Lexical and syntactic variability of languages and text genres – a corpus-based study

Alexandr Rosen
Instytut Czeskiego Korpusu Narodowego
Wydział Filozoficzny Uniwersytetu Karola, Praga Czeska

Seminarium *Przetwarzanie języka naturalnego*
Zespół Inżynierii Lingwistycznej
Instytut Podstaw Informatyki Polskiej Akademii Nauk

Warszawa, 13 października 2024

Outline

1. Linguistic variation within and across languages
2. Metrics of syntactic complexity and lexical diversity
3. The data: InterCorp – a multilingual parallel corpus
4. Accessing the annotation via search interface
5. Using the metrics
6. Perspectives, questions, discussion
7. References

Link to this presentation

- https://drive.google.com/file/d/1PE87XSeWwe8pBqLMA5cJUZgLnLvYKB3-/view?usp=sharing

- https://shorturl.at/SbfPP

Outline

1. Linguistic variation within and across languages
2. Metrics of syntactic complexity and lexical diversity
3. The data: InterCorp – a multilingual parallel corpus
4. Accessing the annotation via search interface
5. Using the metrics
6. Perspectives, questions, discussion
7. References

Why study variation?

- Languages differ ☺
  - Dialects?
  - Language development?
- Within a language:
  - genres, registers (Czech, Arabic, Japanese, Javanese)
  - chat vs. formal address, spoken vs. written, L1 vs. L2, original vs. translated, …
  - also within a single text
- Variation at all levels
- Useful for:
  - L1, L2 learning
  - language technology
  - sociolinguistics, psycholinguistics, language disorders, forensic linguistics
  - linguistic typology, translation studies
  - linguistics in general

Some ways to explore variation

- Linguistic complexity
- Lexical diversity

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 (or other constituents)
- their hierarchy within the sentence

Simplifying complexity

- Complexity is **multi-dimensional,** thus more metrics should be combined (Biber, Larsson & Hancock 2023).
- Metrics are **sensitive** to genre and language.
- Metrics assume (morpho)syntactic **annotation**, consistent across languages

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

Why? Studying syntactic complexity is useful for:

- **Language development** (Givón 2009:4)
- **Monolingual studies** (Mačutek, Čech & Milička 2019; Hudelot 1980; Biber, Larsson & Hancock 2023; …)
- **Contrastive studies** *clause-linking* (Lehmann 1988), *clause-combining* (Cosme 2006, …), *information packaging* (Solfjeld 1996, Fabricius-Hansen 1999), *UD shared tasks* (Berdicevskis et al. 2018, …)
- **Translation studies** (Izquierdo & Marco 2000; Canavese & Mori 2021); *comparable or parallel corpora* (*translation universals* – simplification, *normalisation,* …)
- ***Register variation*** *spoken/written* (Beaman 1984; …), *academic* (Biber & Gray 2017; …)
- ***Typology*** (Levshina 2019, 2021 – *Leipzig Corpora Collection,* comparable, UD)
- **Readability** (Kincaid et al. 1975; Dell'Orletta et al. 2011; Gruszczyński & Ogrodniczuk 2015 ***Jasnopis***).
- **Language acquisition, proficiency assessment** – L1 & L2 (Lu 2010; …)

Lexical diversity

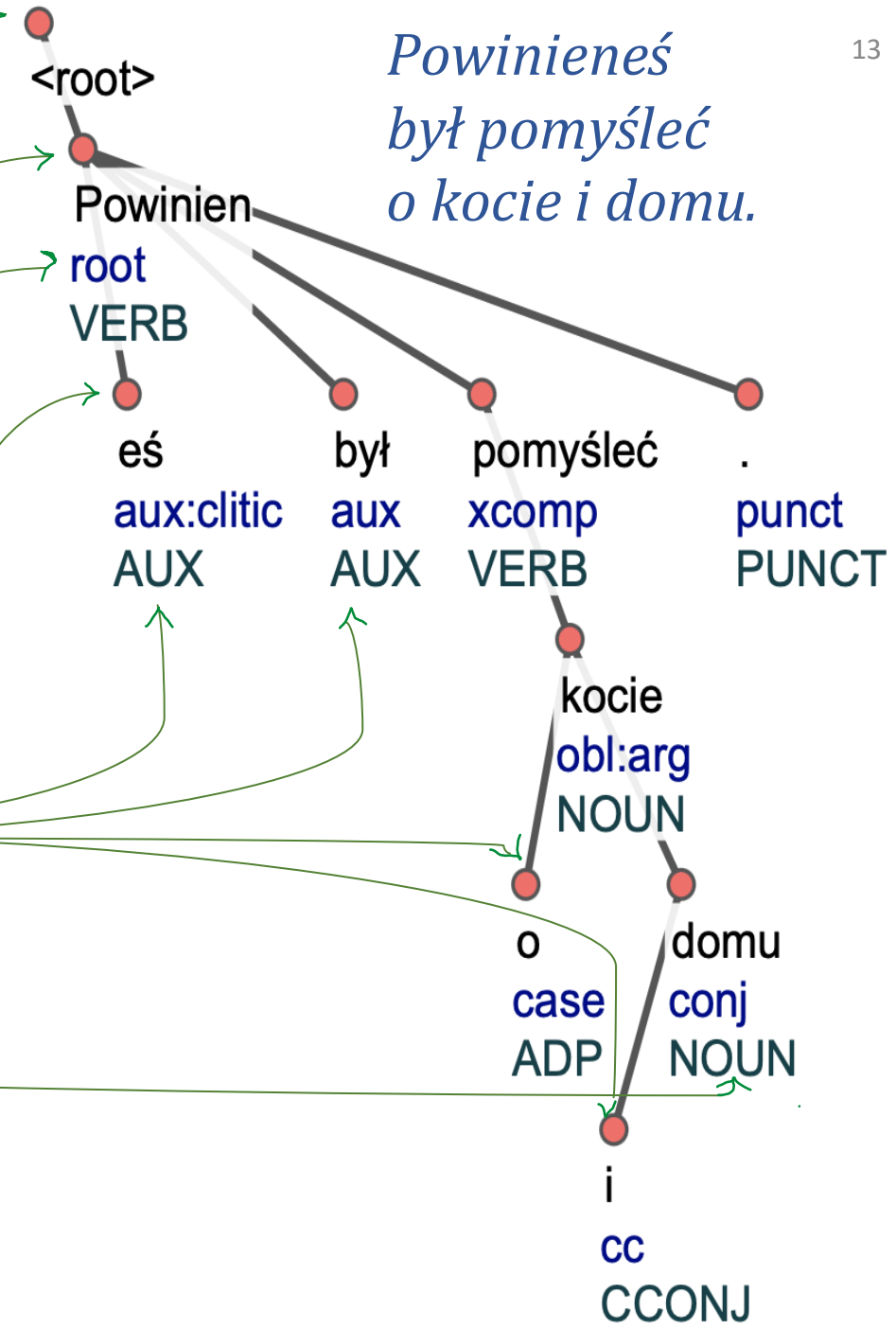
- The more **varied vocabulary** a text possesses, the higher lexical diversity. For a text to be highly lexically diverse, the speaker or writer has to use many different words, with little repetition of the words already used. (Johansson 2008: 62)
- **Type-token ratio (TTR)**: but longer texts give lower TTR and vice versa

Outline

1. Linguistic variation within and across languages
2. Metrics of syntactic complexity and lexical diversity
3. The data: InterCorp – a multilingual parallel corpus
4. Accessing the annotation via search interface
5. Using the metrics
6. Perspectives, questions, discussion
7. References

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



Universal Dependencies (UD)

https://universaldependencies.org

Linguistic categories in UD

- 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`

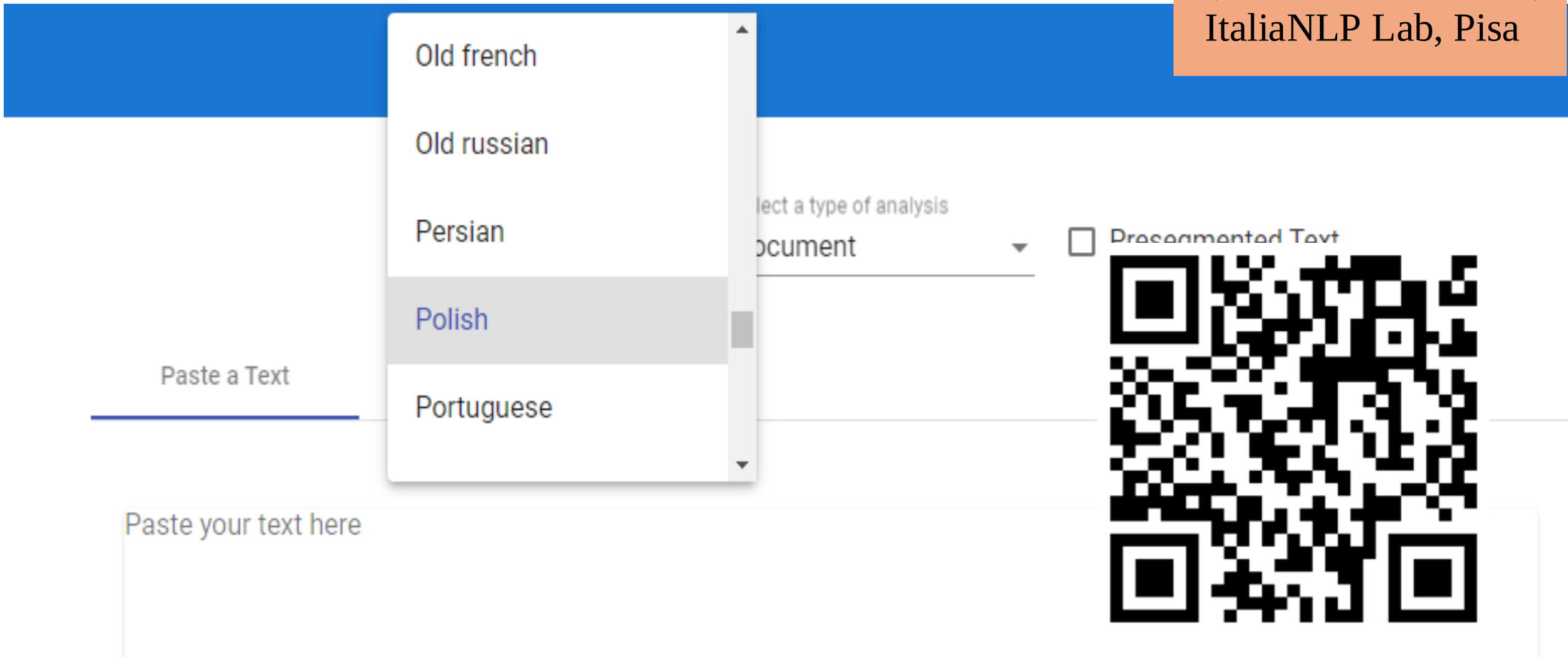
  https://universaldependencies.org/u/feat/index.html

More on UD in InterCorp:
https://wiki.korpus.cz/doku.php/en:pojmy:ud

A tool for computing metrics on line:

Profiling-UD http://linguistic-profiling.italianlp.it/

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



Metrics of syntactic complexity and lexical diversity in InterCorp v16ud

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

More about the metrics in InterCorp:
https://wiki.korpus.cz/doku.php/en:pojmy:syntakticka_komplexita

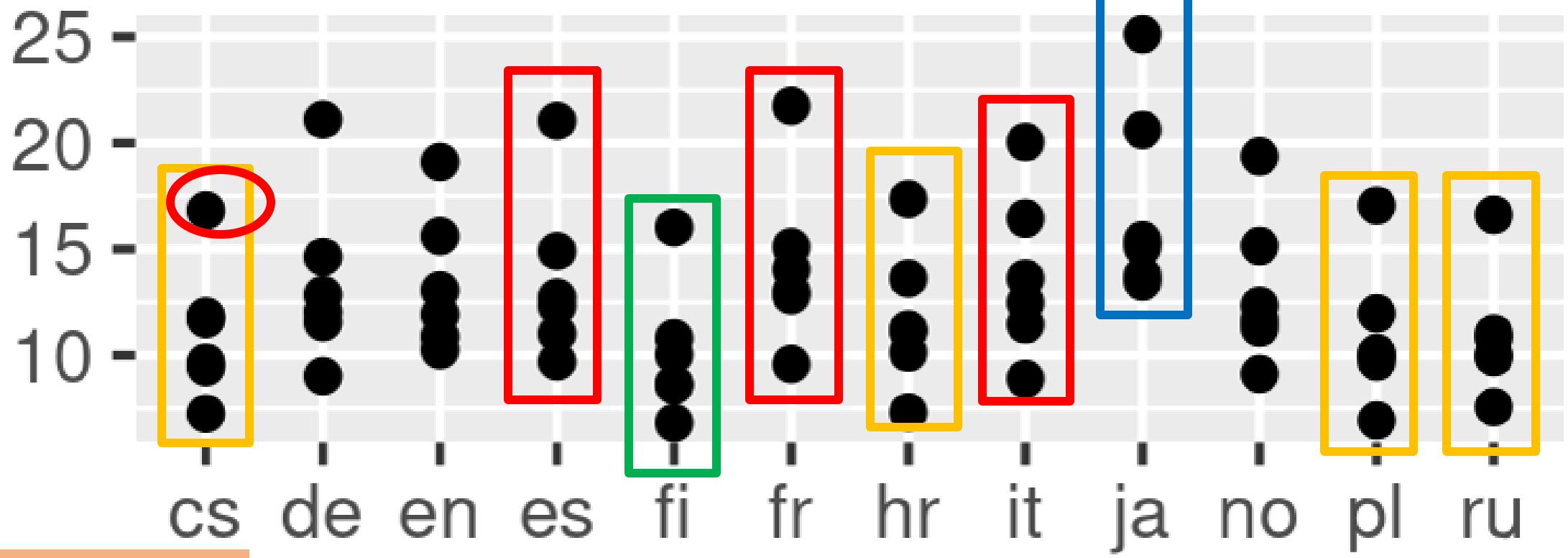
Sentence length across languages

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

Scatterplot of average sentence length

On 6 fiction texts in 12 languages

ja : tokenization



fiction: style matters

Metrics as metadata

Jak tak szli, dżdżownica otrząsnęła się ze swojego przerażenia.

