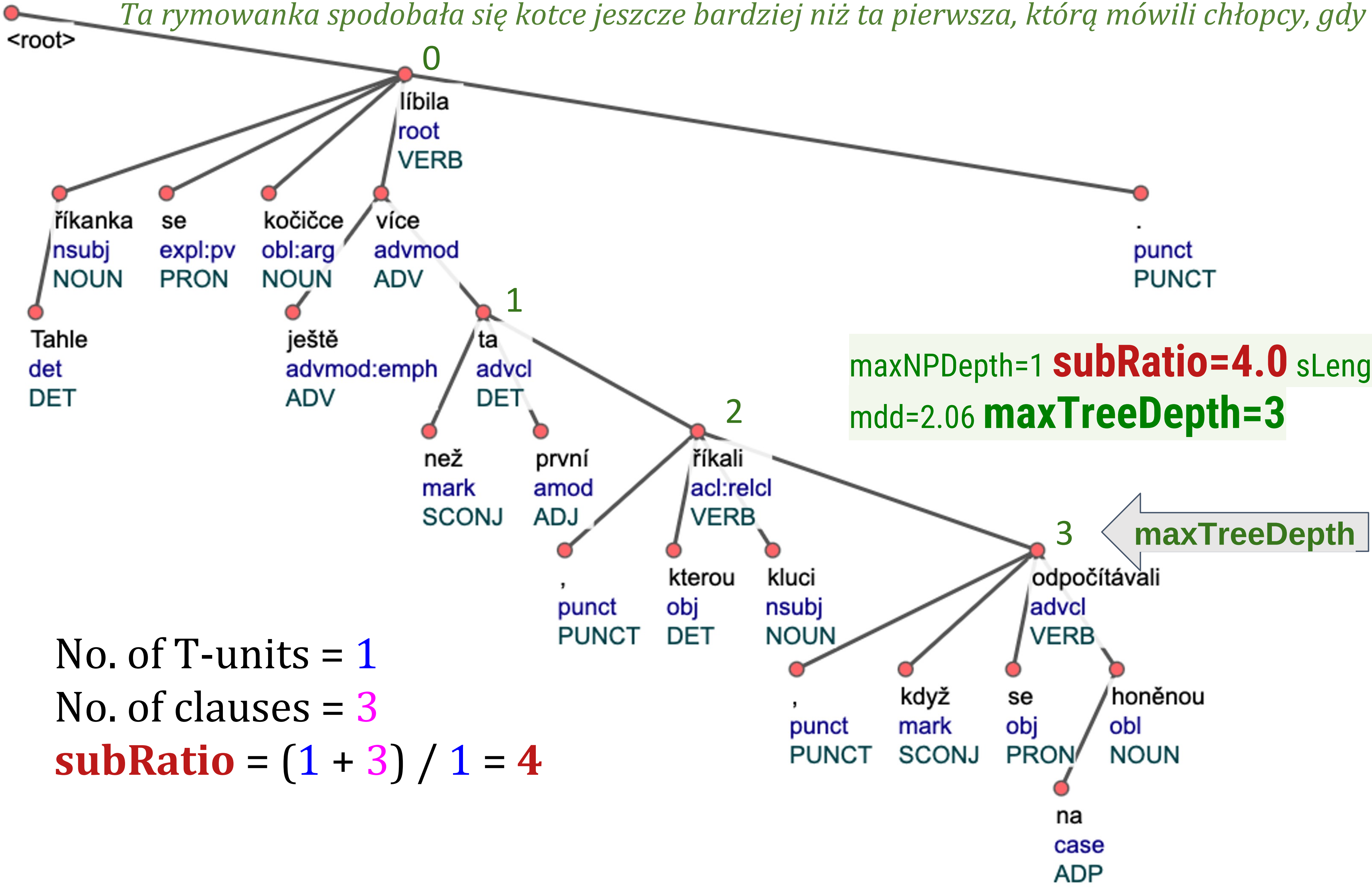


Kočička se chvíli dívala , jak si pejsek s kluky hraje na honěnou , a dostala také chuť si zahrát .

Kotka przez chwilę patrzyła, jak piesek bawi się z chłopcami w berka, i również nabrała ochotę, żeby się pobawić.

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

Ta rymowanka spodobała się kotce jeszcze bardziej niż ta pierwsza, którą mówili chłopcy, gdy odliczali się do berka.



maxNPDepth=1 **subRatio=4.0** sLength=18 maxNPLength=2

mdd=2.06 **maxTreeDepth=3**

No. of T-units = 1

No. of clauses = 3

subRatio = (1 + 3) / 1 = 4

Sentence-level complexity metrics

	Noun phrase	**Sentence/Clause**
horizontal dimension	maxNPLength *maximum length*	sLength *sentence length in words*
		subRatio *subordination ratio*
vertical dimension	maxNPDepth *maximum depth*	maxTreeDepth *maximum tree depth*
cognitive load		mdd *mean dependency distance*

Wersję pilotażową udało się uruchomić dopiero w zeszły czwartek.

Sentence – cognitive load

mdd:

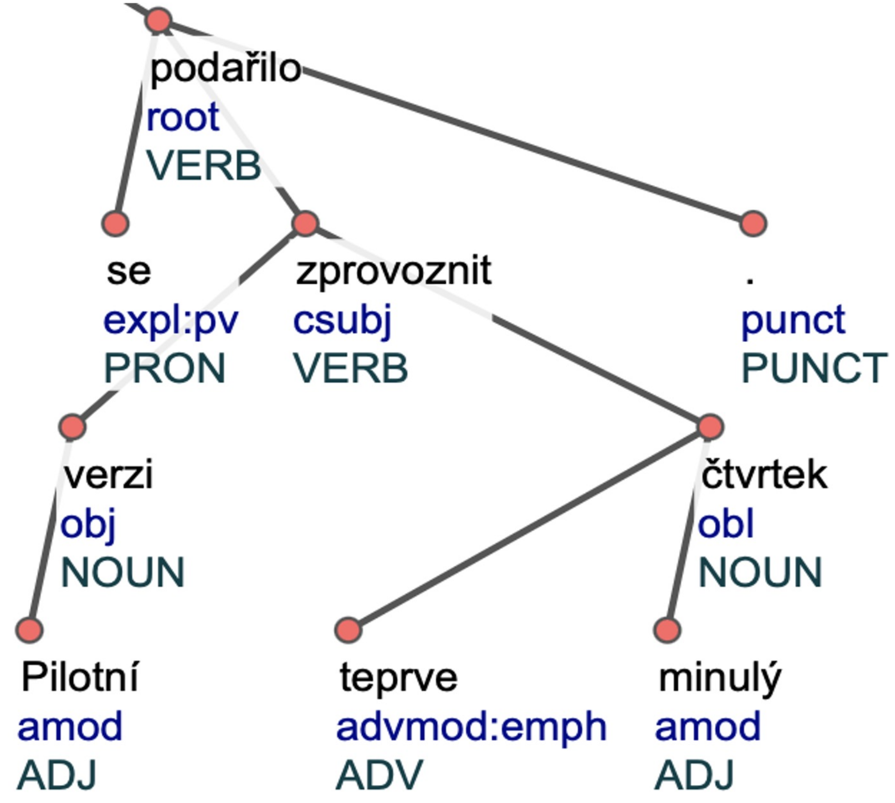
- Mean Dependency Distance (Yan & Li, 2019; Mačutek et al., 2021)
- Average head-daughter distance
- Punctuation is ignored
- calculation (n = 8 ... no. of words in sentence)

