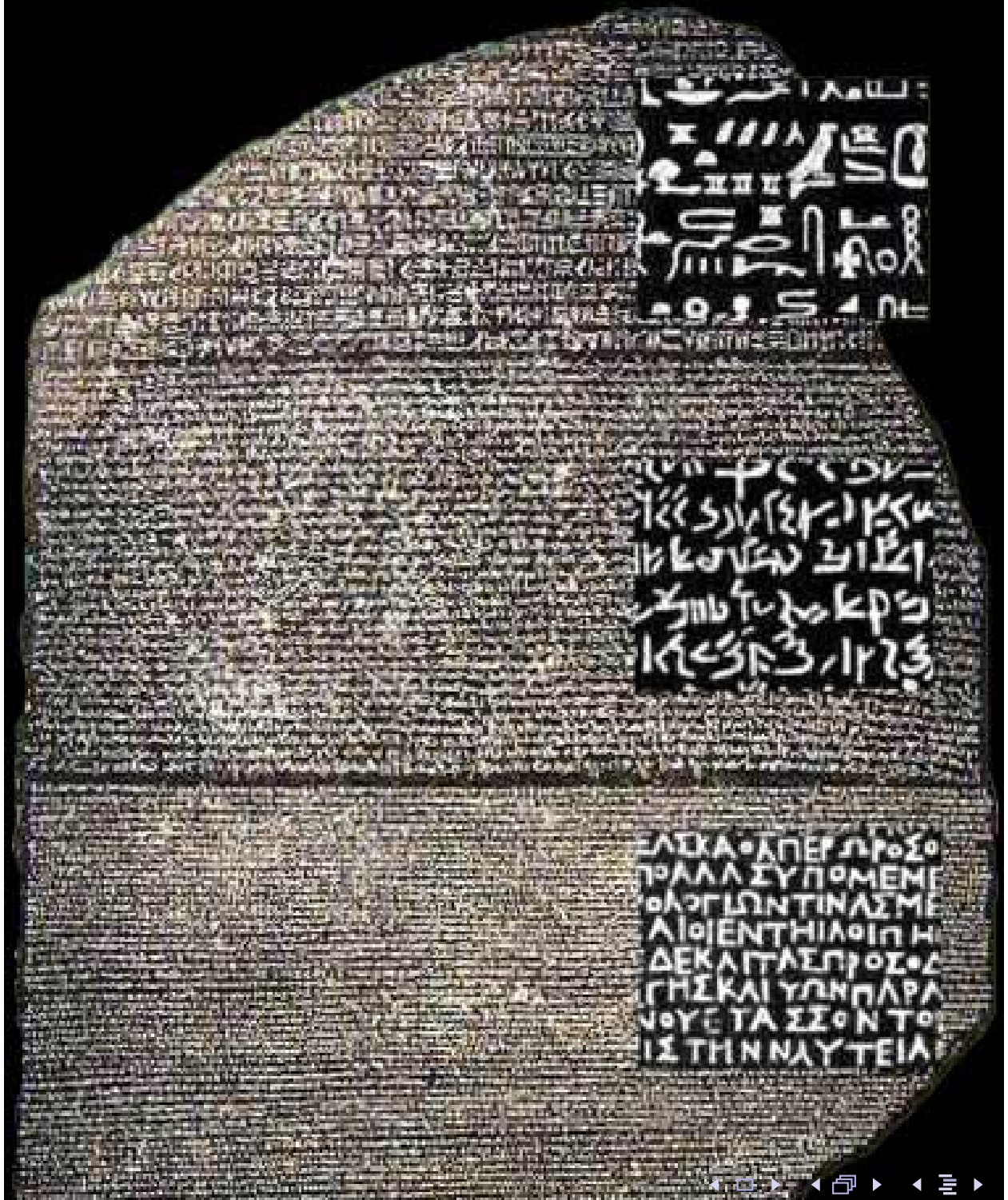


The InterCorp multilingual parallel corpus: representation of grammatical categories

Alexandr Rosen

Institute of Theoretical and Computational Linguistics
Institute of the Czech National Corpus
Charles University, Faculty of Arts

Seminarium *Przetwarzanie języka naturalnego*
Zespół Inżynierii Lingwistycznej
Instytut Podstaw Informatyki Polskiej Akademii Nauk
Warszawa, 18 listopad 2019

Access to the search interface

Access without login

- The Czech National Corpus: kontext.korpus.cz
- InterCorp Polish: https://tinyurl.com/intercorp
- Access is possible without login, but options are restricted
- Same login for other parts of the Czech National Corpus

Several login options

- After registration (free): https://www.korpus.cz/signup
- Using institutional login: https://www.korpus.cz/login
- Using temporary login: https://www.korpus.cz/login
  - Username: seminar
  - Password: kostka

Outline

Outline

What makes a corpus parallel?

- Same text in **multiple versions** (languages, translations, ...)
- **Alignment** by text units (texts, paragraphs, **sentences**, words)
- From **parallel texts**, translated or created as multilingual
- Parallel corpora (in disguise) are **used by**:
  - **NLP** tools: machine translation, information retrieval, projecting annotation, ...
  - **translators** (Computer-Assisted Translation)
  - experts in **Foreign Language Teaching**
  - **lexicographers**
  - **translatologists**
  - ...

Problems

- Authenticity
  - translationese
- Availability
  - not in all languages, genres, text types
  - legal restrictions
- Alignment
  - not error-free
- Specific tools needed
  - aligners
  - parallel concordancers

What a parallel corpus offers

- Translation preserves meaning
- Parallel context
  - → explicit translation equivalence
  - → implicit annotation of meaning
- From meaning to form:
  - find equivalents in another or the same language
  - – translation studies, contrastive linguistics, FLT, MT, CAT
- From form to meaning:
  - find meaning through other languages
  - – text understanding, annotation projection, monolingual lexicography

Beyond alignment ...

- Linguistic annotation for multiple languages

→ implicit meaning equivalence via alignment

→ explicit meaning equivalence by common annotation?

Outline

About InterCorp

- A part of the *Czech National Corpus*,
  built by *The Institute of the* $\sim$, Charles University
- http://www.korpus.cz/intercorp/
- *2005 – kudos to František Čermák
- At first as a service for departments of the Arts Faculty
- On line since 2008
- New releases once per year

The architecture of *InterCorp*

- Alignment: sentence-level
- Each text in Czech and at least one other language
- Tags and lemmas for most languages
- Rich metadata, esp. about the Core part (fiction)

History

- Figures for all languages except Czech – 200 M words in v.12

v.	Year	M Words	Langs	MSD	Milestones
0	2008	25	19	0	ParaConc, Park
1	2009	35	20	10	MSD
2	2009	49	21	10	*Project Syndicate*, monolingual corpora
3	2011	72	22	13	stand-off alignment
4	2011	92	22	13	*Presseurop*
5	2012	543	27	17	*Acquis*
6	2013	867	31	17	*ASPAC*, *Europarl*, Nosketch Engine
7	2014	1,390	38	20	*Subtitles*, KonText
8	2015	1,423	38	20	Treq, Intertext
9	2016	1,460	39	23	text planning
10	2017	1,484	39	23	*The Bible*, Treq v.2
11	2018	1,508	39	26	ja tagged; uk, be by UD tags[1]
12	2019	1,533	40	27	zh added and tagged
					+ a release annotated according to UD