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

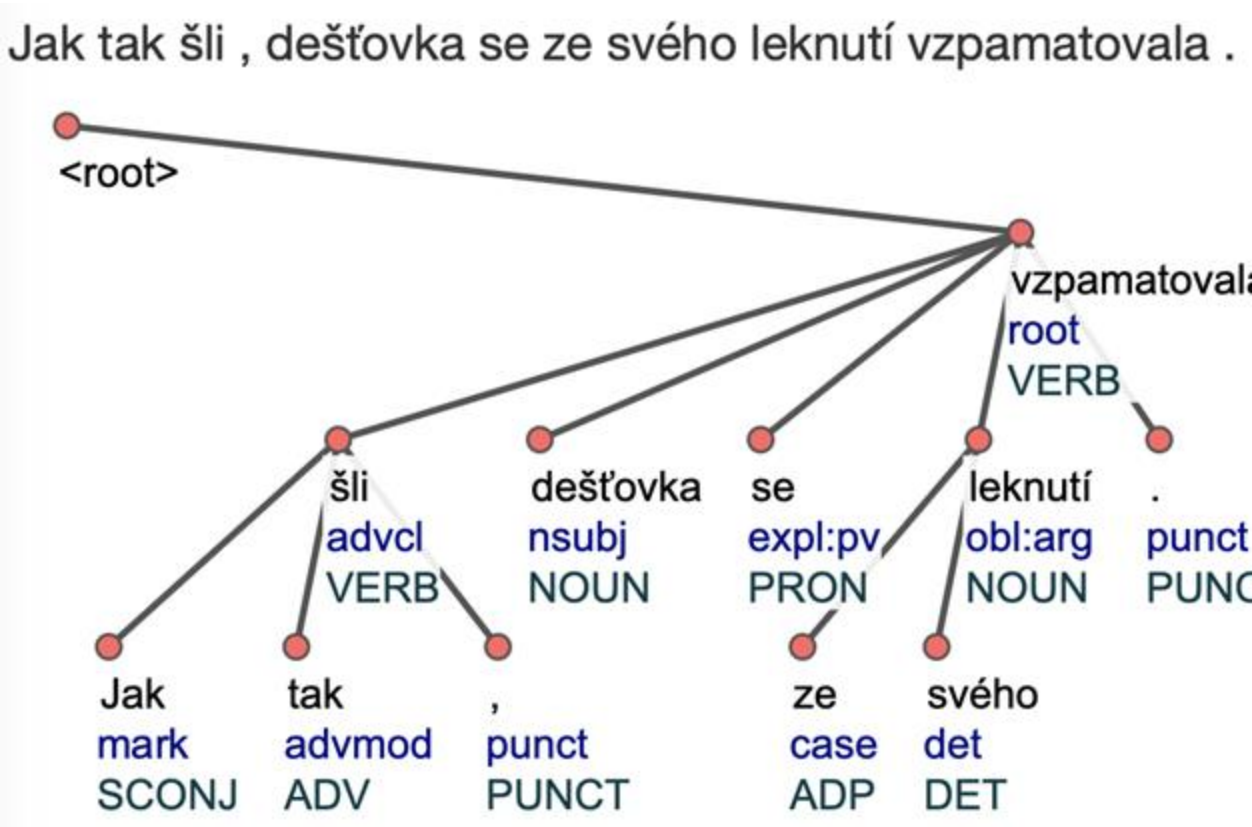
maxNPDepth=1

subRatio=2.0

maxTreeDepth=1

sLength=9

mdd=2.75 **>**

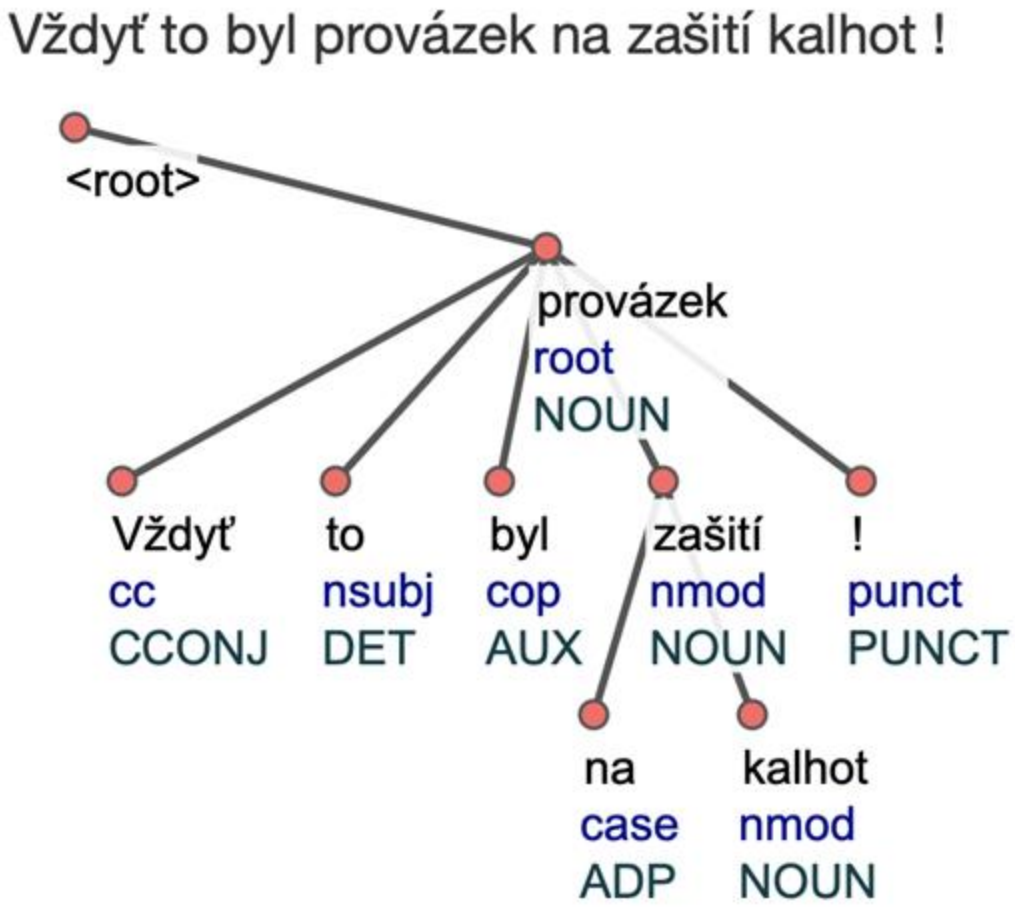


Sentence-level complexity metrics

	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 …*)



Przecież to był ***sznurek do zszycia spodni****!*

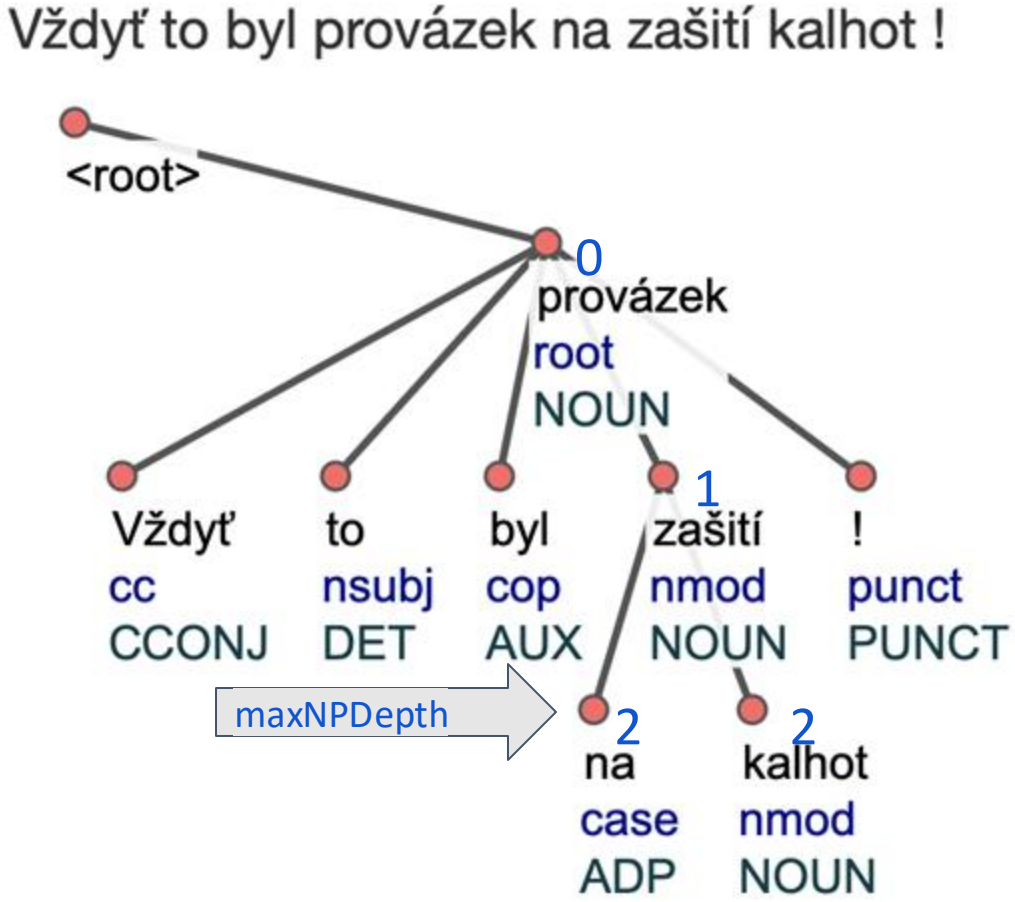
Noun phrase – complexity metrics

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



Przecież to był ***sznurek do zszycia spodni****!*

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*

Sentence – complexity metrics

sLength:

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

MaxTreeDepth:

- maximum number of clausal embeddings in the sentence
- coordination is skipped

subRatio:

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

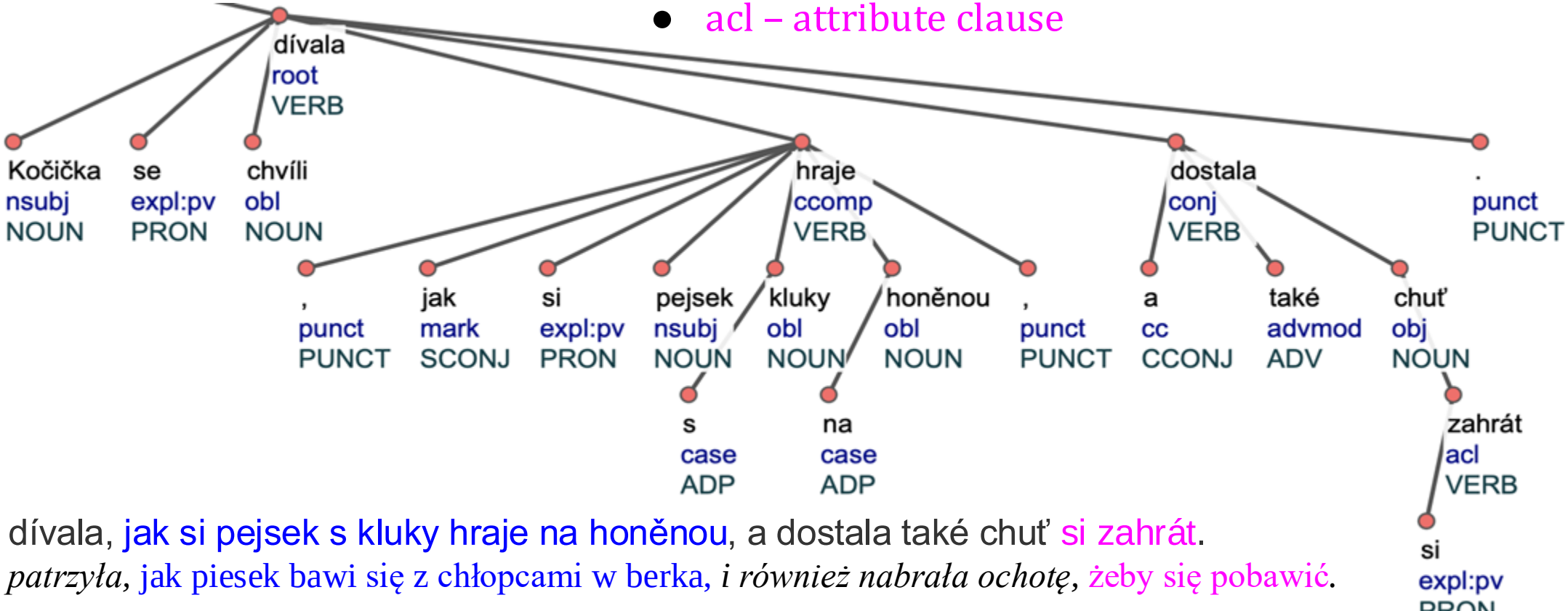
What is a sentence?

T-unit:

- main clause including all dependent clauses (Hunt 1965)
- each main clause conjunct counts

(Subordinate) clause, also 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.

Counting subRatio and maxTreeDepth

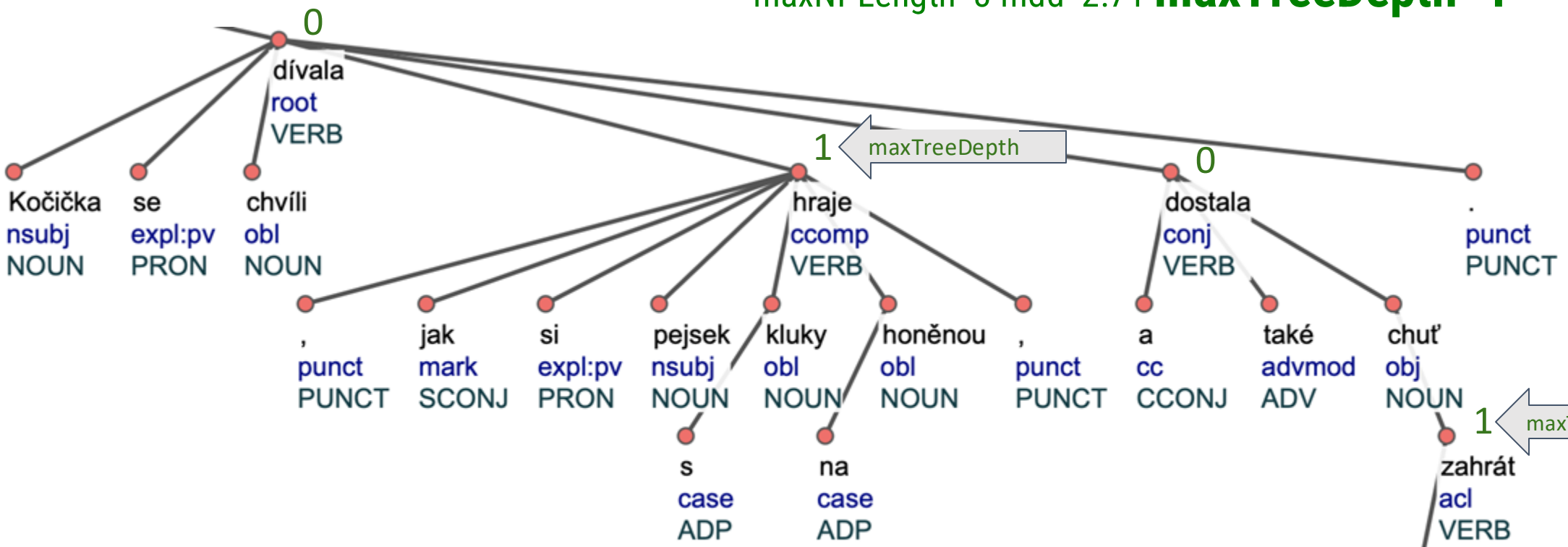
No. of T-units = 2

No. of clauses = 2

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

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

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

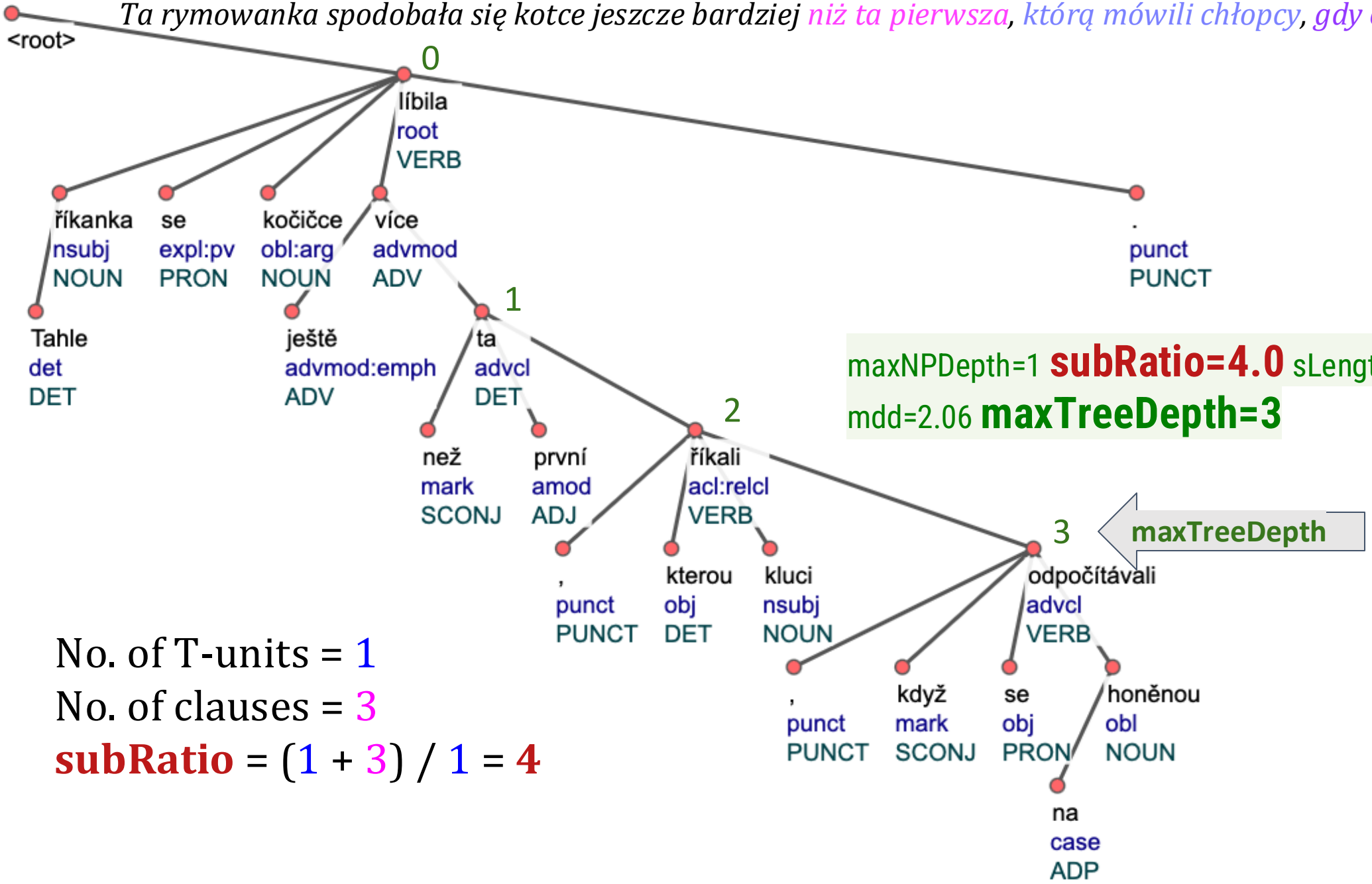


Kočička se chvíli dívala, jak si pejsek s kluky hraje na honěnou, a dostala také chuť si zahrát.