  $\mathrm{DD}_i = |\,\mathrm{ID}_i - \mathrm{head}_i\,|$

  $\mathrm{DD} = \sum_{i=0 \text{ to } n} \mathrm{DD}_i$

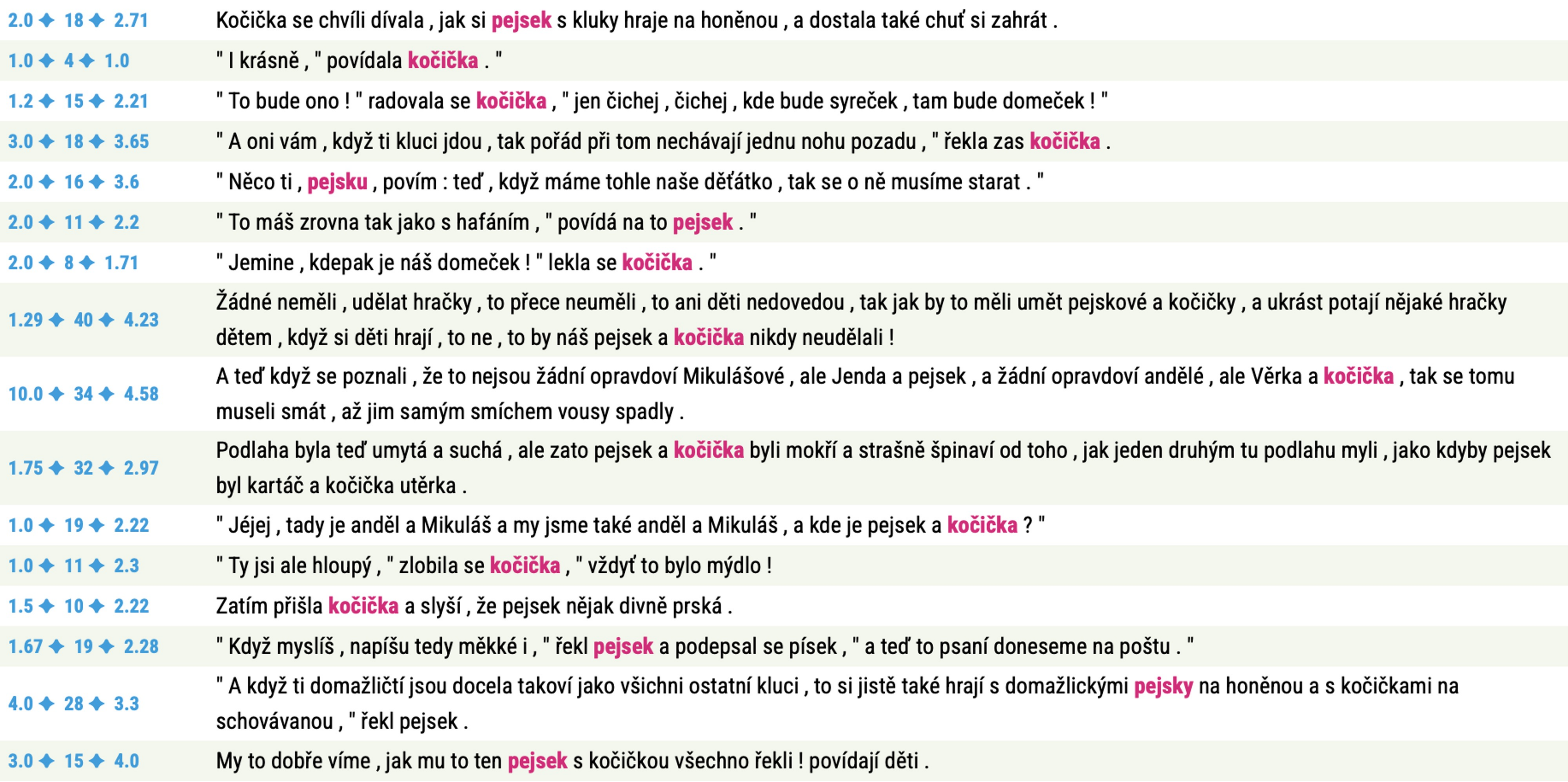
  $\mathrm{mdd} = \mathrm{DD} / (n - 1)$
- DD = 12

  mdd = 12 / 7 ≅ 1,71



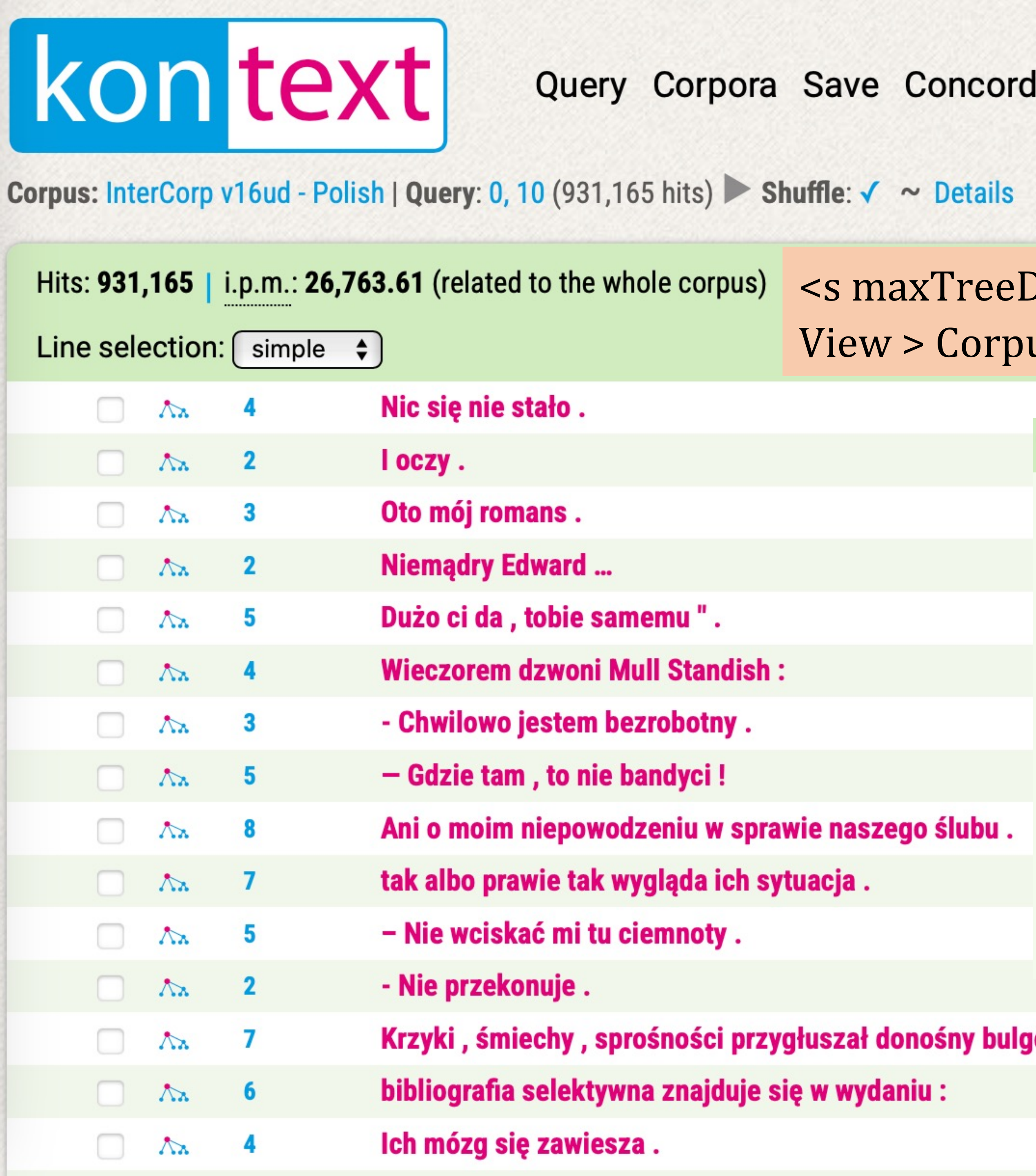
	Pilotní	*verzi*	*se*	*podařilo*	*zprovoznit*	*teprve*	*minulý*	*čtvrtek*
ID (= i)	1	2	3	4	5	6	7	8
head_i	2	5	4	0	4	8	8	5
DD_i	1	3	1	0	1	2	1	3

Comparison: subRatio * sLength * mdd



2.0 ✦ 18 ✦ 2.71	Kočička se chvíli dívala , jak si **pejsek** s kluky hraje na honěnou , a dostala také chuť si zahrát .
1.0 ✦ 4 ✦ 1.0	" I krásně , " povídala **kočička** . "
1.2 ✦ 15 ✦ 2.21	" To bude ono ! " radovala se **kočička** , " jen čichej , čichej , kde bude syreček , tam bude domeček ! "
3.0 ✦ 18 ✦ 3.65	" A oni vám , když ti kluci jdou , tak pořád při tom nechávají jednu nohu pozadu , " řekla zas **kočička** .
2.0 ✦ 16 ✦ 3.6	" Něco ti , **pejsku** , povím : teď , když máme tohle naše děťátko , tak se o ně musíme starat . "
2.0 ✦ 11 ✦ 2.2	" To máš zrovna tak jako s hafáním , " povídá na to **pejsek** . "
2.0 ✦ 8 ✦ 1.71	" Jemine , kdepak je náš domeček ! " lekla se **kočička** . "
1.29 ✦ 40 ✦ 4.23	Žádné neměli , udělat hračky , to přece neuměli , to ani děti nedovedou , tak jak by to měli umět pejskové a kočičky , a ukrást potají nějaké hračky dětem , když si děti hrají , to ne , to by náš pejsek a **kočička** nikdy neudělali !
10.0 ✦ 34 ✦ 4.58	A teď když se poznali , že to nejsou žádní opravdoví Mikulášové , ale Jenda a pejsek , a žádní opravdoví andělé , ale Věrka a **kočička** , tak se tomu museli smát , až jim samým smíchem vousy spadly .
1.75 ✦ 32 ✦ 2.97	Podlaha byla teď umytá a suchá , ale zato pejsek a **kočička** byli mokří a strašně špinaví od toho , jak jeden druhým tu podlahu myli , jako kdyby pejsek byl kartáč a kočička utěrka .
1.0 ✦ 19 ✦ 2.22	" Jéjej , tady je anděl a Mikuláš a my jsme také anděl a Mikuláš , a kde je pejsek a **kočička** ? "
1.0 ✦ 11 ✦ 2.3	" Ty jsi ale hloupý , " zlobila se **kočička** , " vždyť to bylo mýdlo !
1.5 ✦ 10 ✦ 2.22	Zatím přišla **kočička** a slyší , že pejsek nějak divně prská .
1.67 ✦ 19 ✦ 2.28	" Když myslíš , napíšu tedy měkké i , " řekl **pejsek** a podepsal se písek , " a teď to psaní doneseme na poštu . "
4.0 ✦ 28 ✦ 3.3	" A když ti domažličtí jsou docela takoví jako všichni ostatní kluci , to si jistě také hrají s domažlickými **pejsky** na honěnou a s kočičkami na schovávanou , " řekl pejsek .
3.0 ✦ 15 ✦ 4.0	My to dobře víme , jak mu to ten **pejsek** s kočičkou všechno řekli ! povídají děti .