[1] Universal Dependencies https://universaldependencies.org

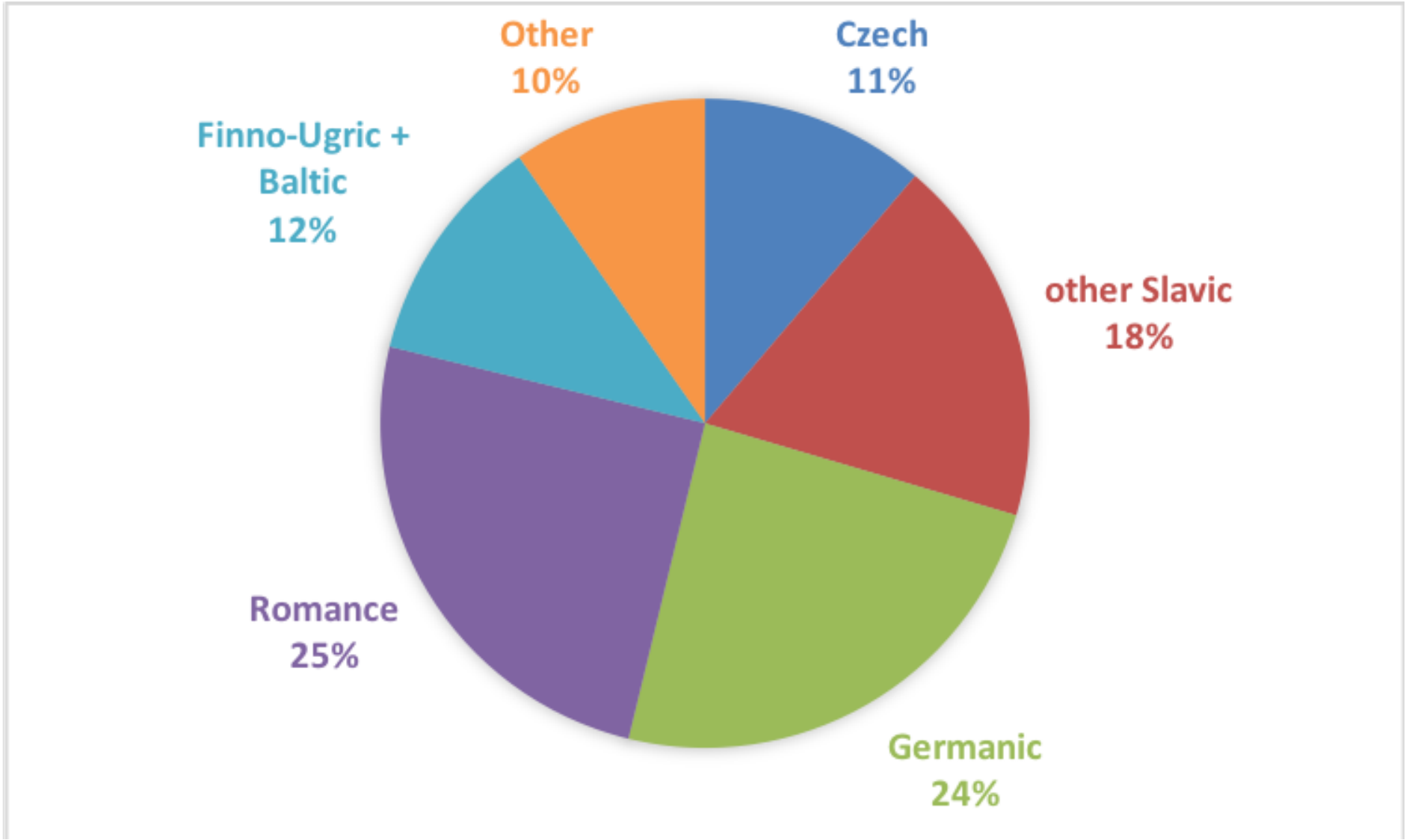
Content (release 12)

40 languages + Czech

- 10 Slavic: be, bg, hr, mk, pl, ru, sk, sl, sr, uk
- 7 Germanic: da, de, en, is, nl, no, sv
- 6 Romance: ca, es, fr, it, pt, ro
- 5 Finno-Ugric + Baltic: et, fi, hu, lt, lv
- 12 other: ar, el, he, hi, ja, ms, mt, rn, sq, tr, vi, zh

- ☛ Only few texts are available in more than 20 languages
- ☛ Languages differ wildly in the volumes of text

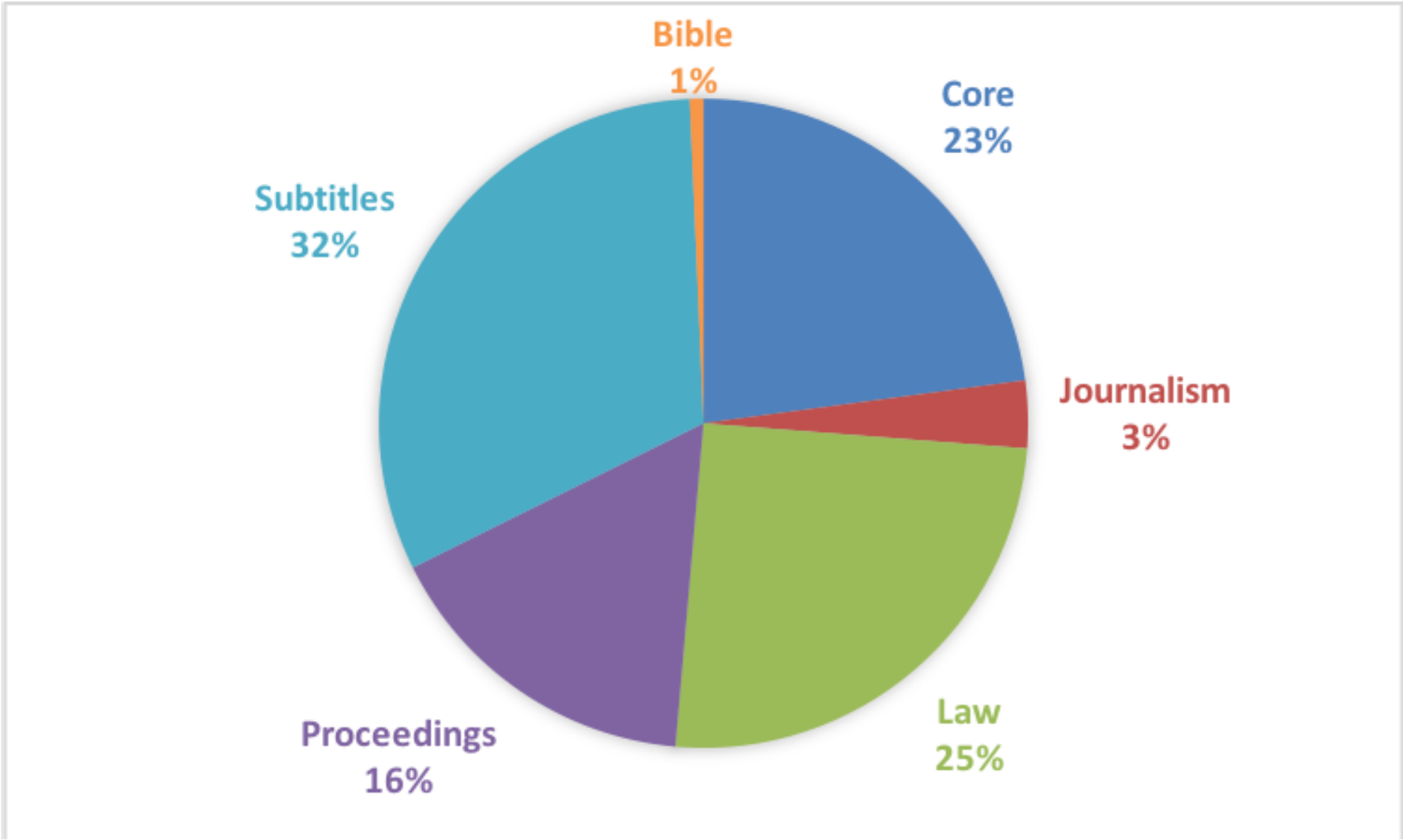
Language groups



Text types

- **Total size** – 1.7 billion words
- **Core** – mostly fiction, proofread
- **Collections** – freely available texts
  - **Journalism**
    Project Syndicate http://www.project-syndicate.org/
    VoxEurope http://www.voxeurop.eu/
  - **Law**
    Acquis Communautaire
    http://langtech.jrc.ec.europa.eu/JRC-Acquis.html
  - **Parliament proceedings**
    Europarl http://www.statmt.org/europarl/
  - **Film subtitles**
    Open Subtitles http://www.opensubtitles.org
  - **The Bible**