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

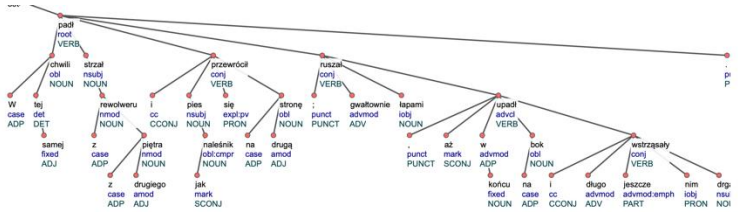
No. of clauses = 3

subRatio = (1 + 3) / 1 = **4**

SubRatio and maxTreeDepth at work

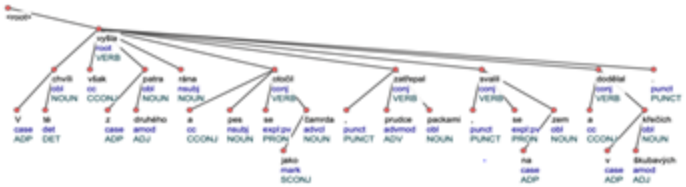
Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts.
(Albert Camus *La Peste*)

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



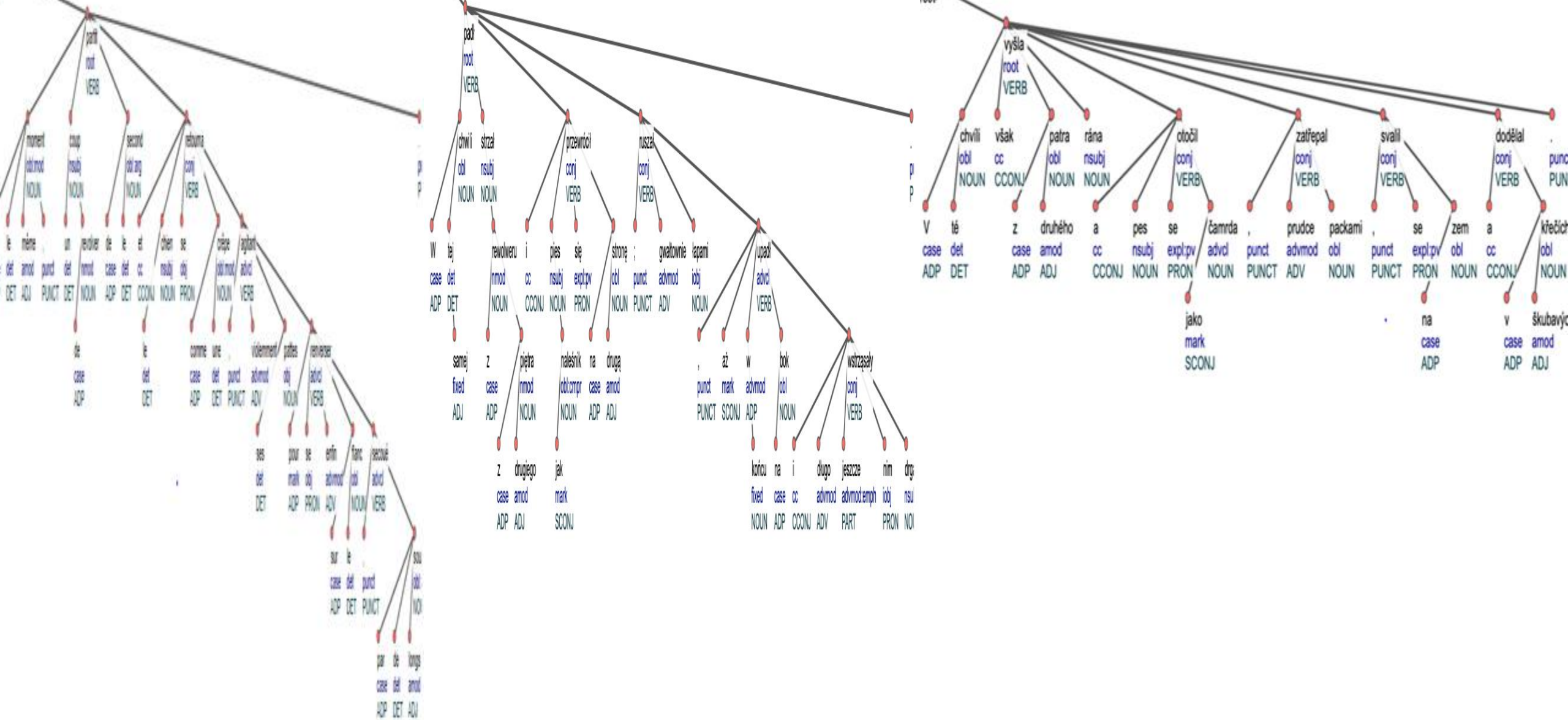
W tej samej chwili padł strzał z rewolweru z drugiego piętra i pies jak naleśnik przewrócił się na drugą stronę; ruszał gwałtownie łapami, aż w końcu upadł na bok i długo jeszcze wstrząsały nim drgawki
(*transl.* Joanna Guze)

Sub.ratio = 1.67 ((3+2)/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


partit
mot
VERB
moment
obl:mod
NOUN
le
det
DET
même
amod
ADJ
,
punct
PUNCT
un
det
DET
coup
nsubj
NOUN
revolver
nmod
NOUN
de
case
ADP
se
det
DET
le
case
ADP
le
det
DET
secondes
obl:arg
NOUN
et
cc
CCONJ
le
det
DET
retourna
conj
VERB
chien
nsubj
NOUN
se
obj
PRON
comme
case
ADP
une
det
DET
,
punct
PUNCT
chèque
obl:mod
NOUN
agitant
advcl
VERB
violemment
advmod
ADV
ses
det
DET
pour
mark
ADP
le
obj
PRON
enfin
advmod
ADV
sur
case
ADP
le
det
DET
,
punct
PUNCT
flanc
obl
NOUN
gauche
advcl
VERB
par
case
ADP
la
det
DET
longu
amod
ADJ
padł
root
VERB
chwili
obl
NOUN
strzał
nsubj
NOUN
W
case
ADP
tej
det
DET
samej
fixed
ADJ
rewolweru
nmod
NOUN
z
case
ADP
z
case
ADP
drugiego
amod
ADJ
pięćdziesięciu
nmod
NOUN
i
cc
CCONJ
pies
nsubj
NOUN
przewrócił
conj
VERB
się
expl:pv
PRON
naleśnik
obl:cmpr
NOUN
na
case
ADP
jak
mark
SCONJ
stronę
obl
NOUN
drugą
amod
ADJ
;
punct
PUNCT
runął
conj
VERB
gwałtownie
advmod
ADV
łapami
iobj
NOUN
,
punct
PUNCT
aż
mark
SCONJ
upadł
advcl
VERB
w
advmod
ADP
końcu
fixed
NOUN
na
case
ADP
bok
obl
NOUN
i
cc
CCONJ
długo
advmod
ADV
jeszcze
advmod:emph
PART
wstrząsały
conj
VERB
nim
iobj
PRON
drg
nsu
.
p
P
vyšla
root
VERB
chvíli
obl
NOUN
V
case
ADP
té
det
DET
však
cc
CCONJ
z
case
ADP
patra
obl
NOUN
druhého
amod
ADJ
rána
nsubj
NOUN
a
cc
CCONJ
pes
nsubj
NOUN
otočil
conj
VERB
se
expl:pv
PRON
jako
mark
SCONJ
čamrda
advcl
NOUN
,
punct
PUNCT
zatřepal
conj
VERB
prudce
advmod
ADV
packami
obl
NOUN
svalil
conj
VERB
,
punct
PUNCT
se
expl:pv
PRON
na
case
ADP
zem
obl
NOUN
dodělal
conj
VERB
a
cc
CCONJ
v
case
ADP
křečích
obl
NOUN
škubavýc
amod
ADJ
.
punc
PUN

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*

Sentence – cognitive load

mdd:

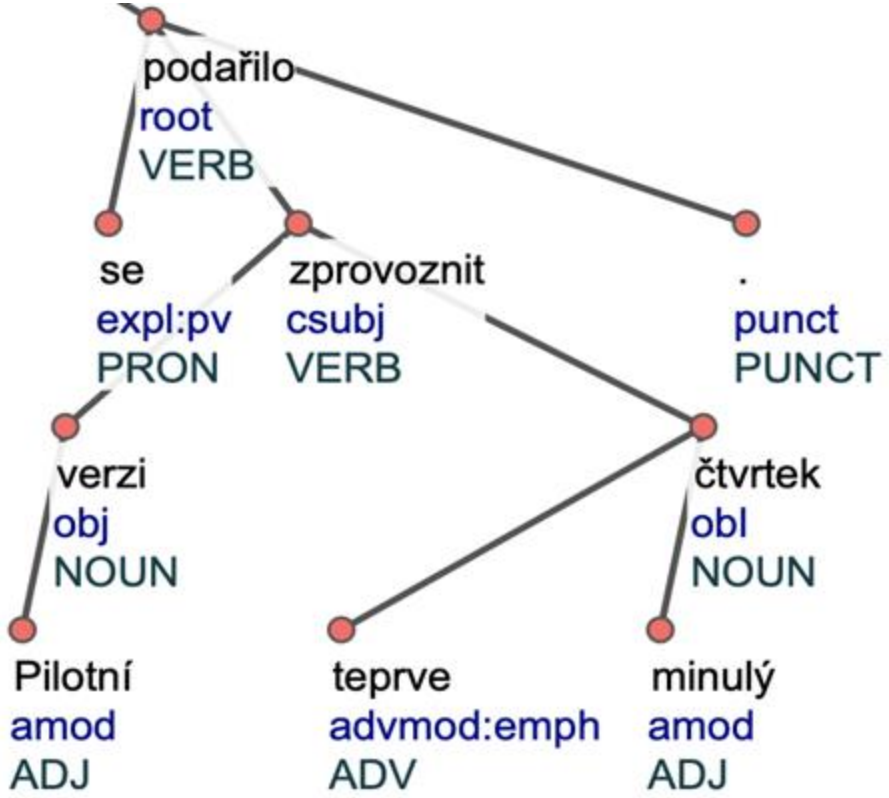
- Mean Dependency Distance (Yan & Li, 2019; Mačutek et al., 2021; Alemany-Puig & Ferrer-i-Cancho 2024)
- Average head-daughter distance
- Punctuation is ignored
- calculation (n = 8 … no. of words in sentence)

  $DD_i = |ID_i - head_i|$

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

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

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*

Lexical diversity

- Only text-level
- 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

Outline

1. Linguistic variation within and across languages
2. Metrics of syntactic complexity and lexical diversity
3. The data: InterCorp – a multilingual parallel corpus
4. Accessing the annotation via search interface
5. Using the metrics
6. Perspectives, questions, discussion
7. References

InterCorp – a multilingual parallel corpus

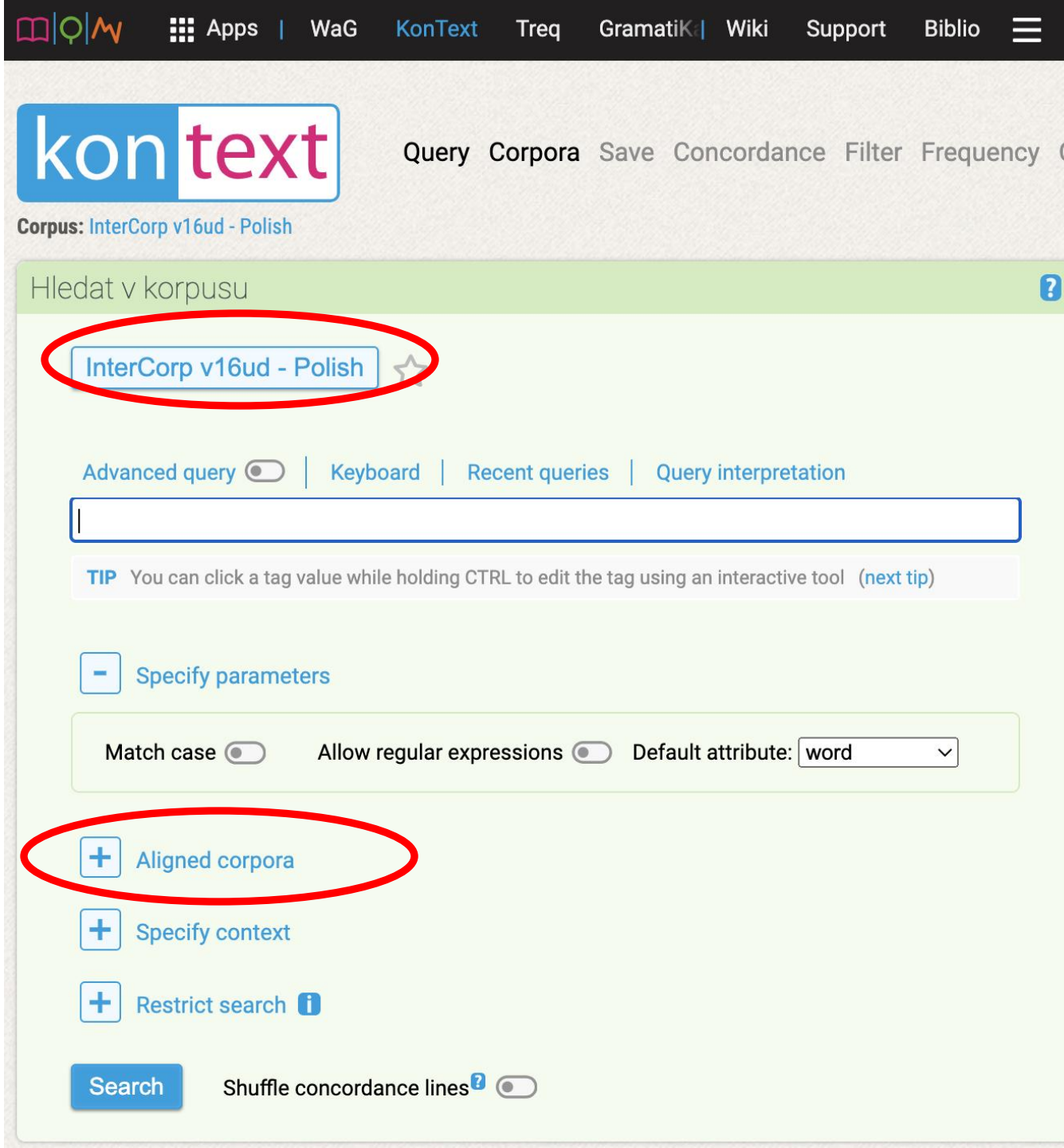
- Part of the *Czech National Corpus*
- Every text in Czech and at least one other language
- 2008: v0 (first online release)
- 2023: v16 (language-specific linguistic annotation)
- 2024: v16ud (linguistic annotation based on Universal Dependencies)
- 62 languages, including 47 UD-annotated
- 5.4 billion words
- Also as monolingual subcorpora
- Polish: 227 mil. words, incl. 27 mil. words in 328 fiction texts