<s maxTreeDepth="0" & sLength <= "10" />
View > Corpus-specific settings > References > s.sLength

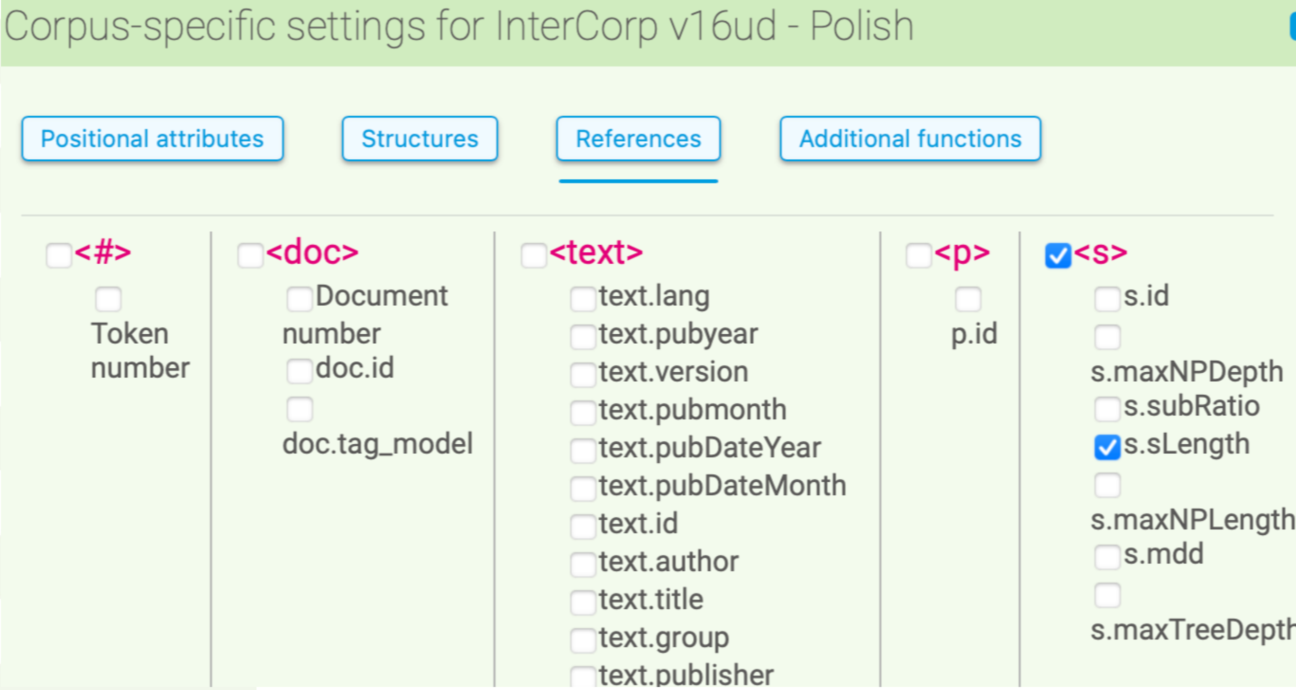


Corpus: InterCorp v16ud - Polish | Query: 0, 10 (931,165 hits) ▶ Shuffle: ✓ ~ Details

Hits: 931,165 | i.p.m.: 26,763.61 (related to the whole corpus)

Line selection: simple

□	4	Nic się nie stało .
□	2	I oczy .
□	3	Oto mój romans .
□	2	Niemądry Edward …
□	5	Dużo ci da , tobie samemu " .
□	4	Wieczorem dzwoni Mull Standish :
□	3	- Chwilowo jestem bezrobotny .
□	5	– Gdzie tam , to nie bandyci !
□	8	Ani o moim niepowodzeniu w sprawie naszego ślubu .
□	7	tak albo prawie tak wygląda ich sytuacja .
□	5	– Nie wciskać mi tu ciemnoty .
□	2	- Nie przekonuje .
□	7	Krzyki , śmiechy , sprośności przygłuszał donośny bulgot wody .
□	6	bibliografia selektywna znajduje się w wydaniu :
□	4	Ich mózg się zawiesza .



Corpus-specific settings for InterCorp v16ud - Polish

□ <#>
□ Token number

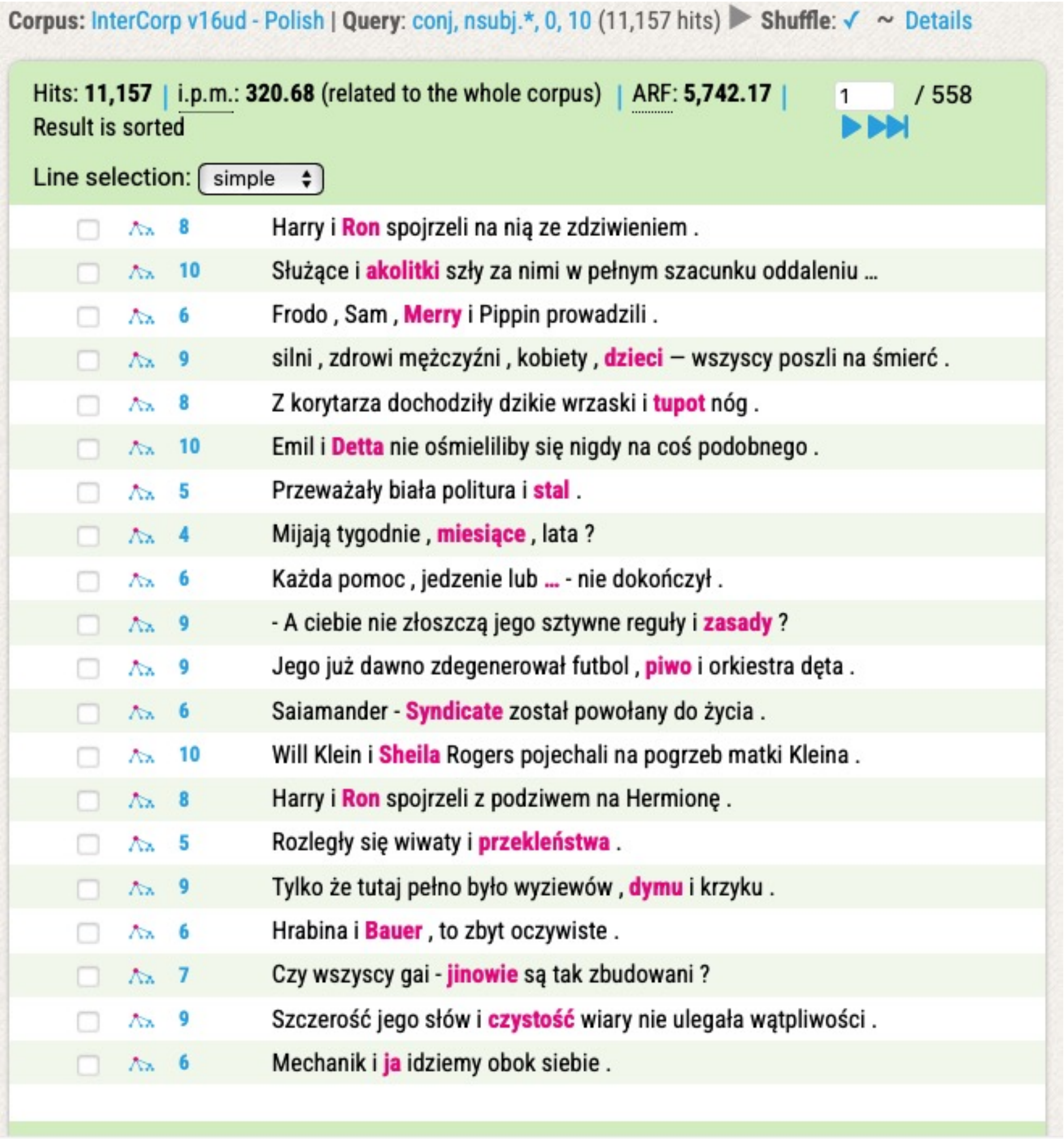
□ <doc>
□ Document number
□ doc.id
□ doc.tag_model

□ <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

☑ <s>
□ s.id
□ s.maxNPDepth
□ s.subRatio
☑ s.sLength
□ s.maxNPLength
□ s.mdd
□ s.maxTreeDepth

```
[deprel="conj" & p_deprel="nsubj.*"]
within
<s maxTreeDepth="0" & sLength <= "10" />
```



Corpus: InterCorp v16ud - Polish | Query: conj, nsubj.*, 0, 10 (11,157 hits) ▶ Shuffle: ✓ ~ Details

Hits: 11,157 | i.p.m.: 320.68 (related to the whole corpus) | ARF: 5,742.17 | 1 / 558