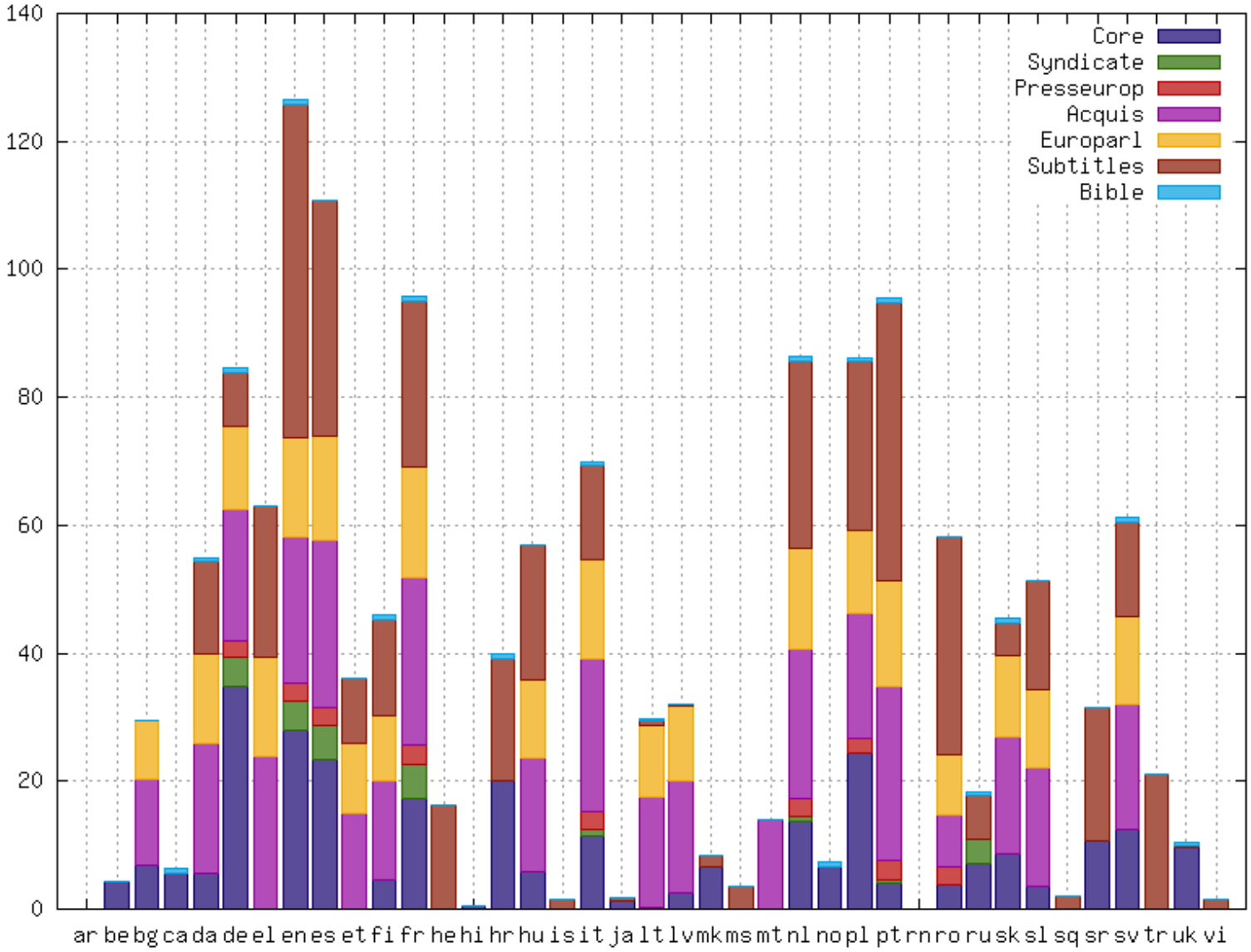
Text types


Bible
1%
Core
23%
Subtitles
32%
Journalism
3%
Proceedings
16%
Law
25%

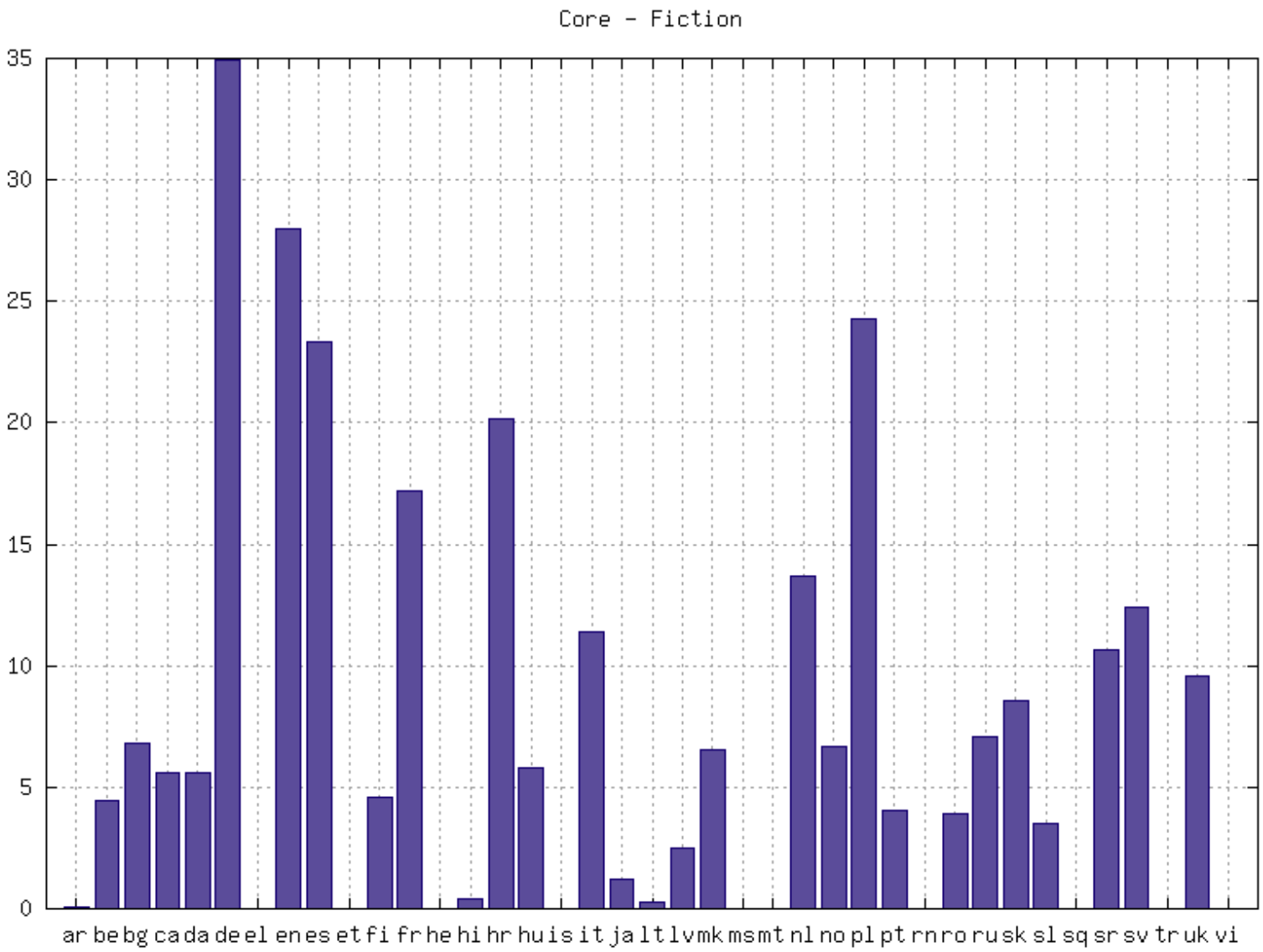
Size in million words (v.11)

	Czech	Polish	Other Slavic	Other foreign	**Total**
Core	106.9	24.3	77.2	186.0	**390.0**
Journalism	6.4	2.4	3.8	44.6	**57.2**
Law	19.0	19.6	50.5	339.3	**428.5**
Proceedings	12.2	12.8	34.1	218.4	**277.5**
Subtitles	50.6	26.6	70.9	391.4	**539.5**
Bible	0.6	0.6	2.3	8.1	**11.6**
Total	**195.8**	**86.2**	**238.8**	**1187.9**	**1704.2**
# core texts	1,564	293	1,083	2,118	**5,058**