Access to:

- InterCorp v16ud

without login
OR
with institutional login (Shibboleth)
OR
with login & pw: ud16test

1. Go to: **korpus.cz**
2. Click on: **KonText**
3. Click on: **syn2020 > All corpora**
4. Select/Type in: **InterCorp v16ud - Polish**



InterCorp v16ud - Polish

[lemma="Czech"]

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

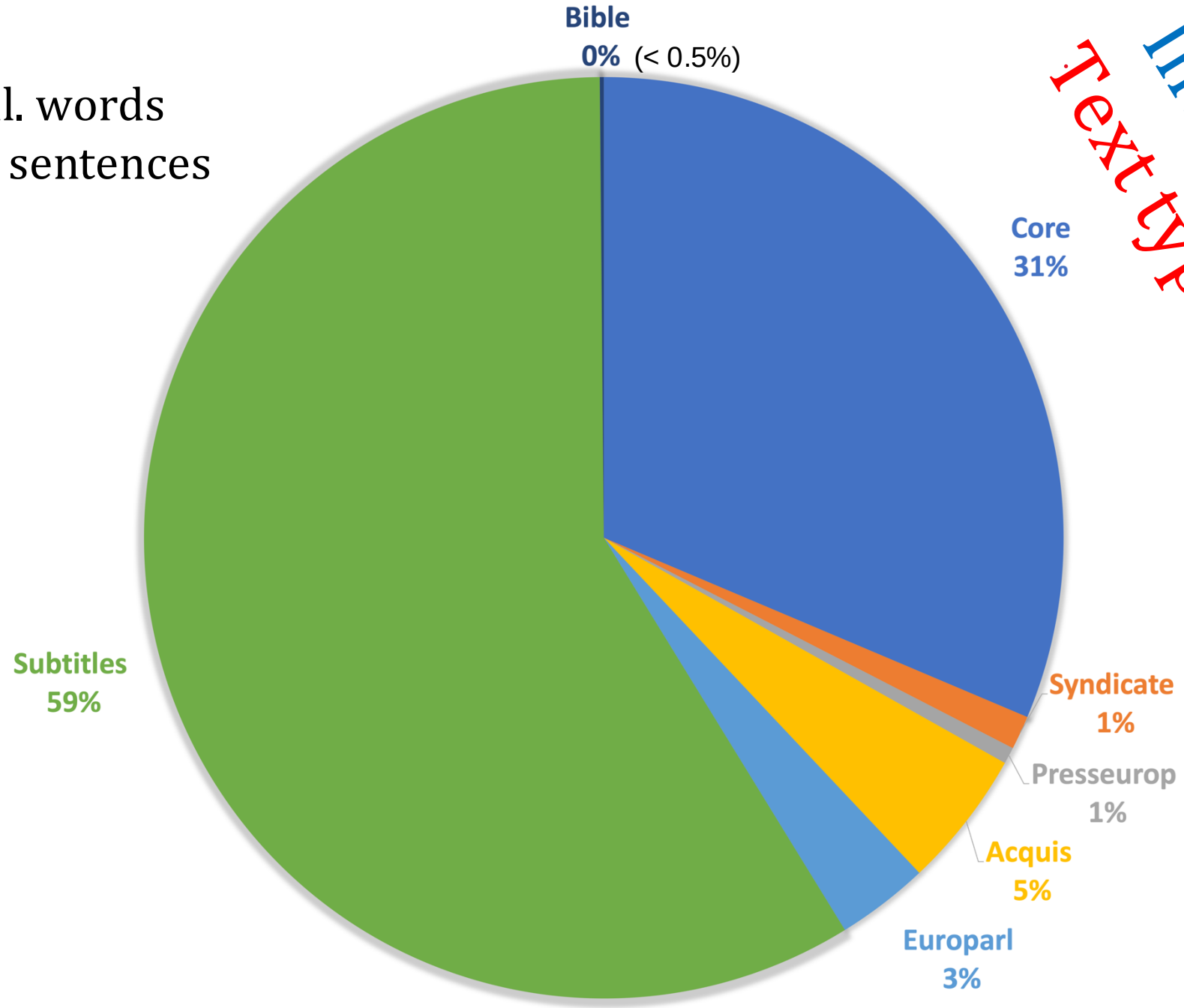
1 / 11

Line selection: simple

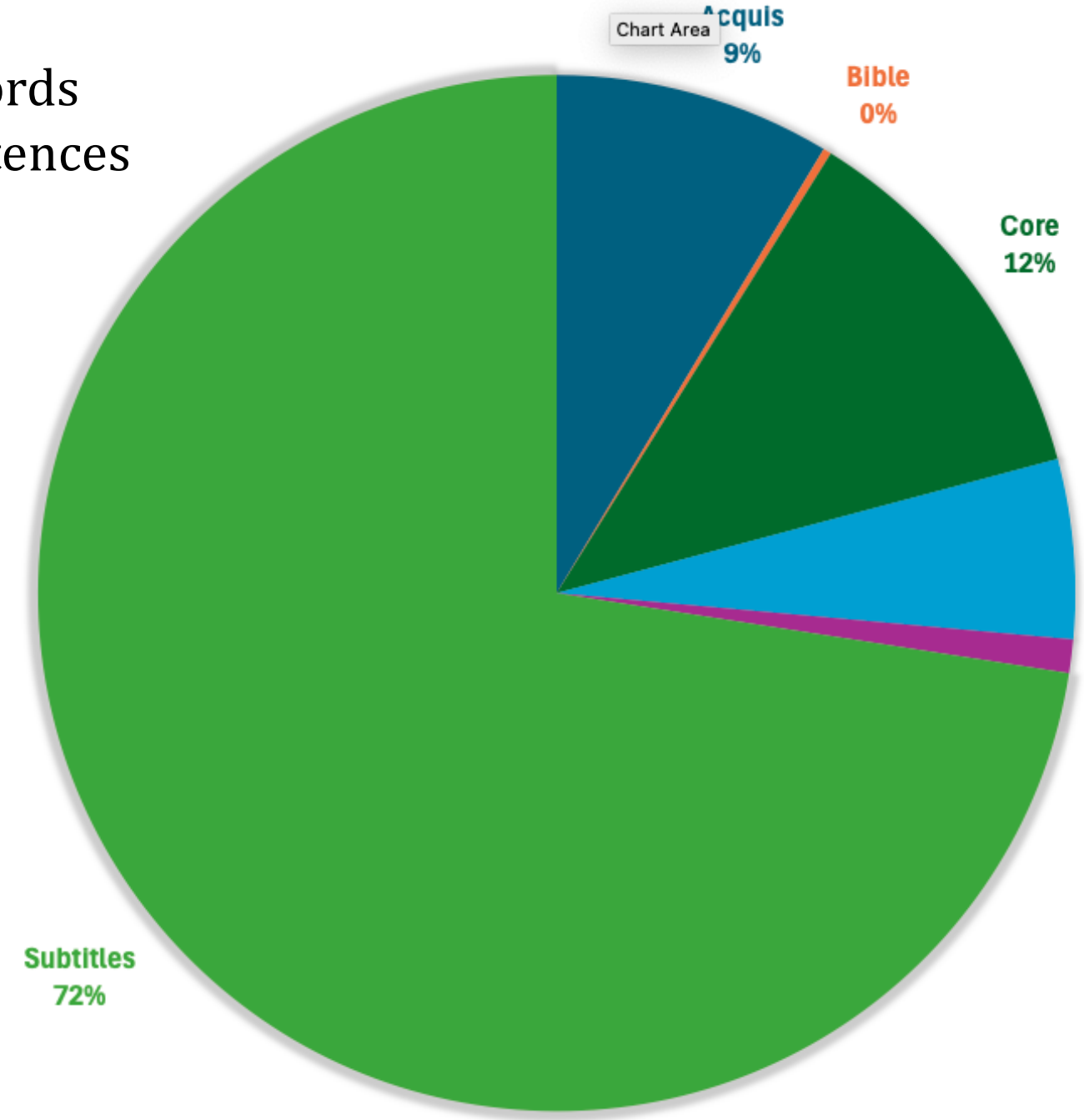
	InterCorp v16ud - Polish	InterCorp v16ud - Czech
	Zasadziłam trochę w donicach , trochę na rabacie , po czym poleciałam do miasta , na obiad ze swoim czytelnikiem , kanadyjskim **Czechem** .	Něco jsem zasadila do truhlíků , něco nechala na záhon a honem do města , kde jsem měla mít oběd se svým čtenářem , 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 .

InterCorp v16ud
Text types (Czech)

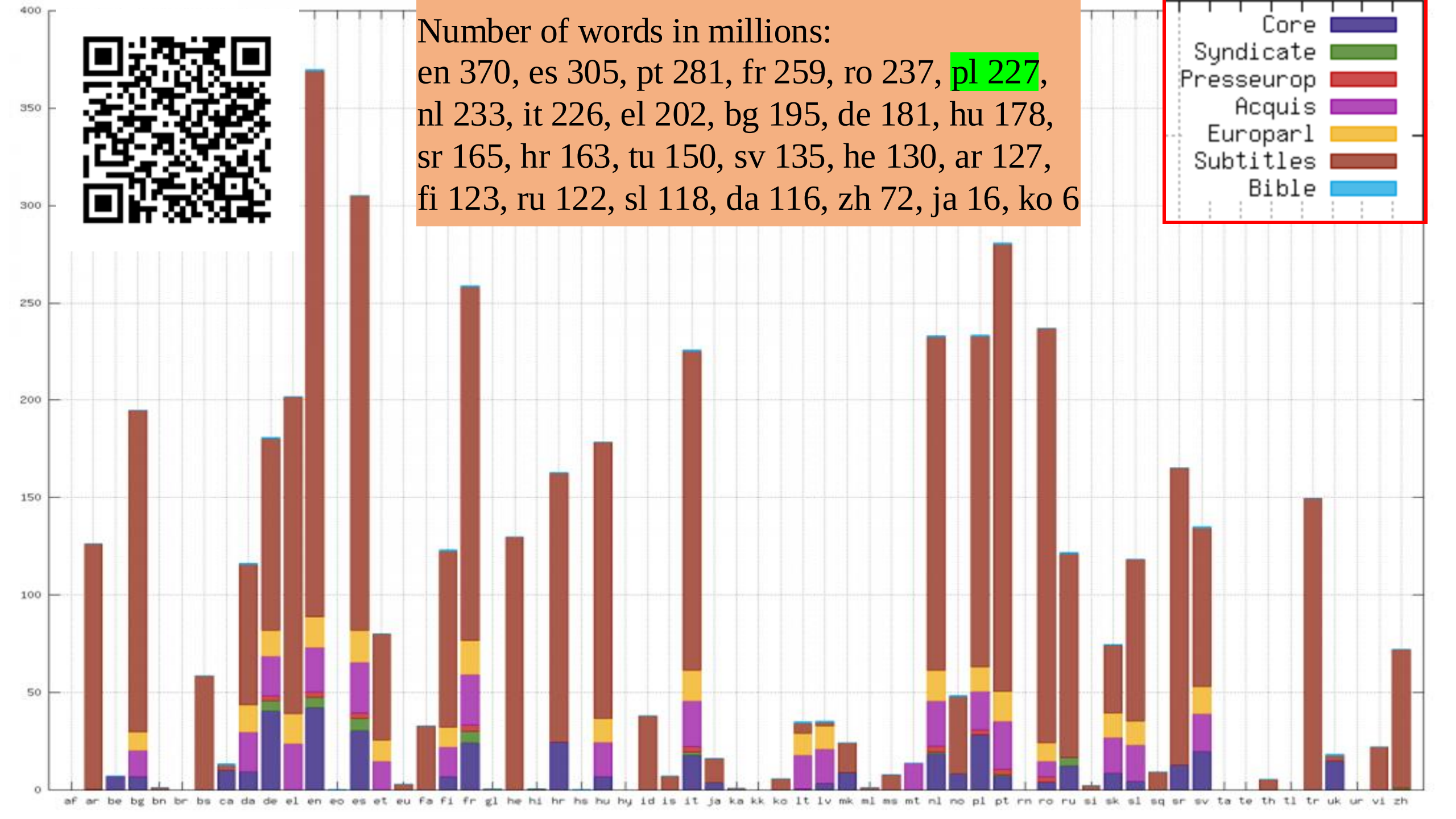
397 mil. words
61 mil. sentences



227 mil. words
41 mil. sentences



InterCorp v16ud
Text types (Polish)


Number of words in millions:
en 370, es 305, pt 281, fr 259, ro 237, pl 227,
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
Core
Syndicate
Presseurop
Acquis
Europarl
Subtitles
Bible
400
350
300
250
200
150
100
50
0
af ar be bg bn br bs ca da de el en eo es et eu fa fi fr gl he hi hr hs hu hy id is it ja ka kk ko lt lv mk ml ms mt nl no pl pt rn ro ru si sk sl sq sr sv ta te th tl tr uk ur vi zh

In preceding releases:
language-specific linguistic annotation

Languages in InterCorp 16ud

With annotation in 16ud

Without annotation in 16ud

Includes fiction

More than 15 texts in the core

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

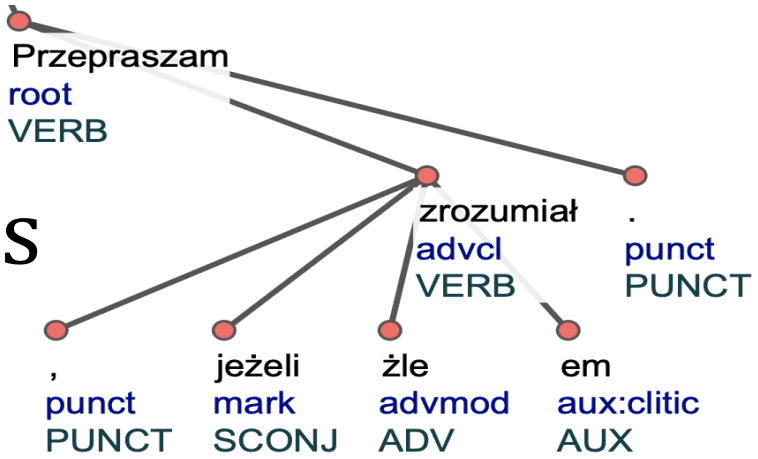
More info:

- All about InterCorp:
  https://wiki.korpus.cz/doku.php/en:cnk:intercorp
- On InterCorp v16ud:
  https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze16ud
- 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