Result is sorted

Line selection: simple

☐	8	Harry i **Ron** spojrzeli na nią ze zdziwieniem .
☐	10	Służące i **akolitki** szły za nimi w pełnym szacunku oddaleniu ...
☐	6	Frodo , Sam , **Merry** i Pippin prowadzili .
☐	9	silni , zdrowi mężczyźni , kobiety , **dzieci** — wszyscy poszli na śmierć .
☐	8	Z korytarza dochodziły dzikie wrzaski i **tupot** nóg .
☐	10	Emil i **Detta** nie ośmieliliby się nigdy na coś podobnego .
☐	5	Przeważały biała politura i **stal** .
☐	4	Mijają tygodnie , **miesiące** , lata ?
☐	6	Każda pomoc , jedzenie lub **...** - nie dokończył .
☐	9	- A ciebie nie złoszczą jego sztywne reguły i **zasady** ?
☐	9	Jego już dawno zdegenerował futbol , **piwo** i orkiestra dęta .
☐	6	Saiamander - **Syndicate** został powołany do życia .
☐	10	Will Klein i **Sheila** Rogers pojechali na pogrzeb matki Kleina .
☐	8	Harry i **Ron** spojrzeli z podziwem na Hermionę .
☐	5	Rozległy się wiwaty i **przekleństwa** .
☐	9	Tylko że tutaj pełno było wyziewów , **dymu** i krzyku .
☐	6	Hrabina i **Bauer** , to zbyt oczywiste .
☐	7	Czy wszyscy gai - **jinowie** są tak zbudowani ?
☐	9	Szczerość jego słów i **czystość** wiary nie ulegała wątpliwości .
☐	6	Mechanik i **ja** idziemy obok siebie .

Text-level complexity metrics

	Noun phrase	**Sentence**
horizontal dimension	maxNPLength**Avg** ***average*** *maximum length*	sLength**Avg** ***average*** *length in no. of words*
		subRatio**Avg** ***average*** *subordination ration*
vertical dimension	maxNPDepth**Avg** ***average*** *maximum depth*	maxTreeDepth**Avg** ***average*** *maximum tree depth*
cognitive load		mdd *mean dependency distance*

Text-level metrics of lexical diversity

- A variant of *type-token ratio*
- Number of different *types* in a moving window 1000 tokens wide
- Undefined if the text is shorter than 1000 tokens
- Average number of different *word forms*: lexDivWord
  - cs: 421–732, en: 350–563
- Average number of different *lexemes*: lexDivLemma
  - cs: 279–629, en: 281–494

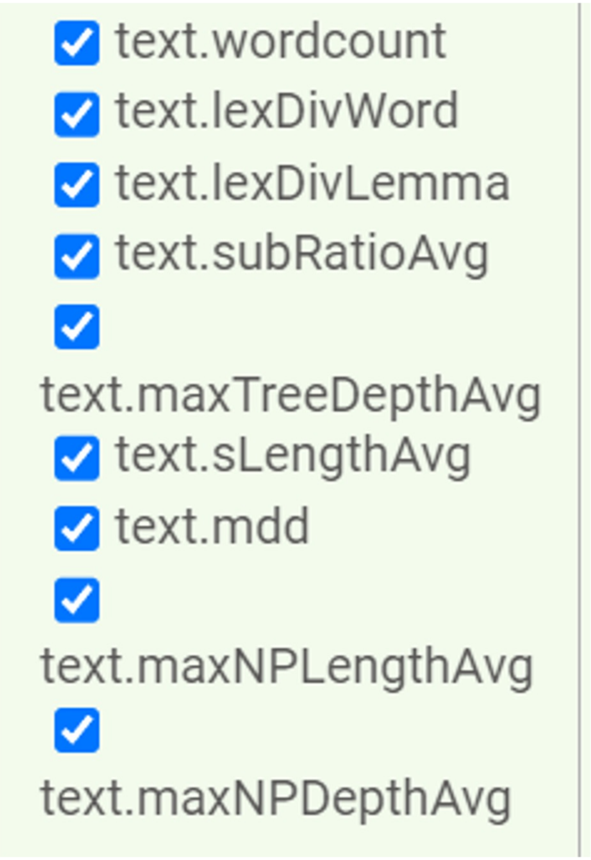
Displaying text-level metrics, downloading results

<text>

View > Corpus-specific settings > References > text.id, text.wordcount, text.lexDivWord, …

Apply View Options

Save > CSV/XLSX



☑ text.wordcount
☑ text.lexDivWord
☑ text.lexDivLemma
☑ text.subRatioAvg
☑ text.maxTreeDepthAvg
☑ text.sLengthAvg
☑ text.mdd
☑ text.maxNPLengthAvg
☑ text.maxNPDepthAvg

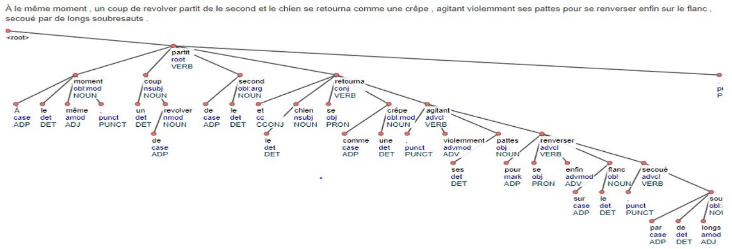
Outline

1. About InterCorp
2. Universal Dependencies (UD)
3. InterCorp with UD
4. Metrics of syntactic complexity and lexical diversity
5. **Using the metrics**
6. Perspectives, questions, discussion

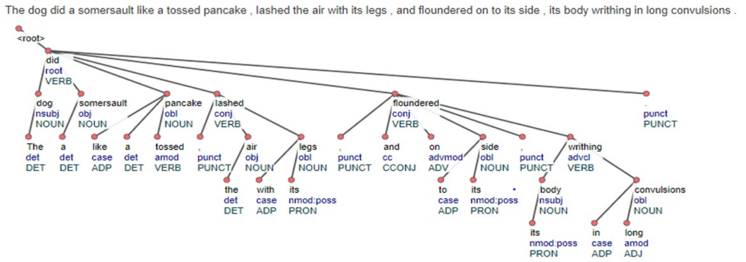
What is it good for?

- Teaching L1/L2
    - Filtering corpus examples
    - Building subcorpora for self-study
    - TODO: evaluation of learner texts online
- Contrastive / typological research of multiple languages
- Translatological research
- Research of text types variability
    - Comparison of metrics for sentences, texts, text types, languages
    - Correlation and comparison of metrics

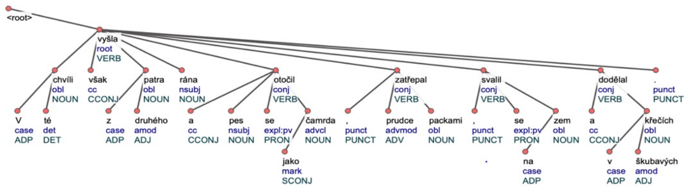
SubRatio a maxTreeDepth



À le même moment , un coup de revolver partit de le 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 .



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 .



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)

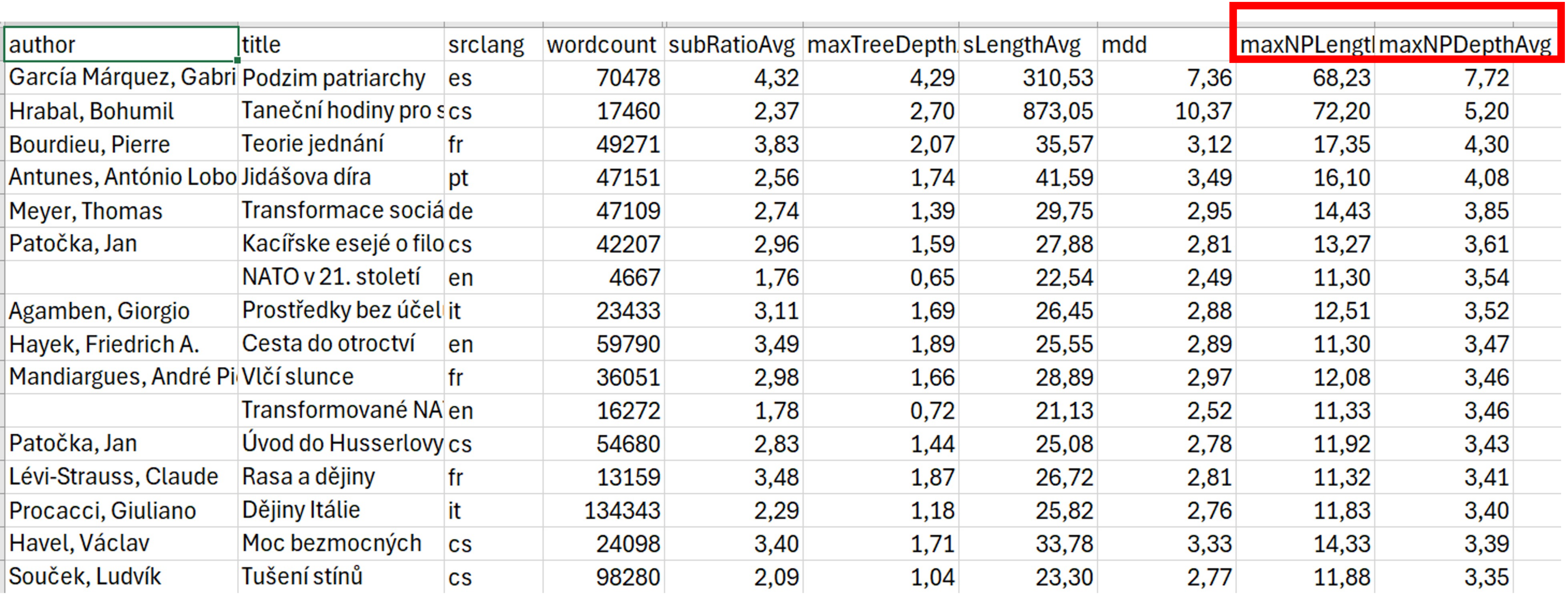
Sub.ratio = 1.33 ((3+1)/3)
Max.Tree.Depth = 1