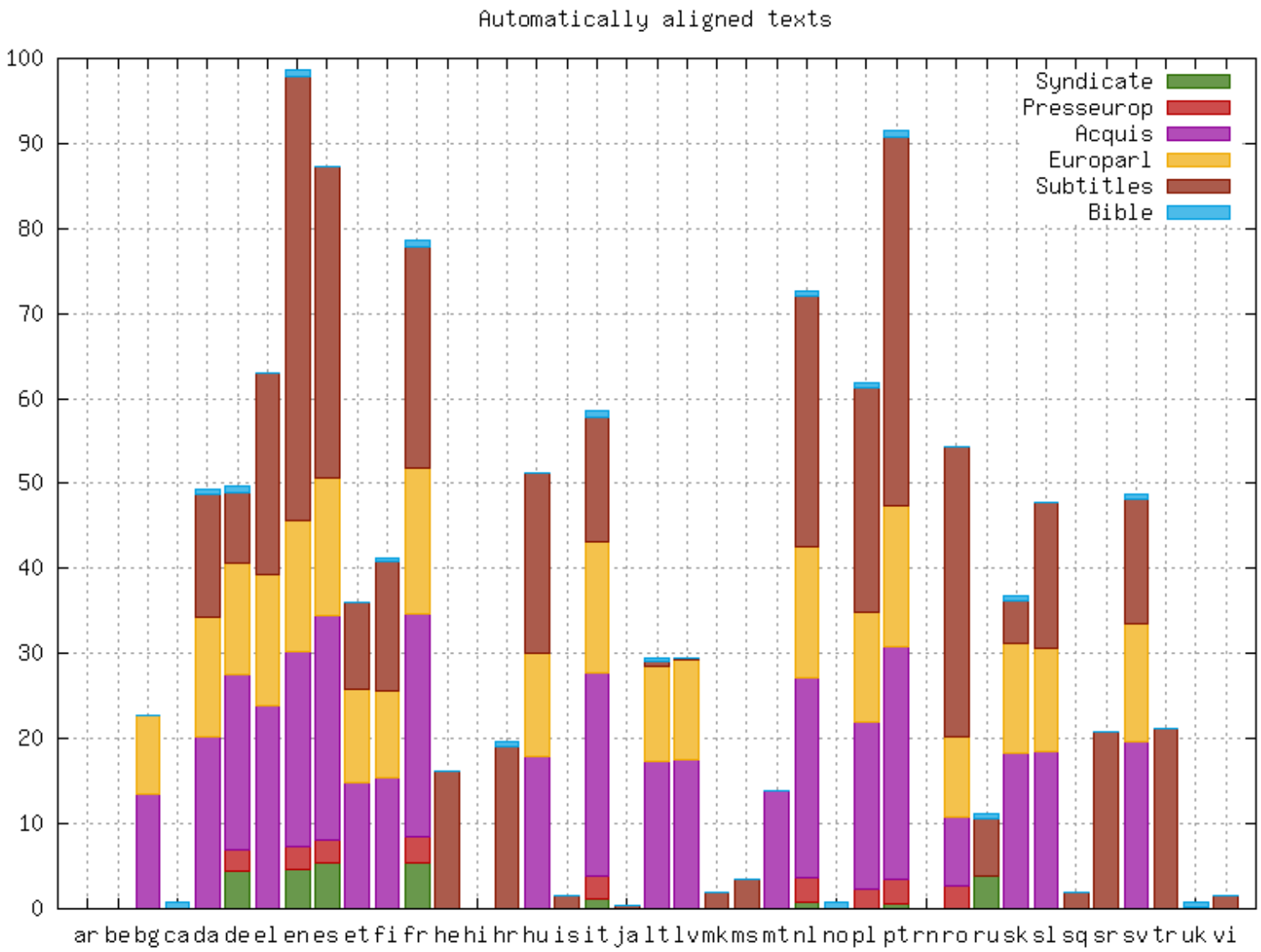
InterCorp by languages and text types



Core (mostly fiction)



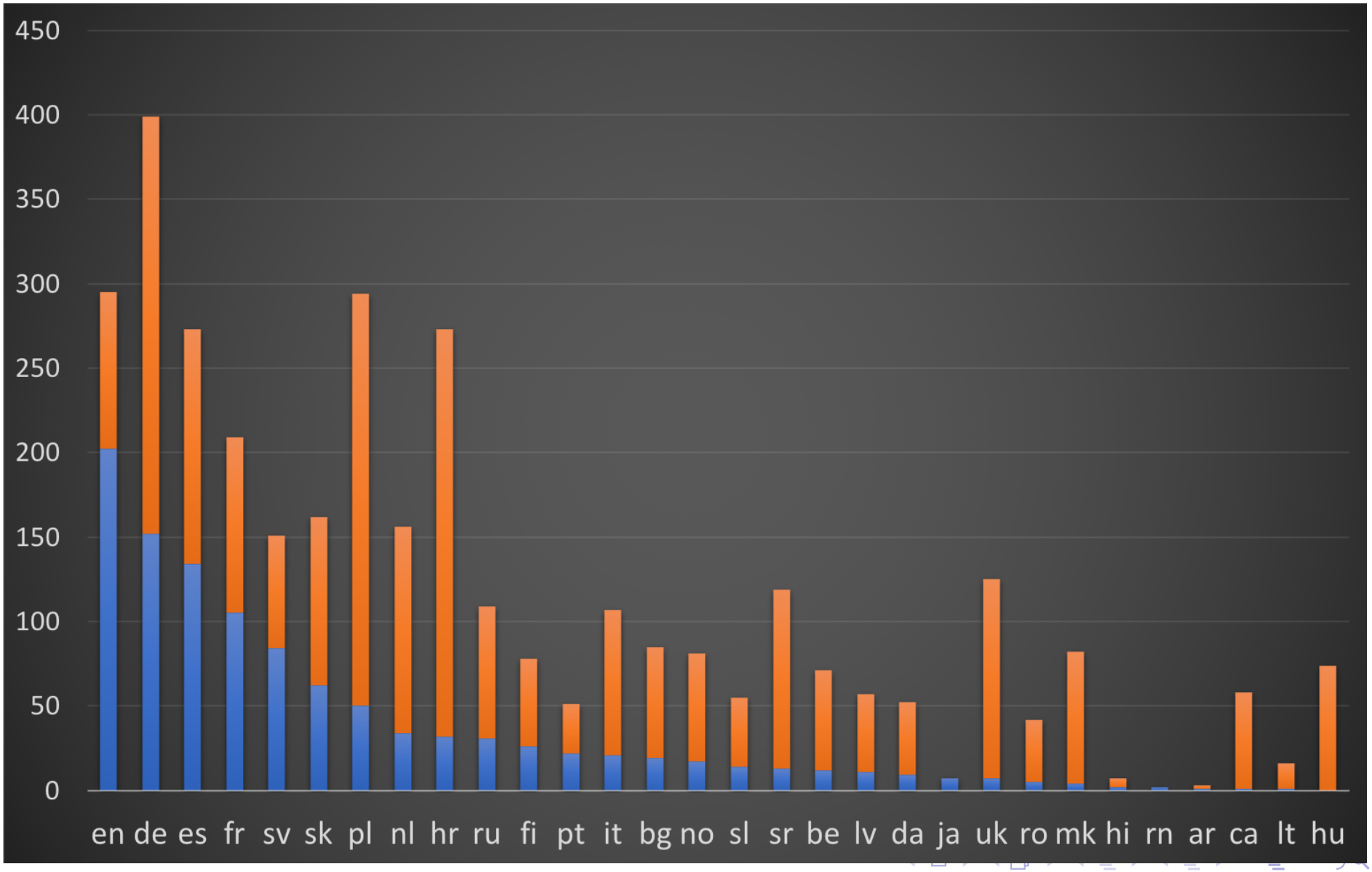
Collections (journalism, law, parliament proceedings)



The Core

- The average number of texts per title: 3.24
- For all languages: 1400 originals (38%), 3657 translations
- For Czech: 320 originals (26%), 1244 translations
- Titles without the original version: 173

Originals and translations in the Core



Slavic	Other	Author	Title
11	15	Rowling	*Harry Potter and the Philosopher's Stone*
11	15	Saint-Exupéry	*The Little Prince*
11	12	Carroll	*Alice in Wonderland*
11	12	Orwell	*1984*
11	12	Tolkien	*The Hobbit or There and Back Again*
11	8	Bulgakov	*The Master and Margarita*
11	7	Milne	*Winnie the Pooh*
11	3	Ostrovsky	*How the Steel Was Tempered*
10	10	Adams	*The Hitch Hiker's Guide to the Galaxy*
10	10	Brown	*The Da Vinci Code*
10	9	Frank	*The Diary of a Young Girl*
10	8	Hašek	*The Good Soldier Švejk*
10	5	Kipling	*The Jungle Book*
10	13	Kundera	*The Unbearable Lightness of Being*
9	12	Coelho	*The Alchemist*
9	11	Tolkien	*The Fellowship of the Ring*
9	11	Tolkien	*The Return of the King*
9	9	Orwell	*Animal Farm*
9	6	Hemingway	*The Old Man and the Sea*
8	12	Rowling	*Harry Potter and the Chamber of Secrets*
8	12	Rowling	*Harry Potter and the Prisoner of Azkaban*
8	11	Kafka	*The Trial*
8	10	Eco	*The Name of Rose*
8	10	Tolkien	*The Two Towers*
8	9	Rowling	*Harry Potter and the Goblet of Fire*
8	8	Brown	*Angels and Demons*
8	3	Lem	*Solaris*
7	10	Hrabal	*I Served the King of England*

Slavic	Other	Author	Title
7	2	Andrić	*The bridge on the Drina*
6	10	Kundera	*Immortality*
6	9	Kundera	*Laughable Loves*
6	5	Ouředník	*Europeana*
6	1	Gombrowicz	*Ferdydurke*
6	0	Tokarczuk	*Primeval and Other Times*
5	14	Kundera	*The Joke*
5	9	Čapek	*War with the Newts*
5	6	Viewegh	*Bringing up Girls in Bohemia*
5	2	Čapek	*Dashenka or the Life of a Puppy*
5	2	Petrov	*The Twelve Chairs*
5	1	Bass	*Klapzuba's Soccer Team*
5	1	Gombrowicz	*Pornografia*
5	0	Dousková	*B. Proudew*
4	10	Kundera	*Farewell Waltz*
4	8	Hrabal	*Too loud a solitude*
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...