Universal in Dependencies 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

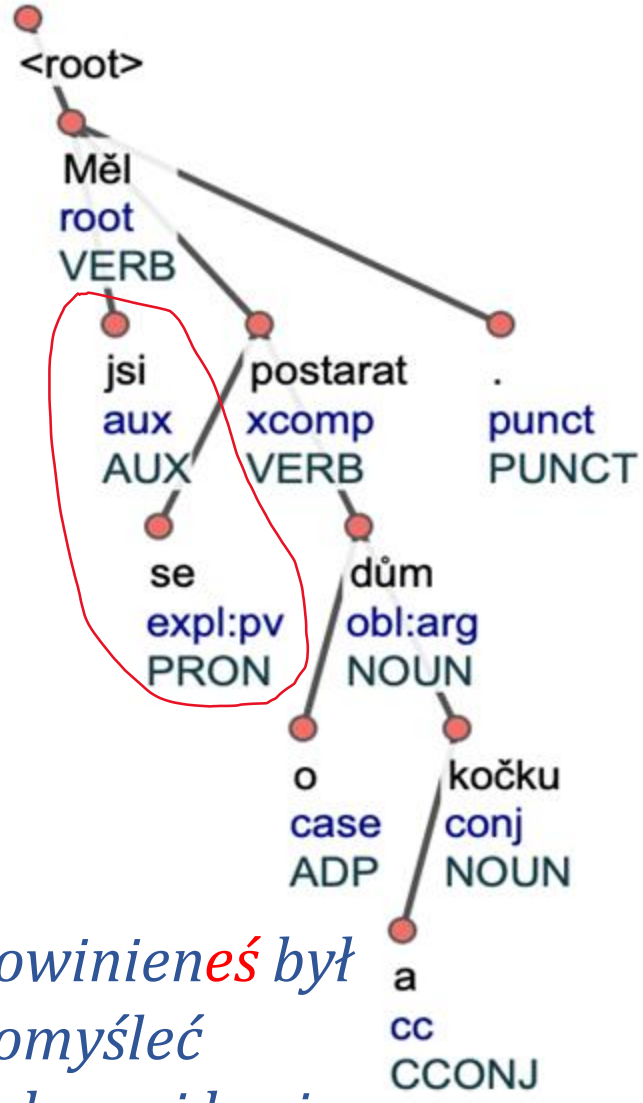
CONLL-U

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



Powinieneś był pomyśleć o domu i kocie.

```
[sword="em"]
[lemma="być"]
[upos="VERB"]
[word="ses"]
[sword="jsi"]
[lemma="být"]
[sword=".*\|.*"]
1:[sword=".*\|.*"] & 1.sword != 1.iword
```

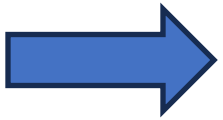
Měl ses postarat o dům a kočku.

UD in InterCorp – syntactic sugar

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

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

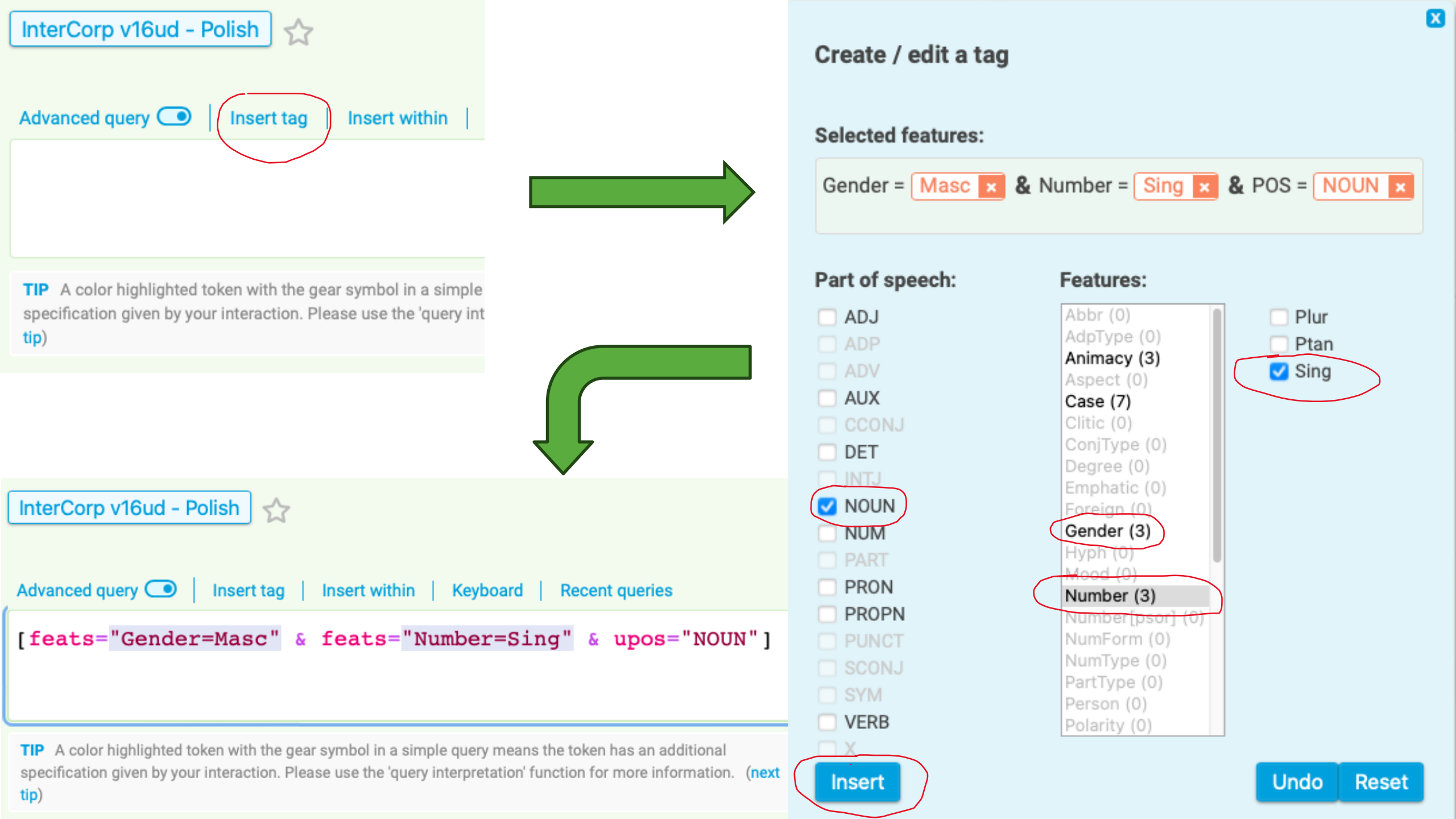
- … navigating syntactic structure (`p_lemma, e_deprel`):
  - lemma, upos, feats, deprel and relative position of the head
  - ID, relative position and deprel of the effective head (for coordination)
- … access to info about function words (`aux_feats, case_lemma`):
  - lemma, upos, feats and deprel 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

Outline

1. Linguistic variation within and across languages
2. Metrics of syntactic complexity and lexical diversity
3. The data: InterCorp – a multilingual parallel corpus
4. Accessing the annotation via search interface
5. Using the metrics
6. Perspectives, questions, discussion
7. References


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
Insert
Undo
Reset
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)

kontext

Query Corpora Save Concordance Filter Frequency

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

Advanced query | Insert tag | Insert within | Keyboard

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

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

To list typical deprels of a lemma

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

kontext Query Corpora Save Concordance Filter Frequency Collocatic

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

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

Frequency distribution

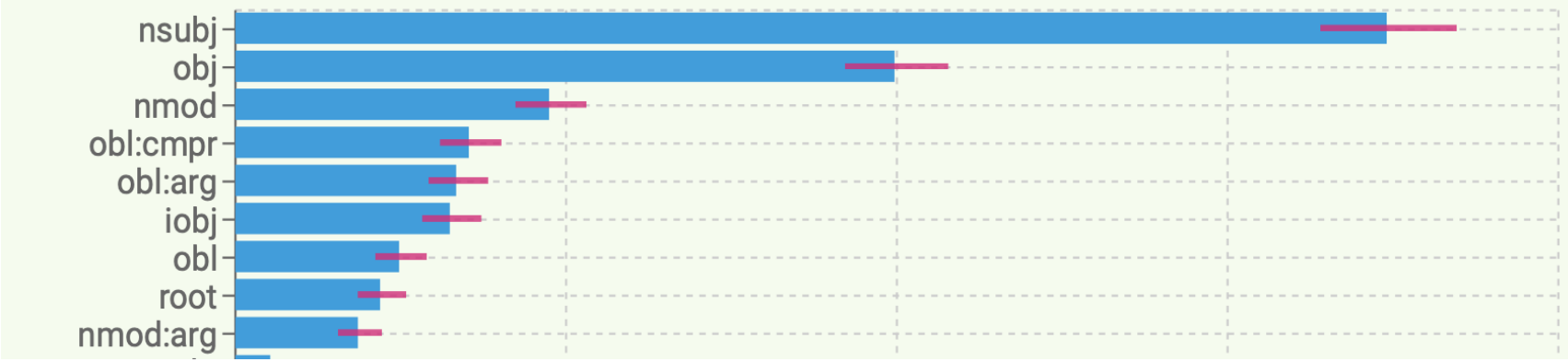
Standard | According to text types | Dispersion | Two-attribute interrelationship

Frequency limit: 1

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

+

Make frequency list



[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

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

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

...in Czech?

To list typical predicates of a subject

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

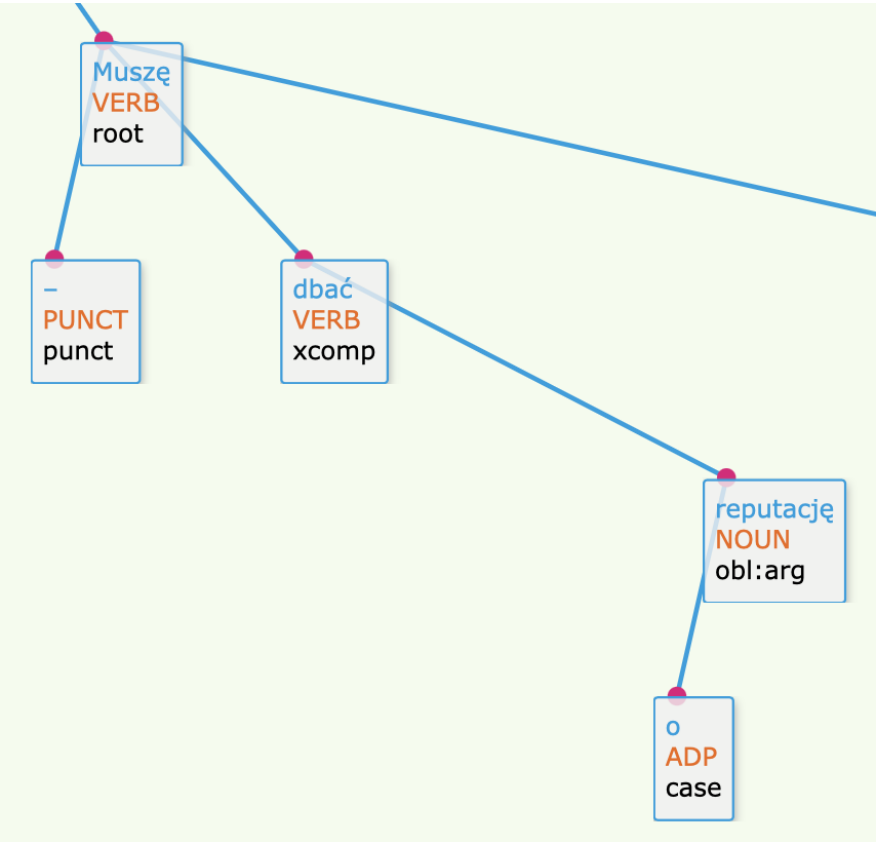
What do the Polish birds do?

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

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

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



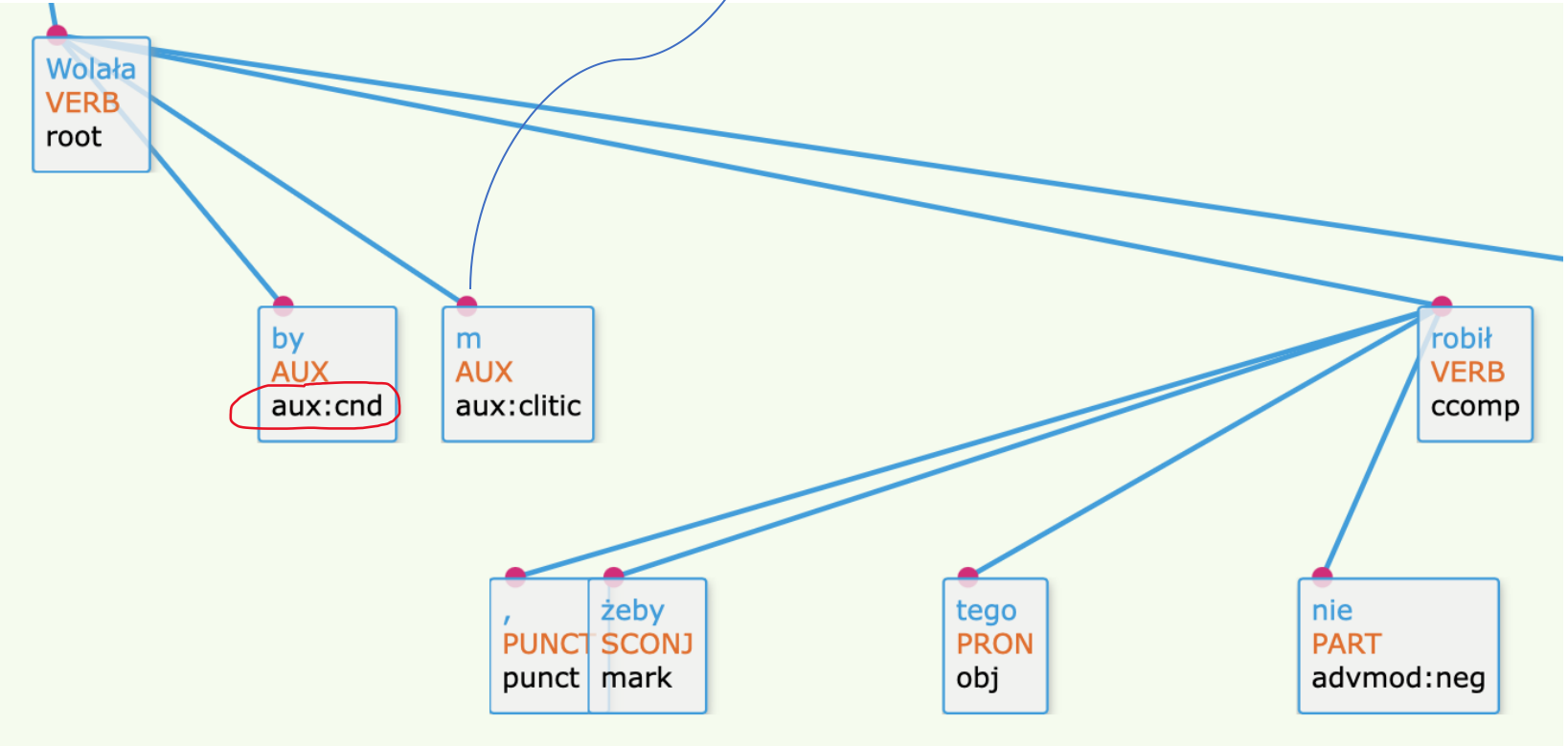
1 / 68 ▶ (total: 3,356 items)