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

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



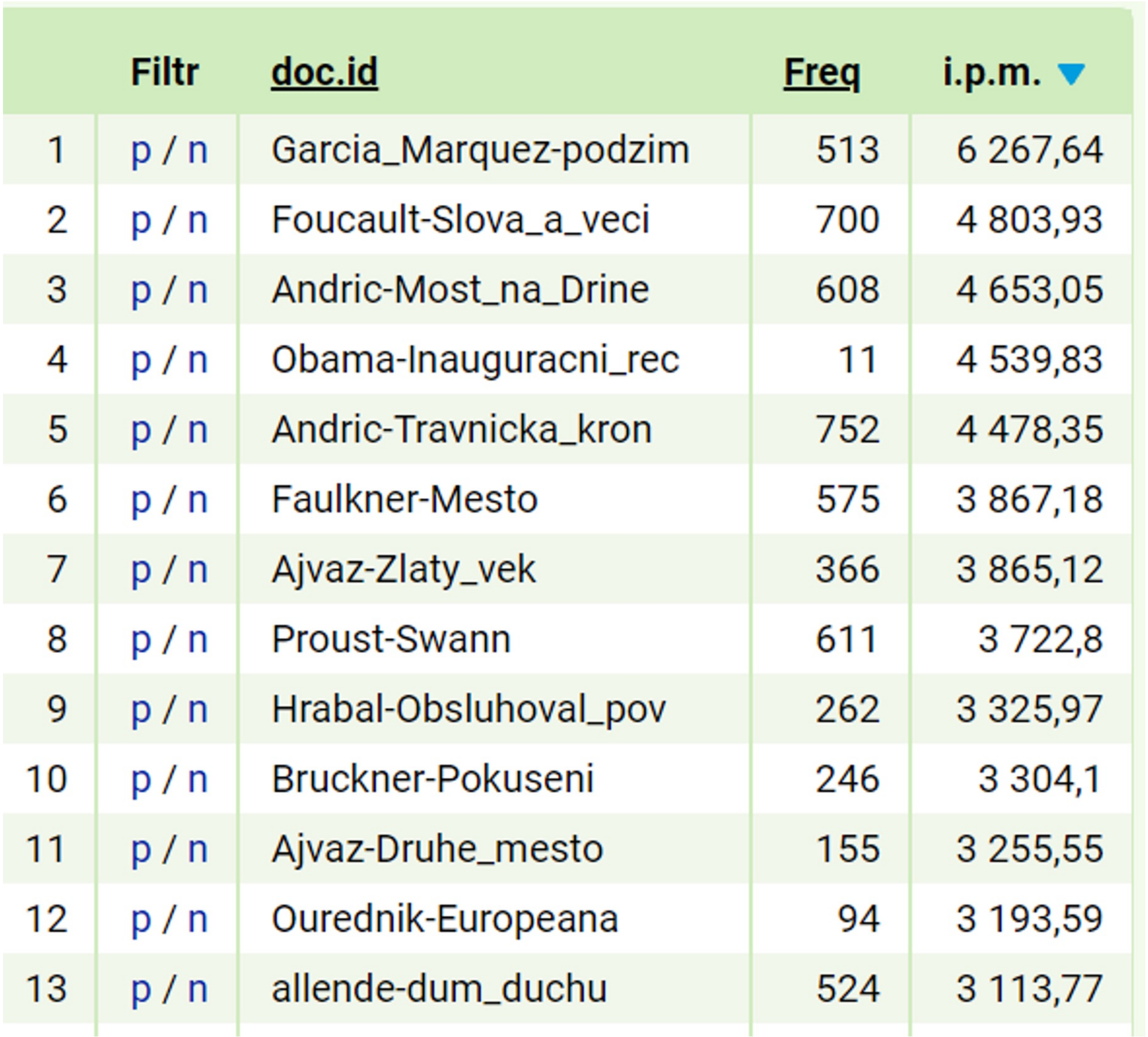
author	title	srclang	wordcount	subRatioAvg	maxTreeDepth	sLengthAvg	mdd	maxNPLengt	maxNPDepthAvg
García Márquez, Gabri	Podzim patriarchy	es	70478	4,32	4,29	310,53	7,36	68,23	7,72
Hrabal, Bohumil	Taneční hodiny pro s	cs	17460	2,37	2,70	873,05	10,37	72,20	5,20
Bourdieu, Pierre	Teorie jednání	fr	49271	3,83	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

Text-level metrics (NP)

text.maxNPDepthAvg
text.maxNPLengthAvg

Any relation with coordinated relative clauses?