Increase in the Core of v.12

	+ th. words	+ %
en	4,185	14.96%
cs	3,675	3.44%
sv	2,924	23.59%
es	2,246	9.62%
ca	1,884	33.67%
it	1,851	16.24%
mk	1,492	22.93%
de	1,458	4.18%
nl	1,438	10.51%
pl	1,315	5.41%
uk	1,246	13.02%
ru	1,061	15.02%

	+ th. words	+ %
da	1,058	18.92%
fr	1,028	5.97%
be	893	20.17%
hr	879	4.37%
fi	745	16.25%
ja	548	45.77%
sr	300	2.81%
bg	288	4.25%
sl	207	5.85%
lt	134	46.77%
lv	122	4.84%
pt	64	1.58%

Total	31,039	7.96%

To do

- More **representative/balanced** core
  genres, periods, originals/translations, authors, translators
  – needed for both contrastive and translatological studies
- **The more the better**
  – the overlap may be too small even for languages such as English or German
- The **original text** should always be included
- **Multiple translations** in a single language

Outline

Name	Types	Langs	Size	Annot	Aligned	Proofread	Search	Download	Metadata
Linguee	legal	25	?	no	S,W	no	yes	no	yes
Glosbe	varia	100+	1Bs	no	S,W	no	yes	no	yes
SKE	varia	38	217M cs	no	S	no	yes	yes	yes
DGT-TM	legal	22	3.7Ms cs	no	S	yes	no	yes	no
Pelcra	varia	31	58M pl	no	S,W	part	no	yes	yes
RNC	varia	6	9M	M	S	part	yes	?	yes
SNK	fiction	7	388M sk	M	S	no	yes	part	yes
CzEng	varia	en, cs	233M en	M,Sy	S	no	sample	yes	no
PCEDT	news	en, cs	1.2M.	M,Sy,Se	S,W	yes	yes	yes	yes
Kačenka	fiction	en, cs	3.3M	no	S	yes	no	yes	yes
Opus	varia	100+	4.7B	M,Sy	S,W	no	yes	yes	no
Parasol	fiction	31	27M	M	S	part	yes	?	yes
ASPAC	fiction	25	68t	no	P	yes	no	?	yes
InterCorp	varia	32	1.6B	M	S	part	yes	yes	yes

- **Linguee**: online search through bilingual texts – http://www.linguee.com
- **Glosbe**: Translation Memory Online – http://glosbe.com/tmem/
- **SKE**: Sketch Engine – http://www.sketchengine.co.uk
- **DGT-TM**: Translation Memory of the European Commission's Directorate-General for Translation – http://ipsc.jrc.ec.europa.eu/?id=197
- **Pelcra**: Polish & English Language Corpora for Research & Applications – http://pelcra.pl/new/
- **RNC**: Russian National Corpus – http://www.ruscorpora.ru
- **SNK**: Slovak National Corpus – http://korpus.juls.savba.sk/par.html
- **CzEng**: Czech-English parallel corpus – http://ufal.mff.cuni.cz/czeng
- **PCEDT**: Prague Czech-English Dependency Treebank – http://ufal.mff.cuni.cz/prague-czech-english-dependency-treebank
- **Kačenka**: English-Czech Corpus of the Department of English Studies, Faculty of Arts, Masaryk University Brno – http://www.phil.muni.cz/angl/kacenka/kachna.html

What makes InterCorp different?

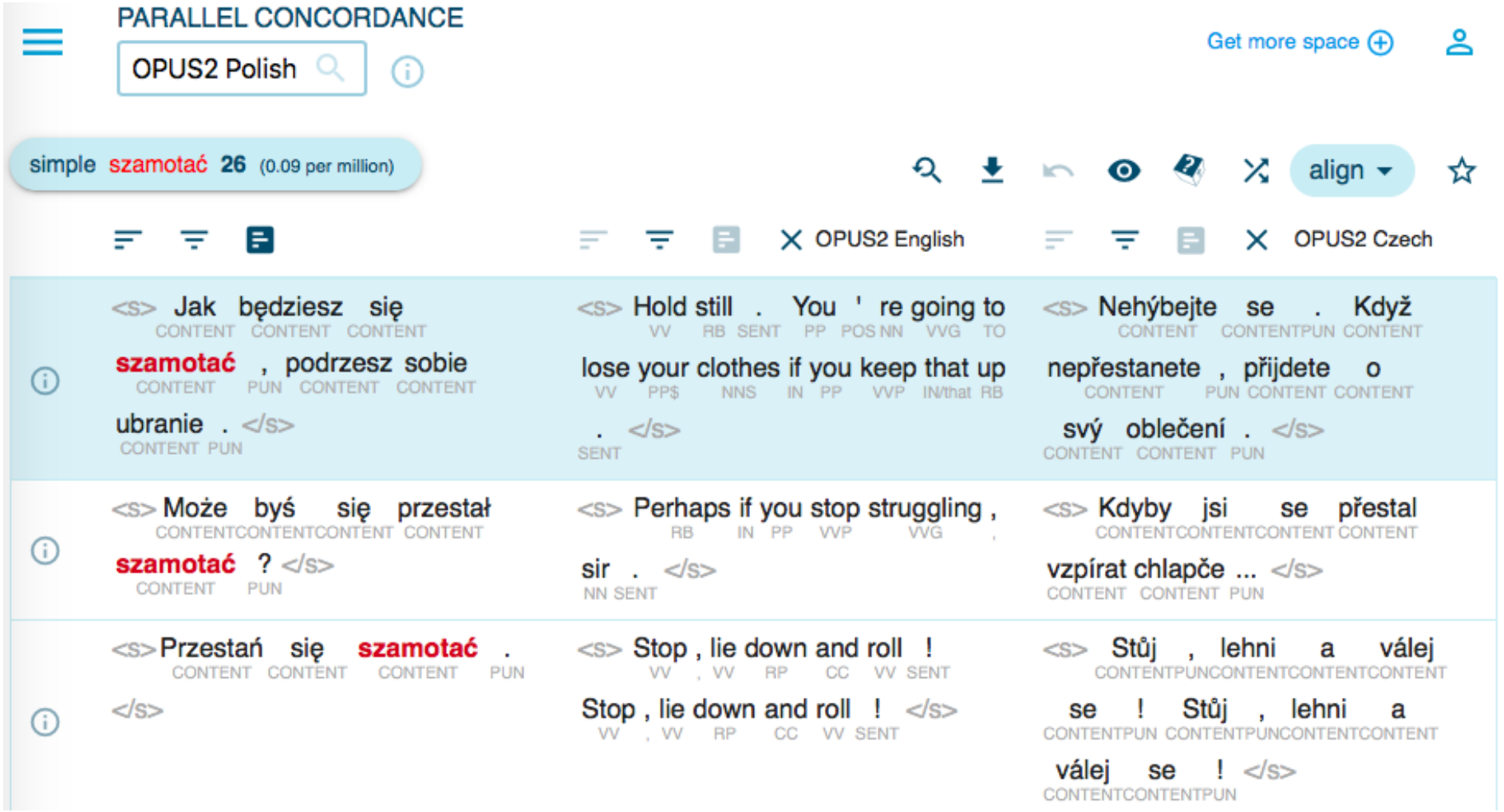
- A substantial share of fiction
- Manually checked
- Rich metadata
- Same search interface as other monolingual corpora of CNC
- Privileged users co-responsible for “their” language

Outline

Concordances

KonText – http://kontext.korpus.cz

- A web search interface based on *Manatee* (like *Sketch Engine*)
- Available from https://github.com/czcorpus/kontext
- Used also by Lindat/Clarin: https://lindat.mff.cuni.cz/services/kontext/
- Search filters:
  - languages, corpus release
  - text, publication year, text type
  - original/translation, language of the original
  - author, translator, ...
- Parallel queries, CQL
- Positive and negative filters on the concordances
- Sorting, frequency distribution, collocations
- Custom subcorpora, export of concordances