	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



Equivalent queries

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

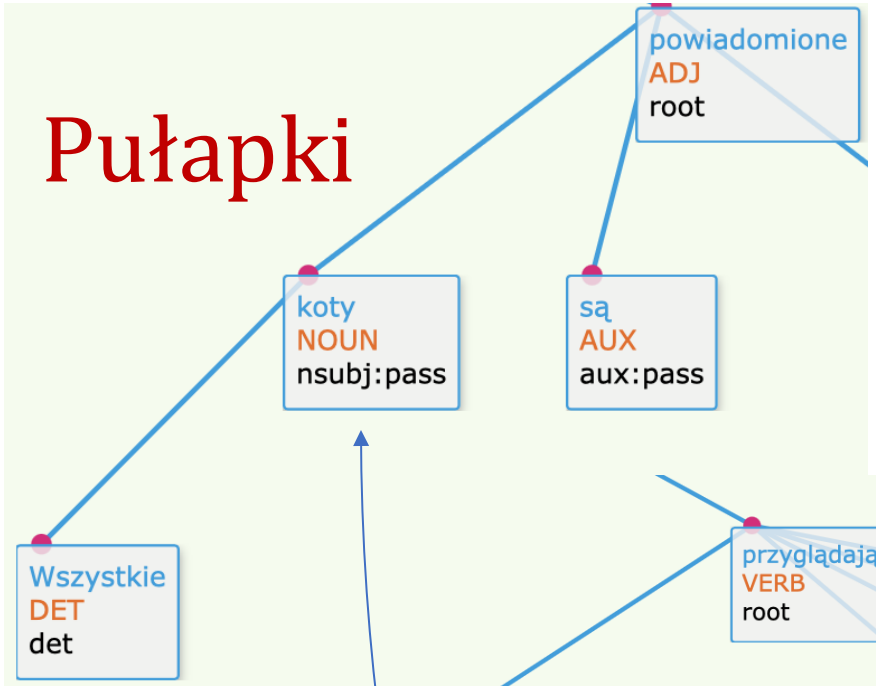
[upos="NOUN" & feats=".*Case=Gen.*Gender=Fem.*"]

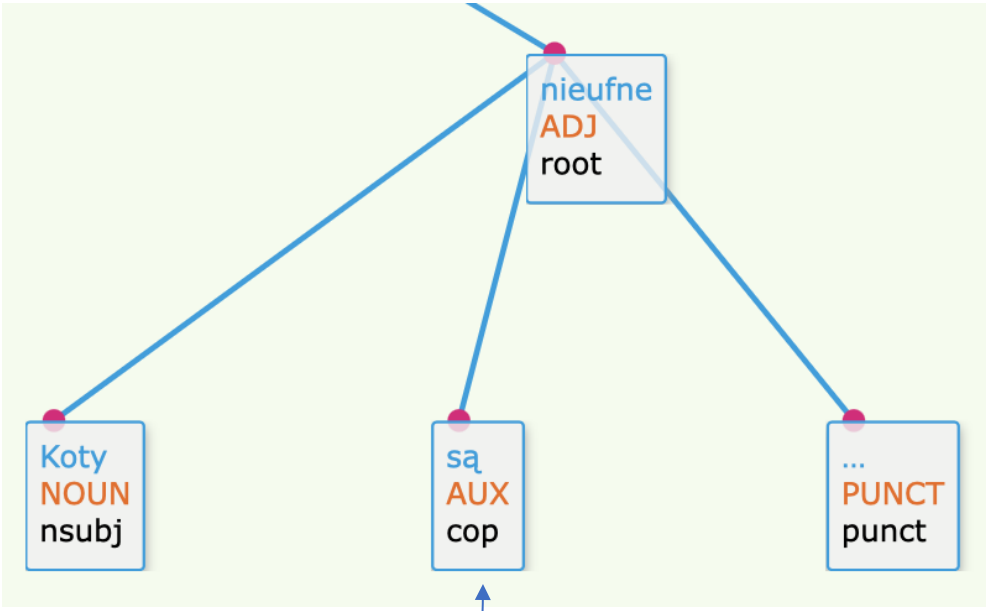
[upos="NOUN" & gender= "Fem" & case="Gen"]

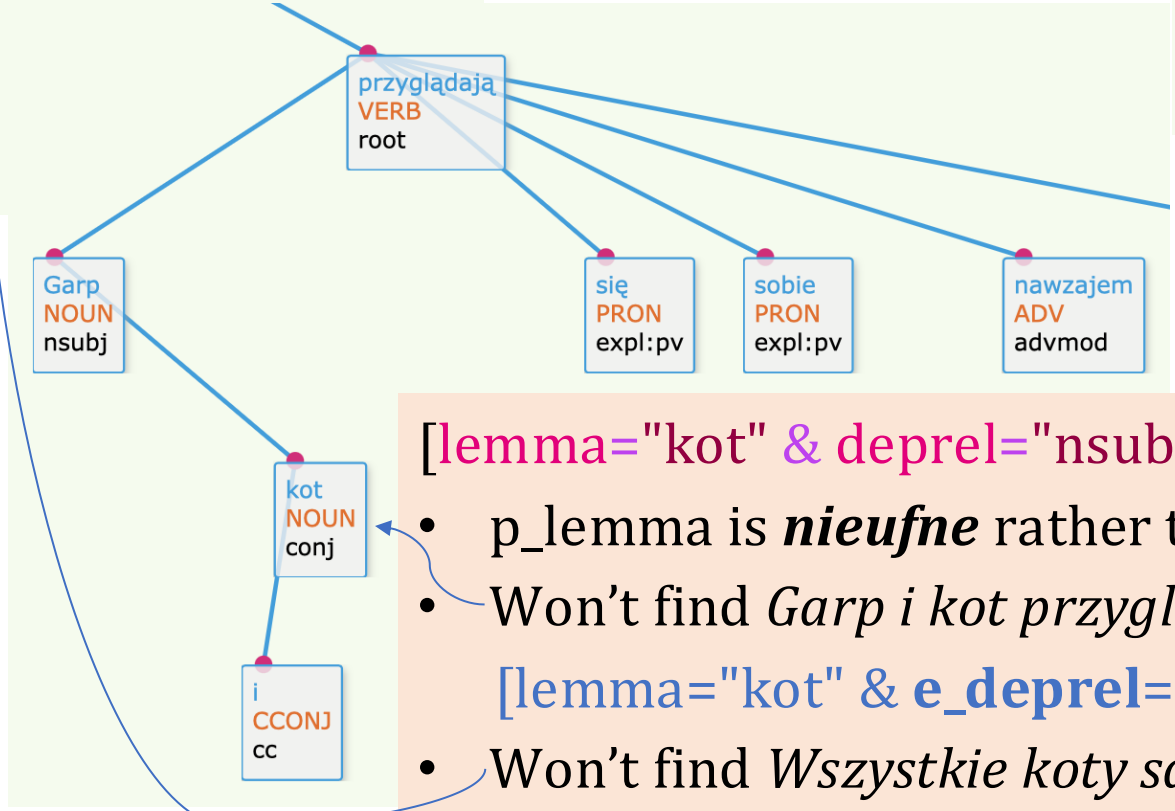
[xpos="subst:..:gen:f"]

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.***"]

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 . "

To view the sentence metrics

View > KWIC/**Sentence**

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

Corpus-specific settings for InterCorp v16ud - Czech

Positional attributes | Structures | References | Additional functions

<doc>	<text>	<p>	<s>	<hi>
id	lang	id	id	rend
tag_model	pubyear		maxNPDepth	
	version		subRatio	
	pubmonth		sLength	
	pubDateYear		maxNPLength	
	pubDateMonth		mdd	
	id		maxTreeDepth	
	author			
	title			
	group			
	publisher			

Select all in all structures

Apply View Options

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

kontext

Query Corpora Save Concordance Filter

To find short and easy sentences

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

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

View > Corpus-specific settings > References > s.sLength

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

Positional attributes | Structures | References | Additional functions

- <#>
  - 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

To find something in short and easy sentences

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

To display text-level metrics & to download results

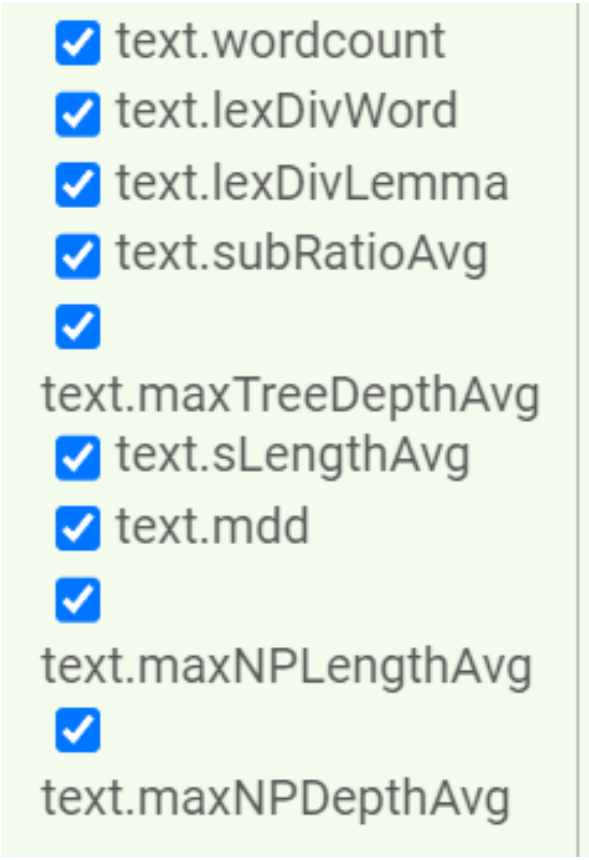
<text> []

View > Sentence

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

Apply View Options

Save > CSV/XLSX



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 are the metrics good for?

- Contrastive / typological research of multiple languages
- Translation studies
- Research of text types variability
  - Comparison of metrics for sentences, texts, text types, languages
  - Correlation and comparison of metrics
- Applications
  - Readability assessment
  - Text simplification
- Teaching L1/L2
  - Filtering corpus examples
  - Building subcorpora for self-study
  - TODO: learner texts assessment

Polish texts with minimal lexDivLemma – Maximum: Nabokov's Lolita (615.81)

Author	Title	text type	src lang.	word count	lexDivword	lexDivword#	lexDivLemma	lexDivLemma#
	Umowy	nonfict	pl	13,076	506.12	1	347.93	1
Milne, Alan Alexander	*Chatka Puchatka*	fiction	en	23,833	509.12	2	367.87	2
Milne, Alan Alexander	*Fredzia Phi-Phi*	fiction	en	18,864	513.88	3	374.46	3
Dousková, Irena	*Bądźżesz*	fiction	cs	38,058	521.26	4	386.24	4
Lindgren, Astrid	*Braciszek i Karlsson ...*	fiction	sv	22,346	544.38	6	400.94	5
Hajduk-Veljkovićowa, L.	*Dolina nad rzeką*	children	hs	2,527	578.56	29	405.26	6
	umowy	other	cs	3,781	527.21	5	409.15	7
Heisenberg, Werner	*Fizyka i filozofia*	nonfict	de	54,966	582.07	34	411.54	8
Grynberg, Henryk	*Żydowska wojna*	fiction	pl	19,623	566.63	14	415.40	9
De Saint-Exupéry, Antoine	*Mały Książę*	fiction	fr	11,041	576.82	25	417.68	10
Hawking, Stephen William	*Krótka historia czasu ...*	nonfict	en	44,042	582.82	35	422.74	11
Fuks, Ladislav	*Wariacje ...*	fiction	cs	118,414	555.92	10	424.35	12
Schmidt, Annie M.G.	*Minu*	fiction	nl	27,563	577.60	27	424.74	13
Hemingway, Ernest	*Wyspy na Golfsztormie*	fiction	en	116,795	575.91	24	425.92	14
Hemingway, Ernest	*Komu bije dzwon*	fiction	en	138,861	573.36	19	426.51	15
Stachura, Edward	*Siekierezada*	fiction	pl	56,242	553.42	8	426.62	16
Carroll, Lewis	*Alicja w Krainie Czarów*	fiction	en	20,848	554.19	9	427.30	17
Gombrowicz, Witold	*Trans-Atlantyk*	fiction	pl	36,945	552.76	7	427.83	18
Carroll, Lewis	*Alicja po tamtej stronie ...*	fiction	en	25,774	558.36	11	428.72	19
Kafka, Franz	*Zamek*	fiction	de	103,195	559.19	12	429.24	20
Grynberg, Henryk	*Zwycięstwo*	fiction	pl	29,295	584.02	38	429.53	21

Polish texts with maximal subRatio – Minimum: Krynicky's Kamień, Szron (1.22)

Author	Title	text type	src lang.	word count	subRatio	subRatio#	maxNP Length	maxNPLength#	maxNP Depth	maxNP Depth#
García Márquez, G.	*Jesień patriarchy*	fiction	es	69,716	7.57	348	96.28	348	8.38	348
Proust, Marcel	*W poszukiwaniu straconego czasu*	fiction	fr	132,689	3.73	347	9.52	339	2.90	330
Hrabal, Bohumil	*Zbyt głośna samotność*	fiction	cs	25,166	3.15	346	20.75	347	3.88	347
Saramago, Jose	*Baltazar i Blimunda*	fiction	pt	109,534	3.06	345	14.50	346	3.27	340
Heisenberg, W.	*Fizyka i filozofia*	nonfict	de	54,966	3.00	344	9.54	340	3.37	343
Lorenz, Konrad	*Tak zwane zło*	nonfict	de	81,019	2.97	343	11.28	343	3.57	346
Hrabal, Bohumil	*Kim jestem*	fiction	cs	13,831	2.95	342	12.79	345	2.93	332
Popper, Karl	*Społeczeństwo otwarte 2*	nonfict	en	91,720	2.91	341	9.00	335	3.07	336
Pamuk, Orhan	*Czarna księga*	fiction	tr	143,389	2.90	340	8.48	325	2.69	324
Schaff, Adam	*Wstęp do semantyki*	nonfict	pl	88,468	2.89	338	10.96	342	3.49	345
Watson, James D.	*Podwójna spirala …*	nonfict	en	38,630	2.89	339	8.56	327	3.31	341
Foucault, Michel	*Słowa i rzeczy. Archeologia nauk..*	nonfict	fr	122,782	2.86	336	11.35	344	3.43	344
García Márquez, G.	*Miłość w czasach zarazy*	fiction	es	122,185	2.86	337	9.63	341	3.08	337
Kuczok, Wojciech	*Gnój*	fiction	pl	34,022	2.82	335	6.61	300	1.96	272
Popper, Karl	*Społeczeństwo otwarte 2*	nonfict	en	66,094	2.79	333	8.45	324	2.97	333
Lem, Stanisław	*Głos Pana*	fiction	pl	51,644	2.79	334	8.57	329	2.93	331

Polish texts with the most varied metrics

… rank 1–348

Author	Krynicki, Ryszard	Heisenberg, Werner	Hawking, Stephen W.	Schaff, Adam	Hrabal, Bohumil	Redliński, Edward	Sofokles	Hrabal, Bohumil
Title	*Kamień, szron*	*Fizyka i filozofia*	*Krótka historia czasu*	*Wstęp do semantyki*	*Różowy kawaler*	*Awans*	*Antygona*	*Obsługiwałem angielskiego…*
text type	fiction	nonfiction	nonfiction	nonfiction	fiction	fiction	drama	fiction
source lang.	pl	de	en	pl	cs	pl		cs
word count	3,598	54,966	44,042	88,468	36,096	31,066	8,118	63,134
lexDivword	723.26	582.07	582.82	584.93	561.98	689.45	667.19	581.10
lexDivword#	348	34	35	41	13	341	306	30
lexDivLemma	599.96	411.54	422.74	430.04	445.20	577.87	514.04	463.65
lexDivLemma#	347	8	11	23	36	343	253	74
subRatio	1.22	3.00	2.76	2.89	2.31	1.65	1.27	2.50
subRatio#	1	344	332	338	297	61	2	317
maxTreeDepth	0.20	1.58	1.49	1.43	1.13	0.60	0.28	1.51
maxTrDepth#	1	341	335	327	287	43	3	337
sLength	4.32	20.57	19.71	22.37	28.26	8.84	5.09	28.84
sLength#	1	324	315	333	342	62	2	343
mdd	1.86	2.45	2.39	2.57	5.78	2.30	1.92	3.11
mdd#	1	230	199	270	342	140	2	329
maxNPLength	2.37	9.54	7.95	10.96	8.93	2.64	2.26	8.60
maxNPLength#	4	340	320	342	333	19	2	330
maxNPDepth	0.91	3.37	3.03	3.49	2.38	1.03	0.87	2.54
maxNPDepth#	4	343	334	345	310	21	2	317
STDEV on #	159.94	143.83	138.52	138.16	137.72	136.87	129.16	128.99