[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

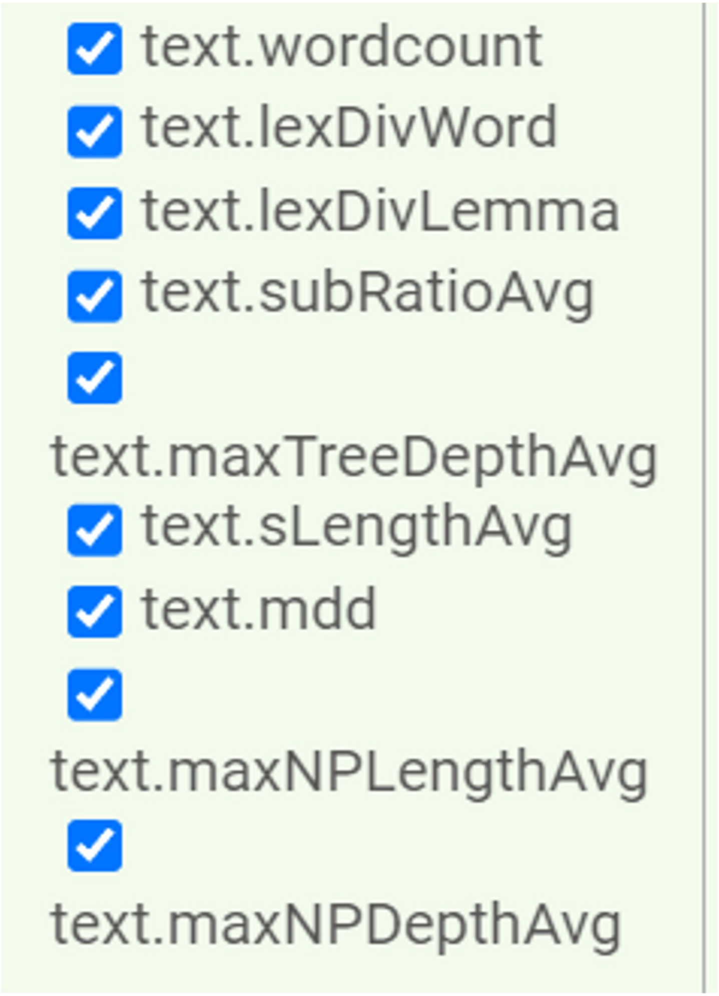
Text-level metrics

text.subRatioAvg

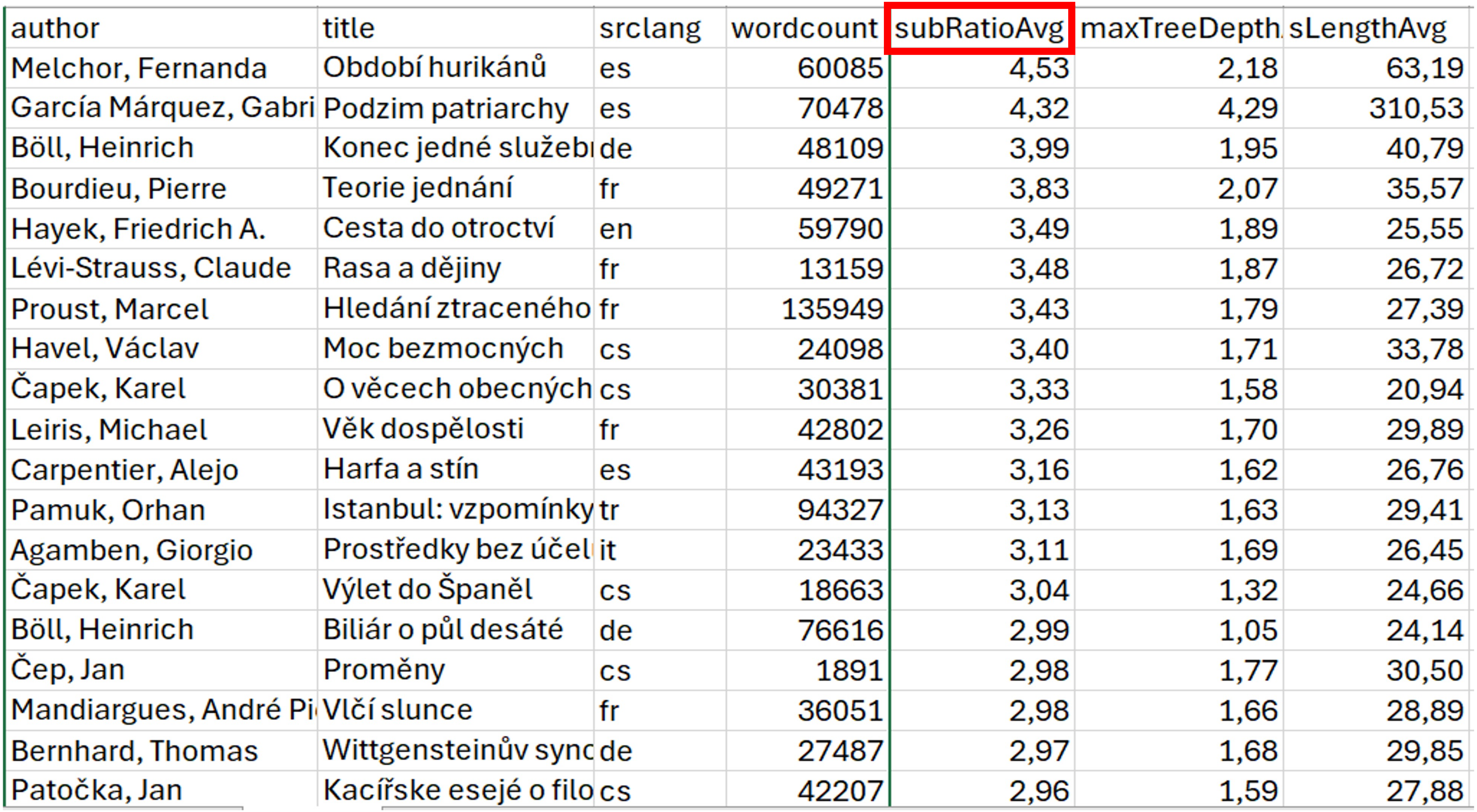
text.maxTreeDepthAvg

MAX for a language?

- Source language?
- Author?
- Text type? (fiction, non-fiction, poetry, drama…)



☑ text.wordcount
☑ text.lexDivWord
☑ text.lexDivLemma
☑ text.subRatioAvg
☑ text.maxTreeDepthAvg
☑ text.sLengthAvg
☑ text.mdd
☑ text.maxNPLengthAvg
☑ text.maxNPDepthAvg

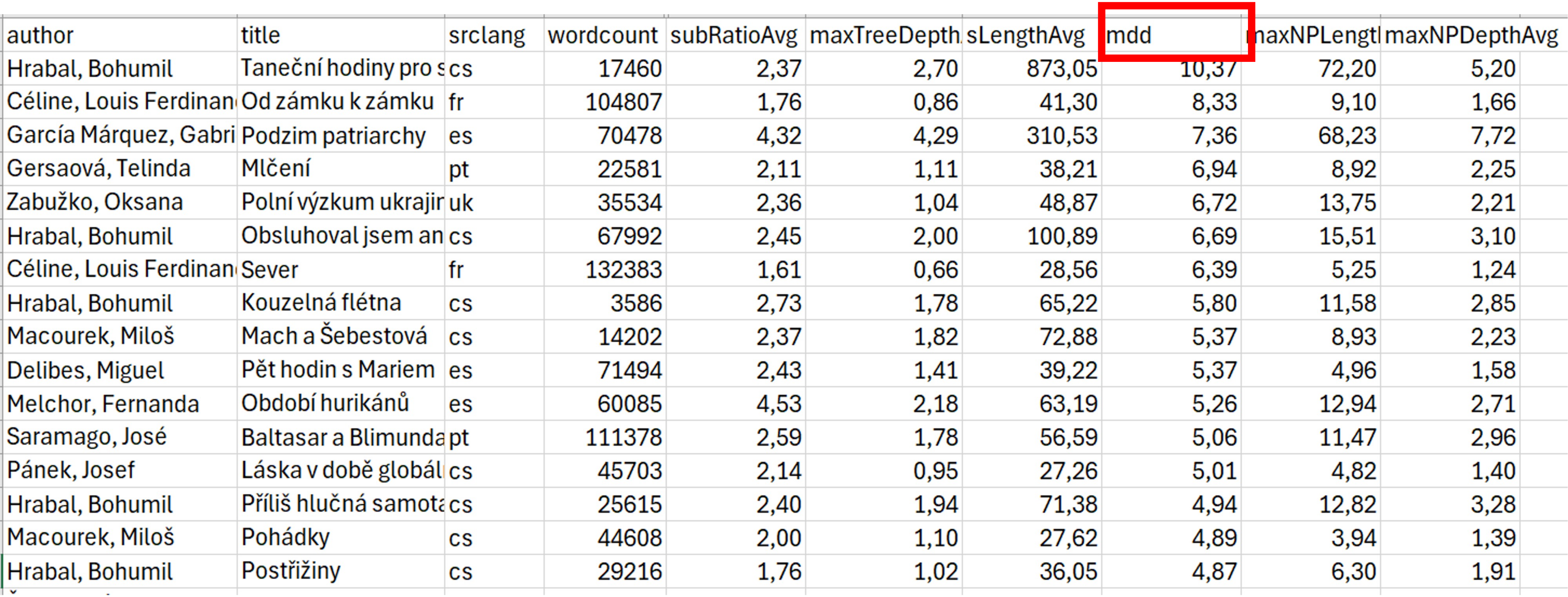


author	title	srclang	wordcount	subRatioAvg	maxTreeDepth	sLengthAvg
Melchor, Fernanda	Období hurikánů	es	60085	4,53	2,18	63,19
García Márquez, Gabri	Podzim patriarchy	es	70478	4,32	4,29	310,53
Böll, Heinrich	Konec jedné služeb	de	48109	3,99	1,95	40,79
Bourdieu, Pierre	Teorie jednání	fr	49271	3,83	2,07	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			1,20	0,21	4,20
Venclova, Tomas	Čas rozpůlil se…			1,20	0,22	6,22
Topol, Josef	Kočka na kolejíc			1,19	0,20	4,33
Ābele, Inga	Ostřice			1,19	0,19	4,13
Goscinny, René; Uderz	Asterix a cesta k			1,19	0,20	4,10
Sofokles	Antigoné			1,19	0,21	4,75
Šotola, Jiří	Podzim v zahrad			1,19	0,19	6,32
Čapek, Karel	Věc Makropulos			1,18	0,18	3,69
Arriaga, Guillermo	Psí lásky			1,18	0,20	4,81
Jarry, Alfred	Ubu			1,18	0,18	4,68
Karvaš, Peter	Antigona a ti druz			1,17	0,16	3,46
Karvaš, Peter	Půlnoční mše			1,17	0,24	5,80
Biebl, Konstantín	Nový Ikaros			1,17	0,16	5,06
Krynicki, Ryszard	Kámen, jinovatka			1,17	0,15	4,44
	Historie města B			1,15	0,19	10,83
Pešková, Vlastimila	Biologie člověka	cs	18634	1,14	0,13	10,81
Fischerová, Daniela	Hodina mezi psem a	cs	17799	1,12	0,13	4,29
Rázusová-Martáková, I	Zatoulané house	sk	170	1,11	0,19	6,54