PARALLEL CONCORDANCE
OPUS2 Polish
Get more space
simple szamotać 26 (0.09 per million)
align
OPUS2 English
OPUS2 Czech
<s> Jak będziesz się
CONTENT CONTENT CONTENT
szamotać , podrzesz sobie
CONTENT PUN CONTENT CONTENT
ubranie . </s>
CONTENT PUN
<s> Hold still . You ' re going to
VV RB SENT PP POS NN VVG TO
lose your clothes if you keep that up
VV PP$ NNS IN PP VVP IN/that RB
. </s>
SENT
<s> Nehýbejte se . Když
CONTENT CONTENTPUN CONTENT
nepřestanete , přijdete o
CONTENT PUN CONTENT CONTENT
svý oblečení . </s>
CONTENT CONTENT PUN
<s> Może byś się przestał
CONTENTCONTENTCONTENT CONTENT
szamotać ? </s>
CONTENT PUN
<s> Perhaps if you stop struggling ,
RB IN PP VVP VVG ,
sir . </s>
NN SENT
<s> Kdyby jsi se přestal
CONTENTCONTENTCONTENT CONTENT
vzpírat chlapče ... </s>
CONTENT CONTENT PUN
<s> Przestań się szamotać .
CONTENT CONTENT CONTENT PUN
</s>
<s> Stop , lie down and roll !
VV , VV RP CC VV SENT
Stop , lie down and roll ! </s>
VV , VV RP CC VV SENT
<s> Stůj , lehni a válej
CONTENTPUNCONTENTCONTENTCONTENT
se ! Stůj , lehni a
CONTENTPUN CONTENTPUNCONTENTCONTENT
válej se ! </s>
CONTENTCONTENTPUN

Hits: **88** | i.p.m.: **0.8** (related to the whole corpus) | ARF: **29.87** | Result is shuffled 1 / 3

Line selection: simple

	InterCorp v11 - Polish	InterCorp v11 - English	InterCorp v11 - Czech
☐	Syrena , ostrzegająca przed mgłą , wyła bez przerwy w cieśninie i na wpół chory **szamotał** em się pomiędzy groteskową rzeczywistością i dzikimi , strasznymi snami .	I COULD N'T sleep all night ; a fog-horn was groaning incessantly on the Sound , and I tossed half-sick between grotesque reality and savage , frightening dreams .	výstražný signál neustále sténal v mlze Průlivu a já se napůl nemocný zmítal mezi groteskní skutečností a divokými , strašidelnými sny .
☐	- Nie ! - zawołał Oliwer **szamocąc** się .	' Do n't , ' cried Oliver , struggling .	" Nechte mě ! " vykřikl Oliver a vzpíral se .
☐	Dopiero wtedy pojęli , że w górze na rusztowaniach sejmeni z kimś się **szamocą** .	Only then did they realize that on the scaffolding above them the guards were struggling with something .	Teprve tehdy rozeznali , že se nahoře na lešení drábi s někým rvou .
☐	Wszyscy cofali się jak na rozkaz , a sprzedawca z Woking dalej **szamotał** się na krawędzi jamy .	There was a general movement backwards . I saw the shopman struggling still on the edge of the pit .	Nastal obecný ústup . Zahlédl jsem příručího , jak se marně drápe ven z jámy .
☐	Łzy lały się strumieniami z jej płonących nienawiścią oczu i **szamotała** się z nim szaleńczo w irracjonalnym napadzie wściekłości , warcząc , bluzgając przekleństwami i wrzeszcząc :	Tears were streaming in single torrents from her flashing , hate-filled eyes as she struggled against him fiercely in an irrational frenzy of maddened might , snarling and cursing savagely and screaming ' Bruto !	Z planoucích , nenávistných očí jí tekly proudy slz , ale nepřestávala se rvát jako saň . Zuřivost jí dodávala neuvěřitelnou sílu .
☐	A jeden z tych , co wpadli , chłop na schwał , morda zbójecka , nie daje się , **szamoce** .	One of them , a great hulking fellow with a mug that just asked for a brick , was kicking and struggling for all he was worth .	Jeden z nich , takový silný kluk , s červenou hubou jak cihla , žebroní , nevzdává se , vzpouzí se .
☐	Lo **szamotała** się jeszcze z ubraniem i klęła mnie słowami , o jakich nigdy by m nie pomyślał , że małe dziewczynki mogą je znać , a co dopiero ich używać .	and with a scrunch and a skid we drove off , Lo still struggling with her clothes and swearing at me in language that I never dreamed little girls could know , let alone use .	a pak jsme skřípavě a smykem vyrazili , zatímco Lo se stále ještě potýkala se šaty a nadávala mi jazykem , o němž se mi jaktěživ nesnilo , že by ho malá děvčátka mohla znát , tím méně používat .
☐	W gabinecie zapadło milczenie , słychać było tylko odgłosy **szamotania** się Ślizgonów z Ronem i resztą więźniów .	There was silence in the office except for the fidgetings and scufflings resulting from the Slytherins ' efforts to keep Ron and the others under control .	V kabinetě bylo ticho až na neklidné a šoupavé zvuky , jak se zmijozelští snažili udržet Rona a ostatní pod kontrolou .

To do

- **biKWiC** – highlighting keyword equivalent
- Info on **alignment**: 1:1 / 2:1 / 1:2 / automatic / manual / confidence score
- Lexical profiles (similar to **Word Sketches** [Kilgarriff et al.(2014)])
- **Alignment** by words, multiword units, constituents
- **Crowdsourcing** to eliminate annotation errors

Lexical lookup

Treq – a database of translation equivalents – http://treq.korpus.cz

- Partial substitute for missing representation of word-to-word alignment in concordances
- Pairs of lexical equivalents extracted from word-aligned parallel texts
- cs/en $\leftrightarrow$ any other language
- Filtering by text groups
- Queries by forms or lemmas
- Support of regular expressions
- Single forms or multi-word expressions
- At the moment still based on texts of *InterCorp* v.9

Source language: Polish | Target language: English | Restrict to: Collection(s): 5

Ver. 2.0

szamotać

Search

Lemma | Multiword | RegEx | A = a

Frequency	Proportion	Polish	English
9	37.5	szamotać	struggle
3	12.5	szamotać	fight
2	8.3	szamotać	tussle
1	4.2	szamotać	scrunch
1	4.2	szamotać	haul
1	4.2	szamotać	flutter
1	4.2	szamotać	fidgetings
1	4.2	szamotać	reel
1	4.2	szamotać	thump
1	4.2	szamotać	scramble
1	4.2	szamotać	grapple
1	4.2	szamotać	flounder
1	4.2	szamotać	blunder
24			

Dissemination of texts

- Technical protection against misuse:
  shuffled order of blocks of translation pairs
- Educational and research licence, no re-distribution

Some usage statistics

Number of queries (2017–2018)

- 553 per day
- 89% select ALL text types
- 2.6% select 3 or more languages, 13.4% a single language
- Most frequent language pairs: en/de↔cs (58%)
- Queries involving Polish: 3.5%

Books, theses, papers based on InterCorp

- https://www.korpus.cz/biblio: 113 entries
- https://ukaz.cuni.cz/: 266 entries
- https://www.researchgate.net/: 74 entries
- https://scholar.google.cz/: 2650 entries

Outline

Pre-processing

1. Text choice
2. Scanning & character recognition
3. Proofreading
4. Segmentation (sentence boundary detection)
5. Alignment
6. Checking of segmentation and alignment
7. Tokenization, morphosyntactic markup
8. Adding metadata
9. Indexing by the corpus search tool

Tools used in pre-processing

1. Bibliographical database
2. *Intertext*[a] – alignment editor
3. *Punkt*[b] – sentence splitter
4. *Hunalign*[c] – aligner
5. Language-specific tokenizers and taggers
6. Coming soon: *UDPipe*[d] – a multilingual tokenizer, tagger and parser, providing (almost) language-universal annotation

[a] https://wanthalf.saga.cz/intertext
[b] https://www.nltk.org/_modules/nltk/tokenize/punkt.html
[c] http://mokk.bme.hu/en/resources/hunalign/
[d] http://ufal.mff.cuni.cz/udpipe

Outline

Linguistic annotation

- tokens
- lemmas
- morphosyntactic tags
- *syntactic functions and structure*

Current strategy – use available taggers, including:

- tokenizers bundled with the tool
- tagsets designed elsewhere
- annotation models trained elsewhere

Tools for tokenization, lemmatization and tagging

Lng	Tool	Preposition Determiner Adjective Noun
be	UDPipe	ADP ADJ Case=Loc Degree=Pos Gender=Masc Number=Sing NOUN Animacy=Inan Case=Loc .
bg	TreeTagger	R Pde-os-n Ansi Ncnsi
ca	TreeTagger	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	TreeTagger	IN DT JJS NNS
es	TreeTagger	PREP ART NC ADJ
et	TreeTagger	P.sg.gen A.pos.sg.gen S.com.sg.kom
fi	TurkuNPP	A:Sg:Gen:Pos N:Sg:Gen Adp:Po
fr	TreeTagger	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	TreeTagger	PRE PRO:demo NOM ADJ
ja	MeCab	連体詞 形容詞 名詞 助詞-格 助詞
lv	LVTagger	spsgy pd0msgn afmsgyp ncmsg1
nl	TreeTagger	prep det__demo adj nounpl
no	VISL	600 370 103 000 prep det adj subst
pl	KRNNT	prep:loc:nwok adj:sg:loc:m3:pos adj:sg:loc:m3:pos subst:sg:loc:m3
pt	TreeTagger	SPS DA0 NCFS AQ0
ru	TreeTagger	Sp-l P--pl Afp-plf Ncmpln
sk	MorphoDita	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	UDPipe	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
zh	ZPar	P DT JJ NN

The problem?