Language or text type? Weighted means for Polish and English

Polish

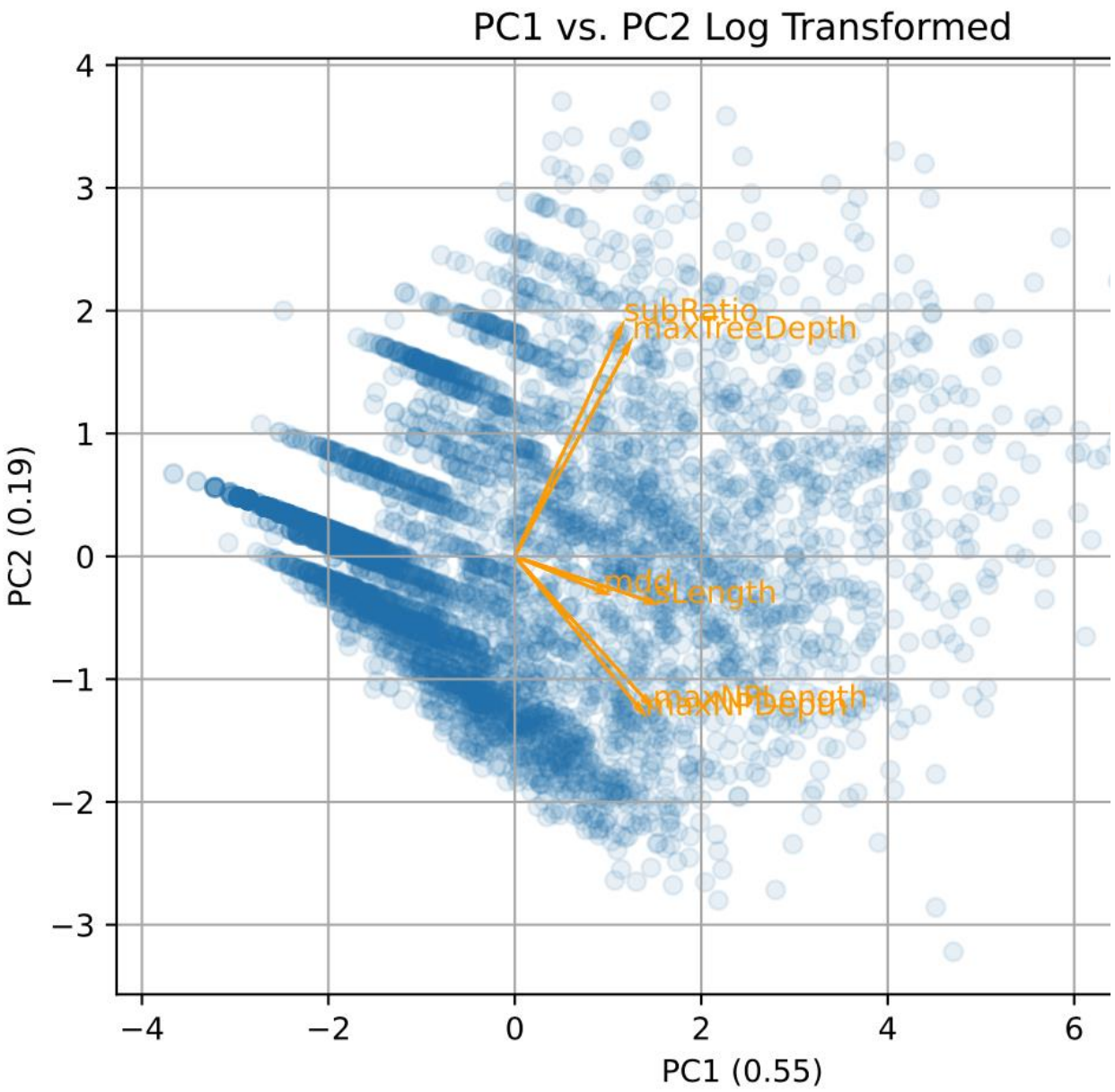
Collection	words	lexDivWord	lexDivLemma	sLength	subRatio	maxTreeDepth	maxNPLength	maxNPDepth	mdd
Core-nonfict	754 K	613.7	460.1	20.825	2.743	1.407	9.385	3.192	2.509
Core-fiction	27,056 K	632.2	499.6	11.498	1.896	0.833	4.162	1.523	2.355
Core-misc	283 K	622.9	471.6	12.263	1.981	0.881	4.978	1.892	2.266
Acquis	19,482 K	481.4	350.6	13.373	2.035	0.714	7.737	2.681	2.622
Bible	576 K	537.0	387.8	12.695	1.724	0.727	4.479	1.725	2.397
Europarl	12,662 K	607.5	447.2	18.340	2.643	1.309	9.387	3.283	2.322
PressEurop	2,367 K	659.8	520.6	14.632	2.143	0.957	7.092	2.645	2.334
Subtitles	164,059 K	602.1	441.5	4.556	1.324	0.319	1.855	0.717	1.832

English

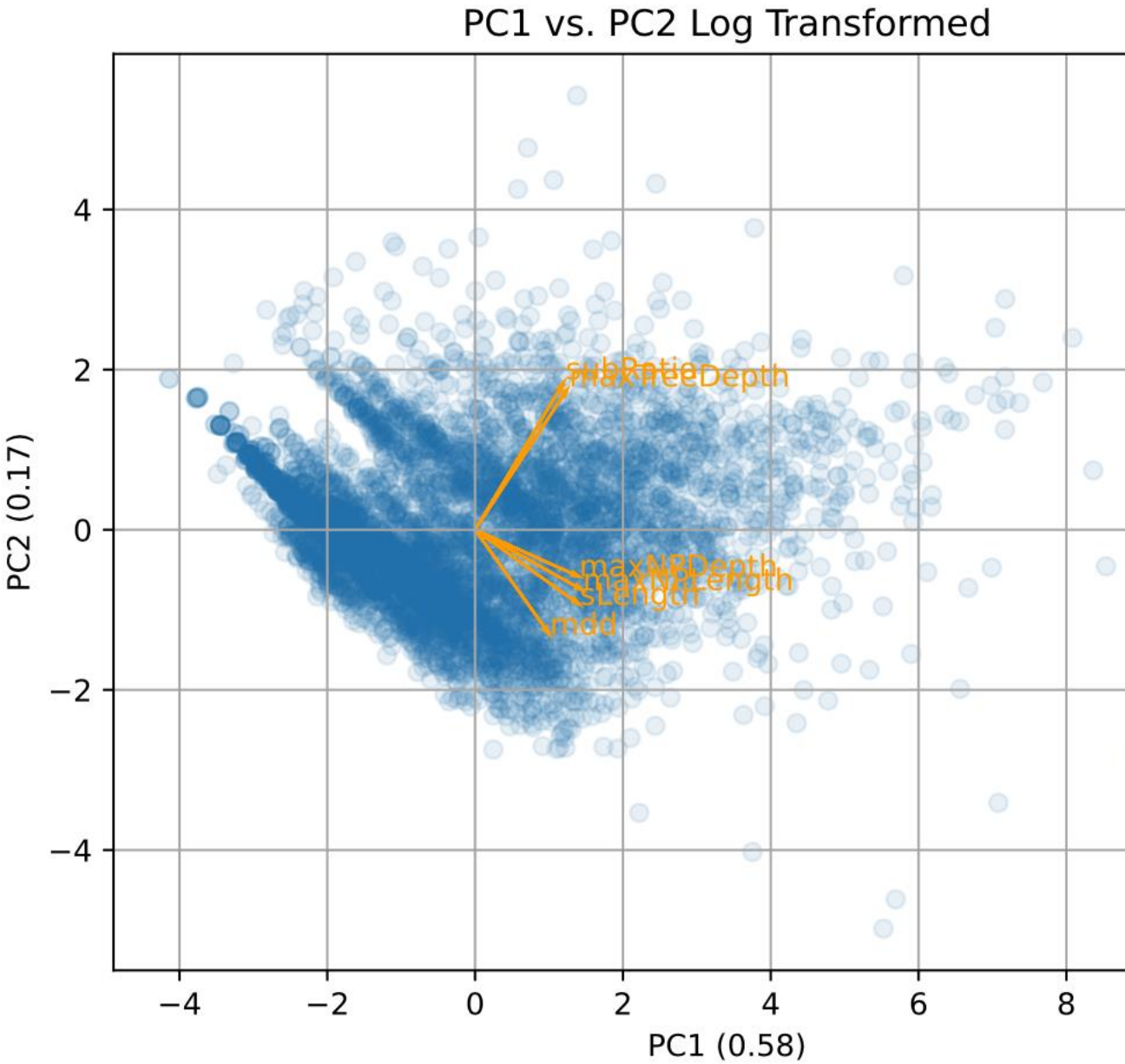
Core-nonfict	4,618 K	466.7	412.4	22.976	2.623	1.292	10.373	2.893	2.793
Core-fiction	36,519 K	466.2	403.2	14.159	2.107	0.945	5.371	1.689	2.576
Core-misc	778 K	455.8	393.7	15.091	2.160	0.967	6.561	1.987	2.557
Acquis	23,062 K	346.1	307.3	20.073	2.193	0.806	11.086	2.912	3.176
Bible	727 K	354.0	296.2	17.458	2.166	1.051	6.271	2.125	2.608
Europarl	15,593 K	411.9	362.9	23.743	2.692	1.402	11.274	3.135	2.736
PressEurop	2,663 K	485.4	431.4	18.016	2.286	1.033	8.828	2.614	2.689
Subtitles	267,843 K	445.1	362.4	5.491	1.401	0.372	2.273	0.811	2.067
Syndicate	5,272 K	494.2	438.7	20.792	2.447	1.186	9.516	2.843	2.733

Principal Component Analysis on sentence samples to find correlations between metrics