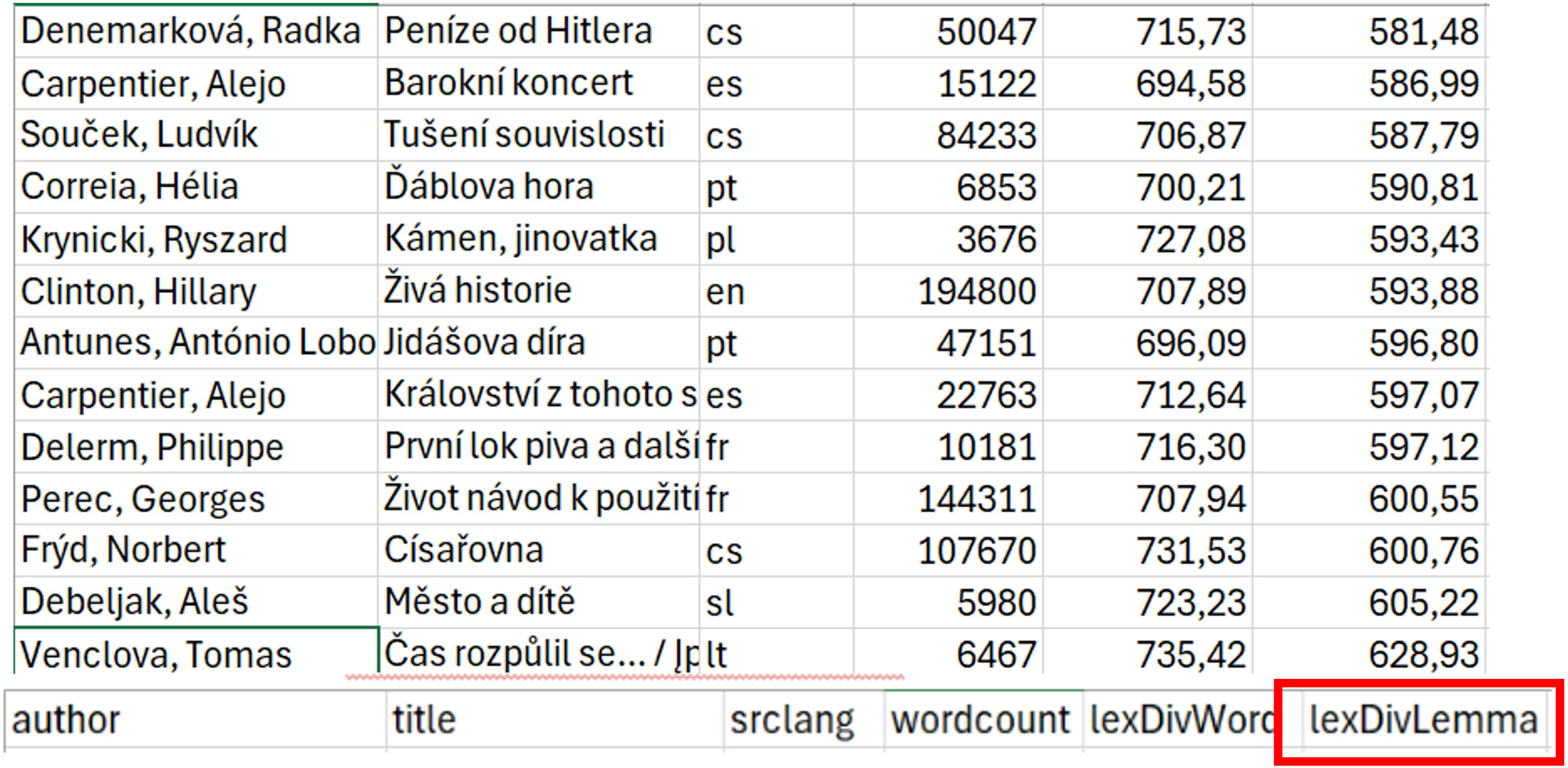
Housata šla dlouhou řadou,
brouzdala se v ranní rose,
a hned kousek za zahradou
jedno house ztratilo se.

mdd – mean dependency distance (MAX)



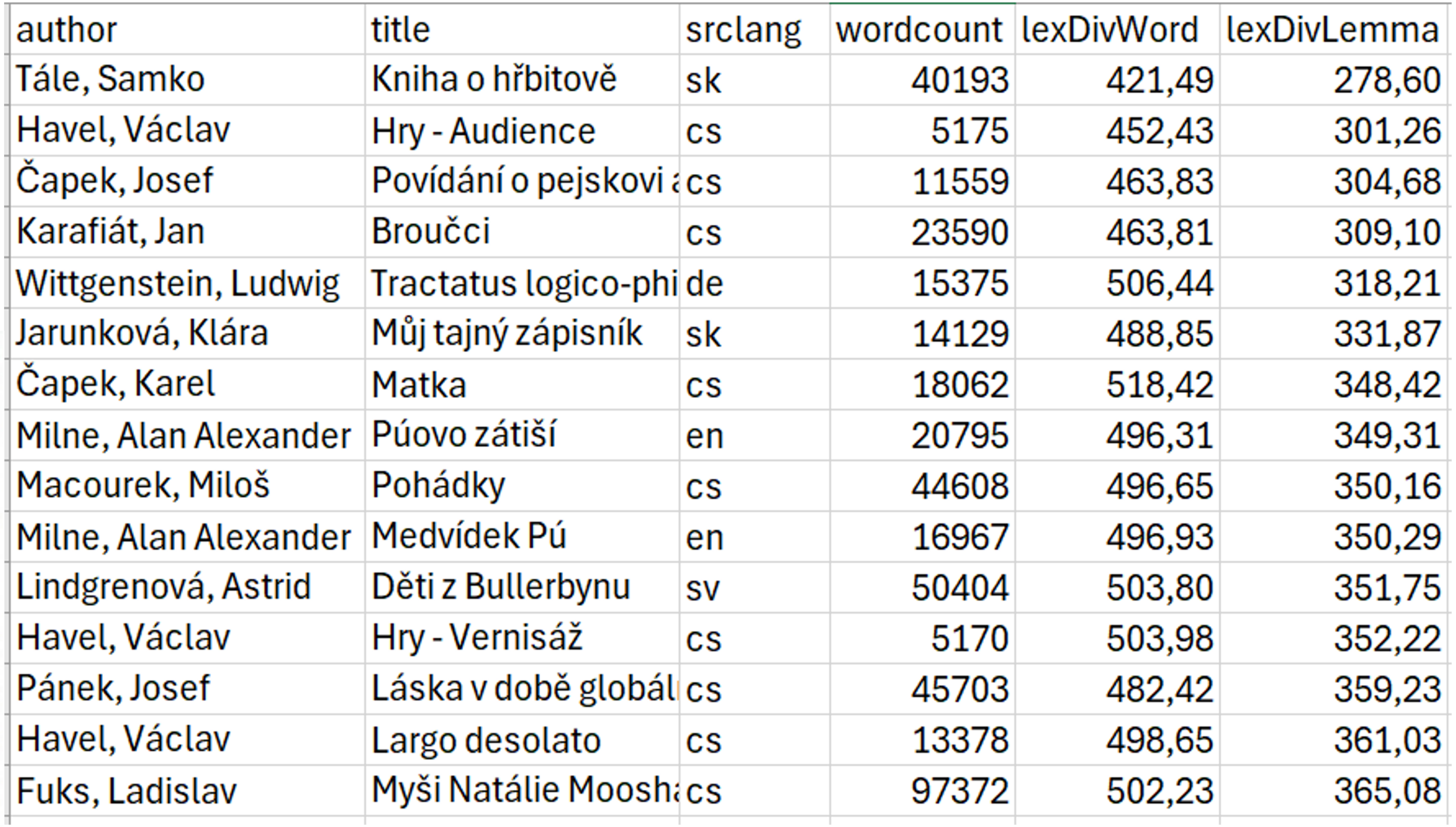
author	title	srclang	wordcount	subRatioAvg	maxTreeDepth	sLengthAvg	mdd	maxNPLengt	maxNPDepthAvg
Hrabal, Bohumil	Taneční hodiny pro s	cs	17460	2,37	2,70	873,05	10,37	72,20	5,20
Céline, Louis Ferdinan	Od zámku k zámku	fr	104807	1,76	0,86	41,30	8,33	9,10	1,66
García Márquez, Gabri	Podzim patriarchy	es	70478	4,32	4,29	310,53	7,36	68,23	7,72
Gersaová, Telinda	Mlčení	pt	22581	2,11	1,11	38,21	6,94	8,92	2,25
Zabužko, Oksana	Polní výzkum ukrajir	uk	35534	2,36	1,04	48,87	6,72	13,75	2,21
Hrabal, Bohumil	Obsluhoval jsem an	cs	67992	2,45	2,00	100,89	6,69	15,51	3,10
Céline, Louis Ferdinan	Sever	fr	132383	1,61	0,66	28,56	6,39	5,25	1,24
Hrabal, Bohumil	Kouzelná flétna	cs	3586	2,73	1,78	65,22	5,80	11,58	2,85
Macourek, Miloš	Mach a Šebestová	cs	14202	2,37	1,82	72,88	5,37	8,93	2,23
Delibes, Miguel	Pět hodin s Mariem	es	71494	2,43	1,41	39,22	5,37	4,96	1,58
Melchor, Fernanda	Období hurikánů	es	60085	4,53	2,18	63,19	5,26	12,94	2,71
Saramago, José	Baltasar a Blimunda	pt	111378	2,59	1,78	56,59	5,06	11,47	2,96
Pánek, Josef	Láska v době globál	cs	45703	2,14	0,95	27,26	5,01	4,82	1,40
Hrabal, Bohumil	Příliš hlučná samota	cs	25615	2,40	1,94	71,38	4,94	12,82	3,28
Macourek, Miloš	Pohádky	cs	44608	2,00	1,10	27,62	4,89	3,94	1,39
Hrabal, Bohumil	Postřižiny	cs	29216	1,76	1,02	36,05	4,87	6,30	1,91

lexDivLemma – lexical diversity (MAX)



author	title	srclang	wordcount	lexDivWord	lexDivLemma
Denemarková, Radka	Peníze od Hitlera	cs	50047	715,73	581,48
Carpentier, Alejo	Barokní koncert	es	15122	694,58	586,99
Souček, Ludvík	Tušení souvislosti	cs	84233	706,87	587,79
Correia, Hélia	Ďáblova hora	pt	6853	700,21	590,81
Krynicki, Ryszard	Kámen, jinovatka	pl	3676	727,08	593,43
Clinton, Hillary	Živá historie	en	194800	707,89	593,88
Antunes, António Lobo	Jidášova díra	pt	47151	696,09	596,80
Carpentier, Alejo	Království z tohoto s	es	22763	712,64	597,07
Delerm, Philippe	První lok piva a další	fr	10181	716,30	597,12
Perec, Georges	Život návod k použití	fr	144311	707,94	600,55
Frýd, Norbert	Císařovna	cs	107670	731,53	600,76
Debeljak, Aleš	Město a dítě	sl	5980	723,23	605,22
Venclova, Tomas	Čas rozpůlil se... / Jp	lt	6467	735,42	628,93