Mismatching tagsets and tokenization

- Conversion into a common tagset?
- Re-tagging by a language-universal tool?
- Tokenization?
- Implementation in *Manatee/KonText*?

Tagsets – not only notational differences

- en: *under*, *because* – **IN** (both as prep and sconj)
- cs: *těch* – **PD** × pl: *tych* – **adj**
- cs: *devátá* – **Cr** × pl: *dziewiąta* – **adj**
- en: *remotest* – **JJS** × de: *abgelegenste* – **ADJA**

Tokenization – similar phenomena treated in opposite ways

- *abyste, udělals, tys, očs, zum, aux*
  × *że by śmy, zrobił eś, ty ś, do ń, gdzieś/gdzie ś, ca n't, I 'm*
  × 豚$_{N}$ みたい$_{AdjN}$ な$_{Pcle}$ 顔$_{N}$ を$_{Pcle}$ する$_{V}$ の$_{Pcle}$ は$_{Pcle}$ よせ$_{V}$!$_{P}$
- *cure-dents, gut-ausgearbeitet, Jelzin-Ära, franco-tedesco, česko-polský, Tchaj-wan*
  × *padne - li, Frýdek - Místek, polsko - czeski, Bielsko - Biała*

Solutions?

- Conversion of language-specific tags into a common taxonomy
  - Ontology of categories – OWL/DL?[a]
  - Multidimensional taxonomy of categories?[b]
  - Existing standard tagset – UD?
- Retagging by a tool trained on a shared tagset
  - Enough training data?
  - Language-specific tags still available in the corpus as an option?

[a] [Chiarcos & Erjavec(2011)], [Chiarcos(2012)]
[b] [Rosen(2014)]

Things to remember:

- Tokenization
- Implementation in in *Manatee/KonText*

Multidimensional taxonomy of word classes

Criteria for distinguishing word classes

- **Semantic** (content-based, lexical)
- **Syntactic** (functional, distributional)
- **Morphological** (inflectional)

Two solutions: ([Komárek et al.(1986)])

- Apply the criteria in parallel, each useful for a purpose: a cross-classification
- Take one of the criteria as the main one, others as complementary
  - Semantics (Jespersen, [Brøndal(1928)], Vinogradov, Tesnière,...)
  - Morphology[2] ([Saloni & Świdziński(1985), 95])
  - Syntax ([Grzegorczykowa et al.(1998), 59])
  - Syntax/Morphology ([Komárek et al.(1986), 13–16])

[2] 'For richly inflected languages morphological criterion is best.'

A parallel corpus in 3D

Cross-classification to embrace existing tagsets:

- Czech (ÚFAL): preference for **semantic** classes
  - *těch* is a **pronoun** rather than adjective or determiner
  - *devátá* is a **numeral** rather than an adjective
- Polish (IPI PAN): **inflectional** classes
  - *tych* and *dziewiąta* are **adjectives** rather than pronouns or numerals
- German (STTS): preference for **syntactic** classes
  - *abgelegenste* is **attributive** adjective rather than a superlative form

- ☛ In distinct dimensions, *dziewiąta* can be numeral and adjective at the same time
- ☛ A tag specifying just one dimension can still be mapped properly in a 3D space of categories

An existing standard

Universal Dependencies – http://universaldependencies.org/

- A de-facto standard also for morphological categories
- Potentially lossy conversion from language-specific tagsets
- Two-level tokenization – orthographical and syntactic words

UD Principles (abbreviated quote)

UD is a very subtle compromise between 6 things, it must be:

1. satisfactory on linguistic analysis grounds for individual languages
2. good for linguistic typology as a basis for bringing out parallelism across languages
3. suitable for rapid, consistent annotation by a human annotator
4. suitable for computer parsing with high accuracy
5. easily comprehended and used by a non-linguist $\rightarrow$ traditional grammar notions preferable
6. supporting downstream language understanding tasks

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD whileC remaining sensitive to all these dimensions.

Problems with UD?

- Nouns as NOUN or PROPN
- Participles as ADJ, NOUN or VERB, depending on the language and context
- Gerunds as VERB or NOUN, depending on the language and context
- Modals as VERB or AUX, depending on the language
- Ordinal numbers as ADJ or ADV
- DET for all quantifiers and pronouns in pre-nominal position (demonstrative, possessive, interrogative, relative, indefinite) pronouns

Why conversion into a common tagset could be problematic

PROS:

- Language-specific annotation (tags) can be retained

CONS:

- Conversion specifications for some tagsets are not available
- Lack or loss of information
- Potentially incompatible tokenization
- Possibly still conceptual differences in the result

Why retagging by UD could be easier

PROS:

- Toolchain exists for nearly all currently tagged and some additional languages
- Perspective for better coverage and results
- Annotation based on the same principles across languages
- Incompatible tokenization is not an issue
- Syntactic annotation as a bonus

CONS:

- Small training data resulting in poor quality for some languages
- Unrelated to existing language-specific annotation in IC, important to many users

Reconciling Universal Dependencies with Manatee

- UD distinguishes orthographic and syntactic words while Manatee has a single level of tokenization
  - syntactic or othographic words as Manatee tokens?
  - both options tested[3]
- UD represents categories as attributes and values while the standard in Manatee is a positional tagset
  - menu in the query inteface for queries concerning categories
- Manatee is not suited to represent syntactic structure
  - pointers to the position, form and lemma of the head
  - image of the syntax tree with parallel highlighting of text and tree nodes

[3] See, e.g. https://lindat.mff.cuni.cz/services/kontext/first_form?corpname=ud_23_cs_fictree_a

A fused word in the UD format (CoNLL-U), modified for Manatee

(1) Kdybychom nepřišli ...
'Gdybyśmy nie przyszli ...'

ID	1-2
WORD	Kdybychom
CFORM	kdy|bychom
IFORM	když|bychom
LEMMA	když|být
UPOS	SCONJ|AUX
FEATS	1:|2:|Mood=Cnd|Number=Plur|Person=1|VerbForm=Fin
HEAD	3|3
DEPREL	mark|aux

(2) Ale měla by **sis** uspořádat život.
ale miała by sobie+ś ułozić życie
'Ale miała byś sobie ułozić życie'

https://tinyurl.com/fusedsis

kontext Query Corpora Save Concordance Filter Frequency Collocations View Help

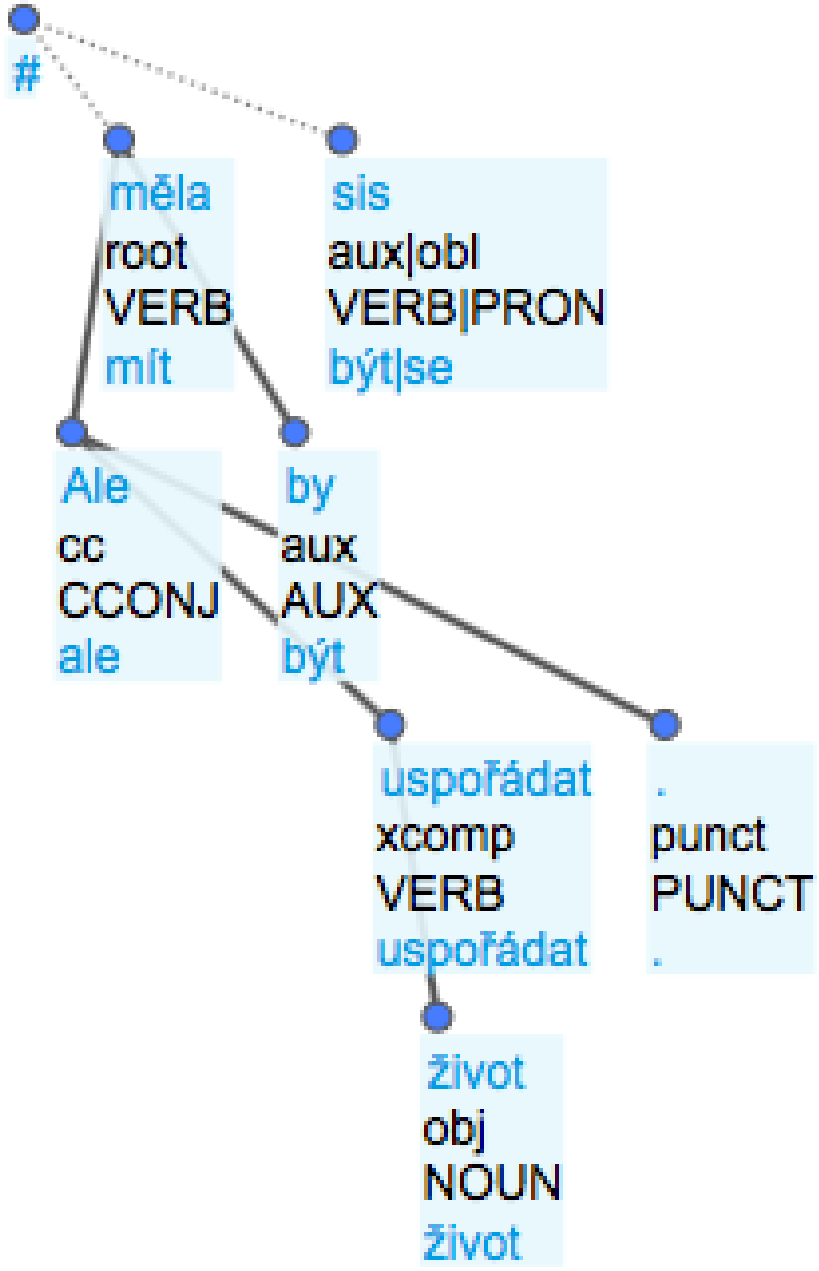
Corpus: UD 2.3 - Czech FicTree | **Query:** sis (14 hits) ▶ **Shuffle:** ✓