Polish

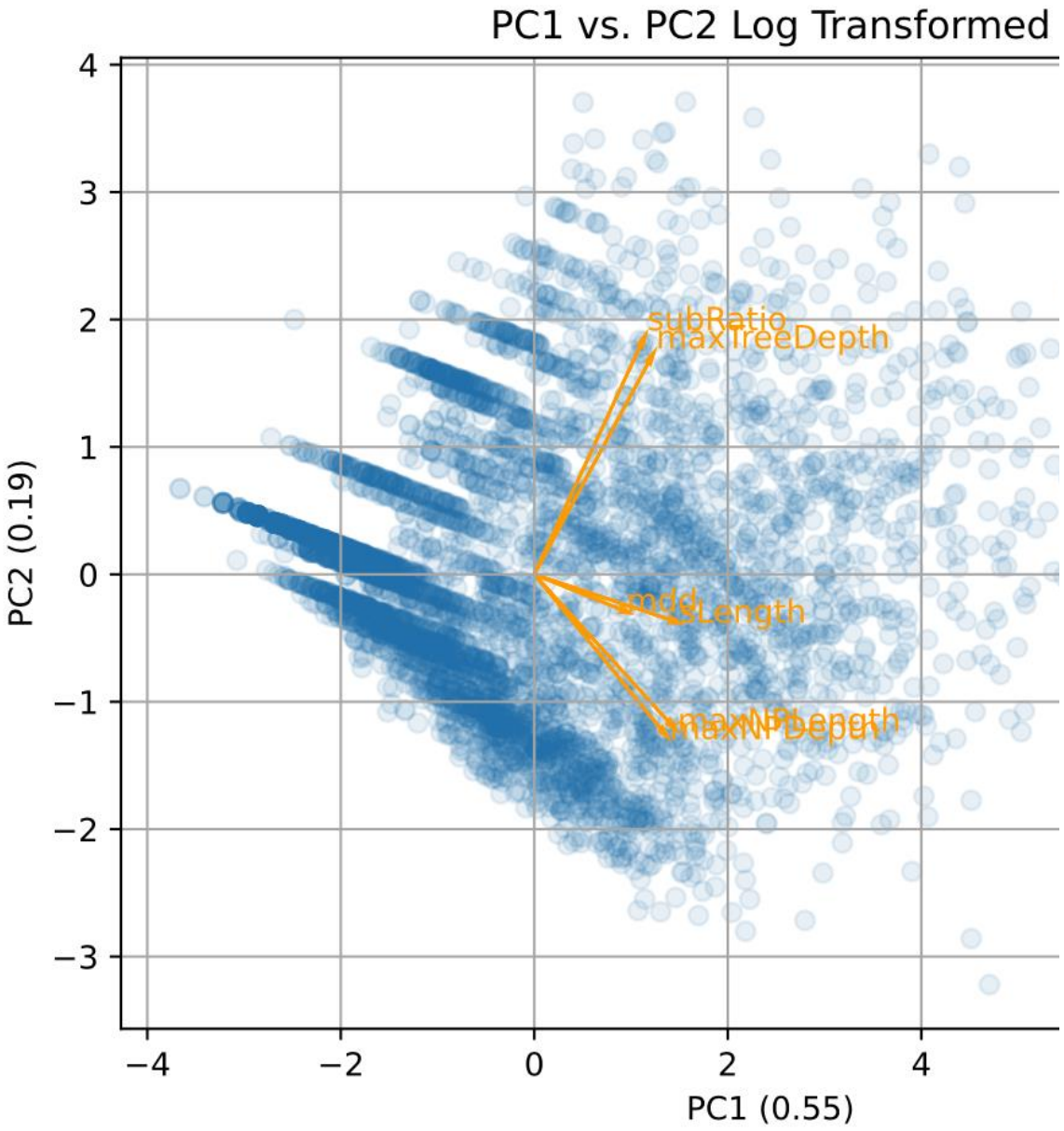


French

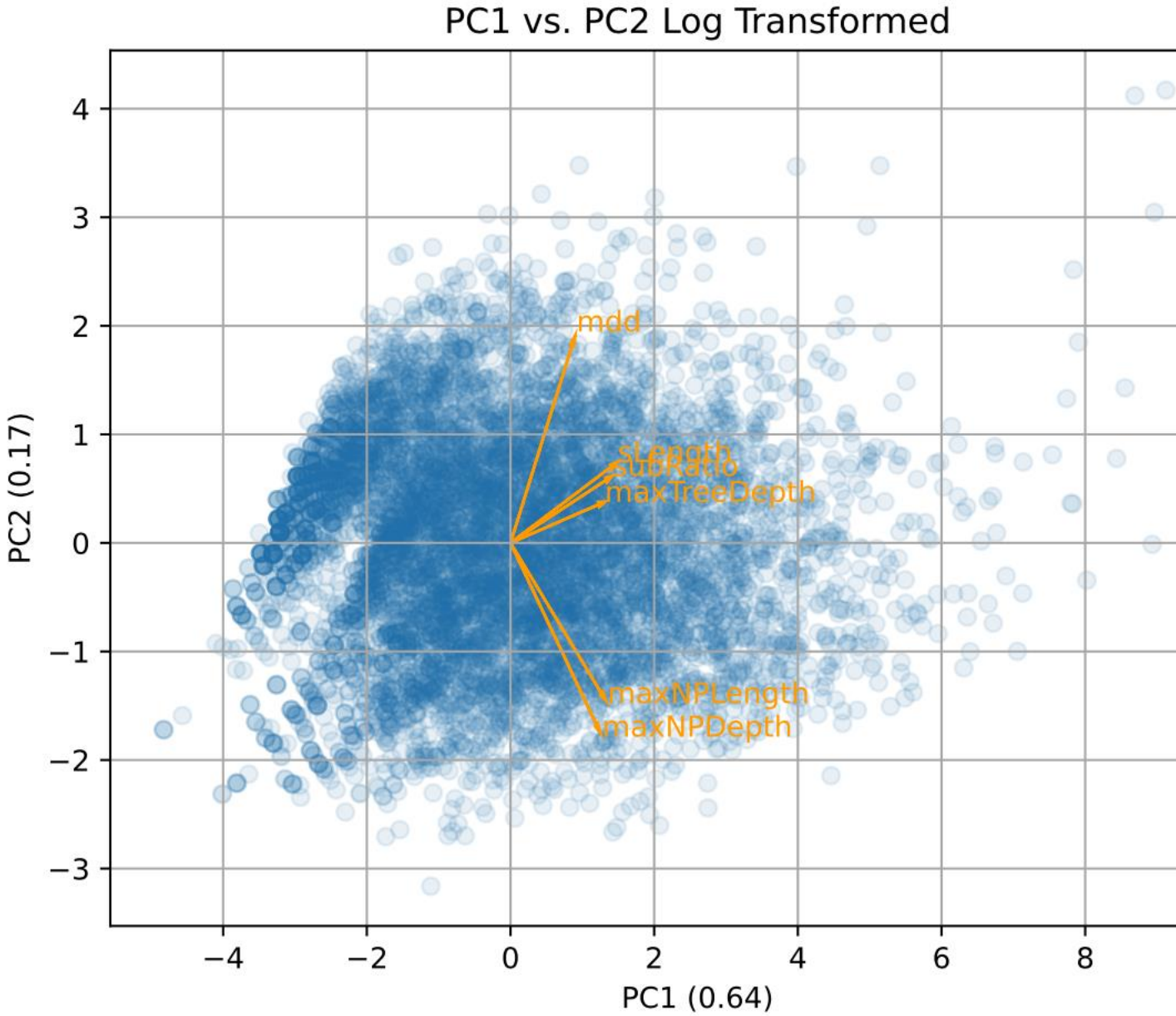


Principal Component Analysis on sentence samples to find correlations between metrics

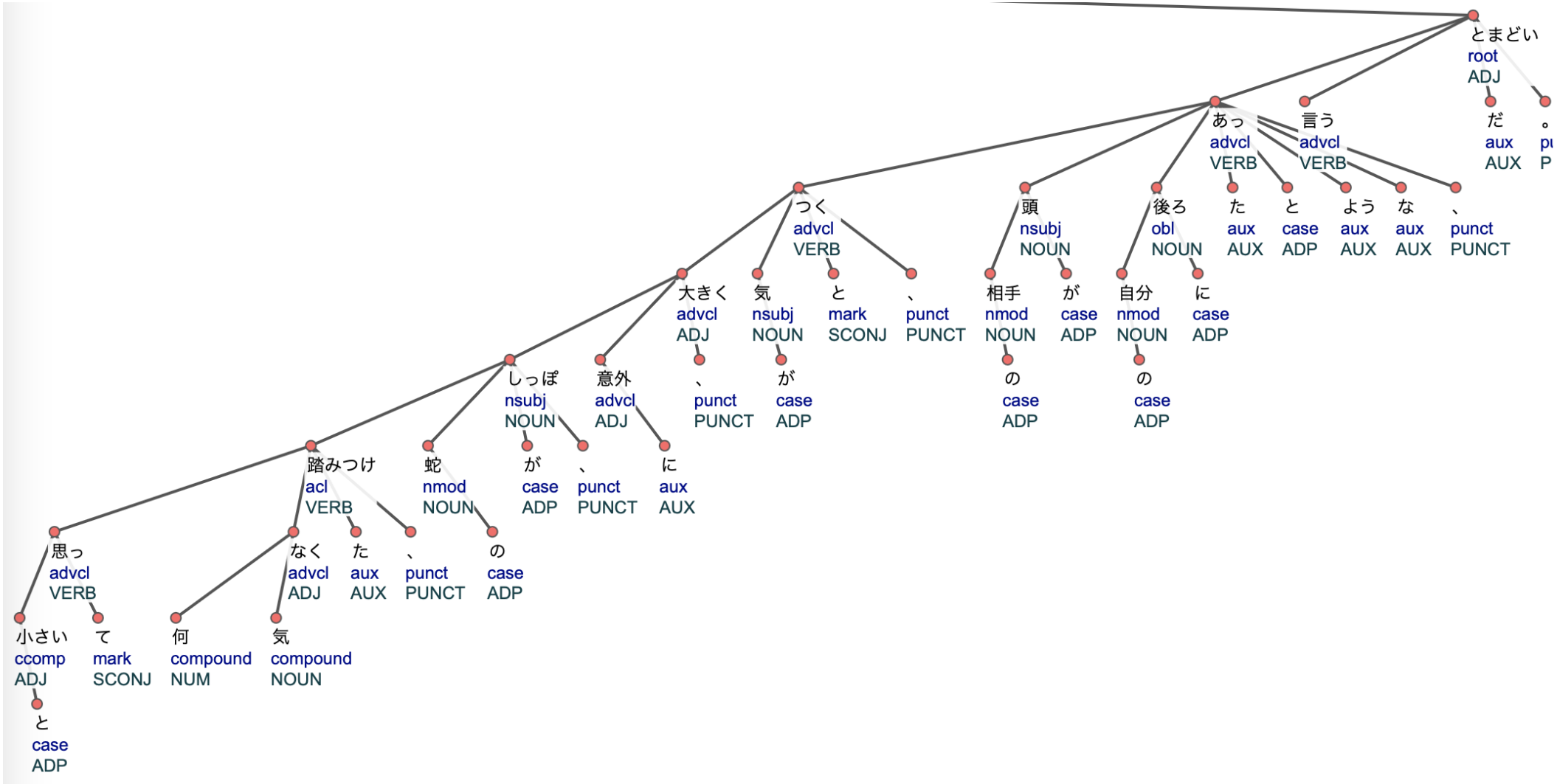
Polish



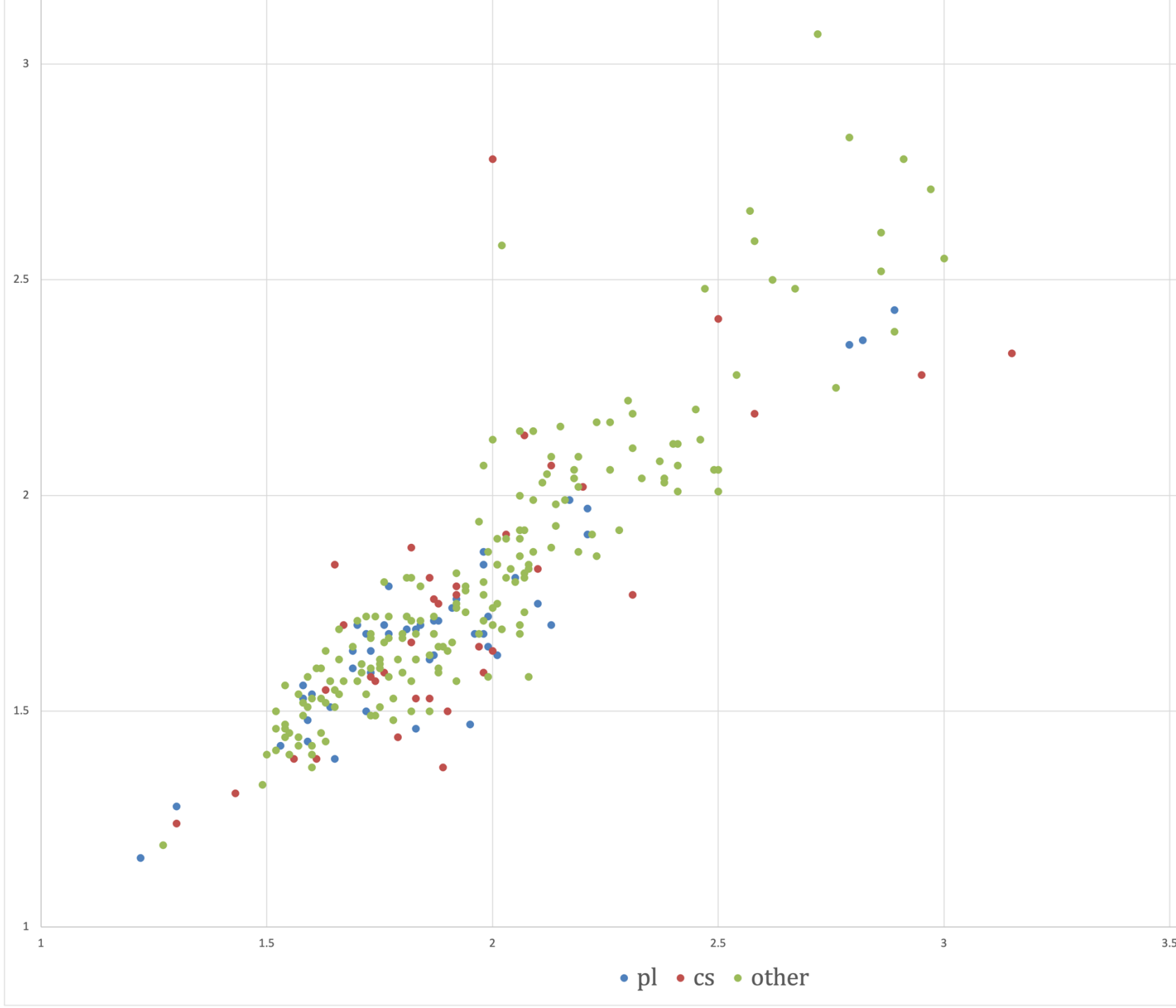
Japanese



Structure and tokenisation of Japanese



Scatter plot:
subRatio of Polish (x) and Czech (y) texts with Polish, Czech or other source language



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
7. References

Explaining the differences

Stylistic:

les disparités opposant [deprel=acl] *les classes populaires et les classes moyennes*
rozdíly mezi [deprel=case] *lidovou a střední vrstvou*
różnice między warstwą ludową a średnią

Normalization:

J'ai bu. J'ai eu alors envie de fumer.
Vypil jsem ji a dostal jsem chuť si zakouřit.
Wypiłem ją i nabrałem ochoty zapalić.

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*

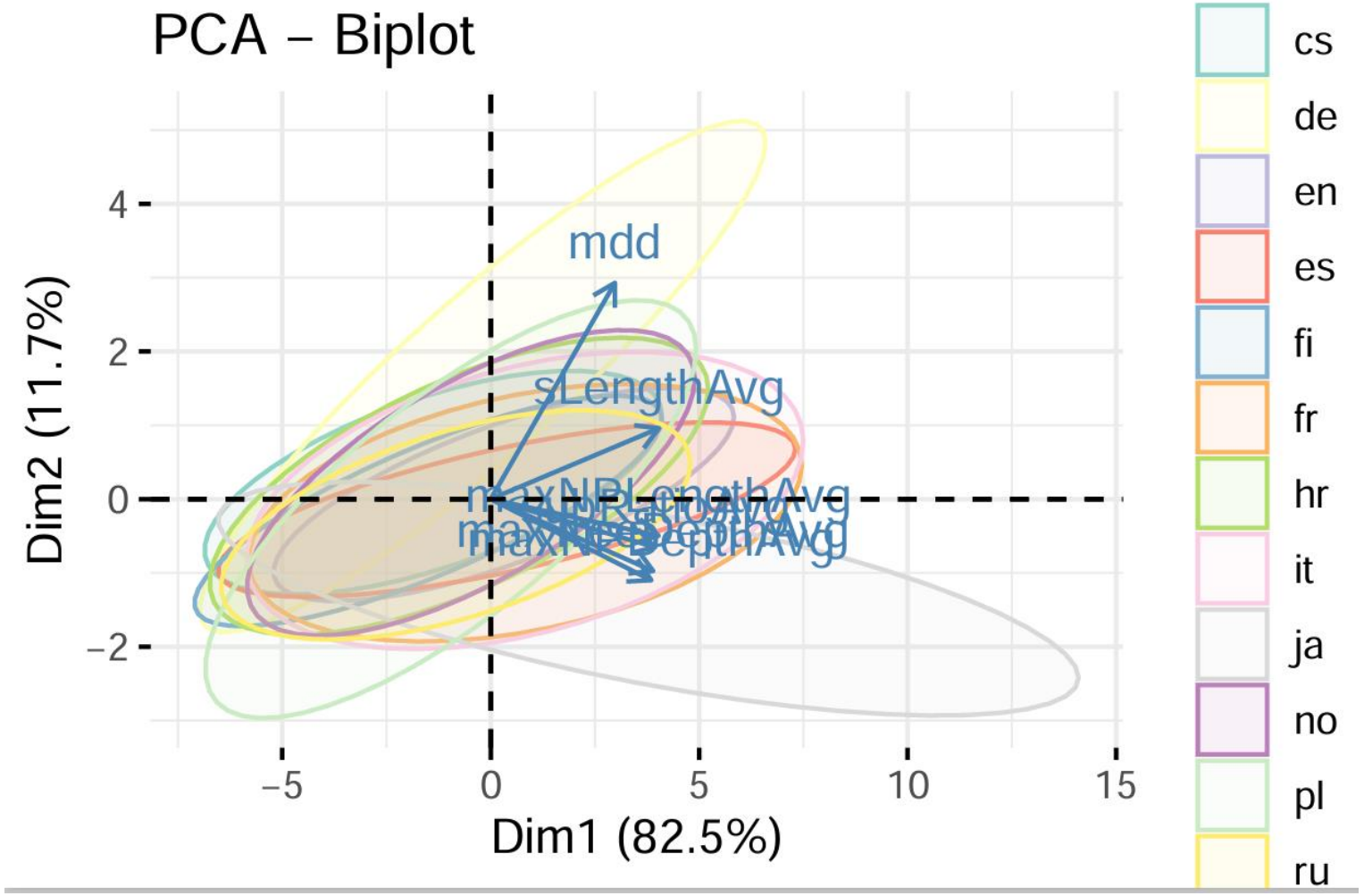
CONSEQUENCE: false differences in subRatio. BUT: only 5% clauses are **acl.**

PCA of 6 texts in 12 languages

- subRatio, maxTreeDepth, maxNPdepth & -length **vs** s_length & mdd
- mdd higher in **ja** and **de** (SOV?)

TODO:

- Larger, more varied sample
- More representative in terms of languages, language families and text types



Perspectives – what next?

- More reliable statistics
- More explanations
- Metrics as a web service

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

I am grateful to:

- *Olga Nádvorníková*
- *Martin Vavřín*
- *Bohumil Šimčík*
- *Jiří Milička (lexical diversity)*

for the idea, design and implementation

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

Outline

1. Linguistic variation within and across languages
2. Metrics of syntactic complexity and lexical diversity
3. The data: InterCorp – a multilingual parallel corpus
4. Accessing the annotation via search interface
5. Using the metrics
6. Perspectives, questions, discussion
7. References

Alemany-Puig, L., and Ferrer-i-Cancho, R. (2024). The expected sum of edge lengths in planar linearizations of trees. *Journal of Language Modelling* 12(1): 1–42.

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*, 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. 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)*, 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): 288-313.
Fabricius-Hansen, C. (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.
Johansson, V. (2008) Lexical diversity and lexical density in speech and writing: a developmental perspective, *Working Papers* 53, 61-79, Lund University, Dept. of Linguistics and Phonetics
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
Mačutek, J., Čech, R., and Milička J. (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, O. and Šotolová, J. (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*

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.

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.