Lexical diversity – lexDivLemma (MIN)



author	title	srclang	wordcount	lexDivWord	lexDivLemma
Tále, Samko	Kniha o hřbitově	sk	40193	421,49	278,60
Havel, Václav	Hry - Audience	cs	5175	452,43	301,26
Čapek, Josef	Povídání o pejskovi a	cs	11559	463,83	304,68
Karafiát, Jan	Broučci	cs	23590	463,81	309,10
Wittgenstein, Ludwig	Tractatus logico-phi	de	15375	506,44	318,21
Jarunková, Klára	Můj tajný zápisník	sk	14129	488,85	331,87
Čapek, Karel	Matka	cs	18062	518,42	348,42
Milne, Alan Alexander	Púovo zátiší	en	20795	496,31	349,31
Macourek, Miloš	Pohádky	cs	44608	496,65	350,16
Milne, Alan Alexander	Medvídek Pú	en	16967	496,93	350,29
Lindgrenová, Astrid	Děti z Bullerbynu	sv	50404	503,80	351,75
Havel, Václav	Hry - Vernisáž	cs	5170	503,98	352,22
Pánek, Josef	Láska v době globál	cs	45703	482,42	359,23
Havel, Václav	Largo desolato	cs	13378	498,65	361,03
Fuks, Ladislav	Myši Natálie Moosha	cs	97372	502,23	365,08

Explaining the differences

Stylistic:

les disparités opposant [deprel=acl] *les classes populaires et les classes moyennes*

rozdíly mezi [deprel=case] *lidovou a střední vrstvou*

Normalization:

J'ai bu. J'ai eu alors envie de fumer.

Vypil jsem ji a dostal jsem chuť si zakouřit.

Differences in annotation:

next slide…

Categorial differences (linguistic traditions)

FR participles [deprel=**acl**] … **+clause**

FR:
des formes dérivées [deprel=**acl**] *des idées suprêmes du Bien*

EN like FR:
forms derived [deprel=acl] *from the utmost ideas of Good*

PL like FR and EN:
argumenty przytoczone [deprel=acl] *przez Platona*

CS participles [deprel=**amod**] … **–clause**

CS:
formy odvozené [deprel=**amod**] *od nejzazší ideje Dobra*

argumenty odvozené [deprel=**amod**] *z Platónova naturalismu*

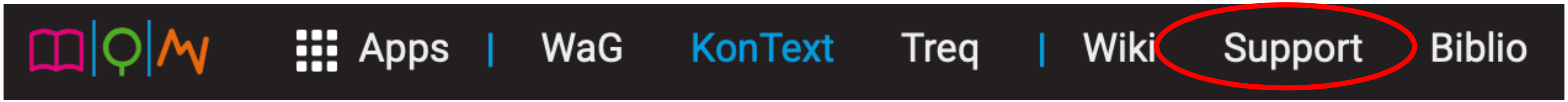
CONSEQUENCE: false differences in subRatio. BUT: only 5% clauses are **acl.**

Outline

1. About InterCorp
2. Universal Dependencies (UD)
3. InterCorp with UD
4. Metrics of syntactic complexity and lexical diversity
5. Using the metrics
6. Perspectives, questions, discussion

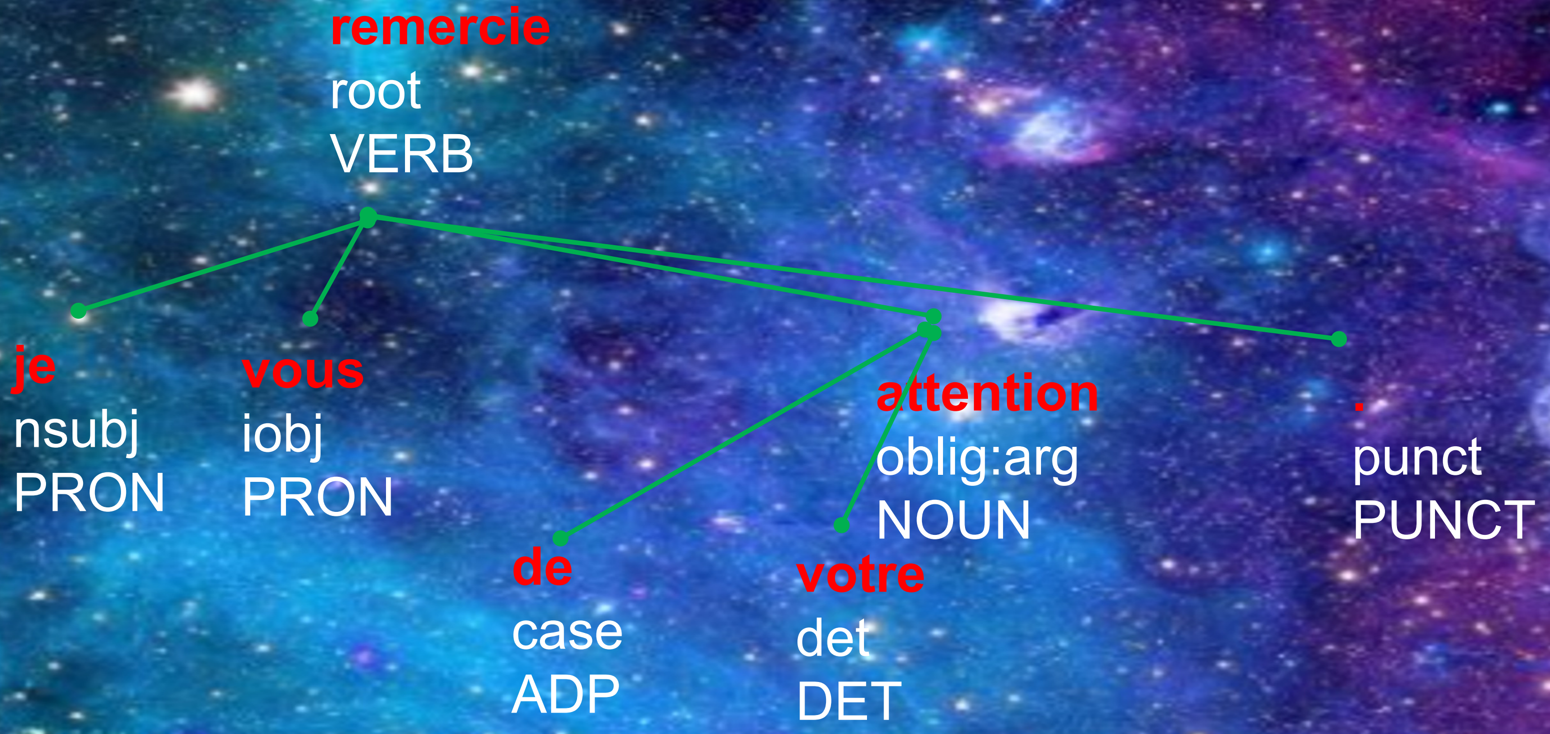
Perspectives – what next?

- Documentation (wiki)
- Whole corpus (all text types)
- Workshop on Biennial of Czech Linguistics (17–20 Sep 2024)
- Fixing bugs – please report
- Contrastive / typological research



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

Grazie mille della vostra attenzione.
Labai dėkoju už dėmesį.
Liels paldies par uzmanību.
Dank u zeer voor uw aandacht.
Dziękuję bardzo Państwu za uwagę.
Muito obrigado pela vossa atenção.
非常感您的注。
Veľmi pekne vám ďakujem za pozornosť.
Najlepša hvala za vašo pozornost.
Tack så mycket för er uppmärksamhet.
Mange tak for Deres opmærksomhed.
Vielen Dank für Ihre Aufmerksamkeit.
Thank you very much for your attention.
Muchísimas gracias por su atención.
Suur tänu tähelepanu eest.
ご清聴ありがとうございました。
Oikein paljon kiitoksia mielenkiinnostanne.
Je vous remercie de votre attention.
Nagyon szépen köszönöm a figyelmüket.
Velice vám děkuji za pozornost.



remercie
root
VERB

je
nsubj
PRON

vous
iobj
PRON

de
case
ADP

votre
det
DET

attention
oblig:arg
NOUN

.
punct
PUNCT

Pytania

Dyskusja

?

References, resources

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.

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.

Kim Gerdes, 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

Jan Macutek, Radek Cech, and Jiri Milicka. 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.

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.

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

Nádvorníková, Olga and 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

Joakim Nivre, 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.

Przepiórkowski, A. and Patejuk, A. 2019. Nested coordination in Universal Dependencies. In Alexandre Rademaker and Francis Tyers, editors, *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 58–69. Association for Computational Linguistics, 2019.

Przepiórkowski, A., Borysiak, M. and Głowacki, A. 2024. An argument for symmetric coordination from Dependency Length Minimization: A replication study. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors,), pages 1021–1033, Torino, Italy, 2024. ELRA and ICCL. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024*

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.

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.

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

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

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

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.

WEB:

https://universaldependencies.org/guidelines.html

Lindat UD Corpora (online search): https://lindat.mff.cuni.cz/services/kontext/corpora/corplist

Lindat UDPipe: https://lindat.mff.cuni.cz/services/udpipe/

Daniel Zeman: Universal Dependencies and the Slavic Languages. Warszawa, 19.11.2018.

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.