Hits: **14** | i.p.m.: **84.12** (related to the whole corpus) | ARF: **9.04** | Result is shuffled 1 / 1

Line selection: simple

test	pravil Leoš konejšivě . " Kdybys byla doma , mohla	sis	toho všimnout už minulý týden , když jsme byli spolu
train	zapomněla . Poslední dobou je čím dál senilnější , nevšim	sis	? pravil druhý . Jednoho dne si nás zavolal náš
train	, Simono , varoval . Včera tě chytly , jak	sis	nesla domů oběd . Abych to upřesnil , nejednalo se
train	tedy v nebi , " oslovil ho , " jak	sis	přál . Ale jak vidím , pořád čteš . "
train	. Ale ráda mne nemáš a já nechci , aby	sis	to namlouvala . " Šli pořád dál , mlčeli a
train	To jsem , " přiznala jsem stručně . " Zřejmě	sis	našla náhradu , " pokračoval kategoricky . " Ne .
dev	mně , ale kvůli mým malým dětem , jejichž matku	sis	už vzal . Dej mi ještě čas , abych vydělal
train	byl důrazně ohlášen . Nesnaž se teď tvrdit , že	sis	ničeho nevšiml . A teď jsem k tobě přišel já
train	ty budeš bezpochyby postupovat jinak než tvůj otec , abys	sis	zajistil budoucnost . Ale každý den jsem byl tady ,
train	ne ? To je logické . Ale vidím , že	sis	poradil sám . " Bez vyzvání se posadil na plastové
train	radost a pomoci v nouzi . " Tu dalekou cestu	sis	mohl ušetřit , " odpověděl mistr , " a stejně
train	aby si odřihla . " Nevadí . Ale měla by	sis	uspořádat život . Uvědomuješ si , že jsi vdaná ?
train	člověče , svého života a dbej na to , aby	sis	byl vědom jeho ceny . " Spravedlnost Dva přátelé celý
train	ruce , a malá Janička se rozplakala . " Tak	sis	měl vzít moji mámu , když jí tak obdivuješ !

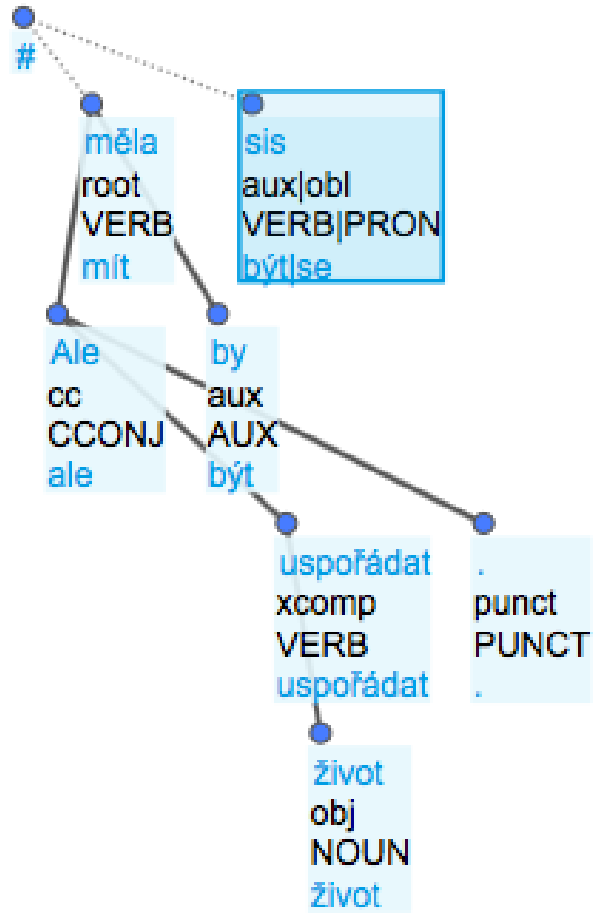
#Ale měla by sis uspořádat život .



#Ale měla by sis uspořádat život .

Hide empty attributes

deprel	aux|obl
deps	2:aux|6:obl
feats	Mood=Ind|Number=Sing|Person=2|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act|Case=Dat||PronType=Prs|Reflex=Yes|Variant=Short
id	4-5
lc	sis
lemma	být|se
p_deprel	root|xcomp
p_feats	Gender=Fem|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act||VerbForm=Inf
p_id	2|6
p_lemma	mít|uspořádat
p_upos	VERB
p_word	měla|uspořádat
p_xpos	VpFS----R-AA---|Vf--------A----
parent	0
upos	VERB|PRON
word	sis
xpos	VB-S---2P-AA---|P7-S3--2-------



The UD format (CoNLL-U, abbreviated, from the Fictree treebank):

ID	WORD	LEMMA	UPOS	FEATS	HEAD	DEPREL
1	Ale	ale	CCONJ	_	2	cc
2	měla	mít	VERB	Gender=Fem|Number=Sing|Polarity=Pos|Tense=Past|...	0	root
3	by	být	AUX	Mood=Cnd|VerbForm=Fin	2	aux
4-5	sis	_	_	_	_	_
4	jsi	být	AUX	Mood=Ind|Number=Sing|Person=2|Polarity=Pos|Tense=Pres|...	2	aux
5	si	se	PRON	Case=Dat|Number=Sing|PronType=Prs|Reflex=Yes|Variant=Short	6	obl
6	uspořádat	uspořádat	VERB	Polarity=Pos|VerbForm=Inf	2	xcomp
7	život	život	NOUN	Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing|Polarity=Pos	6	obj
8	.	.	PUNCT	_	2	punct

The UD format modified for Manatee, only the fused word *sis* shown:

ID	WORD	CFORM	IFORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
4-5	sis	si|s	si|jsi	se|být	PRON|AUX	1:|Case=Dat|Number=Sing|PronType=Prs|...|	6|2	obl|aux
						2:|Mood=Ind|Number=Sing|Person=2|...		

Stay tuned!

https://intercorp.korpus.cz/

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.

Outline

Brøndal, V. (1928).
Ordklasserne. Partes Orationis.
G. E. C. Gad, København.

Chiarcos, C. (2012).
Ontologies of linguistic annotation: Survey and perspectives.
In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 303–310, Istanbul, Turkey. European Language Resources Association (ELRA).

Chiarcos, C. & Erjavec, T. (2011).
OWL/DL formalization of the MULTEXT-East morphosyntactic specifications.
In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 11–20, Portland, Oregon, USA. Association for Computational Linguistics.


Grzegorczykowa, R., Laskowski, R., & Wróbel, H., editors (1998).
Gramatyka współczesnego języka polskiego – Morfologia, volume 1.
Wydawniczwo Naukowe PWN.


Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014).
The Sketch Engine: ten years on.
Lexicography, **1**(1), 7–36.


Komárek, M., Kořenský, J., Petr, J., & Veselková, J., editors (1986).
Mluvnice češtiny 2 – Tvarosloví.
Academia, Praha.


Rosen, A. (2014).
A 3D taxonomy of word classes at work.
In L. Veselovská and M. Janebová, editors, *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*, volume 4 of *Olomouc Modern Language Series*, pages 575–590, Olomouc. Palacký University.


Saloni, Z. & Świdziński, M. (1985).
Składnia współczesnego języka polskiego.
Państwowe Wydawnictwo Naukowe, Warszawa.